



MONTCLAIR STATE
UNIVERSITY

Montclair State University
**Montclair State University Digital
Commons**

Theses, Dissertations and Culminating Projects

1-2021

Mining Social Media and Structured Data in Urban Environmental Management to Develop Smart Cities

Xu Du

Montclair State University

Follow this and additional works at: <https://digitalcommons.montclair.edu/etd>



Part of the [Environmental Sciences Commons](#)

Recommended Citation

Du, Xu, "Mining Social Media and Structured Data in Urban Environmental Management to Develop Smart Cities" (2021). *Theses, Dissertations and Culminating Projects*. 697.

<https://digitalcommons.montclair.edu/etd/697>

This Dissertation is brought to you for free and open access by Montclair State University Digital Commons. It has been accepted for inclusion in Theses, Dissertations and Culminating Projects by an authorized administrator of Montclair State University Digital Commons. For more information, please contact digitalcommons@montclair.edu.

**Mining Social Media and Structured Data in Urban Environmental Management to
Develop Smart Cities**

A DISSERTATION

Submitted to the Faculty of
Montclair State University in partial fulfillment
of the requirements
for the degree of Doctor of Philosophy

by

Xu Du

Montclair State University

Montclair, NJ

January 2021

Dissertation Chair: Dr. Aparna Varde

DISSERTATION APPROVAL

We hereby approve the Dissertation

MINING SOCIAL MEDIA AND STRUCTURED DATA IN URBAN
ENVIRONMENTAL MANAGEMENT TO DEVELOP SMART CITIES

of

Xu Du

Candidate for the Degree:

Doctor of Philosophy

Graduate Program:
Environmental Management

Dissertation Committee:

Certified by:

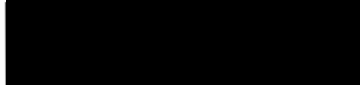

Dr. Aparna Varde
Dissertation Chair


Vice Provost for Research and
Dean of the Graduate School


Dr. Robert Taylor

Date:

January 24, 2021


Dr. Clement Alo


Dr. Vineet Chaoji

Copyright © 2021 by *Xu Du*. All rights reserved.

Abstract

This research presented the deployment of data mining on social media and structured data in urban studies. We analyzed urban relocation, air quality and traffic parameters on multicity data as early work. We applied the data mining techniques of association rules, clustering and classification on urban legislative history. Results showed that data mining could produce meaningful knowledge to support urban management. We treated ordinances (local laws) and the tweets about them as indicators to assess urban policy and public opinion. Hence, we conducted ordinance and tweet mining including sentiment analysis of tweets. This part of the study focused on NYC with a goal of assessing how well it heads towards a smart city. We built domain-specific knowledge bases according to widely accepted smart city characteristics, incorporating commonsense knowledge sources for ordinance-tweet mapping. We developed decision support tools on multiple platforms using the knowledge discovered to guide urban management. Our research is a concrete step in harnessing the power of data mining in urban studies to enhance smart city development.

Keywords: data mining, text mining, ordinances, urban policy, sentiment analysis.

Acknowledgments

I would first like to thank my advisor Dr. Aparna Varde for her support, guidance, patience, and encouragement on this academic journey. I would also like to thank my committee members, Dr. Robert Taylor, Dr. Clement Alo, and Dr. Vineet Chaoji, for their assistance. I would also thank all the coauthors of my papers; their support and cooperation help me a lot during the research. A special thanks to all the Facilities personals who help me from the first day at Montclair Stated University.

I would like to thank the department of EAES and the Graduate School for my doctoral assistantship. I also would like to thank Drs. Brachfeld, Chopping, and Pope for their support as program director and department chairperson. I would like to thank all the professors who work as my teaching assistant supervisor; their trust supports my lab instructor job.

I would like to thank my parents for their support, both financially and emotionally. Without their support, I would never be here. Finally, thank all my friends, classmates, and colleagues for their help during this academic journey.

Table of Contents

Abstract.....	i
Acknowledgments.....	ii
Table of Contents	iii
List of Tables.....	xiv
List of Figures.....	xv
Chapter 1	1
1. Introduction.....	1
1. 1 General Introduction	1
1. 2 Research Objectives.....	5
1. 3 Connect the Ordinances and Tweets	6
1. 3. 1 Ordinance Source and Preprocessing.....	6
1. 3. 2 Tweets Data Source and Preprocessing	7
1. 3. 3 CSK and Domain KBs	9
1. 3. 4 SCC Mapping Approach	10
1. 3. 5 Sentiment Analysis.....	11
1. 4 Decision Support Tools for Urban Management	12
1. 4. 1 Air Quality Prediction Tool.....	12

1. 4. 2 SCC Mapping Tool	13
1. 4. 3 Mobile Application and Web Platform for Dissemination.....	13
1. 5 Broader Impact.....	15
1. 6 Summary	16
1. 7 Organization of Dissertation	17
1. 8 References.....	20
Chapter 2.....	23
2. Early Work	23
2. 1 Mining Multicity Urban Data for Sustainable Population Relocation	23
2. 1. 1 Introduction.....	24
2. 1. 2 Problem Definition.....	28
2. 1. 3 Proposed Solution	29
2. 1. 4 Urban Big Data	31
2. 1. 5 Non-Linear Relationship Analysis	37
2. 1. 6 Experimentation.....	39
2. 1. 7 Related Work.....	45
2. 1. 8 Conclusions and Ongoing Work	47
2. 1. 9 Acknowledgement.....	48
2. 1. 10 Reference	49

2. 2 Mining PM2.5 and Traffic Conditions for Air Quality	52
2. 2. 1 Introduction.....	53
2. 2. 2 Problem Definition.....	56
2. 2. 3 Proposed Solution	58
2. 2. 4 Experimentation.....	61
2. 2. 5 Related Work.....	68
2. 2. 6 Conclusions and Future Work.....	71
2. 2. 7 Acknowledgment	72
2. 2. 8 Reference	73
Chapter 3.....	76
3. Ordinance Mining	76
3. 1 Urban Legislation Assessment by Data Analytics with Smart City Characteristics.....	76
3. 1. 1 Introduction.....	77
3. 1. 2 Data Description	78
3. 1. 3 Analytical Methods	81
3. 1. 3. 1 Data Warehousing	81
3. 1. 3. 2 XML Data Management	86
3. 1. 3. 3 Data Mining	90

3. 1. 4 Results and Discussion	92
3. 1. 5 Related Work.....	96
3. 1. 6 Conclusion	98
3. 1. 7 Acknowledgments.....	99
3. 1. 8 References.....	99
3. 2. Mining Ordinance Data from the Web for Smart City Development	102
3. 2. 1 Introduction.....	102
3. 2. 2 Approach for Ordinance Mining	106
3. 2. 2. 1 Overview of Approach.....	106
3. 2. 2. 2 Harnessing Common Sense Knowledge.....	107
3. 2. 2. 3 Data Processing and Smart City Mapping.....	110
3. 2. 2. 4 Deployment of Mining Methods.....	111
3. 2. 3 Experimental Results	112
3. 2. 3. 1 Statistical Analysis with Clustering.....	113
3. 2. 3. 2 Association Rule Mining.....	116
3. 2. 3. 3 Decision Tree Classification	119
3. 2. 3. 4 Summary of Observations.....	120
3. 2. 4 Related Work.....	122

3. 2. 5 Conclusion	124
3. 2. 6 References.....	127
Chapter 4.....	130
4. Social Media Text Mining.....	130
4. 1 Air Quality Assessment from Social Media and Structured Data....	130
4. 1. 1 Introduction.....	131
4. 1. 2 Mining Structured Data on Pollutants.....	132
4. 1. 2. 1 Background and Goals.....	132
4. 1. 2. 2 Data and Standards	133
4. 1. 2. 3 Approach and Experiments	134
4. 1. 3 Opinion Mining on Pollution from Social Media	138
4. 1. 3. 1 Motivation and Problem Definition	138
4. 1. 3. 2 Proposed Methodology	138
4. 1. 3. 3 Experiments and Observations	144
4. 1. 3 Predictive Analysis and Discussion	147
4. 1. 4 Related Work.....	149
4. 1. 5 Conclusions.....	150
4. 1. 6 Acknowledgment	152
4. 1. 7 References.....	152

4. 2 Mapping Ordinances and Tweets Using Smart City Characteristics to Aid Opinion Mining.....	155
4. 2. 1 Introduction.....	156
4. 2. 2 Related Work.....	159
4. 2. 3 Proposed Mapping Approach.....	162
4. 2. 3. 1 CSK — SCC based KB Development.....	163
4. 2. 3. 2 Linking using SCCs and CSK.....	166
4. 2. 4 Evaluation of The Mapping	168
4. 2. 4. 1 Ordinance to SCC Mapping.....	168
4. 2. 4. 2 Tweet to SCC Mapping.....	172
4. 2. 4. 3 Assessment and Discussion.....	174
4. 2. 5 Conclusion	179
4. 2. 6 Acknowledgments.....	180
4. 2. 7 References.....	181
4. 3 Smart Governance through Opinion Mining of Public Reactions on Ordinances	188
4. 3. 1 Introduction.....	189
4. 3. 2 Related Work.....	193
4. 3. 3 Approach for Mapping.....	195

4. 3. 3. 1 SCC Based Mapping Process.....	195
4. 3. 3. 2 Role of Commonsense Knowledge.....	197
4. 3. 3. 3 Mapping with Single or Multiple SCCs	199
4. 3. 4 Sentiment Analysis of Tweets	201
4. 3. 4. 1 Process of Sentiment Analysis.....	201
4. 3. 4. 2 Opinion Mining using CSK	201
4. 3. 4. 3 Algorithm for Polarity Classification.....	202
4. 3. 5 Experimental Evaluation.....	203
4. 3. 5. 1 Ordinance to SCC Mapping.....	203
4. 3. 5. 2 Tweet to SCC Mapping.....	205
4. 3. 5. 3 SCC Mapping Assessment.....	207
4. 3. 5. 4 Ordinance to Tweet Mapping Output.....	209
4. 3. 5. 5 Results of Sentiment Analysis on Tweets	210
4. 3. 6 Discussion and Challenges	213
4. 3. 7 Conclusions.....	215
4. 3. 8 Acknowledgments.....	216
4. 3. 9 References.....	216
Chapter 5.....	221
5. Result Dissemination and Application.....	221

5. 1 LSOMP: Large Scale Ordinance Mining Portal	221
5. 1. 1 Problem Definition.....	221
5. 1. 2 Proposed Solution	222
5. 1. 3 Demo and Experiments.....	224
5. 1. 4 Scalable Extension: Historical Analysis	230
5. 1. 5 Related Work.....	232
5. 1. 6 Discussion and Roadmap.....	233
5. 1. 7 References.....	233
5. 2 An Ordinance-Tweet Mining App to Disseminate Urban Policy	
Knowledge for Smart Governance.....	235
5. 2. 1 Introduction.....	236
5. 2. 2 Overview of Ordinance and Tweet Mining.....	238
5. 2. 3 Approach for App Design	242
5. 2. 4 Implementation of the App	244
5. 2. 5 Experiments and Discussion.....	247
5. 2. 6 Related Work.....	251
5. 2. 7 Conclusions.....	253
5. 2. 8 References.....	254
5. 3 Sentiment Analysis of Twitter Data with Hybrid Learning for	

Recommender Applications	259
5. 3. 1 Introduction.....	260
5. 3. 2 Related Work.....	262
5. 3. 3 Sentiment Analysis Models and Methods.....	264
5. 3. 3. 1 Document Level Model	264
5. 3. 3. 2 Sentence Level Model.....	265
5. 3. 3. 3 Supervised Learning Method for Analysis	266
5. 3. 3. 4 Unsupervised Learning Method for Analysis	267
5. 3. 4 Proposed Approach: Hybrid Learning	267
5. 3. 4. 1 Overview of Approach.....	268
5. 3. 4. 2 Steps of Sentiment Analyzer.....	269
5. 3. 5 Implementation of Sentiment Analysis Approach	271
5. 3. 6 Experimental Evaluation.....	275
5. 3. 6. 1 Data on iPhone 6.....	276
5. 3. 6. 2 Data on Peatland Fires	278
5. 3. 6. 3 Data on NYC Ordinances	279
5. 3. 7 Recommender Applications.....	281
5. 3. 7. 1 Product Reviews	281
5. 3. 7. 2 Political Elections	282

5. 3. 7. 3 Search Engine Optimization	283
5. 3. 7. 4 Stock Market.....	283
5. 3. 7. 5 Urban Policy	284
5. 3. 8 Conclusions.....	285
5. 3. 9 References.....	286
Chapter 6.....	289
6. Related Work.....	289
6. 1 Public Opinion Matters: Mining Social Media Text for Environmental Management.....	289
6. 1. 1 Introduction.....	290
6. 1. 2 Environmental Applications.....	291
6. 1. 2. 1 Climate Change and Global Warming	291
6. 1. 2. 2 Urban Policy and Local Laws.....	295
6. 1. 2. 3 Traffic and Mobility Issues	300
6. 1. 2. 4 Energy and Resource Conservation	305
6. 1. 2. 5 Disaster and Resilience	309
6. 1. 3 Discussion on Open Issues.....	312
6. 1. 4 Conclusion	316
6. 1. 5 References.....	317

Chapter 7.....	321
7. Conclusions and Future Work.....	321
7. 1 Conclusions.....	321
7. 2 Future Work	326
7. 3 References.....	331
Publications from Dissertation.....	333
References List.....	336

List of Tables

Table 1.1: Total Ordinance Enactment of Each Session and Yearly Average	16
Table 3.1: Raw Data Extracted from Websites	80
Table 3.2: Additional Ordinances Data	80
Table 3.3: Smart City Characteristics	81
Table 3.4: Apriori Data Attributes.....	90
Table 3.5: Processed Data on Ordinances.....	111
Table 3.6: Ordinance Distribution W.R.T. Smart City	120
Table 4.1: AQI Standards for Health based on Pm2.5	134
Table 4.2: Partial Snapshot of Clustering	136
Table 4.3: Potential List of Domains (Partial Snapshot).....	140
Table 4.4: Curated List of Relevant Domains for KB Slicing	140
Table 4.5: A Sample Ordinance and its SCC Mapping.....	170
Table 4.6: Accuracy of Ordinance and Tweet Mapping.....	177
Table 4.7: Mapping of a Sample Ordinance to SCC(s)	205
Table 4.8: Public Contentment for Each SCC	212
Table 5.1 Mapping Accuracy of Ordinances and Tweets.....	225

List of Figures

Figure 1.1: Role of Social Media and Ordinance	2
Figure 1.2: Six Dimensions and Related Working Areas of Smart City	4
Figure 1.3: Ordinance Website Information	7
Figure 1.4: Raw Tweets Json File Screenshot.....	8
Figure 1.5: The SCC Mapping Flowchart.....	11
Figure 1.6: Interactive FAQ example.....	15
Figure 2.1: Proposed Solution for Analysis	30
Figure 2.2: From the data table to the spatial map.....	35
Figure 2.3: Examples of Association Rules	41
Figure 2.4: Clustering Result Example.....	42
Figure 2.5: Visualization of Clusters.....	43
Figure 2.6: Visualization of Decision Tree	44
Figure 2.7: PM2.5 penetration in lungs and harmful effects of the pollutant	54
Figure 2.8: AQI values for PM2.5 as per health standards	55
Figure 2.9: Proposed Approach for PM2.5 Analysis	59
Figure 2.10: Raw Data on PM2.5 from Worldwide Sources	60
Figure 2.11: Sample Output of Clustering	62

Figure 2.12: Example of predicted output with safe PM2.5 range	65
Figure 2.13: Example of moderate PM2.5 range as predicted ouput.....	66
Figure 2.14: Example of moderate to potentially unsafe PM2.5 range prediction...	66
Figure 3.1: The Data Source Page	79
Figure 3.2: Star schema of the ordinances database	83
Figure 3.3: Star schema of the committee database.....	84
Figure 3.4: Star schema of the meeting database.....	85
Figure 3.5: Partial snapshot of fact constellation.....	86
Figure 3.6: XML DB structure for ordinances.....	87
Figure 3.7: XML DB structure for committees.....	88
Figure 3.8: XML DB structure for meetings.....	89
Figure 3.9: Ordinance data after filtering	91
Figure 3.10: Ordinance distribution in sessions as per smart city characteristics	94
Figure 3.11: Typical smart city characteristics	103
Figure 3.12: Smart city example – Amsterdam	105
Figure 3.13: Illustration of ordinance mining approach	107
Figure 3.14: Snapshot of WebChild browser	108
Figure 3.15: Relevant domains selected in KB.....	109
Figure 3.16: Concepts entered in domains.....	109

Figure 3.17: Example of NYC ordinance data.....	113
Figure 3.18: Statistical plot of enacted and initialized ordinances	114
Figure 3.19: Visualization of ordinances clustered from 2006-2009.....	114
Figure 3.20: Visualization of ordinances clustered from 2010-2013.....	115
Figure 4.1: Example of populating domain specific KB	141
Figure 4.2: Code snippet of functions for cleaning tweets	146
Figure 4.3: Example of visualizing opinion mining results.....	147
Figure 4.4: Evaluation example with good PM2.5 range	148
Figure 4.5: Evaluation example with moderate PM2.5 range	148
Figure 4.6: Smart City Characteristics – Highlights.....	157
Figure 4.7: Proposed approach for SCC mapping	163
Figure 4.8: Relevant partial screenshot of WebChild	164
Figure 4.9: Part of Domain KBs with SCC (Subset of Smart Environment and Smart Mobility terms)	166
Figure 4.10: Sample of NYC Council website	169
Figure 4.11: Summary plot of ordinance SCC mapping.....	171
Figure 4.12: Subset of tweets analyzed from NYC sites	173
Figure 4.13: Partial snapshot of tweet to SCC mapping.....	174
Figure 4.14: Example of SCC mapping identified.....	175

Figure 4.15: Example of no matches for any SCC	175
Figure 4.16: Smart City Characteristics [3]	190
Figure 4.17: Illustration of the Mapping Process.....	196
Figure 4.18: Partial Screenshot of WebChild.....	198
Figure 4.19: Sample SCC Domains	199
Figure 4.20: NYC Ordinance Website	204
Figure 4.21: Screenshot of GUI for Tweet to SCC Mapping.....	207
Figure 4.22: Summary of Mapping Assessment.....	208
Figure 4.23: Polarity Classification of Tweets on all SCCs.....	211
Figure 5.1 Widely accepted SCCs: Smart City Characteristics (adapted from [1])	223
Figure 5.2 System architecture of the Web portal: LSOMP	224
Figure 5.3 Portal depiction of mapping ordinances to SCCs.....	225
Figure 5.4 Real-time graphics: tweet to single SCC mapping.....	227
Figure 5.5 Real-time graphics: tweet to multiple SCCs with equal mapping.....	228
Figure 5.6 Real-time graphics: tweet to multiple SCCs with unequal mapping.....	229
Figure 5.7 Example query in Web QA: pertaining to public satisfaction	230
Figure 5.8 Extension: Historical data analysis of ordinances and tweets	232
Figure 5.9: NYC (New York City) as a prominent smart city [left] and NYC council website [right]	238

Figure 5.10: App navigation flowchart	244
Figure 5.11: Implementation process of the app using the Android platform	246
Figure 5.12: Layout screens of the Ordinance-Tweet Mining App	246
Figure 5.13: FAQs for various selection categories in the app	247
Figure 5.14: Responses to Q1: “Do you find the app quick and easy to use?”	248
Figure 5.15: Responses to Q2: “Does this work increase public awareness of urban policy?”	248
Figure 5.16: Responses to Q3: “Do you feel NYC is getting better as a smart city?	249
Figure 5.17: Example of document level model.....	265
Figure 5.18: Example of sentence level model.....	265
Figure 5.19: Example of training set in supervised learning for sentiment analysis	266
Figure 5.20: Functions to clean the tweets.....	272
Figure 5.21: Feature List Example.....	273
Figure 5.22: Partial snapshot of a sentiword table.....	274
Figure 5.23: Example of output file.....	275
Figure 5.24: Graph plotted from sentiment analysis for iPhone 6 product review.	277
Figure 5.25: Area Chart for Sentiment Analysis of IPF Impact.....	279

Figure 5.26: Pie Chart on Public Reactions to NYC Ordinances	280
Figure 6.1: Topics about pollution learned from a probabilistic topic model [Wang et al. 2015]	294
Figure 6.2: NYC Council ordinance website (left) and approach for ordinance– tweet mapping (right) [Puri et al. 2018].....	296
Figure 6.3: Workflow of Twitter data acquisition, processing and analysis for traffic problems [Gu et al. 2016]	301
Figure 6.4: Interactive map of New York State with display of sprawl affected regions generated from GIS data [Pampoore-Thampi et al. 2014].....	303
Figure 6.5: Summary of sentiment analysis over social media posts about the Case Carbon Capture and Storage system [Nuortimo 2018].....	307
Figure 6.6: Workflow for a disaster event database [Wang et al. 2018]	310
Figure 7.1: Input Interface Concept	328
Figure 7.2: Output Interface Concept	328
Figure 7.3: Proposed Enhancement	330

Chapter 1

1. Introduction

1.1 General Introduction

This dissertation focuses on applying social media text mining on municipal level policy making to support Smart City development. The concept of a Smart City supports current urban areas for competitive and sustainable development by addressing the environmental influence and increasing management efficiency. Social media text mining provides a unique view to examine the interaction between urban policy and resident opinions. The social media posts of the common public contain information about the users' reactions to their daily lives, highly influenced by urban policy. Text mining and sentiment analysis can extract information from raw data. This could provide useful insight into the effectiveness of the policies.

There are multiple social media platforms. For the purpose of this dissertation, Twitter, which is one of the most popular social media platforms with each tweet being limited to 280 characters, is considered the most suitable for efficient and effective analysis. Tweets published by urban residents are suitable for text mining since they are limited in text size. Historical tweets are collected by researchers as 1% of the total Twitter stream data (Scott, 2012), called the "Spritzer" version. Those historical tweets that contain geoinformation and other location

indicators ensure the long term and extensive coverage of the local public opinions. For local policy, we choose the ordinances as the research target; municipal level legislatures initial and enact the ordinances, making the ordinances highly related to the local public opinions. The information of most ordinances is publicly available and free to acquire. In our research, we get all the New York City (NYC) ordinances data from the NYC city council website ("The New York City Council - Legislation", 2020). We selected NYC because it is one of the most populated urban areas in the United States, it would bring enormous benefit if our research could increase the effectiveness of NYC's urban management. Figure 1.1 shows how public opinion and urban policy interact while social media and ordinances serve as suitable indicators of public opinion and urban policy.

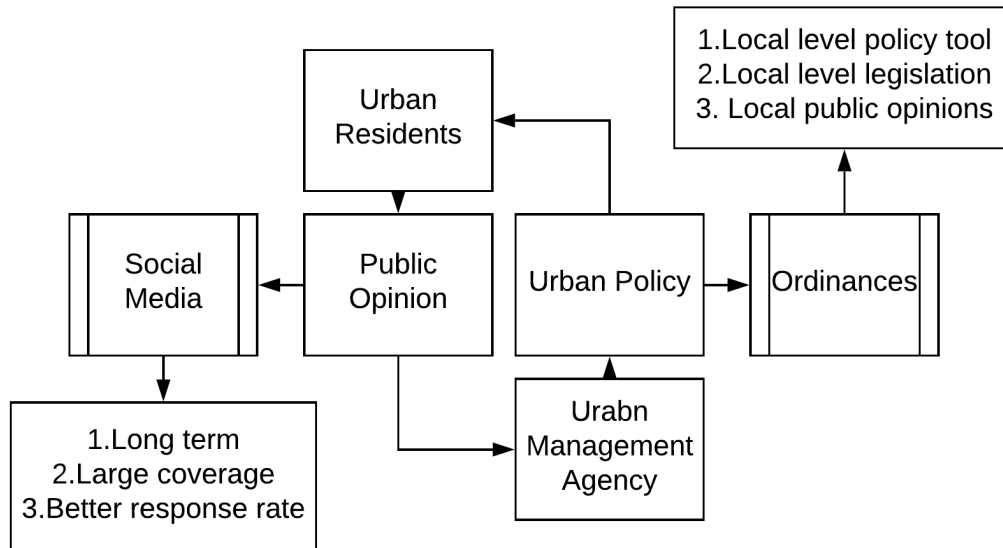


Figure 1.1: Role of Social Media and Ordinance

One of the main challenges when analyzing the interaction between urban ordinances and tweets is the automation of the processes. The number of tweets is vast; extracting the related tweets and analyzing their sentiment would take a seemingly infinite amount of time if processed by humans. On the other hand, computers are very efficient in processing large amounts of data due to hardware and software advancements. However, a computer cannot comprehend and connect the tweets and ordinances without proper programming. This research utilizes commonsense knowledge (CSK) to build the domain knowledge bases (KB), allowing the computer program to identify related tweets containing similar information via the domain KB. CSK could be treated as the bridge between human comprehension and computer program logic or in other words, a tool to expand the human knowledge so the domain KBs could have enough coverage to support the program for identification.

The mapping between ordinances and tweets also presents the challenge of reducing dimensions of connections. Every ordinance should have its domain KB for the most precise extraction of related tweets in the ideal condition. However, this is difficult to achieve since building domain KBs requires direct inputs at the first step. At this point, a CSK-based tool can expand the input to form the final applicable KB. This would be overly time consuming since we have hundreds of ordinances. The same approach is even harder for the tweet to ordinance mapping due to the vast number of tweets. This research provides an approach to overcome this problem: Build a medium of domain KBs between the ordinances and tweets, then categorize

them under the same conditions. This approach would reduce the computing and time consumption of the program. This research is conducted to support Smart City development. It would thus be suitable if the domain KBs also indicate the development of Smart City. This research utilizes a widely accepted system in the literature which divides the Smart City development into six dimensions (Giffinger & Pichler-Milanović, 2007). They are Smart Governance, Smart Economy, Smart Mobility, Smart Environment, Smart People, and Smart Living. Figure 1.2 (Cohen, 2020) shows the six dimensions and related working areas. This Smart City concept system enhances the connection between the research outcome and Smart City development. It is a comprehensive system that considers environmental, economic, and social aspects, which has the potential to form political objectives.

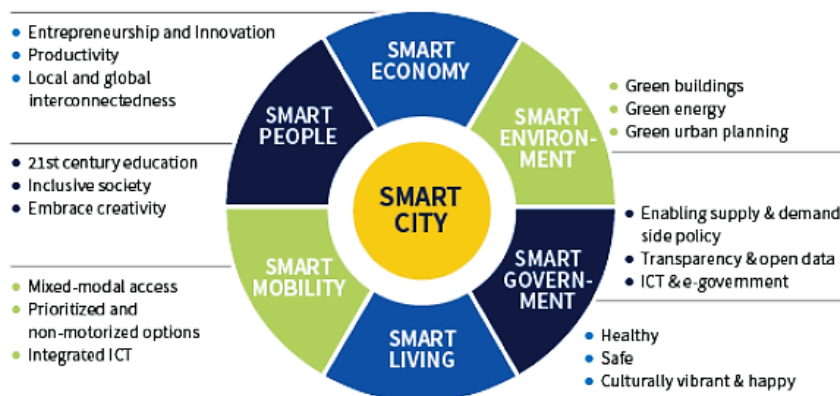


Figure 1.2: Six Dimensions and Related Working Areas of Smart City

Other structural data analysis could also support the Smart City development, e.g., the population data, the urban sprawl parameters, air quality data, and traffic conditions. Data mining

there could also provide fruitful results. Accordingly, this research also deploys multiple data mining methods on urban parameters to access the driving force of urban sprawl. Association rule mining, clustering analysis, and decision tree learning all produce interesting conclusions. The data mining results of air quality data and traffic conditions support the development of an air quality prediction tool based on decision tree learning. This is an important outcome of this dissertation.

1. 2 Research Objectives

This research proposes to support Smart City development by analyzing tweets and ordinances. Text mining and CSK support the mapping via domain KBs guided by Smart City Characteristics (SCCs). The approach can be divided into two primary objectives;

1. Connect the ordinances to the tweets
2. Develop decision support tools for urban management

We propose to develop an approach to connect the tweets and ordinances, which could support efficient analysis of urban policy. We propose to develop tools that support the decision making of urban management. We gather all data from free public accessible sources, such as the NYC city council website and Twitter website.

1.3 Connect the Ordinances and Tweets

Ordinances and tweets contain information that humans can understand but this is not easily comprehensible by computers. However, the large amount of tweets makes it infeasible for the analysis to be conducted directly by humans. A proper approach design that combines human comprehension and computer computing power is a good choice based on current technology.

1.3.1 Ordinance Source and Preprocessing

The ordinance source is the New York City council website ("The New York City Council - Legislation", 2020); it provides the function to download CSV format ordinance data, which contains the information about each ordinance. The file does not contain the initial and enactment dates, which are only available on the website. We either manually input them or use a web crawler to collect the data. Figure 1.3 shows how to get the ordinance data downloaded and where to get the initial and enactment dates.

The screenshot shows the website for The New York City Council, headed by Corey Johnson, Speaker. The navigation menu includes Council Home, Legislation, Calendar, City Council, and Committees. A search bar is present with filters for Session (2018 to 2021) and Local Law. Search options are checked for file #, text, attachments, and other info. A table of 467 records is displayed, with the first two records visible:

File #	Law Number	Committee	Prime Sponsor	Council Member Sponsors	Title
Int 0001-2018	2018/084	Committee on Finance	Daniel Dromm	1	A Local Law the preliminary date of su
Int 0600-2018	2018/085	Committee on Housing	Corey D.	11	A Local Law

Below the table, details for Int 0001-2018 are shown:

File #:	Int 0001-2018	Version: *	Name:	Budget Extender
Type:	Introduction		Status:	Enacted
On agenda:	1/16/2018		Committee:	Committee on Finance
Enactment date:	2/16/2018	Law number:	2018/084	

Figure 1.3: Ordinance Website Information

The title section contains information on what the given ordinance indicates. By utilizing the domain experts' manual input along with the domain KB, we can decide which SCC aspect is the primary concern of this ordinance. Because the number of ordinances is not overwhelmingly huge, this process can be conducted either by human experts or computer programs guided by well-designed domain KBs. Data mining that only analyzes the temporal distribution of different SCC categorized ordinances could provide interesting knowledge.

1.3.2 Tweets Data Source and Preprocessing

Twitter, as the most popular social media website, provides a free search application

programming interface (API) to acquire tweets of the past seven days, which is too short for this research. The historical tweets are collected via a third party collector called Archive Team (Scott, 2012). They provide free historical tweets from 2011 to June-2020 with only one disadvantage: This collection is named the Spritzer version and only contains one percent of the whole Twitter stream. It could influence the accuracy of the result because it is not the complete data; however, this minor negative impact is acceptable; the correlation coefficient r (linear) of the Tweets' percentage on the same topic between two datasets (Spritzer and the whole stream) is around 0.94 for most topics (Leetaru, 2019)

The raw tweets data, except those cut-and-pasted from the Twitter website, is stored in JSON format. This format can be handled by multiple programming languages and can have excellent compatibility. The raw tweets are very hard to read by humans since they are in a very compact format. Figure 1.4 shows what this looks like in a text editor.

```
{\"created_at\":\"Wed Jan 01 07:00:00 +0000 2020\",\"id\":12122672112654581
a\\u003e\",\"truncated\":false,\"in_reply_to_status_id\":null,\"in_reply_to_
PS02\\n\\u305f\\u307e\\u306b\\u7d75\\u3092\\u63cf\\u304f\",\"translator_type\":
2017\",\"utc_offset\":null,\"time_zone\":null,\"geo_enabled\":false,\"lang\":n
profile_images\\/1206239768876478464\\/oHRC29Nk_normal.jpg\",\"profile_im
1570544770\",\"default_profile\":true,\"default_profile_image\":false,\"fol
{\"created_at\":\"Wed Jan 01 07:00:00 +0000 2020\",\"id\":12122672112612802
a\\u003e\",\"truncated\":false,\"in_reply_to_status_id\":null,\"in_reply_to_
Fujimori\\u306ebeatmania\\u66f2\\u3092\\u545f\\u304fbot\\u3067\\u3059\\u3002\\
2010\",\"utc_offset\":null,\"time_zone\":null,\"geo_enabled\":false,\"lang\":n
bg.png\",\"profile_background_tile\":false,\"profile_link_color\":\"1DA1F2\"
img_pro_synthesized_normal.jpg\",\"default_profile\":true,\"default_profi
{\"created_at\":\"Wed Jan 01 07:00:00 +0000 2020\",\"id\":12122672112906240
a\\u003e\",\"truncated\":false,\"in_reply_to_status_id\":null,\"in_reply_to_
e1\\u53f7\\u3002\\u901a\\u5e33\\u306f\\u307f\\u306a\\u3044\\u3088\\u3046\\u306b\\
```

Figure 1.4: Raw Tweets Json File Screenshot

The raw tweets contain much interesting information, such as the created date, user location, and text. This research uses a python program to preprocess this raw data to be convenient for further analysis. However, only a limited number of tweets have spatial information and user location. This problem could be improved if we had access to complete Twitter data with more information or other location identification methods.

1.3.3 CSK and Domain KBs

The domain KBs are the core component of SCC mapping between ordinances and tweets. They are built by researchers with domain knowledge and other existing domain information. Since this research is about Smart City development, we consider the existing well-defined Smart City development dimensions as the six major domains. They are also called SCCs in this research. There are many indicators for those six domains (Giffinger & Pichler-Milanović, 2007). We transfer them into the domain KBs with adjustments. The domain expert also modify the domain KB with a Topic Model analysis of the ordinance data. The Topic Model analysis counts the word frequency in ordinance text (by Python program) and identifies useful topic keywords.

We need to utilize CSK to expand our domain KBs prototype. Those prototype domain KBs only contain the words we selected. The CSK expands the coverage of those words to cover the most words in the text of the real ordinances and tweets. The CSK source we primarily use is

WebChild, which contains commonsense knowledge automatically extracted from Web contents ("Max-Planck-Institut für Informatik: WebChild", 2018). In addition, we also use WordNet as a complementary source for semantic matching. The final domain KBs contain text-based terms, which are relevant to six SCCs.

1.3.4 SCC Mapping Approach

Figure 1.5 shows the whole approach of SCC mapping. The SCC mapping's core function is the SCC identification by the six domain KBs, each about one of the six aspects of Smart City development. We currently use a computer program to count the number of words related to the domain KBs in the text of ordinances or tweets. We assign SCC scores based on the counts of terms that match the domain KBs. If the same term appears in multiple SCC domain KBs, the count is incremented by one for each related SCC. For each ordinance, this approach connects the tweets with similar SCC scores. We will discuss the details of the mapping algorithm in Chapter 4. Our research implements this method, which maintains a broad connection based on SCCs because the primary research purpose is supporting the Smart City development. Further work can address how to build a method to assign weighting factors to different terms in the SCC domain KBs.

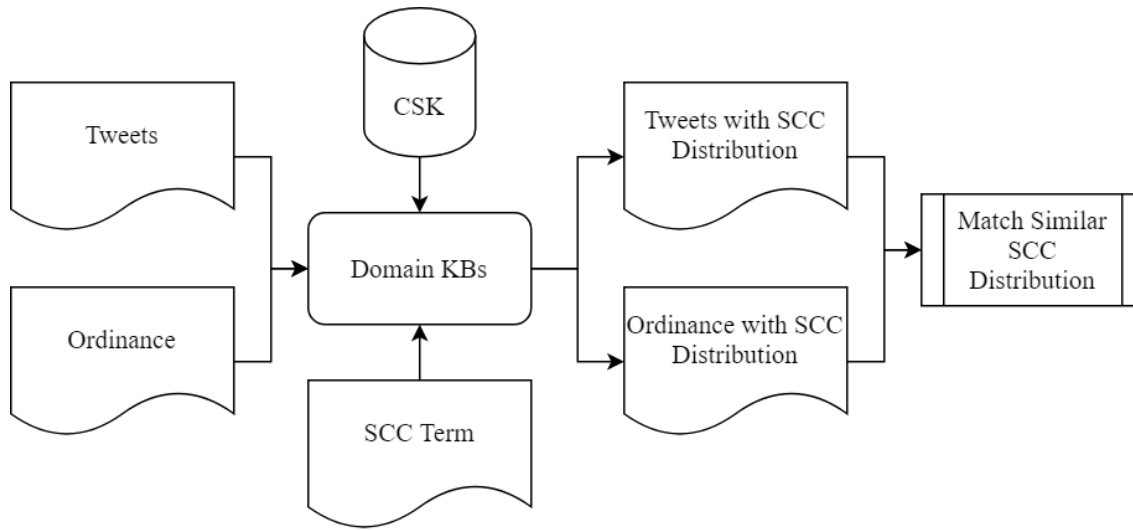


Figure 1.5: The SCC Mapping Flowchart

1.3.5 Sentiment Analysis

To further access the relationship between ordinances and tweets, we need to conduct tweet sentiment analysis. Since the ordinances are objective while the tweets can either be objective or subjective, the tweets go through the same preprocessing based on the SCC mapping. The computer program handles the filtered tweets' text with some existing libraries. For Python, there are two primary libraries, Natural Language Tool Kit (NLTK) and Pattern. We will discuss the details in Chapter 4. The method we choose is polarity classification, which assigns the whole text a score range of $[-1,+1]$. The "-1" means the most negative and "+1" means the most positive ("0" =neutral). The principal of the scoring system is based on our SCC mapping domain KBs. There are two domain KBs; one is the positive KB; the other is the negative KB. Once the

collected tweets go through the polarity classification process, we analyze how the sentiment distribution in those tweets.

1. 4 Decision Support Tools for Urban Management

The ultimate goal of this research is to support Smart City Development. Mining social media and structural data provide useful knowledge, which leads to the development of decision support tools. We transfer that knowledge into decision support tools, which could help related fields such as urban management and legislature. We will explain the detail of those tools in the forthcoming chapters of this dissertation.

1. 4. 1 Air Quality Prediction Tool

This tool is designed based on data mining of urban traffic condition data and air quality data. We select Particulate Matter (PM2.5) as the indicator of air quality. The traffic data selection is based on World Bank data. We will discuss the details in Chapter 2. The decision support technique of this tool is decision tree learning. We have built a graphical user interface (GUI) to provide convenient access even for novice users. This tool's objective is to support multiple level users' comprehension of the relationship between air quality and traffic conditions. The decision tree learning provides a transparent explanation of the air quality determination. We have also built a questionnaire to determine the users' knowledge level and provide different user

support. The demo tool, which only utilizes limited sample data, would achieve an accuracy of around 83%. We could improve this tool by using sufficient data resolution and introduce other air quality parameters. This tool can bring a broader impact on urban management, such as more efficient traffic design and air quality control.

1. 4. 2 SCC Mapping Tool

We utilize the domain KBs and CSK knowledge to categorize the tweets and ordinances and build connections between them. One of the key components of this process is determining the SCC attributes of the target ordinance or tweet. We design a GUI tool, which allows the user to input the ordinance or tweet text to determine its SCC relativeness. The current version of our tool could identify multiple SCCs from the input text. This tool has the potential to help other related studies. Based on expert evaluation, the accuracy is around 80%. We would aim to improve the function and accuracy in future research. The technique details will be discussed in Chapter 4.

1. 4. 3 Mobile Application and Web Platform for Dissemination

The previous SCC mapping tool is only available for computer users. It would bring more benefits if we can reach users from other platforms. Thus, a mobile application platform version or mobile app could create a significant impact due to the extensive coverage and convenient

access to smartphones. Hence, we have built an Android App, which (to the best of our knowledge) is the first app disseminating ordinance-tweet text mining with the consideration of Human-Computer Interaction(HCI), e.g., the conceptualization of actions, fast and accurate navigation, and ubiquitous access. This app would provide users of different knowledge levels (e.g., novice, intermediate, expert) convenient legislative information access. This app's broader impact contributes to the Smart City development, especially Smart Governance, by making urban policy information more transparent and comprehensible. The current version of our app only provides an analysis of the NYC area. We can expand the coverage with quick modification. We will discuss the technical details in section 5.1.

We have developed a Web portal to provide user experiences similar to the mobile app. In our prototype, we have successfully integrated the SCC mapping function and ordinance results. We are working on designing an interactive frequently asked questions (FAQ) system, which allows the user to enter questions and get the most relevant answers, as Figure 1.6 shows. This FAQ system is based on the Natural language processing (NLP) technique. This portal would be further enhanced to accommodate more developments based on future work.

Frequently Asked Questions

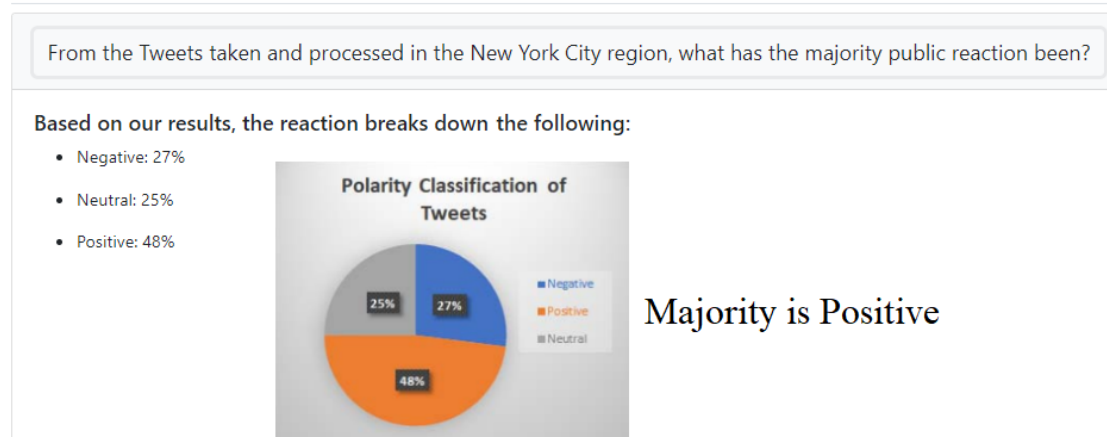


Figure 1.6: Interactive FAQ example

1.5 Broader Impact

The ordinance is the local policy tool that serves a significant role in urban management. The number of ordinances keeps increasing with each session ("The New York City Council - Legislation", 2020); as Table 1.1 shows, the number of ordinances has increased drastically and contributes more to urban management in the most recent three complete sessions. Social media posts are also increasing year by year. The daily tweets total to almost 500 million per day (Sayce, 2020). There is a demand to utilize this valuable big data to discover useful knowledge supporting urban management. This research contributes to the assessment between ordinances and tweets, which would support the Smart City development from several aspects. The knowledge discovered by this research would help urban management agencies identify how

their policies affect each Smart City development dimension. This research is trans-disciplinary between Environmental management and Computer Science. It could foster future research in the related fields.

Table 1.1: Total Ordinance Enactment of Each Session and Yearly Average

Session	2006-2009	2010-2013	2014-2017
Total	287	365	708
Average	71.75	91.25	177

The various decision support tools will help the urban management agencies by providing novel and useful knowledge discovered by data mining and other techniques. Our research has the potential to form a comprehensive decision support system with the power of data mining, CSK, and other techniques. This proposed system will be cross-platform capable of handling multiple social media data sources and other structural data. In the future, urban management will involve more data mining techniques to improve their efficiency; our work is a concrete step to achieve this goal.

1.6 Summary

This dissertation utilizes data mining, CSK, sentiment analysis, and other techniques to analyze the relationship between urban policy and social media. It is a firm first step for the full evaluation of ordinance efficiency via social media mining. The decision support tools reveal the

potential power of the knowledge discovered via our work. Future research can subsequently improve these tools with enhanced data. This research will lead to more fruitful future works to support urban sustainability and Smart City development. Soon, local governments will utilize more data mining techniques for efficient management. This research will be one of the pillars that supports the coming new era of Smart City development..

1. 7 Organization of Dissertation

- The research completes the research objectives. The Chapters consist of papers; they are either accepted or published in Journals or Conference proceedings, except Chapter 1 and Chapter 7. Here is a brief description of the content of each Chapter.
- Chapter 2 is "Early Work" and represents two studies. The first study (Du & Varde, 2015) applies data mining techniques on population relocation data to discover novel knowledge related to urban sprawl. The second study (Du & Varde, 2016) is about data mining on PM2.5 data and traffic conditions to find useful knowledge, which leads to a prediction tool based on decision tree learning.
- Chapter 3 is "Ordinance Mining" and represents two studies. They are both about data mining on ordinance data to discover useful knowledge. The first study (Du et al., 2017) focuses on applying database management techniques to manage the data. The second study's (Du et al., 2017) novel point is applying CSK to categorize the

ordinance for data mining.

- Chapter 4 is "Social Media Text Mining" and represents three studies. The first study (Du et al., 2016) is Chapter 2's second study's follow-up; we conduct tweets sentiment analysis related to peatland fire air pollution. The second study (Puri et al., 2018) focuses on SCC mapping. This study utilizes text mining and CSK to build domain KBs according to the Smart City aspects. We connect the tweets and ordinances based on their relatedness to each SCC as per the domain KBs. The third study (Puri et al., 2018) improves the SCC mapping technique, allowing us to assign multiple SCC types to ordinances and tweets (instead of a single type) while maintaining similar accuracy. In the last two studies, we design and improve the SCC mapping tool.
- Chapter 5 is "Result Dissemination and Application" and represents three studies. The first study (Du et al., 2020) builds a prototype web platform, with the purpose of including the SCC mapping and interactive FAQ functions. The demo website already has the capabilities for SCC mapping. The second study (Varghese et al., 2020) is about integrating the SCC mapping function into a mobile app. This study discusses the HCI guided app design and the benefit of mobile apps for convenient access. The third study (Gandhe et al., 2018) discusses the possible applications of sentiment analysis on various topics that include urban management. This study introduces a hybrid approach for sentiment analysis.

- Chapter 6 is "Related Work" and represents one study. It is a literature review (Du et al., 2020) of text mining studies related to Environmental Management. This study discusses different social media text mining researches on Environmental Management related topics, such as climate change and global warming, urban policy and local laws, traffic and mobility issues, and etc.
- Chapter 7 is about the Conclusion; we also discuss the future works in this chapter.

1. 8 References

- Cohen, B. (2020). *Transportation Systems Management and Operations in Smart Connected Communities - Chapter 1. Transportation Systems Management and Operations (TSMO) in Smart Connected Communities - FHWA Office of Operations*. Ops.fhwa.dot.gov. Retrieved 10 October 2020, from <https://ops.fhwa.dot.gov/publications/fhwahop19004/ch1.htm>.
- Du, X., & Varde, A. (2015). Mining Multicity Urban Data for Sustainable Population Relocation. *International Journal On Computer, Electrical, Automation, Control And Information Engineering*, 9(12), 2441-2448. <https://doi.org/doi.org/10.5281/zenodo.1110816>
- Du, X., & Varde, A. (2016). Mining PM2.5 and traffic conditions for air quality. *2016 7Th International Conference On Information And Communication Systems (ICICS)*. <https://doi.org/10.1109/iacs.2016.7476082>
- Du, X., Emebo, O., Varde, A., Tandon, N., Chowdhury, S., & Weikum, G. (2016). Air quality assessment from social media and structured data: Pollutants and health impacts in urban planning. *2016 IEEE 32Nd International Conference On Data Engineering Workshops (ICDEW)*. <https://doi.org/10.1109/icdew.2016.7495616>
- Du, X., Kowalski, M., & Varde, A. (2020). LSOMP: Large Scale Ordinance Mining Portal. In *IEEE International Conference on Big Data (IEEE BigData 2020)*. Atlanta, GA.
- Du, X., Kowalski, M., Varde, A., de Melo, G., & Taylor, R. (2020). Public opinion matters. *ACM*

-
- SIGWEB Newsletter*, (Autumn), 1-15. <https://doi.org/10.1145/3352683.3352688>
- Du, X., Liporace, D., & Varde, A. (2017). Urban legislation assessment by data analytics with smart city characteristics. *2017 IEEE 8Th Annual Ubiquitous Computing, Electronics And Mobile Communication Conference (UEMCON)*. <https://doi.org/10.1109/uemcon.2017.8248972>
- Du, X., Varde, A., & Taylor, R. (2017). Mining Ordinance Data From the Web for Smart City Development. In *International Conference on Data Mining DMIN* (pp. 84-90). Las Vegas; CSREA press.
- Gandhe, K., Varde, A., & Du, X. (2018). Sentiment Analysis of Twitter Data with Hybrid Learning for Recommender Applications. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 57-63). New York City, NY, USA. <https://doi.org/10.1109/UEMCON.2018.8796661>.
- Giffinger, R., & Pichler-Milanović, N. (2007). *Smart cities*. Centre of Regional Science, Vienna University of Technology.
- Leetaru, K. (2019). *Is Twitter's Spritzer Stream Really A Nearly Perfect 1% Sample Of Its Firehose?*. Forbes. Retrieved 10 October 2020, from <https://www.forbes.com/sites/kalevleetaru/2019/02/27/is-twitthers-spritzer-stream-really-a-nearly-perfect-1-sample-of-its-firehose/#5aa45ef35401>.
- Max-Planck-Institut für Informatik: WebChild*. [Mpi-inf.mpg.de](http://mpi-inf.mpg.de). (2018). Retrieved 10 October 2020, from <https://www.mpi-inf.mpg.de/departments/databases-and-information->

systems/research/yago-naga/commonsense/webchild.

Puri, M., Du, X., Varde, A., & de Melo, G. (2018). Mapping Ordinances and Tweets using Smart City Characteristics to Aid Opinion Mining. *Companion Of The The Web Conference 2018 On The Web Conference 2018 - WWW '18*. <https://doi.org/10.1145/3184558.3191632>

Puri, M., Varde, A., Du, X., & de Melo, G. (2018). Smart Governance Through Opinion Mining of Public Reactions on Ordinances. *2018 IEEE 30Th International Conference On Tools With Artificial Intelligence (ICTAI)*. <https://doi.org/10.1109/ictai.2018.00131>

Sayce, D. (2020). *The Number of tweets per day in 2020 | David Sayce*. David Sayce. Retrieved 18 October 2020, from <https://www.dsayce.com/social-media/tweets-day/#:~:text=Every%20second%2C%20on%20average%2C%20around%206%2C000%20tweets%20are%20tweeted%20on,200%20billion%20tweets%20per%20year>.

Scott, J. (2012). *Archive Team: The Twitter Stream Grab*. Archive.org. Retrieved 10 October 2020, from <https://archive.org/details/twitterstream?tab=about>.

The New York City Council - Legislation. Legistar.council.nyc.gov. (2020). Retrieved 10 October 2020, from <https://legistar.council.nyc.gov/Legislation.aspx>.

Varghese, C., Varde, A., & Du, X. (2020). An Ordinance-Tweet Mining App to Disseminate Urban Policy Knowledge for Smart Governance. *Lecture Notes In Computer Science*, 389-401. https://doi.org/10.1007/978-3-030-45002-1_34

Chapter 2

2. Early Work

2.1 Mining Multicity Urban Data for Sustainable Population Relocation

Abstract: In this research, we propose to conduct diagnostic and predictive analysis about the key factors and consequences of urban population relocation. To achieve this goal, urban simulation models extract the urban development trends as land use change patterns from a variety of data sources. The results are treated as part of urban big data with other information such as population change and economic conditions. Multiple data mining methods are deployed on this data to analyze nonlinear relationships between parameters. The result determines the driving force of population relocation with respect to urban sprawl and urban sustainability and their related parameters. This work sets the stage for developing a comprehensive urban simulation model for catering to specific questions by targeted users. It contributes towards achieving sustainability as a whole.

Keywords: *Data Mining, Environmental Modeling, Sustainability, Urban Planning*

(Chapter 2.1 reused the previously published paper Du, X., & Varde, A. (2015), Mining Multicity Urban Data for Sustainable Population Relocation, *International Journal on Computer, Electrical, Automation, Control And Information Engineering*, 9(12), 2441-2448.

<https://doi.org/doi.org/10.5281/zenodo.1110816>).

2. 1. 1 Introduction

Intrusive unorganized land use change that happens around the boundaries of urban areas is urban decentralization. It is also called as urban sprawl, which leads to the relocation of population, employment, transportation, and land use types. The process and result of urban decentralization causes many negative outcomes, for example, functional open space shortage, farmland loss and habitat fragmentation, traffic congestion and accidents, air pollution and fossil fuel consumption, incline of management costs, and lack of social capital [1]. Population dynamics play a highly significant role in urban decentralization. Low population density is a major phenomenon of urban decentralization policy implementation to reduce urban decentralization like smart growth focuses on supporting high density communities and regulating low density communities [2]. There are various factors which would influence urban population relocation. The traditional theory believes that economic factors, social amenities, health services, traffic, employment, and other variables could drive population changes [3]. The relationship between them is non-linear and changes among different cities. Recently, the population growth in urban areas has been higher compared with rural areas. From 2000 to 2010, the urban population in the U.S. grew 12.1%, while the rural population growth rate was just 0.7% (the total population growth of the U.S. was 9% during that time). This is a significant phenomenon due to the reverse direction of the population decentralization, which is the major

cause of urban sprawl [2]. Identifying the key parameters of this process would be significant for urban sustainability. The relationship between population relocation, urban land use change and other conditions would be the major focus of this research. It would provide valuable information for urban management and planning agencies to promote more compact urban areas.

Data Mining involves the discovery of novel, useful, and interesting patterns and trends from huge volumes of data. It usually involves large data sets and computing. Previous research showed that some methods of data mining can be applied to the calibration of cellular automata transition rules [4]. Data mining is a very broad field, which involves statistics, machine learning, pattern recognition, numeric search, and scientific visualization [5]. Recently, there are many applications of data mining in various research fields due to three major reasons: Firstly, the amount of available data is increasing; secondly, there are more powerful computers; thirdly, there are pertinent advances in statistical and machine learning algorithms [5]. Data mining is suitable for the nonlinear relationship analysis between urban sustainability and urban conditions [6]. To process the data mining for urban sustainability research, a proper data set must be established. The empirical urban databases provide large amounts of data. However, there are no data directly related to urban development trends, usually represented as urban land use change. To replenish this data, the research herewith integrates urban simulation models and data mining. The urban simulation models would extract urban development trends from raw data in the form of indicators, matrix, or rules.

Urban simulation models are the simplified, computed form of the real urban areas. Firstly, the goal of simulation models was to determine transportation capacity needs by predicted land use trends. Then it transferred to policy objectives like reducing the air pollution. Currently, the objectives are predicting and explaining the development trends to support the urban management and planning.

In this research, the major simulation models are the land use change models. There are various land use change models: such as the cellular automata based models, statistical analysis models, Markov chain models, artificial neural network models, economic-based models, and agent-based models. Most of these models have the ability to predict the future change of land use, it also means there can extract the development trend and utilize them (land use change matrix-Markov chain, transition rules- cellular automata, statistical indicators- statistical analysis models and its). These extracted trends can be utilized by the data mining to discover interesting knowledge.

Different cities show different development patterns, and there is no universal pattern of urban growth [7]. Various factors influence urban growth and there is no direct linear connection between the factors and responses [8]. Due to this, urban simulation models need adjustments for each specific city for proper results. Previous research works usually consider ambient variables as weighted indexes. The weighted value is determined by statistical methods in single area simulations. A proper weighted value brings more accurate results. However, when the location

changes, the weight needs to be adjusted. Few urban simulations have involved multiple urban areas [9]. For this, illustrating the non-linear connection between urban conditions and urban development patterns is helpful to build a proper simulation model, which would provide more accurate information for urban environmental management. More importantly, the single urban simulation would not be able to provide enough data for the data mining process, both in terms of the quantity as well as the quality.

Population dynamics have a highly significant role in urban decentralization. Low population density is a major phenomenon in the urban decentralization process [2]. The population ratio between the urban areas and rural areas are an indicator of urban compactness [1]. Policy implementation to reduce the urban decentralization like smart growth focuses on supporting high density communities and regulating low density communities. There are various factors which would influence urban population relocation and existing theories state that economic factors, social amenities, health services, traffic, employment and other variables could drive population change. The relationship between them is non-linear and changes among different cities. To analyze this change, especially the population re-centralization in the United States between 2000-2010, the urban simulation models and data mining methods must be integrated to provide useful information, which enhances the urban sustainability.

2. 1. 2 Problem Definition

Decision makers and management agencies need the information and knowledge about urban population dynamics, urban land use change and urban development to frame proper policies, which could ensure sustainable growth patterns. Urban systems are very complicated and different cities have different development patterns (Schneider & Woodcock, 2008). Thus, the main problem of this research is to build scalable and flexible urban simulation models in multicity environments for conducting predictive and diagnostic analysis on relationships between spatio-temporal changes of population, land use, and urban development to enhance urban sustainability. We define the following sub-problems in this work

1. Propose methods to analyze multicity big data
 - a. Generate complex urban data capturing the required spatial-temporal features and process this big data
 - b. Develop data mining approaches to reveal relationships between urban conditions and urban population relocation
2. Diagnose the key factors causing urban sprawl
 - a. What are the major parameters causing sprawl?
 - b. How do these parameters affect each other?
 - c. How does sprawl itself impact the parameters again?

3. Predict the indicators to enhance sustainability

- a. What are the primary urban sustainability goals?
- b. How do specific changes affect each other?
- c. How exactly do sprawl and sustainability correlate?

4. Set the stage for a comprehensive urban simulation model to answer potential user

questions such as

- a. What is the relationship between parameters causing sprawl?
- b. If size of city is a parameter, how does the model alter based on the size?
- c. What is the quantitative relationship between individual factors affecting sustainability?

(e.g., between population density and number of doctors).

2. 1. 3 Proposed Solution

In this research, the ultimate goal is to enhance urban sustainability in a multicity environment catering to various objectives such as minimizing sprawl, offering valuable information for urban management and planning agencies and improving the environmental management aspect of urban areas. In order to achieve this goal, this research aims to conduct data mining on complex urban data, which is suitable for analysis of nonlinear relationships between urban development activities and urban conditions. It proposes to perform urban simulation modeling, which helps us understand the urban development. The proposed approach

involves urban land use change simulation modeling by data mining in a multicity environment and is depicted in Figure 2.1.

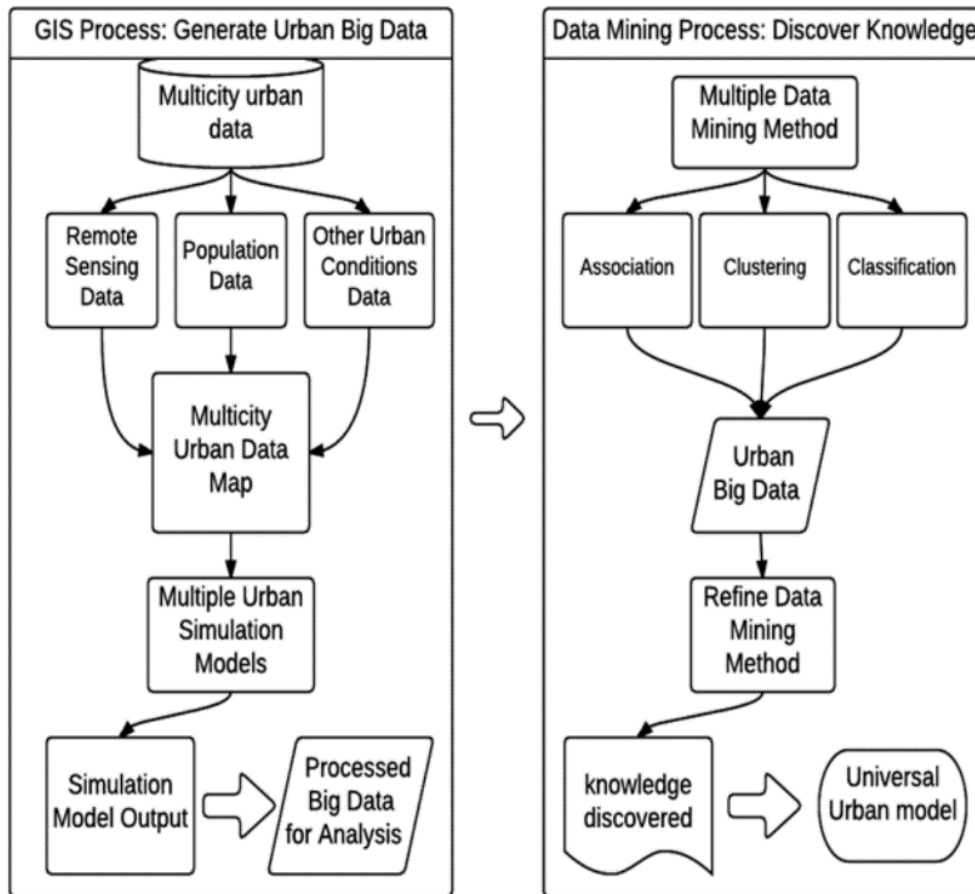


Figure 2.1: Proposed Solution for Analysis

Firstly, urban data from multicity environments is gathered. For now the most important data are the land use data, the population data and other data such as economic and policy data. All these data are preprocessed into an urban data map, which is a spatial form of data derived from multiple data sources. These provide the stationary information of different time periods. To conduct further analysis, such as population relocation and land use change, other information is

required. The urban simulation model gathers this information. The detailed process is explained later. The original urban data and intermediate model output are combined as the urban big data for further data mining analysis.

This research utilizes multiple data mining methods such as association rules, clustering, and classification to analyze the nonlinear relationships in the urban big data. Association rule mining addresses issues such as the driving force and consequence of population relocation by identifying suitable antecedents and consequents through relationships of the type A implies B. Cluster analysis identifies various forms of urban population relocation by determining the groups or clusters of urban population and their respective relocation. Classification is able to predict targets such as the estimated growth of the urban population over a certain period.

Once the knowledge of the relationships between land use change, population relocation and other parameters are found through the proposed approach. It sets the stage to combine them together in a comprehensive urban simulation model which has the ability to predict and explain the population relocation with universal capability, since this model is based on the knowledge from multicity urban big data. The following sections are the detailed explanation of these steps.

2. 1. 4 Urban Big Data

This research identifies the causes and consequences of population relocation. To analyze the non-linear relationship among population dynamics, land use change, and other urban conditions by the data mining method, a

proper data set must be generated. The data set in this research not only has a large number of urban areas, but also a large amount of attributes and indicators. It is a multi-dimensional data set of complex urban information, which constitutes urban big data.

A. Multicity System

The urban system this research aims to analyze is a system with multiple cities around the United States [10]. Different urban areas have different urban population dynamics, urban development trends, and conditions. To achieve the goal of this research, a single urban area is not suitable, since a single urban area cannot provide enough data for the nonlinear analysis, in terms of both quality and quantity. This research intends to produce a comprehensive urban simulation model, hence the knowledge from single urban area is not sufficient due to different urban have different conditions.

Due to the data available and the time limitation, it is not feasible to analyze all the urban areas in the United States. In 2014, there was research conducted on a nationwide survey of urban sprawls, they analyzed about 200 urban areas [1]. We use some of the results from this analysis for further work.

B. Data and Databases Description

The form of the urban area influences its conditions, which are measured from the empirical databases. To analyze the relationship between them, researchers point out many

indicator systems to identify and analyze the dynamic of urban development.

To analyze the urban areas population dynamic and development trend, there is a large requirement of various empirical databases. These data would be obtained from the database about population, the database about employment, the database about land use, and the database about street distribution. In the USA: The population database is U.S. Census of Population and Housing. The employment databases are the Census Transportation Planning Package (CTPP) data on employment and the Local Employment Dynamics (LED) database. The land use database is the National Land Cover Database (NLCD). The street distribution database is the national dataset of street centerlines by TomTom.

C. Urban Data Map Description

All the data collected from the previous steps are combined for data mining analysis, which is presented as a multicity urban data map.

The land use data from remote sensing is based on the emission and reflection of radiation. It has only accounted the impervious surfaces percentage and constructed materials. It would not fully represent the land use types. Census data which has the spatial distribution of population, income and employment, by interaction with these two layers of data, would generate a more reasonable urban data map. Once the land use map is generated, the other information can also be easily associated with each of the urban areas. The land use map is part of the urban big data. With this explanation, we now describe urban development activities captured by urban

simulation models.

Different urban simulation models require different forms of data, the economic-base model, and statistical analysis model may just need the general report of target areas. Some other simulation models like cellular automata and agent-based models require GIS software to provide available information to process further analysis. The data table must be presented in spatial form for analysis. Figure 2.2 shows an example of transferring population density data table to spatial form. The map is from sociaexplorer.com.

We prefer to use the ArcGIS platform to process the raw data from different databases. ArcGIS is the most popular GIS platform and could perform analysis of the data. The urban land use simulation models require land use map of different time periods, which can be easily achieved by the normal function of ArcGIS.

The ArcGIS platform preprocesses the raw data in this research. The raw data may not always contain all the information we need. For example, some of the raw land use maps may just contain remote sensing data with the information of impervious rate and land cover to identify the related land use type. No population data are contained in this kind of map. Thus if we need to process any urban simulation which requires population, ArcGIS is useful to connect all the population information from other databases to the raw map. This process can generate more relevant land use maps. ArcGIS can easily achieve this goal by the joint function in the attribute table.

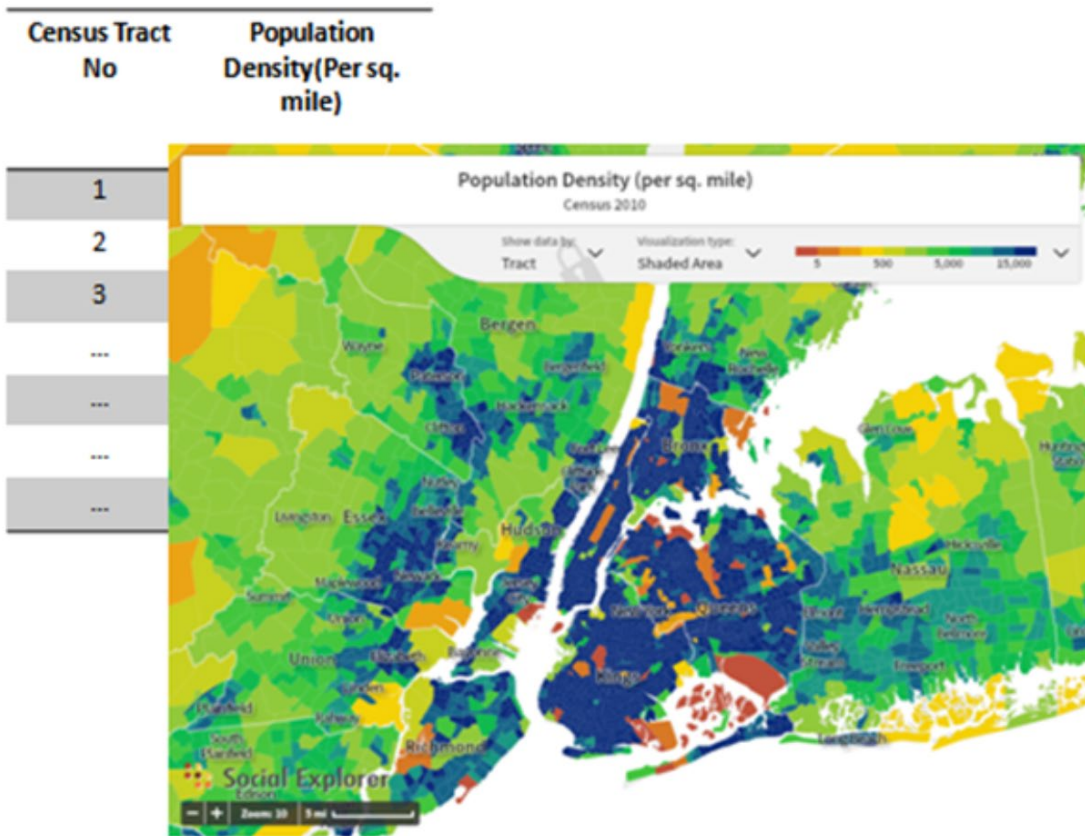


Figure 2.2: From the data table to the spatial map

The ArcGIS platform not only serves the purpose of processing data for the simulation, but in addition the visualization function could also be utilized for result checking and presentation.

D. Urban Data Map to Urban Big Data

The urban data map just contains the static information of urban areas. However, the urban development trends that are usually represented as the land use changes play a significant role in the population dynamics. These can only be captured by applying urban simulation models.

Different urban simulations measure different indicators and utilize them to explain current urban

development and predict future trends. These indicators have different forms and would be valuable for data mining analysis since they contain information about the urban development trends. The simulation model involved with artificial neural networks is not utilized much here due to the fact that it is a black box process. On the other hand, we find that the simulation model based on empirical assessment provides valuable knowledge. Examples include the land use transition matrix in Markov-chain model and the transition rules of cellular automata model. They contain the information about land use change trends. The combination of urban data maps and development trends would be treated as urban big data. The data mining methods would be applied on these. This following example of development trends extraction is using cellular automata method. In 2004, Xia Li, Anthony Gar-On Yeh [4] applied the decision tree learning model on the calibration of historical observation data to generate transition rules. This method has its own limitation due to being highly adapted to the sample areas. On the other hand, it means that the transition rules obtained by this method are highly associated with the sample areas and contain the information about local development trends.

$$Entropy = - \sum_{j=1}^R \frac{C_j}{S} * \log_2 \frac{C_j}{S}$$

If the division is efficient, it will get a smaller entropy value than the previous one. The efficient decision tree learning would ensure “the gain ratio is maximized at each node of the tree” [4]. This principle would also prevent generating too many transition rules.

We now describe the analysis conducted on this data after generating the urban big data through data maps and existing models including transition rules.

2. 1. 5 Non-Linear Relationship Analysis

The urban big data from previous step is utilized to discover knowledge about the relationships between urban population relocation, urban land use change and other urban conditions. These parameters are considered as constraints in the urban simulation and data mining. For example, the urban simulation model could generate a data set with population data as a constraint, e.g., population above a certain number, the land use type changes etc.

Furthermore, the data mining methods could discover knowledge about the relationships between urban land use change trend and population dynamics.

This analysis helps to answer questions such as: “What are the reasons for the differences between urban development trends of different cities?” We could find quantifying information, which is valuable for a comprehensive urban simulation model.

We find that association rule mining is suitable for analysis of nonlinear relationships between urban growth and urban conditions. Association rule mining is the technique of detecting rules among data sets, the rules are typically of the type: $A \Rightarrow B$ where A is the antecedent and B is the consequent [11]. This means that if a trend A occurs, then B is likely to occur. These rules have interestingness measures called Confidence and Support. The

Confidence C of a rule $A \Rightarrow B$ is the probability of B given A [i.e., $C = P(B|A)$], while the Support S is the probability of A and B occurring together in the entire data set [i.e., $S = P(A \wedge B)$]. This is standard terminology in association rule mining.

Now consider this with reference to our work. The urban big data as a data set is:

$$D = \{I_{p,q}\}$$

Here, the variable I relates to a certain parameter or trend, while p marks cities, and q marks certain attributes. Thus, in our context we define C as the confidence between parameters x and y which is given as:

$$C = \frac{\{I_{1,x}, I_{3,x}, \dots, I_{n,x} \cap I_{1,y}, I_{3,y}, \dots, I_{n,y}\}}{\{\{I_{1,x}, I_{3,x}, \dots, I_{m,x}\} \cap D\}}$$

This is the ratio between number of urban areas in which both x and y occur and the number of urban areas in which only x occurs. When we measure the confidence as defined herewith, it could be utilized for prediction, thus the more similar the urban condition is to the association rule, the more likely it is to occur.

Cluster analysis is helpful in urban simulation as follows. Clustering is a data mining technique that divides the entire data set of different objects into groups based on their similarity [12]. The urban big data contains information that could be used as indicators of similarity. For example, a transition rule may be described as: in city A , non-urban lands have $B\%$ chance to change into the urban lands when it has C distance to the center of the cities and $D\%$ of the

neighborhood is urban land. The A, B, C and D are the indicators, and by using cluster analysis, a lot of significant knowledge would be revealed. For example: in City A, X...Z, non-urban lands have the similar chance to change into urban lands under similar neighborhood conditions. Thus, when we build the comprehensive urban model, in the cities similar to City A, X...Z, this transition rule is applicable.

The classification analysis intends to produce rules which discover the relationships between urban land use change, population relocation and other indicators, and help to predict a target. For example, it could predict that under certain land use change and economic conditions, the amount of population change would be within a certain range. There are various of classification analysis methods. A common one is J4.8 decision tree learning [13]. Decision tree learning follows an inductive approach to learn from an existing data set and build a stem and leaf structure such that root represents a starting point, the intermediate nodes represent certain parameters and leaf nodes represent the final outcomes, e.g., in our case this could be urban sprawl.

With this description of the approach used in our work, we now proceed with a summary of our experimental evaluation.

2. 1. 6 Experimentation

In 2014, Hamidi and Ewing conducted a research to measure the urban decentralization

around the USA from 2000 to 2010 [14]. They analyzed 163 census urbanized areas. We use their results as one source of input. The data sets are publicly accessible [15]. They contain four main types of numeric indicators as follows [14].

- *Density factor*: The density factor refers to population, employment and build-up land density.
- *Mix Factor*: The mix factor or mix use factor pertains to the condition of population and employment land use type mixture.
- *Center Factor*: The center factor or centering factor determines the condition of how the population and employment concentrates in the urban center.
- *Street Factor*: The street factor is the condition of accessibility of the urban area.

In addition, there is also a *Numeric Composite Factor* which relates to the urban sprawl. It contains information about population and employment (density factor), urban land use/urban form (mix and centering factor), and accessibility (street factor). They utilize the statistical models to output indicators of population dynamics and urban development, which is suitable as a form of urban big data. Thus, we use these factors in our analysis with data mining methods.

Based on these inputs and the urban big data that we have generated, we run association rule mining, clustering and classification as described next.

The association rule mining finds several rules from which we can infer some interesting facts as follows. We find that the greater the mix factor, the lower is the tendency of the urban

sprawl occurrence. This is due to the composite index being “compact”. This implies that as there is a better mix between population and employment land use, the city tends to be more compact and less sprawl-prone. Likewise, we discover other interesting trends. Examples of association rules discovered from this analysis are shown in Figure 2.3.

1. mix factor10=high composite index10=compact 63 ==> mix factor00=high 63 conf:(1)
2. mix factor10=high composite index10=compact 63 ==> mix factor00=high 63 conf:(1)
3. mix factor10=high composite index10=compact composite index00=compact 61 ==> mix factor00=high 61 conf:(1)
4. mix factor10=high street factor10=high 58 ==> mix factor00=high 58 conf:(1)
5. mix factor10=high 85 ==> mix factor00=high 84 conf:(0.99)
6. density factor10=high 64 ==> density factor00=high 62 conf:(0.97)
7. mix factor10=high composite index00=compact 63 ==> composite index10=compact 61 conf:(0.97)
8. mix factor10=high composite index10=compact 63 ==> composite index00=compact 61 conf:(0.97)
9. mix factor10=high mix factor00=high composite index00=compact 63 ==> composite index10=compact 61 conf:(0.97)
10. mix factor10=high composite index10=compact mix factor00=high 63 ==> composite index00=compact 61 conf:(0.97)

Figure 2.3: Examples of Association Rules

Cluster analysis is then performed on the data set. Snapshots of the results appear in Figure 2.4 and Figure 2.5. In these figures, Cluster 0 and 3 are compact clusters, Cluster 1 and 2 are sprawl clusters, Cluster 4 is the sprawl+ cluster. Based on these figures, it is observed that urban areas with high centering factor are more compact, and those with low centering factor tend to be sprawl. Thus, urban centering, i.e., concentration of employment and population in the urban center would reduce the urban sprawl. Also, we notice that when the street factor improves, sprawl would reduce. This can be interpreted as follows. Proper design of streets would limit

edge development by stimulating the growth of developed urban areas, thereby decreasing the urban sprawl.

Attribute	Full Data (162)	0 (46)	1 (35)	2 (30)	3 (34)	4 (17)
density factor10	mid	high	mid	mid	high	mid
mix factor10	high	high	mid	high	high	low
centering factor10	high	mid	high	mid	high	mid
street factor10	high	high	mid	mid	high	low
composite index10	compact	compact	sprawl	sprawl	compact	sprawl+
density factor00	high	high	mid	mid	high	mid
mix factor00	high	high	high	high	high	low
centering factor00	high	mid	high	mid	high	mid
street factor00	mid	high	low	mid	high	low
composite index00	compact	compact	sprawl	sprawl	compact	sprawl+
density factor change	reduce	reduce	reduce	reduce	stable	stable
mix factor change	reduce	reduce	reduce	reduce	reduce	reduce+
centering factor change	stable	stable	stable	reduce	increase	stable
street factor change	increase	increase	increase+	increase	increase	increase+
composite index change	reduce	reduce	increase	reduce	increase	increase

Figure 2.4: Clustering Result Example

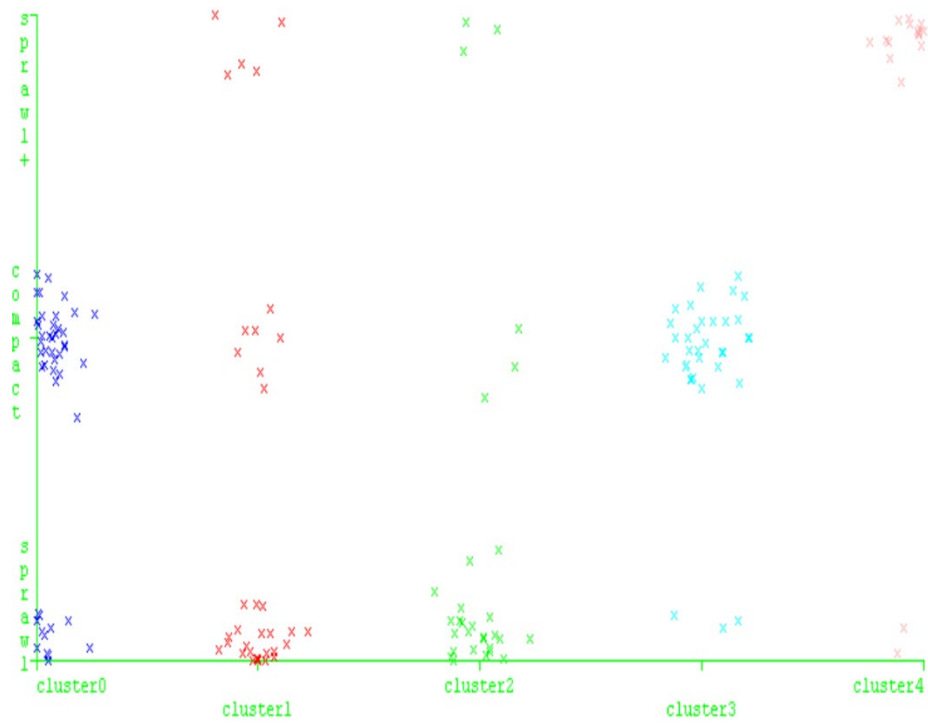


Figure 2.5: Visualization of Clusters

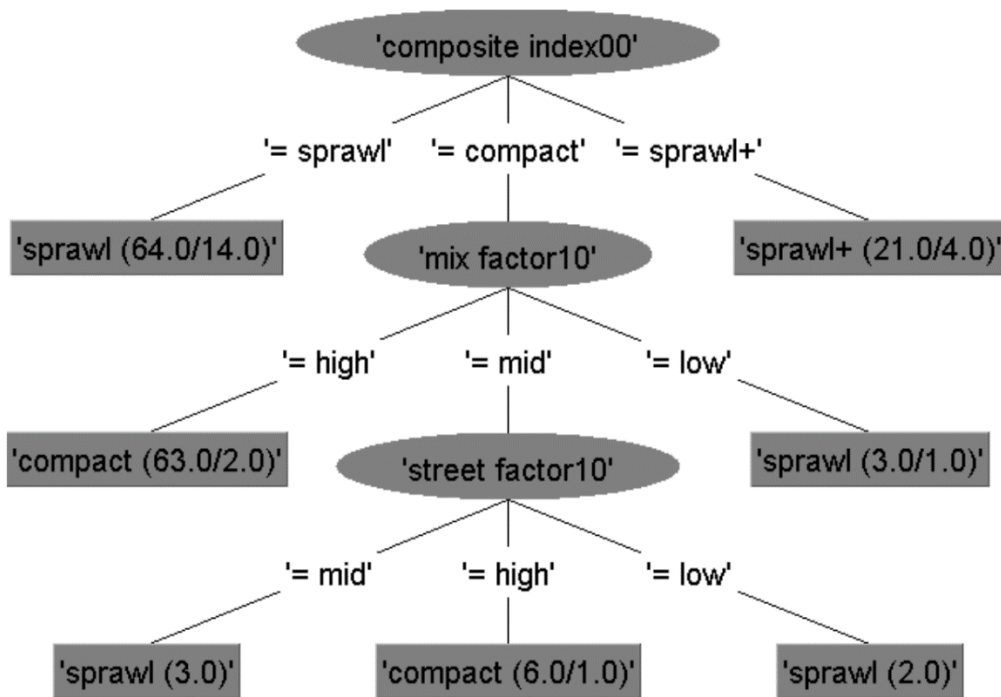


Figure 2.6: Visualization of Decision Tree

Classification analysis is performed with J4.8 decision tree learning for the same data set. A partial snapshot of example results appears in Figure 2.6. This example gives a result with 82.716% correctly classified instances. This result shows that most of the urban areas maintain their development conditions in these 10 years and the mix factor and street factor showed significant influence on the composite index. Low mix factor urban areas have the tendency to become sprawl. The low street factor with medium mix factor would also lead to sprawl. This result also follows the previous result as the mix factor and street factor have strong influence of the urban development. However, it does not include any influence of the changing trends of the factors. This along with other issues is being addressed in ongoing work.

2. 1. 7 Related Work

There is interest in the field of urban sustainability today from several perspectives. Urban decentralization has negative outcomes, e.g., traffic congestion, air pollution, lack of social capital [1]. Low population density is feature of urban decentralization. Policy implementation to reduce decentralization, e.g., smart growth supports high density communities, regulates low density communities [2]. Theories claim that economic factors, social amenities, health services etc. drive population changes [3]. Recently, there is applied data mining research in many fields since amount of available data is increasing, there are more powerful computers and there are advances in statistics & machine learning [5]. Prior research showed that data mining can be applied to calibration of cellular automata transition rules [4]. Data mining is suitable for the nonlinear relationship analysis between urban sustainability and urban conditions [6].

We address issues that have not been predominant in earlier works, e.g., factors affecting sprawl and sustainability and the relationships between them. Also, existing research typically has single-city environments while we consider a multicity global context. Our work also entails analysis of complex urban big data with the generation of the data itself involving multiple procedures including using GIS, remote sensing, and data from existing simulation models.

Our earlier work on Mining GIS Data to Predict Urban Sprawl [16] appeared in ACM KDD 2014. We analyzed data on *urban sprawl* (overgrowth & expansion of low-density areas

with issues like car dependency and segregation of residential & commercial use).

Spatiotemporal features on real GIS data e.g., population growth & demographics were mined using Apriori for association rules [12] and J4.8 for decision tree classification [13], adapted to geospatial analysis, with ArcGIS for mapping. Knowledge discovered was used to build a spatial decision support system (SDSS) to predict whether “urban sprawl” was likely to occur with reasons. In our current work, we delve deeper into specific aspects of urban sprawl and sustainability and head towards generating a comprehensive urban simulation model to cater to various interesting user questions. Our proposed research activity would contribute to the state-of-the-art by discovering knowledge useful to environmental scientists, urban planners and other interested users.

Recently, there is much interest in the development of Smart Cities [17]-[19]. These entail several characteristics, among which our work would potentially make contributions to Smart Governance and Smart Environment. The Smart Governance aspect includes transparent governance and participation in decision-making, where our work on providing useful information pertaining to sprawl and sustainability could play a role. The Smart Environment aspect includes features such as greenness and energy efficiency [20], conserving natural resources and living sustainably [21]. Thus, our work has the effect of contributing in that avenue due to the analysis of sustainability parameters and goals of sustainable population relocation as a whole. Hence, this work has a broader impact in the context of Smart Cities [17]-[19].

2. 1. 8 Conclusions and Ongoing Work

In this research, we address the issue of multicity urban simulation. We propose to integrate urban simulation and data mining to conduct predictive and diagnostic analysis about the relationship between population dynamics, land use change, and urban development. Our experimentation reveals that data mining methods have the ability to discover knowledge from the national level urban data sets that contain urban development trends and urban conditions. The following are some interesting findings from this work.

- Greater the mix factor, lower the tendency of urban sprawl occurrence
- Urban areas with high centering factor are more compact and those with low centering factor tend to cause sprawl
- Mix factor and street factor combined have a significant influence on sustainable urban development
- Proper design of streets is an important indicator of sustainability

The outcomes obtained from some experiments could be even further improved by future work in this research. With respect to the techniques, we could potentially consider other methods such as: an ensemble of classifiers constituting a mixture of experts scenario for prediction in the real world; discovering associations and using them to build classifiers; clustering followed by classification and more.

With respect to the data, we could enhance the data set itself so that it includes text and image data in addition to the sources already considered. We could also mine opinions from social media data. This would be very useful given that the public satisfaction is very important in aspects such as urban development and population relocation. Public opinions are often expressed over social media and hence it would be useful to capture them in the mining process. The data on social media itself could consist of textual, numeric and image data. This would need more advanced techniques for mining. Thus, we could conduct further analysis with enhanced data sets and use that to generate a comprehensive urban simulation model. Mining over such data could potentially yield even more interesting results.

In order to address this, we need to solve various sub-tasks in this research, e.g., defining precise interestingness measures for association rules, selecting appropriate classifiers in an ensemble, pre-processing the urban big data to extract relevant information for mining, extracting and interpreting important social media data etc. All of this constitutes our ongoing research. We claim that this would discover even more interesting knowledge that would be of greater value to urban planners and other users

2. 1. 9 Acknowledgement

This research is supported by a Doctoral Assistantship from the Environmental Management Program at Montclair State University. The authors thank the former program

director Dr. Dibyendu Sarkar and the current program director Dr Stefanie Brachfeld for the research funding.

2. 1. 10 Reference

- [1] Ewing, R., & Hamidi, S. (2014). Measuring Sprawl 2014. Retrieved from <http://www.smartgrowthamerica.org/documents/measuring-sprawl-2014.pdf>
- [2] Smartgrowthamerica.org, 'What is "smart growth?" | Smart Growth America', 2015. (Online). Available: <http://www.smartgrowthamerica.org/what-is-smart-growth>.
- [3] Nagy, R., & Lockaby, B. (2010). Urbanization in the Southeastern United States: Socioeconomic forces and ecological responses along an urban-rural gradient. *Urban Ecosystems*, 14(1), 71-86. doi:10.1007/s11252-010-0143-6
- [4] Li, X., & Gar-On Yeh, A. (2004). Data mining of cellular automata's transition rules. *International Journal Of Geographical Information Science*, 18(8), 723-744. doi:10.1080/13658810410001705325
- [5] Miller, H., & Han, J. (2001). *Geographic data mining and knowledge discovery*. London: Taylor & Francis.
- [6] Rajasekar, U., & Weng, Q. (2009). Application of Association Rule Mining for Exploring the Relationship between Urban Land Surface Temperature and Biophysical/Social Parameters. *Photogrammetric Engineering & Remote Sensing*, 75(4), 385-396.

doi:10.14358/pers.75.4.385

[7] Schneider, A., & Woodcock, C. (2008). Compact, Dispersed, Fragmented, Extensive? A Comparison of Urban Growth in Twenty-five Global Cities using Remotely Sensed Data, Pattern Metrics and Census Information. *Urban Studies*, 45(3), 659-692.

doi:10.1177/0042098007087340

[8] Göktuğ, M. (2012). Urban Sprawl and Public Policy: A Complexity Theory Perspective. *Emergence: Complexity & Organization*, 14(4), 1-16.

[9] Santé, I., García, A., Miranda, D., & Crecente, R. (2010). Cellular automata models for the simulation of real-world urban processes: A review and analysis. *Landscape And Urban Planning*, 96(2), 108-122. doi:10.1016/j.landurbplan.2010.03.001

[10] Branch, G. (2015). 2010 Urban Area Facts - Geography - U.S. Census Bureau. Census.gov. Retrieved 18 February 2015, <https://www.census.gov/geo/reference/ua/uafacts.html>

[11] Agrawal R., Imieliński T. and Swami A., 'Mining association rules between sets of items in large databases', *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207-216, 1993.

[12] MacQueen, J. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1: Statistics, 281—297.

[13] Quinlan J.R., *C4.5*. San Mateo, Calif.: Morgan Kaufmann Publishers, 1993.

[14] Hamidi, S., & Ewing, R. (2014). A longitudinal study of changes in urban sprawl

between 2000 and 2010 in the United States. *Landscape And Urban Planning*, 128, 72-82.

doi:10.1016/j.landurbplan.2014.04.021

[15] [Gis.cancer.gov](http://gis.cancer.gov),. (2015). County Level Urban Sprawl Indices - Geographic Information Systems & Science. Retrieved 25 April 2015, from <http://gis.cancer.gov/tools/urban-sprawl/>

[16] Pampoore-Thampi, A. & Varde, A. (2014). Mining GIS Data to Predict Urban Sprawl, ACM conference on Knowledge Discovery and Data Mining (KDD Bloomberg Track), New York City, NY, pp 118-125.

[17] IEEE Smart Cities, <http://smartcities.ieee.org/>

[18] Vienna University of Technology et al., European Smart Cities, www.smart-cities.eu

[19] Smartcitiescouncil.com, "Smart Cities Council | Definitions and overviews", 2015. (Online). Available: <http://smartcitiescouncil.com/smart-cities-information-center/definitions-and-overviews>.

[20] Pawlish, M., Varde, A., Robila, S. and Ranganathan, A. (2014). A Call for Energy Efficiency in Data Centers, *Journal of ACM's Special Interest Group on Management of Data Record (SIGMOD Record)*, 2014, Vol. 43, No. 1, pp. 45-51.

[21] Varde A., and Du X., Multicity Simulation with Data Mining for Urban Sustainability, Presentation at Bloomberg Data Science Labs, March 2015

2. 2 Mining PM2.5 and Traffic Conditions for Air Quality

Abstract: Fine particle pollution is related to road traffic conditions. In this work, we analyze *Particulate Matter with a diameter less than 2.5 micrometers, called PM2.5*, along with traffic conditions. This is done for multicity data to study the relationships in the context of environmental modeling. The goal behind this modeling is to support prediction of PM2.5 concentration and resulting air quality. We deploy data mining algorithms in association rules, clustering and classification to discover knowledge from the concerned data sets. The results are used to develop a prototype tool for the prediction of PM2.5 and hence air quality for public health and safety. This paper describes our approach and experiments with examples of PM2.5 prediction that would be helpful for decision support to potential users in a smart cities context. These users include city dwellers, environmental scientists and urban planners. Novel aspects of this work are *multicity PM2.5 analysis by data mining and the resulting air quality prediction tool*, the first of its kind, to the best of our knowledge.

Keywords: *Air Pollution; Data Mining; Environmental Modeling; Fine Particles; Predictive Analysis; Public Health*

(Chapter 2.2 reused the previously published paper Du, X., & Varde, A. (2016), Mining PM2.5 and traffic conditions for air quality, *7th International Conference on Information and Communication Systems (ICICS)*, <https://doi.org/10.1109/iacs.2016.7476082>).

2. 2. 1 Introduction

In order to mitigate the negative effect of airborne fine particles on human health, air visibility and global climate, it is useful to have a good prediction tool. This would be in line with the modern day concept of making cities smart for prospective users by aiding in decision-making scenarios.

Road traffic in cities is the major source of airborne fine particles while the burning of fossil fuel produces both fine particles and its precursors [1]. The traffic would somehow relate to economic conditions, which is also significant with respect to fine particle pollution. In traffic sites, i.e., areas with high traffic volume, the air has a higher concentration of fine particle pollutants [2]. This motivates the development of regulations and standards for heading towards a cleaner environment [3]. The Clean Air Act regards particulate matter as a harmful pollutant to public health and requires the United States Environmental Protection Agency (EPA) [4] to set national air quality standards for PM_{2.5} and PM₁₀. Note that the term PM refers to *particulate matter* while the adjacent number refers to its maximum diameter in micrometers. Hence, PM_{2.5} is particulate matter with diameter less than 2.5 μ m. This is found to be particularly harmful to the human body since it is relatively harder for the respiratory system to filter this out. Figure 2.7 shows the penetration of PM_{2.5} into the lungs (left) and its harmful effects demonstrated by observing rat lungs (right). From this figure, it can be seen that pollutants with diameter around

10 μ m penetrate into the nose while those with diameter less than 2.5 μ m penetrate really deep into the lungs. Short term exposure to PM_{2.5} can cause problems such as asthma attacks and acute bronchitis while its long term exposure can cause reduced lung function, chronic bronchitis and possibly premature death. Hence, it is important to set standards for PM_{2.5} pertaining to health and safety.

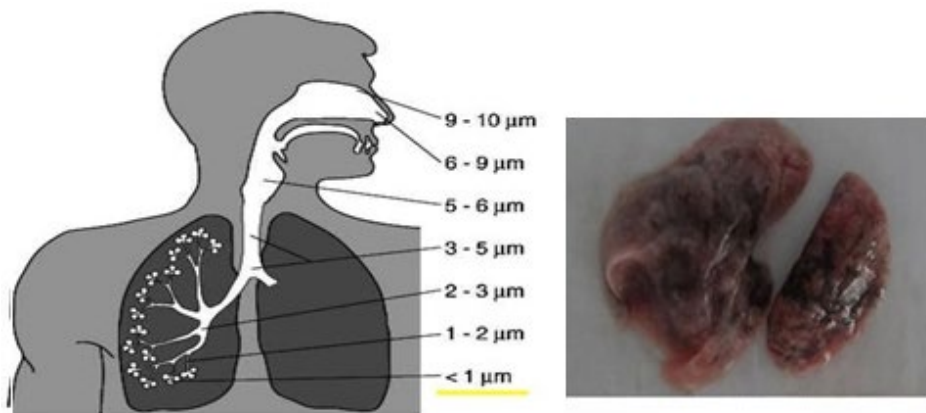


Figure 2.7: PM_{2.5} penetration in lungs and harmful effects of the pollutant

In 2012, EPA updated the PM_{2.5} standards range from 15 to 12 microgram per cubic meter as the safe limit [4, 5]. Thus, for the environment to be safe from a health standpoint, the amount of particulate matter of the 2.5 type should be no more than 12 μ g/m³ for a 24-hour period.

AQI Category	Index Values	Revised Breakpoints ($\mu\text{g}/\text{m}^3$, 24-hour average)
Good	0 - 50	0.0 – 12.0
Moderate	51 - 100	12.1 – 35.4
Unhealthy for Sensitive Groups	101 – 150	35.5 – 55.4
Unhealthy	151 – 200	55.5 – 150.4
Very Unhealthy	201 – 300	150.5 – 250.4
Hazardous	301 – 400	250.5 – 350.4
	401 – 500	350.5 – 500

Figure 2.8: AQI values for PM2.5 as per health standards

The EPA has an Air Quality Index (AQI) system [5] built for daily prediction and record for PM2.5. In Figure 2.8, there is brief depiction of how AQI is related to health impact and the PM2.5 concentration. In this figure, AQI Category refers to its description from a health standpoint, Index Values define the actual numerical value for AQI, and Revised Breakpoints pertain to the PM2.5 concentration for the given index value.

Data mining, the process of discovering knowledge from data, provides a good approach for modeling the relationships between environmental parameters. We consider multiple data mining methods in this research for analysis of PM2.5 and traffic conditions with respect to air quality. There are multiple sources of PM2.5 and a vast number of factors, which would influence the concentration [1, 2, 3]. Traffic is a major source of PM2.5 and automobile exhaust, abrasion, re-

suspension sources are all traffic related. The analysis between traffic indicators and PM2.5 concentration would thus reveal useful knowledge and would be valuable for the prediction of PM2.5 concentration. This in turn can be used to predict the air quality and its suitability for public use along with health and safety issues. This forms the focus of our work and would potentially help in making a contribution to smart cities [6] through predictive analysis of air quality.

The rest of this paper is organized as follows. Section II gives a description of the problem we address with its research questions. Section III describes our proposed solution. Section IV provides the experimental evaluation along with a prototype tool for PM2.5 prediction. Section V overviews related work in the area. Section VI states the conclusions and ongoing work.

2. 2. 2 Problem Definition

PM2.5 refers to air pollutants consisting of fine particles having a diameter less than 2.5 micrometers. High PM2.5 concentration would cause some damage to human health while long term exposure to PM2.5 could possibly lead to cardiovascular and respiratory disease and also genotoxicity, mutagenicity and cancer. All this could occur due to its high penetration into the human body [6]. Since PM2.5 has highly negative effects, it is desirable to avoid it, and thus it is smarter to live in an environment with negligible PM2.5 concentration. In this research, we model the relationship between PM2.5 and traffic conditions with respect to air quality. Thus, the

problem addressed is divided into 3 research questions as follows.

Q1. What is the relationship between traffic parameters and PM2.5 concentration?

Traffic contributes to PM2.5 emission via multiple ways: direct emission, abrasion and re-suspension. Gas pollutants lead to secondary PM2.5. Adequate road density would mitigate traffic congestion, which would reduce the exhaust by decreasing the running time. Traffic conditions are also related to economic conditions, which would influence the quality of fossil fuel and consumption. A proper combination of traffic parameters would lead to reduced PM concentration. This research would model these relationships via data mining.

Q2. What are the main traffic parameters pertaining to PM2.5 concentration?

There are various traffic parameters. It is important to estimate which of these would be the major indicators of PM2.5 concentration. It is also useful to understand how the parameters interact with other parameters.

Q3. How can we utilize the knowledge discovered by data mining for prediction of PM2.5 concentration?

To produce environmental management benefits, the knowledge that has been discovered by mining data pertaining to PM2.5 should be used for prediction. It is helpful be able to predict PM2.5 concentration based on various environmental factors. A prototype tool would be built to address this.

2. 2. 3 Proposed Solution

We propose to utilize data mining methods to discover relationships between PM2.5 and traffic conditions. We focus on association rule mining, cluster analysis and classification techniques [7]. The proposed approach in our analysis is illustrated in Figure 2.9. As seen here, we first collect data on traffic conditions and PM2.5 and preprocessing is conducted on it by applying suitable filters and other operations, e.g., attribute selection, instance sampling, discretization etc. The resulting data is stored in a preprocessed database. This is then mined using the following techniques.

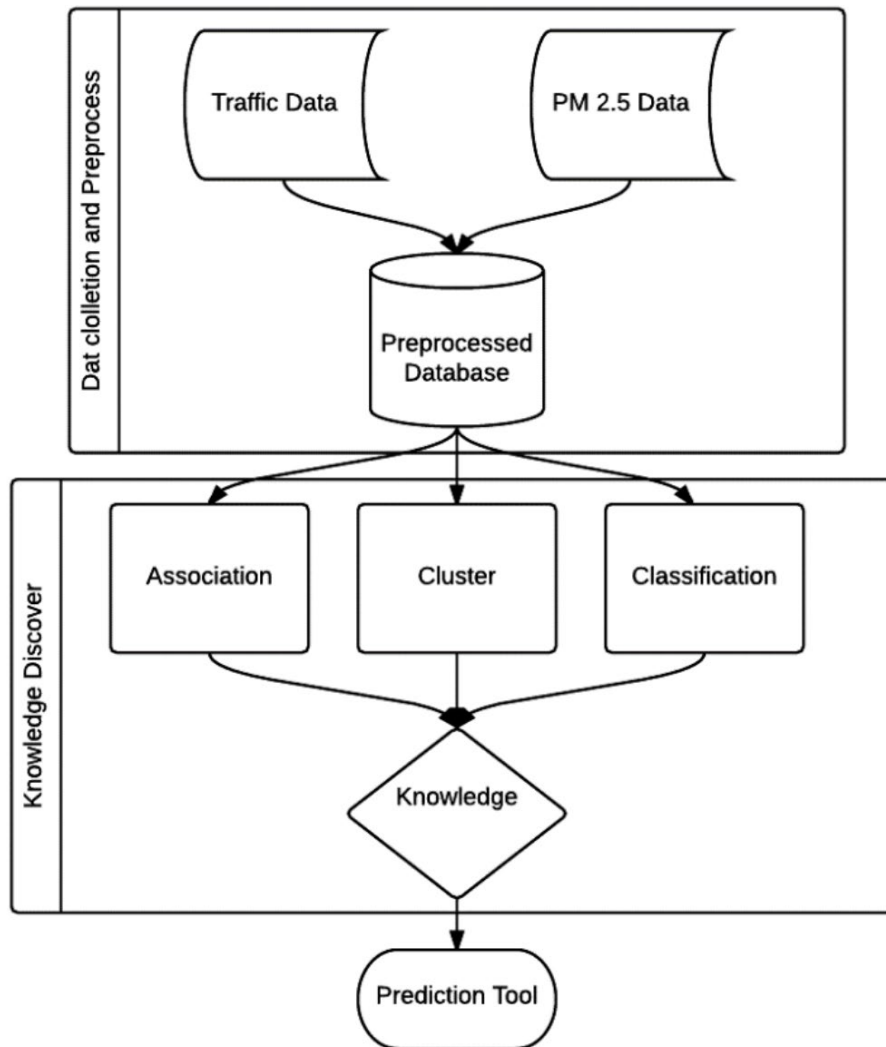


Figure 2.9: Proposed Approach for PM2.5 Analysis

Association rule mining helps discover relationships of the type $A \Rightarrow B$ [7]. Hence, it can discover how one parameter on PM2.5 affects another. Clustering helps grouping instances based on similarity [7]. Thus it would form categories based on similar ranges of PM2.5 and related parameters. Classification helps in the prediction of a target [7]. This could therefore be used to predict the range of PM2.5 given various other attributes. Knowledge discovered by mining is

then used to build a prototype prediction tool for decision support. This estimates the PM2.5 ranges and thereby the air quality based on user inputs.

Our data sources here are mainly from World Bank [8] and World Health Organization [9] online databases. Data gathered from here is combined into a comprehensive data set for data mining. In our work, we have joined the PM2.5 concentration data based on country code. The original raw data has the following significant aspects as shown in Figure 2.10.

Series Code	Series Name
PM25	PM2.5 pollution, mean annual exposure (micrograms per cubic meter)
IS.ROD.DNST.K2	Road density (km of road per 100 sq. km of land area)
IS.ROD.DESL.PC	Road sector diesel fuel consumption per capita (kg of oil equivalent)
IS.ROD.ENGY.PC	Road sector energy consumption per capita (kg of oil equivalent)
IS.ROD.SGAS.PC	Road sector gasoline fuel consumption per capita (kg of oil equivalent)
IS.VEH.NVEH.P3	Motor vehicles (per 1,000 people)
IS.VEH.PCAR.P3	Passenger cars (per 1,000 people)
IS.VEH.ROAD.K1	Vehicles (per km of road)

Figure 2.10: Raw Data on PM2.5 from Worldwide Sources

The original data are all numerical variables. These might not be suitable for some data mining methods, thus we perform discretization to convert them into nominal data. The US EPA sets the standard of 12 $\mu\text{g}/\text{m}^3$ for PM2.5 as being safe. We thus use this standard in our analysis. This is for knowledge discovery by data mining methods as well as prediction in the prototype decision support tool.

2. 2. 4 Experimentation

We provide a summary of the experimental evaluation we conducted with our proposed approach. Association analysis was the first data mining method we used to detect the correlation among the attributes. We used the well-known Apriori algorithm [10] to conduct association rule mining. We discretized the numeric data on PM2.5 using the equal frequency binning method.

After conducting analysis, we got some interesting inferences. For example, we found regions that have strong connection with PM2.5 concentration. There were rules showing that income groups could influence the other traffic conditions. This was reasonable due to the fact that the economic conditions directly influence the traffic facility construction. It was also found that high diesel consumption was not directly related to high PM2.5 concentration. Examples of interesting association rules obtained as the output of Apriori are shown herewith:

*Region=Europe & Central Asia Vehicles_Per_KM=VERY LOW => PM25_Class=GOOD
conf:(1)*

*Gasoline_Consumption=VERYLOW Road_Density=VERY LOW
Cars_Per_K_People=LOW => PM25_CLASS=MODERATE conf:(0.91)*

The terms GOOD and MODERATE here, pertain to the PM2.5 ranges with respect to their impact on air quality index (as shown in Figure 2.8). For example, PM2.5 class = GOOD implies that the resulting AQI category is good since its index value is in the range of 0-50, which would

occur with a PM2.5 concentration of 0.0 to 12.0 $\mu\text{g}/\text{m}^3$ as a 24-hour average. This is with reference to the first row in the table in Figure 3-2. Likewise, we can interpret other ranges.

Attribute	Full Data (142)	Cluster#			
		0 (58)	1 (36)	2 (26)	3 (22)
IncomeGroup	High income: OECD	Upper middle income	High income: OECD	High income: nonOECD	High income: OECD
Diesel_Consumption	229.4782	108.6279	416.6128	208.7012	266.4182
Gasoline_Consumption	207.3129	96.9031	341.0744	286.0335	186.4773
Road_Density	94.2376	39.3336	140.8336	149.4238	97.5164
Cars_Per_K_People	262.1393	120.5369	493.2781	234.1404	290.3173
Vehicles_Per_K_People	317.5645	151.9903	588.3836	288.6919	345.0418
Vehicles_Per_KM	48.3194	37.9036	50.2311	86.2058	27.8759
PM2.5_RANGE	'(-inf-5.845]'	'(15.12-18.43]'	'(-inf-5.845]'	'(21.755-inf]'	'(5.845-11.98]'

Figure 2.11: Sample Output of Clustering

During cluster analysis, the classical simple k-means algorithm [11] was used. A sample output is shown in Figure 2.11. Here the value of $k=4$, i.e., there are 4 clusters. It is observed that Cluster 0 has relatively low traffic indicators, however the medium PM2.5 range has been already over safe PM2.5 standards. It showed that in these countries, the traffic is not the major source of PM2.5 and the income of this cluster is the lowest. Cluster 2 has the highest PM2.5 concentration, yet it is not the highest traffic indicator, the countries in this cluster also have other significant PM2.5 sources or poor regulation of automobile emission. Cluster 1 and Cluster 3 both have PM2.5 under the safe standards and also pertain to OECD (Organization for Economic Cooperation and Development) countries. It shows that the PM2.5 presence is also affected by factors other than the actual traffic concentration.

Finally, we conducted classification analysis on the data. We used the J4.8 decision tree algorithm, the Java version of the C4.5 algorithm for learning by decision tree induction [12].

The main theory behind it is to maximize the information gain of each node of the tree. By analyzing the rules of the J4.8 classifier output, it was found that the region attributes have the strongest influence. It was also discovered that the PM2.5 pollution is highly associated with local conditions.

An interesting result was the fact that high gasoline and diesel consumption does not directly lead to the higher PM2.5 concentration. In fact, contrary to the popular belief it was found that medium gas consumption causes greater PM2.5 concentration than high gas consumption. After further analysis, it was found that this could be reasoned as follows. High gas consumption usually associates with much better economic conditions and better pollutant regulations. Further, the income attribute is also significant to these rule distributions. In other words, high income groups and high gas consumption groups have better regulatory facilities due to which PM2.5 concentration does not increase significantly. Therefore, while low gas consumption causes low PM2.5 concentration, the relationship is not linear since other factors also influence PM2.5 presence. A partial snapshot of a decision tree obtained in our experiments is shown below.

Region = East Asia & Pacific

| Gasoline_Consumption <= 427.7

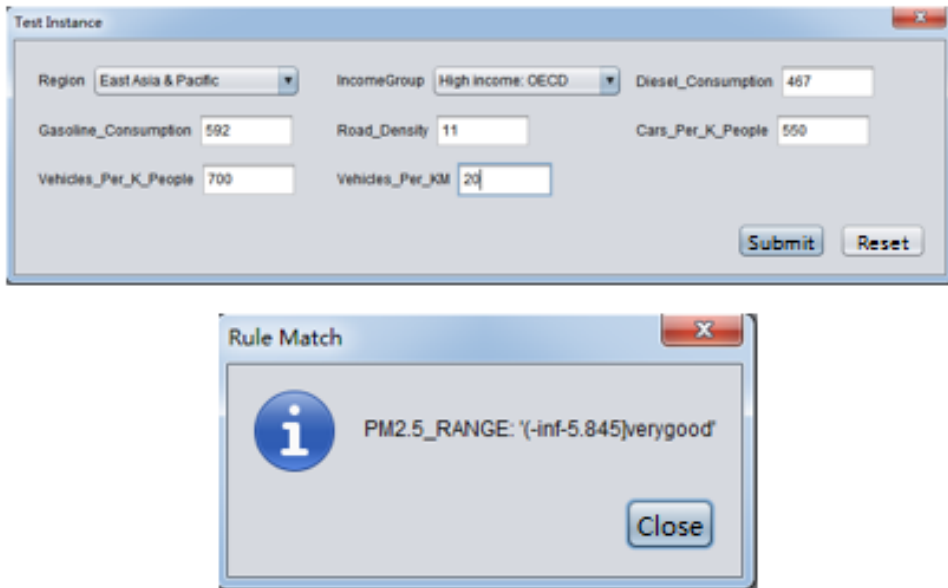
|| IncomeGroup = High income: nonOECD: '(18.43-21.755]' (2.0)

|| IncomeGroup = High income: OECD: '(21.755-inf)' (2.0)

```
|| IncomeGroup = Low income: '(18.43-21.755]' (2.0/1.0)
|| IncomeGroup = Lower middle income: '(11.98-15.12]' (2.0)
|| IncomeGroup = Upper middle income
||| Diesel_Consumption <= 114.38: '(21.755-inf)' (2.0)
||| Diesel_Consumption > 114.38: '(11.98-15.12]' (2.0)
| Gasoline_Consumption > 427.7: '(-inf-5.845]' (5.0)
```

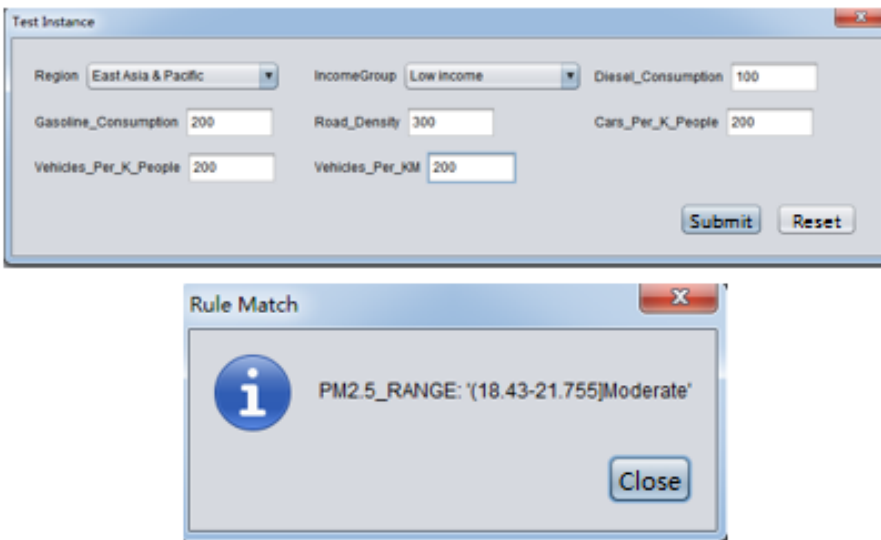
In this tree we can see some of the interesting findings mentioned herewith. The region and income have significant influence on PM2.5 concentration. Diesel consumption seems to have a reverse connection with PM2.5 concentration. Gas consumption has an effect but is not directly proportional to the concentration of the PM2.5 pollutants.

Results from the data mining analysis were then used to develop a prototype prediction tool. In our work, the programming for the prediction tool was done in Java. We used the output of the experiments conducted herewith to design the tool, useful in decision support.



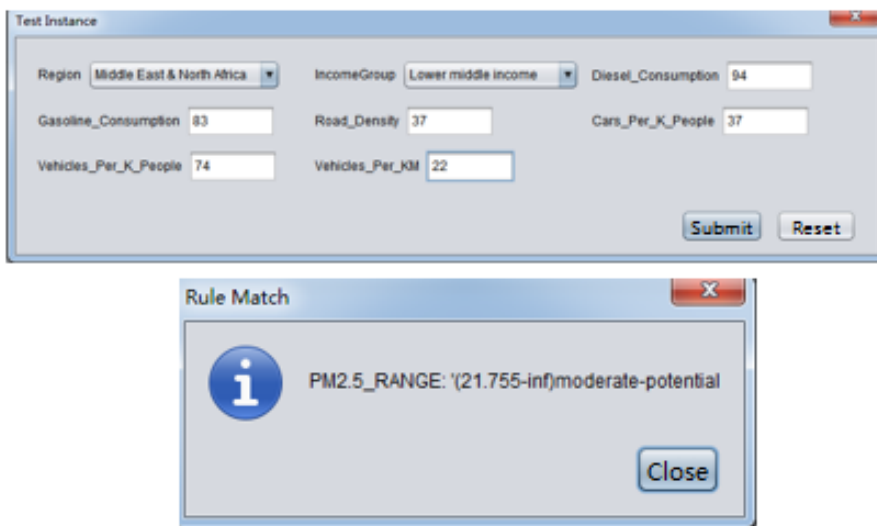
Much less than the 12 $\mu\text{g}/\text{m}^3$ regulation, minimal negative impact

Figure 2.12: Example of predicted output with safe PM_{2.5} range



More than the 12ug/m3 regulation,
moderate chance to have negative impact

Figure 2.13: Example of moderate PM2.5 range as predicted output



Much more than the 12ug/m3 regulation, more chance to have
bad day, potential damage to human health.

Figure 2.14: Example of moderate to potentially unsafe PM2.5 range prediction

This tool asks users to input the relevant data and predicts the PM2.5 range as the output with suggestions on its health and safety consequences. This Java-based tool has a GUI with an interactive, menu-driven screen. Users are allowed to enter the input conditions for prediction of PM2.5 ranges. The output helps users even without much professional knowledge, to fathom the result, namely, estimated range of PM2.5 based on given inputs with respect to health impacts.

The results with sample executions of user inputs are shown in Figure 2.12, Figure 2.13 and Figure 2.14. The terms “very good”, “moderate” and so on describe the PM2.5 safety range in air quality as per the chance of affecting public health. For example, consider Figure 2.12. If the user entered inputs for East Asia and Pacific as gas consumption: 182, vehicle concentration: 700, high income group, road density: 11, vehicles per kilometer: 20, diesel consumption: 467 and car concentration: 180, the tool would predict that the PM2.5 range is “very good”. This would mean that the range is between 0.0 to 12.0 $\mu\text{g}/\text{m}^3$, which is well within the safe limits for good health, with reference to AQI standards in Figure 2-8. Likewise, the other figures can also be interpreted.

With many such experiments, we found that there were useful predictions provided by this prototype tool, as evaluated by domain experts in Environmental Management. The results of these predictions would be useful to government bodies in order to estimate PM2.5 levels based on various factors and regulate policies accordingly. They would also be useful to urban dwellers and prospective residents to get an idea of pollutant concentrations and make decisions about

current lifestyles and potential relocation.

Furthermore, this tool would provide inputs to scientists in Environmental Management to conduct further research. For example, it would propel them to analyze the detailed causes of correlations between specific traffic conditions and PM_{2.5} concentrations. It would also help to promote the discovery of approaches for reducing the harmful effects of pollutants and making improvements from a health standpoint.

While these results are good for a prototype, the data in this paper is World Bank and WHO data and the data scale is too big for precise analysis. Since the PM_{2.5} concentration is a major concern for urban environmental health, city scale data would be better. Data collected by global remote sensing would also be useful. Further, the PM_{2.5} data only has two time periods: 2005 and 2010, while the global remote sensing data could yield PM_{2.5} data for each year. Yet another aspect is that the road density data does not incorporate specific details of the road conditions. This could also have an impact on the PM_{2.5} concentration. Finally, some regions have significantly less data than the other regions, and streamlining this data to make it more uniform could also lead to more accurate prediction. These and other issues provide the potential for further research.

2. 2. 5 Related Work

Urbanization though desirable has its negative outcomes, e.g., traffic congestion, air

pollution and lack of social capital [13]. Policy implementation in this area e.g., smart growth supports high density communities and regulates low density communities [14]. This is in line with the concept of smart cities that aim to provide better urban facilities including prior analysis useful to potential residents, catering to their smart environment and smart governance characteristics [15]. Furthermore, there are various theories claiming that economic factors, social amenities and health services are some of the factors that drive urban population changes [16]. This motivates conducting research on such factors with the ultimate goal of enhancing urban sustainability.

Recently, applied data mining research has been found useful in many fields including Environmental Management since amount of available data is increasing, there are more powerful computers and there are advances in statistics & machine learning [17]. Prior research shows that data mining can be applied to the calibration of cellular automata transition rules that could potentially relate to specific theories in urban relocation [18]. Moreover, data mining techniques such as association rules have been applied for conducting nonlinear relationship analysis between various urban conditions [19].

Given this general background on the role of data mining in urban sustainability research, we specifically address the issues that have not been the focus of earlier works. Hence, in this paper we deal with fine particle air pollutants, more specifically, particulate matter with diameter less than $2.5\mu\text{m}$. We focus on these due to the fact that the human body cannot easily filter such

fine particles and thus they penetrate deep into the respiratory system, thereby being particularly harmful. We analyze PM_{2.5} impact on air quality with respect to the effects on public health. Also, much of the existing pollution research caters to single-city environments while in our work we consider a multicity global context for pollutant analysis, focusing on real data from worldwide sources.

Our earlier research on mining GIS data in the context of urban sprawl prediction appeared in ACM KDD 2014 (Bloomberg Track) [20]. We analyzed data pertaining to urban sprawl (overgrowth & expansion of low-density areas with issues like car dependency and segregation of residential & commercial use). Conducting further work in the area, we analyzed urban big data considering parameters such as population density, street factors and employment rate, in order to discover knowledge useful for sustainable population relocation [21]. This was found to be useful from a geoinformatics standpoint.

In our current research, we delve deeper into specific aspects of urban sprawl and sustainability to head towards smart cities [5]. We thus analyze climate change, a hot topic intriguing environmental scientists. Within that, a specific subtopic is air pollution and that brings us to analyzing the effects of fine particle pollutants. We also address health and safety consequences, to provide suggestions for sustainable population relocation. Our research would contribute to the state-of-the-art by discovering useful knowledge on air pollution, climate change, its health impacts and the effects on urban population. This knowledge would be very

useful to environmental scientists, urban planning agencies and city residents. It would therefore have the impact of contributing to the smart cities initiative [5, 14, 15] by helping to provide a smart environment that is clean and healthy. It would also help in smart governance through better decision support based on predictive analysis in urban planning.

2. 2. 6 Conclusions and Future Work

In this paper, we have addressed the issue of modeling the relationships between fine particle air pollutants PM_{2.5} and traffic conditions in urban locations worldwide. This has been done with the goal of predicting air quality with respect to the presence of PM_{2.5} and its impact on public safety from a health standpoint. We have used online environmental data on PM_{2.5} from cities in a global context and conducted data mining using association rules, clustering and classification to model the relationships between various PM_{2.5} related parameters. The knowledge discovered by this environmental modeling has been used to build a prototype tool for the prediction of PM_{2.5} based on environmental conditions entered as user inputs. The tool predicts the range of PM_{2.5} as relevant to air quality with respect to public health.

This prototype prediction tool is helpful in analyzing PM_{2.5} occurrence and its impacts in the broad context of smart cities. It would be useful in decision support for existing city dwellers, potential residents of urban areas and government agencies such as urban planning departments. Additionally, this would provide inputs to environmental scientists for further research. It would

also help data mining professionals in real-world case studies. To the best of our knowledge, ours is one of the first works to build a prediction tool for air quality. This, along with the fact that we delve into multicity PM2.5 research with data mining constitutes the novelty of our initiative.

Further research in this area includes expanding this prototype into a large-scale predictive analytics tool. This would involve detailed analysis with remote sensing data, social media sites and other sources. It would also involve addressing more specific analytical issues with respect to the research questions in the pertinent areas to enhance the development of smart cities. It is expected this full-fledged predictive analytics tool would have the broader impact of enhancing urban sustainability.

2. 2. 7 Acknowledgment

This research is supported by a Doctoral Assistantship for the PhD student Mr. Xu Du from the Environmental Management Program in the College of Science and Mathematics at Montclair State University. The authors convey thanks to the former Program Director Dr. Dibyendu Sarkar and the current Program Director Dr Stefanie Brachfeld for the research funding.

The authors would also like to thank Dr. Robert Taylor and Dr. Clement Alo, professors from the Department of Earth and Environmental Studies in CSAM at MSU and Dr. Vineet Chaoji, research scientist at Amazon for their inputs during various stages of this research.

2. 2. 8 Reference

© 2016 IEEE. Reprinted, with permission, from Xu Du, Aparna Varde, Mining PM2.5 and Traffic Conditions for Air Quality, IEEE International Conference on Information and Communication Systems, Apr 2016, pp. 33 - 38.

[1] S. Zauli Sajani, I. Ricciardelli, A. Trentini, D. Bacco, C. Maccone, S. Castellazzi, P. Lauriola, V. Poluzzi and R. Harrison, “Spatial and indoor/outdoor gradients in urban concentrations of ultrafine particles and PM2.5 mass and chemical components”, *Atmospheric Environment*, Vol. 103, pp. 307-320, 2015.

[2] P. Pant and R. Harrison, “Estimation of the contribution of road traffic emissions to particulate matter concentrations from field measurements: a review”, *Atmospheric Environment*, Vol. 77, pp. 78-97, 2013.

[3] P. Kumar, A. Robins, S. Vardoulakis and R. Britter, “A review of the characteristics of nanoparticles in the urban atmosphere and the prospects for developing regulatory controls”, *Atmospheric Environment*, Vol. 44, no. 39, pp. 5035-5052, 2010.

[4] US EPA (United States Environmental Protection Agency), “Particulate Matter (PM) Standards”, epa.gov, 2015.

[5] US EPA, “Policy Assessment for Review of the Particulate Matter National Ambient Air Quality Standards (NAAQS)”, epa.gov, 2015.

-
- [6] <http://smartcities.ieee.org/>, IEEE Smart Cities, 2015.
- [7] J. Han, M.Kamber and J. Pei, Data Mining Concepts and Techniques, Morgan Kaufmann, San Francisco, CA, 2011.
- [8] <http://data.worldbank.gov>, The World Bank, Data By Country, 2015.
- [9] <http://www.who.int/gho/en/>, World Health Organization, Data Repository, 2015.
- [10] R. Agrawal, T. Imieliński and A. Swami, "Mining association rules between sets of items in large databases", ACM SIGMOD Record, Vol. 22, no. 2, pp. 207-216, 1993.
- [11] J. MacQueen, "Some methods for classification and analysis of multivariate observations", Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol 1, 281—297, 1967.
- [12] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.
- [13] S. Hamidi and R. Ewing, "A longitudinal study of changes in urban sprawl between 2000 and 2010 in the United States". Landscape And Urban Planning, vol. 128, pp. 72-82, 2014.
- [14] Smartgrowthamerica.org, "What is "smart growth?", Smart Growth America", 2015
- [15] Vienna University of Technology (TU Wien), "European Smart Cities", Technical Report, Vienna, Austria, 2015.
- [16] R. Nagy and B. Lockaby, "Urbanization in the Southeastern United States: socioeconomic forces and ecological responses along an urbanrural gradient. urban ecosystems",

Vol. 14, No. 1, pp. 71-86.

[17] H. Miller and J. Han, *Geographic Data Mining and Knowledge Discovery*, Taylor & Francis, London, UK, 2001.

[18] X. Li and A. Gar-On Yeh, “Data mining of cellular automata's transition rules”, *International Journal Of Geographical Information Science*, Vol. 18, no. 8, pp. 723-744, 2004.

[19] U. Rajasekar and Q. Weng, “Application of association rule mining for exploring the relationship between urban land surface temperature and biophysical / social parameters”, *Photogrammetric Engineering and Remote Sensing*, Vol. 75, No. 4, pp. 385-396.

[20] A. Pampoore-Thampi, A. Varde and D. Yu, “Mining GIS Data to Predict Urban Sprawl”, *ACM conference on Knowledge Discovery and Data Mining (KDD)*, New York City, NY, pp. 118-125, August 2014.

[21] X. Du and A. Varde, Mining multicity urban data for sustainable population relocation, To appear in *International Conference on Geo Informatics (ICGI)*, Dubai, UAE, December 2014.

Chapter 3

3. Ordinance Mining

3.1 Urban Legislation Assessment by Data Analytics with Smart City

Characteristics

Abstract: Smart cities receive great attention today especially in conjunction with ubiquitous computing. People feel the need to access information about their cities anywhere anytime. They wish to be actively involved with local government bodies for policy decisions affecting urban lifestyle. Accordingly, this paper describes our research on urban policy management. We analyze urban legislation, more specifically, ordinances or local laws. We categorize ordinances based on smart city characteristics they address. This work deploys data warehousing, XML data management and data mining over categorized ordinances. Interesting findings include relative importance of smart city characteristics considering the focus given by urban agencies. This research helps agencies assess their current ordinance policies with decision support for the future. It also provides urban residents at-a-glance information about their cities and policies with analysis. This work has broader impacts of enhancing smart cities and ubiquitous computing by making useful information widely accessible with suitable inferences.

Keywords: Data Mining; Data Warehousing; Ordinances; Smart City; Urban Policy; XML

Data Management

(Chapter 3.1 reused the previously published paper Du, X., Liporace, D., & Varde, A. (2017), Urban legislation assessment by data analytics with smart city characteristics, *IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, <https://doi.org/10.1109/uemcon.2017.8248972>).

3. 1. 1 Introduction

urban management agencies use local level policy tools to perform and support their own work. In the USA, laws passed by local level jurisdictions are called ordinances. The ordinances enhance and complement federal and state laws. In this paper, we focus on the problem of urban policy assessment based on publicly available ordinance data. The goals of this work are twofold. First, we aim to provide meaningful access to urban policy data for city residents and urban agencies such that they can conceptualize and interpret their legislative activities. The second and more important goal is to evaluate the effectiveness of the urban legislative policies with respect to smart city characteristics they address.

These two goals thus constitute our problem definition. In order to address the first goal herewith, we deploy data warehousing and XML data management strategies to store and exchange local legislature data along with visualization. This helps to make the information easily accessible and understandable in a ubiquitous manner. For the second goal, we apply data mining techniques on the ordinances to extract useful information in urban management for

evidence-based decision-making catering to the characteristics of smart cities. Since data mining consists of techniques used to discover novel and useful knowledge from large data sets, it can be utilized to find interesting patterns. Data mining is thus very helpful for nonlinear relationship analysis [1] which makes it suitable in our work to assess relationships between ordinance data and smart city characteristics.

In this paper we focus on New York City data, since it is the most populated city in the United States [2] and its effective urban management is critical. The NYC Council has 35 standing committees covering many urban administrative aspects [3]. This council has its election every four years. The two most recent full sessions are 2006-2009 and 2010-2013 and each involve the enactment of hundreds of ordinances. Our research provides urban management agencies a unique view of their legislative activity. Connecting each ordinance with one of the six smart city characteristics: *smart living, governance, economy, mobility, people and environment* increases awareness and ability in developing a smart city. By assessing the relationships of ordinances with smart city characteristics and making that easily accessible to the public, this work contributes to urban sustainable development and smart city research with ubiquitous access.

3. 1. 2 Data Description

Legislative data in our research is found on the NYC council website [4]. Details on

ordinances, meetings and committees can be viewed by the general public. Additionally, the website provides an export function to allow users to download parts of the raw data. Users can select the sessions and types of files in the webpage as Figure 3.1 shows. If the ordinance is related to more than one characteristic, we select the closest one.

The screenshot shows the header of the New York City Council website with the speaker's name, Melissa Mark-Viverito. Below the header are navigation tabs for Council Home, Legislation, Calendar, City Council, and Committees. A search bar is present with filters for 'Session 2014-2017' and 'Local Law'. Checkboxes for 'file #', 'text', 'attachments', and 'other info' are visible. A 'Search Legislation' button and a 'Help' link are also shown. Below this is a table with 323 records. The table has columns for File #, Law Number, Committee, Prime Sponsor, Council Member Sponsors, and Title. The first row shows 'Int 0001-2014' with a law number of '2014/007'. An 'Export' menu is open over the table, showing options for 'Export to Excel', 'Export to PDF', and 'Export to Word'.

File #	Law Number	Committee	Prime Sponsor	Council Member Sponsors	Title
Int 0001-2014	2014/007	Introduction - Enacted - Committee on	Marqaret S.	41	A Local Law to amend

Figure 3.1: The Data Source Page

The raw data is as described in Table 3.1. Other data, as seen in Table 3.2 is obtained from the website and added to our system manually. In order to connect smart city characteristics with ordinances, we first need to understand their definition. We utilize the European Smart Cities 4.0 standard [5] which suits NYC conditions most closely. This system divides the smart city characteristics into six big categories with various sub-categories [5]. As Table 3.3 describes, we

associate each individual ordinance with only one of the six smart city characteristics [5] based on its content. We select the most closely related one and proceed with the analysis.

Table 3.1: Raw Data Extracted from Websites

Name of Attributes	Explanation
File #	This is given when the ordinance is initialized
Law Number	This is given when the ordinance is enacted
Committee	The committee that processed this ordinance
Prime Sponsor	The name of main sponsor for this ordinance
Council Member Sponsors	Number of committee members supporting the ordinance
Title	Short introduction of the ordinance

Table 3.2: Additional Ordinances Data

Name of Attributes	Explanation
IniDay / EnDay	Day of the month the ordinance was initialized / enacted
IniMonth / EnMonth	The month the ordinance was initialized / enacted
IniYear / EnYear	The year the ordinance was initialized / enacted
IniDate / EnDate	The initialized / enacted date of the ordinance
TimeSpan	Number of days between initialization and enactment
Smart City Characteristics	The most relevant smart city characteristic
Meeting	Number of meetings held by related committees between initialization and enactment

Table 3.3: Smart City Characteristics

Smart City Characteristic	Contained Concepts
Smart Economy	Innovative spirit, Entrepreneurship, City image, Productivity, Labor Market, International integration
Smart Environment	Air quality (no pollution), Ecological awareness, Sustainable resource management
Smart People	Education, Lifelong learning, Ethnic plurality, Open-mindedness
Smart Living	Cultural and leisure facilities, Health conditions, Individual security, Housing quality, Education facilities, Touristic attractiveness, Social cohesion
Smart Governance	Political awareness, Public and social services, Efficient and transparent administration
Smart Mobility	Local transport system, Transport Management, International accessibility, Sustainability

3. 1. 3 Analytical Methods

We describe the data analysis conducted with respect to data warehousing, XML data management and data mining. This aids storage, exchange and knowledge discovery respectively.

3. 1. 3. 1 Data Warehousing

As a widely accepted definition [6], “A data warehouse is a subject-oriented, integrated, time-variant, non-volatile data collection in support of decision-making processes”. There are

many terms in data warehouse design. Among these, the star schema has fact tables with the actual facts, i.e., core content of data being analyzed; while dimension tables store data on the concerned dimensions, i.e., relevant features [6].

Urban legislation fits the data warehouse model on several levels. The major subject is legislative data. Historical data on past legislative sessions is collected and key structures contain the elements of time (see Table 3.2). Also, data is gathered from multiple sources. OLAP (online analytical processing) supports the decision-making process of targeted users, mainly urban management agencies.

Raw data is collected from NYC council websites and three databases are formed as the foundations of NYC legislation data warehouses. The warehouses in our work are developed using the free open source tool phpMyAdmin [7]. The databases in this warehouse are described next.

Database 1 on Ordinances: This database contains all the related information about each individual ordinance, e.g., the initialization and enactment dates. It has one fact table and three dimension tables. The fact table (Ordinance_Fact) contains three keys: TimeKey, File Key, and ContentKey. Each key points back to a dimension table.

TimeKey → *Time_Dimension*

FileKey → *File_Dimension*

ContentKey → *Content Dimension*

The Time_Dimension table has date related information about ordinances. We store data with details on initialization and enactment dates of ordinances. The File_Dimension table has information specific to file number and file type. The Content_Dimension table has information on the smart city characteristics and the actual content or description of the ordinances. Each dimension table also contains its associated fact table key. This is illustrated in Figure 3.2 below.

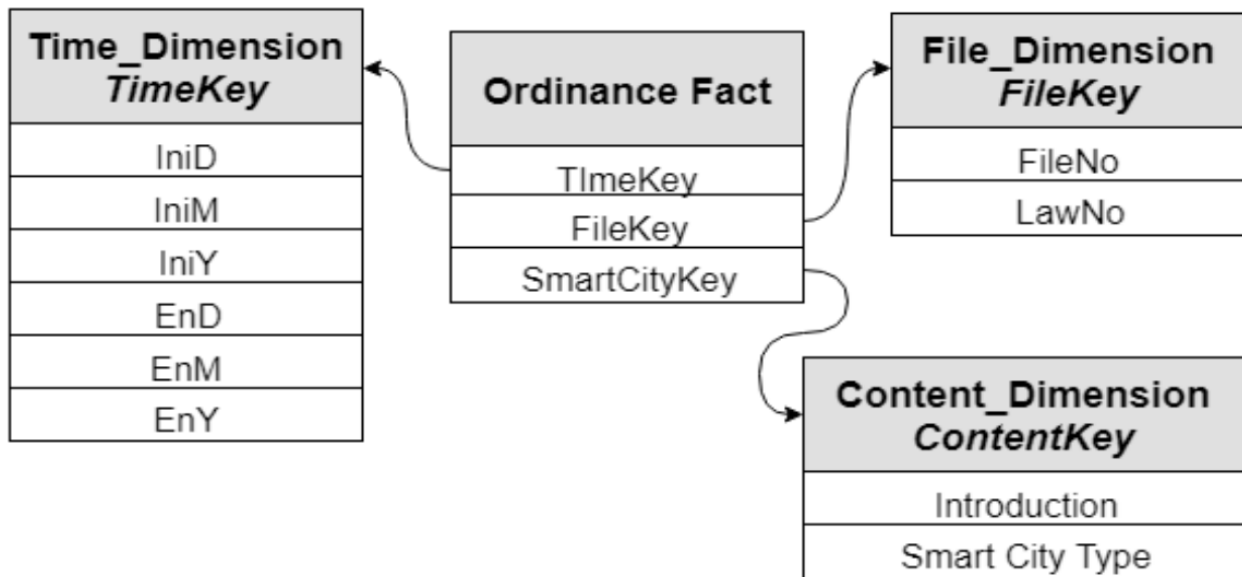


Figure 3.2: Star schema of the ordinances database

Database 2 on Committees: This design has a fact table and two dimension tables. The fact table (Committee_Fact) has two keys: CommitteeKey and MemberKey, each pointing back to a dimension table as shown next. Figure 3.3 is a star schema of the committee database.

CommitteeKey → *Committee_Dimension*

MemberKey → *Member_Dimension*

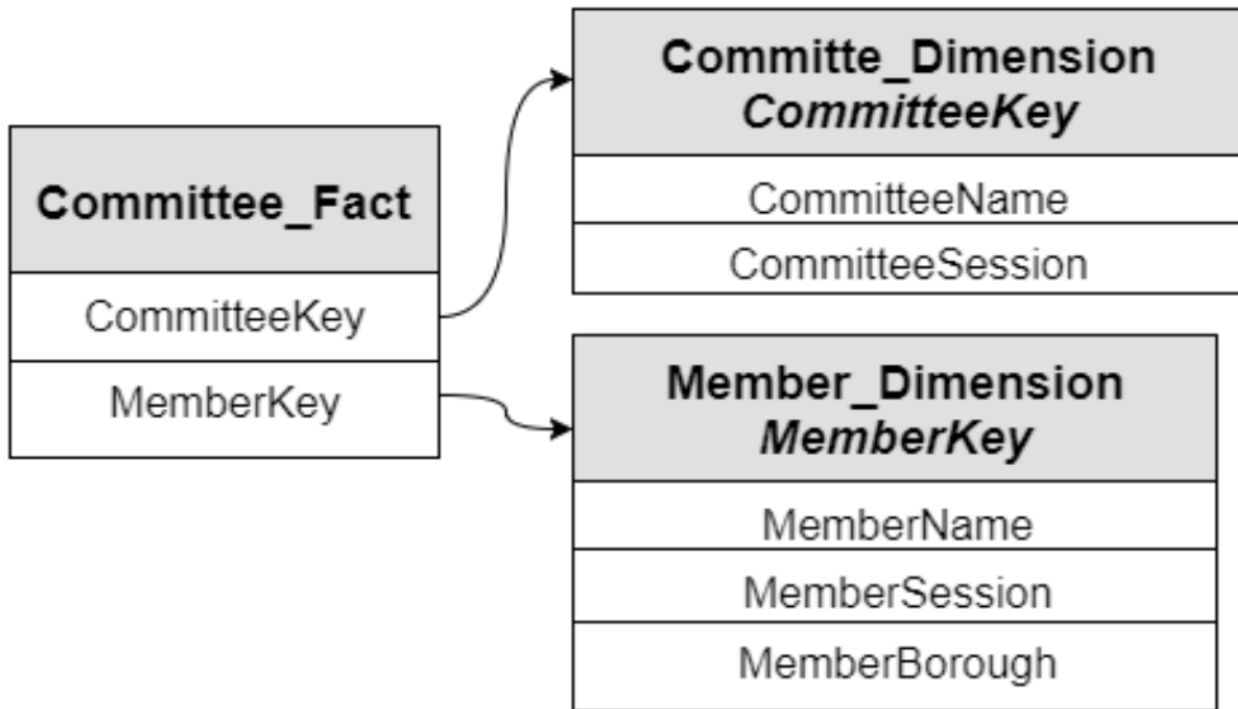


Figure 3.3: Star schema of the committee database

The `Committee_Dimension` table contains the committee name and the committee session.

The `Member_Dimension` contains the names of the committee members, the member sessions and their boroughs.

Database 3 on Meetings: This database has information on meetings held by the council committees such as dates and committees. Figure 3.4 illustrates the star schema of the Meeting Database. Using a star schema provides an excellent view of the original data with the potential of converting to a snowflake schema, i.e., a refinement of a star schema with some dimension tables normalized to reduce redundancy [6]

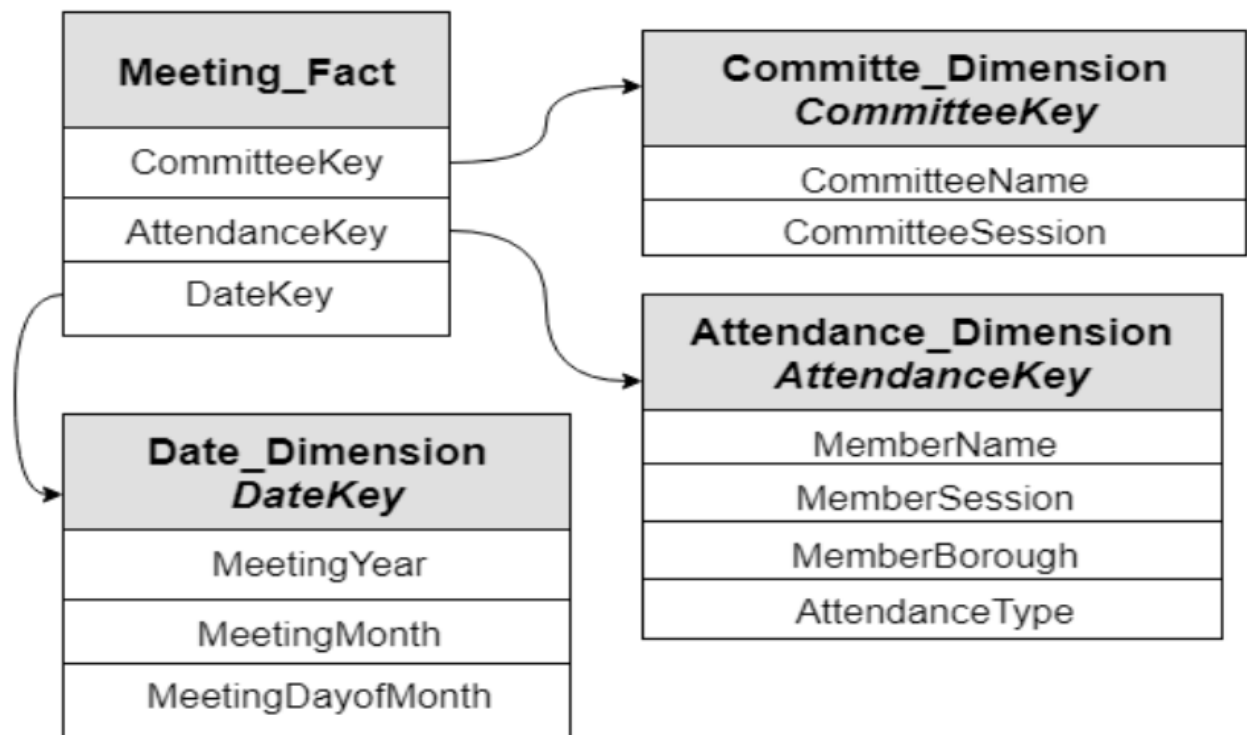


Figure 3.4: Star schema of the meeting database

This can be further enhanced into a fact constellation, i.e., a conceptual model in database design where multiple fact tables share dimension tables [6]. In the example depicted in Figure 3.5, we see relevant parts of the fact constellation for our given star schemas. This structure helps us infer how many ordinances about a given characteristic have been initialized in a certain time frame via the fact constellation functions. Thus, the conceptual modeling of databases through star / snowflake schemas and fact constellations enhance the visualization and at-a-glance analysis of the data for ubiquitous access.

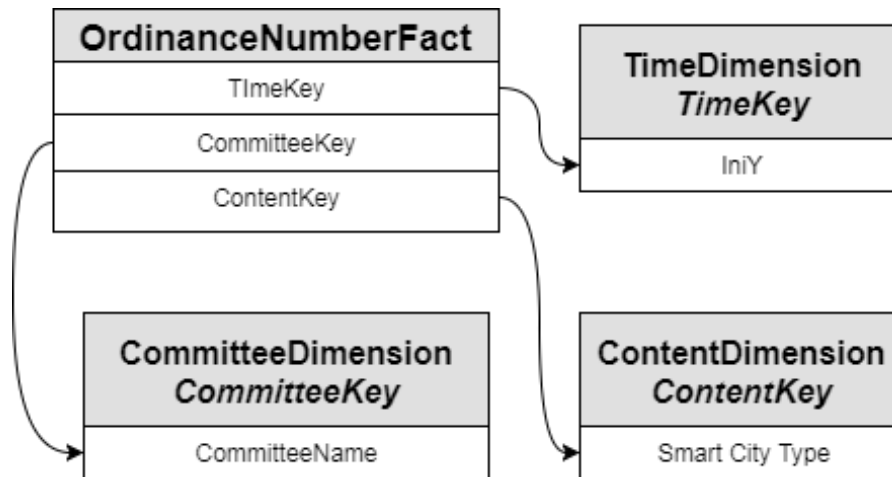


Figure 3.5: Partial snapshot of fact constellation

3. 1. 3. 2 XML Data Management

The eXtensible Markup Language (XML), an industry standard developed by W3C: World Wide Web Consortium, uses tree structures to store data. Descriptive tags called elements capture the semantics. The tags can be extended by adding attributes to include more information. In our work, XMLSpy [8] is used to create three XML databases in the same categories as the data warehouse. These XML DB structures are illustrated next.

XML DB1 on Ordinances: In the XML DB for Ordinances, the root element in the conceptual model is “Ordinance”. There are eight child elements: Initial Time, Enacted Time, Related Committee, Description, Sponsor, Smart City Related Aspects, File Number and Law Number. Initial Time is a complex element and is further divided into Year, Month, and Day. This is illustrated in Figure 3.6.

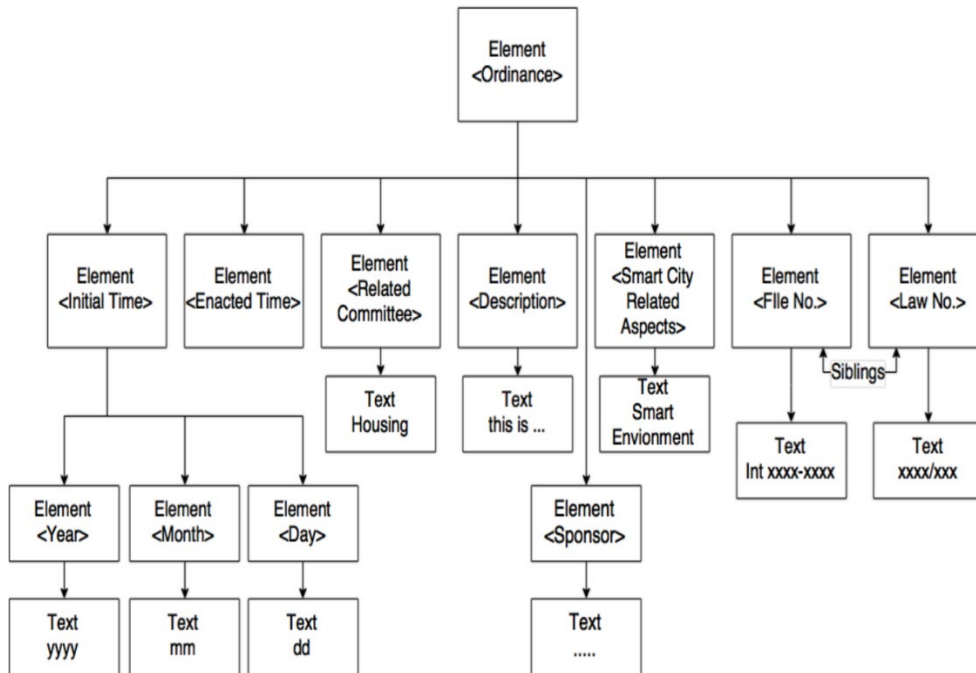


Figure 3.6: XML DB structure for ordinances

XML DB2 on Committees: In the Committees XML DB structure, the root element (conceptual model) is “Committee”. There are three child elements: Committee Name, Session and Member. Member is a complex element, consisting of name attributes used in XML to provide further information. The structure of this XML DB appears in Figure 3.7.

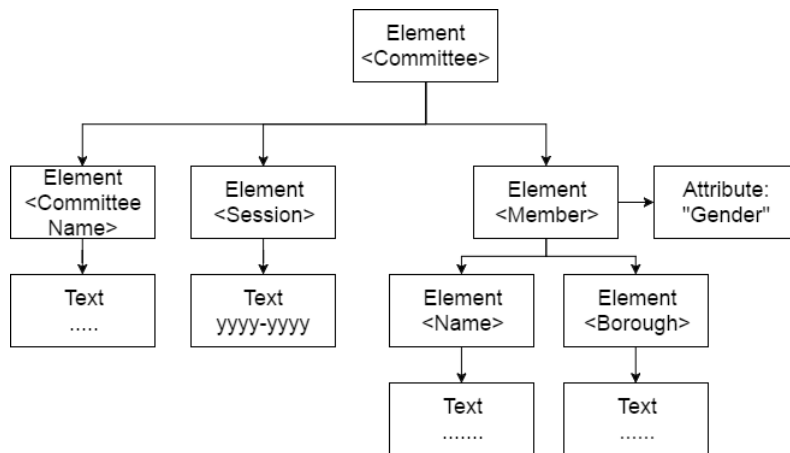


Figure 3.7: XML DB structure for committees

XML DB3 on Meetings: In the Meetings XML DB structure, the root element is “Meeting”. There are four child elements: Date, Committee, Attendance, and Session. Date and Attendance are complex type elements and are further broken down to Year, Month, Day and Attend Member, Absent Member. There can be one or more names associated with a meetings attendance. This is shown in Figure 3.8.

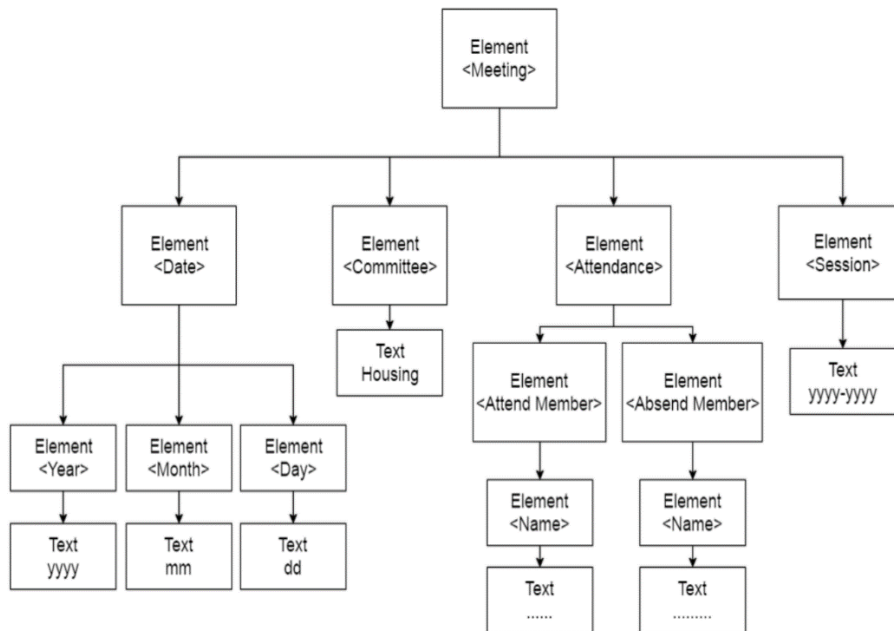


Figure 3.8: XML DB structure for meetings

Compared with data warehousing, XML databases cannot directly provide a very deep view of the data. However, XML storage facilitates worldwide data sharing. It makes the data easily publishable and enhances information exchange through a lingua franca for the Web. XML files can be utilized by different platforms, thus expanding the potential application of the datasets. Hence, for deeper analytical operations on urban legislative data, we prefer data warehouse design while for easy data exchange and publishing among data analysts, legislators and other users, we prefer XML formats. Both these methods enhance the ubiquitous aspects of data storage and processing with respect to urban legislature.

3.1.3.3 Data Mining

The data warehouses and XML databases serve as the basis for data mining by selecting, processing and pre-analyzing the relevant data. In our analysis herewith, we have selected ten significant attributes to perform association rule mining with the classical Apriori algorithm. Apriori is used to discover association rules such as $A \Rightarrow B$ (A implies B) by the analysis of the frequent items in the given data sets. It is useful in our analysis since it helps to discover relationships among various features of legislative activities and corresponding smart city characteristics. In our work we use the implementation of Apriori in the WEKA tool [9]. The ten selected attributes as listed in Table 3.4 are obtained using WEKA filters, further guided by domain knowledge.

Table 3.4: Apriori Data Attributes

Attribute Name	Explanation
IniY	The year of initialization
IniM	The month of initialization
IniD	The day in the month of initialization
EnactY	The year of enactment
EnactM	The month of enactment
EnactD	The day in the month of enactment
TimeSpan	The timespan between initial and enactment
Committee	The committee that processed the ordinance
Meeting	Number of meetings of the respective committee in the given session
SCC	The most relevant smart city characteristic

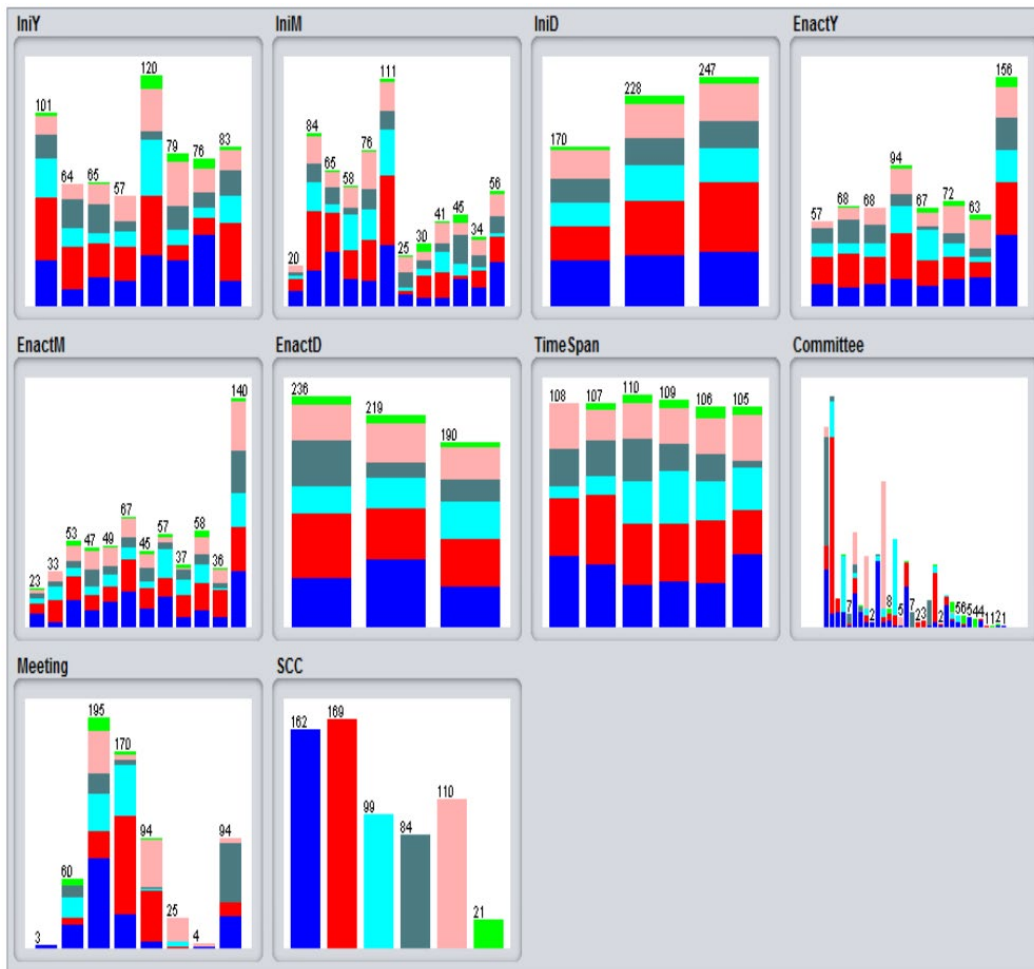


Figure 3.9: Ordinance data after filtering

The visualization of attributes in ordinance data after filtering is shown in Figure 3.9 herewith. By observing this filtered data, we find that the first year of each session usually has the highest number of initialized ordinances while the last year of each session has the highest number of enacted ordinances. Further, the first half of each year within a given session has a higher percentage of initialized ordinances whereas December has the highest percentage of enacted ordinances. Such views of the data are useful for ubiquitous analysis.

3. 1. 4 Results and Discussion

In our analysis with urban legislative data, we build the structured data warehouses and XML databases as described herewith. Thereafter, we generate the relevant ten-attribute table for conducting data mining.

We use rule confidence and rule support as the metrics to discover interesting rules. In a rule $A \Rightarrow B$, rule confidence is the number of times B occurs, given A occurs while rule support is total number of times A and B both occur in the whole data set [9]. Given these details and all the experiments we conduct, association rule mining in our work produces the following interesting rules.

1. *Meeting*='(143.625-inf)' 94 \Rightarrow *Committee*=*Committee on Finance* 94 <conf:(1)> lift:(6.86) lev:(0.12) [80] conv:(80.3)
2. *TimeSpan*='(-inf-39.5]' *Committee*=*Committee on Finance* 35 \Rightarrow *Meeting*='(143.625-inf)' 35 <conf:(1)> lift:(6.86) lev:(0.05) [29] conv:(29.9)
3. *Committee*=*Committee on Sanitation and Solid Waste Management* 34 \Rightarrow *Meeting*='(46.75-66.125]' 34 <conf:(1)> lift:(3.31) lev:(0.04) [23] conv:(23.72)
4. *Committee*=*Committee on Parks and Recreation* 33 \Rightarrow *Meeting*='(46.75- 66.125]' 33 <conf:(1)> lift:(3.31) lev:(0.04) [23] conv:(23.02)
5. *Committee*=*Committee on Transportation* 68 \Rightarrow *SCC*=*Mobility* 60 <conf:(0.88)> lift:(5.17) lev:(0.08) [48] conv:(6.27)
6. *Committee*=*Committee on Environmental Protection* 42 \Rightarrow *SCC*=*Environment* 36 <conf:(0.86)> lift:(5.58) lev:(0.05) [29] conv:(5.08)
7. *Committee*=*Committee on Housing and Buildings* 108 \Rightarrow *SCC*=*Living* 83 <conf:(0.77)> lift:(2.93) lev:(0.08) [54] conv:(3.07) 23
8. *Committee*=*Committee on Transportation* *SCC*=*Mobility* 60 \Rightarrow *Meeting*='(85.5-104.875]' 41 <conf:(0.68)> lift:(4.69) lev:(0.05) [32] conv:(2.56)
9. *Committee*=*Committee on Transportation* 68 \Rightarrow *Meeting*='(85.5-104.875]' 44 <conf:(0.65)> lift:(4.44) lev:(0.05) [34] conv:(2.32)
10. *SCC*=*Economy* 84 \Rightarrow *Committee*=*Committee on Finance* 51 <conf:(0.61)> lift:(4.17) lev:(0.06) [38] conv:(2.11)
11. *SCC*=*Economy* 84 \Rightarrow *Meeting*='(143.625-inf)' 51 <conf:(0.61)> lift:(4.17) lev:(0.06) [38] conv:(2.11)
12. *SCC*=*Economy* 84 \Rightarrow *Committee*=*Committee on Finance* *Meeting*='(143.625-inf)' 51 <conf:(0.61)> lift:(4.17) lev:(0.06) [38] conv:(2.11)
13. *Committee*=*Committee on Transportation* 68 \Rightarrow *Meeting*='(85.5- 104.875]' *SCC*=*Mobility* 41 <conf:(0.6)> lift:(9.49) lev:(0.06) [36] conv:(2.27)

Note that *SCC* here is *Smart City Characteristic*. Measures seen near the rules (e.g. *conf* for confidence) with their values denote experimental parameters in our work. By analyzing these rules, we discover some patterns as listed next.

1. Ordinances of some committees are focused on specific smart city characteristics. Rules 5, 6, 7 and 10 support this claim. For instance, the committee on transportation enacts ordinances mainly about mobility. This corroborates the link between smart mobility and transportation ordinances.

2. Some committees have almost the same number of meetings across two sessions. Rules 1, 3, 4, 8, 9, 11, 12 and 13 corroborate this claim. For example, the committee on finance has over 143 meetings in each session.

3. Ordinances of some smart city characteristics have shorter time spans. Rule 2 favors this claim, e.g., the committee on finance has some ordinances passed in as few as 40 days.

Likewise, other inferences can be drawn from the analysis of this data. Association rule mining in our work helps to relate smart city characteristics to ordinances and to make other observations about certain patterns in the data.

Based on all the detailed ordinance analysis conducted, examples of which are shown herewith, we now summarize the results pertaining to sessions and smart city characteristics. Figure 3.10 illustrates a distribution of ordinances addressed in the two legislative sessions analyzed here, namely, Session 2006- 2009 and Session 2010-2013, with respect to the addressed

six characteristics of smart cities

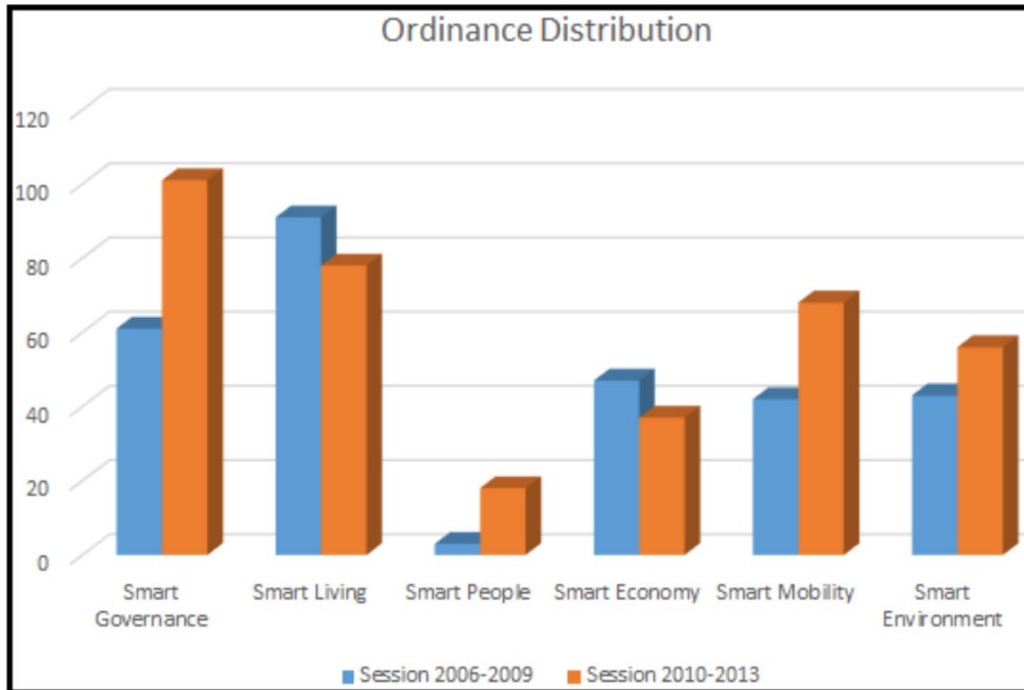


Figure 3.10: Ordinance distribution in sessions as per smart city characteristics

From this figure, we can infer that four characteristics get greater focus in the later session: *smart governance, people, mobility and environment*; while the two characteristics of *smart living and economy* have more focus in the earlier session. Another interesting observation is that the smart city characteristic with maximum attention in the earlier session is *smart living* and that with minimum attention is *smart people*. In the later session, most ordinances addressed are on *smart governance* while the fewest ones are on *smart people*.

Based on all these findings, one suggestion to potentially offer urban management agencies is that they could focus more on issues related to *smart people*, e.g., education, lifelong learning,

ethnic plurality etc. Note that this characteristic does receive a little more attention in the later session than the earlier one, though still less compared to other characteristics. It is to be noted that providing such analysis with visualization makes it very convenient for several urban users including residents and agencies to analyze performance of the legislature at-a-glance in a ubiquitous manner.

Our analysis on ordinance data therefore serves useful as evident from our methods and results. This work helps to conceptualize legislative activities by modeling them with respect to parameters such as ordinances, committees and meetings, thus providing suitable views of data for ubiquitous access by city residents and urban agencies. Furthermore, it helps to assess how much the urban policy legislations lead to smart city development. This analysis, in addition to giving urban residents overviews of how well their legislators perform, also helps the urban management agencies evaluate their own effectiveness and support future decision-making. For instance, as stated earlier, the agencies could start focusing more on legislations that have received relatively less attention on a certain smart city characteristic (smart people here). Conversely, they could continue to give greater focus to those that have already received significant attention on a specific characteristic, hence aiming to make their city among the smartest in that perspective. For example, considering the analysis herewith, urban legislators in NYC could continue to focus more on smart governance ordinances aiming to make NYC among the leaders in that aspect. Since this entails transparency and public involvement in governing the

city, it is corroborated by the fact that the current NYC mayor reaches out to city residents through a mailing list to gather their opinion on various issues. Thus, a continued focus by urban agencies on smart governance would be good.

3. 1. 5 Related Work

Data mining techniques have been applied to urban sustainability research in multiple ways. Earlier research has deployed association rule mining on various types of urban data: i.e. population data, traffic data and social media data [10, 11, 12] and produced reliable results. However, to the best of our knowledge, this paper is among the first to conduct data analytics on urban legislature with various perspectives that include data warehousing, mining and XML processing.

Researchers of legislative events usually try to extract the patterns of legislative activity and establish prediction models. They regard legislators and bills as points in the representative spatial model [13]. The prediction is performed by analyzing relative positions of legislators and bills. Multiple researchers improve the accuracy and data utilization based on this type of prediction model considering legislators' political profiles and other suitable data [14, 15]. These research works produce reliable prediction models for bills with respect to their chance to be passed. However, research about the whole legislative body activity, i.e., types of bills and the number of those types to be passed are not fully addressed. We address these and other issues in

our work with the aim of providing inputs potentially useful to urban agencies.

Smart city research is attracting much interest in recent years. Melbourne in Australia is supposedly among the finest “knowledge cities” [5, 16]. The term *knowledge city* is many times used interchangeably with *smart city*, yet there are distinguishing factors [16]. Emphasis is more on ubiquitous knowledge dissemination in the first term while it is on several smart city characteristics in the second term. In this work, we focus on the knowledge city as well as smart city aspects since we provide ubiquitous access with convenient interpretation and data visualization; while also catering to the individual characteristics of smart cities in our analysis.

There are many smart cities in Europe. For instance, buses from Barcelona in Spain run on routes that maximize energy efficiency [5]. In some cities, e.g., Saarbruecken in Germany, customers are given money back for returning recyclable items such as empty plastic bottles, thus motivating them with respect to the financial as well as environmental factors to head towards smart city goals [17]. Also, solar panels are installed on rooftops in many cities, e.g., Paris in France, as a means to achieve greater energy efficiency mechanism [5, 17]. Our work in this paper adds to such smart city contributions from a computational perspective. It assists in dissemination of useful information to city residents and urban agencies and also helps to assess the effectiveness of urban policies for the smart city paradigm as a whole. Including common sense knowledge [17] to further enhance smart cities from various perspectives is part of our future work.

3. 1. 6 Conclusion

In this paper, we address urban legislative research from a smart city angle. The deployment of data warehousing and XML databases provides convenient sources for data mining of ordinances. Using these methods, researchers can generate multiple data tables for mining. By utilizing the relationship analysis through mining, researchers can extract patterns of legislative activity in datasets. This makes data available for ubiquitous access with easy interpretation and visualization.

Data mining yields valuable knowledge for urban residents to better understand and judge their government policies. It guides urban management agencies by helping them conduct self-assessment, especially with reference to smart cities. This work thus provides a novel method to analyze activities of urban management agencies to support decision-making. It could be further augmented with other research, e.g., public opinion surveys, which constitutes future work. Note that in this research we assign each ordinance to the most relevant smart city characteristic. An improved score system could assign ordinances based on multiple smart city characteristics with relative importance. This could provide enhanced results for decision-making, another aspect of future work.

This paper would be interesting to urban researchers and data analysts. To the best of our knowledge, it is among the first works to conduct data analytics on urban legislative activity

from a smart city angle with data warehousing, XML databases and data mining. It contributes on the whole to the realms of ubiquitous computing and smart city development.

3. 1. 7 Acknowledgments

This work is funded by a Doctoral Research Assistantship for Xu Du by the PhD program in Environmental Management at Montclair State University where Aparna Varde is Doctoral Faculty. We thank Dr. Robert Taylor from our Department of Earth and Environmental Studies for his inputs.

3. 1. 8 References

© 2017 IEEE. Reprinted, with permission, from Xu Du, Diane Liporace and Aparna Varde, Urban Legislation Assessment by Data Analytics with Smart City Characteristics, IEEE Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON, Oct 2017, pp. 20-25

[1] R. Nagy and B. Lockaby, "Urbanization in the Southeastern United States: Socioeconomic forces and ecological responses along an urban-rural gradient", Urban Ecosystems, vol. 14, no. 1, pp. 71-86, 2010.

[2] "New York City (NYC) Population Facts", www1.nyc.gov, 2017, <http://www1.nyc.gov/site/planning/data-maps/nyc-population/populationfacts.page>.

-
- [3] "The New York City Council - Committees", Legistar.council.nyc.gov, 2017, <http://legistar.council.nyc.gov/Departments.aspx>.
- [4] Vienna University of Technology (TU Wien) "European smart cities 4.0", Technical Report, Vienna, Austria, 2015.
- [5] W. Inmon, W., Building the data warehouse, 1st ed. Indianapolis: Wiley, 2011.
- [6] "phpMyAdmin", phpMyAdmin, 2017, <https://www.phpmyadmin.net/>.
- [7] "XMLSpy: The XML Editor", Altova.com, 2017, <https://www.altova.com/xmlspy.html>.
- [8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten, "The WEKA data mining software", ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10-18, 2009.
- [9] X. Du and A. Varde, "Mining Multicity Urban Data for Sustainable Population Relocation", International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol. 9, no. 12, pp. 2441- 2448, 2015.
- [10] X. Du and A. Varde, "Mining PM2.5 and traffic conditions for air quality", IEEE International Conference on Information and Communication Systems, pp. 33-38, 2016.
- [11] X. Du, O. Emebo, A. Varde, N. Tandon, S. Nag Chowdhury and G. Weikum, "Air quality assessment from social media and structured data: Pollutants and health impacts in urban planning", IEEE International Conference on Data Engineering Workshops, pp. 54-59, 2016.
- [12] J. Clinton, S. Jackman and D. Rivers, "The Statistical Analysis of Roll Call Data", American Political Science Review, vol. 98, no. 02, pp. 355-370, 2004.

-
- [13] J. Wang, K. Varshney and A. Mojsilović, "Legislative Prediction via Random Walks over a Heterogeneous Graph", SIAM International Conference on Data Mining, pp. 1095-1106, 2012.
- [14] Y. Cheng, A. Agrawal, H. Liu and A. Choudhary, "Legislative Prediction with Dual Uncertainty Minimization from Heterogeneous Information", SIAM Conference on Data Mining, pp. 361-369, 2015.
- [15] M. de Jong, S. Joss, D. Schraven, C. Zhan and M. Weijnen, "SustainableSmart-Resilient-Low Carbon Eco-Knowledge Cities; Making sense of a multitude of concepts promoting sustainable urbanization", Journal of Cleaner Production, Elsevier, vol.109, pp. 25-38, 2015.
- [16] A. Varde, N. Tandon, S. Nag Chowdhury and G. Weikum, "Common Sense Knowledge in Domain-Specific Knowledge Bases", Tech. Rep., Max Planck Institute for Informatics, Saarbruecken, Germany, 2015.

3. 2. Mining Ordinance Data from the Web for Smart City Development

Abstract: In this research, we aim to discover knowledge from ordinances, i.e., local laws on urban policy. This is useful in policy assessment which we address especially with respect to smart cities. To analyze the publicly available ordinance data from websites guided by human judgment, we use common sense knowledge from a repository called WebChild and its domain-specific knowledge bases in relevant areas, e.g., town planning. Much of this ordinance data maps to smart city characteristics, e.g., smart environment. Hence, based on mining using association rules and other methods, we give feedback to urban agencies for decision support, particularly in a smart city context. To the best of our knowledge, this is among the first works to conduct ordinance mining.

Keywords: *Association Rules; Classification; Common Sense Knowledge; Decision Support; Urban Policy*

(Chapter 3.2 reused the previously published paper Du, X., Varde, A., & Taylor, R. (2017), Mining Ordinance Data From the Web for Smart City Development, In *CSREA Press, International Conference on Data Mining DMIN* (pp. 84-90), Las Vegas, NV).

3. 2. 1 Introduction

Public policy has produced many laws that support the goals of environmental management. *Ordinances*, i.e., local laws at municipal levels, are direct policy tools developed by urban

management agencies and passed by local-level jurisdictions. The legislation and amendment of those laws are interactive and related with local public opinions [1, 2].

Analyzing the relationship between ordinances and conducting related studies would thus support efficient urban management. We address this issue, particularly with the intention of heading towards the development of smart cities. A smart city is typically expected to have the characteristics [3] as shown in Figure 3.11.

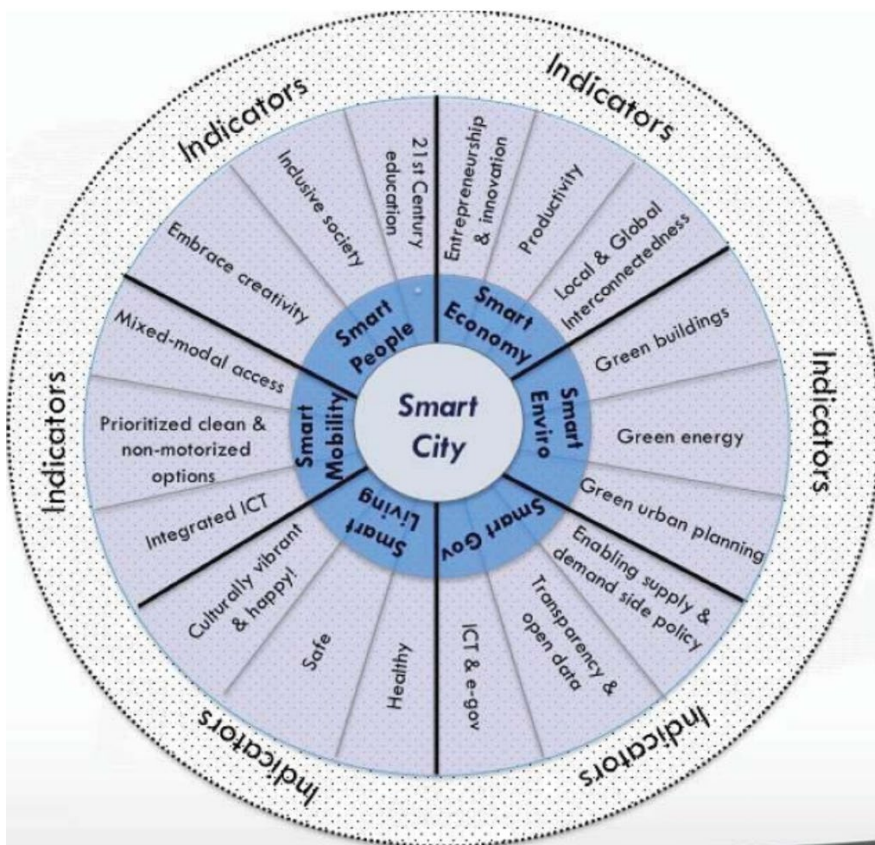


Figure 3.11: Typical smart city characteristics

These characteristics are smart governance, smart environment, smart mobility, smart

living, smart people, and smart economy [3, 4]. *Smart Governance* pertains to government effectiveness, including transparency and public participation in decisions. *Smart Environment* is concerned with energy efficiency, pollution control, sustainable resources etc. *Smart Mobility* focuses on transport issues such as local accessibility with sustainable and safe systems. *Smart Economy* is concerned mainly with competitiveness, innovative spirit, productivity and maintaining cost savings while meeting imperative demands. *Smart Living* deals with public health, safety, housing quality etc. The *Smart People* characteristic entails social and human capital, qualification, creativity and related aspects.

There are many smart cities all over the world. Figure 3.12 shows an example of a smart city Amsterdam in the Netherlands where street lamps allow municipal councils to dim and brighten lights based on pedestrian usage [4]. This would certainly enhance transportation by providing more sustainable systems, thus catering to the smart mobility characteristic of smart cities.



Figure 3.12: Smart city example – Amsterdam

Given this overall framework, the issues in urban policy can also be divided into these different categories in order to address a smart city context. Considering this, our problem goals are as follows.

- Investigate ordinances passed by urban agencies in a given location over multiple time spans based on enactment, initialization and other relevant aspects.
- Gauge the effectiveness of ordinances with respect to urban policy considering the respective smart city characteristics they address

Our source of data for ordinances is public websites. We consider these ordinances over multiple time spans. We aim to conduct mining over the data that can be used to answer questions of interest to urban agencies, e.g., *“Which ordinances in a given year cater to smart environment?”*; *“What is the average time span of an ordinance legislation in a given session?”*;

“Which smart city characteristic has received the greatest attention over all the years?”; “What is the relationship between the initialization and enactment of an ordinance over a certain time period?”; “How have ordinances on smart mobility changed in the last five years?” etc. This would help urban agencies investigate their overall performance and also assess where they stand in developing a smart city, i.e., which characteristics are considered and how they are addressed.

The rest of this paper is organized as follows. Section II describes our approach on mining of ordinances. Section III summarizes the experimental results we obtain. Section IV outlines related work in the area. Section V states the conclusions and ongoing research.

3. 2. 2 Approach for Ordinance Mining

3. 2. 2. 1 Overview of Approach

The approach we deploy for mining location-specific temporal ordinance data is illustrated in Figure 3.13. We propose the use of common sense knowledge (CSK) since it helps in mapping the ordinances to smart city characteristics. For instance, an ordinance on an aspect of energy consumption may not have the words “smart environment” or its terms in Figure 3.11. By using CSK, it is possible to find this mapping and hence enhance the mining on ordinances pertaining to smart city characteristics.

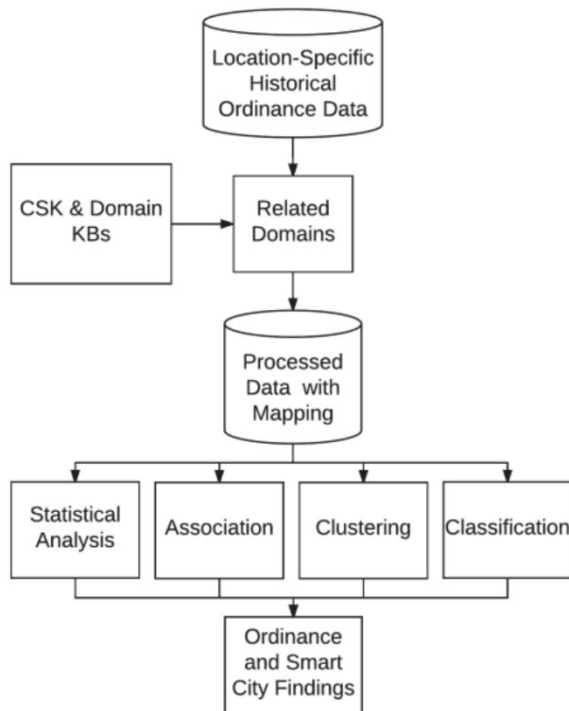


Figure 3.13: Illustration of ordinance mining approach

Data on ordinances obtained from the Web is subjected to processing guided by common sense knowledge. Hence, this is mapped to relevant smart city characteristics. It is then subject to data mining using statistical approaches, association rules, clustering and classification. The knowledge discovered is reported as ordinance and smart city findings and can be used to answer questions useful to urban management agencies for decision support. We now explain its detailed steps.

3.2.2.2 Harnessing Common Sense Knowledge

In order to utilize CSK, we use WebChild, a huge commonsense knowledge base built from

Web contents [5]. It has a browser with which users can search information on real world concepts, their common properties and related terms with pictures. Figure 3.14 is a snapshot of the WebChild browser [5]. This has been used to create domain-specific knowledge bases (domain KBs) in relevant areas [6] with ground truth constituting common sense concepts on urban policy. Given a file with terms from a probabilistic domain classifier, relevant domains are selected and concepts in those domains entered as elaborated in [6] and briefly illustrated in Figure 3.15 and Figure 3.16 respectively.

The screenshot shows the 'WEBCHILD Commonsense Browser' interface. At the top, there is a search bar with the text 'city' and a magnifying glass icon. Below the search bar, the interface is divided into several sections:

- Guess the concept:** A section with a dropdown menu for 'Domain' and other categories like 'Comparable', 'Physical Part', 'Activity', 'Property', and 'Location'. Below the menu is a button labeled 'Ask me!'. To the right of the menu is a small image of a city skyline and a definition: 'a large and densely populated urban area; may include several independent administrative districts; 'Ancient Troy was a great city''.
- TYPE OF:** A section with a dropdown menu set to 'municipality'. Below it, text reads 'Related to location, under the category of town_planning'.
- COMPARABLES:** A section with a row of buttons: 'city,suburb', 'town,city', 'city,asylum', 'mexico,city', 'london,city', and 'More'.
- ACTIVITIES:** A section with a row of buttons: 'leave city', 'enter city', 'build city', 'reach city', and 'see city'.
- HAS PHYSICAL PARTS:** A section with a row of buttons: 'concrete jungle', 'city center', 'financial center', 'civic center', 'inner city', and 'More'.
- IS MEMBER OF:** A section with a row of buttons: 'people', 'company', 'peoples', 'age group', 'ancients', and 'More'.
- IN SPATIAL PROXIMITY WITH:** A section with a row of buttons: 'way', 'place', 'person', 'life', 'land', and 'More'.
- Examples:** A section with a list of examples: 'cup', 'truck,car', 'tiger-n-2', and 'a.drive car'.
- Related Concepts:** A section with a button labeled 'nicaea'.
- Download Dataset!:** A button with a blue background and white text.

Figure 3.14: Snapshot of WebChild browser

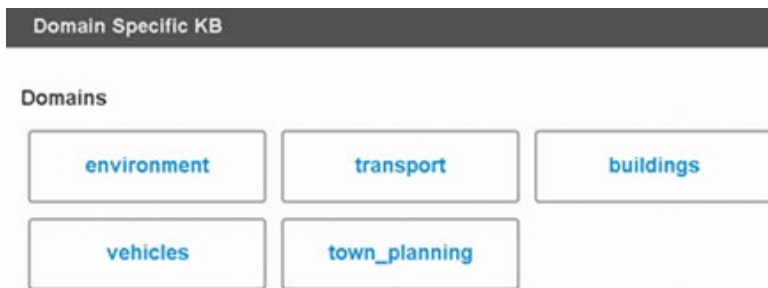


Figure 3.15: Relevant domains selected in KB

Domain : buildings
Concept : apartment complex

CSK

LOCATION	▼	specific zipcode	▼
PARTOF	▼	city, town	▼
SIZE	▼	tall	▼
SHAPE	▼	cuboid	▼
COLOR	▼	white, cream, brick-red;	▼
ACTIVITY	▼	buy or sell apartment, rer	▼
WEIGHT	▼	heavy	▼
MOTION	▼	always stationary	▼
LENGTH	▼	less than 1 block	▼
COMPARABLE	▼	office building, hotel, scho	▼

Figure 3.16: Concepts entered in domains

3. 2. 2. 3 Data Processing and Smart City Mapping

Considering CSK and domain KBs, we proceed as follows. As shown in Figure 3.13, we start with the publicly available ordinance data comprising the location-specific historical information over different time spans. This raw data on ordinances is processed by our program using the terms in WebChild [5] and its related domain KBs [6].

Note that common sense knowledge and related domain specific knowledge bases play a twofold role here. First, they drastically reduce the data set size for mining by selecting only pertinent data on “Related Domains” (see Figure 3.13) and filtering out the rest, thus making the process more effectual. This is done by incorporating relevant CSK and domain KB terms into the programs that access the respective websites to produce the ordinance data in a format suitable for mining. Second, they also map the ordinances to their smart city characteristics analogous to a human, but more efficiently. For example, consider the following ordinance found on a website [2]: “A local law to amend the New York City building code in relation to requiring carbon monoxide detectors in certain apartments is hereby passed.” This would be found relevant to the domain KB on “buildings” with its specific concept being “apartment complex’. Hence, the concerned program would map this to the smart city characteristic of smart living which comprises public health, safety, housing quality etc., taking into account CSK which relates “building” to “housing”. Details of all mappings are not shown herewith but occur on similar

lines, thus helping to generate processed data in the format shown in Table 3.5.

Table 3.5: Processed Data on Ordinances

Time	Ordinance	Related Policies	Smart City Characteristics
MM/DD/YY	Actual content in the website	Energy/ Transportation/	Smart Environment/ Smart Mobility
...

This table depicts a data set for a specific location over multiple timespan. It exemplifies the processed data stored as intermediate output from Web based ordinances and is maintained in databases for further analysis.

3. 2. 2. 4 Deployment of Mining Methods

The processed data sets are subjected to exploratory data mining with statistical analysis [7] including temporal factors, median calculations, minimum and maximum value observations and other aspects. Association rules, clustering and classification are also conducted over the data [7].

We select these data mining techniques for the following reasons.

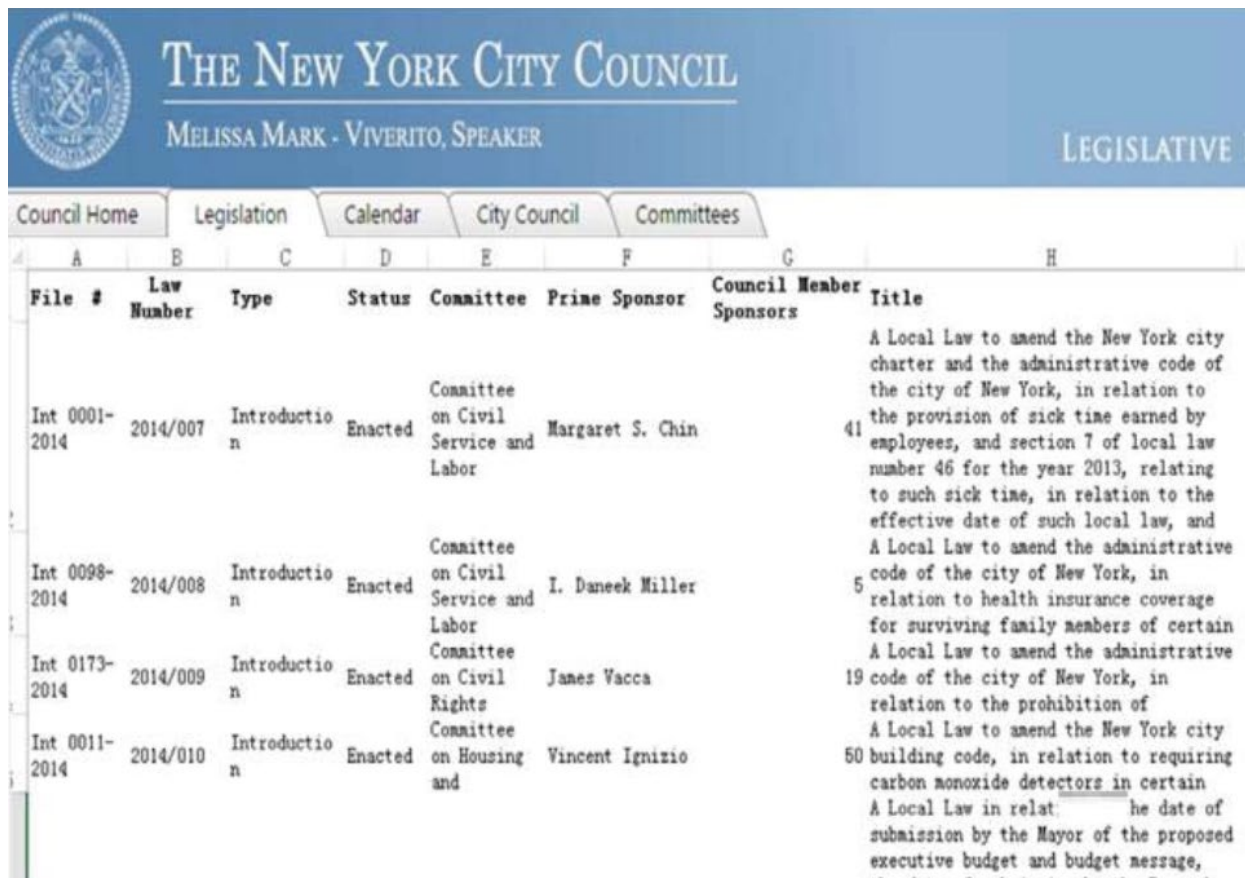
Since association rule mining finds relationships of the type $A \Rightarrow B$, it is expected to be useful in identifying how one feature of the urban policy relates to another. Clustering places items in groups based on their similarity and hence is likely to help in finding similarities among the ordinances by grouping the relevant ones together. Classification predicts a target based on

analysis of existing data and thus is found potentially suitable with respect to categorization. In other words, it would help to specifically categorize an ordinance based on its smart city characteristic addressed.

Outputs provided by data mining are therefore expected to help in understanding relationships between various aspects of the ordinances passed by urban management agencies and in assessing them with reference to smart city characteristics. This would guide decision support for these urban agencies.

3. 2. 3 Experimental Results

We conducted experiments using our approach for ordinance mining as described herewith in the subsections on statistical analysis with clustering; association rules and classification respectively. In this paper, we focused on New York City as the location and considered its council data [2]. We chose NYC since it is the most populous city in the USA, has systematic ordinance data publicly available on the Web and also has many urban policy issues addressed. An example of the NYC council data used in our work appears in Figure 3.17.



File #	Law Number	Type	Status	Committee	Prime Sponsor	Council Member Sponsors	Title
Int 0001-2014	2014/007	Introduction	Enacted	Committee on Civil Service and Labor	Margaret S. Chin	41	A Local Law to amend the New York city charter and the administrative code of the city of New York, in relation to the provision of sick time earned by employees, and section 7 of local law number 46 for the year 2013, relating to such sick time, in relation to the effective date of such local law, and
Int 0098-2014	2014/008	Introduction	Enacted	Committee on Civil Service and Labor	I. Daneek Miller	5	A Local Law to amend the administrative code of the city of New York, in relation to health insurance coverage for surviving family members of certain
Int 0173-2014	2014/009	Introduction	Enacted	Committee on Civil Rights	James Vacca	19	A Local Law to amend the administrative code of the city of New York, in relation to the prohibition of
Int 0011-2014	2014/010	Introduction	Enacted	Committee on Housing and	Vincent Ignizio	50	A Local Law to amend the New York city building code, in relation to requiring carbon monoxide detectors in certain
							A Local Law in relat. he date of submission by the Mayor of the proposed executive budget and budget message,

Figure 3.17: Example of NYC ordinance data

3. 2. 3. 1 Statistical Analysis with Clustering

We conducted exploratory data mining on the legislative activity related to the ordinances by the NYC council from 2006 to 2013 using statistical analysis [7] taking into account the time factor. The time period corresponded to the two latest full NYC city council sessions. We thereby analyzed the distribution of the ordinances by different urban committees over time. We also calculated the days that the city council spent to enact each ordinance. The results of this analysis

are summarized in Figure 3.18. Here the dotted line indicates the ordinances initialized in the respective years and the solid line indicates the ones enacted that year.

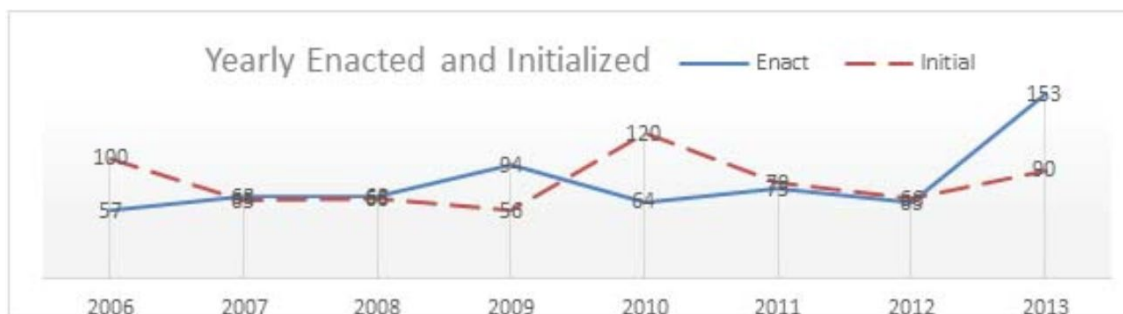


Figure 3.18: Statistical plot of enacted and initialized ordinances

Furthermore, these ordinances were subjected to a simple clustering [7] by grouping them with respect to sessions between 2006-2009 and 2010-2013. The results of this process are visualized in Figure 3.19 and Figure 3.20 respectively. The dotted and solid lines represent the same aspects as in Figure 3.18.

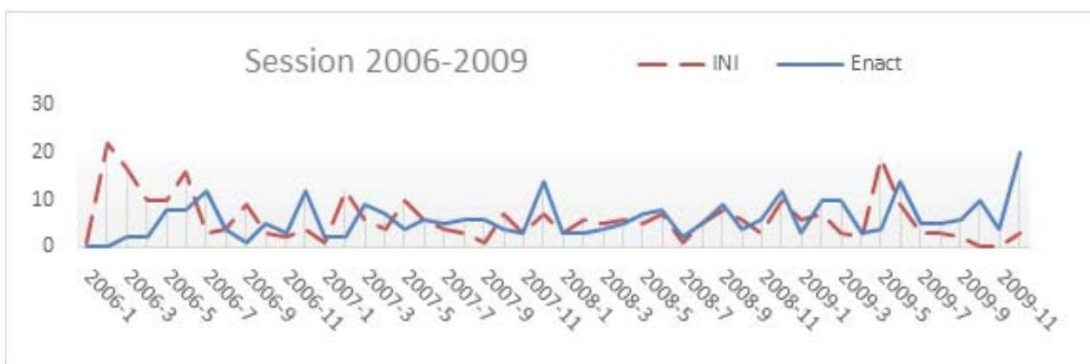


Figure 3.19: Visualization of ordinances clustered from 2006-2009

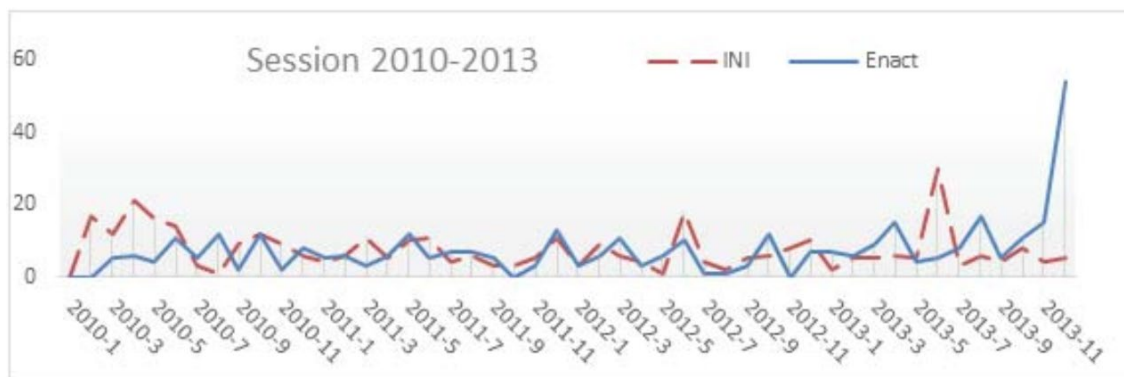


Figure 3.20: Visualization of ordinances clustered from 2010-2013

From this temporal statistical analysis and basic session related clustering we found that the first year of each session had the highest number of initialized ordinances while the last year had the highest number of enacted ordinances. Also, some additional observations from the statistical analysis revealed that the total number of ordinances increased from 287 in earlier the time period of 2006-2009 to 358 in the later time period of 2010-2013. *This indicated that urban agencies passed more ordinances as time progressed.*

Temporal statistical analysis of the legislation time span presented other interesting results. It was found that the average time span of ordinance legislation increased from 204 to 222 days in the two respective sessions. The median number had a drastic change from 89 to 131 days. *These results indicated that, on the whole, ordinance legislation took a longer time in the later sessions.*

Considering the distribution of ordinances by committees, the top three committees of session 2006-2009 were found to be the Committee on Finance, Committee on Housing and

Buildings and Committee on Environmental Protection, the percentages being 17.07%, 16.72% and 9.41% respectively. The top three committees of session 2010-2013 were the Committee on Housing and Buildings, Committee on Finance and Committee on Transportation, the percentages being 16.76%, 12.57% and 12.29% respectively. *This showed an interesting change, i.e., an increase in the number of transportation ordinances being passed in the 2010-2013 session.*

The average legislation time span of the top three committees from 2006-2009 were 89, 192 and 406 days respectively while the time span of the 2010-2013 respective committees were 176, 133 and 283 days. We noticed that the average time span of all the ordinances in those two sessions were 204 and 222 days respectively. This indicated that the ordinances of the Committee on Finance had a much shorter time span of legislation (89d&204d, 133d&222d) respectively while those of the Committee of Environment Protection had a longer time span (406d&204d, 355d&222d) respectively. *Thus, financial legislations were found to be faster while the environmental ones were much slower, probably indicating that there was a significantly greater demand to speed up financial policies.*

3. 2. 3. 2 Association Rule Mining

We conducted association rule mining using the classical Apriori algorithm [7] which follows the principle of frequent item sets and their supersets. The mining was done in two steps.

We first mined the plain committee-based data. This was done by discretizing the time span into ten bins based on equal width and using the Apriori algorithm with suitable parameters.

Examples of rules obtained are shown below.

1. *Committee=Finance* 94 => *TimeSpan='(-Inf-143.6]'* 77 <conf:(0.82) > lift:(1.4) Lev:(0.03) [22] Conv:(2.17)
2. *EnactM=7* 45 => *TimeSpan='(-inf-143.6]'* 35 <conf:(0.78)> lift:(1.33) lev:(0.01) [8] conv:(1.7)
3. *EnactM=5* 49 => *TimeSpan='(-inf-143.6]'* 34 <conf:(0.69)> lift:(1.19) lev:(0.01) [5] conv:(1.27)
4. *EnactM=10* 58 => *TimeSpan='(-inf-143.6]'* 40 <conf:(0.69)> lift:(1.18) lev:(0.01) [6] conv:(1.27)
5. *EnactM=6* 67 => *TimeSpan='(-inf-143.6]'* 46 <conf:(0.69)> lift:(1.17) lev:(0.01) [6] conv:(1.27)
6. *IniM=12* 56 => *TimeSpan='(-inf-143.6]'* 35 <conf:(0.63)> lift:(1.07) lev:(0) [2] conv:(1.06)
7. *IniM=5* 76 => *TimeSpan='(-inf-143.6]'* 46 <conf:(0.61)> lift:(1.04) lev:(0) [1] conv:(1.02)
8. *IniM=3* 65 => *TimeSpan='(-inf-143.6]'* 38 <conf:(0.58)> lift:(1) lev:(0) [0] conv:(0.96)
9. *IniM=6* 111 => *TimeSpan='(-inf-143.6]'* 63 <conf:(0.57)> lift:(0.97) lev:(-0) [-1] conv:(0.94)
10. *IniM=4* 58 => *TimeSpan='(-inf-143.6]'* 32 <conf:(0.55)> lift:(0.94) lev:(-0) [-1] conv:(0.89)

The useful knowledge extracted from here was that the ordinances initialized and enacted during the middle of the year had a relatively short legislation time span. The rules 2, 3, 5, 7 and 8 seemed to support this fact. Also rule 1 supported our previous inference from the statistical analysis that the ordinances passed by the Committee of Finance had a shorter legislation time span.

As a next step, we added more dimensions to the data for association rule mining, i.e., the

relevance to smart city characteristics (pertaining to Table 3.5). Accordingly, we conducted association rule mining on the processed data incorporating the respective smart city terms. It generated a different set of rules, examples of which are shown next.

11. *Committee=Transportation 68 => Concept=Mobility 60 <conf:(0.88)> lift:(5.17) lev:(0.08) [48] conv:(6.27)*
12. *Committee=Environmental Protection 42 => Concept=Environment 36 <conf:(0.86)> lift:(5.58) lev:(0.05) [29] conv:(5.08)*
13. *TimeSpan='(-inf-143.6]' Committee=Housing and Buildings 59 => Concept=Living 47 <conf:(0.8)> lift:(3.06) lev:(0.05) [31] conv:(3.36)*
14. *Committee=Finance Concept=Economy 51 => TimeSpan='(-inf- 143.6]' 40 <conf:(0.78)> lift:(1.34) lev:(0.02) [10] conv:(1.77)*
15. *Committee=Housing and Buildings 108 => Concept=Living 83 <conf:(0.77)> lift:(2.95) lev:(0.09) [54] conv:(3.07)*
16. *TimeSpan='(-inf-143.6]' Concept=Economy 61 => Committee=Finance 40 <conf:(0.66)> lift:(4.5) lev:(0.05) [31] conv:(2.37)*
17. *Concept=Economy 84 => Committee=Finance 51 <conf:(0.61)> lift:(4.17) lev:(0.06) [38] conv:(2.11)*
18. *Concept=Living 168 => TimeSpan='(-inf-143.6]' 101 <conf:(0.6)> lift:(1.03) lev:(0) [2] conv:(1.03)*
19. *Concept=Governance 162 => TimeSpan='(-inf-143.6]' 92 <conf:(0.57)> lift:(0.97) lev:(-0) [-2] conv:(0.95)*
20. *Committee=Housing and Buildings Concept=Living 83 => TimeSpan='(-inf-143.6]' 47 <conf:(0.57)> lift:(0.97) lev:(-0) [-1] conv:(0.93)*
21. *Concept=Mobility 110 => TimeSpan='(-inf-143.6]' 61 <conf:(0.55)> lift:(0.95) lev:(-0.01) [-3] conv:(0.91)*

These rules depicted the relationships between the respective committees and smart city characteristics. It was thus found that the Committee on Transportation, Committee on Environmental Protection, Committee on Housing and Building and Committee on Finance had a strong correlation with smart mobility, smart environment, smart living and smart economy respectively, which was not surprising. These rules also showed the connection between time

span as observed in the statistical analysis with respect to the relevant smart city concepts. It was found that the ordinances related to smart economy and smart living took a shorter time to enact. The ordinances related to smart governance and smart mobility, on the other hand, had a relatively longer time span. *It could thus be inferred that ordinances related to smart economy and smart living were probably found to be more demanding and thus needed faster legislation.*

3. 2. 3. 3 Decision Tree Classification

We conducted classification analysis of the data using J4.8 decision tree classifiers [7]. As is well known in the data mining community, decision trees provide a stem and leaf structure with the stems representing the paths based on attributes of the data and the leaves representing the decisions or the classification targets. The J4.8 algorithm for classification is a Java based extension of C4.5 which follows the principle of entropy in inducing a decision tree given a data set [7]. In our data sets, the classification targets were designed to be the smart city characteristics. A summary of the findings is listed below as observed.

Committee = Committee on Finance: Economy (94.0/43.0)

Committee = Committee on Housing and Buildings: Living (108.0/25.0)

Committee = Committee on Sanitation and Solid Waste Management: Environment (34.0/8.0)

Committee = Committee on Contracts: Governance (10.0/3.0)

Committee = Committee on Parks and Recreation: Mobility (33.0/8.0)

Committee = Committee on Standards and Ethics: Governance (2.0)

Committee = Committee on Governmental Operations: Governance (34.0/3.0)

Committee = Committee on Transportation: Mobility (68.0/8.0)

Committee = Committee on Mental Health, Developmental Disability, Alcoholism,

Substance Abuse and Disability Services: Living (2.0)

The numbers here can be interpreted as follows. Consider the first finding. Here, among the 94 ordinances passed by the Committee on Finance, 43 were classified as addressing the smart economy characteristic of smart cities. Likewise, it can be inferred from the results of the overall classification analysis seen here that various smart city characteristics were addressed to some extent in the urban policy ordinance data mined herewith.

3. 2. 3. 4 Summary of Observations

Based on the ordinance data mining conducted so far, we tabulated the results with reference to the smart city characteristics as shown next. Table 3.6 herewith depicts an overall distribution of the ordinances addressed in each session with respect to the characteristics of smart cities.

Table 3.6: Ordinance Distribution W.R.T. Smart City

Smart City Concept	Session 2006–2009	Percent	Session 2010–2013	Percent
Governance	61	21.25%	101	28.21%
Living	91	31.71%	78	21.79%
People	3	1.05%	18	5.03%
Economy	47	16.38%	37	10.34%
Mobility	42	14.63%	68	18.99%
Environment	43	14.98%	56	15.64%

Thus, we found that the smart city characteristic achieving the greatest attention in the

2006-2009 session was “smart living” while that with the least attention was “smart people”. Likewise, in 2010-2013, the maximum ordinances passed were on the “smart governance” characteristic, while the minimum ordinances were on “smart people”. *Thus, the urban management agencies can potentially be provided with the suggestion that they need to focus more on urban policy issues related to the “smart people” characteristic such as social and human capital, 21st century education etc.* Note that this characteristic did receive somewhat more attention in the later time period than the earlier one, though still significantly less compared to the others.

Based on our experiments we can conclude the following. The data mining on the ordinances does reveal useful information. First, it helps to explore the statistical aspects of the ordinances with respect to time, e.g., trends in the enacted versus initialized ordinances over the years, maximum number of ordinances in a given session etc. Second, it helps to determine how much the urban policy issues head towards developing a smart city. This mining would thus help to answer some questions useful to urban management agencies as stated in the introduction. It would help the agencies assess their effectiveness and enable them to gauge how close they are in catering to smart city characteristics. It also would have the future impact of making them pass ordinances that head towards making their city smarter. Thus, data mining on the ordinances would potentially guide decision support for the urban management agencies in the overall development of smart cities.

3. 2. 4 Related Work

The paradigm of smart cities is receiving tremendous attention today. The city of Melbourne in Australia is considered to be one of the finest “knowledge cities” as gathered from the literature, e.g., [4, 8]. The term knowledge city is often used synonymously with smart city, however, it does have subtle differences [8], the focus being more on ubiquitous knowledge dissemination in the first case versus several smart city characteristics in the second one. Many smart cities are found in Europe catering to several characteristics. For example, buses in Barcelona operate on routes designed to optimize energy efficiency [4]. Recycling is encouraged in some cities by giving customers money back for returning recyclable items such as empty plastic bottles. Solar panels are installed on rooftops in many places as an energy efficiency mechanism.

Consequently smart city research is certainly motivated. A framework for automating implicit requirements in software engineering has been built [9] based on common sense knowledge along with text mining and ontology and has an application in the development of smart city tools. Since these requirements are implicit as opposed to explicit ones, they take into account subtle aspects that users often desire but not state upfront, therefore they are found to be crucial in the adequate functioning of software systems and would be particularly useful in smart city applications, catering to various characteristics. Global sustainability has been addressed

from an agricultural perspective [10]. Issues such as food security; urban versus rural agriculture; and carbon footprints are discussed with the important conclusion that under-used roof space in large global cities can be used to grow food. This heads towards making a city smarter by more effective use of resources for meeting population needs. In our work in this paper we are in line with these general paradigms, addressing the specific issue of urban policy ordinances.

Computational analysis and data mining have helped significantly in geographic studies. Nagy et al. [11] analyze urban and rural gradients in the USA. They consider various social, economic and environmental aspects along with some relevant responses from an ecological perspective. Many of these have been found useful in geographic data analysis. Pampoore-Thampi et al. address the issue of predicting urban sprawl based on data in geographic information systems (GIS) [12]. They estimate factors causing urban sprawl considering the state of NY with sprawl affected areas over different time spans. They consider factors such as population, employment and transportation with respect to the bidirectional impact on sprawl. In [13], assessing air quality by mining data on fine particle pollutants and related attributes is conducted, especially with respect to public health and safety standards recommended by EPA, the Environmental Protection Agency of the USA. Our work in this paper falls under the same broad realm. We conduct mining on Web based temporal and location-specific urban policy ordinance data with respect to the characteristics of smart cities, use common sense knowledge in the overall process and aim to provide inputs for urban management agencies based on the

results.

Researchers have conducted several studies to test the capability of social media mining and sentiment analysis, such as preferences for candidates, interests on certain topics or goods and political opinions [14]. Often, positive correlations have been found between the mining results and reality outcomes. On the other hand, some researchers criticized the method [15] indicating that social media mining and sentiment analysis is not perfect and still has room for improvement.

However, the critics have still asserted that this will be a very good complement to the traditional methods. The supporters as well as the critics agree that full-fledged user surveys in the real world are extremely time-consuming. Therefore, extracting useful knowledge from public opinion expressed in cyberspace seems a better alternative. We intend to address this in our future work by mining social media data on public reaction to ordinances. This would help to assess public satisfaction on urban policy issues through opinion mining, thus providing additional suggestions for decision support.

3. 2. 5 Conclusion

In this paper, we mine data on location-specific ordinances over different time spans in order to assess the effectiveness of urban policy in a smart city context. We deploy common sense knowledge along with related domain specific knowledge bases for selecting pertinent

ordinances, and also for mapping them to the concerned smart city characteristics.

The analysis conducted in our work would potentially help in answering questions to guide urban management agencies in decision support for urban policy in general and especially for the development of smart cities. To the best of our understanding, this paper is among the first works to perform data mining on urban policy ordinance data in particular, thereby presenting interesting applied research.

- A few interesting findings from this research are listed herewith with respect to NYC ordinance mining.
- Urban agencies passed more ordinances during the 2010-2013 time span than during 2006-2009, hence indicating an increase in the need for urban policies as time progressed.
- Finance-related ordinances were passed in the shortest span of time, thus implying a greater focus on speeding up policies in “smart economy” so far.
- Ordinances initialized and enacted around the middle of a year seemed to progress faster in legislation, thereby providing a potentially useful suggestion to urban agencies to pass more ordinances around that time to ensure faster progress in the future.
- The smart city characteristic receiving the least attention in both sessions was on “smart people”, which could serve as an input to urban agencies to give greater attention to its aspects such as 21st century education.

-
- The characteristic of “smart living” got the maximum attention in the 2006-2009 session, but dropped to 2nd place in the 2010-2013 session having fewer ordinances on it passed than in the earlier session, thus offering a potential suggestion to give it more importance unless the public seems really satisfied.
 - The “smart governance” characteristic topped the list overall, receiving greater attention in the 2010-2013 session with 101 ordinances passed, which is a good observation and should be well-maintained by urban agencies henceforth.

Note that these observations and the related suggestions are intended to support the future decisions of the urban management agencies while helping them assess their current performance. The disclaimer is that the analysis in this paper does not actually translate to making decisions for these agencies, it would be up to their discretion. However, these suggestions would help in heading more towards smart cities, further corroborated with public opinion as needed.

It is also to be noted that the methods in our analysis here relate each ordinance with just one smart city characteristic. While this discovers interesting knowledge, it presents some limitations, as many ordinances could potentially relate to multiple smart city characteristics. This would be addressed in future work. An integration of clustering and association rule mining could probably be helpful here.

Future work would also include mining social media data to discover knowledge from

public opinion on ordinances. This would constitute sentiment analysis to assess the satisfaction of the public on urban policy in a general context and with particular emphasis on smart city characteristics. All this work is geared towards decision support for urban management agencies, especially in providing inputs to build and enhance smart cities.

3. 2. 6 References

- [1] *Law vs Ordinance | Difference Between*, Differencebetween.info, 2016
- [2] *New York City Council*, Legistar.council.nyc.gov, 2016
- [3] *Smart Cities Council | Definitions and Overviews*, Smartcitiescouncil.com, 2015.
- [4] Vienna University of Technology (TU Wien), *European Smart Cities*, Technical Report, Vienna, Austria, 2015.
- [5] N. Tandon, G. de Melo, F. Suchanek, G. Weikum, “WebChild: Harvesting and Organizing Commonsense Knowledge from the Web”, *ACM WSDM*, Feb 2014, pp. 523-532.
- [6] A. Varde, N. Tandon, S. Nag Chowdhury, G. Weikum, *Common Sense Knowledge in Domain-Specific Knowledge Bases*, Technical Report, Max Planck Institute for Informatics, Saarbruecken, Germany, Aug 2015.
- [7] J. Han, M. Kamber and J. Pei, *Data Mining Concepts and Techniques*, Morgan Kaufmann, 3rd Edition, © Elsevier 2012.
- [8] M. de Jong, S. Joss, D. Schraven, C. Zhan and M. Weijnen, “Sustainable-Smart-

Resilient-Low Carbon Eco-Knowledge Cities; Making sense of a multitude of concepts promoting sustainable urbanization”, *Journal of Cleaner Production*, Elsevier, Dec 2015, 109:25-38.

[9] O. Emebo, A. Varde, D. Olawande, “Common Sense Knowledge, Ontology and Text Mining for Implicit Requirements”, *DMIN*, Jul 2016, ISBN: 1-60132-431-6, CSREA Press, pp. 146 – 152.

[10] R. Taylor, J. Carandang, C. Alexander, J. Callega, “Making Global Cities Sustainable: Urban Rooftop Hydroponics for Diversified Agriculture in Emerging Economies”, *OIDA International Journal of Sustainable Development*, Dec 2012, 5(7):11-28.

[11] R. Nagy, B. Lockaby, “Urbanization in the Southeastern United States: Socioeconomic Forces and Ecological Responses along an UrbanRural Gradient”. *Journal of Urban Ecosystems*, 2010, 14(1): 71-86.

[12] A. Pampoore-Thampi, A. Varde, D. Yu, “Mining GIS Data to Predict Urban Sprawl”, *ACM KDD (Bloomberg Track)* Aug 2014, pp. 118-125.

[13] X. Du, O. Emebo, A. Varde, N. Tandon, S. Nag Chowdhury, G. Weikum, “Air Quality Assessment from Social Media and Structured Data”, *IEEE ICDE Workshops*, May 2016, pp. 54-59.

[14] S. Baccianella, A. Esuli, F. Sebastiani, “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining”, *LREC*, May 2010, 10:2200-2204.

-
- [15] F. Diaz, M. Gamon, J. Hofman, E. Kacaman, D. Rothschild, “Online and Social Media Data as an Imperfect Continuous Panel Survey”, *PLOS ONE*, 2016, 11(1):e0145406.

Chapter 4

4. Social Media Text Mining

4.1 Air Quality Assessment from Social Media and Structured Data

Abstract: This paper describes our work on mining pollutant data to assess air quality in urban areas. Notable aspects of this work are that we mine social media and structured data in a domain-specific context, incorporate commonsense knowledge in mining media opinions and focus on the urban planning domain in a multicity environment. The results of mining are useful for predictive analysis in urbanization. A significant contribution is that we provide useful information on urban health impacts.

Keywords: *Air Pollution, Commonsense Knowledge, Health Impacts, Opinion Mining, Predictive Analysis, Urban Planning*

(Chapter 4.1 reused the previously published paper Du, X., Emebo, O., Varde, A., Tandon, N., Chowdhury, S., & Weikum, G. (2016), Air quality assessment from social media and structured data: Pollutants and health impacts in urban planning, *IEEE 32nd International Conference on Data Engineering - Workshops (ICDE Workshops)*, <https://doi.org/10.1109/icdew.2016.7495616>).

4. 1. 1 Introduction

The quality of air in urban regions is important with respect to health impacts. A significant aspect of air quality is the presence of pollutants and their effects on human health [1]. Given this, an important sub-problem in our work is to mine real data on pollutants from *structured repositories* to assess air quality. We propose an approach entailing the classical data mining paradigms of association rules, clustering and classification for this purpose.

Another important aspect today is public reaction typically expressed through social media. Opinions entered by urban residents on sites such as Twitter give an idea of user satisfaction. This brings us to another interesting sub-problem, i.e., mining *social media* data on pollutants to assess air quality. One of the biggest challenges here is to review relevant information intuitively as a human would. We thus incorporate commonsense knowledge [2] in this process and develop domain-specific knowledge bases in order to guide the social media mining. We also incorporate lexical databases [3] of words with sentiments to mine public opinions.

The results of these mining processes can be used to help urban residents plan lifestyles, assist government bodies in urban policies and give inputs to environmental scientists for research. Accordingly, we conduct predictive analysis based on the results of mining. The broader impact of this work includes developing *smart cities* catering to the *smart environment* characteristic [4] by monitoring air quality, enhancing greenness and improving health. Domain

KBs developed here can be useful in *smart governance* [4] by promoting automation and providing at-a-glance information for decision support. To the best of our knowledge, this is one of the first works to incorporate structured data mining and public opinion mining for urban planning.

4. 1. 2 Mining Structured Data on Pollutants

4. 1. 2. 1 Background and Goals

In the first sub-problem, we focus on mining pollutant data. More specifically, we consider fine particle pollutants PM_{2.5} (Particulate Matter, diameter < 2.5 μm). Finer pollutants are worse as the human respiratory system cannot easily filter them [1]. High PM_{2.5} concentration could cause severe health problems; long term exposure to it could lead to cardiovascular and respiratory diseases, genotoxicity, mutagenicity and cancer. Since PM_{2.5} has highly negative effects, it is desirable to avoid it, thus it is *smart* to live in a *city* with negligible PM_{2.5} concentration [1]. A major source of PM_{2.5} is traffic in urban areas. Hence, we collect real data on traffic conditions from structured sources and mine it with the following goals:

- Analyze the causes of PM_{2.5} occurrence in air based on multicity traffic conditions
- Predict the impact of PM_{2.5} presence on air quality with respect to health standards

4.1.2.2 Data and Standards

We propose to use the AQI (Air Quality Index) by EPA (Environmental Protection Agency, USA) [5] as *ground truth*. This is because it is a widely accepted global standard and is recommended by experts in Environmental Management for health impacts. This is shown in Table 4.1. For example, an index of 401-500 implies that PM_{2.5} concentration is between 350.5 and 500 $\mu\text{g}/\text{m}^3$. This is “Hazardous” for health. Note that color coding is significant (e.g., green: good, red: unhealthy).

The structured data sources for PM_{2.5} used here are from WHO (World Health Organization) [6] and World Bank [7]. The time frame of this data is mainly the last ten years and the geographic scope is worldwide. Attributes analyzed are: *Region, Income Group, Diesel Consumption, Gasoline Consumption, Road Density, Cars per k people, Vehicles per k people, Vehicles per km and PM_{2.5} Range ($\mu\text{g}/\text{m}^3$)*. *Region* is the area analyzed, e.g., East Asia, Middle East etc. *Income Group* is categorical: it considers OECD (Organization for Economic Cooperation & Development) countries and others.

Table 4.1: AQI Standards for Health based on Pm2.5

AQI Category	Index Value	Breakpoints (mcg/m³, 24- hour average)
Good	0-50	0-12
Moderate	51-100	12.1-35.4
Unhealthy for sensitive groups	101-150	35.5-55.4
Unhealthy	151-200	55.5-150.4
Very Unhealthy	201-300	150.5-250.4
Hazardous	301-400	250.5-350.4
	401-500	350.5-500

4. 1. 2. 3 Approach and Experiments

We propose an approach of combined analysis with classical mining paradigms. We deploy Apriori for association rules, k-means for clustering and decision trees for classification.

We mine association rules with Apriori, as we need to study potential impact of parameters on each other. For this, we discretize numeric data with equal frequency binning. After discretizing continuous data into ranges, we assign categorical values to a few variables, e.g.,

“high”, “low” etc. for gas consumption using domain-specific mapping [5]. After running experiments with Apriori, we get useful inferences. There are rules showing that income groups could influence other traffic conditions. This is reasonable as economic conditions affect traffic facility construction. It is also found that high diesel consumption is not directly related to high concentration of PM2.5 in air. Examples of interesting rules are shown below.

Region=Europe & Central Asia Vehicles_Per_KM=VERY LOW => PM25_Class=GOOD
conf:(1)

Gasoline_Consumption=VERYLOW Road_Density=VERY LOW
Cars_Per_K_People=LOW => PM25_CLASS=MODERATE conf: (0.91)

The terms GOOD and MODERATE, pertain to the PM2.5 ranges with respect to their impact on air quality index (see Table 4.1). For example, PM2.5 class = GOOD implies that the resulting AQI category is good since its index value is in the safe range of 0-50, which occurs with PM2.5 concentration of 0.0 to 12.0 µg/m³. Likewise, we can interpret the other ranges.

Clustering is performed with k-means, an algorithm well-suited to numerical attributes, as found in this data set. We disregard the Region attribute here to avoid obvious clusters. An example of experimental results with clustering is shown in Table 4.2. The numbers in brackets are the number of items in each cluster. We note a few interesting observations as listed next.

- Cluster 0 has relatively low traffic indicators, yet its PM2.5 range is not within safe standards

- Income of Cluster 0 is the lowest
- Cluster 2 has the highest PM2.5 concentration, yet it is not the highest traffic indicator
- Countries in Cluster 2 may have other significant PM2.5 sources or poor regulation of car emission
- Cluster 1 and cluster 3 both have the PM2.5 within safe standards and are OECD countries

Table 4.2: Partial Snapshot of Clustering

Attribute	Cluster0 (58)	Cluster1 (36)	Cluster2 (26)	Cluster3 (22)
Income Group	Upper Middle	High: OECD	High: non-OECD	High OECD
Diesel Consumption	108.63	416.61	208.7	266.42
Gasoline Consumption	96.9	341.07	286.03	186.47
Road Density	39.33	140.83	149.42	97.51
Cars per k people	120.54	493.28	234.14	290.31
Vehicles per k people	151.99	588.38	288.69	345.04
Vehicles per km	37.9	50.23	86.21	27.88
PM2.5 Range	15.12 - 18.43	-inf - 5.85	21.76 - inf	5.85 - 11.98

In these observations, it is significant that high gas consumption does not associate with high PM2.5 concentration. In fact, *medium gas consumption is associated with higher PM2.5 concentration*. With further analysis, this can be reasoned as:

- High gas consumption usually associates with better economic conditions and better pollutant regulations

- The *Income* attribute is also significant
- High income groups & high gas consumption groups have better regulatory facilities, so PM2.5 concentration does not rise much

Decision tree classification is conducted with J4.8, the Java version of the classical C4.5 algorithm, to inductively learn a decision tree from categorical attributes. This is useful because we aim to learn potential causes of the *PM2.5 range*, which thus forms the classification target. Mapping from numeric to categorical attributes is done in a manner similar to that for association rules. A partial snapshot of results is shown below. It is found that the *Region* attribute has the strongest influence here. It is also discovered that PM2.5 pollution is highly associated with local conditions.

```

Region = East Asia & Pacific
| Gasoline_Consumption <= 427.7
|| IncomeGroup = High income: nonOECD: '(18.43- 21.755]' (2.0)
|| IncomeGroup = High income: OECD: '(21.755-inf)' (2.0)
|| IncomeGroup = Low income: '(18.43-21.755]' (2.0/1.0)
|| IncomeGroup = Lower middle income: '(11.98-15.12]' (2.0)
|| IncomeGroup = Upper middle income
||| Diesel_Consumption <= 114.38: '(21.755-inf)' (2.0)
||| Diesel_Consumption > 114.38: '(11.98-15.12]' (2.0)
| Gasoline_Consumption > 427.7: '(-inf-5.845]' (5.0)

```

Thus, we have analyzed the causes of PM2.5 occurrence in air based on traffic conditions, which caters to the first goal of this sub-problem. The results of this are used for predictive analysis to address the second goal, i.e., predicting the health impact of PM2.5 on air quality, as elaborated in Section 4.1.4.

4. 1. 3 Opinion Mining on Pollution from Social Media

4. 1. 3. 1 Motivation and Problem Definition

Opinion mining or sentiment analysis deals with automated discovery of knowledge about public reactions from sites such as weblogs, review pages etc. This is important to assess user satisfaction. It motivates us to mine social media based on entries relevant to our issue, i.e., pollution and air quality. We focus on Twitter here, since it is a micro-blogging site with concise information. Thus, the goals of this sub-problem are:

Analyze tweets on pollutants and related terms to discover knowledge useful in air quality assessment

Use the discovered knowledge to predict potential health impacts in the context of urban planning

4. 1. 3. 2 Proposed Methodology

We propose a two-phase approach for opinion mining. Phase 1 involves developing domain-specific knowledge bases (domain KBs) bootstrapped from Commonsense Knowledge (CSK). These provide the *background knowledge* to classify domain specific information. This background knowledge comprises the concepts and instances (named entities) within our domain. Phase 2 involves a domain-specific tweet crawler using the background knowledge of

phase 1 (e.g., spotting concepts and instances in a tweet), and analyzing sentiments in the crawled tweets, followed by data visualization.

1) *Developing Domain-Specific Knowledge Bases*: We propose using domain KBs to employ background knowledge like an expert. The KB creation is outlined in the steps below.

a) *Harnessing Commonsense Knowledge*: Humans possess the ability to tell apart relevant content (in our case, relevant tweets) due to CSK. On the other hand, machines do not possess such knowledge. We propose to provide this background commonsense knowledge through a large, automatically mined commonsense knowledge repository, WebChild [2], which contains commonsense facts about concepts. WebChild provides a mapping from a domain to concepts and commonsense properties of these concepts.

b) *Slicing WebChild*: WebChild comprises a large list of domains (illustrated in Table 4.3) however, we require a subset of these domains. We thus manually specify a smaller list of WebChild domains that are relevant to our context (*urban planning*). This is depicted in Table 4.4. It is conceivable to automate this process via a *probabilistic domain classifier* [1, 8] to derive a subset of domains, but would be an overkill for our usecase herewith. Thus, are now left with a slice of WebChild that contains concepts relevant to *urban planning*.

Table 4.3: Potential List of Domains (Partial Snapshot)

acoustics	administration	agriculture	anatomy	animals
archery	architecture	Art	astrology	aeronautics
biology	banking	buildings	chemistry	cinema ...

Table 4.4: Curated List of Relevant Domains for KB Slicing

environment	transport	buildings	vehicles	town_planning
--------------------	------------------	------------------	-----------------	----------------------

c) Curating the sliced WebChild: The selected domains provide us with a list of concepts for the given domain (e.g., *pollutant* for the domain *environment*). The sliced WebChild can be incomplete or noisy for certain concepts. We curate this slice of WebChild by designing a smart GUI (see Figure 4.1) that assists the curator by automatically proposing relevant attribute values. For example, using the WebChild knowledge, the GUI knows that *small* is a size and that a pollutant is comparable to a toxin. Figure 4.1 shows an example of curation for the concept *pollutant* in the domain *environment*. As discussed in [8], this curated knowledge about our *urban planning* domain is used to propose relevant Wikipedia categories. These Wikipedia categories lead to the Wikipedia entries where the categories appear, enabling the compilation of encyclopedic entries for concepts, e.g., PM2.5.

Domain : environment
Concept : pollutant

Common Sense Encyclopedic Social

SIZE	▼	small,
WEIGHT	▼	light,
LOCATION	▼	air
PARTOF	▼	smoke, vehicle emissions
ACTIVITY	▼	spoil the air, damage health
EMOTION	▼	discomfort
COLOR	▼	multicolored
TASTE	▼	inedible
SMELL	▼	disgusting
COMPARABLE	▼	toxin, allergen

Submit Common Sense Entries

Figure 4.1: Example of populating domain specific KB

d) From domain KB to tweets: We propose a mapping from *Commonsense Concept Classes* → *Wiki Categories* → *Wiki entries* → *Hashtags* to set the stage for mining social media [8].

In essence, we spot the presence of a domain relevant encyclopedic entry (e.g., PM2.5) or a domain relevant commonsense concept (e.g., pollutant) in a tweet’s hashtag which highlights the main topic or subject of a tweet. If there is an overlap of the tweet’s hashtag in our domain vocabulary, we consider that the tweet is relevant to our domain. As explained in [8], it is conceivable to make a more sophisticated model (e.g., a language model over our domain) that estimates whether a given tweet can be generated by the language model representing the big context (urban planning). Note that besides being useful in opinion mining from social media,

domain KBs can be helpful in broader settings. This includes giving inputs to smart cities for smart environment and smart governance; utility in machine learning to automate various learning processes; and providing domain knowledge to mine visual commonsense from multimodal content [4, 8].

2) Building a Sentiment Analyzer: Using the domain KBs, NLP and other resources, the analyzer is built as follows.

a) Tweet Collection with Hashtags: To collect tweets, we use a Twitter API and a script written in Python. The Twitter API gives us access to user tweets using the OAuth, while the Python script collects tweets with keyword combinations and hashtags. These hashtags are derived from domain KBs, currently using the domain-specific commonsense concepts and encyclopedic entities as a dictionary. We have a tunable support threshold, the higher the support sup (at least sup number of dictionary entries are expected in the tweet), the higher the accuracy and lower the coverage. As described in [8], an alternative approach is to construct language models over domain-specific data to estimate the likelihood of the language model to generate the tweet. This step is crucial in filtering tweets and collecting only pertinent ones. For example, from 750 million tweets, we got 2.5 million urban domain-specific tweets, with sup being set to 1.

b) Storage and Cleaning: Once pertinent raw tweets are collected, the file is downloaded, converted into CSV and imported to a MySQL database for further computation. The data on

tweets is then cleaned before further processing. Unnecessary characters, hashtags, usernames are removed. Any duplicate posts such as retweets and identical tweets are removed as well. It is important to clean the tweets to enhance classification accuracy by removing unwanted details that do not contribute to sentiment analysis. Consequently, any URLs in tweets are also removed. It is possible to design a more complex system that deeply analyzes the content of URLs. However, our design decision was simplicity and efficiency, as this is a pre-processing step. Also, we do not want URL content to affect polarity classification through sentiwords.

c) Text Processing of Tweets: We use a sentence level model for processing (not document level) because Twitter is a microblogging site where a tweet is at most 140 characters, therefore a sentence level model is preferable over a document level model. Text processing of tweets is conducted with TextBlob, a Python library that provides a consistent API for common NLP tasks including part-of-speech tagging, noun phrase extraction, classification, translation and more.

d) Polarity Classification with Sentiwords: The Sentiwords lexicon is used to analyze sentiments expressed in tweets. Sentiwords are words pertaining to emotions. We use SentiWordNet 3.0 from LREC [3] for this purpose. The sentiwords are mapped to the content of the tweet to determine whether it is closest to expressing a positive or negative or neutral sentiment. Thus, the polarity of tweets is classified into one of the three categories and used for further analysis.

e) Analysis and Visualization: The information about the polarity of each tweet is computed

and stored in a json file. A Python script is written to aggregate this information, thus as an output, we get a set of positive, negative and neutral tweets. Using this polarity information, we plot graphs with the given data (discussed in the next section). Graph plotting is done using IPython Notebook. The plotted results displayed in graphical form allow users to see public reaction at-a-glance.

4. 1. 3. 3 Experiments and Observations

We summarize our experiments pertaining to tweets in South East Asia on pollution caused by peatland fires [9]. We briefly explain the background for our experiments here. Peatlands have vast organic matter due to low decomposition of plant residue. Indonesia has the most peatlands in South East Asia. Pollutants due to these fires also affect neighboring countries, e.g., Malaysia and Singapore. Thus, Indonesian Peatland Fires (IPFs) are considered to be an international problem in Environmental Management. Pollution caused by airborne particulates is of primary concern. Studies show that rhinitis, asthma, and respiratory infections increase when particulate concentration is of hazardous level [10]. Singapore has built an air quality system called Pollutant Standards Index (PSI), which incorporates six pollutants: sulphur dioxide (SO₂), particulate matter (PM₁₀), fine particulate matter (PM_{2.5}), nitrogen dioxide (NO₂), carbon monoxide (CO) and ozone (O₃). The Singapore national environment agency publicly publishes the PSI level hourly through websites (e.g., haze.gov.sg). Twitter is one of the most visited social

media sites. People get the information about PSI levels through this, and more importantly, express their reaction to the daily PSI level and air quality. We thus use this Twitter data in the experiments shown here. Note that it is important to conduct this analysis, since it also has the broader impact of catering to smart cities. Public opinion expressed through social media is useful for the smart governance characteristic. Also, counterbalancing the effect of hazards to maintain public health and safety is important in the smart environment characteristic.

In the experiments shown here, we collect pertinent tweets using hashtags and store them in a MySQL database. Based on KB knowledge, some hashtags used in collection of these tweets are: CO2, clean air, air pollution, Singapore, climate change, etc. The tweet collection parameters are as follows:

- i) `q=air+pollution+singapore+%22air+pollution%22+%23+Singapore`; This shows the query used
- ii) `lang=en`; This is the language, which in our case is English
- iii) `count=100`; The number of tweets to return per page, up to a max of 100
- iv) `until=2015-10-01`; Returns tweets generated before the given date.
- v) `since_id=?` Returns results with an ID greater than (i.e., more recent than) the specified ID

These tweets are limited by geographical range, in our case, Singapore (though we consider a multicity context, tweets are collected from Singapore for experiments here; yet they reflect reactions of people in other cities also, constituting multicity analysis). The date range for these tweets is the end of October to the first week of December, 2015. The tweets are then fetched from the table one by one for cleaning. Figure 4.2 shows a code snippet of the functions used for

cleaning the tweets. Once the cleaning is completed, the clean tweets undergo classification either as positive, negative or neutral tweets. These results of sentiment analysis are then visualized by graphical plotting as the last step of the analyzer. Figure 4.3 shows an example of visualization. This provides an at-a-glance view of mining public opinion in the area.

```
def replaceDuplictes(s);
    #replace repetitions
    pattern = re.compile(r"(\1{1,})", re.DOTALL)
    return pattern.sub(r"\1",s)

def processTweet(tweet);
    #clean the tweets
    #Convert to lower case
    tweet = tweet.lower()
    #Remove www.* or https?/*
    tweet = re.sub('((www\.[\s]+)|(https?://[\s]+))', '', tweet)
    #Remove @username
    tweet = re.sub('@[\s]+', '', tweet)
    #Remove additional white spaces
    tweet = re.sub('[\s]+', ' ', tweet)
    #Replace hashtags with word
    tweet = re.sub(r'#([\s]+)', r'\1', tweet)
    #trim
    tweet = tweet.strip('\n')
    return tweet
#end
```

Figure 4.2: Code snippet of functions for cleaning tweets

From this figure, it appears that policies to counter-balance the effect of pollution seem fairly satisfactory since 61% of users have expressed positive sentiments. However, there is scope for improvement due to 25% of the users being neutral and 14% being negative. This opinion mining thus provides useful inputs to government bodies in urban planning and also to prospective residents and environmental scientists.

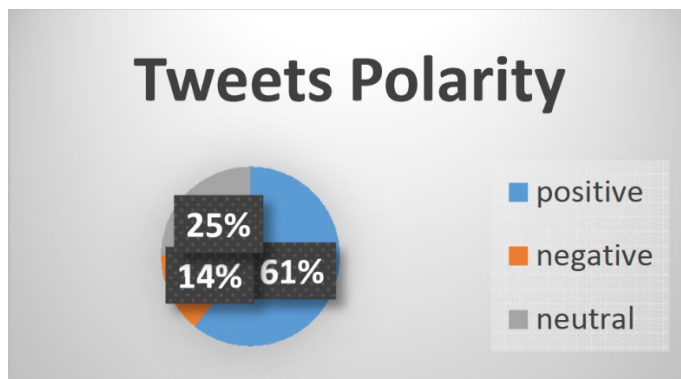


Figure 4.3: Example of visualizing opinion mining results

4. 1. 3 Predictive Analysis and Discussion

Results from the mining can be used for predictive analysis in Environmental Management, more specifically urban planning. To demonstrate this, we develop a prototype prediction tool. Programming for this tool is done in Java. We summarize the evaluation herewith. Sample executions are shown in Figure 4.4 and Figure 4.5. Users enter input conditions and the tool estimates the range of PM_{2.5} based on health impacts. We use terms “very good”, “moderate” etc. to describe PM_{2.5} safety range as per the chance of affecting public health based on AQI (see Table 4.1). For example in Figure 4.4, if a user enters East Asia & Pacific with gas consumption: 582, vehicles per k people: 700, high income OECD group, road density: 11, vehicles per km: 20, diesel consumption: 467 and cars per k people: 550, the tool predicts that PM_{2.5} range is “very good”. It means that, as learned by mining over existing data, the PM_{2.5} range for the given user entry is predicted as 0 - 12.0 µg/m³, which is within safe limits for good

health. Similarly, we can interpret Figure 4.5.

Many experiments are conducted with the prototype tool and useful predictions are obtained. This tool is evaluated by scientists in Environmental Management who consider it to be helpful in urban planning. For example, government bodies can get an idea of how PM2.5 concentration is affected by change in traffic conditions with respect to health impacts. This can help them plan policies. Residents can estimate air quality based on various inputs to plan their current lifestyles and prospective future moves.

The screenshot shows a web-based evaluation tool with the following inputs and outputs:

Region	IncomeGroup	Diesel_Consumption
East Asia & Pacific	High income: OECD	467
Gasoline_Consumption	Road_Density	Cars_Per_K_People
592	11	550
Vehicles_Per_K_People	Vehicles_Per_KM	
700	20	

Below the input fields, an information icon (i) is displayed next to the output: `PM2.5_RANGE: '(-inf-5.845]verygood'`

Figure 4.4: Evaluation example with good PM2.5 range

The screenshot shows the same evaluation tool with different inputs and outputs:

Region	IncomeGroup	Diesel_Consumption
Middle East & North Africa	Upper middle income	134
Gasoline_Consumption	Road_Density	Cars_Per_K_People
45	12	87
Vehicles_Per_K_People	Vehicles_Per_KM	
125	68	

Below the input fields, an information icon (i) is displayed next to the output: `PM2.5_RANGE: '(18.43-21.755]Moderate'`

Figure 4.5: Evaluation example with moderate PM2.5 range

Likewise, the polarity classification of tweets on air quality is also very useful in predictive analysis. As an output of the social media mining, the tweets are stored in a database along with

their polarities. Visualization of opinion mining results is also stored. This serves as the basis to perform predictive analysis. For example, in the specific scenario here, it enables studying the correlations between users' sentiments and the actual PSI (Pollutants Standards Index) level.

Furthermore, this helps predict potential concern of users given certain PSI levels (based on opinion mining of existing data and correlation). In other words, if a particular PSI level is maintained, it helps estimate whether user sentiments would be positive, negative or neutral. This predictive analysis is useful in urban planning by allowing government bodies to estimate public opinion in advance while making regulations. It helps in catering to the satisfaction of current and future residents. It also provides inputs to environmental scientists for research, e.g., factors leading to PSI and potential measures for improvements from a health standpoint.

4. 1. 4 Related Work

Applied data mining research appears in many fields today as the amount of available data increases and there is also a need to automate analysis from a domain perspective, e.g., in Environmental Management [11]. In urban planning, mining is applied in calibration of cellular automata transition rules that potentially relate to theories on relocation [12]. In this paper, we address issues that are not the focus of earlier works. We consider fine particle pollutants as these are especially harmful due to not being easily filtered by the respiratory system. Also, prior research focuses mostly on single cities while we consider a multicity global context.

An overview of sentiment analysis appears in [13]. They describe approaches for opinion-oriented IR. In SentiWordNet 3.0, a lexical resource to support sentiment classification is developed [3]. It is the result of annotating WordNet synsets by degrees of positivity, negativity and neutrality. In [14] they use an approach to extract sentiments with polarities for specific subjects from a document. They have a syntactic parser and sentiment lexicon for finding sentiments in Web pages and news. Our work fits in this category, orthogonal to the existing literature. We do opinion mining in a domain specific context, incorporating commonsense knowledge to extract concepts from social media as a human expert would. We build domain KBs useful for other tasks as well.

Studies have been conducted on pollutants. Zhou et al. analyze relationships of indoor and outdoor pollutant concentration, finding that they depend on individuals' situations [10]. Forsyth analyzes articles from representative newspapers in affected nations to help provide public opinions to pollutant problems [15]. This shows that public reaction is significant to develop urban regulations. Our research takes a step ahead and mines public reaction from online social media. Since this reaction is crucial in the urban planning area, our paper makes an important contribution here through public opinion mining.

4.1.5 Conclusions

In this paper, we conduct mining on pollutant data from social media and structured sources

to discover knowledge on air quality from a health standpoint. We use association rules, clustering and classification to mine structured data from global sources on urban air pollution. In social media mining we use Twitter, incorporate CSK and build domain KBs to guide extraction as a human expert would. We use this domain knowledge, lexical databases and text processing for polarity classification of tweets and visualize the results. Knowledge discovered by mining is useful in predictive analysis. To demonstrate this, we build a prototype tool to estimate air quality with respect to health standards. This is evaluated by domain experts and found useful in urban planning. Estimation from predictive analysis can be helpful to government bodies for urban polices, residents for lifestyle decisions and environmental scientists for further research. Notable contributions of this work include: *mining social media and structured data in a domain-specific context; using CSK for mining tweets; addressing a multicity environment in urban planning; and conducting predictive analysis on air quality for human health*. Ongoing work includes enhancing domain KBs to provide inputs to smart cities, using CSK in the automation of learning processes and potentially deploying CSK with domain KBs for mining from photo-blogs. Another ongoing task is the use of CSK and social media mining to automate identification of IMRs (Implicit Requirements) in Software Requirement Specifications for inputs to AI tools.

4. 1. 6 Acknowledgment

Xu Du is funded by a Doctoral Assistantship from Montclair State University. Onyeka Emebo is supported by a Fulbright Scholarship for International PhD students. Aparna Varde has been a Visiting Senior Researcher at Max Planck Institute of Informatics during a part of this research.

4. 1. 7 References

© 2016 IEEE. Reprinted, with permission, from Xu Du, Onyeka Emebo, Aparna Varde, Niket Tandon, Sreyasi Nag Chowdhury and Gerhard Weikum, Air Quality Assessment from Social Media and Structured Data, IEEE International Conference on Data Engineering, ICDE, HDMM Workshop, May 2016, pp. 54 - 59

[1] S. Zauli Sajani, I. Ricciardelli, A. Trentini, D. Bacco, C. Maccone, S. Castellazzi, P. Lauriola, V. Poluzzi and R. Harrison, “Spatial and indoor/outdoor gradients in urban concentrations of ultrafine particles and PM_{2.5} mass and chemical components”, *Atmospheric Environment*, 2015, Vol. 103, 307-320.

[2] N. Tandon, G de Melo, F. Suchanek and G. Weikum, “WebChild: harvesting and organizing commonsense knowledge from the Web”, *ACM WSDM*, Feb 2014, pp. 523-532.

[3] S. Baccianella, A. Esuli and F. Sebastiani, “SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining”, May 2010, *LREC*, Vol. 10, pp. 2200-2204.

-
- [4] Vienna University of Technology (TU Wien), “European Smart Cities”, Tech Rep, Vienna, Austria, 2015.
- [5] US EPA, “Policy Assessment for Review of the Particulate Matter National Ambient Air Quality Standards (NAAQS)”, epa.gov, 2015.
- [6] <http://www.who.int/gho/en/>, World Health Organization, Data Repository, 2015.
- [7] <http://data.worldbank.gov>, The World Bank, Data By Country, 2015.
- [8] A Varde, N. Tandon, S. Nag Chowdhury and G. Weikum, “Commonsense knowledge in domain-specific knowledge bases”, Technical Report, Max Planck Institute for Informatics, Saarbruecken, Germany, Aug 2015.
- [9] Y. Fujii, S. Tohno, N. Amil, M. T. Latif, M. Oda, J. Matsumoto, and A. Mizohata. "Annual Variations of Carbonaceous PM2.5 in Malaysia: Influence by Indonesian Peatland Fires." *Atmospheric Chemistry and Physics Discussions*, 2015, Vol. 15, pp. 22419-22449.
- [10] J. Zhou, A. Chen, Q. Cao, B. Yang, W. Victor, C. Chang, and W. Nazaroff. "Particle Exposure during the 2013 Haze in Singapore: importance of the built environment." *Bldg and Env.*, 2015, pp. 14-23.
- [11] A. Pampoore-Thampi, A. Varde and D. Yu, “Mining GIS data to predict urban sprawl”, ACM KDD (Bloomberg Track), 2014, pp. 118-125.
- [12] X. Li, Y. Gar-On and A. Yeh, “Data mining of cellular automata's transition rules”. *International Journal Of Geographical Information Science*, 2004, Vol. 18, No. 8, pp. 723-744.

[13] B. Pang and L. Lee, “Opinion mining and sentiment analysis”, Foundations & Trends in Information Retrieval, 2008, Vol. 2, pp. 1-135.

[14] T. Nasukawa and J. Yi, “Sentiment analysis: capturing favorability using natural language processing”, ACM K-Cap, New York City, NY, Oct 2003, pp. 70-77.

[15] T. Forsyth, “Public concerns about transboundary haze: a comparison of Indonesia, Singapore and Malaysia.”, Global Environmental Change, 2014, Vol. 25, pp. 76-86.

4.2 Mapping Ordinances and Tweets Using Smart City Characteristics to Aid Opinion Mining

Abstract: This research focuses on mining ordinances (local laws) and public reactions to them expressed on social media. We place particular emphasis on ordinances and tweets relating to Smart City Characteristics (SCCs), since an important aim of our work is to assess how well a given region heads towards a Smart City. We rely on SCCs as a nexus between a seemingly infinite number of ordinances and tweets to be able to map them, and also to facilitate SCC-based opinion mining later for providing feedback to urban agencies based on public reactions. Common sense knowledge is harnessed in our approach to reflect human judgment in mapping. This paper presents our research in ordinance and tweet mapping with SCCs, including the proposed mapping approach, our initial experiments, related discussion, and future work emerging therein. To the best of our knowledge, ours is among the first works to conduct mining on ordinances and tweets for Smart Cities. This work has a broader impact with a vision to enhance Smart City growth.

CCS Concepts • Information systems → **Data mining**; *Content analysis and feature selection; Clustering and classification;*

Keywords: *Social media, Enterprise Intelligence, Knowledge bases, Local laws, NLP, Sentiment analysis, Text mining*

(Chapter 4.2 resued the previously published paper Puri, M., Du, X., Varde, A., & de Melo, G. (2018), Mapping Ordinances and Tweets using Smart City Characteristics to Aid Opinion Mining. *Companion Volume of The Web Conference - WWW '18*, <https://doi.org/10.1145/3184558.3191632>).

4. 2. 1 Introduction

This research addresses the task of mining urban policy. Our vision is to analyze ordinances or local laws from websites with respect to the public reaction to them expressed on social media. This enables tangential surveys to assess opinions of residents, reflecting their satisfaction and views on urban policies. An important focus in our work is to determine to what extent such ordinances contribute to establishing the relevant urban region as a Smart City. Hence, we aim to categorize these ordinances based on their pertinent Smart City Characteristics (SCCs), of which a small snapshot with highlights is shown in Figure 4.6 (image source [19]). Public opinion is gathered from Twitter, given its role as a micro-blogging site with over 330 million active users. The specific objective of the present research is to relate the ordinances to the respective tweets on Twitter that express the public reaction to them.

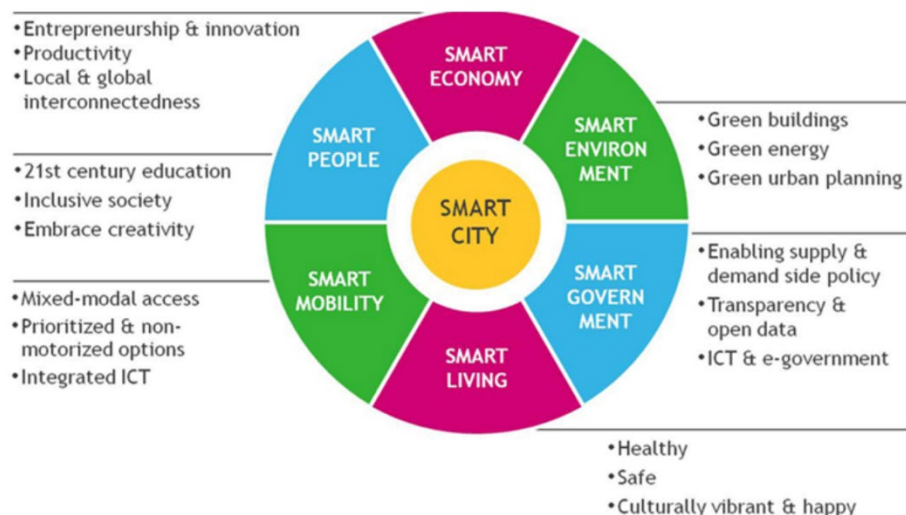


Figure 4.6: Smart City Characteristics – Highlights

We aim to connect ordinances to relevant tweets by drawing on their semantic relatedness. This is non-trivial, as ordinances and tweets both involve highly intricate and rather heterogeneous natural language, so simple keyword matching does not suffice. Traditional machine learning techniques [36] and related advances are not found suitable for learning this sort of mapping, as they require vast amounts of training data.

Since ours is pioneering work in ordinance mining, we do not have such prior training data. To overcome these challenges, we propose a two-step approach for mapping that exploits the transitive nature of the connection between ordinances and tweets considering their relationship with SCCs. Specifically, the transitive property we invoke is that: if the ordinance relates to a given SCC and any tweet relates to the same SCC, then the ordinance bears a connection to the tweet. This approach is proposed because classical sources of SCC data, e.g. [16, 19] are finite

and are restricted to a limited set of identifying features that can be relied upon for mapping (see Figure 4.6). Thus, this transitive approach is more feasible than attempting to directly relate a seemingly infinite number of tweets to ordinances from various websites.

As a first step, we discover connections between SCCs and ordinances using classical SCC sources guided by common sense knowledge (CSK) from web-based repositories. In a second step, we consider the mapping of tweets to SCCs, again drawing on such CSK. This approach then enables us to directly relate ordinances and the tweets to the pertinent aspects of Smart Cities and also sets the stage for sentiment polarity classification [10, 23] and sentiment aspect analysis [34] of pertinent tweets using suitable methods to assess public opinion.

This work aims for broader impact by contributing to the development towards Smart Cities. If we identify which SCCs are being addressed by the local laws or ordinances passed by urban agencies, we are able to provide feedback on how well their urban policies head towards Smart City development across various categories.

Moreover, this work relates to the theme of Social Sensing. Public reactions inferred from opinion mining (to be conducted after connecting ordinances to tweets using SCCs) can further enable involved urban councils and management agencies to judge public satisfaction. This can allow for assessing the appeal of Smart City ordinances from a public opinion standpoint, thus providing useful feedback to the agencies that may enable them to enhance their policies for Smart City development. To achieve this sort of analysis, we draw on artificial intelligence

aspects of text mining, natural language processing, and common sense knowledge. The rest of this paper is organized as follows. Section 2 describes pertinent related work. Section 3 explains our proposed mapping approach to connect ordinances and social media postings. Section 4 summarizes its evaluation through experiments and discussion. Section 5 gives the conclusions, including our findings and a description of ongoing research.

4. 2. 2 Related Work

While there has been ample work on mining social media, most previous work differs substantially from the task we consider here.

There is a long history of research on link prediction in social networks [2, 35]. These methods, however, are geared towards creating links between homogeneous sorts of nodes, such as predicting friendship connections between pairs of social network users. The same applies to most of the research on the even longer standing problems of entity resolution [7] and alignment between resources [8]. Only few approaches have targeted open-domain linking between arbitrary entities and concepts [1, 4, 9, 25, 26]. However, these typically assume structured data as input, i.e. entities with a series of attributes. In our case, we are attempting to connect two forms of unstructured natural language text. On the one side, we have public ordinances expressed using highly formal language, replete with legal terminology. On the other side, we have social media posts consisting of text that is typically very informal in nature, including

embedded hashtags, URLs, etc.

For social media text, one important line of inquiry has focused on unsupervised topic modeling and trend detection in social media [15]. In [38], a fuzzy-based approach is used to preprocess and analyze hashtags in Twitter with the resulting fuzzy clusters being studied to investigate temporal trends on hashtag popularity. Such works however cannot easily be applied to the task of mapping tweets to a pre-existing set of ordinances, which we consider in our research. Neural vector-based representations of documents [5] also fail when the two items are as heterogeneous as in our case.

Some recent approaches on linking social media text have relied on supervised classification. While standard methods can be applied to predict links between heterogeneous items [36], an important challenge is that large training sets are required to accurately cope with the short length (leading to data sparsity) and variability of tweets. To overcome this, the TweetSift system [18] classifies tweets by topic while exploiting external entity knowledge and topic-enhanced word embeddings. The latter leads to topic-specific word embeddings such that the different senses of ambiguous words obtain different representations. However, this assumes that the knowledge base can provide highly pertinent signals about entities such as specific Twitter users. Our model in contrast exploits generic common sense knowledge and does not require a detailed labeled training set.

Furthermore, previous work has not considered the setting of ordinances (with tweets) and

Smart City Characteristics, along with their challenging use of language. To the best of our knowledge, our work is therefore among the pioneering research in this area.

Much attention is being given to Smart Cities in recent years. Buses in Barcelona are designed to run on routes optimal for power consumption [19]. Canal lights in Amsterdam automatically brighten and dim based on pedestrian usage [19]. The work in [22] addresses the potential enhancement of automated vehicles by embedding them with common sense knowledge. Such initiatives contribute mainly to the *Smart Mobility* characteristic. There is also significant research on making use of technology in fighting crime, e.g. the monitoring system to identify and categorize crime-related events in text documents [24] that was developed within the EU ePOOLICE project. Such research contributes to the *Smart Living* characteristic. The work in [21] targets the *Smart Environment* characteristic through cloud computing solutions for data centers (instead of on-premise servers). They analyze scenarios where cloud models provide greater energy efficiency, yet meeting productivity targets. Security, privacy, and availability issues are discussed for cloud usage in the greening of data centers. Free cooling for data centers as addressed in [20] by considering temperature, humidity and other parameters, also contributes to *Smart Environment*. The work in [37] has a tangential influence on *Smart Economy*. The authors propose a mathematical model to minimize trips in scheduled pickups and deliveries by cooperation. This is a cost-effective method useful in urban delivery systems to reduce operational expenses in a cooperative mode. Likewise, the research conducted in [33] has an

indirect impact on the *Smart People* characteristic by addressing an aspect of 21st century education through collocation-based writing aids for second language learners of English, as they constitute a large part of the population in cities worldwide. The work of [12] while primarily impacting *Smart Environment* through its estimation of air quality by analyzing pollutant data, also has a secondary impact on *Smart Living* since it addresses issues from a health standpoint. Thus, several researchers are conducting studies to augment the characteristics of Smart Cities.

Our work in this paper seeks to make a notable impact here, by advocating for the deployment of common sense knowledge in the realm of Smart Cities. While works such as [11, 17] motivate the need for common sense in the areas of *Smart Mobility* and *Economy*, respectively, the actual use of such knowledge in these paradigms remains at the stage of inception, e.g. [22]. As addressed in several works on common sense in machine intelligence (acquisition, representation, and application) surveyed in [31], the increased usage of CSK in many areas would promote much smarter machines. Our research in this paper aims to take a significant step along this avenue, with the overall goal of enhancing Smart Cities.

4. 2. 3 Proposed Mapping Approach

The approach we propose for ordinance to tweet mapping through Smart City Characteristics (SCCs) is illustrated in Figure 4.7. It is described in detail in the following subsections.

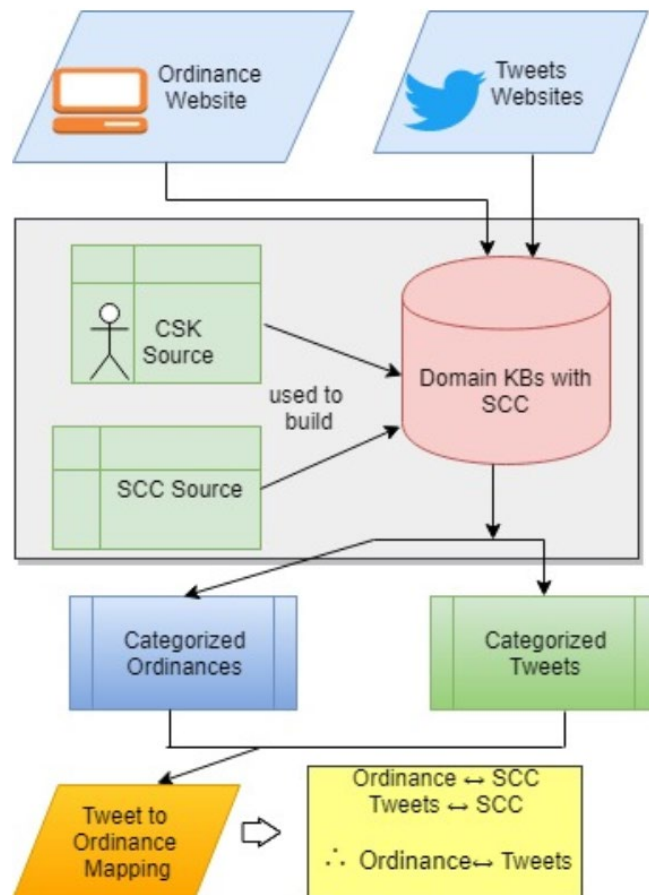


Figure 4.7: Proposed approach for SCC mapping

4.2.3.1 CSK — SCC based KB Development

The *SCC source* used in our approach is derived from the widely accepted technical report from TU Wien [19], which enumerates six SCCs. These are *Smart Governance* (or *Government*), *Smart Economy*, *Smart Mobility*, *Smart Environment*, *Smart People*, and *Smart Living*, respectively.

Consider, for instance, the SCC *Smart Governance*. This encompasses the features listed next, some of which are also included among the highlights listed in Figure 4.6.

- Transparency in government
- Optimizing public service and administration
- Direct involvement in public policies
- Citizen participation
- Positive and open communication channel with citizens
- More informed decisions by feedback and engagement

The screenshot shows the WEBCHILD Commonsense Browser interface. At the top, there is a search bar with the word "economy" entered and a search icon. Below the search bar, the interface is divided into two main sections. On the left, there is a sidebar titled "Guess the concept" with several expandable categories: Domain, Comparable, Physical Part, Activity, Property, and Location. Below these categories is a button labeled "Ask me!". On the right, the search results for "economy" are displayed. The results include a definition: "economy" is "the system of production and distribution and consumption". Below this, there are several sections of related terms and concepts, each with a "More" link:

- TYPE OF**: system
- Related to group, under the category of social
- COMPARABLES**: economy.wall_street, economy.market, economy.Indicator, service.economy
- ACTIVITIES**: create economy, call economy, make economy, help economy, practice economy
- PHYSICAL PROPERTIES**: open, fast, weak, large, strong, More
- ABSTRACT PROPERTIES**: romantic, attractive, serious, healthy, private, More
- OTHER PROPERTIES**: big, productive, loud, robust, cheap, More

Figure 4.8: Relevant partial screenshot of WebChild

Thus, if ordinances reference any of the above features, we infer that they likely relate to Smart Governance. However, these expressions are not particularly likely to be observed in the

ordinances literally. If human users were to inspect these ordinances, they could draw relevant connections, which are often quite subtle, by relying on linguistic knowledge and common sense. To automate this process, we draw on common sense knowledge (CSK) web sources, specifically, the large WebChild repository [28, 30] with common sense concepts mined from vast amounts of data on the Web along with their *properties* and *relationships*. A partial screenshot of the WebChild browser appears in Figure 4.8. This depicts a relevant concept *economy*, which pertains to a specific SCC.

Using WebChild as the main *CSK source* along with requisite information for knowledge base development [32] and other common sense related sources such as the lexical database WordNet [14], we build domain-specific knowledge bases on Smart City Characteristics (Domain KBs with SCC). These KBs are text-based and contain terms relevant to specific Smart City Characteristics derived from CSK repositories and SCC sources, using NLP and semantic matching. Note that one could also apply techniques such as knowledge base extraction from text [27, 29] and rule mining [6] to increase the size of these domain KBs. Figure 4.9 shows a subset of our Domain KBs with terms relevant to the characteristics of *Smart Environment* and *Smart Mobility*

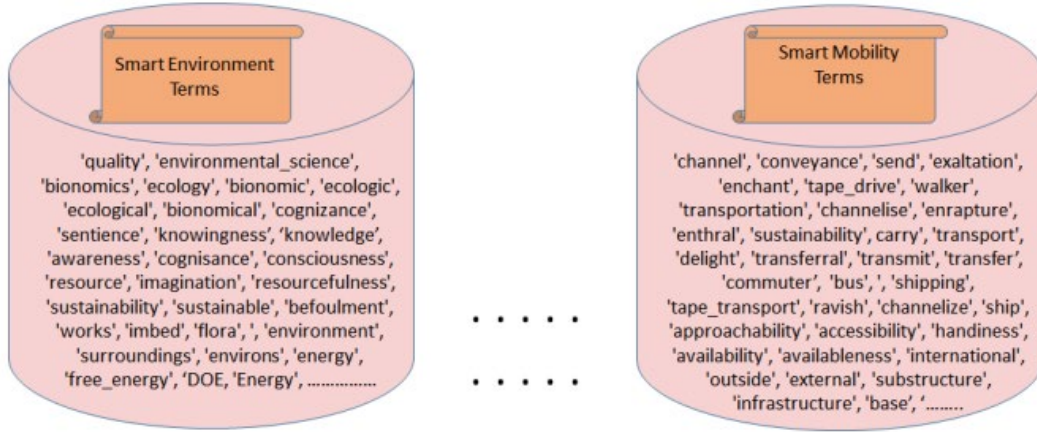


Figure 4.9: Part of Domain KBs with SCC (Subset of Smart Environment and Smart Mobility terms)

Algorithm 1 Linking algorithm

```

1: for each SCC  $S_j$  do
2:   Build domain knowledge base  $K_j$ 
3: for each ordinance  $O_i$  do ▷ ordinance linking
4:   for each SCC  $S_j$  do
5:      $L_{i,j} \leftarrow \sum_{x \in K_j} C(O_i, x)$ 
6:   Assign  $O_i$  to the  $S_j$  with  $j = \operatorname{argmax}_j L_{i,j}$ 
7: for each social media posting  $T_i$  do ▷ social media linking
8:   for each SCC  $S_j$  do
9:      $M_{i,j} \leftarrow \sum_{x \in K_j} C(T_i, x)$ 
10:  Assign  $T_i$  to the  $S_j$  with  $j = \operatorname{argmax}_j M_{i,j}$ 
11:  $\mathcal{O} \leftarrow \{(O_i, T_k) \mid \exists S_j : (O_i \text{ assigned to } S_j) \wedge (T_k \text{ assigned to } S_j)\}$ 
12: return  $\mathcal{O}$  ▷ Links between ordinances and social media

```

Algorithm 1: Linking algorithm

4. 2. 3. 2 Linking using SCCs and CSK

Using these domain KBs, CSK concepts are deployed to semantically relate terms x in

ordinance text T to SCCs. We denote this as $C(T, x)$. For example, if the ordinance text includes the term *smoke detector*, then CSK concepts help to semantically relate this with the SCC *Smart Environment* through the CSK properties of *smoke detector* that have features relevant to this SCC. This information is found in the domain KBs derived from the SCC and CSK sources.

The same ordinance can also have features that relate to other SCCs. It is possible that some terms in ordinances may overlap with multiple SCCs. In that case, they would be observed in the KBs of each of those SCCs. If such concept terms are discovered in the ordinances, their occurrences are counted towards multiple categories. For example, if a term such as *sustainability* occurs in an ordinance, then that ordinance would be counted under the characteristics of *Smart Mobility* as well as *Smart Environment* (see Figure 4.9). Thus, the counts for both of these SCCs would be updated in this particular example. Finally, all the aggregate SCC counts are examined and each ordinance is accordingly linked to the SCC with the maximum number of relevant features. CSK plays a crucial role in finding semantic relatedness for this mapping through concepts, properties, etc. Likewise, we map tweets to SCCs following a similar CSK-guided procedure. Using this, we finally aim to output the linkages between ordinances and tweets via mutual SCC connections. Thus, we emphasize that: *an ordinance broadly links to a particular tweet if they both map to the same SCC.*

This mapping approach used for linking them is summarized in Algorithm 1 herewith. As of now, for simplicity, we emit only the closest matching SCC for the ordinances and tweets as

output.

4. 2. 4 Evaluation of The Mapping

We conduct an evaluation of mapping ordinances and tweets with SCCs using large amounts of real data from publicly accessible websites on ordinances and tweets. A summary of our experimental evaluation is presented in the following.

4. 2. 4. 1 Ordinance to SCC Mapping

Large amounts of historical data on ordinances are gathered from the website of the NYC council [3], which is openly available to the public. A small portion of this is shown in the screenshot that appears in Figure 4.10. These ordinances are first extracted into a machine-readable form and then subjected to a preprocessing step such that only their textual content is retained. The other attributes such as “Prime Sponsor”, “Council Member Sponsor”, etc. (see Figure 4.10) are filtered out during this preprocessing phase. The textual content of the ordinances then serves as input to our algorithm that conducts the ordinance to SCC mapping. The algorithm interfaces with the SCC KB and uses the relevant terms for mapping. This inking procedure is formalized within Algorithm 1. It accordingly counts all such matches to output the SCC with the maximum counts as the closest matching one.



The screenshot shows the NYC Council website interface. At the top, it says 'THE NEW YORK CITY COUNCIL' with 'MELISSA MARK - VIVERITO, SPEAKER' and 'LEGISLATIVE RESEARCH CENTER'. Below this is a navigation menu with 'Council Home', 'Legislation', 'Calendar', 'City Council', and 'Committees'. A search bar is present with 'Session 2014-2017' and 'Local Law' selected. The search results show 323 records. The first three records are visible in the table below.

File #	Law Number	Introduction	Enacted	Committee	Prime Sponsor	Council Member Sponsors	Title
Int.0001-2014	2014/007			Committee on Civil Service and Labor	Margaret S. Chin	41	A Local Law to amend the New York city charter and the administrative code of the city of New York, in relation to the provision of sick time earned by employees, and section 7 of local law number 46 for the year 2013, relating to such sick time, in relation to the effective date of such local law, and to repeal section 6 of local law number 46 for the year 2013, relating to a determination of the Independent Budget Office.
Int.0098-2014	2014/008	Introduction	Enacted	Committee on Civil Service and Labor	I. Daneek Miller	5	A Local Law to amend the administrative code of the city of New York, in relation to health insurance coverage for surviving family members of certain deceased employees of the department of environmental protection.
Int.0173-2014	2014/009	Introduction	Enacted	Committee on Civil Rights	James Vacca	19	A Local Law to amend the administrative code of the city of New York, in relation to the prohibition of

Figure 4.10: Sample of NYC Council website

Shown herewith is an excerpt from an ordinance (Ord. 1) from the aforementioned NYC council website, along with its closest matching SCC (Table 4.5) based on quantifying relevant ordinance terms with SCC features.

Ord. 1: *A Local Law to amend the administrative code of the city of New York, in relation to amending the district plan of the Downtown – Lower Manhattan business improvement district to change...*

With reference to this ordinance excerpt, our algorithm relies on the SCC Domain KBs and comes to the conclusion that the only term relevant to the *Smart Economy* characteristic is *business*, while many terms are relevant to *Smart Governance*, including, among others, *law*, *administrative*, *district plan*, *improvement*, etc., as summarized in Table 4.5. Thus, the SCC that

is returned as the closest matching one in this example is *Smart Governance*.

Table 4.5: A Sample Ordinance and its SCC Mapping

Smart City Characteristic	Count of Terms
Economy	1
Environment	0
Governance	15
People	0
Mobility	0
Living	0

Numerous further ordinances are analyzed following the same pattern. Note that in our execution so far, only the closest matching SCC is offered as the ordinance mapping output, for simplicity. The same holds for the mapping of tweets to SCCs.

We evaluate of a subset of NYC council data that encompasses two recent ordinance sessions, namely, 2006 to 2009 and 2010 to 2013. Based on this evaluation, we obtain a summary plot of ordinance to SCC mappings given in Figure 4.11.

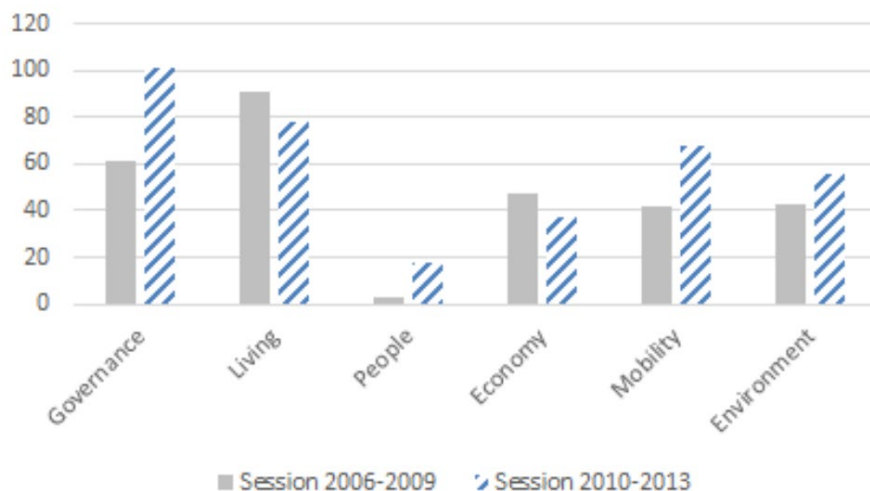


Figure 4.11: Summary plot of ordinance SCC mapping

The observations in this summary plot are useful to provide some feedback to urban management agencies on the extent to which their ordinances cater to various aspects of Smart Cities. For example, from the results, one can conclude that the Smart City Characteristic receiving the greatest attention is *Smart Living* in the first session and *Smart Governance* in the second session. In both of the sessions, the SCC supposedly receiving the least attention is *Smart People*. This may help the urban agencies to plan their future policies such that they make progress on policies pertaining also to those characteristics that have been received comparably little attention so far, in this case the *Smart People* characteristic. Details on various aspects of urban legislation impacts with respect to such analysis appear in [13] catering mainly to a domain-specific angle. This is an important motivation for our current research with ordinances, tweets and SCCs.

4.2.4.2 Tweet to SCC Mapping

We extract thousands of tweets posted by the public on Twitter pertaining to NYC location-specific data. The Twitter Streaming API feature labeled *Filter Realtime Tweets* is used for conducting the extraction. The tweets are extracted to a text file and further processed using NLP techniques such as regular expressions. The relevant parts of the tweets such as their textual content and hyperlinks are retained. These are stored as cleaned tweets. The SCC mapping is then performed on the cleaned tweets using the concerned part of our approach as depicted in Algorithm 1. In Figure 4.12, we show only a small subset of cleaned tweets used among over 1,000 tweets extracted in our experiments.

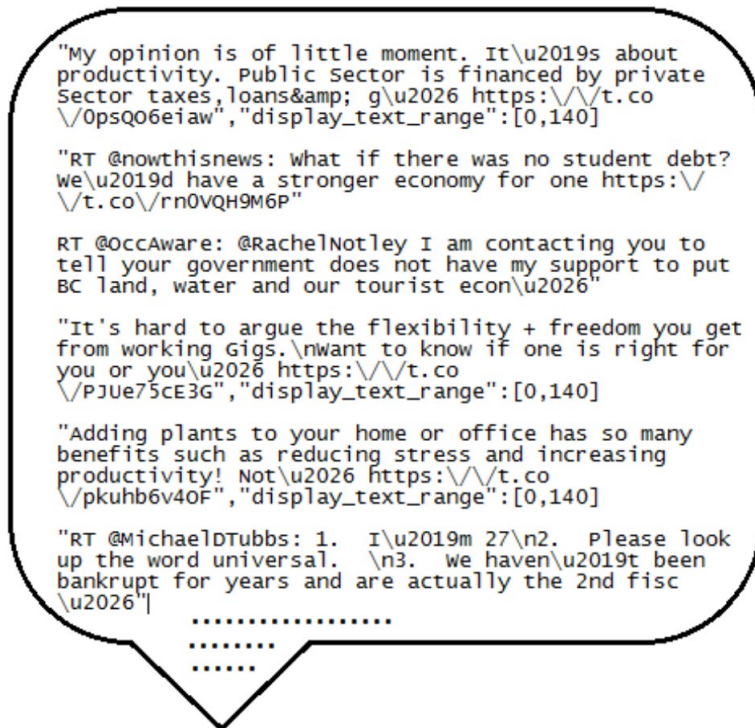


Figure 4.12: Subset of tweets analyzed from NYC sites

Based on these tweets, Figure 4.13 depicts a partial snapshot of our program mapping these cleaned tweets to their most relevant SCC, with reference to the relevant part of the process in Algorithm 1. This is interpreted as follows. Among tweets processed herewith, the overall mapping indicates that 37 of them are on *Smart Economy*, 25 are on *Smart Environment*, 208 on *Smart Living*, etc. These are obtained by the processing shown in the figure, e.g., features of the *Smart Living* SCC include the terms: *home*, *benefits*, *tourist*, *building*, etc., while those of *Smart Environment* include: *energy*, *sustainable*, etc. (The terms are obtained from KBs built using CSK and SCC sources). It is observed in this figure that, overall, 352 tweets are mapped to SCCs

(37+25+. . .+208). Hence, many tweets among approximately 1000 cleaned ones analyzed in these experiments are not mapped to any SCC. This could be due to the fact that not all tweets published by users pertain to SCCs. It could also be that some mappings are not precisely identified in the initial experiments conducted herewith.

```

('Smart City characteristic      occurrences')
(' Smart Economy                ', 37)
(' Smart Environment            ', 25)
(' Smart Governance             ', 45)
(' Smart People                 ', 19)
(' Smart Mobility               ', 38)
(' Smart Living                 ', 208)

('SMART LIVING FEATURE: TERM ', 'home', 1)
('SMART GOVERNANCE FEATURE: TERM ', 'office', 1)
('SMART LIVING FEATURE: TERM ', 'benefits', 2)
('SMART ENVIRONMENT FEATURE: TERM ', 'reducing', 1)
('SMART ECONOMY FEATURE: TERM ', 'economy', 1)
('SMART ECONOMY FEATURE: TERM ', 'economy', 2)
('SMART LIVING FEATURE: TERM ', 'tourist', 3)
('SMART ECONOMY FEATURE: TERM ', 'economy', 3)
('SMART ECONOMY FEATURE: TERM ', 'economy', 4)
('SMART LIVING FEATURE: TERM ', 'tourist', 4)
('SMART LIVING FEATURE: TERM ', 'building', 5)
('SMART MOBILITY FEATURE: TERM ', 'car', 1)
('SMART ENVIRONMENT FEATURE: TERM ', 'energy', 2)
('SMART ENVIRONMENT FEATURE: TERM ', 'sustainable', 3)
.....
.....

```

Figure 4.13: Partial snapshot of tweet to SCC mapping

4.2.4.3 Assessment and Discussion

In order to facilitate judging the correctness of the mappings, we have developed very simple GUIs in our initial execution. We illustrate a relevant part of our *Tweet Mapping GUI* next. This accepts a tweet as the input and emits the closest matching SCC from the user as its

output, or “*No matches*” if none gets matched. Figure 4.14 shows an example of a tweet and its SCC identified as *Smart Environment*, while an example of a non-matching tweet is given in Figure 4.15. Both of these are partial GUI screenshots.

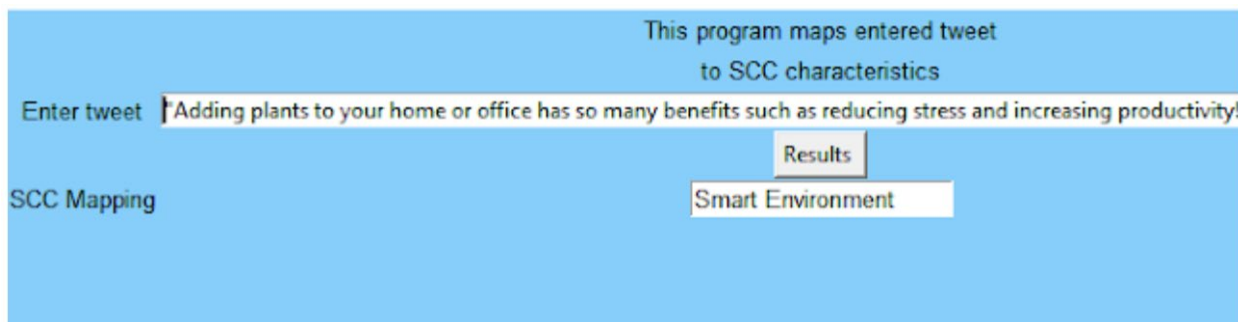


Figure 4.14: Example of SCC mapping identified

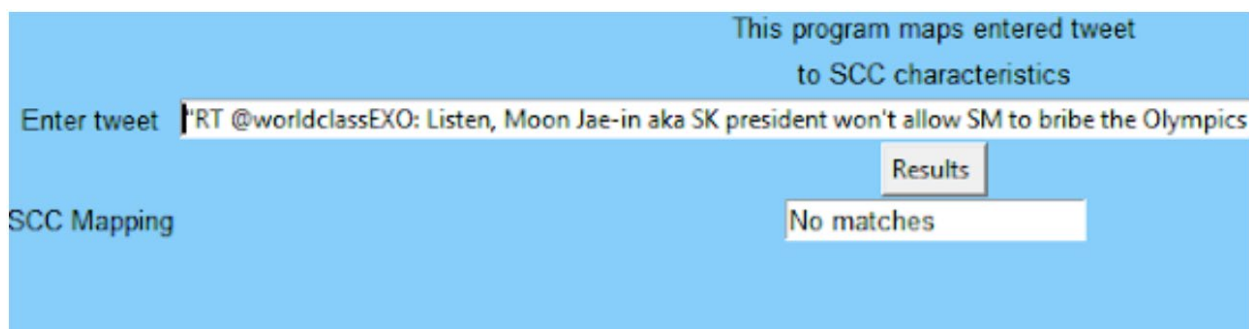


Figure 4.15: Example of no matches for any SCC

Considering several tweets entered and SCCs identified through this GUI, the correctness of these mappings is assessed by domain experts from Earth and Environmental Studies. A similar *Ordinance Mapping GUI* is provided for the ordinance to SCC mappings for enabling at-a-glance displays. These mappings are also assessed by the domain experts.

The actual calculation of accuracy is done using an *Accuracy* metric as follows. Considering the judgment provided by domain experts, if an ordinance or tweet is mapped to a given SCC by

our proposed approach (or if it returns a *No Match*) and this is verified as correct by the expert, it is considered a *True Mapping (TM)*. If the expert labels this mapping as incorrect, it is a *False Mapping (FM)*. For example, if the approach indicates that the SCC is *Smart Governance*, but the expert states that it is *Smart Economy* or that it is a *No Match*, it would be a *False Mapping*. Also, if the approach indicates a *No Match*, but the expert states that it maps to a given SCC, it is still considered a *False Mapping*. In other words, the ground truth is defined by experts for the data analyzed herewith.

With this justification, we proceed to calculate

$$Accuracy = \frac{TM}{TM + FM}$$

and this is used for measuring the effectiveness of our proposed mapping approach. This is analogous to the classical notion of true positives and false positives in data mining and machine learning techniques [36]. (We do not consider true negatives and false negatives at this point in our research, since their appropriate definition needs further insights and discussions with domain experts. This is an aspect of future work). Based on the given definition of Accuracy herewith, we obtain the evaluation scores as listed in Table 4.6.

Table 4.6: Accuracy of Ordinance and Tweet Mapping

	Ordinances	Tweets
Expert 1	84%	72%
Expert 2	86%	69%
Expert 3	81%	70%

Thus, the ordinance to SCC mapping, as verified by domain experts, is found to be accurate for around 85% of the ordinances. This is considered satisfactory on the whole, although there is scope for improvement. The main reasons for the difference are that some ordinances can actually map almost equally to multiple SCCs, and hence it is possible that our approach identifies one particular SCC as the top match, while an expert identifies another.

The accuracy of the tweet to SCC mapping is in the range of around 70%, which seems fairly reasonable for a start. However, it is much lower than that of the ordinance to SCC mapping. We present a few examples of tweets below that are classified incorrectly or return no match, thereby adversely affecting the performance of the tweet to SCC mapping in our approach.

- "Wind 0.0 mph N. Barometer 30.134 in, Falling slowly. Temperature 26.1 °F. Rain today 0.00in. Humidity 94%"
- "RT @worldclassEXO: Listen, Moon Jae-in aka SK president won't allow SM to bribe the

Olympics bec that's gonna ruin the country's reputation. . ."

- "@FoxNews No DACA until Wall is built."
- "Our February STEM Hero is... #STEMed #STEM #SciEd #ScienceEd @polyprep

[https://9O7Gf5sZP7..."](https://9O7Gf5sZP7...)

Inspecting such examples, an important observation is that the problem of inaccurate mapping (or that of no matches being found) occurs mainly due to challenges such as ambiguity, informal language, excessive use of acronyms and hashtags, etc. These issues pose significant challenges in the execution of the mapping. This calls for further research on the tweet to SCC mapping process. We have encountered the challenges listed next in the tweet mapping part of our research.

(1) Tweets use informal language, which makes their extraction and analysis difficult.

(2) The length restrictions imposed on tweets results in users resorting to an excessive use of acronyms.

(3) There is limited coverage, e.g., 1/3 of mentions on the web cannot be linked to Wikipedia (around 30% loss).

(4) NEE (Named Entity Extraction) and NED (Named Entity Disambiguation) involve many degrees of uncertainty.

Addressing these non-trivial challenges, while also aiming to improve the ordinance to SCC mapping accuracy and considering the semantic proximity to multiple SCCs via rankings,

constitutes our ongoing work. We further aim to drill down to a finer level in the mapping, which would entail identifying fine-grained aspects of the individual features in the SCCs as opposed to the entire SCC per se. This is likely to yield better performance.

4. 2. 5 Conclusion

This paper proposes an approach to map ordinances to tweets expressing public opinion, based on Smart City Characteristics (SCCs) relevant to both of them. The execution of our approach with initial experiments yields an accuracy of ordinance to SCC mapping of around 85%, while that of the tweet to SCC mapping is approximately 70%, both confirmed by domain experts.

Ongoing research includes addressing challenges in the tweet to SCC mapping, improving the accuracy of the ordinance to SCC mapping, considering the mapping of both ordinances and tweets to multiple SCCs with ranking, and also attaining a finer granularity in the mappings besides the broad categorizations considered here. This would enable us to draw more specific conclusions from the results, particularly when they are used for polarity classification of tweets to assess the public reaction.

The long term vision of our research is to provide urban agencies useful feedback on how well they are doing in policy decisions (based on this mining) and hence indicate how closely the given urban region heads towards a Smart City. In summary, our research makes the following

contributions:

(1) addressing the mining of local laws or ordinances and their public reaction through tweets to give urban agencies useful feedback, which is pioneering work in the area;

(2) proposing an approach for ordinance to tweet mapping using Smart City Characteristics as a nexus, deploying the transitive property of semantic relatedness between them;

(3) conducting a study with genuine ordinance data from the NYC council, with mapping accuracy of around 85% (ordinance to SCC) and 70% (tweet to SCC) respectively;

(4) motivating the need for mapping ordinances and tweets to multiple SCCs with ranking, and dealing with finer levels of granularity in SCC features for enhanced performance

Ultimately, we envision our work as contributing to the development of Smart Cities on the whole as a broader impact.

4. 2. 6 Acknowledgments

We acknowledge Niket Tandon from the Allen Institute for Artificial Intelligence (Seattle, WA) for his valuable inputs. We thank Robert Taylor and Clement Alo from Earth and Environmental Studies at Montclair State University for their feedback. Early work on this project started while Aparna Varde was a visiting researcher at the Max Planck Institute for Informatics (Saarbrücken, Germany) in the group of Gerhard Weikum in August 2015. Gerard de Melo's research at Rutgers University is funded in part by ARO grant no. W911NF-17-C-0098 as

part of the DARPA SocialSim program. Manish Puri and Xu Du are funded by a Research Assistantship (Computer Science) and a Doctoral Assistantship (Environmental Management) respectively at Montclair State University.

4.2.7 References

- [1] Christoph Böhm, Gerard de Melo, Felix Naumann, and Gerhard Weikum. 2012. LINDA: Distributed Web-of-Data-Scale Entity Matching. In *Proceedings of the 21st ACM Conference on Information and Knowledge Management (CIKM 2012)*, Xue-wen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki (Eds.). ACM, New York, NY, USA.
- [2] Zhu Cao, Linlin Wang, and Gerard de Melo. 2018. Link Prediction via Subgraph Embedding-Based Convex Matrix Completion. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*. AAAI Press.
- [3] The New York City Council. 2018. Legislative Research Center Web Page. (2018). Retrieved March 4, 2018 from <http://legistar.council.nyc.gov/>
- [4] Gerard de Melo. 2013. Not Quite the Same: Identity Constraints for the Web of Linked Data. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI 2013)*, Marie desJardins and Michael L. Littman (Eds.). AAAI Press, Menlo Park, CA, USA, 1092–1098. <http://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/view/6491>
- [5] Gerard de Melo. 2017. Multilingual Vector Representations of Words, Sentences, and

Documents. In *Proceedings of IJCNLP 2017*. <http://www.aclweb.org/anthology/I17-5002>

[6] Gerard de Melo, Mouna Kacimi, and Aparna Varde. 2015. Dissertation Research Problems in Data Management and Related Areas. *SIGMOD Record* 44, 4 (December 2015). http://sigmod.org/publications/sigmodRecord/1512/pdfs/09_reports_Melo.pdf

[7] Gerard de Melo and Gerhard Weikum. 2008. Language as a Foundation of the Semantic Web. In *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC 2008) (CEUR WS)*, Christian Bizer and Anupam Joshi (Eds.), Vol. 401. CEUR, Karlsruhe, Germany.

[8] Gerard de Melo and Gerhard Weikum. 2008. Mapping Roget's Thesaurus and WordNet to French. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*. ELRA, Paris, France, 3306–3313.

[9] Gerard de Melo and Gerhard Weikum. 2010. Untangling the Cross-Lingual Link Structure of Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA.

[10] Xin Dong and Gerard de Melo. 2018. Cross-Lingual Propagation for Deep Sentiment Analysis. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*. AAAI Press.

[11] Katherine Rose Driggs-Campbell, Victor Shia, and Ruzena Bajcsy. 2014. Decisions for Autonomous Vehicles: Integrating Sensors, Communication, and Control. In *Proceedings of the*

3rd International Conference on High Confidence Networked Systems (HiCoNS '14). ACM, New York, NY, USA, 59–60. <https://doi.org/10.1145/2566468.2576850>

[12] Xu Du, Onyeka Emebo, Aparna Varde, Niket Tandon, Sreyasi Nag Chowdhury, and Gerard Weikum. 2016. Air Quality Assessment from Social Media and Structured Data: Pollutants and Health Impacts in Urban Planning. In *IEEE International Conference on Data Engineering (ICDE) - Workshops*. 54–59.

[13] Xu Du, Diane Liporace, and Aparna Varde. 2017. Urban Legislation Assessment by Data Analytics with Smart City Characteristics. In *IEEE Ubiquitous Computing, Electronics and Mobile Communications Conference (UEMCON)*. 20–25.

[14] Christiane Fellbaum (Ed.). 1998. *WordNet: An Electronic Lexical Database*. The MIT Press. <http://www.amazon.ca/exec/obidos/redirect?tag=citeulike04-20{&}path=ASIN/026206197X>

[15] Liangjie Hong and Brian D. Davison. 2010. Empirical Study of Topic Modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics (SOMA '10)*. ACM, New York, NY, USA, 80–88. <https://doi.org/10.1145/1964858.1964870>

[16] IEEE. 2018. IEEE Smart Cities Technical Community. (2018). Retrieved March 4, 2018 from <https://smartcities.ieee.org/>

[17] G. Leef. 2007. *Smart Economics: Commonsense Answers to 50 Questions about Government, Taxes, Business and Households*. Foundation for Economic Education.

[18] Quanzhi Li, Sameena Shah, Xiaomo Liu, Armineh Nourbakhsh, and Rui Fang. 2016. TweetSift: Tweet Topic Classification Based on Entity Knowledge Base and Topic Enhanced Word Embedding. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM '16)*. ACM, New York, NY, USA, 2429–2432.

<https://doi.org/10.1145/2983323.2983325>

[19] TU Wien (Vienna University of Technology). 2015. *European Smart Cities, Technical Report*. Technical Report. Vienna, Austria.

[20] Michael Pawlish and Aparna Varde. 2010. Free Cooling: A Paradigm Shift in Data Centers. In *IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*. 347–352.

[21] Michael J. Pawlish, Aparna S. Varde, and Stefan A. Robila. 2015. The Greening of Data Centers with Cloud Technology. *Int. J. Cloud Appl. Comput.* 5, 4 (Oct. 2015), 1–23.

<https://doi.org/10.4018/IJCAC.2015100101>

[22] Priya Persaud, Aparna Varde, and Stefan Robila. 2017. Enhancing Autonomous Vehicles with Commonsense: Smart Mobility in Smart Cities. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI) - Smart Cities Workshop*.

[23] Andi Rexha, Mark Kröll, Mauro Dragoni, and Roman Kern. 2016. Polarity Classification for Target Phrases in Tweets: A Word2Vec Approach. In *The Semantic Web - ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 - June 2, 2016, Revised Selected*

Papers (Lecture Notes in Computer Science), Harald Sack, Giuseppe Rizzo, Nadine Steinmetz, Dunja Mladenic, Sören Auer, and Christoph Lange (Eds.), Vol. 9989. 217–223.

https://doi.org/10.1007/978-3-319-47602-5_40

[24] Jacobo Rouces, Gerard de Melo, and Katja Hose. 2015. Representing Specialized Events with FrameBase. In *Proceedings of the 4th International Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2015) at ESWC 2015*. http://ceur-ws.org/Vol-1363/paper_7.pdf

[25] Jacobo Rouces, Gerard de Melo, and Katja Hose. 2016. Complex Schema Mapping and Linking Data: Beyond Binary Predicates. In *Proceedings of the WWW 2016 Workshop on Linked Data on the Web (LDOW 2016)*.

[26] Jacobo Rouces, Gerard de Melo, and Katja Hose. 2016. Heuristics for Connecting Heterogeneous Knowledge via FrameBase. In *Proceedings of ESWC 2016 (Lecture Notes in Computer Science)*. Springer.

[27] Niket Tandon and Gerard de Melo. 2010. Information Extraction from WebScale N-Gram Data. In *Web N-gram Workshop. Workshop of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Chengxiang Zhai, David Yarowsky, Evelyne Viegas, Kuansan Wang, and Stephan Vogel (Eds.), Vol. 5803. ACM, 8–15.

[28] Niket Tandon, Gerard de Melo, Fabian Suchanek, and Gerhard Weikum. 2014. and Organizing Commonsense Knowledge from the Web. In *Proceedings of the 7th ACM*

International Conference on Web Search and Data Mining (WSDM '14). ACM, New York, NY, USA, 523–532. <https://doi.org/10.1145/2556195.2556245>

[29] Niket Tandon, Gerard de Melo, and Gerhard Weikum. 2011. Deriving a Web-Scale Common Sense Fact Database. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI 2011)*. AAAI Press, Palo Alto, CA, USA, 152–157.

[30] Niket Tandon, Gerard de Melo, and Gerhard Weikum. 2017. WebChild 2.0 : Fine-Grained Commonsense Knowledge Distillation. In *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, Vancouver, Canada, 115–120. <http://aclweb.org/anthology/P17-4020>

[31] Niket Tandon, Aparna S. Varde, and Gerard de Melo. 2018. Commonsense Knowledge in Machine Intelligence. *SIGMOD Rec.* 46, 4 (Feb. 2018), 49–52. <https://doi.org/10.1145/3186549.3186562>

[32] Aparna Varde, Niket Tandon, Sreyasi Nag Chowdhury, and Gerhard Weikum. 2015. *Common Sense Knowledge in Domain-Specific Knowledge Bases*. Technical Report. Max-Planck-Institut für Informatik, Saarbrücken, Germany.

[33] Alan Varghese, Aparna Varde, Jing Peng, and Eileen Fitzpatrick. 2015. A Framework for Collocation Error Correction in Web Pages and Text Documents. *ACM SIGKDD Explorations* 17, 1 (June 2015), 14–23.

[34] Linlin Wang, Kang Liu, Zhu Cao, Jun Zhao, and Gerard de Melo. 2015.

SentimentAspect Extraction based on Restricted Boltzmann Machines. In *Proceedings of ACL 2015*. 616–625.

[35] Liqiang Wang, Yafang Wang, Bin Liu, Lirong He, Shijun Liu, Gerard de Melo, and Zenglin Xu. 2017. Link Prediction by Exploiting Network Formation Games in Exchangeable Graphs. In *Proceedings of IJCNN 2017*.

[36] Ian H. Witten, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[37] Yang Yu, Qi Lou, Jiafu Tang, Junwei Wang, and XiaoHang Yue. 2017. An exact decomposition method to save trips in cooperative pickup and delivery based on scheduled trips and profit distribution. *Computers Operations Research* 87 (2017), 245 – 257.

<https://doi.org/10.1016/j.cor.2017.02.015>

[38] L. A. Zadeh, A. M. Abbasov, and S. N. Shahbazova. 2015. Analysis of Twitter hashtags: Fuzzy clustering approach. In *2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC)*. 1–6. <https://doi.org/10.1109/NAFIPS-WConSC.2015.7284196>

4.3 Smart Governance through Opinion Mining of Public Reactions on Ordinances

Abstract: This work focuses on the area of Smart Governance in Smart Cities, which entails transparency in government through public involvement. Specifically, it describes our research on mining urban ordinances or local laws and the public reactions to them expressed on the social media site Twitter. We mine ordinances and tweets related to each other through their mutual connection with Smart City Characteristics (SCCs) and conduct sentiment analysis of relevant tweets for analyzing opinions of the public on local laws in the given urban region. This helps assess how well that region heads towards a Smart City based on (1) how closely ordinances map to the respective SCCs and (2) the extent of public satisfaction on ordinances related to those SCCs. The mining process relies on Commonsense Knowledge (CSK), i.e., knowledge that is obvious to humans but needs to be explicitly fed into machines for automation. CSK is useful in filtering during tweet selection, conducting SCC-based ordinance tweet mapping and performing sentiment analysis of tweets. This paper presents our work in mapping ordinances to tweets through single or multiple SCCs and opinion mining of tweets along with an experimental evaluation and a discussion with useful recommendations.

Keywords: *Big Data; Classification; Commonsense Knowledge; Data Mining; Local Laws; Machine Learning; Sentiment Analysis; Smart Cities; Social Media; Urban Policy*

(Chapter 4.3 reused the previously published paper Puri, M., Varde, A., Du, X., & de Melo, G. (2018), Smart Governance Through Opinion Mining of Public Reactions on Ordinances, *IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, <https://doi.org/10.1109/ictai.2018.00131>).

4.3.1 Introduction

This paper centers on enhancing Smart Governance, which falls under the umbrella of Smart Cities. Specifically, we seek to analyze tweets about ordinances or local laws in a given urban region, which represent opinions of people on related topics. This aids in understanding people's reactions to the respective urban policies addressed in these ordinances. An important goal of this work is to assess how well the concerned urban area is progressing towards being a Smart City based on the ordinances passed and the public reactions to them. Figure 4.16 shows different Smart City Characteristics, as widely accepted in the literature [1, 2].



Figure 4.16: Smart City Characteristics [3]

Twitter is one of the biggest sources for data mining, with about 350 million users and over 500 million tweets sent per day on a variety of topics. Hence, Twitter is a valuable source of data on public reactions to ordinances. Its micro-blogging nature is useful with respect to the brevity of the information to be analyzed, since each tweet is limited to 280 characters

We map tweets to corresponding ordinances through their mutual connection with respective Smart City Characteristics based on a measure of semantic relatedness. A trivial keyword matching approach of trying to connect ordinances directly to tweets does not suffice as they both contain intricate and heterogeneous natural language. Moreover, ordinances and tweets both constitute big data, as there are thousands of ordinances and millions of tweets. Hence, obtaining a direct mapping is challenging. Existing techniques from the field of machine learning

[4] are not useful to learn these mappings, as they need significant volumes of data to train the models. Ours is pioneering work in the area of ordinance mining and hence we do not possess such large volumes of training data for our mapping task.

Instead, our proposed mapping technique takes into account generic connections between ordinances and tweets, through SCCs. We rely on a transitive property: “If ordinances map to one or more SCCs and if tweets map to the same SCCs, the ordinances are likely to be broadly related to the respective tweets”. This is due to the finite nature of classical sources of SCC data [1, 2] which possess a limited set of identifying features that can be used for mapping. Hence, this transitive mapping approach seems more feasible, since by using this, we can bypass having to map millions of tweets to thousands of ordinances directly.

Since a single ordinance or tweet can map to one or more SCCs, we develop an algorithm for SCC mapping accordingly. In the process of ordinance–tweet mapping, we make use of Commonsense Knowledge (CSK) from sources such as WebChild [5] and WordNet [6]. The use of CSK is vital to measure semantic relatedness in a more informed way. The terms encountered in classical SCC sources may not directly appear in relevant ordinances and tweets. For example, an ordinance or tweet may contain the term “Pre-Kindergarten for all”, which pertains to the characteristic of Smart People (since one feature of this SCC is “21st century education”). Humans can intuitively make the connection upon reading the content of the ordinances or tweets and the SCC features. However, to automate the mapping, this knowledge needs to be

explicitly fed to the algorithm, which is done by deploying concepts, properties and relationships in the CSK source WebChild [5].

Likewise, CSK also helps in filtering out unwanted tweets from among the millions of tweets initially obtained through the use of terms in SCC Domain KBs (derived using CSK and SCC sources). Accordingly, after finding relevant tweets and mapping them to their respective ordinances, our next step is sentiment analysis. Here we conduct a sentiment polarity based classification of tweets, using SentiWordNet [7] derived from the CSK source WordNet [6], in order to gauge public opinion. CSK plays a role here by connecting relevant terms to appropriate sentiment words, thereby capturing subtle human judgment. The outcome of the sentiment analysis can be used to provide feedback to urban agencies on their policies.

Our research is thus potentially useful to urban agencies for assessment in relevance to Smart Cities. By the identification of SCCs addressed in local laws, information can be provided on how well urban policies are aiming towards Smart City development. Further, the knowledge discovered by mining data on Twitter and mapping ordinances to tweets is helpful to urban agencies to evaluate public contentment. This would help them determine their Smart City public appeal and accordingly enforce appropriate legislation to enhance urban management as needed. This work thus falls under the characteristic of Smart Governance (or Smart Government), which embodies the involvement of the public in decision making and transparency of the whole governing process.

The rest of this paper is organized as follows. Section II overviews related work in the area. Section III describes our approach on mapping ordinances and tweets through SCCs. Section IV focuses on sentiment analysis of tweets. Section V summarizes our experimental evaluation, while Section VI presents discussion and challenges. Section VII states the conclusions and ongoing work.

4.3.2 Related Work

Recently, there has been significant interest in Smart Cities, and a number of developments have occurred in this field. For example, in Barcelona, buses are now configured to run on optimal routes for better power consumption [2]. In Amsterdam, there are canal lights that adjust their brightness automatically depending on how often they are used by pedestrians [2]. These initiatives fall under the category of Smart Mobility, whereas other works in this area consider decisions for autonomous vehicles [8] or saving trips in delivery and pickup [9]. Copenhagen (ranked number one in the 2017 Smart City Index) has buildings with sensors for air quality and climate control and smart meters for intelligent control of energy consumption [10]. Health and safety issues, considering AQI (Air Quality Index) standards for human health, are addressed in [11]. Such works bridge the areas of Smart Living and Smart Environment. Further research affects the Smart Environment domain by relying on cloud computing solutions rather than on-site storage solutions for mid-sized data centers [12]. In Smart Economy, cost savings and profit

distribution has been studied [9], as well as important issues for business and household taxation [13] and data center cost savings [14]. Another study [15] considers urban policy data, modeling it using data warehousing and XML databases, and conducts preliminary data mining using classical techniques such as association rules and decision tree classifiers [4]. This heads towards Smart Governance.

Various studies have been conducted on data mining from social media. An important piece of work in this field focuses on unsupervised modeling and trend detection in social media [16]. In an experiment in 2015, a fuzzy-based method was used to pre-process and analyze Twitter hashtags so as to study trends in hashtag popularity [17]. However, such approaches are not feasible in our current project, which involves mapping tweets to a pre-existing set of ordinances without existing training data. Another example of research in this field focuses on supervised classification. While standard methodologies exist [4], these require massive training sets to account for the short length and variability of tweets. Another approach is in the TweetSift system [18], which classifies tweets by topic and uses external entity knowledge and word embeddings that are topic-enhanced. These lead to topic-specific word embeddings such that different senses of equivocal words obtain different representations. This process also comes with the assumption that the knowledge bases provide very relevant signals regarding different entities, such as particular uses on Twitter. In contrast to this, our approach uses generic Commonsense Knowledge and does not need labeled training data. In addition, works by other

researchers have not considered the setting of ordinances and Smart City Characteristics. Hence, our work is among the first to explore this area.

The use of Commonsense Knowledge in the domain of Smart Cities is in its early stages as well, with important opportunities to enhance a number of research areas. Works in the field [19, 20] describe the use of CSK for various tasks such as: enhancement of autonomous vehicles that can contribute to Smart Mobility; and machine translation in writing aids that have the potential to impact the Smart People characteristic. The use of CSK would enhance the reasoning capabilities of machines, which would enable them to solve several challenging problems. The use of such CSK-enabled machines in various aspects of automation in Smart Cities would make them even smarter. Our work in this paper addresses the theme of Commonsense Knowledge in Smart Cities. We make use of CSK in various parts of the ordinance and tweet mining process, including mapping as well as sentiment analysis.

4.3.3 Approach for Mapping

4.3.3.1 SCC Based Mapping Process

Our proposed approach for mapping tweets and ordinances to each other through SCCs is illustrated in Figure 4.17. The main source for the SCCs in our approach is a technical report [2] published by the Vienna University of Technology. This report describes Smart Cities as having six different Smart City Characteristics: Smart Mobility, Governance, People, Living, Economy

and Environment.

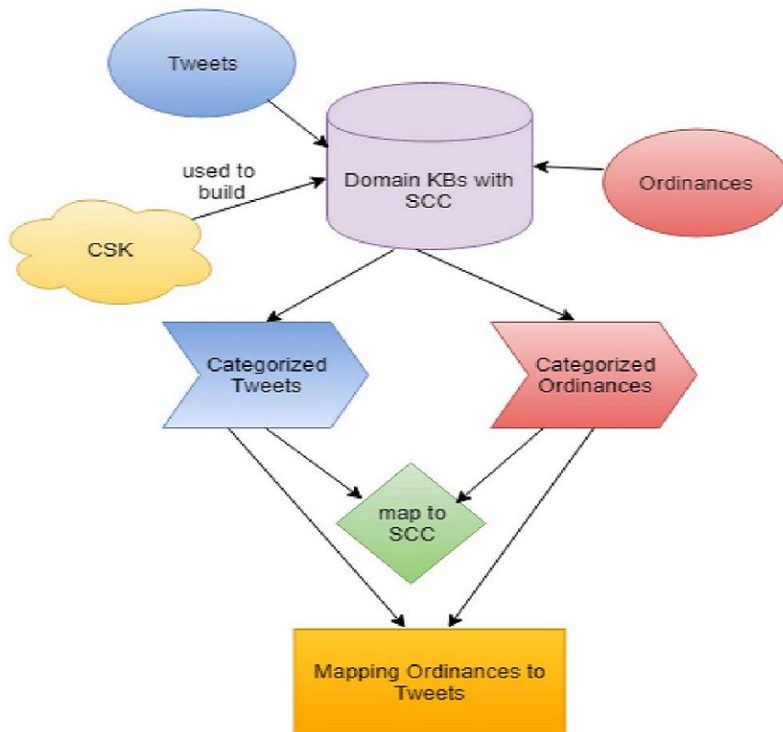


Figure 4.17: Illustration of the Mapping Process

Our own work in this research project is on transparency in government through public involvement, which falls under the realm of *Smart Governance*. Hence, we enumerate its specifics, just to give an example of SCC features.

Smart Governance

- Transparency in government
- Optimization of public service and administration
- Direct involvement in public policies
- Citizen participation

-
- Positive and open communication channel with citizens
 - More informed decisions by feedback and engagement

4.3.3.2 Role of Commonsense Knowledge

Considering the SCC Smart Governance elaborated herewith, if ordinances or tweets relate to any of the specifics described above, they are most likely related to Smart Governance. However, most tweets and ordinances may not directly contain these specific features such as *citizen participation* in their descriptions and hence machines may not be able to recognize them. Humans, through common sense, will be able to spot such features and thus infer that an ordinance or tweet maps to one or more SCCs. Hence, to automate the mapping process, we make use of a commonsense knowledge (CSK) repository called WebChild [5]. This consists of various commonsense concepts derived from vast quantities of data available on the Web, as well as their properties and relationships. A snapshot of the WebChild browser is given in Figure 4.18, which depicts results related to the concept

The screenshot shows the Commonsense Browser interface for the word "nation". At the top, the title "Commonsense Browser" is displayed in a dark header, with the word "nation" in a search box to its right. Below the header, the word "nation" is listed in blue. A definition follows: "a politically organized body of people under a single government; 'the state has elected a new president'; 'come to the nation's capitol'; 'the country's largest manufacturer'; 'an industrialized land'". Below the definition is a table with six rows, each representing a different semantic category for the word "nation".

TYPE OF	political_unit
	Related to group , under the category of diplomacy
COMPARABLES	nation,war policy,nation government,nation canada,nation
ACTIVITIES	make nation build nation represent nation lead nation serve nati
HAS MEMBER	quorum membership man
PHYSICAL PROPERTIES	considerable fast large strong slow More

Figure 4.18: Partial Screenshot of WebChild

In order to induce a mapping of ordinances and tweets using Smart City Characteristics, we construct domain specific knowledge bases (Domain KBs) using various SCC sources guided by CSK. Sample Domain KBs for three SCCs are shown in Figure 4.19. WebChild is the primary source for CSK in our research. We also use the lexical database WordNet [6] while building the KBs. By using these SCC Domain KBs, CSK principles are used to find connections between terms x in every ordinance text O and SCCs. This is denoted by $C(O, x)$. For example, if the

ordinance text includes the term *unemployment*, the CSK concepts are useful to find its semantic relatedness with the SCC Smart Economy. This is because the terms in the corresponding SCC Domain KB have been derived from classical SCC sources as well as the CSK source WebChild to help make the connection. These Domain KBs guide the mapping of ordinances and tweets through SCCs in our algorithm, as outlined next.

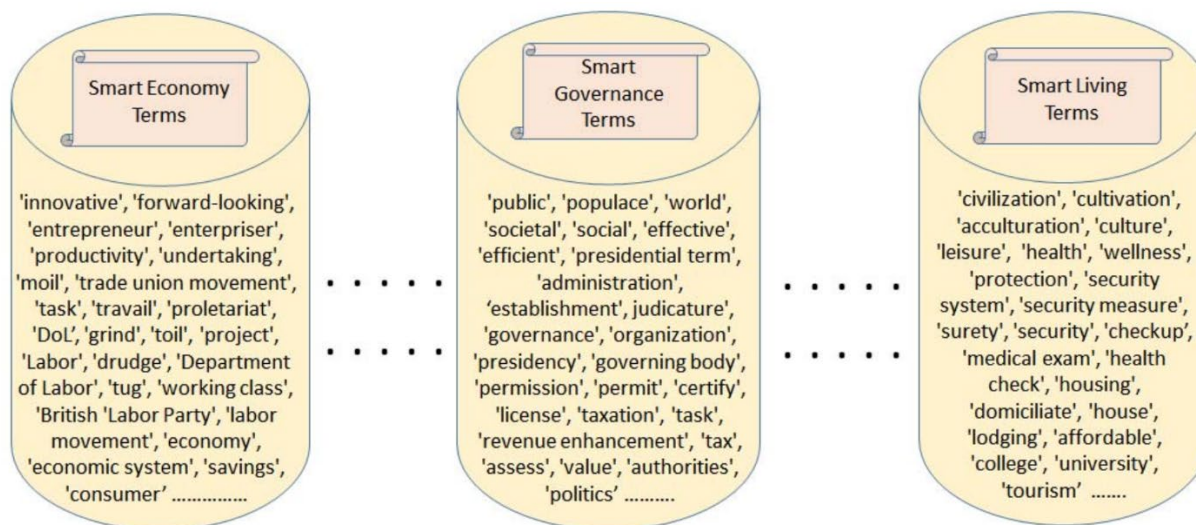


Figure 4.19: Sample SCC Domains

4.3.3.3 Mapping with Single or Multiple SCCs

An ordinance or a tweet can relate to one or more SCCs. Accordingly, each ordinance / tweet is mapped to SCCs with the most relevant number of features. Mapping to different SCCs is based on the weights of relevant terms assigned to them. For instance, if an ordinance/tweet has two terms related to Smart Economy and one related to Smart Governance, the ratio of

mapping Smart Economy:Smart Governance is 2:1. If an ordinance or tweet consists of a significant number of terms related to a single SCC, the other terms may be ignored as they may not be relevant to the intention of the tweet. This is determined by a threshold, which can be adjusted as needed.

Based on this discussion, our proposed algorithm for mapping ordinances tweets through SCCs is as follows.

ALGORITHM 1: ORDINANCE-TWEET-SCC MAPPING

1. **for each** SCC S_i **do**:
 2. build domain KB K_i
 3. $A \leftarrow \emptyset$
 4. **for each** ordinance O_i **do**:
 5. **for each** SCC S_j **do**:
 6. $L_{i,j} \leftarrow \sum_{x \in K_j} C(O_i, x)$
 7. $A \leftarrow A \cup \{(O_i, S_j) \mid j = \operatorname{argmax}_j L_{i,j}\}$
 8. **for each** tweet T_i **do**:
 9. **for each** SCC S_j **do**:
 10. $M_{i,j} \leftarrow \sum_{x \in K_j} C(T_i, x)$
 11. $A \leftarrow A \cup \{(T_i, S_j) \mid j = \operatorname{argmax}_j M_{i,j}\}$
 12. $\theta \leftarrow \{(O_i, T_k) \mid \exists S_j : (O_i, S_j) \in A \wedge (T_k, S_j) \in A\}$
 13. **return** θ
-

Hence, this algorithm incorporates the transitive property such that if ordinances map to one or more SCC(s) and tweets map to the same SCC(s), then the ordinances broadly map to the tweets. This can thus be used to determine the actual ordinance to tweet mapping on a broad scale, which is the final output of our mapping process. Note that we do not deal with the finest levels of granularity in this ordinance-tweet mapping as of now. We maintain a more generic connection at the level of relevance to SCCs, since an important aspect of this work entails heading towards Smart Cities.

4.3.4 Sentiment Analysis of Tweets

4.3.4.1 Process of Sentiment Analysis

Sentiment analysis of the text involves determining the emotion expressed in a particular piece of writing. The specific task of sentiment polarity classification focuses on assessing whether the sentiment is “positive”, “negative” or “neutral”, and sometimes the extent to which it heads in that respective direction, i.e., “strongly positive” etc. This often serves the purpose of opinion mining, i.e., discovering knowledge from people’s opinions or reactions. Sentiment analysis has a number of applications in different areas as follows.

- **Business:** Used by companies to obtain product, service and brand satisfaction from customers
- **Politics:** To gauge the population’s interest in various political events
- **Social events:** To understand people’s reactions to economic and social events in general

4.3.4.2 Opinion Mining using CSK

In our research, we conduct sentiment analysis to discover knowledge specifically with respect to opinion mining of tweets on ordinances. This is conducted after the mapping of ordinances to tweets (as explained in the previous subsection). The primary database used for Sentiment Analysis in this work is SentiWordNet [7]. The SentiWordNet source has been built

for guiding sentiment classification and opinion mining. This is an enhanced version of the CSK source WordNet [6]. It groups words into synonym sets (synsets) annotated by how positive the terms are. Accordingly, words are classified as positive, negative or neutral based on polarity of terms.

In SentiWordNet, different meanings exhibited by the same word can have different sentiment scores. For example, the word *estimable* when relating to computation has a neutral score of 0.0, while the same word in the sense of *deserving respect* is assigned a positive score of 0.75. The process we deploy for sentiment analysis of tweets constitutes a semisupervised learning method using SentiWordNet. Through this, subtle human judgment through commonsense in understanding emotions is embodied in the mining processes with specific reference to context.

4.3.4.3 Algorithm for Polarity Classification

Based on the given discussion, our proposed algorithm for sentiment analysis of tweets through polarity classification is as follows.

ALGORITHM 2: TWEET POLARITY CLASSIFICATION

1. **for each** tweet t_i **do**:
 2. **if** not (t_i relevant according to SCC KB):
 3. **continue** (with next tweet)
 4. map t_i to ordinances using Algorithm 1
 5. $W_i \leftarrow$ set of words in t_i
 6. **for each** $w \in W_i$ **do**:
 7. $s_w \leftarrow$ polarity score of w in SentiWordNet
 8. $s_i \leftarrow \sum_{w \in W_i} s_w$
 9. **return** final polarity scores s_i for relevant t_i
-

Based on this algorithm, we classify thousands of tweets that we obtain from Twitter. Note that the selection of relevant tweets and also the mapping of tweets to their respective ordinances is guided by CSK. The SCC Domain KBs derived from WebChild and WordNet serve to filter out unwanted tweets as a first step, followed by the mapping of tweets to relevant ordinances using SCCs as a next step. The results of our polarity classification using this approach are presented in the experimental evaluation section, after the results of ordinance and tweet mapping through SCCs.

4.3.5 Experimental Evaluation

4.3.5.1 Ordinance to SCC Mapping

The source of the ordinances in our experiments herewith is the NYC metropolitan legislative council website [21]. A partial snapshot of this is seen in Figure 4.20. Consider the following sample ordinance obtained from this NYC source.

The screenshot shows the website for The New York City Council, led by Corey Johnson, Speaker. The navigation menu includes Council Home, Legislation, Calendar, City Council, and Committees. A search bar is present with filters for Session (2018-2021) and Local Law. Below the search bar, there are buttons for 'Search Legislation' and 'Help'. The main content is a table of 20 records, with columns for File #, Law Number, Type, Status, Committee, Prime Sponsor, Council Member Sponsors, and Title. The table lists five entries, each representing a local law introduced in 2018 and enacted.

File #	Law Number	Type	Status	Committee	Prime Sponsor	Council Member Sponsors	Title
Int 0001-2018	2018/084	Introduction	Enacted	Committee on Finance	Daniel Dromm	1	A Local Law in relation and the date prior to v council shall submit a the preliminary certific capital projects, the di publication by the dire expenditures, the date
Int 0600-2018	2018/085	Introduction	Enacted	Committee on Housing and Buildings	Corey D. Johnson	11	A Local Law to amend rent stabilization laws
Int 0410-2018	2018/086	Introduction	Enacted	Committee on Youth Services	Corey D. Johnson	8	A Local Law to amend runaway and homeless
Int 0490-2018	2018/087	Introduction	Enacted	Committee on Youth Services	Vanessa L. Gibson	9	A Local Law to the adi runaway and homeless
Int 0556-2018	2018/088	Introduction	Enacted	Committee on	Ritchie L.	10	A Local Law to amend

Figure 4.20: NYC Ordinance Website

Sample Ordinance: *A Local Law to amend the administrative code of the city of New York, in relation to recycling outreach, education and enforcement; and to repeal subdivisions d and e of section 16-305 and section 16-311 of the administrative code of the city of New York, relating to source separation of recyclable materials and recycling centers.*

The mapping of this sample ordinance to its relevant SCC(s) is shown in Table 4.7.

Table 4.7: Mapping of a Sample Ordinance to SCC(s)

Smart City Characteristics	Occurrences
Economy	0
Environment	4
Governance	0
People	0
Mobility	0
Living	1

Thus, it can be seen that we get the following mapping results for this ordinance: Smart Economy = 0%, Smart Environment = 80%, Smart Governance = 0%, Smart People = 0%, Smart Mobility = 0%, Smart Living = 20%. Hence, we observe that the given ordinance maps most closely with Smart Environment, but also to some extent with Smart Living. Based on the threshold used in this execution, it is mapped to Smart Environment:Smart Living with a ratio of 80:20, stating the outputs as percentages.

4.3.5.2 Tweet to SCC Mapping

We extract location-specific geo-tagged tweets posted by the general public from NYC and map these to SCCs using the process described in the previous section. We mine approximately 5,000 tweets in the experiments shown here. A small random subset of tweets as mined before mapping (regardless of their relevance to ordinances) is given below.

Tweet T1: *"RT @OccAware: @RachelNotley I am contacting you to tell your government does not have my support to put BC land, water and our tourist economy at risk while trashing*

indigenous rights”

Tweet T2: *”RT @verge: Elon Musk made history launching a car into space. Did he make art too?”*

Tweet T3: *”Concerns about HB1319 the predatory lending bill. High interest rates, short-term loans for poor. Veterans groups social service gps have spoken against.”*

Tweet T4: *”Adding plants to your home or office has so many benefits such as reducing stress and increasing productivity! Not to mention they add a personal touch to your surroundings. ”*

Tweet T5: *”RT @CoachKCullen: AI a strong non-digital learning environment is needed before you introduce digital learning environment. Bad practices can just multiply the more technology is introduced satchat”*

These tweets are subjected to mapping using Algorithm 1. Each tweet can have terms related to one or more SCC(s). Different SCC terms from the input tweets are considered and accordingly weights are assigned to each SCC. These determine the relevance of the tweets to the SCC(s) based on the threshold levels used in the experiments.

We have developed a GUI that accepts a tweet as the input and determines the closest matching SCC(s) as the output, or returns *“No matches”* if absolutely no SCC gets matched during the mapping process. Figure 4.21 shows an example of tweet to SCC mapping using this

GUI.

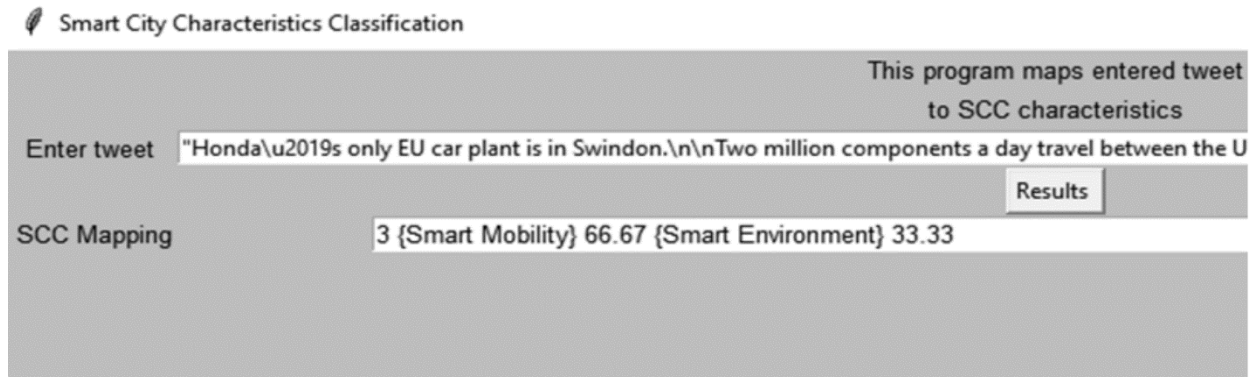


Figure 4.21: Screenshot of GUI for Tweet to SCC Mapping

4.3.5.3 SCC Mapping Assessment

After the mapping is conducted, its outcomes are assessed by domain experts from the department of Earth and Environmental Studies. Ground truth is defined by the experts such that a mapping is identified as correct if it agrees with the judgment of the expert and incorrect otherwise. For example, if the system identifies the mapping as Smart Economy:Smart Governance with a ratio of 60:40 and the expert agrees, this is considered to be a correct mapping. On the other hand, if the expert disagrees with the single or multiple SCC(s) identified in the mapping, or considers their ratios to be highly inappropriate, this is treated as an incorrect mapping. Assessment is then conducted using the standard *Precision* metric [4] as the ratio of correct mappings to all mappings. Thus, $Precision = \text{Correct Mappings} / (\text{Correct Mappings} + \text{Incorrect Mappings})$. Based on this evaluation, Figure 4.22 shows a chart summarizing the

domain expert assessment of tweet and ordinance mappings with respect to the SCCs.

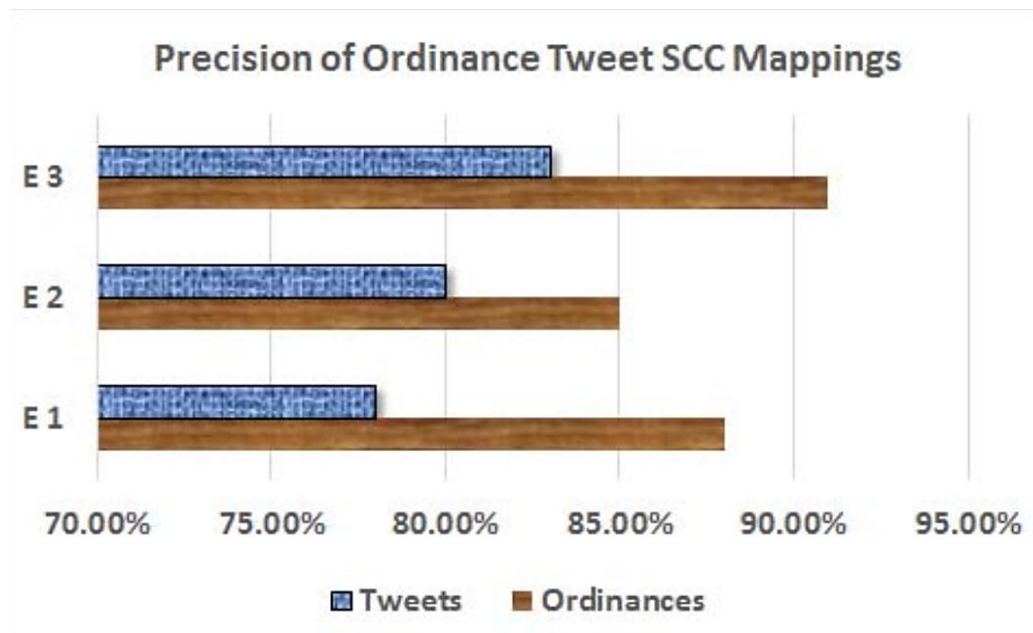


Figure 4.22: Summary of Mapping Assessment

The evaluation of the mapping results obtained can be interpreted as follows. With respect to Expert 1, around 88% of the ordinances are correctly mapped to their respective multiple SCCs, while approximately 78% of the tweets are appropriately mapped. Hence, the ordinance to tweet mapping precision on a broad range would be at best 78%, considering their mutual connection with multiple SCCs, as per the ground truth defined by this expert. Likewise, on the whole, the ordinance to multiple SCC mapping precision is observed to be in the range of the higher 80s, while that of the tweets to multiple SCC is around the lower 80s. Thus, the ordinance to tweet mapping precision through mutually relevant SCCs would be at best in the lower 80s.

Note that this is an enhancement over our early work [22], where we considered mapping of

ordinances and tweets to single SCCs only, based on the closest match. It was concluded therein [22] that the mapping precision needs to be higher to conduct opinion mining of tweets with respect to relevant ordinances and SCCs. Thus, we proposed enhanced algorithms in this paper and also refined domain KBs on SCCs with more intricate CSK concepts

4.3.5.4 Ordinance to Tweet Mapping Output

The precision ranges obtained in the ordinance to SCC and tweet to SCC mappings are considered acceptable by our domain experts to proceed with further work. Hence, it is feasible to use these mappings in order to output the ordinance–tweet mappings. We present examples herewith of broadly related correct ordinance to tweet mappings as determined through their mutual SCC connections.

Mapping Example 1: In this excerpt, both the ordinance and tweet map to Smart Economy.

***Ordinance:** A Local Law to amend the administrative code of the city of New York, in relation to authorizing an increase in the amount to be expended annually in seven business improvement districts and two special assessment districts.*

***Tweet:** @DowntownNYC is one of NYCs largest Business Improvement Districts, which works to enhance the quality of life in #LowerManhattan. Attend our #YLG #SecretsofSuccess on 5/11 ft. @JessLappin for an exclusive talk about what it's like to lead a #BID.*

Mapping Example 2: Here, the ordinance and tweet both map to Smart Mobility as well as

Smart Living.

Ordinance: *A Local Law to amend the administrative code of the city of New York, in relation to parking violations issued for the failure to display a muni-meter receipt.*

Tweet: *The new NYC Parking Ticket: Pay or Dispute app makes the process of paying or disputing a violation easier and faster. #nycpayordispute*

Likewise, various such ordinance-tweet mappings through their semantic relatedness with the SCC(s) set the stage for sentiment analysis of tweets to analyze public opinion.

4.3.5.5 Results of Sentiment Analysis on Tweets

In the experiments shown here, we use tweets from the NYC region after filtering out unwanted ones guided by SCC domain KBs derived through CSK. We map these to ordinances through our proposed mapping approach in Algorithm 1. Sentiment analysis of these tweets is then conducted using Algorithm 2. Examples of these are summarized next.

NYC Tweet Example 1:

"#FairFares is just common sense. What it will do is level the playing field so every #NYC resident can access the @MTA. Pretty simple idea. Not only is it the right thing to do but it will also help to grow our economy. #nowisthetime"

This tweet maps to Smart Mobility and Smart Economy. For this tweet, the net score is 0.43. It obtains a positive score of 0.72 and a negative score of -0.29.

NYC Tweet Example 2:

"#NewYork's statue of Liberty was RED before pollution turned it green"

This tweet maps to Smart Environment. For this tweet, the net score is -0.21. Its positive score is 0.15, while its negative score is -0.36.

Likewise, based on several examples, the combined results of sentiment analysis on tweets are obtained. These are illustrated in Figure 4.23. The pie chart in this figure summarizes the overall public reactions as determined from tweets across all SCCs. Since the percentage of positive tweets is the greatest among these, we can conclude that the people of NYC seem to approve of the concerned policies (to a greater extent than complaining about them or being neutral). However, positive sentiments are expressed by less than half the public, which means that there is potential for improvement with respect to heading towards a Smart City.

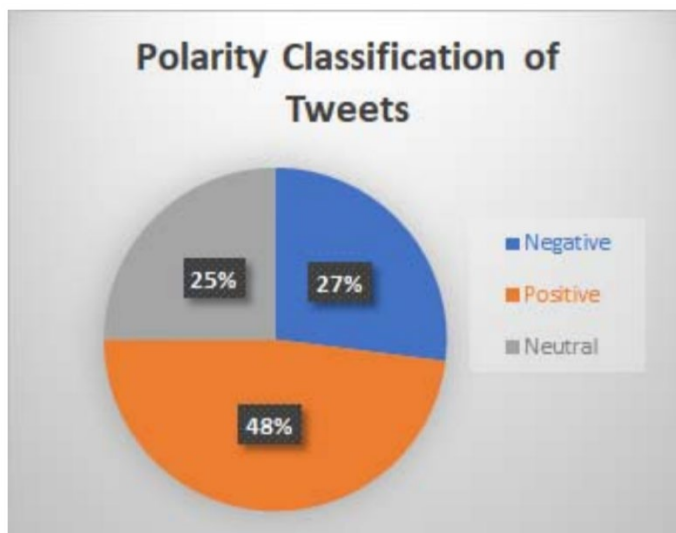


Figure 4.23: Polarity Classification of Tweets on all SCCs

Additionally, we analyze specific public contentment for each SCC and obtain the results as shown in Table 4.8. These numbers help urban agencies assess how satisfied people are on matters pertaining to each individual SCC. For example, here the public seems least satisfied with issues related to Smart Environment. Thus, feedback can be provided to urban agencies that they need to address policies to make the environment smarter, e.g., develop more energy efficient systems.

Table 4.8: Public Contentment for Each SCC

Aspect	Positive Polarity
Smart Economy	47%
Smart Mobility	48%
Smart Environment	33%
Smart Governance	45%
Smart People	52%
Smart Living	56%

On the whole, public contentment seems fine considering the various SCCs in our sample. Hence, from this analysis, we can infer that based on our data used and the accuracy of our results that the metropolitan region of NYC seems to tend fairly well towards being a Smart City, with scope for further enhancement. A summary of these results can be provided as feedback to urban agencies to help them make decisions in outlining further legislative policies.

4.3.6 Discussion and Challenges

The overall evaluation in this paper reveals that many ordinances and tweets get correctly mapped to each other on a broad level of semantic relatedness. Domain experts confirm that this is acceptable for current use, although it could get better in the future. Accordingly, upon closer observation we find that a few mappings of ordinances and tweets are imprecise with respect to their SCC identification. Tweet mapping particularly needs further refinement. Moreover, some ordinance–tweet mapping outputs are found to be imprecise, in spite of their correct mutual SCC connection being identified in the ordinance–SCC and tweet–SCC mappings.

In order to address these issues, we need further research with respect to the refinement of the content in the SCC Domain KBs and also the levels of granularity in ordinance to tweet mappings. SCCs can still be used as a mutual connection; however, this could be done with more specific features that refer to their actual details. Direct mapping of ordinances to tweets can be a next step addressing intricate natural language and other aspects. Note that our proposed SCC-based mapping approach substantially narrows down the sample space for potential direct ordinance–tweet mappings. In the absence of this approach, the large quadratic number of pairs is very challenging, given ordinances and tweets in the order of thousands and millions, respectively.

Sentiment analysis of tweets has also been assessed by our domain experts through a

process similar to that of mappings (details not shown herewith). The polarity classification is found to be accurate for the most part, i.e., if the overall classification of a tweet is positive as identified by our approach, it indeed expresses a positive sentiment with respect to ground truth defined by experts etc. Due to this, we have considered it feasible to use the results of sentiment analysis for summarizing the public reactions on ordinances as shown in Figs. 10 and 11. We can therefore use these outcomes to provide recommendations to urban agencies. However, it is to be noted that we have found a few incorrect polarity classifications. This can be attributed to the fact that tweets in general do not follow a systematic grammar structure, making it difficult to derive semantic patterns in some cases.

Based on our work, we outline the following challenges in dealing with tweets for mapping and polarity classification.

- The usage of informal language in tweets makes it difficult for pre-processing using basic Natural Language Processing (NLP) techniques.
- There is rampant usage of acronyms in Twitter, which do not relate to standard vocabulary sources.
- Tweets show ambiguous characteristics with respect to NEE (Named Entity Extraction) as well as NED (Named Entity Disambiguation).

Addressing these challenges and the other ongoing tasks is nontrivial and constitutes further work. Ongoing research involves addressing these issues for enhanced knowledge discovery

from ordinance and tweets in the future.

4.3.7 Conclusions

In this paper, we mine ordinances and their public reactions to gauge how well a given urban region tends towards a Smart City. The novelty of this work includes: (1) being among the first to address ordinance mining, especially for Smart Cities; (2) implementing single and multiple SCC-based ordinance to tweet mapping; and (3) deploying commonsense knowledge in mining (for tweet selection, mapping processes and polarity classification). The overall challenges include: (1) dealing with intricate natural language in ordinances and tweets; (2) handling big data of the order of thousands and millions respectively; and (3) considering further issues in tweets, e.g., acronyms and ambiguity.

We evaluate our work with real data from NYC sources. The ordinance to SCC mapping precision is found to be in the higher 80% range while tweet to SCC mapping precision is in the lower 80s on an average. The polarity classification of tweets suggests that the majority of the public is satisfied with the topics that the ordinances cover. Yet, positive sentiment amount to lower than 50%, implying scope for improvement. Through this analysis, feedback can be provided to urban agencies for policy decisions. This work contributes to Smart Governance, by public involvement entailing transparent decision-making.

Our ongoing work seeks to enhance the mapping precision and to address finer levels of

granularity in the ordinance– tweet mapping. We aim to improve tweet sentiment analysis, incorporating advanced concepts in NLP and Machine Learning. The long-term goal of our research is to aid urban regions in enhancing legislation related to Smart Cities. This constitutes multidisciplinary work in Artificial Intelligence and Environmental Management.

4.3.8 Acknowledgments

We thank Robert Taylor, Clement Alo and their PhD students from the Department of Earth and Environmental Studies at Montclair State University (NJ) for useful inputs from a domain-specific angle. Early work in this area started when Aparna Varde was a visiting researcher at Max-Planck-Institut für Informatik (Saarbrücken, Germany) in Gerhard Weikum’s Databases and Information Systems group. Gerard de Melo’s research at Rutgers University (NJ) is funded in part by ARO grant no. W911NF-17-C-0098 (DARPA SocialSim program).

4.3.9 References

© 2018 IEEE. Reprinted, with permission, from Manish Puri, Aparna Varde, Xu Du and Gerard de Melo, Smart Governance through Opinion Mining of Public Reactions on Ordinances, IEEE International Conference on Tools with Artificial Intelligence (IEEE ICTAI 2018), Volos, Greece, Nov 2018, pp. 838 - 845.

[1] The IEEE Smart Cities Technical Community. <https://smartcities.ieee.org/>, 2018.

-
- [2] TU Wien (Vienna University of Technology). European smart cities. Technical report, Vienna, Austria, 2015.
- [3] Smart City Consortium. Background.
<https://smartcity.org.hk/index.php/aboutus/background>.
- [4] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA, 2011.
- [5] Niket Tandon, Gerard de Melo, Fabian Suchanek, and Gerhard Weikum. Webchild: Harvesting and organizing commonsense knowledge from the web. In *ACM International Conference on Web Search and Data Mining, WSDM*, pages 523–532, 2014.
- [6] George Miller and Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, 1998.
- [7] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *International Conference on Language Resources and Evaluation, LREC*, May 2010.
- [8] Katherine Rose Driggs-Campbell, Victor Shia, and Ruzena Bajcsy. Decisions for autonomous vehicles: Integrating sensors, communication, and control. In *International Conference on High Confidence Networked Systems, HiCoNS '14*, pages 59– 60. ACM, 2014.
- [9] Yang Yu, Qi Lou, Jiafu Tang, Junwei Wang, and XiaoHang Yue. An exact decomposition method to save trips in cooperative pickup and delivery based on scheduled trips and profit

distribution. *Computers and Operations Research Journal*, 87:245 – 257, 2017.

[10] EasyPark Group. The 2017 smart cities index. <https://easyparkgroup.com/smart-cities-index>, 2017.

[11] Xu Du, Onyeka Emebo, Aparna Varde, Niket Tandon, Sreyasi Nag Chowdhury, and Gerard Weikum. Air quality assessment from social media and structured data: Pollutants and health impacts in urban planning. In *IEEE International Conference on Data Engineering (ICDE) - Workshops*, pages 54–59, 2016.

[12] Michael J. Pawlish, Aparna S. Varde, and Stefan A. Robila. The greening of data centers with cloud technology. *International Journal of Cloud Applications and Computing*, 5(4):1–23, October 2015.

[13] G. Leef. *Smart Economics: Commonsense Answers to 50 Questions about Government, Taxes, Business and Households*. Foundation for Economic Education, 2007.

[14] Michael Pawlish and Aparna Varde. Free cooling: A paradigm shift in data centers. In *IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 347–352, 2010.

[15] Xu Du, Diane Liporace, and Aparna Varde. Urban legislation assessment by data analytics with smart city characteristics. In *IEEE Ubiquitous Computing, Electronics and Mobile Communications Conference (UEMCON)*, pages 20–25, 2017.

[16] Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in twitter. In

Workshop on Social Media Analytics, SOMA '10, pages 80–88. ACM, 2010.

[17] L. A. Zadeh, A. M. Abbasov, and S. N. Shahbazova. Analysis of twitter hashtags: Fuzzy clustering approach. In *2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC)*, pages 1–6, 2015.

[18] Quanzhi Li, Sameena Shah, Xiaomo Liu, Armineh Nourbakhsh, and Rui Fang. Tweetsift: Tweet topic classification based on entity knowledge base and topic enhanced word embedding. In *ACM International on Conference on Information and Knowledge Management, CIKM*, 2016.

[19] Priya Persaud, Aparna Varde, and Stefan Robila. Enhancing autonomous vehicles with commonsense: Smart mobility in smart cities. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI) - Smart Cities Workshop*, pages 1008–1012, 2017.

[20] Niket Tandon, Aparna S. Varde, and Gerard de Melo. Commonsense knowledge in machine intelligence. *ACM SIGMOD Record*, 46(4):49–52, 2017.

[21] The New York City Council. Legislative research center web page.
<http://legistar.council.nyc.gov/>, 2018.

[22] Manish Puri, Xu Du, Aparna Varde, and Gerard de Melo. Mapping ordinances and tweets using smart city characteristics to aid opinion mining. In *The Web Conference, WWW Companion Volume, Satellite Track on Social Sensing and Enterprise Intelligence SSEI*, pages

1721–1728. W3C, 2018.

Chapter 5

5. Result Dissemination and Application

5.1 LSOMP: Large Scale Ordinance Mining Portal

Abstract: We propose a novel scalable Web portal called *LSOMP (Large Scale Ordinance Mining Portal)* to analyze ordinances and their tweets (of the order of thousands and millions). It entails commonsense knowledge (CSK) and natural language processing (NLP), disseminating ordinance-tweet mining results via interactive graphics and Question Answering (QA).

(Chapter 5.1 reused the previously published paper Du, X., Kowalski, M., & Varde, A. (2020), LSOMP: Large Scale Ordinance Mining Portal. In *IEEE International Conference on Big Data (IEEE BigData 2020)*, Atlanta, GA).

5.1.1 Problem Definition

The complex legalese in ordinances (local laws) on urban policy and their informally toned public reactions (often expressed on Twitter) need analysis so that a broad spectrum of users can comprehend them. This motivates mining big data on ordinances and their tweets to discover knowledge, e.g., how well the enacted ordinances enhance the urban area into a smart city and to what extent the public is satisfied. A reliable Web portal is beneficial to ensure convenient access for users to analyze the results, allowing for seamless extensions.

5. 1. 2 Proposed Solution

We propose a Web portal called LSOMP to disseminate results of mining ordinances and tweets. We have designed a mining approach leveraging domain-specific knowledge bases (KBs) built for ordinances and smart city characteristics (SCCs) using publicly available sources, e.g. [1]. These, in addition to commonsense knowledge (CSK) from sources such as WordNet and WebChild [2] serve as guidelines to map ordinances and tweets, SCCs being the nexus.

Figure 5.1 portrays SCCs used in this work as widely accepted worldwide [1]. For example, if an ordinance includes “air pollution”, then CSK and domain KBs help to identify this ordinance as relevant to “Smart Environment” analogous to human commonsense reasoning. The same process applies to tweets. We claim there is a link between tweets and ordinances that share a similar SCC, and fine-grained analysis is conducted for direct mapping with minimal linkage complexity in our approach called TOLCS (Tweet Ordinance Linkage by Commonsense and Semantics) [3]. Sentiment analysis on tweets occurs by CSK-based polarity classification to gauge the effectiveness of ordinances via public opinion. A single ordinance or tweet can map to one or more SCC(s).

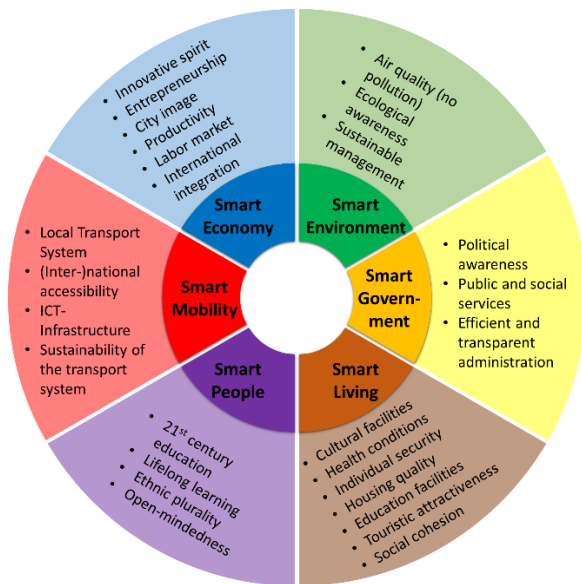


Figure 5.1 Widely accepted SCCs: Smart City Characteristics (adapted from [1])

LSOMP serves as our dissemination center with interactive graphics and QA, entailing NLP and CSK to fathom the text and proffer responses based on the mining results. *To the best of our knowledge, LSOMP is the first Web portal on ordinance-tweet mining.* Figure 5.2 portrays its system architecture.

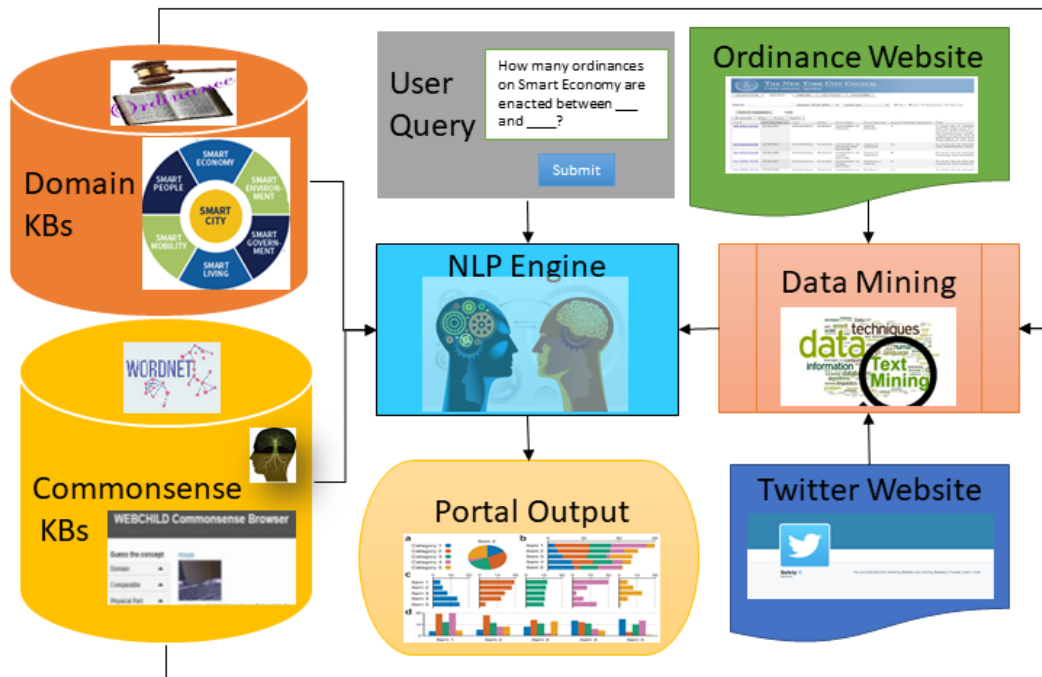


Figure 5.2 System architecture of the Web portal: LSOMP

5.1.3 Demo and Experiments

We conduct experiments [4], our source being ordinance data from the New York City council public website. We present herewith the analysis of ordinance sessions 2006- 2009 and 2010-2013. Ordinances and tweets are mapped with SCCs to obtain their mutual semantic relatedness. This serves as the basis for opinion miming of tweets pertaining to the ordinances via polarity classification of the tweets. Domain experts have evaluated the accuracy of our ordinance-tweet SCC mappings. They have confirmed that the accuracy is very good as tabulated here (see Table 5.1).

Table 5.1 Mapping Accuracy of Ordinances and Tweets

Item	Expert 1	Expert 2	Expert 3
Ordinance-SCC	92%	89%	92%
Tweet-SCC	82%	83%	80%

We have built a prototype of LSOMP, integrating its mapping functionality. This incorporates interactive graphics and Web QA. In Figure 5.3 we present a summary of the overall ordinance-SCC mapping results of two recent NYC ordinance sessions, i.e. 2006-2009 and 2010-2013, as disseminated in this portal based on our corresponding analysis [5]. We later aim to expand data coverage on ordinances and tweets.

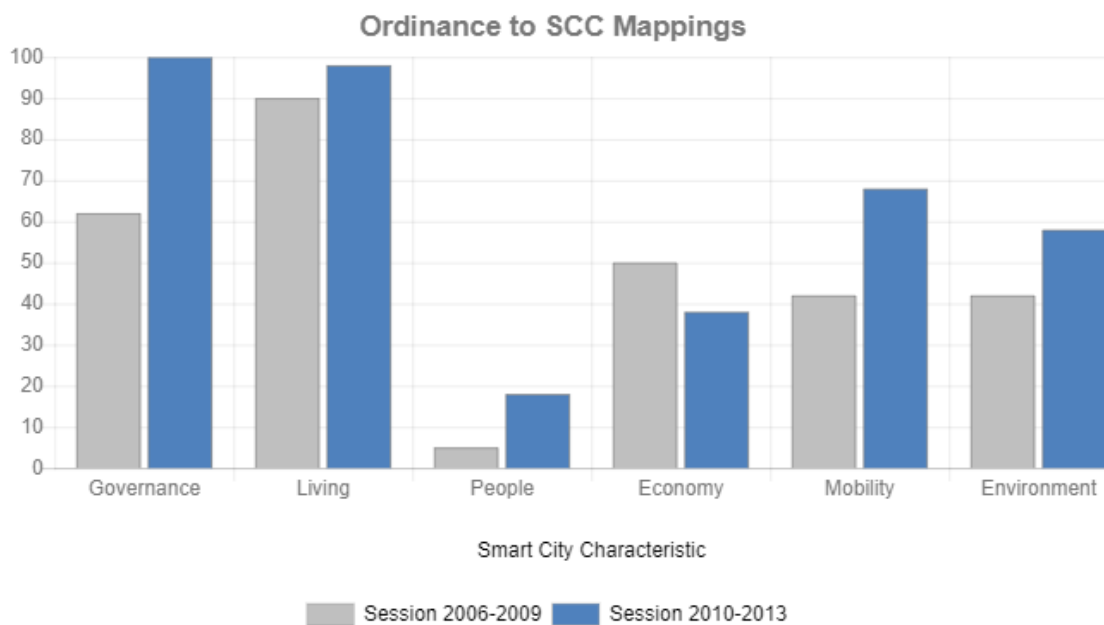


Figure 5.3 Portal depiction of mapping ordinances to SCCs

Mapping of tweets to SCCs is illustrated in this portal via real-time pie-chart generation.

Figure 5.4 portrays results of tweet-mapping with a single SCC while Figure 5.5 and Figure 5.6

depict mappings with multiple SCCs. The extent of the tweet's mapping with each SCC is illustrated within the chart (e.g. this is equal for all SCCs in Figure 5.5 while there is a greater mapping with one SCC compared to another in Figure 5.6). Such graphics are automatically constructed in real-time when a tweet is entered in the portal. These mappings are derived by incorporating knowledge from commonsense KBs and domain KBs as depicted in the system architecture (see Figure 5.2). An important functionality in LSOMP is Web QA as exemplified in Figure 5.7. Here, the user inputs a question and the portal finds the corresponding answer. Such answers are based on knowledge discovered via ordinance-tweet mining in our work [3], [4], [5], embedded in this portal for dissemination.

Tweet/Phrase to Map

This is awesome! Fair fares for subways make so much sense

Enter a Tweet/Phrase and then click your return ("Enter") key to see the results

Max Length: 280 characters

Possible Smart City Characteristics

- Smart Economy
- Smart Mobility
- Smart Environment
- Smart Governance
- Smart People
- Smart Living

Tweet to SCC Mapping Results

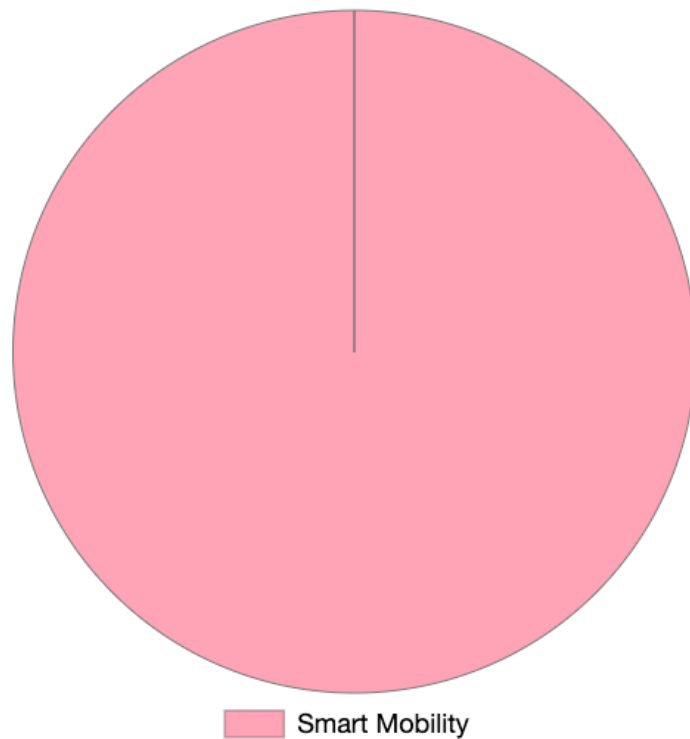


Figure 5.4 Real-time graphics: tweet to single SCC mapping

Tweet/Phrase to Map

Your government does not have my support to put land, water and our tourist economy at risk while trashing indigenous rights

Enter a Tweet/Phrase and then click your return ("Enter") key to see the results

Max Length: 280 characters

Possible Smart City Characteristics

- Smart Economy
- Smart Mobility
- Smart Environment
- Smart Governance
- Smart People
- Smart Living

Tweet to SCC Mapping Results

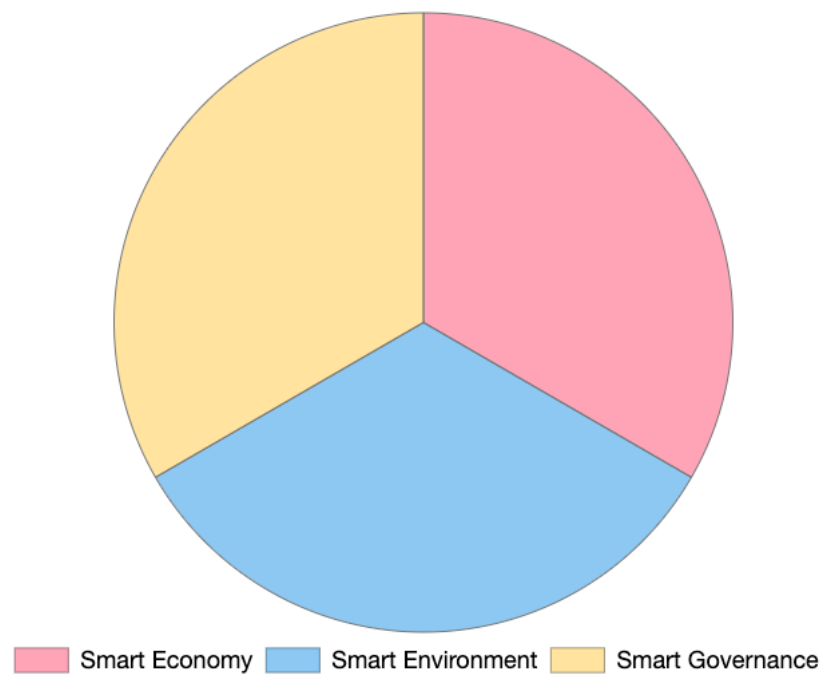


Figure 5.5 Real-time graphics: tweet to multiple SCCs with equal mapping

Tweet/Phrase to Map

The rent is too damn high, property taxes are too damn high, the subways are in a mess

Enter a Tweet/Phrase and then click your return ("Enter") key to see the results

Max Length: 280 characters

Possible Smart City Characteristics

- Smart Economy
- Smart Mobility
- Smart Environment
- Smart Governance
- Smart People
- Smart Living

Tweet to SCC Mapping Results

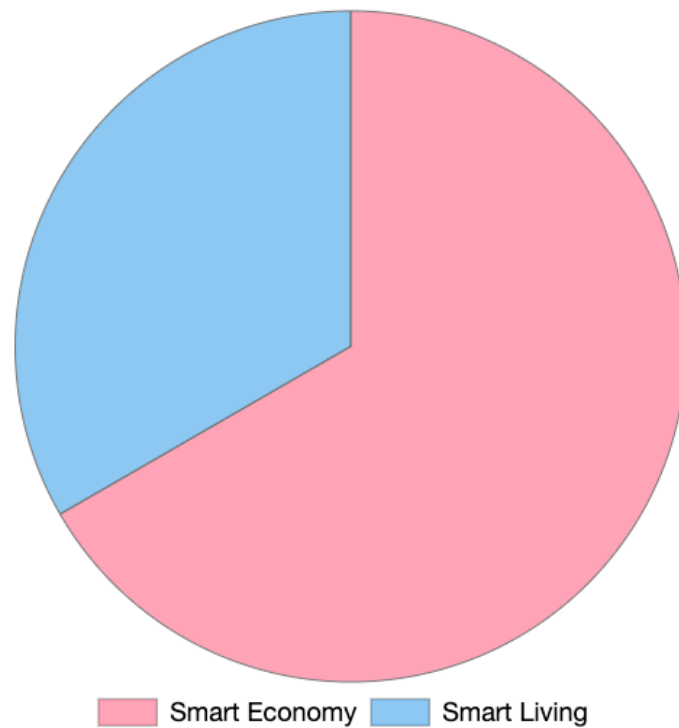
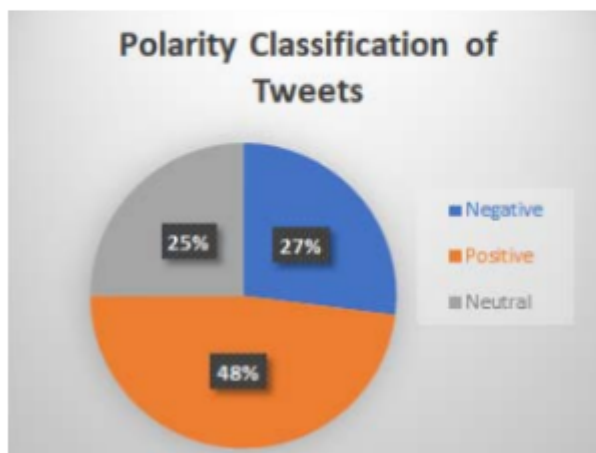


Figure 5.6 Real-time graphics: tweet to multiple SCCs with unequal mapping

From the Tweets taken and processed in the New York City region

what has the majority public reaction been?

Based on our results, the reaction breaks down the following:



Majority
is Positive

Figure 5.7 Example query in Web QA: pertaining to public satisfaction

5.1.4 Scalable Extension: Historical Analysis

Motivated by the success of our research in this area, we hereby propose a scalable extension to this work that entails the analysis of historical data on ordinances and tweets. This proposed extension is summarized in Figure 5.8. Users would enter a specific ordinance number or search by keywords. The matching ordinance would be displayed as a result of the ordinance search functionality. Users would be allowed to select the duration of tweet analysis before and after the given ordinance. Accordingly, the results of the polarity classification of the respective tweets that map to the given ordinance would be displayed to express the distribution based on

the sentiment analysis. Note that this diagram simply depicts the summary of the extension in one snapshot to conceptualize the process. In practice, there would be multiple screens for user interaction that would convey the ordinance search results, tweet polarity classification based on opinion mining etc. This scalable extension leverages approaches similar to those described in this work for ordinance-tweet mapping and tweet sentiment analysis guided by CSK and domain knowledge on ordinances and SCCs. It would also encompass NLP for interaction between the user and the system, as well as for deciphering the ordinances and tweets, given their complex legalese and informal tone respectively. The TOLCS approach [3] that provides tweet ordinance linkage by significantly reducing mapping complexity for the big data on ordinances and tweets would be useful here. We anticipate that this extension would be even more useful in ordinance-tweet mining, taking into account smart city perspectives. Accordingly, the LSOMP portal for dissemination would also be enhanced based on this extended analysis.

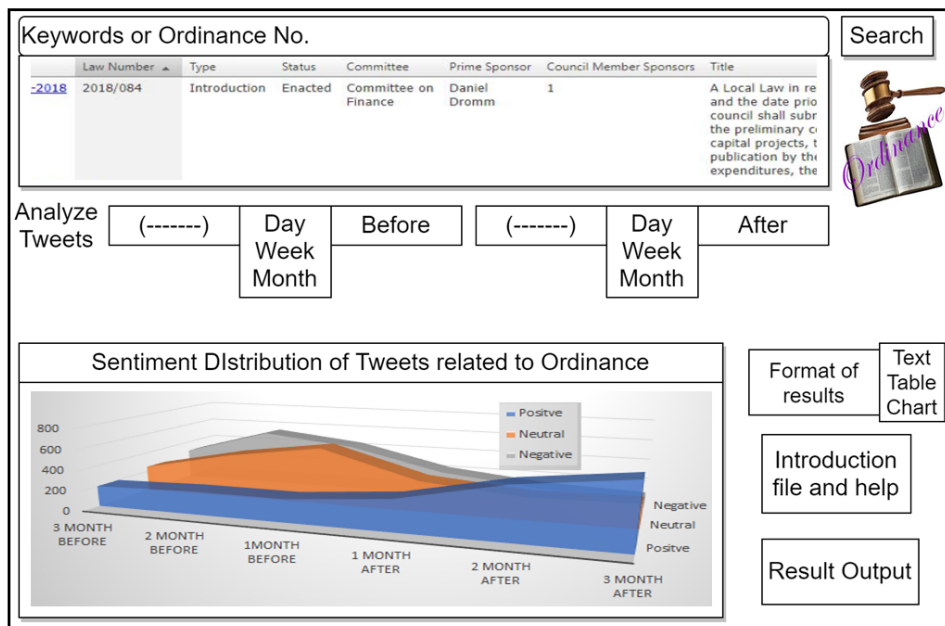


Figure 5.8 Extension: Historical data analysis of ordinances and tweets

5. 1. 5 Related Work

There is much work in sentiment analysis of social media data. Nuortimo [6] conducts opinion mining to gauge public acceptance of a system “Case Carbon Capture and Storage” for energy conservation. Huang et al. [7] synthesize social media posts, remote sensing data and Wikipedia with spatial data mining cum text mining for disaster analysis of past and future events. Gu et al. [8] propose a classifier-based method for tweet mining to get traffic incident data for highways and small roads. While these works address social media, as far as we know, ours is pioneering work on ordinance mining per se in conjunction with tweets, deploying CSK. It is in line with recently surveyed works on commonsense knowledge [9].

5. 1. 6 Discussion and Roadmap

The Web portal LSOMP is seamlessly extensible for expansion, e.g. KBs require updates for good mapping, and such functions are inbuilt. Historical analysis of ordinances and tweets, based on data mining guided by CSK and NLP incorporating smart city perspectives, is on our immediate roadmap. Further enhancement includes deep learning for better mapping accuracy and advanced NLP features for Web QA. Future work entails analysis of big data on COVID-related ordinances and tweets, e.g. post-COVID policy reshaping. This would incur substantial big data on ordinances and tweets where our approaches can be scaled with modifications.

5. 1. 7 References

© 2020 IEEE. Reprinted, with permission, from Xu Du, Matthew Kowalski, Aparna Varde and Boxiang Dong, LSOMP: Large Scale Ordinance Mining Portal, 2020 IEEE International Conference on Big Data, Dec 2020.

[1] TU-Wien, European smart cities, technical report, Tech. Rep., Vienna University of Technology, Austria, 2015.

[2] N. Tandon, G. de Melo, G. Weikum, Acquiring comparative commonsense Knowledge from the Web, AAI 2014, 166–172.

[3] M.Puri, A. Varde, B. Dong. Pragmatics and semantics to connect specific local laws with public reactions, IEEE Big Data 2018, 5433–5435.

- [4] M. Puri, A. Varde, X. Du, G. de Melo, Smart governance through opinion mining of public reactions on ordinances, *IEEE ICTAI 2018*, 838–845.
- [5] X. Du, D. Liporace, A. Varde, Urban legislation assessment by data analytics with smart city characteristics, *IEEE UEMCON 2017*, 20–25.
- [6] K. Nuortimo, Measuring public acceptance with opinion mining, *Journal of Intelligence Studies in Business*, 2018, 8(2);6–22.
- [7] Q. Huang, G. Cervone, G. Zhang, A cloud-enabled automatic disaster analysis system of multi-sourced data streams, *Computers, Environment & Urban Systems*, 2017, 66:23–37.
- [8] Y. Gu, Z. Qian, F. Chen, From Twitter to detector, *Transportation Research Part C: Emerging Technologies*, 2016, 67:321–342.
- [9] N. Tandon, A. Varde, G. de Melo, Commonsense knowledge in machine intelligence, *ACM SIGMOD Record 2017*, 46(4):49–52.

5.2 An Ordinance-Tweet Mining App to Disseminate Urban Policy

Knowledge for Smart Governance

Abstract: This paper focuses on how populations by the use of technology, more specifically an app, can comprehend the enactment of ordinances (local laws) in an urban area along with their public reactions expressed as tweets. Furthermore, they can understand how well the area is developing and enhancing as a Smart City. The main goal of this research is to develop an Ordinance-Tweet Mining App that disseminates the results of analyzing ordinances and tweets about them, especially related to Smart City Characteristics such as Smart Environment, Smart Mobility etc. This app would be beneficial to various users such as environmental scientists, policy makers, city committees as well as the common public in becoming more aware of legislative bodies, and possibly contributing in different aspects to make the urban area improve as a Smart City. This work fits the realm of Smart Governance due to transparency via public involvement.

Keywords: *App development, Human Computer Interaction (HCI), Legislative information, Opinion mining, Twitter data, Smart cities, Urban policy*

(Chapter 5.2 reused the previously published paper Varghese, C., Varde, A., & Du, X. (2020), An Ordinance-Tweet Mining App to Disseminate Urban Policy Knowledge for Smart Governance, *Lecture Notes in Computer Science, Conference on e-Business, e-Services and e-*

Society, I3E 2020, Vol. 12067, 389-401. https://doi.org/10.1007/978-3-030-45002-1_34).

5.2.1 Introduction

The paradigm of Smart Governance leverages transparency through public inclusion. This is gaining impetus with Smart Cities [1, 2]. It is thus important to disseminate knowledge on urban policy to the public. Ordinances (local laws) are often publicly available, e.g. [3], yet many residents view them as impervious due to complex legalese. Thus, ordinances need mining to discover knowledge that a wide spectrum of users can comprehend. A related aspect is public opinion on social media. It is important to gauge public reactions on issues related to ordinances by opinion mining. The results of mining ordinances and their public reactions need dissemination to be easily accessible and understandable. This is the focus of our overall research [4, 5]. We address ordinances from publicly available sites, and reactions to them expressed on Twitter. We conduct mining on these to discover knowledge on how well the ordinances enhance the given area into a Smart City, and to what extent the public is satisfied with the ordinances.

We conduct such ordinance and tweet mining, guided by commonsense knowledge (CSK) [6] to capture subtle human reasoning. We use CSK sources, along with the well-known word2vec [7] for ordinance-tweet mapping, and with sentiment polarity classification for mining. Based on this, we develop an Ordinance-Tweet Mining App, to disseminate the analysis. In this paper, we describe the app development using principles from HCI (Human Computer

Interaction), e.g. [8, 9]. We explain the mining of ordinances and tweets leading to useful legislative information disseminated by the app. This app provides QA (Question Answering) on interesting issues in urban policy. It is user-friendly, including interactive graphs and FAQs (Frequently Asked Questions) that facilitate comprehension by the public and experts. Targeted users, including the common public and domain experts in environmental management evaluate this app. In this paper, we focus on disseminating the results of mining NYC ordinances available on its public legislative council website [3]. We collect tweets from NYC through location-based data available on Twitter. We consider NYC since it is the financial capital of the USA, one of the major metropolitan cities in the world and a leading Smart City (see Figure 5.9) [3, 10]. As per world rankings, NYC is among the top 25 smart cities worldwide [11]. This is a good achievement. Yet there is scope for enhancement. We address this in the mining of ordinances and tweets, and in the corresponding app.

We present the development and experimentation of our Ordinance-Tweet Mining App herewith. To the best of our knowledge, ours is the first app disseminating the outcomes of ordinance-tweet mining, leveraging HCI. This app contributes to Smart Governance by making information on urban policy ubiquitous and comprehensible.



Figure 5.9: NYC (New York City) as a prominent smart city [left] and NYC council website [right]

5.2.2 Overview of Ordinance and Tweet Mining

In our earlier works, we propose methods for mining of ordinances from websites [4, 12] and furthermore, the mining of public opinions about them expressed on Twitter [5, 13, 14]. We use these methods for ordinance-tweet mining within NYC and conduct evaluation as included in our papers [4, 5, 12–14]. As stated in [5, pp. 1721–1722], “an important focus in our work is to determine to what extent such ordinances contribute to establishing the relevant urban region as a Smart City. Hence, we categorize ordinances based on their pertinent Smart City Characteristics (SCCs). We aim to connect ordinances to relevant tweets by drawing on their semantic relatedness. This is nontrivial, as ordinances and tweets both involve highly intricate and rather heterogeneous natural language, so simple keyword matching does not suffice. We propose a two-step approach for mapping that exploits the transitive nature of the connection between ordinances and tweets considering their relationship with SCCs. Specifically, the transitive

property we invoke is that: if the ordinance relates to a given SCC and any tweet relates to the same SCC, then the ordinance bears a connection to the tweet. This approach is proposed because classical sources of SCC data e.g., [1, 2] are finite and are restricted to a limited set of identifying features that can be relied upon for mapping. Thus, this transitive approach is more feasible than attempting to directly relate a seemingly infinite number of tweets to ordinances from various websites. As a first step, we discover connections between SCCs and ordinances using classical SCC sources [2] guided by commonsense knowledge (CSK) from web-based repositories [15, 16]. In the second step, we consider the mapping of tweets to SCCs, again drawing on such CSK. This approach then enables us to directly relate ordinances and the tweets to the pertinent aspects of Smart Cities and also sets the stage for sentiment polarity classification” [5, pp. 1721–1722]. Based on this, our mapping algorithm is Algorithm 1, as presented herewith [5].

ALGORITHM 1: ORDINANCE-TWEET-SCC MAPPING

1. **for each** SCC S_i **do**:
 2. **build** domain KB K_i
 3. $A \leftarrow \emptyset$
 4. **for each** ordinance O_i **do**:
 5. **for each** SCC S_j **do**:
 6. $L_{i,j} \leftarrow \sum_{x \in K_j} C(O_i, x)$
 7. $A \leftarrow A \cup \{(O_i, S_j) \mid j = \operatorname{argmax}_j L_{i,j}\}$
 8. **for each** tweet T_i **do**:
 9. **for each** SCC S_j **do**:
 10. $M_{i,j} \leftarrow \sum_{x \in K_j} C(T_i, x)$
 11. $A \leftarrow A \cup \{(T_i, S_j) \mid j = \operatorname{argmax}_j M_{i,j}\}$
 12. $\theta \leftarrow \{(O_i, T_k) \mid \exists S_j : (O_i, S_j) \in A \wedge (T_k, S_j) \in A\}$
 13. **return** θ
-

As stated in [13, pp. 841–842], “we conduct sentiment analysis to discover knowledge specifically with respect to opinion mining of tweets on ordinances. This is conducted after the mapping of ordinances to tweets. The primary database used for An Ordinance-Tweet Mining App to Disseminate Urban Policy Knowledge 391 Sentiment Analysis in this work is SentiWordNet [17]. This is an enhanced version of the CSK source WordNet [15]. It groups words into synonym sets (synsets) annotated by how positive the terms are. Accordingly, words are classified as positive, negative or neutral based on polarity of terms. In SentiWordNet, different meanings exhibited by the same word can have different sentiment scores. For example, the word estimable when relating to computation has a neutral score of 0.0, while the same word in the sense of deserving respect is assigned a positive score of 0.75. The process we deploy for sentiment analysis of tweets constitutes a semi-supervised learning method using SentiWordNet. Through this, subtle human judgment through commonsense in understanding emotions is embodied in the mining processes with specific reference to context” [13, pp. 841–842]. Accordingly, our algorithm for polarity classification of tweets is Algorithm 2 as presented next [13].

ALGORITHM 2: TWEET POLARITY CLASSIFICATION

1. **for each tweet t_i do:**
 2. **if not (t_i relevant according to SCC KB):**
 3. **continue** (with next tweet)
 4. map t_i to ordinances using Algorithm 1
 5. $W_i \leftarrow$ set of words in t_i
 6. **for each $w \in W_i$ do:**
 7. $s_w \leftarrow$ polarity score of w in SentiWordNet
 8. $s_i \leftarrow \sum_{w \in W_i} s_w$
 9. **return** final polarity scores s_i for relevant t_i
-

As further stated in [13, pp. 841–842], “based on this algorithm, we classify thousands of tweets that we obtain from Twitter. Note that the selection of relevant tweets and also the mapping of tweets to their respective ordinances is guided by CSK. We construct SCC-based Domain KBs (Domain-Specific Knowledge Bases) [5, 13, 14] derived from WebChild [16] and WordNet [15], to filter out unwanted tweets as a first step, followed by the mapping of tweets to relevant ordinances using SCCs as a next step” [13, pp. 841–842]. We start mapping groups of ordinances to tweets [5, 13], and then connect them with each other at finer levels of granularity [14] by mapping individual ordinances to tweets. This is via ordinance KBs in addition to SCC KBs, incorporating pragmatics and semantics through CSK and domain knowledge respectively [14]. The mining in our work is on ordinances and tweets from NYC, using a public council site [3] and Twitter location-based data. The results of the SCCbased ordinance analysis and tweet polarity classification comprise significant inputs for building the Ordinance-Tweet Mining App. We now describe its design process.

5. 2. 3 Approach for App Design

With the advancement in technology over the years, we are now benefiting from the “The Digital Age” which gave rise to the Internet and various mobile devices. This has digitized many institutions over the past few decades with governments processing applications via e-government websites that help people process their applications faster and avoid long wait periods. Retail stores today have e-commerce websites that help them reach a global audience. This progress has evolved the manner in which we do business on a higher scale. New technologies emerging in AI will help improve urban policies significantly [18], especially with respect to outreach initiatives. Users today often wish to have ubiquitous access to information. This leads to the development of apps. Accordingly, in the realm of e-government, an app for disseminating the results of ordinance-tweet mining is useful.

Based on this background and the overview of our ordinance and tweet mining research, we explain our proposed approach to design an Ordinance-Tweet Mining App. We comprehensively use principles from Human Computer Interaction (HCI) [8, 9, 19, 20] to create a user experience that would be all encompassing for various users. An important concept in HCI is Fitt’s Law [9, pp. 518–519], i.e., “ $T = k \log_2 (D/S + 1.0)$ where T = time to move the pointer to a target, D = distance between the pointer and the target, S = size of the target, k is a constant of approximately 200 ms/bit”. We incorporate Fitt’s Law in our app design. Thus, we design items

in the app such that they are big enough to enable users to spot them fast, especially as navigation proceeds further from the opening screen. Yet, these items are small enough to fit on the required screens, and hence users do not need to spend much time while searching.

For the layout of the app, we use another HCI concept, i.e. “mockup designs”. Mockups are multiple designs created by interviews with “stakeholders” that give an overview of the blueprint of the app [8]. Stakeholders in HCI terminology are various influential groups such as domain experts who have a thorough knowledge of the field; students/working professionals involved in providing data; and end-users ranging from novice to expert, and casual to frequent [8]. Among the mockups, we select the best designs for the app layout and screens. Other HCI aspects we incorporate are the simplicity and efficiency of the interface [20]. These include navigation, the time spent to find an answer to a question asked by the user, and the analytical graphs displayed. Considering HCI, some principles used in the app adhere to the guidelines of Google’s material design [19] to ensure that the app is up to date with the current and latest software, and to warrant that the app runs smoothly without errors or bugs. Another HCI aspect is a “metaphor” [9]. The term metaphor refers to conceptualization of actions, typically for the interface. For example, a shopping cart is an interface metaphor used for checkout in online shopping. We incorporate metaphors in our app for various purposes, e.g. to depict different ordinance departments and smart city characteristics.

Good design of this app entails navigating pages without spending too much time on

locating the options that the users need [20]. Based on HCI principles, Figure 5.10 illustrates the navigation of the app in an efficient manner to access the given information. For details, please see [21]. This navigation keeps the users engaged and alert. Another important factor is the specific usage to assess the emotions of users while navigating the app and seeing the results. In order to incorporate this, we make sure that the intermediate pages and end-results focus on inclusion and interactive design elements [19].

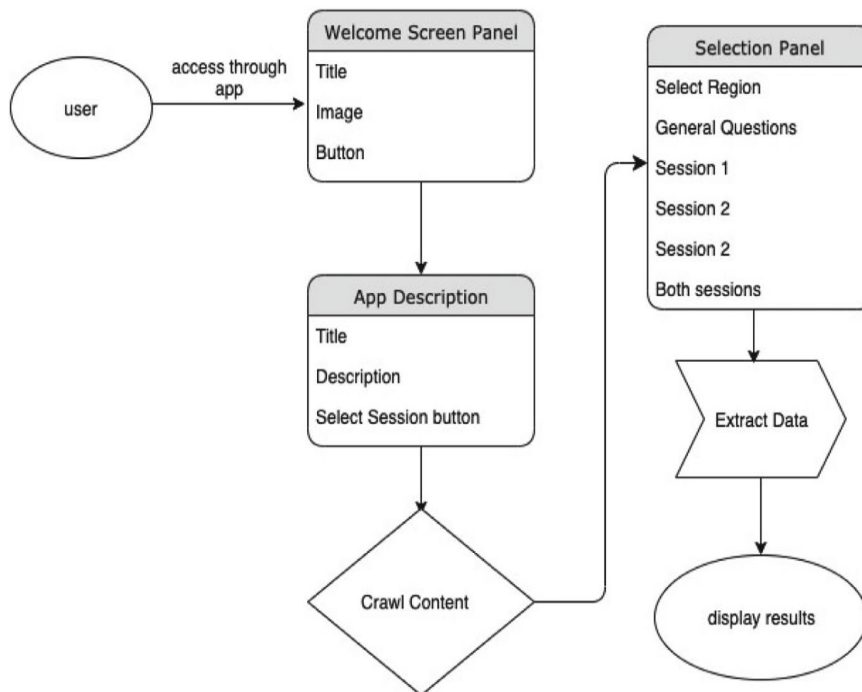


Figure 5.10: App navigation flowchart

5. 2. 4 Implementation of the App

For the implementation of the Ordinance-Tweet Mining App, we use Android Studio [22,

23] as the Integrated Development Environment (IDE). Android Studio provides useful features that make the actual programming of app very convenient and efficient. The IDE facilitates development and use for anyone ranging from a beginner to a fullfledged software engineer [24].

Some features of Android Studio are:

- Code Completion
- Creation of Templates
- Instant App Run
- Fast Emulator
- Smart Code Editor
- Kotlin Programming Language Support

Based on these features, we implement the Ordinance-Tweet Mining App deploying Android Studio. Figure 5.11 illustrates the implementation process using UML for Android Development [23] with a self-explanatory diagram. For a more detailed explanation, please see [21]. Given this implementation, Figure 5.12 shows a snapshot of the app layout. The left screen in the figure serves as a landing and welcome page with a call to action button labeled “Get Started”. This lets the users know that they can access the app and navigate to the desired location. The center screen gives a brief description of the app. Once the users are ready to select an option, they can click on “Select Ordinance Session”. The right screen serves as an action sheet. Action sheets are helpful whenever there are multiple actions [20]. They work well on this

screen, since it shows the different options for the region and the sessions that the users can select.

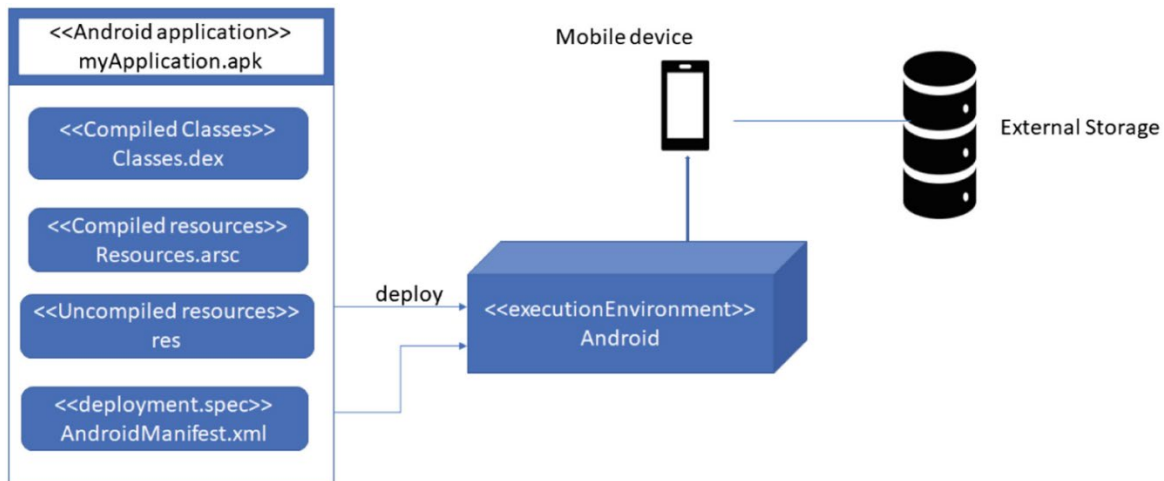


Figure 5.11: Implementation process of the app using the Android platform

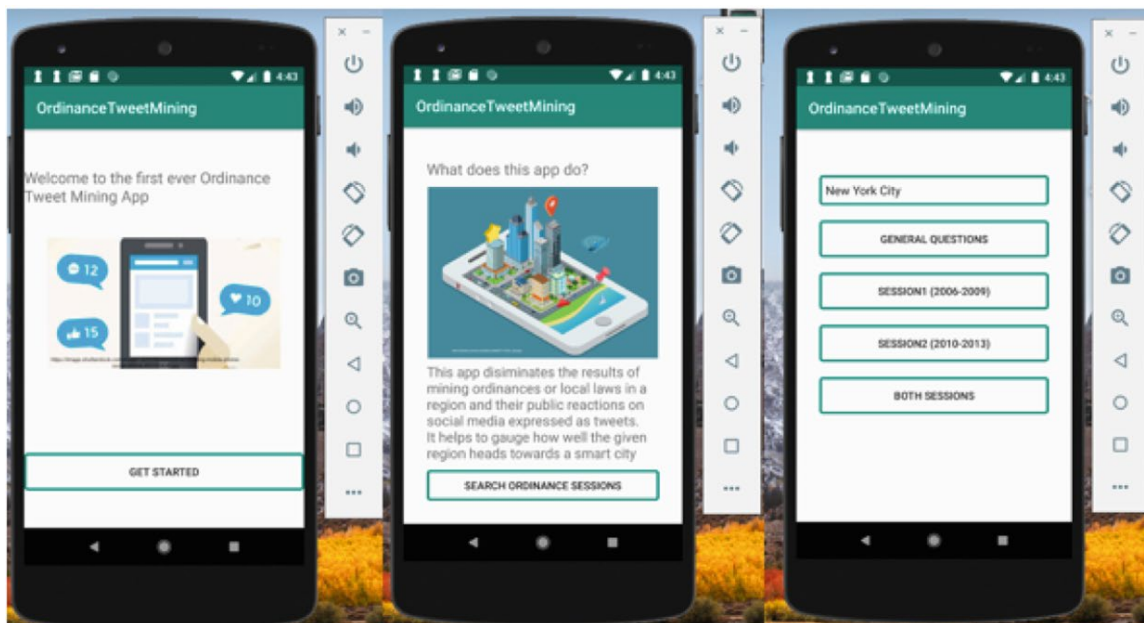


Figure 5.12: Layout screens of the Ordinance-Tweet Mining App

Figure 5.13 depicts the screens with FAQs and corresponding results of ordinance mining

for the NYC ordinance passing sessions considered [3], i.e. Session 1 (2006– 2009) and Session 2 (2010–2103). The leftmost screen includes general questions that users may have for example, “What is a smart city?” “What specific characteristics does a smart city have?” and so on. The other screens depict the outputs from the mining of the specific sessions and the combined results from mining both the sessions.

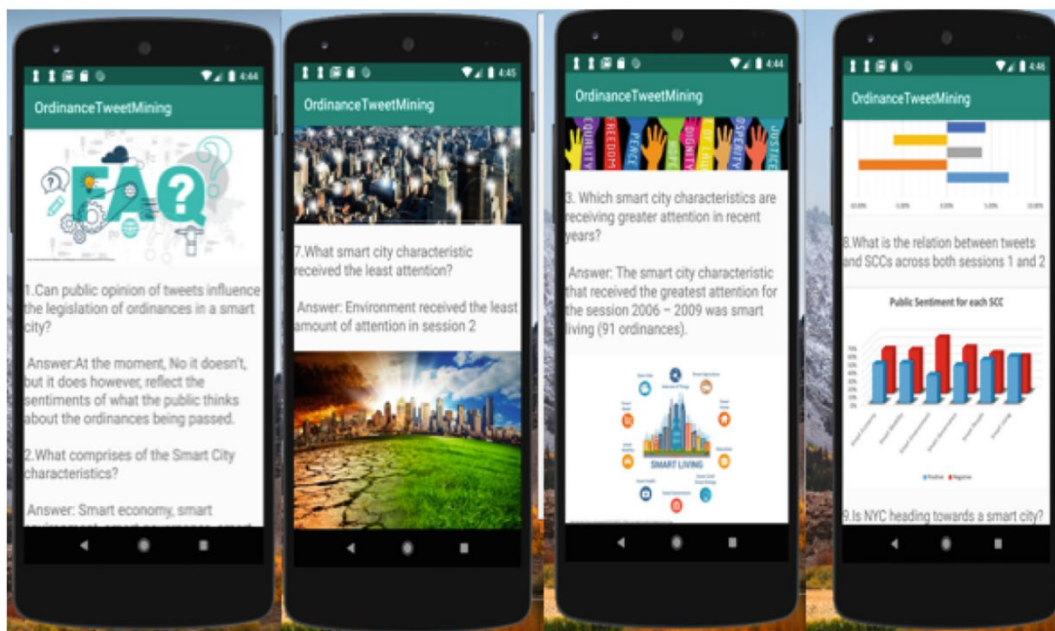


Figure 5.13: FAQs for various selection categories in the app

5. 2. 5 Experiments and Discussion

In order to evaluate the app, we conduct user surveys [21]. We create survey questions using a Likert Scale format [9]. We summarize the results in Figure 5.14, Figure 5.15 and Figure 5.16.

These encompass the feedback of 34 participants with an assortment of computer and

environmental scientists, students, lawyers, policymakers and researchers. The main questions are as follows with responses on a scale of 1–5 [1: Strongly Agree... 5: Strongly Disagree]

- Q1: Do you find this app, quick and easy to use?
- Q2: Does this work increase public awareness of urban policy?
- Q3: Do you feel NYC is getting better as a smart city?

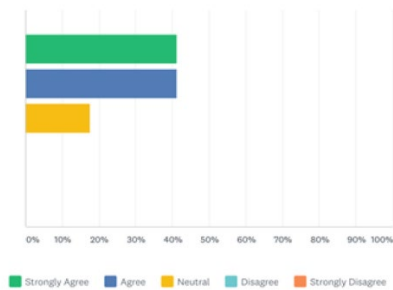


Figure 5.14: Responses to Q1: “Do you find the app quick and easy to use?”

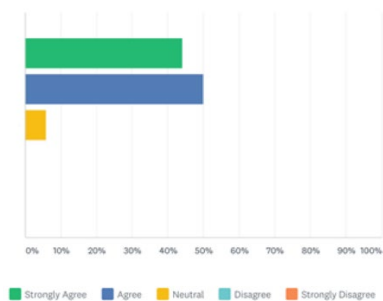


Figure 5.15: Responses to Q2: “Does this work increase public awareness of urban policy?”

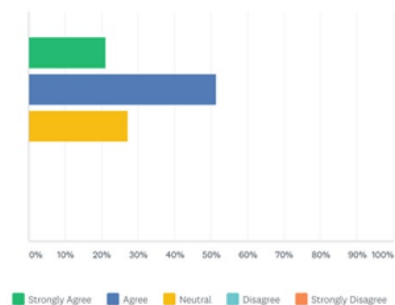


Figure 5.16: Responses to Q3: “Do you feel NYC is getting better as a smart city?”

As seen in these figures, the user survey results indicate favorable responses towards this app. After collecting the data from the surveys, it is evident that the app is feasible for use and makes users more aware of urban policy. Many users feel that NYC is getting better as a smart city while some are neutral. In addition to the Likert scale evaluation, some comments included by the users in the survey are as follows.

- “This can be useful in courses on urban policy”
- “This app looks great – simple and informative is what New Yorkers need. I would have liked to have access to links to the researcher’s published work, if available or maybe lists of smart city ordinances as a resource.”
- “Very good usage of graphics, makes data easier to understand”
- “Great work you are doing. Well-done”
- “I am interested in understanding the technology behind this product”

Based on these survey results and the general feedback received from the users we can infer that the work on extracting ordinance data and conducting sentiment analysis through mining of

tweets have proven useful by reaching a larger audience and by disseminating information that may not have been readily available to the public earlier. Please note that our research on ordinance-tweet mining as well as the development of this app both constitute pioneering work in the area, to the best of our knowledge. Hence, we do not conduct comparative studies for the mining or the app development.

All this work is in line with the recent concept of greater awareness and more transparency in governance. Some local celebrities have played their part by influencing laws that improve society. Tim Tebow, a quarterback for NY Jets inspired South Carolina legislators to pass the Equal Access to Interscholastic Activities act in May 2012. Also known as the “Tim Tebow Law” [25], this allows homeschooled children to participate in public school extracurricular activities. Candy Lightner founded the “Mothers against Drunk Driving (MADD)” organization to ensure sterner laws against drunk driving [26]. Due to the media attention this organization received with Candy’s story, it raised more awareness of driving responsibly. Likewise, our Ordinance-Tweet Mining App, the first of its kind, is likely to play an important role in making the public more aware of legislative policies and take action needed for enhancement. This concept fits the realm of Smart Governance [1, 2], a characteristic of Smart Cities that leverages greater transparency in governing processes via more public involvement.

5. 2. 6 Related Work

Social media users generate massive amounts of information in their daily life. Scientists consider that as valuable data for various studies on urban policy, traffic, energy conservation, climate change, disaster management etc. Social media text mining is therefore a powerful tool to extract useful knowledge.

Gu et al. [27] propose a method to use tweets for gathering road incident information. They build an API (Application Programming Interface) to compare incidents recorded on social media and in traditional databases. This not only validates existing incident information but also finds new incidents, thus supplementing the data in databases. Gandhe et al. [28] conduct sentiment analysis of data from Twitter on political scenarios, urban events etc. As stated in [28, p. 57], the “proposed approach entails a hybrid learning method for classification of tweets based on a Bayesian probabilistic method for sentence level models given partially labeled training data”. The advantage is that the approach is semi-supervised, and works even with partly labeled data. Nuortimo [29] studies social media data from multiple platforms to understand public reactions for a system called Case Carbon Capture and Storage, to control carbon dioxide emissions. Results show that the overall reactions are positive. This study indicates that social media mining could be a great tool to measure public awareness and acceptance for topics related to energy and climate. Huang et al. [30] assess disaster analysis of historical and future

events. They gather social media, remote sensing and Wikipedia data, performing spatial analysis and social media mining. Their results show that social media mining enhances disaster analysis and provides real-time tracking.

All these works demonstrate the potential of social media mining. Accordingly, if an app disseminates knowledge from such mining for public outreach, it would have broader impacts on sustainability and Smart Cities [1, 2, 11], especially with reference to Smart Governance [2]. It would foster building other related apps. The sharing of data via such apps would benefit pertinent research. Various useful mobile apps for Androids exist in the literature, as described in recent work [31]. Our design and development of the Ordinance-Tweet Mining App contributes to this overall realm.

Various aspects of AI can make an impact on Smart Cities as surveyed in the literature [2, 6, 18, 32–34]. AI can help record car activity, foot traffic, types of shoppers that go to different retailers, their preferences, availability of parking spots etc. These are minor details yet they make big impacts to create efficient solutions [18]. AI can contribute to autonomous and semi-autonomous driving through incorporation of CSK-based techniques for enhanced decision-making [6, 32]. AI can play a role in augmenting object detection for Smart Mobility in Smart Cities with neural models, deep learning, CSK and adversarial datasets [6, 33, 35]. Another major aspect is AI in lighting. NYC has bright lights, which imply significant energy consumption. Thus, if lamppost design occurs with sensors, these can adjust their brightness

depending on the amount of traffic within the area. In some cities such as Amsterdam, canal lights dim and brighten based on pedestrian usage [2]. Likewise, AI can contribute to several aspects of Smart Cities [34]. Our work in this paper is a step in this direction, using HCI-based app design and disseminating the results of mining ordinances along with their public reactions. Hence, this paper makes an impact on Smart Governance in Smart Cities.

5. 2. 7 Conclusions

This paper addresses the development of an Ordinance-Tweet Mining App that disseminates knowledge discovered by mining ordinances in a given region and tweets about them, especially relevant to Smart Cities. Through this app, users from various backgrounds can obtain quick and easy access to legislative information. Via the app the public can make better decisions and contributions, e.g. by understanding policies and public reactions, they can participate city council committees, or support their region through financial means and community outreach. In addition, this app can provide decision support to lawmakers by providing ubiquitous information, and can enhance the scope of study for researchers through future issues emerging from the work here. To the best of our knowledge, ours is the first ever Ordinance-Tweet Mining App. While this app focuses on NYC in particular, it can foster the development of similar or related apps for other cities, by reuse of the approaches and data with modification.

Future work includes embedding intricate NLP (Natural Language Processing) along with

semantics and pragmatics in order to facilitate direct QA (Question Answering) beyond static FAQs (Frequently Asked Questions) and keywords. This would encompass advances in the field of CSK (Commonsense Knowledge) to fathom the QA text and give enhanced responses in the app based on the mining results. This An Ordinance-Tweet Mining App to Disseminate Urban Policy Knowledge 399 constitutes part of our ongoing research. In general, our work in this paper makes a broader impact on Smart Governance.

5. 2. 8 References

1. IEEE Smart Cities: What Makes a City Smart (2020). <https://smartcities.ieee.org/>
2. TU Wien: European Smart Cities, Technical Report, Vienna University of Technology, Austria, August 2015
3. The New York City Council: Legislative Research Center (2018).
<http://legistar.council.nyc.gov/>
4. Du, X., Liporace, D., Varde, A.: Urban legislation assessment by data analytics with smart city characteristics. In: IEEE UEMCON, pp. 20–25, October 2017
5. Puri, M., Du, X., Varde, A., de Melo, G.: Mapping ordinances and tweets using smart city characteristics to aid opinion mining. In: WWW Companion, pp. 1721–1728, April 2018
6. Tandon, N., Varde, A., de Melo, G.: Commonsense knowledge in machine intelligence. ACM SIGMOD Rec. 46(4), 49–52 (2017)

-
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS, pp. 3111–3119, December 2013
 8. Harper, R., Rodden, T., Rogers, Y., Sellen, A.: Being Human: Human-Computer Interaction in the Year 2020. Microsoft Research (2008)
 9. Rogers, Y., Sharp, H., Preece, J.: Interaction Design: Beyond Human-Computer Interaction, 4th edn. Wiley, Hoboken (2015). ISBN 978-1-119-02075-2
 10. NYC Smart City: Eventa. <https://www.eventa.us/events/new-york-ny/nyc-smart-city-trek>
 11. The EasyPark Group: 2017 Smart Cities Index (2017).
<https://easyparkgroup.com/smartcities-index>
 12. Du, X., Varde, A., Taylor, R.: Mining ordinance data from the web for smart city development. In: DMIN, pp. 84–90. CRSEA Press, July 2017. ISBN 1-60132-453-7
 13. Puri, M., Varde, A., Du, X., de Melo, G.: Smart governance through opinion mining of public reactions on ordinances. In: IEEE ICTAI, pp. 838–845, November 2018
 14. Puri, M., Varde, A., Dong, B.: Pragmatics and semantics to connect specific local laws with public reactions. In: IEEE Big Data, pp. 5433–5435, December 2019
 15. Miller, G.: WordNet: a lexical database for English. *Commun. ACM* 38(1), 39–41 (1995)
 16. Tandon, N., de Melo, G., Suchanek, F., Weikum, G.: WebChild: harvesting and

organizing commonsense knowledge from the web. In: ACM WSDM, pp. 523–532, February 2014

17. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: LREC, May 2010

18. Dirican, C.: The impacts of robotics and artificial intelligence on business and economics. *Soc. Behav. Sci.* 195, 564–573 (2015)

19. Soni, S.: Google Material Design’s Impact on Mobile App Design, February 2019. [https:// appinventiv.com/blog/mobile-app-designers-guide-on-material-design/](https://appinventiv.com/blog/mobile-app-designers-guide-on-material-design/)

20. So, Y.: Designing for Mobile Apps: Overall Principles, Common Patterns, and Interface Guidelines, May 2012. <https://medium.com/blueprint-by-intuit/native-mobile-app-designoverall-principles-and-common-patterns-26edee8ced10>

21. Varghese, C., Varde, A.: Disseminating results of mining ordinances and their tweets by Android App development. Technical report, Montclair State University, December 2019

22. Android Developers: The Android Studio. <https://developer.android.com/studio>

23. Android Application: UML Deployment. <https://www.uml-diagrams.org/android-applications-on-uml-deployment-diagram-example.html>

24. ADMEC Institute: Top 10 Features of Android, July 2018. <https://www.admecindia.co.in/blog/top-10-features-android-studio-developers-not-miss>

25. The Week: 7 Celebrities Who Inspired New Laws. Business Insider, September 2012.

<https://www.businessinsider.com/7-laws-named-for-celebrities-2012-9#tim-tebow-2>

26. O’Connell, C.: 15 Ordinary People Who Changed History. Reader’s Digest:

<https://www.rd.com/true-stories/inspiring/inspiring-stories-9-ordinary-people-who-changed-history/>

27. Gu, Y., Qian, Z., Chen, F.: From Twitter to detector: real-time traffic incident detection using social media data. *Transp. Res. Part C: Emerg. Technol.* 67, 321–342 (2016)

28. Gandhe, K., Varde, A., Du, X.: Sentiment analysis of Twitter data with hybrid learning for recommender applications. In: *IEEE UEMCON*, pp. 57–63, October 2018

29. Nuortimo, K.: Measuring public acceptance with opinion mining: the case of the energy industry with long-term coal R&D investment projects. *J. Intell. Stud. Bus.* 8(2), 6–22 (2018)

30. Huang, Q., Cervone, G., Zhang, G.: A cloud-enabled automatic disaster analysis system of multi-sourced data streams: an example synthesizing social media, remote sensing and Wikipedia data. *Comput. Environ. Urban Syst.* 66, 23–37 (2017)

31. Basavaraju, P., Varde, A.: Supervised learning techniques in mobile device apps for Androids. *ACM SIGKDD Explor.* 18(2), 18–29 (2016)

32. Persaud, P., Varde, A., Robila, S.: Enhancing autonomous vehicles with commonsense: smart mobility in smart cities. In: *IEEE ICTAI*, pp. 1008–1012, November 2017

33. Pandey, A., Puri, M., Varde, A.: Object detection with neural models, deep learning and common sense to aid smart mobility. In: *IEEE ICTAI*, pp. 859–863, November 2018

34. Packt Publishing: Artificial Intelligence for Smart Cities. Becoming Human: AI Magazine. <https://becominghuman.ai/artificial-intelligence-for-smart-cities-64e6774808f8>

35. Garg, A., Tandon, N., Varde, A.: I am guessing you can't recognize this: generating adversarial images for object detection using spatial commonsense. In: AAA Conference, February 2020

5.3 Sentiment Analysis of Twitter Data with Hybrid Learning for Recommender Applications

Abstract: This paper proposes a sentiment analysis approach to extract sentiments of tweets based on their polarity and subjectivity, classify them and visualize results graphically. This helps to understand opinions of existing users that can be helpful in future recommendations. Our proposed approach entails a hybrid learning method for classification of tweets based on a Bayesian probabilistic method for sentence level models given partially labeled training data. For implementation, we use AWS to extract data from Twitter, store extracted data in MySQL databases and code Python scripts in order to implement the analyzer. The graphical models are viewed using IPython Notebook. The results of this work would be helpful in providing recommendations to users for product reviews, political campaigns, stock predictions, urban policy decisions etc. The novelty of this research lies mainly in the hybrid learning method for sentiment analysis. We present our approach along with its implementation, evaluation and applications.

Keywords: *Data Analytics; Hybrid Learning; Recommenders; Opinion Mining; Social Media; Twitter; Urban Policy*

(Chapter 5.3 reused the previously published paper Gandhe, K., Varde, A., & Du, X. (2018), Sentiment Analysis of Twitter Data with Hybrid Learning for Recommender Applications,

In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 57-63), New York City, NY, USA, <https://doi.org/10.1109/UEMCON.2018.8796661>).

5.3.1 Introduction

Sentiment Analysis refers to the automated study and investigation of evaluative text and tracking of predictive judgments therein. [1]. Due to the advent of social media, people enter their feeds on various events, products, current affairs and so on. These feeds are the actual opinions of influential people who feel free to express their views on social networking sites such as Twitter, Facebook etc. If we analyze these feeds, we are often likely to get truer, clearer opinions of people than from a guided survey [2]. Results of such analysis can be useful in various areas as follows.

Product Reviews: “What do people think about a particular product?” We can find reviews of any product/event/movie etc. from sentiment analysis. These results can be useful to the buyer as well as the seller. The buyer can pick up the best product as per their requirement. The seller can keep a closer watch on the product reviews, which can be used to improve the quality of the product.

Political Elections: “What are the public sentiments about the candidate/campaign?” The mood of the public is especially important during political campaigns. The results in this case

can be extremely useful for candidates to design or alter their campaigns. Candidates can get a clear picture of the issues that really concern people. Also, the results can be used for prediction of the winner.

Search Engine Optimization (SEO): “What are people talking about as trending news?”

Creating SEO friendly content is the key for any Website to get a high rank. From sentiment analysis, we can figure out the hot topics that interest people and that should be displayed on search engine home pages as headlines.

Stock Market: “What stocks should we invest in to maximize our gains? Market sentiment is the attitude of investors. Nowadays, investors are known to measure market sentiment through the use of news analytics, which include sentiment analysis on textual stories about companies and sectors. Thus, sentiment analysis can be used to find the market sentiment and hence predict the price development in stock markets.

Urban Policy: “What is the reaction of the city residents to various policies implemented by their legislators? These policies could pertain to the urban legislations are passed with respect to issues such as managing the environment, making information accessible to the public, imparting education, making healthcare more affordable and providing a good quality of life to residents on the whole. They could be related to general legislation on the whole, or to specific action pertaining to certain significant events. The opinions expressed by the public on Twitter can offer insights into their extent of satisfaction with urban policy matters. The analysis of these public

sentiments can be useful for providing recommendations to urban agencies for decision making.

Given this background, we focus on our goals. We consider Twitter to get social media feeds.

Tweets constitute microblogging with maximum character limits. Thus, we find that tweets being compact are very good for efficient sentiment analysis. Our goals are thus to:

- Analyze tweets and discover useful knowledge for various areas such as product reviews
- Provide inputs that could be potentially useful for recommender applications in these areas

5.3.2 Related Work

A broad overview of the existing work in sentiment analysis is presented in [3]. The authors describe existing techniques and approaches for an opinion-oriented information retrieval. In [4], the authors use Weblogs as datasets for sentiment analysis and use emoticons assigned to blog posts as indicators of users' moods. The authors apply SVM (support vector machines) and CRF (conditional random field) learners to classify sentiments at the sentence level and then investigate several strategies to determine the sentiment.

The work in [5] uses emoticons to form a training set for sentiment classification. They collect texts containing emoticons from Usenet newsgroups. Datasets are divided into "positive" (texts with happy emoticons) and "negative" (texts with sad or angry emoticons) classes.

Emoticon-trained classifiers, SVM and Naive Bayes, are able to obtain up to 70% accuracy on the test set.

The authors [6] use Twitter to collect training data and perform a sentiment search. They construct corpora by using emoticons to obtain “positive” and “negative” samples, and then use various classifiers. The best result is obtained by the Naive Bayes classifier. The authors are able to obtain up to 81% accuracy on their test set. In the work [7], the authors use the Twitter corpus to predict political elections in Germany. Their results show that Twitter is indeed used extensively for political deliberation.

In the research by [8], the authors augment accuracy of sentiment analysis by properly identifying semantic relationships between sentiment expressions and subjects. They are able to achieve precision of 75%-95% depending on inputs. The dataset they use consists of about a half million Web pages and a quarter million news articles. They do not use Twitter datasets.

In [9], the focus is on Twitter for the task of sentiment analysis. They use a method for the automatic collection of a corpus that can be applied to train a sentiment classifier. They use TreeTagger for POS-tagging and observe the difference in distributions among positive, negative and neutral sets. The classifier they use is based on the multinomial Naive Bayes that uses N-gram and POS (part-of-speech) tags as features.

Our work is somewhat orthogonal to the existing literature. We consider sentiment analysis on tweets in particular. The novelty of our research is that we propose a hybrid learning approach

that works even when pre-labeled training data is not fully available. The goals are specifically to mine the opinions of users with the intention of providing good recommenders for applications.

5.3.3 Sentiment Analysis Models and Methods

We first describe the models and methods that exist in the literature on sentiment analysis in order to propose our specific approach in this paper.

5.3.3.1 Document Level Model

In this model, whole document is classified as positive or negative. Documents can be opinionated. A document may be a review on a Website such as Trip-advisor or another source such as a company whitepaper. Consider the example shown in Figure 5.17. In this example, the user has posted review of a hotel. This review has multiple sentences. It includes some positives (“Beautiful hotel”, “great location”) and some negatives (“sad service”, “just painful”). The overall opinion of the review is calculated in document level classification. Due to multiple types of sentiments, classifying a document can be challenging task.



Figure 5.17: Example of document level model

5.3.3.2 Sentence Level Model

This model classifies a single sentence as positive, negative or neutral. Since a single sentence is generally likely to have only one sentiment (positive or negative), it is easier to classify than document. Also, it is more accurate than document level classification. An example of this model appears in Figure 5.18.



Figure 5.18: Example of sentence level model

In sentence level classification, two sub-tasks are performed:

1. *Subjectivity classification*: Determine whether the sentence is a subjective or an objective

sentence,

2. *Sentence-level sentiment classification*: If the sentence is subjective, determine whether it expresses a positive or negative opinion.

For instance, in Figure 5.18 the tweet is highly subjective and has a positive sentiment.

5.3.3.3 Supervised Learning Method for Analysis

In sentiment analysis, supervised learning involves techniques of learning from human-annotated labeled training data (as in supervised learning elsewhere). The training datasets have examples with a *text and label* pair. Consider the example shown in Figure 5.19.

```
"after years with that carrier's expensive plans and horrible customer service ,
portability seemed heaven-sent","pos"

"here's the brief synopsis : the phone is tiny , cute , feels kind of plastic-like ( as if
it might break ) , but seems pretty sturdy", "pos"

"it has lots of little cute features , my favorite being the games and the pim ( personal
information manager -- i.e. organizer ) , and the radio !", "pos"

"i spent hours setting up the stations ( accepts about 13-14 , i believe ) , though the
reception is unpredictable", "neg"

"the headset that comes with the phone has good sound volume but it hurts the ears like
you cannot imagine!", "neg"
```

Figure 5.19: Example of training set in supervised learning for sentiment analysis

In this example, the reviews and the polarity of reviews are used as training data. These training data reviews are classified manually. When these datasets are used for training, the learning technique uses the concerned functions to map these to new unseen examples (input). The respective algorithm would then classify the unseen input correctly. We can collect the

correct datasets, determine the input features, select the algorithm to be used and run the algorithm on training data. Once this is done, the accuracy of the algorithm can be evaluated using test data.

5. 3. 3. 4 Unsupervised Learning Method for Analysis

This type of learning for sentiment analysis does not need human-annotated data. It uses lexical methods to classify the unlabeled data. Opinion words and phrases are the dominating indicators for sentiment classification. Unsupervised learning in sentiment analysis is generally based on such opinion words and phrases. It performs classification based on some fixed syntactic phrases likely to be used to express opinions. (e.g., noun phrases) Unsupervised learning overcomes the limitation of supervised methods (where pre-labeled training datasets are essential).

5. 3. 4 Proposed Approach: Hybrid Learning

We propose a hybrid approach combining supervised and unsupervised learning in sentiment analysis. This is because we intend to take advantage of labeled training data whenever available but also need to classify tweets that lack specific labels. The approach is described next.

5.3.4.1 Overview of Approach

We strongly prefer sentence level models in this proposed approach of hybrid learning for sentiment analysis. This is because are useful for microblogging sites such as Twitter, since the maximum limit of a tweet is 280 characters. Thus, a sentence level model would fit better as a sentence would typically have less than 140 characters (very few documents are that small). Sentence level models would also be more precise for sentiment classification due to likely having only one sentiment per tweet.

We propose to build a classifier for sentiment analysis deploying the classical Naive Bayes concept [10]. The Naive Bayes algorithm uses some probability theory aspects explained as follows.

$$P(C_j|D) = P(D|C_j) P(C_j) / P(D)$$

where, $P(C_j|D)$ = probability of instance D being in class C_j

$P(D|C_j)$ = probability of generating instance d given class C_j

$P(C_j)$ = probability of occurrence of class C_j

$P(D)$ = probability of instance D occurring

Thus, in the case of unlabeled samples in the training data, Naive Bayes can find the probability of them being either positive or negative based on similar pre-classified data. In many practical applications, parameter estimation for Naive Bayes models uses the method of

maximum likelihood. Thus, it calculates all the probabilities of a feature being positive or negative using the training dataset. The probability of a sentence in test data to be positive or negative is calculated based on the formula herewith. For multiple feature data sets, Naive Bayes assumes that each feature is independent of other features in the dataset.

Thus, in our context, Naive Bayes would be interpreted with an example as follows. *Given that a person expresses an opinion in a university Website tweet, we need to know whether the person is male or female, furthermore whether he/she is a professor or a student.* This classification is performed based on learning from pre-classified datasets of tweets with the gender and occupation included, by applying the probability concepts herewith.

Based on these concepts, we build a classifier to conduct sentiment analysis, focusing on specific words in the tweets that correspond to features in the item of interest with respect to the given domain. The steps of our hybrid approach for sentiment analysis are explained next.

5.3.4.2 Steps of Sentiment Analyzer

We build the sentiment analyzer with the following steps:

- 1. Create a Twitter Developer Account:* Twitter requires authentication by OAuth (Open Authorization) to use the Twitter API for any application. To collect tweets using this, the user needs to authenticate requests. We thus create a developer account to get the authentication.
- 2. Collect the Tweets:* To collect tweets, we use Twitter API and Amazon Web Services

[AWS]. For this, we create the S3 bucket and then code Python scripts to collect tweets with keywords, e.g., iPhone, Samsung Galaxy, Amazon Fire etc.

3. Store the Tweets in a Database: Once the desired data is in the S3 bucket, we download the files, convert to CSV files and import them to a MySQL Database for further computation.

4. Implement the Analyzer: We implement the sentiment analyzer using TextBlob, a Python library for processing textual data. The details of the implementation are explained in the next section.

5. Analyze the Results: We get the information about the polarity of each feature in a tweet, which is stored in json file. The Python script calculates and classifies the features with polarities from this file. Thus, as an output, we get the number of positive and negative tweets about features of a product, e.g., for a given model of the iPhone, it gives the average polarity of tweets for its battery, camera etc.

6. Visualize the Output: Using polarity information, we visualize the data and present it in a user-friendly manner. Graph plotting can be done using IPython Notebook, MatLab etc. GUI development can be done as needed. This extends the console-based approach to *interactive* computing in a qualitatively new direction, providing Web-based applications suitable for capturing the computation process: developing, documenting, executing code and communicating the results.

After building the sentiment analyzer using these steps, the results plotted in graphical form

allow the end users to easily detect which features are good or bad in an item. This can be helpful for making decisions.

5.3.5 Implementation of Sentiment Analysis Approach

We implement the sentiment analyzer using TextBlob. This is a Python library for text data processing that provides a consistent application programming interface (API) for diving into common NLP (natural language processing) tasks such as POS (part-of-speech) tagging, noun phrase extraction and further analysis. TextBlob stands on the giant shoulders of NLTK and Pattern. NLTK is the Natural Language Tool Kit for Python that helps to build Python programs to work with human language data. Pattern is Python's Web mining module with tools for machine learning, data mining, network analysis and more.

In this implementation, we use the TextBlob classifier module to classify the tweets as positive or negative. Tweets are stored in MySQL database. MySQL DB module of Python is used to communicate with database. Using this, MySQL connection is established, tweets are fetched from table and each tweet is processed as follows.

First, the tweet is cleaned. For example, consider a tweet from 2014: "Yes it's true, the revolutionary iPhone6 is up for launch, finally! \Have you Pre-registered? \#iPhone6india @MehekMahtani". We need to remove hashtags, usernames etc. If there is any URL, we should remove that as well. Also, extra spaces, multiple characters should be removed. We use following

two functions as shown in Figure 5.20 for tweet cleaning.

```
def replaceDuplicates(s):
    #replace repetitions
    pattern = re.compile(r"(\.)\1{1,}" , re.DOTALL)
    return pattern.sub(r"\1\1", s)

def processTweet(tweet):
    # clean the tweets
    #Convert to lower case
    tweet = tweet.lower()
    #Remove www.* or https://.*
    tweet = re.sub('((www\.[\s]+)|(https?://[\s]+))', '', tweet)
    #Remove @username
    tweet = re.sub('@[\s]+', '', tweet)
    #Remove additional white spaces
    tweet = re.sub('[\s]+', ' ', tweet)
    #Replace hashtags with word
    tweet = re.sub(r'#([\s]+)', r'\1', tweet)
    #trim
    tweet = tweet.strip('\ ')
    return tweet
#end
```

Figure 5.20: Functions to clean the tweets

First, we convert each tweet into lowercase. Then we check if it contains any URL by searching for *www* and *https* in the tweet. If found, we just remove the URL. Similarly if we find *@username*, we remove it as we do not need the username to classify polarity. We also replace hashtag with a normal word that describes the hashtag. By processing each tweet in this manner, we minimize clutter and provide clean tweets to Textblob for better accuracy. Pattern, the Web mining module of Python is used to find repetitions of particular characters in a tweet. We replace the repeated characters for better accuracy. After cleaning the tweet, we pass the tweet to Textblob for classification.

Once the tweet is classified, we extract features from the tweet (e.g., Battery Life, Size, Looks etc.). For this, we make use of Sentiword [11] which is a huge lexical source of words and their respective sentiments. We extract Nouns and Verbs from the tweet and check the feature list to find out whether any of the features are present in the tweet. An example of a feature list appears in Figure 5.21

```
pFeatures=["iphone6","iphone6","phone","battery","size","bluetooth","software","price","usage","use","keypad","look","looking","camera","display","color","handsfree","upgrade","screen","feel","feels","shape","audio","construction","apps","memory","video","games","security","service","rebate","internet","quality","OS"]
```

Figure 5.21: Feature List Example

If a token matches, we store the features with polarity, subjectivity and sentiwords (if any) of tweet in json file. Thus, we get the class (positive / negative) and the feature of the tweet. For this we download the positive and negative words and import them in the database. We find that adjectives and adverbs in the tweet (ADJ and ADV) show the actual sentiment / opinion about the feature (Noun / Verb) in the tweet. To store sentiwords, we classify them in two categories. The sentiment is 1 if the word is positive and it is 0 if the word is negative. Figure 5.22 shows a partial snapshot of a sentiword table in a MySQL database.

Showing rows 0 - 29 (~6,788 total), Query took 0.0007 sec

```
SELECT *
FROM `Sentiwords`
LIMIT 0, 30
```

Profiling [Inline] [Edit] [Expl

1 > >> | Show : Start row: 30 Number of rows: 30 Headers every 100 rows

+ Options

				words	sentiment
<input type="checkbox"/>				a+	1
<input type="checkbox"/>				abound	1
<input type="checkbox"/>				abounds	1
<input type="checkbox"/>				abundance	1
<input type="checkbox"/>				abundant	1
<input type="checkbox"/>				accessible	1
<input type="checkbox"/>				accessible	1
<input type="checkbox"/>				acclaim	1
<input type="checkbox"/>				acclaimed	1
<input type="checkbox"/>				acclamation	1
<input type="checkbox"/>				accolade	1

Figure 5.22: Partial snapshot of a sentiword table

After conducting this implementation, we get an output file which stores the polarities of the tweets. Figure 5.23 shows an example of a *json* file used to store the output.

```
{
  "polarity": 0.175,
  "sentiments": [
    "revolutionary"
  ],
  "features": [
    "iphone6"
  ],
  "subjectivity": 0.825
},
{
  "polarity": 0.5,
  "sentiments": [
    "unimaginable"
  ],
  "features": [
    "iphone6"
  ],
  "subjectivity": 0.5
},
}
```

Figure 5.23: Example of output file

This output is with reference to the iPhone 6 product example from 2014. A Polarity of 1.0 means that the statement is 100% positive while negative polarity is denoted by -1.0. Also, the statement is either subjective or objective. If the tweet is objective, it is denoted by 0.0 whereas the extent of subjectivity is expressed by the term “*subjectivity*” in output file. Thus, higher the number, higher is the subjectivity of the tweet. The output can be used for visualizing the results. Visualization can be done by graph plotting and helps to make the output more appealing to the end users at-a-glance. An example of graph plotting is shown in the experimental results as described next.

5.3.6 Experimental Evaluation

We conduct the performance evaluation of our sentiment analyzer with real data from

Twitter. A summary of our evaluation is presented herewith.

An important category of our experiments involves collecting tweets in the smartphone domain. This is with the goal of providing recommendations to buyers and sellers of various smartphones and helping product launches. We collect real data from tweets and process it as follows.

5.3.6.1 Data on iPhone 6

The datasets we download here consist of around 7000 tweets (2000-iPhone6, 2000-iPhone6 plus, 1000-Samsung Galaxy, 1000-Amazon Fire, 1000-HTC) from the year 2014. These tweets are analyzed and used to find the sentiments about the products. We plot the results of sentiment analysis in graphical form, so as to enable the end users to easily analyze which features are good or bad in the products.

An example of such a graphical plot appears in Figure 5.24. This figure shows that according to tweets, sentiments are positive for iPhone6 model in general (0.2) but are negative specifically for its Camera (-0.8) and its Battery (-0.4). Thus, we can infer that influential users entering these tweets are more interested in the overall outlook of the iPhone 6 per se than in some of its individual features. Such information can be used in product recommendations.

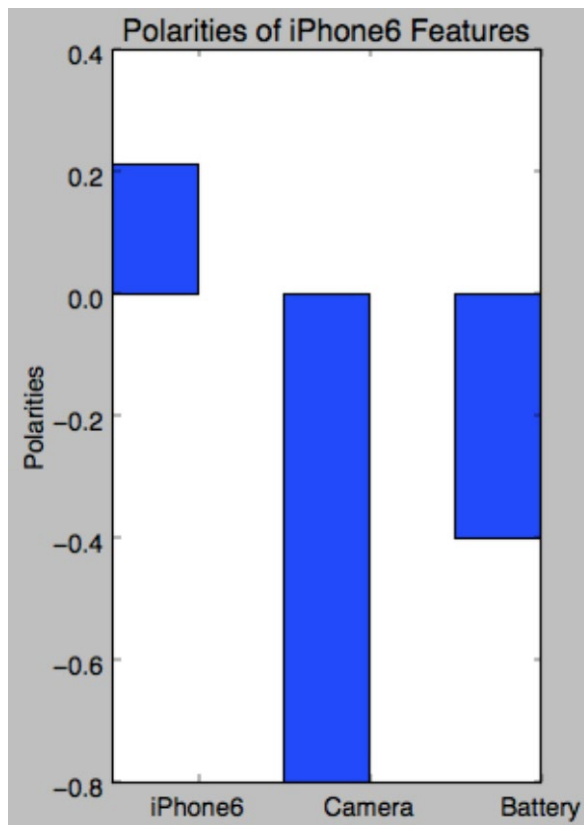


Figure 5.24: Graph plotted from sentiment analysis for iPhone 6 product review

In retrospect, we find that these recommendations have actually been useful. The iPhone6 in 2014 did indeed get very well-received by customers overall, however its battery seemed to pose some problems and its camera was often rated relatively low as compared to that of other products (such as Samsung Galaxy). Thus, the experimental evaluation presented herewith based on sentiment analysis conducted corroborates the real reception of the product. This confirms the validity of the sentiment analysis.

5.3.6.2 Data on Peatland Fires

Peatlands have much organic matter caused by decomposition of plant residue. Indonesia has the maximum peatlands in South East Asia. Pollutants from these fires affect neighboring areas, e.g., Singapore due to which Indonesian Peatland Fires (IPFs) are considered international hazards in Environmental Management. Airborne particulates pollution is a major concern. Research shows that rhinitis, asthma, and respiratory infections increase if particulate concentration is of hazardous level [12]. Thus, regulatory policies have been passed by Singaporean urban agencies to counterbalance the hazardous impact of IPFs. Singapore has an air quality system PSI (Pollutant Standards Index) with 6 pollutants: sulphur dioxide (SO₂), particulate matter (PM₁₀), fine particulate matter (PM_{2.5}), nitrogen dioxide (NO₂), carbon monoxide (CO) and ozone (O₃). Their environmental agency publishes PSI levels hourly through websites such as haze.gov.sg. People get this PSI information through and tweet their reaction on daily PSI level and air quality.

We use this Twitter data in the experiments shown here and analyze it based on the approach described in this paper. This gauges the sentiments of the public expressing their opinion on the policies taken by their agencies to deal with this event, namely Peatland fires. The results of the sentiment analysis are summarized in Figure 5.25. The data shown in this chart is based on two different user groups in the same region but different time periods, separated by six

months.

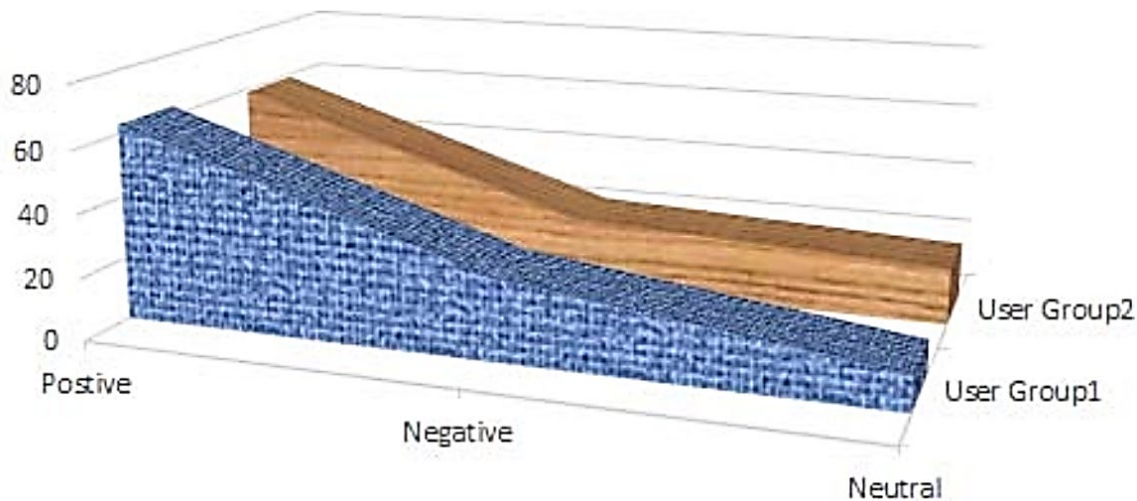


Figure 5.25: Area Chart for Sentiment Analysis of IPF Impact

The chart shows that policies to counterbalance the effect of IPF-based pollution appear to be fairly good since around 60% of users express positive sentiments. Yet, there is potential for improvement as around 25% of users are neutral and 15% are negative in their sentiments. This opinion mining thus provides useful inputs to government bodies in the respective region and also to its prospective residents.

5.3.6.3 Data on NYC Ordinances

We investigate data on ordinances or local laws in the NYC metropolitan area [13]. In the experiments shown here, we collect around 5000 tweets posted by the public in NYC pertaining to ordinances on various general policies pertaining to the economy, transportation etc. These

tweets are subject to sentiment analysis using the approach explained herewith. The results are summarized in the pie chart in Figure 5.26.

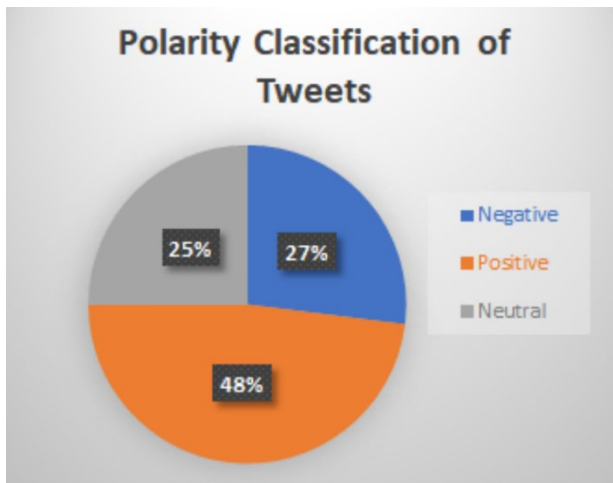


Figure 5.26: Pie Chart on Public Reactions to NYC Ordinances

It is clear from this pie chart that the percentage of positive tweets is the highest of all, thus we can infer that residents of NYC seem to approve of their urban policies on the whole. However, note that positive sentiments are conveyed by less than half the total number of residents here, which means that there is scope for further enhancement with respect to urban policy. Such charts can serve as recommendations to the urban management agencies. Further details can be provided on the specific aspects in which there is need for improvement, e.g., similar charts expressing sentiments on policies in the economic sector alone or transportation sector alone etc. Thus, if the greatest percentage of negative tweets come from a particular sector, it can be inferred that there is need for better policies in that sector. Conversely, if the public seems highly satisfied with policies on a given aspect, e.g., environment, that serves as positive

feedback in recommendations.

Likewise, several experiments have been conducted. The results of these experiments have actually been found useful in providing recommendations to prospective users.

5.3.7 Recommender Applications

Based on sentiment analysis conducted with real data, we now describe its targeted applications in recommenders.

5.3.7.1 Product Reviews

We consider reviews from the angles of sellers, buyers and product launches.

Seller: Our sentiment analysis approach can predict how the product is doing in the market. For example, by studying the graphical plot shown in Figure 5.24, a seller can determine that users are interested in the iPhone6 overall. Likewise, with other such plots, Apple can get an idea of how well a given iPhone model is received by public, what features should be incorporated in next model etc.

Buyer: From a buying angle, our approach can help in comparing features from various sellers to decide which product or service best suits their needs, e.g., if features pertaining to hotel reviews from Websites like tripadvisor.com are visually depicted (analogous to Figure 5.24), users can select a hotel based on previous reviews.

Product Launch: Our approach can make existing users affect new product launches. For instance, controversies related to movies can lead to a good start. Referring to Figure 5.24, if there were negative tweets about a movie (as for cameras and batteries here), its launch would likely succeed (as the iPhone 6 launch did). This is because existing viewers would convince new ones to watch the movie to find controversies.

5.3.7.2 Political Elections

We focus on outcome prediction and campaigning processes in political elections where recommendations matter.

Outcome Prediction: The sentiments expressed by influential users on social media can be used to predict victory in elections. For example, if a candidate is as well-received as an iPhone 6 (in Figure Figure 5.24), he or she is quite likely to win. This is because people freely express their views about candidates / parties on social media, so if their sentiments are positive, that reflects well about candidates.

Campaigning Processes: Influential users create awareness among people about positive and negative changes. For instance, candidates not leading based on tweets, can outline strategies referring to positive sentiments expressed about their opponents, and build campaigns accordingly.

5.3.7.3 Search Engine Optimization

In applications for SEO, i.e., Search Engine Optimization, we consider market trends and blogging.

Market Trends: Trends in the market pertaining to hot topics can be captured using our sentiment analysis approach. For example, with reference to our experiments, if we find that *iPhone* occurs more frequently than *HTC*, we can conclude that people are more excited about the iPhone than the HTC smartphone. Adding these keywords in Website contents will lead to more hits which in turn will help in increasing the page rank, through SEO.

Blogs: Several blogs become popular if the content is interesting. Our sentiment analysis approach can help find topics in which people are interested. Thus, bloggers can get information about people's choices. This information can be used to create new blogs and improve existing ones.

5.3.7.4 Stock Market

In the stock market area, we focus on two aspects, namely, bulls & bears, and price changes. We see how recommenders impact these applications.

Bulls & Bears: When a high proportion of investors express a bearish (negative) sentiment, some analysts consider it to be a strong signal that a market bottom may be near. Likewise, if sentiment is bullish (positive), analysts consider that market will go up. Our sentiment analysis

experiments taking into account such polarities with graphical plots (see Figure 5.24) can thus be helpful in estimating bulls & bears in the stock market.

Price Changes: Investors and stockholders measure sentiments by analyzing and mining textual stories about companies and sectors. Positive sentiments could lead to increase in stock prices whereas negative sentiments about the company could lead to decrease in prices. Our sentiment analysis approach could thus cause influential users to have an impact on stock prices.

5.3.7.5 Urban Policy

In this work, we consider urban policy issues related to specific events and general legislation.

Specific Events: Policy makers often take certain measures to act upon significant events that have occurred, e.g., flood, famine, fire etc. The sentiments expressed by users on social media sites such as Twitter enable us to gauge the reaction of the public on the satisfaction with these policies, with respect to how much they cater to addressing the respective issues pertaining to the corresponding events. Sentiment analysis on this can provide recommendations to policy makers on these specific aspects, so they can enhance current policies as needed and plan for future occurrences accordingly.

General Legislation: Urban regions often have local laws that affect the general lifestyle of the public on a daily basis. These could pertain to transportation systems, healthcare issues, use

of mobile devices, education facilities and so on. Residents often express their opinion on such policy matters through social media to make their voice heard. Sentiment analysis of such data can enable legislators to get a good idea of the impact their legislation makes on the common public. This can be useful for recommendations to enable better decision making through public involvement.

5.3.8 Conclusions

Microblogging has emerged as a major type of communication today. Large amounts of data in microblogging sites provide useful inputs for sentiment analysis. In our work on sentiment analysis in this paper, we make the following contributions: Present a method for opinion mining that would be useful in various recommender applications.

- Propose a hybrid learning approach with Naive Bayes for sentiment analysis, using probabilistic estimates where exact labels are not available
- Conduct evaluation on real data relevant to product reviews and urban policy
- Visualize the results of sentiment analysis for easy depiction to end users to facilitate recommendations

Ongoing work includes conducting more experiments with other domains. Some of our ongoing research also entails incorporating commonsense knowledge in sentiment analysis to simulate human judgment in opinion mining [14, 15]. As future work, we could consider

enhancing our approach further to include a combination of classifiers in an ensemble.

5.3.9 References

© 2018 IEEE. Reprinted, with permission, from Ketaki Gandhe, Aparna Varde and Xu Du, Sentiment Analysis of Twitter Data with Hybrid Learning for Recommender Applications, IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (IEEE UEMCON 2018), Nov 2018, DOI: 10.1109/UEMCON.2018.879666.1

[1] Das, S. and Chen, M. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. *Asia Pacific Finance Association Annual Conference (APFA)*, pp. 79-86.

[2] Whitelaw, C., Garg, N., Argamon, S. 2005. Using appraisal groups for sentiment analysis. *ACM CIKM*, pp. 625-631.

[3] Pang, B. and Lee, L. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. *ACL*, Article 271.

[4] Yang, C., Lin, K. H., and Chen, H. 2007. Emotion classification using web blog corpora. *IEEE/WIC/ACM Intl. Conf. Web Intelligence (WI)*, pp. 275-278.

[5] Read, J. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *ACL Student Workshop*, pp. 43-48.

[6] Go, A., Bhayani, R., and Huang, L. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford Digital Library Technologies Project.

-
- [7] Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *AAAI Conference on Weblogs and Social Media*, pp.178-185.
- [8] Nasukawa, T., and Yi, J. 2003. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. *K-CAP*, pp. 70-77.
- [9] Pak, A., and Paroubek, P. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREC*, pp. 19-21.
- [10] Russell, S. and Norvig, P. 2003. *Artificial Intelligence: A Modern Approach*. 2nd Edition. Prentice Hall.
- [11] Baccianella, S., Esuli, A. and Sebastiani, F. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *LREC*.
- [12] Zhou, J., Chen, A., Cao, Q., Yang, B., Victor, W., Chang, C. and Nazaroff, W. 2015. Particle Exposure during the 2013 Haze in Singapore: importance of the built environment. *J. of Building & Environment*, pp. 14-23.
- [13] NYC Council.Legislature <http://legistar.council.nyc.gov/>, 2018.
- [14] Tandon, N., Varde, A. and de Melo, G. 2017. Commonsense Knowledge in Machine Intelligence. *ACM SIGMOD Record*, 46(4): 49-52.
- [15] Puri, M., Du, X., Varde, A. and de Melo, G. 2018. Mapping Ordinances and Tweets using Smart City Characteristics to Aid Opinion Mining. *The Web Conference WWW Companion*

Volume, pp. 1721-1728.

Chapter 6

6. Related Work

6.1 Public Opinion Matters: Mining Social Media Text for Environmental Management

Abstract: Social media mining has proven useful in multiple research fields as a tool for public opinion extraction and analysis. Such mining can discover knowledge from unstructured data in booming social media sources that provide instant public responses and also capture long-term data. Environmental scientists have realized its potential and conducted various studies where *public opinion matters*. We focus our discussion in this article on mining social media text on environmental issues, with particular emphasis on sentiment analysis, fitting the theme of *Data Science and Sustainability*. The data science community today is interested in topics that overlap with environmental issues and their broader impacts on sustainability. Such work appeals to scientists focusing on areas such as smart cities, climate change and geo-informatics. Future issues emerging from this research include domain-specific multilingual mining, and advanced geo-location tagging with demographically focused sentiment analysis.

(Chapter 6 reused the previously published paper Du, X., Kowalski, M., Varde, A., de Melo, G., & Taylor, R. (2020), Public opinion matters, *ACM SIGWEB Newsletter*, (Autumn 2019), pp. 1-15, <https://doi.org/10.1145/3352683.3352688>).

6. 1. 1 Introduction

The rapid growth of the social media industry has attracted vast numbers of users, offering valuable insights to researchers. The total number of social media users is over 2 billion worldwide. The data therein have proven pivotal due to the vast amounts of timely information as well as of long-term data for longitudinal studies. Social media mining provides opportunities to extract opinions on topics such as political issues, consumer products, emergency incidents and environmental concerns. In light of this, many environmental scientists emphasize the power of opinions expressed on social media. These include researchers on urban policy, energy conservation, transportation aspects, health issues, sustainable living, and smart cities [Zou et al. 2018; Taylor 2012; Gandhe et al. 2018].

This survey article aims to provide a review on social media text mining applications from an environmental perspective. This encompasses aspects such as climate change, smart cities, traffic management, urban policy and energy conservation, thereby fitting *Data Science and Sustainability*, a prevalent theme today. For instance, ACM KDD 2014 had *Data Science for Social Good* as its the theme, while ACM CIKM 2017 had the theme *Smart Cities, Smart Nations*, which is closely related.

6. 1. 2 Environmental Applications

6. 1. 2. 1 Climate Change and Global Warming

The public often expresses concern via social media posts related to *climate change* and its adverse impacts on sustainable living. Recent NYC TV News during climate week in September 2019 showed numerous students from local schools joining peaceful protest marches on climate change related issues. They carried banners with slogans such as "There is no Planet B" voicing their views about having nowhere to go if planet Earth deteriorated drastically in the future due to climate change and global warming. Likewise, many users post climate change-related opinions on social media platforms such as Twitter. These hence facilitate public opinion mining across time and space, since geo-tagged postings contain timestamps and geographic coordinates of latitude and longitude.

In a recent study [Dahal et al. 2019], geo-tagged tweets with keywords on climate change are mined using topic modeling and sentiment analysis. LDA is deployed for topic modeling to draw inferences from various discussion issues, while a Valence Aware Dictionary and Sentiment Reasoner are applied to conduct sentiment analysis for gauging the feelings and attitudes in the tweets.

LDA (Latent Dirichlet Allocation) is a well-known technique for topic modeling, widely surveyed in many studies [Rozeva and Zerkova 2017]. It is a generative statistical model used for

sets of observations to be explained by unobserved groups that describe why some portions of data are similar to others. For example, if observations are words gathered within documents, LDA postulates that each document is a combination of a few topics and that the presence of each word is caused by one of the topics in the document. LDA can thus be used for topic modeling in environmental studies to find the prevalence of subjects in public posts.

Sentiment analysis often takes the form of polarity classification, i.e., judging whether the concerned sentiment in the text is *positive*, *negative* or *neutral*, and often the extent to which it heads in that direction, e.g., *strongly positive* etc. For instance, with reference to environmental management, users may state that they are “dissatisfied” with a climatic occurrence or legislative policy, which is a negative emotion, versus that they are “infuriated”, which is a much stronger negative emotion.

The authors perform a comparison between climate change discussions across several countries over different time periods. Not surprisingly, the overall sentiment in the tweets is *negative*. This negative sentiment is even more emphatic when users express opinions on “political situations” affecting climate change or on events related to “extreme weather conditions”. Interestingly, the study reveals that the climate change discussion is diverse, yet some topics are more prevalent, e.g., climate change posts in the USA are less focused on policy-related topics than corresponding posts in other countries. A broader impact of this study could entail further investigation on policy-related matters in climate change. It is important to fathom

why the public currently expresses fewer opinions on policy-related issues in the USA in comparison with other parts of the world.

In another interesting study [Wang et al. 2015], a supervised classification method is designed to process 76 million Weibo posts on climate change, aiming to discover connections between public responses and air pollution levels. The authors collect 93 million messages from 74 cities, finding that the volume of relevant posts is connected to pollution levels. Potentially relevant words on "pollution" from a probabilistic topic model are shown in Figure 6.1. These words are utilized to filter the Weibo posts. The study builds a 2-level classifier with randomly selected messages as the training data: the 1st level is designed to distinguish between related and unrelated messages; the 2nd level is meant to classify the related messages into "request-for-action" category versus a "pollution-experience" type. The authors suggest that combining a supervised method with an unsupervised method using LDA for topic modeling can yield higher correlations. As a broader impact, their research indicates that social media mining can be effective for air pollution monitoring even with a light-weight method.



Figure 6.1: Topics about pollution learned from a probabilistic topic model [Wang et al. 2015]

There is interesting work on air quality assessment by mining over structured data and unstructured social media text [Du et al. 2016]. In this work, the authors apply association rules, clustering, and decision tree classification over structured data sources on fine particle air pollutants. They also conduct opinion mining on tweets about Indonesian Peatland Fires (IPF) and their impact on the nearby country of Singapore, since these may affect the climate. The results are used for air quality analysis from a health standpoint by using worldwide AQI (Air Quality Index) standards. A commonsense knowledge repository called *WebChild* [Tandon et al. 2014] is consulted to build domain-specific knowledge bases that capture useful domain knowledge and enable subtle human reasoning in opinion mining. A sentiment polarity classification of tweets is conducted to analyze public responses using SentiWordNet 3.0

[Baccianella et al. 2010], a lexical resource. Methods in this work can be applied to social media text mining on related topics, e.g., water quality (analogous to air quality) gauging crucial sentiments of the public, with demographics.

A related study [Sachdeva et al. 2016] conducts research on social media activity in response to the 2014 King fire in northern California. The authors induce topic models with unsupervised feature selection methods to scrutinize users' behavior on Twitter. They compare spatial and temporal variations of the most frequent topics in tweets. The results show that there are significant differences between tweets of users from regions closer to vs. further from the fire. Also, discussions about arson and threatened houses are not as persistent as air quality concerns and potential health impacts. The authors conclude that a deeper sentiment analysis of tweets with wider data coverage could yield better results. As broader impacts, they suggest that combining social media text mining and spatio-temporal analysis can support inferences on related issues about the environment.

6. 1. 2. 2 Urban Policy and Local Laws

Social media serves as a powerful tool for urban residents to express views on policies passed by their legislature. Likewise, the public also expresses opinions on various urban trends, including population growth/decline, and associated policies.

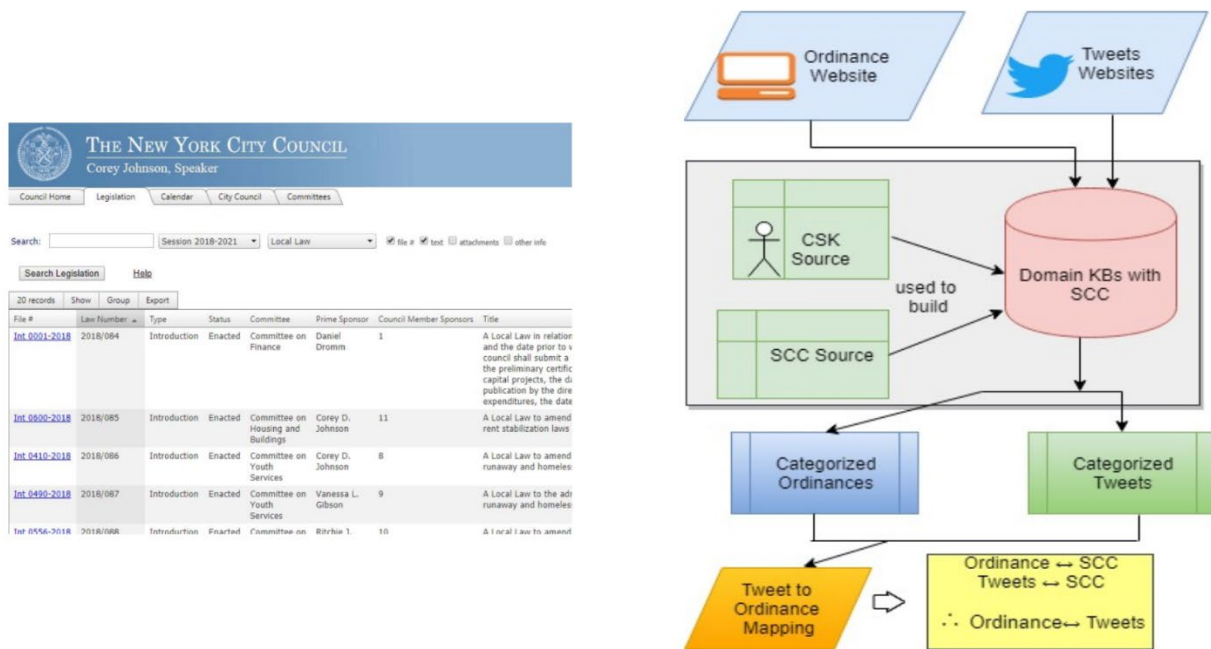


Figure 6.2: NYC Council ordinance website (left) and approach for ordinance–tweet mapping (right) [Puri et al. 2018]

In urban policy mining, a process for analyzing ordinances (local laws) and their public reactions expressed via tweets has been proposed in recent work [Puri et al. 2018]. Ordinances in this study stem from the publicly available NYC Council ordinance website. The authors aim to analyze how closely a given urban region heads towards developing into a smart city by mapping groups of ordinances and tweets to smart city characteristics (SCCs) and conducting sentiment analysis of tweets on the respective SCCs to assess public satisfaction. They consider a set of six SCCs: *Smart Environment*, *Smart Governance*, *Smart Living*, *Smart Mobility*, *Smart People*, and *Smart Economy*, as defined in the literature [TU-Wien 2015]. The mapping process is depicted in

Figure 6.2. It entails adopting commonsense knowledge from sources such as WebChild [Tandon et al. 2014] and WordNet [Fellbaum 1998] to harness human judgment involved in mapping, in line with the concept of deploying humanoid common sense within the realm of machine intelligence [Tandon et al. 2017]. The authors map groups of ordinances with tweets using the transitive property: "If ordinances map to SCCs and if the tweets map to the same SCCs, the ordinances are likely to be broadly related to the respective tweets". This substantially reduces the sample space for ordinance–tweet mapping, since ordinances and tweets are of the order of thousands and millions, respectively. Further mapping is conducted with the *word2vec* approach (widely used by many researchers, see Rozeva and Zerkova [2017]), for finding contextual similarity in the reduced mapping space.

This mapping sets the stage for sentiment analysis on the tweets by polarity classification encompassing commonsense knowledge [Tandon et al. 2017]. Results of the ordinance–tweet mining reveal the overall public satisfaction on ordinances related to various SCCs. The results suggest a positive public sentiment towards New York City as a smart city. The authors also analyze avenues for potential improvement based on public feedback. This information can be useful to urban agencies to adjust policies accordingly. This concept addresses *smart governance*, a smart city characteristic, that leverages transparency in urban decision-making through public involvement. More generally, the proposed approach [Puri et al. 2018] can be useful to map other data for opinion mining, e.g. *News and tweets*, since it is desirable to assess

public reactions to various news articles, current and historical. The approach herein for mapping formal legalese in ordinances to informal acronym-ridden tweets, is potentially helpful for mapping news and tweets since these also feature formal and informal text, respectively. This mining of public opinion on news leverages *smart governance* to a considerable extent, since it entails news scrutiny in order to assess public feedback.

In a relatively recent working paper [Hollander and Renski 2015], the authors conduct exploratory research on attitudes of people in urban areas. They focus on a study in the Urban Attitudes Lab, where micro-blogging data from Twitter are assessed with quantitative and qualitative methods, such as content analysis and advanced multivariate statistics. These methods are used for a detailed study on urban experience and its implications for public policy. The authors apply a propensity scoring mechanism to create matched pairs of mid-sized cities in the Northeast and Midwest United States, where the most significant difference between each pair is that of *population decline*. The outcome is a group of 50 declining cities paired with 50 growing/stable cities. More than 300,000 tweets over a 2-month time span are analyzed, for *positive* or *negative* sentiment. The authors conduct difference of means tests, concluding that the sentiment in declining cities does not vary much in a statistically significant manner compared to that in stable and growing cities. These findings, though rather surprising, present the scope for further research. They indicate that opportunities are available to enhance the comprehension of urban attitudes based on sentiment analysis of tweets from the respective areas. Reasons for a

lack of significant differences among attitudes of growing vs. declining cities would potentially be interesting to urban planning agencies, environmentalists and data scientists. Hence, this exploratory research presents promising avenues for future work on studies related to urban population growth/decline and urban policy.

The proposition of a sentiment analysis approach to extract emotions in tweets based on polarity and subjectivity, using partially labeled data, is described in recent work [Gandhe et al. 2018]. The authors put forth a hybrid approach combining supervised and unsupervised learning to take advantage of labeled training data if available, while also classifying tweets that lack specific labels. They build a classifier for sentiment analysis based on the Naive Bayes machine learning algorithm. They analyze tweets on issues such as political elections, stock markets and urban policy. The urban policy tweets are on general legislation as a whole, and on specific actions related to significant events, e.g., disasters. They implement this using TextBlob, a software library for text data processing that builds on NLTK (the Natural Language Toolkit) to better handle human language data. This research contributes to the idea of sustainable urban development being made *smarter* by mining social media data. This is achieved using hybrid approaches that produce useful results even when fully labeled training data are not easily available.

6. 1. 2. 3 Traffic and Mobility Issues

Social media users often post reactions about incidents that occur on the road. Issues on traffic and mobility also pertain to population relocation. Thus, sentiment analysis of the text in these posts can produce valuable results to support sustainable development and traffic optimization.

A classification based method is proposed by Gu et al. [2016] for mining tweets to extract incident information for highways and smaller roads. This method offers cost-effective data collection (see Figure 6.3) based on the Twitter API to obtain tweets and related information, especially location data. Using these data and metadata, the authors compare the tweets to an existing dataset to observe whether any discussed traffic incident matches with regard to the details and specifications, i.e., to authenticate its validity based on facts vs. opinions. They are also able to use this process to find additional incidents absent in the dataset but commonly discussed in the tweets. This sets the stage for further investigation on such incidents. This research is a leap in the direction of sustainable development and optimization of traffic management, by offering cost-effective social media text collection and propelling investigatory studies based on the workflow.

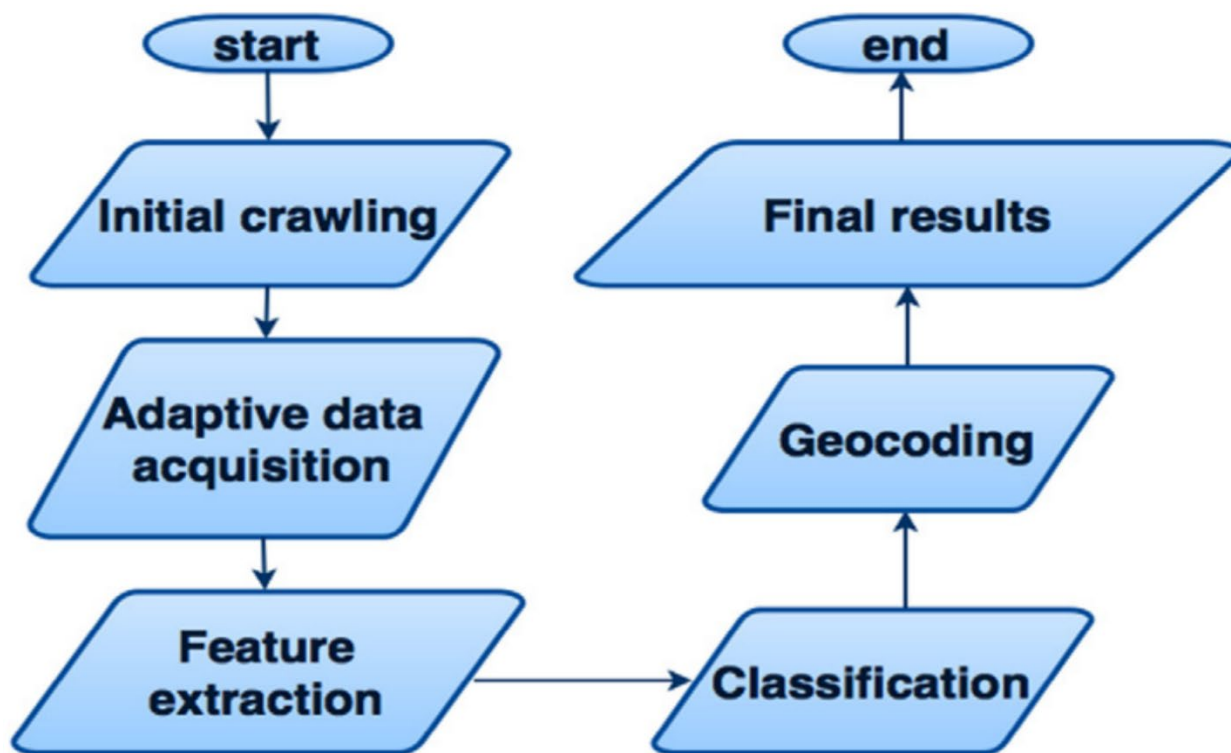


Figure 6.3: Workflow of Twitter data acquisition, processing and analysis for traffic problems [Gu et al. 2016]

The idea of utilizing large-scale social media data and valuable information therein (geolocations, times, dates and places) to infer land use within a given area has been the basis of appealing research [Zhan et al. 2014]. The researchers focus on collecting tweets having geolocations in NYC and deploying a 3rd-party location-based service *Foursquare* to get more accurate location information from the tweets. If a user on Foursquare performs a "check-in" at a specific location, the user can share that information on Twitter. Then, by referencing the geolocation of the tweet alongside the Foursquare data within the tweet, the proposed approach can draw inferences about the content of a tweet and its neighborhood. As these tweets are

continuously collected, they are categorized with respect to one of the following categories:

home, work, eating, entertainment, recreation, shopping, social service, education, and travel.

After this sorting, the approach obtains specific details from each category, and thus performs intricate land use inference. The authors suggest that similar approaches would allow cities of varying sizes to analyze land use and interaction in a given area, thus providing greater insights into the precise activities therein. This fits the theme of *smart living* in smart cities [TU-Wien 2015].

[Wang et al. 2017] present a new method to report traffic conditions, addressing shortcomings of prior approaches. The authors base their research on the idea that while GPS (Global Positioning System) probe data are extremely useful in our everyday lives, they prove rather inadequate at fully estimating traffic conditions due to the low sampling frequency. To overcome this problem, the authors propose using social media to collect further information on traffic events not common within the geographical area. To correlate the GPS probe data with social media, they focus on a deep analysis of the incoming social media data. This entails dissecting the social media posting text such that significant traffic-related phrases and locations can be separated and stored. Using these processed texts, they take the GPS probe data and fill in the missing data. From here, interesting patterns can be discovered about traffic conditions, which can be used to gain a deeper understanding of traffic commonalities in a given area. This research highlights that the process of collecting texts from social media in conjunction with

GPS probes, and utilizing them to analyze specific issues, bears substantial potential in our modern world. Although the study focuses on GPS data, it exemplifies the fact that social media data can be used to augment other data sources for enhanced mining. This work caters to the *smart mobility* aspect of smart cities [TU-Wien 2015] due to the relevance of smart monitoring of traffic

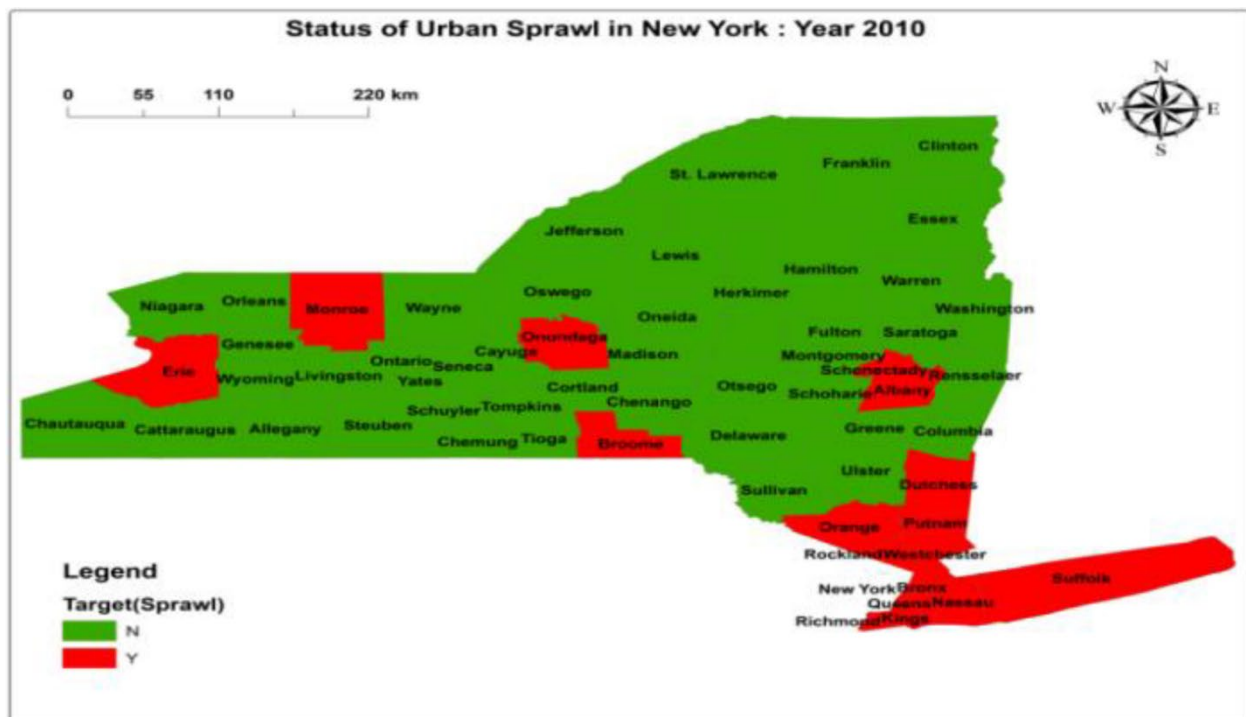


Figure 6.4: Interactive map of New York State with display of sprawl affected regions generated from GIS data

[Pampoore-Thampi et al. 2014]

The works of [2016], Zhan et al. [2014], and Wang et al. [2017] additionally highlight the potential for a deeper analysis on related issues such as urban sprawl. The term *urban sprawl* mainly implies the unrestricted growth of housing, transportation and commercial development

over vast expanses of urban land. Researchers aim to study sprawl-causing parameters to mitigate its effects. The study by Pampoore-Thampi et al. [2014] investigates sprawl using association rules and decision tree classifiers with GIS (Geographic Information System) sources. The authors generate interactive maps (using the classical ArcGIS software) that superimpose sprawl data on the respective geographic areas to provide at-a-glance views of sprawl-affected regions (see Figure 6.4 for New York State). Based on these data, they analyze the impact of various sprawl-related parameters on each other and on the sprawl itself. These parameters are various spatio-temporal features related to real GIS data, e.g., population growth, travel time to work, number of vehicles etc. These are mined to discover knowledge for a spatial decision support system (SDSS). This SDSS provides a predicted output on whether urban sprawl is likely to occur, given input parameters. It also estimates values of pertinent sprawl-related parameters to help understand their mutual impacts.

Research such as this can potentially benefit from sentiment analysis of social media text relevant to sprawl. For example, in addition to mining sprawl-related parameters, the mining of pertinent social media posts may yield further information to augment knowledge discovery. The reactions of the common public, legislators and scientists on sprawl, its causes and effects can be beneficial in understanding the gravity of certain aspects and assessing the relative importance of sprawl-related parameters. Such information can potentially be used to enhance systems such as the SDSS therein, to provide more well-informed decision support based on opinion mining.

These studies thus relate to sustainable living

6. 1. 2. 4 Energy and Resource Conservation

Fossil fuel combustion is a major source of ambient carbon dioxide (CO₂) concentrations. The increasing public awareness of greenhouse gas emissions leads to concerns about energy types and conservation of natural resources. Social media mining can provide valuable information about public opinions on energy usage and natural resources.

Understanding public reactions on energy and resources can be extremely powerful. In the thought-provoking study by Nuortimo [2018], the authors argue that collecting data from various social media platforms is beneficial when insight on a given topic is needed, especially if that topic is rather complex. They propose a system called Case Carbon Capture and Storage to reduce harmful CO₂ emissions. Such emissions can cause widespread environmental problems. The stages in their proposed system focus on: capture and compression of CO₂ from power stations; transportation of CO₂; and storage of this captured CO₂ in a manner that keeps it out of the atmosphere. While the introduction and development of the system in this work is not being actively investigated by regulators due to low incentives, the research makes use of public reactions to emphasize the need for such a system, hoping that it will allow the project to gain more traction. In order to obtain these public reactions, text mining of social media is performed. As the social media texts are entered, the processing is conducted such that only information on

the concerned system (Case Carbon Capture and Storage) is retained. From here, an analysis is conducted to observe public opinions on this environmental system.

Figure 6.5 provides a snapshot of opinion mining results on this system, based on posts from Social Media (SoMe). As seen here, the majority of the posts convey positive sentiments, however this majority is less than 50% of the total. While the number of negative posts are somewhat less than positive ones, they outnumber the mixed and neutral posts. This could potentially yield the inference that considerable further work might be needed for a much more widespread acceptance of the proposed technology by the public. Since concerns of many people expressed via social media are heavily opinionated, critical and often specific, important insights are gained into specific aspects needed to increase the public acceptance of the given system. Similar arguments can be applied to other such systems. Hence, in this work, sentiment analysis of social media text serves as a tool to increase the *public awareness* and *potential acceptance* of a new technology in environmental management, geared towards solving critical energy related problems.

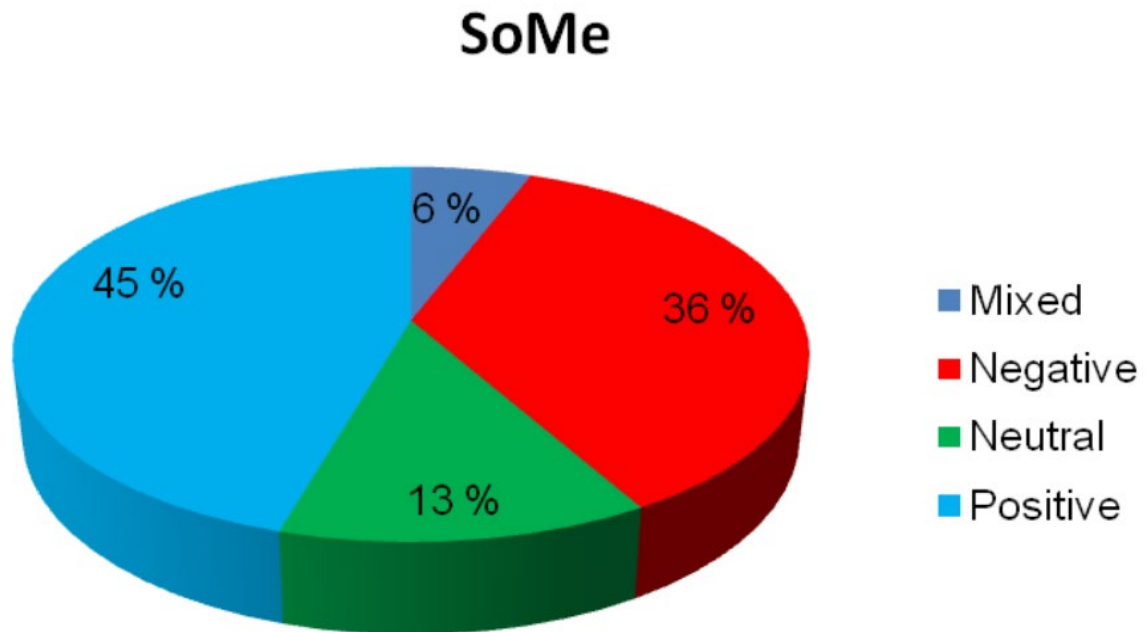


Figure 6.5: Summary of sentiment analysis over social media posts about the Case Carbon Capture and Storage system [Nuortimo 2018]

Researchers often focus on opinion mining to better understand the psychological determinants of social acceptance of environment-based technologies. In useful energy-related research, Nuortimo and Harkonen [2018] focus on an analysis of *failed technologies* in which *social acceptance has been a primary factor* in the failure. This allows for better predictability during the introduction of new technology. This work emphasizes that early acceptance of a technology by the public is extremely important as new work is developed. Public opinions are obtained via machine-based social media data mining and analysis. This research furthers the idea that social media mining offers valuable insights in assessing eco-friendly technologies.

This can be applied to *sustainable computing*, where the public often has mixed views about “greenness and energy conservation” versus “productivity and efficiency”. Social media mining can reveal highly useful results here on the universal acceptance of eco-friendly technologies and policies.

In order to gain awareness for wildlife conservation, environmental scientists in China [Wu et al. 2018] applied social media to their research. They considered WeChat, one of China’s largest social networking platforms, studying online news and relevant public comments in media posts about “*Sousa chinensis*” (Indo-Pacific humpback dolphin), a flagship species in China. They analyze media releases on dolphins straying into the Dongping, Beijiang and Baisha rivers of China. They deploy Content Analysis (CA) to discover knowledge from wildlife conservation information found in articles and public opinions. Their results suggest that the public feels highly doubtful about conservation efforts proposed by government bodies and experts. This is a useful and rather unfortunate observation. An interesting finding of this study is that greater efforts are needed to promote awareness on wildlife conservation, e.g., rescue operations, so as to reduce public misunderstanding. This seems quite a debatable issue on whether the public is right in expressing views on governmental lack of concern for conservation efforts, or whether the government is right in issuing appropriate conservation measures that simply need better dissemination. This work has broader impacts on *sustainable living*. Findings from this research prove that social media posts are valuable in analyzing wildlife conservation,

where the public is highly opinionated, and consequently various debates continue.

6. 1. 2 .5 Disaster and Resilience

Researchers of environmental management pay attention to the issue of disasters, since it influences natural resources and human society. Disaster-related events often propel further social media activities. By analyzing these, knowledge about efficient disaster responses can be discovered to support environmental management. The concerned studies can also help in enhancing resilience, which enables a faster recovery and may mitigate the damage brought by the disaster.

Wang et al. [2018] remark that the scarcity of hyper-resolution data for urban flooding prevents a detailed flood risk analysis. To address this issue, the authors introduce social media and crowdsourcing data into the mix. They apply NLP (Natural Language Processing) and computer vision techniques to the data they collect from Twitter and from a crowdsourcing app called MyCoast. From there, they utilize the processed data to complement existing data. In particular, they validate the extracted information against precipitation data and road closure reports to examine the quality of the data, and then utilize the results as required. The introduction of this approach for the procurement of fresh and easy-to-collect data is extremely beneficial to current environmental management techniques, in terms of its broader impacts. The application of social media in conjunction with crowdsourcing to augment data collection of

otherwise rare datasets, is a useful contribution.

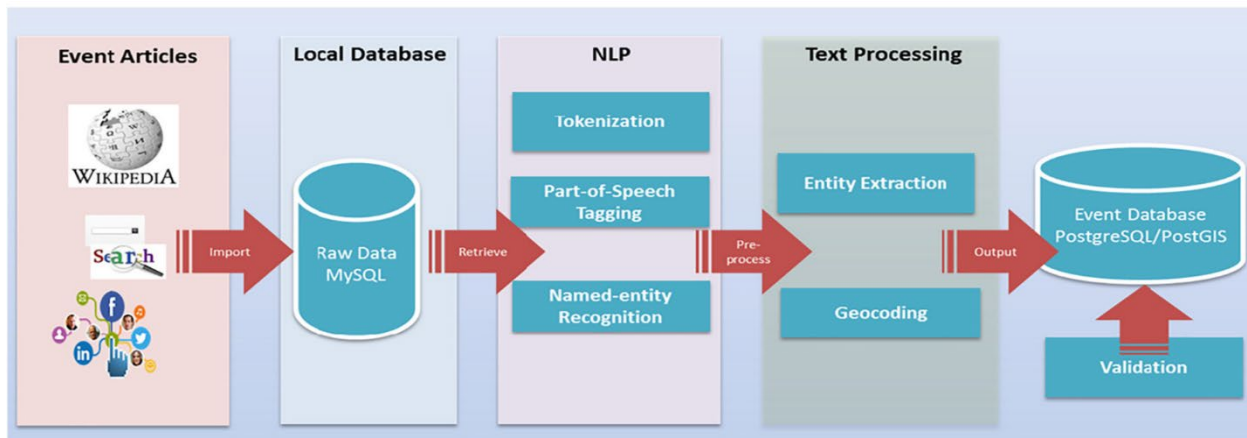


Figure 6.6: Workflow for a disaster event database [Wang et al. 2018]

The framework of Huang et al. [2017] synthesizes multi-sourced data (including social media postings, remote sensing data and Wikipedia) with spatial data mining and text mining for a solution that supports disaster analysis of historical and future events. This is illustrated in Figure 6.6. While Wikipedia is a primary source in their work, data from Twitter and other social media platforms are also utilized to obtain more information on disasters. Using all the collected data enables the discovery of patterns in disasters through various pattern mining methods. This allows us to obtain further information that may be missing from historical reports. This framework offers advantages for disaster analysis, since data sources are added through social media and other platforms in addition to historical reports on disasters. The authors claim that a more intricate analysis and processing can facilitate real-time event tracking, highly useful for enhanced performance in disaster management and recovery.

In disaster resilience on hurricane activity, there is research [Zou et al. 2018] that focuses on mining public reactions via Twitter. The authors focus on spatio-temporal patterns of Twitter activities during Hurricane Sandy that impacted the Northeastern USA in October 2012. The study leverages 126 counties impacted by Sandy. An important finding is that social and geographic disparities are prevalent in Twitter usage. Public communities with higher socioeconomic status are found to post more hurricane-related tweets. This study also derives common indexes from Twitter data including normalized ratios to facilitate comparison across regions, and to aid in emergency management and resilience analysis. Adding Twitter indexes to a damage estimation model is found to enhance the performance. The authors thus conclude that social media data can benefit post-disaster damage estimation, provided other pertinent environmental and socioeconomic parameters are also included. Although their research addresses one particular hurricane, their results and the knowledge gained from the study can yield extremely valuable insights into strategies for using social media to increase disaster resilience. This ability is imperative in understanding how a disaster truly affects the public and how social media can be used for a deep analysis of the concerned reactions. Disaster control and resilience are by far the most critical aspects of environmental management. The works surveyed here demonstrate that social media mining plays a vital role in providing additional data for analysis beyond other recorded sources. Moreover, it helps in monitoring public reactions on disaster repercussions and the availability of recovery mechanisms, which constitute the true

tests of a good disaster management system.

6. 1. 3 Discussion on Open Issues

Based on the above survey of the literature, we outline several thought-provoking ideas on social media text mining that offer the scope for future work from a generic web and text mining standpoint, as well as a domain-specific angle.

Demographics of posts: In the works on urban policy and local laws, it is useful to address the demographics of location-based social media posts in order to analyze public reactions to urban policies based on the backgrounds of people that post online. This could pertain to their educational, social and cultural background, as well as age and gender.

Historical diagnosis: Urban policy research can potentially entail the diagnosis of information pertaining to the historical analysis of various ordinances and their respective media posts, i.e., gauging public opinion before and after ordinance passing to assess the opinions of the public. Similar issues apply to the analysis of news and social media, i.e., these posts can be analyzed before and after a given news item is published.

Levels of granularity: While addressing the relevance of various social media posts to specific aspects of interest (e.g. tweets with regard to smart city characteristics [Puri et al. 2018]), it would be beneficial to consider the posts at a finer level of granularity. For instance, one might consider posts relating to the notion of *smart environment* in response to a given news

item or a local law. It would be interesting to focus on a specific aspect within smart environment, such as green energy, and thereby assess its impacts. The same reasoning can apply to analyzing media posts in response to news etc. considering other aspects, such as climate change or disaster recovery.

Automated geo-tagging: Third-party services for collecting data on social media mining (e.g. [Zhan et al. 2014]) may not be as effective as utilizing the social media platform itself. Not all media posts have geo-locations attached to them. Thus, methodologies can be formulated for better approaches in automatically geo-tagging the posts, which would further help in more precisely mapping a mined post to a given location. This would also propel advanced demographic analysis.

Crowdsourcing and monitoring: If the utilization of crowdsourcing apps and social media is required in the enhancement of hyper-resolution monitoring in some applications (e.g. [Wang et al. 2018]), issues would arise if a certain geographical area does not have a multitude of people providing constant updates via these apps. This motivates further research in methods used for crowdsourcing, with the goal of promoting better monitoring and analysis.

Veracity of posts: While investigating matters such as disaster recovery and resilience, data being collected via social media must be filtered and verified to avoid data manipulation. False data on such highly sensitive topics can result in misunderstandings. Hence, more research is needed on addressing the veracity of each social media post, especially pertaining to sensitive

issues such as disaster repercussions and resilience. While there is much research on veracity in general, and it constitutes one of the *Vs* of big data, some of this research needs to target a more domain-specific angle, especially for sensitive subjects.

Multilingual and multicultural issues: In current studies on social media text mining, the influence of language and culture has not been an item of significant focus. There is a lack of research providing comparisons of social media posts on a given topic among people speaking different languages and emerging from various cultures. This could be a potential topic of ongoing and future research, that forms multilingual and multicultural domain-specific social media analysis, driven by recent advances in cross-lingual natural language processing [de Melo 2017; Dong and de Melo 2019]. This is particularly relevant in our current era of increasing globalization.

Irony and sarcasm: Sentiment identification in social media text mining does not particularly emphasize linguistic subtleties such as irony and sarcasm. These aspects are rather difficult to measure in media posts. Also, expressions of irony as well as sarcasm may vary across different languages and cultures. These present avenues for further research, where the state-of-the-art in idiomatic expressions and emotion detection from formal written texts can play a significant role. Such analysis in informal texts, especially on domainspecific aspects such as environmental issues, can be quite challenging. This calls for further research.

Error correction tools: Social media text is usually noisy, with both spelling errors and

erroneous or non-standard grammar. Hence, progress on techniques to cope with these issues has the potential to benefit social media analytics in numerous different domains.

Abbreviations and acronyms: Excessive usage of abbreviations and acronyms in social media text often presents difficulties in mining. This challenge is even more pronounced when multiple domains are involved, each with different meanings of acronyms, thus leading to greater degrees of ambiguity and adding to the confusion caused by colloquial terms in public posts. Existing techniques in Named Entity Extraction (NEE), Named Entity Disambiguation (NED) and related areas need further research to be applicable to such informal language in social media posts, particularly with reference to context.

Big versus small data: Much of social media mining utilizes only small parts of the big data on social media. In many published studies, there is a lack of discussion about whether the small sample of data used is sufficiently robust and whether the exclusion of the bulk of remaining data can lead to adverse impacts and incorrect inferences with respect to the result interpretation. This calls for further research and discussion. For example, big data is useful to understand the big picture in a given context along with its hidden correlations. Small data may be too specific here. Hence, focusing on analyzing big data in social media with respect to the several *Vs* such as volume, velocity, variety etc. could present more interesting insights into social media mining. Some of these could be useful in domain-specific applications, where obtaining the big data itself could pose considerable challenges.

Multiple media sources: Very few studies use data from multiple sources of social media to address a single common topic. The comparison between posts on different social media sources (e.g., Twitter and Facebook) on the same topic could potentially be addressed in greater depth. This may yield even more meaningful and interesting results than analyzing each source individually, since the sources could provide a broader perspective on the opinions expressed. Such in-depth text mining over multiple media across a common thread of topics could be an aspect of future work.

6. 1. 4 Conclusion

This survey paper disseminates an overview of social media text mining applications in the environmental management area. It covers a number of facets, including climate change and global warming, urban policy and local laws, traffic and mobility issues, energy and resource conservation, as well as disaster and resilience. The topics discussed herein have significant broader impacts, as outlined in the respective subsections. These encompass news scrutiny, pollution monitoring, healthcare related decision-making, legislative transparency, traffic safety, climate change investigation, disaster repercussions, dataset enhancement for analysis, public acceptance of policies, sustainable computing issues, and the development of smart cities.

The papers surveyed in this article present the scope for future research on several topics such as multilingual domain-specific social media mining, enhanced geo-location tagging with

advanced demographic analysis, subtle issues such as irony and sarcasm in social media, veracity related to sensitive subjects, crowdsourcing research in conjunction with social media mining etc. We anticipate that addressing such topics for future research can make social media text mining an even more impactful area, with greater benefits to the data science community and various application domains.

6. 1. 5 References

- Baccianella, S., Esuli, A., and Sebastiani, F. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Intl. Conf. on Language Resources and Evaluation, LREC*. Valletta, Malta.
- Dahal, B., Kumar, S., and Li, Z. 2019. Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining, Springer 9*, 24.
- De Melo, G. 2017. Inducing conceptual embedding spaces from Wikipedia. In *Proceedings of WWW 2017*. ACM.
- Dong, X. and De Melo, G. 2019. A robust self-learning framework for cross-lingual text classification. In *Proceedings of EMNLP-IJCNLP 2019*. ACL.
- Du, X., Emebo, O., Varde, A., Tandon, N., Nag Chowdhury, S., and Weikum, G. 2016. Air quality assessment from social media and structured data. In *IEEE Intl. Conf. on Data Engineering - Workshop on Health Data Management and Mining (ICDE-HDMM)*. Helsinki,

Finland, 54–59.

Fellbaum, C., Ed. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Gandhe, K., Varde, A., and Du, X. 2018. Sentiment analysis of Twitter data with hybrid learning for recommender applications. In *IEEE Ubiquitous Computing, Electronics and Mobile Communications Conference (UEMCON)*. New York, NY, 57–63.

Gu, Y., Qian, Z., and Chen, F. 2016. From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies* 67, 321–342.

Hollander, J. and Renski, H. 2015. Measuring urban attitudes using Twitter: An exploratory study. Working Paper WP15JH1, Lincoln Institute of Land Policy, USA.

Huang, Q., Cervone, G., and Zhang, G. 2017. A cloud-enabled automatic disaster analysis system of multi-sourced data streams: An example synthesizing social media, remote sensing and Wikipedia data. *Computers, Environment and Urban Systems* 66, 23–37.

Nuortimo, K. 2018. Measuring public acceptance with opinion mining: The case of the energy industry with long-term coal R&D investment projects. *Journal of Intelligence Studies in Business* 8, 2, 6–22.

Nuortimo, K. and Harkonen, J. 2018. Opinion mining approach to study media-image of energy production - implications to public acceptance and market deployment. *Renewable and Sustainable Energy Reviews* 96, 210–217. PAMPOORE-THAMPI, A., VARDE, A., AND YU, D. 2014. Mining GIS data to predict urban sprawl. In *ACM Conference on Knowledge*

-
- Discovery and Data Mining (KDD), Bloomberg Track*. NYC, New York, 118–125.
- Puri, M., Du, X., Varde, A., and De Melo, G. 2018. Mapping ordinances and tweets using smart city characteristics to aid opinion mining. In *The Web Conference (WWW) Companion Volume*. Lyon, France, 1721–1728.
- Rozeva, A. and Zerkova, S. 2017. Assessing semantic similarity of texts - methods and algorithms. In *Intl. Conf. on Applications of Mathematics in Engineering and Economics (AIP Conf. Proc)*. 1–8.
- Sachdeva, S., Mccaffrey, S., and Locke, D. 2016. Social media approaches to modeling wildfire smoke dispersion: Spatiotemporal and social scientific investigations. *Information, Communication and Society* 20, 8, 1146–1161.
- Tandon, N., De Melo, G., Suchanek, F., and Weikum, G. 2014. WebChild: Harvesting and organizing commonsense knowledge from the web. In *ACM International Conference on Web Search and Data Mining (WSDM)*. NYC, New York, 523–532.
- Tandon, N., Varde, A., and De Melo, G. 2017. Commonsense knowledge in machine intelligence. *ACM SIGMOD Record* 46, 49–52.
- Taylor, R. 2012. Urbanization, local government and planning for sustainability. *Sustainability Science: the Emerging Paradigm and the Urban Environment*.
- TU-WIEN. 2015. European smart cities, technical report. Vienna University of Technology, Vienna Austria.

-
- WANG, R., MAO, H., WANG, Y., RAE, C., AND SHAW, W. 2018. Hyper-resolution monitoring of urban flooding with social media and crowdsourcing. *Data, Computers and Geosciences* 111, 139–147.
- Wang, S., Paul, M., and Dredze, M. 2015. Social media as a sensor of air quality and public response in China. *Journal of Medical Internet Research* 17, 3.
- Wang, S., Zhang, X., Cao, J., He, L., Stenneth, L., Yu, P., Li, Z., and Huang, Z. 2017. Computing urban traffic congestions by incorporating sparse GPS probe data and social media data. *ACM Transactions on Information Systems* 35, 4, 1–30.
- Wu, Y., Xie, L., Huang, S., Li, P., Yuan, Z., and Liu, W. 2018. Using social media to strengthen public awareness of wildlife conservation. *Ocean and Coastal Management* 153, 76–83.
- Zhan, X., Ukkusuri, S., and Zhu, F. 2014. Inferring urban land use using large-scale social media check-in data. *Networks and Spatial Economics* 14, 3, 647–667.
- Zou, L., Lam, N., Cai, H., and Qiang, Y. 2018. Mining Twitter data for improved understanding of disaster resilience. *Annals of the American Association of Geographers* 108, 5, 1422–1441.

Chapter 7

7. Conclusions and Future Work

7.1 Conclusions

In this research, we conducted data mining on urban policy based on structured data and social media. The data mining on air quality data and traffic conditions showed that the data mining techniques could serve prediction purposes and decision support. The mining on ordinance data showed that data mining could reveal interesting urban legislative activity patterns, potentially supporting urban policy decision-making processes. Because we connected the Smart City development aspect to the ordinance, the results could also improve sustainable city development.

In short the main contributions of this dissertation based on our major tasks are as follows:

1. Analyzing multicity urban traffic data pertaining to sustainable population relocation as early work.
2. Conducting data mining on PM2.5 fine particle air pollutants for prediction of air quality incorporating health standards and building a tool for result dissemination.
3. Modeling and mining publicly available ordinance data from NYC deploying data mining techniques of association rules, clustering and classification and assessing how well they make the region head towards a Smart City.

-
4. Mapping ordinances with their pertinent tweets expressed on social media, using Smart City Characteristics as a nexus, incorporating commonsense knowledge and text mining.
 5. Performing sentiment analysis on the tweets for opinion mining to gauge to the reactions of the public on their respective ordinances.
 6. Disseminating the results of our analysis through the development of an Android app for ordinance-tweet mining and a Web portal for QA (question answering) along with interactive graphics.

This dissertation would be the good step for further analysis of the interaction between urban policy and public opinions. Sentiment analysis also yielded interesting results, proving it can support future work, such as analyzing the change in sentiment of tweets related to the specific ordinance. The decision support tools developed based on the knowledge discovered through the research have acceptable accuracy; however, there would be a massive improvement in accuracy and application range with advanced methods and extended data sources. Thus, we propose future works to enhance the impact of this research, further contributing to Smart City development.

The Important findings of this dissertation are as follows:

1. The journal about multicity population relocation (Du & Varde, 2015) had findings that proper street design could mitigate urban sprawl; the mix between residential and employment land use types could lead to a compact city, and urban areas with

population and employment concentrated in the center had less chance of sprawl.

2. The ICICS 2016 paper about traffic conditions and air quality (Du & Varde, 2016) had findings that higher gas consumption usually indicated better economic conditions and more strict pollutant regulations, thus contributing to lower PM2.5 concentration; high income also contributed to low PM2.5 concentration; high diesel consumption did not always associate with high PM2.5 concentration.
3. The IEEE ICDE 2016 workshop paper about social media mining of peatland fires (Du et al., 2016) had findings based on sample tweets about Singapore's air pollution where 61% of collected tweets had a positive sentiment; yet there was room for improvement since 25% of the tweets were neutral, and 14% were negative.
4. The two papers about ordinance mining (Du et al., 2017; Du et al., 2017) had findings that data mining could discover meaningful patterns from the legislative activities data e. g., some committees focused on ordinances with specific Smart City Characteristics (SCCs), and ordinances on Smart Economy had shorter timespans while ordinances on Smart Environment had more extended timespans.
5. The IEEE UEMON 2017 paper about ordinance mining and database management (Du et al., 2017) had findings that the two sessions focused on different SCCs, and that the Smart People characteristic received the least legislative attention overall since the number of ordinances related to the SCC of Smart People was observed to be the lowest

for both sessions.

6. The DMIN 2017 paper about ordinance mining and applying CSK for efficient ordinance categorization (Du et al., 2017) had findings that the overall number of ordinances increased from 287 in session 2006-2009 to 358 in session 2010-2013; the average timespan of ordinances increased from 204 to 222 days; the first year of both sessions had the highest number of initialized ordinances in each session while the last year had the highest number of enacted ordinances.
7. The WWW 2018 paper about applying CSK on mapping ordinances and tweets (Puri et al., 2018) had findings that only 35% of tweets could be categorized to different SCCs (which shows that not all the tweets published by users relate to any SCCs); among the ones that did relate, the SCC mapping method achieved an accuracy of around 80% as confirmed by domain experts.
8. The IEEE ICTAI 2018 paper on mapping ordinances and tweets (Puri et al., 2018) had findings that there were 48% positive tweets and 27% negative tweets based on the collected samples, which showed the overall public opinion was positive. Moreover, Smart Living had the highest positive percentage (52%) for each specific SCC while Smart Environment had the lowest positive percentage (33%). In this paper we improved the mapping method such that it allowed the SCC mapping tool to assign multiple SCCs to ordinances and tweets (instead of a single SCC) leading to equal or

higher mapping accuracy.

9. The I3E 2020 paper about Android App design with the SCC mapping function (Varghese et al., 2020) had findings that it was the first app about ordinance-tweet mining with respect to Smart City development; it helped users acquire useful Smart City knowledge conveniently at-a-glance with ubiquitous access; and the positive user feedback on the app proved that it could increase public awareness of urban policy hence broadly contributing to Smart City development.
10. The IEEE Big Data 2020 poster paper about a web portal prototype (Du et al., 2020) with the SCC mapping had findings that the approach in our web portal could provide real-time SCC mapping results with integrated SCC mapping functions; the portal was capable of updating with new data while maintaining accuracy as long as there were research improvements; and that the portal is a good step towards real-world applications of our research with contributions to Smart City development.
11. The recently submitted journal paper titled “Prediction Tool on Fine Particle Pollutants and Air Quality for Environmental Engineering” (Du et al.) has findings that discover valuable relationships between urban traffic and PM2.5 data, that e.g. traffic volume per se is not directly proportional to PM2.5 emissions; "region" attributes have significant effects on PM2.5 concentrations; high gasoline and diesel consumption does not always cause unsafe PM2.5 ranges; and economic conditions highly influence the presence of

PM2.5 in air. Some of these are supplementary to our findings from conference papers on this topic.

12. In general, a very important finding of this dissertation is that ordinance-tweet mining is pioneering work conducted by our team of researchers, and it leads to essential contributions from Smart City perspectives. This dissertation establishes the claim that this method of mining public reactions on ordinances or local laws is a significant aspect of urban policy research. Overall, this dissertation has tremendous scope for further research and development, such as conducting the historical analysis of ordinances and tweets, addressing subtle nuances in the language within them, and pursuing potential social media mining in multilingual contexts as well. Some future issues are listed in this dissertation.

7.2 Future Work

Method Improvement and Extended Data Coverage: This research conducted SCC mapping and decision support tools development. We counted the appearances of terms that belonged to domain KBs in ordinances or tweets to assign SCC scores. The current method defines every term with the same score. However, we know that some terms should be more significant. We need to design a weighting factor system to assign different SCC scores to the terms in the domain KBs. This improvement will enhance the accuracy of our mapping and

contribution to Smart City development.

The data we used are primarily NYC data and only two sessions of ordinances. It will improve the impact of environmental management if we apply the same technique to other areas. The air quality tool uses PM2.5 as the air quality indicator. For the air quality tool, we proposed to include other common air pollution indicators, e.g., PM10, for a more comprehensive prediction and decision support, which benefits the Smart City development.

Before and After Analysis of Tweets Sentiments: We utilized social media mining and CSK to identify the related tweets. The sentiment changes of related tweets, primarily those published before and after the enactment date of the ordinance, could be considered highly associated with that ordinance. We proposed to design a website or software that automatically identifies the ordinance and output sentiment changes of the related tweets. Ordinances, the powerful tools of urban governments, can be evaluated accurately and conveniently. The efficiency of urban management will be enhanced.

The foundation of our approach is that we believe that Twitter users express their feelings about different aspects of their daily lives by posting tweets. The ordinances are also related to the daily lives of urban residents. By building connections with tweets and ordinances, we can connect urban management with urban residents. We have designed the concept program, as Figure 7.1 and Figure 7.2 portray herewith.

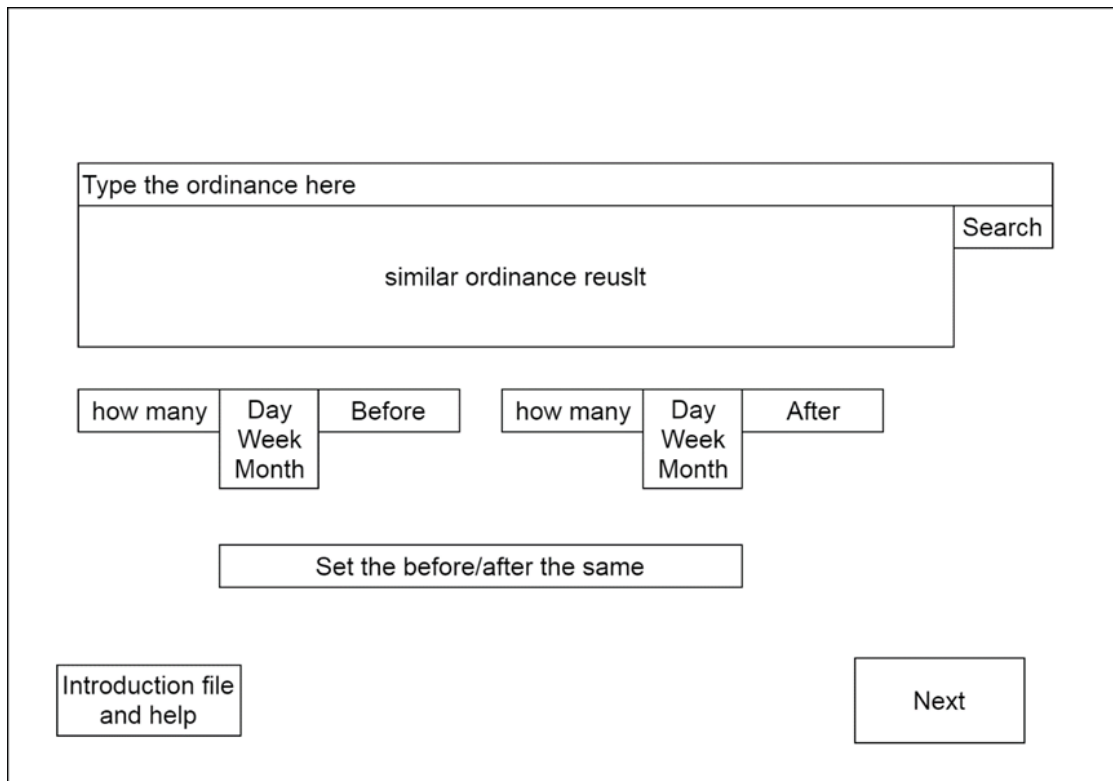


Figure 7.1: Input Interface Concept

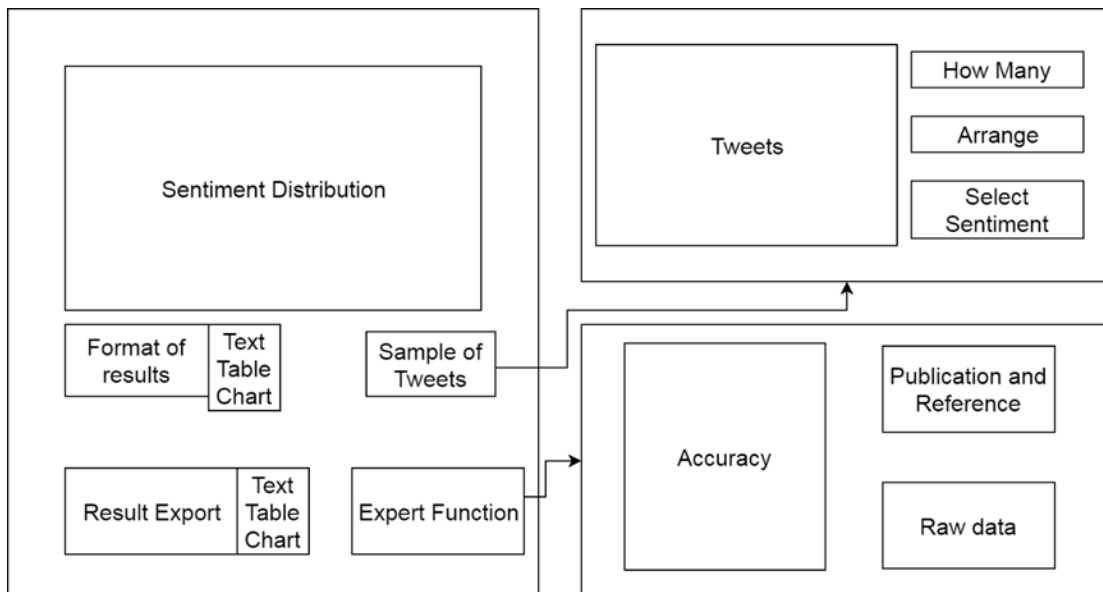


Figure 7.2: Output Interface Concept

We plan to ask the Ph.D. students of Earth and Environmental Studies to test this program as the first step. We will test the tool with an urban management agency after this test. We plan to let the users operate the program for 15 minutes each and give their opinions based on the user experience. An informal survey will help in the design of the metrics. There will be two groups of questions: 1. Easy to use. 2. How useful it is. Each group will have 2-5 questions with a score of 1-5. After finishing the test, we will analyze the overall score of the tool and improve it.

Enhance and Combine the Decision Support Tools: The Internet of Things (IoT) connects and exchanges data between different devices, systems, and platforms via the Internet. This means cross-platform information gathering and sharing. We have designed decision support tools with various data sources and platforms (Desktop system, Android system, Website). We could combine all the functions from our multiple tools and provide access to multiple platforms. It will contribute even more to Environmental Management and Smart City Development. To fully achieve IoT of decision support tools, we also need an automated information gathering agent, such as the web crawler to collect data since ordinances and tweets update constantly. The concept of a proposed enhancement is shown in Figure 7.3 here.

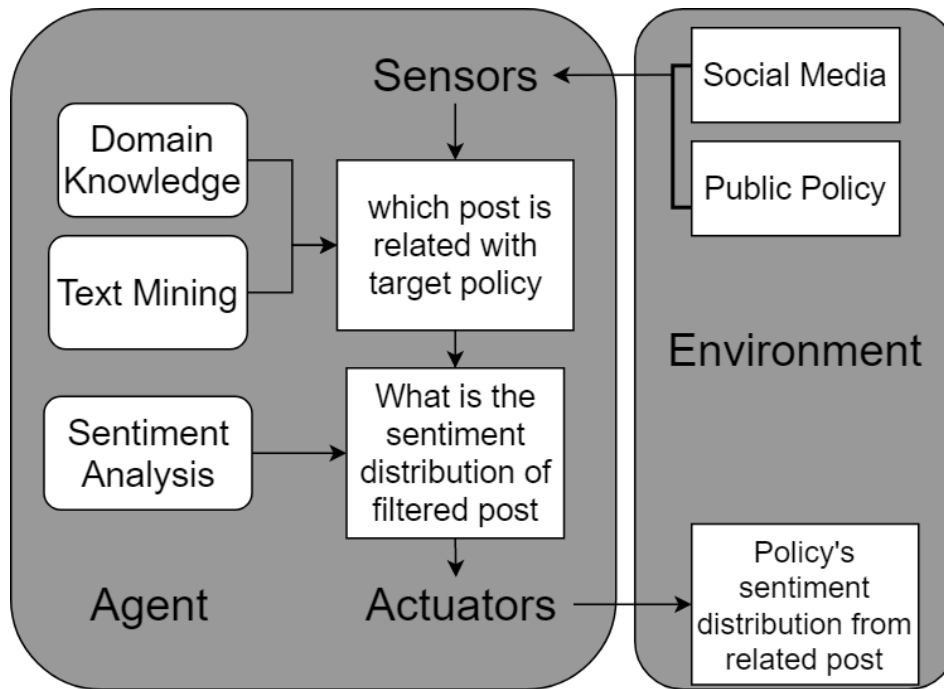


Figure 7.3: Proposed Enhancement

With all these potential open avenues presenting the scope for further research, considerable future work emerges from this dissertation. This can entail areas such as decision support, urban policy, data mining, and Smart Cities.

7.3 References

- Du, X., & Varde, A. (2015), Mining Multicity Urban Data for Sustainable Population Relocation, *International Journal on Computer, Electrical, Automation, Control and Information Engineering*, 9(12), 2441-2448. <https://doi.org/doi.org/10.5281/zenodo.1110816>.
- Du, X., & Varde, A. (2016), Mining PM2.5 and traffic conditions for air quality, *7th International Conference on Information and Communication Systems (ICICS)*, <https://doi.org/10.1109/iacs.2016.7476082>.
- Du, X., Emebo, O., Varde, A., Tandon, N., Chowdhury, S., & Weikum, G. (2016), Air quality assessment from social media and structured data: Pollutants and health impacts in urban planning, *IEEE 32nd International Conference On Data Engineering - Workshops (ICDE-Workshops)*. <https://doi.org/10.1109/icdew.2016.7495616>.
- Du, X., Kowalski, M., & Varde, A. (2020), LSOMP: Large Scale Ordinance Mining Portal. In *IEEE International Conference on Big Data (IEEE BigData 2020)*, Atlanta, GA,
- Du, X., Kowalski, M., Varde, A., de Melo, G., & Taylor, R. (2020), Public opinion matters. *ACM SIGWEB Newsletter*, (Autumn 2019), 1-15, <https://doi.org/10.1145/3352683.3352688>.
- Du, X., Liporace, D., & Varde, A. (2017), Urban legislation assessment by data analytics with smart city characteristics, *IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, <https://doi.org/10.1109/uemcon.2017.8248972>.

-
- Du, X., Pandey, A., & Varde, A. (2020), *Prediction Tool on Fine Particle Pollutants and Air Quality for Environmental Engineering*, Manuscript submitted for publication.
- Du, X., Varde, A., & Taylor, R. (2017), Mining Ordinance Data From the Web for Smart City Development, In *CSREA Press, International Conference on Data Mining DMIN* (pp. 84-90). Las Vegas, NV.
- Puri, M., Du, X., Varde, A., & de Melo, G. (2018), Mapping Ordinances and Tweets using Smart City Characteristics to Aid Opinion Mining, *Companion Volume of The Web Conference 2018 - WWW '18*. <https://doi.org/10.1145/3184558.3191632>.
- Puri, M., Varde, A., Du, X., & de Melo, G. (2018), Smart Governance Through Opinion Mining of Public Reactions on Ordinances, *IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. <https://doi.org/10.1109/ictai.2018.00131>.
- Varghese, C., Varde, A., & Du, X. (2020), An Ordinance-Tweet Mining App to Disseminate Urban Policy Knowledge for Smart Governance, *Lecture Notes in Computer Science, Conference on e-Business, e-Services and e-Society, I3E 2020, for Responsible Design, Implementation and Use of Information and Communication Technology*, Vol. 12067, 389-401. https://doi.org/10.1007/978-3-030-45002-1_34.

Publications from Dissertation

Peer-reviewed Journal Articles:

1. Xu Du, Matthew Kowalski, Aparna Varde, Gerard de Melo and Robert Taylor --- Public Opinion Matters: Mining Social Media Text for Environmental Management --- ACM SIGWEB Journal, Autumn 2019, ISSN: 1931-1745, Article No. 5, pp. 1-15, doi.org/10.1145/3352683.3352688.
2. Xu Du, Aparna Varde --- Mining Multicity Urban Data for Sustainable Population Relocation --- In International Journal on Computer, Electrical, Automation, Control and Information Engineering, 2015, Volume 9, Issue 12, pp. 2441-2448.
3. Xu Du, Abidha Pandey and Aparna Varde --- Prediction Tool on Fine Particle Pollutants and Air Quality for Environmental Engineering --- In Journal Submission.

Peer-reviewed Conference Papers:

1. Xu Du, Matthew Kowalski, Aparna Varde and Boxiang Dong --- LSOMP: Large Scale Ordinance Mining Portal --- In 2020 IEEE International Conference on Big Data (accepted for publication, to appear).

-
2. Christina Varghese, Aparna Varde and Xu Du --- An Ordinance-Tweet Mining App to Disseminate Urban Policy Knowledge for Smart Governance --- Springer LNCS Conference on e-Business, e-Services and e-Society for Responsible Design, Implementation and Use of Information and Communication Technology, I3E, Apr 2020, Vol. 12067, pp. 389-401.
 3. Manish Puri, Aparna Varde, Xu Du and Gerard de Melo --- Smart Governance through Opinion Mining of Public Reactions on Ordinances --- IEEE International Conference on Tools with Artificial Intelligence (IEEE ICTAI 2018), Volos, Greece, Nov 2018, pp. 838 - 845.
 4. Ketaki Gandhe, Aparna Varde and Xu Du --- Sentiment Analysis of Twitter Data with Hybrid Learning for Recommender Applications --- IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (IEEE UEMCON 2018), Nov 2018, DOI: 10.1109/UEMCON.2018.879666.1.
 5. Manish Puri, Xu Du, Aparna Varde and Gerard de Melo --- Mapping Ordinances and Tweets using Smart City Characteristics to Aid Opinion Mining --- In W3C's WWW

(The Web Conference), Companion Vol., Apr 2018, Lyon, France, pp. 1721 - 1728.

6. Xu Du, Diane Liporace and Aparna Varde --- Urban Legislation Assessment by Data Analytics with Smart City Characteristics --- In IEEE Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON, Oct 2017, pp. 20-25.
7. Xu Du, Aparna Varde and Robert Taylor --- Mining Ordinance Data From the Web for Smart City Development --- In CSREA Press, International Conference on Data Mining DMIN, Jul 2017, pp. 84-90, ISBN: 1-60132-453-7.
8. Xu Du, Onyeka Emebo, Aparna Varde, Niket Tandon, Sreyasi Nag Chowdhury and Gerhard Weikum --- Air Quality Assessment from Social Media and Structured Data --- In IEEE International Conference on Data Engineering, ICDE, HDMM Workshop, May 2016, pp. 54 - 59.
9. Xu Du, Aparna Varde --- Mining PM2.5 and Traffic Conditions for Air Quality --- In IEEE International Conference on Information and Communication Systems, Apr 2016, pp. 33 - 38.

References List

7 Celebrities Who Inspired New Laws. Business Insider. (2012). Retrieved from

<https://www.businessinsider.com/7-laws-named-for-celebrities-2012-9#tim-tebow-2>.

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2), 207-216.

<https://doi.org/10.1145/170036.170072>

Android Developers: The Android Studio. developer.android.com. (2021). Retrieved from

<https://developer.android.com/studio>.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta; European Language

Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf.

Background | Smart City Consortium (SCC). Smartcity.org.hk. (2021). Retrieved from

<https://smartcity.org.hk/en/about-background.php>.

Basavaraju, P., & Varde, A. (2017). Supervised Learning Techniques in Mobile Device Apps for Androids. *ACM SIGKDD Explorations Newsletter*, 18(2), 18-29.

<https://doi.org/10.1145/3068777.3068782>

Böhm, C., De Melo, G., Naumann, F., & Weikum, G. (2012). LINDA: Distributed Web-of-Data-Scale Entity Matching. *Proceedings Of The 21St ACM International Conference On Information And Knowledge Management - CIKM '12*.

<https://doi.org/10.1145/2396761.2398582>

Bureau, U. (2021). *2010 Urban Area FAQs*. The United States Census Bureau. Retrieved from <https://www.census.gov/programs-surveys/geography/about/faq/2010-urban-area-faq.html>.

Cao, Z., Wang, L., & De Melo, G. (2018). Link Prediction via Subgraph Embedding-Based Convex Matrix Completion. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018* (pp. 2803-2810). New Orleans, United States; AAAI press.

Cheng, Y., Agrawal, A., Liu, H., & Choudhary, A. (2015). Legislative prediction with dual uncertainty minimization from heterogeneous information. In *EventSIAM International Conference on Data Mining 2015, SDM 2015* (pp. 361-369). Vancouver, Canada; Society for Industrial and Applied Mathematics Publications.

<https://asu.pure.elsevier.com/en/publications/legislative-prediction-with-dual-uncertainty-minimization-from-he>.

Clinton, J., Jackman, S., & Rivers, D. (2004). The Statistical Analysis of Roll Call Data. *American Political Science Review*, *98*(2), 355-370.

<https://doi.org/10.1017/s0003055404001194>

Cohen, B. (2020). *Transportation Systems Management and Operations in Smart Connected*

-
- Communities - Chapter 1. Transportation Systems Management and Operations (TSMO) in Smart Connected Communities - FHWA Office of Operations*. Ops.fhwa.dot.gov. Retrieved 10 October 2020, from <https://ops.fhwa.dot.gov/publications/fhwahop19004/ch1.htm>.
- Dahal, B., Kumar, S., & Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis And Mining*, 9, 1-20.
- Das, S., & Chen, M. (2001). Yahoo! for Amazon: Sentiment Parsing from Small Talk on the Web. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.276189>
- de Jong, M., Joss, S., Schraven, D., Zhan, C., & Weijnen, M. (2015). Sustainable–smart–resilient–low carbon–eco–knowledge cities; making sense of a multitude of concepts promoting sustainable urbanization. *Journal Of Cleaner Production*, 109, 25-38.
<https://doi.org/10.1016/j.jclepro.2015.02.004>
- De Melo, G. (2008). Language as a Foundation of the Semantic Web. *Proceedings Of The Poster And Demonstration Session At The 7Th International Semantic Web Conference (ISWC 2008) (CEUR WS)*, 401.
- De Melo, G. (2013). Not Quite the Same: Identity Constraints for the Web of Linked Data. *Proceedings Of The 27Th AAAI Conference On Artificial Intelligence (AAAI 2013)*, 1092–1098.
- De Melo, G. (2017). Inducing Conceptual Embedding Spaces from Wikipedia. *Proceedings Of The 26Th International Conference On World Wide Web Companion - WWW '17 Companion*, 43-50. <https://doi.org/10.1145/3041021.3054144>

-
- De Melo, G. (2017). Multilingual Vector Representations of Words, Sentences, and Documents. *Proceedings Of IJCNLP 2017*, 3-5. <https://www.aclweb.org/anthology/I17-5002.pdf>.
- De Melo, G., & Weikum, G. (2008). Mapping Roget's Thesaurus and WordNet to French. *Proceedings Of The 6Th Language Resources And Evaluation Conference (LREC 2008)*, 3306–3313.
- De Melo, G., & Weikum, G. (2010). Untangling the Cross-Lingual Link Structure of Wikipedia. *Proceedings Of The 48Th Annual Meeting Of The Association For Computational Linguistics*.
- De Melo, G., Kacimi, M., & Varde, A. (2016). Dissertation Research Problems in Data Management and Related Areas. *ACM SIGMOD Record*, 44(4), 53-56.
<https://doi.org/10.1145/2935694.2935707>
- Diaz, F., Gamon, M., Hofman, J., Kıcıman, E., & Rothschild, D. (2016). Online and Social Media Data As an Imperfect Continuous Panel Survey. *PLOS ONE*, 11(1), e0145406.
<https://doi.org/10.1371/journal.pone.0145406>
- Difference between Law and Ordinance | Law vs Ordinance*. Differencebetween.info. (2021). Retrieved from <http://www.differencebetween.info/difference-between-law-and-ordinance>.
- Dirican, C. (2015). The Impacts of Robotics, Artificial Intelligence On Business and Economics. *Procedia - Social And Behavioral Sciences*, 195, 564-573.
<https://doi.org/10.1016/j.sbspro.2015.06.134>
- Dong, X., & De Melo, G. (2018). Cross-Lingual Propagation for Deep Sentiment Analysis.

Proceedings Of The 32Nd AAAI Conference On Artificial Intelligence (AAAI 2018).

Dong, X., & De Melo, G. (2019). A Robust Self-Learning Framework for Cross-Lingual Text Classification. *Proceedings Of The 2019 Conference On Empirical Methods In Natural Language Processing And The 9Th International Joint Conference On Natural Language Processing (EMNLP-IJCNLP)*, 6306–6310. <https://doi.org/10.18653/v1/d19-1658>

Driggs-Campbell, K., Shia, V., & Bajcsy, R. (2014). Decisions for autonomous vehicles. *Proceedings Of The 3Rd International Conference On High Confidence Networked Systems*. <https://doi.org/10.1145/2566468.2576850>

Du, X., & Varde, A. (2015). Mining Multicity Urban Data for Sustainable Population Relocation. *International Journal On Computer, Electrical, Automation, Control And Information Engineering*, 9(12), 2441-2448. <https://doi.org/doi.org/10.5281/zenodo.1110816>

Du, X., & Varde, A. (2016). Mining PM2.5 and traffic conditions for air quality. *2016 7Th International Conference On Information And Communication Systems (ICICS)*. <https://doi.org/10.1109/iacs.2016.7476082>

Du, X., Emebo, O., Varde, A., Tandon, N., Chowdhury, S., & Weikum, G. (2016). Air quality assessment from social media and structured data: Pollutants and health impacts in urban planning. *2016 IEEE 32Nd International Conference On Data Engineering Workshops (ICDEW)*. <https://doi.org/10.1109/icdew.2016.7495616>

Du, X., Kowalski, M., & Varde, A. (2020). LSOMP: Large Scale Ordinance Mining Portal. In

-
- IEEE International Conference on Big Data (IEEE BigData 2020)*. Atlanta, GA.
- Du, X., Kowalski, M., Varde, A., De Melo, G., & Taylor, R. (2020). Public opinion matters. *ACM SIGWEB Newsletter*, (Autumn), 1-15. <https://doi.org/10.1145/3352683.3352688>
- Du, X., Liporace, D., & Varde, A. (2017). Urban legislation assessment by data analytics with smart city characteristics. *2017 IEEE 8Th Annual Ubiquitous Computing, Electronics And Mobile Communication Conference (UEMCON)*. <https://doi.org/10.1109/uemcon.2017.8248972>
- Du, X., Varde, A., & Taylor, R. (2017). Mining Ordinance Data From the Web for Smart City Development. In *International Conference on Data Mining DMIN* (pp. 84-90). Las Vegas; CSREA press.
- Emebo, O., Varde, A., & Daramola, O. (2016). Common Sense Knowledge, Ontology and Text Mining for Implicit Requirements. In *International Conference on Data Mining 2016 (DMIN'16)* (pp. 146–152). Las Vegas, Nevada, USA; CSREA Press.
<https://core.ac.uk/download/pdf/189859043.pdf>.
- European Smart Cities*. Smart-cities.eu. (2015). Retrieved from <http://www.smart-cities.eu/>.
- Ewing, R. (2014). *Geographic Information Systems & Science - County Level Urban Sprawl Indices*. Gis.cancer.gov. Retrieved from <https://gis.cancer.gov/tools/urban-sprawl/>.
- Ewing, R., & Hamidi, S. (2014). *Measuring Sprawl 2014*. smartgrowthamerica.org. Retrieved from <https://smartgrowthamerica.org/wp-content/uploads/2016/08/measuring-sprawl-2014.pdf>
- Fellbaum, C., & Miller, G. (1998). *WordNet - An Electronic Lexical Database*. MIT Press.

-
- Forsyth, T. (2014). Public concerns about transboundary haze: A comparison of Indonesia, Singapore, and Malaysia. *Global Environmental Change*, 25, 76-86.
<https://doi.org/10.1016/j.gloenvcha.2014.01.013>
- Fujii, Y., Tohno, S., Amil, N., Latif, M., Oda, M., Matsumoto, J., & Mizohata, A. (2015). Annual variations of carbonaceous PM_{2.5} in Malaysia: Influence by Indonesian peatland fires. *Atmospheric Chemistry And Physics*, 15(23), 13319-13329. <https://doi.org/10.5194/acp-15-13319-2015>
- Gandhe, K., Varde, A., & Du, X. (2018). Sentiment Analysis of Twitter Data with Hybrid Learning for Recommender Applications. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 57-63). New York City, NY, USA. <https://doi.org/10.1109/UEMCON.2018.8796661>.
- Garg, A., Tandon, N., & Varde, A. (2020). I Am Guessing You Can't Recognize This: Generating Adversarial Images for Object Detection Using Spatial Commonsense (Student Abstract). *Proceedings Of The AAAI Conference On Artificial Intelligence*, 34(10), 13789-13790.
<https://doi.org/10.1609/aaai.v34i10.7166>
- Giffinger, R., & Pichler-Milanović, N. (2007). *Smart cities*. Centre of Regional Science, Vienna University of Technology.
- Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Milanović, N., & Meijers, E. (2007). *Smart cities Ranking of European medium-sized cities*. Vienna University of Technology.

-
- Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *Processing*, 1-6.
<http://www.stanford.edu/~alecmgo/papers/TwitterDistantSupervision09.pdf>.
- Gu, Y., Qian, Z., & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, 67, 321-342.
<https://doi.org/10.1016/j.trc.2016.02.011>
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1), 10-18.
<https://doi.org/10.1145/1656274.1656278>
- Hamidi, S., & Ewing, R. (2014). A longitudinal study of changes in urban sprawl between 2000 and 2010 in the United States. *Landscape And Urban Planning*, 128, 72-82.
<https://doi.org/10.1016/j.landurbplan.2014.04.021>
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann Publishers.
- Harper, R., Rodden, T., Rogers, Y., & Sellen, A. (2020). *Being Human: Human-Computer Interaction in the Year 2020*. Microsoft Research.
- Hollander, J., & Renski, H. (2015). *Measuring Urban Attitudes Using Twitter: An Exploratory Study*. Lincoln Institute of Land Policy. Retrieved from
https://www.lincolninst.edu/sites/default/files/pubfiles/3607_2954_Hollander%20WP15JH1.pdf

Home - IEEE Smart Cities. Smartcities.ieee.org. (2014). Retrieved from

<http://smartcities.ieee.org/>.

Hong, L., & Davison, B. (2010). Empirical study of topic modeling in Twitter. *Proceedings Of The First Workshop On Social Media Analytics - SOMA '10*.

<https://doi.org/10.1145/1964858.1964870>

Huang, Q., Cervone, G., & Zhang, G. (2017). A cloud-enabled automatic disaster analysis system of multi-sourced data streams: An example synthesizing social media, remote sensing and Wikipedia data. *Computers, Environment And Urban Systems*, 66, 23-37.

<https://doi.org/10.1016/j.compenvurbsys.2017.06.004>

Inmon, W. (2011). *Building the data warehouse*. Wiley.

Jong, M., Joss, S., Schraven, D., Zhan, C., & Weijnen, M. (2015). Sustainable–smart–resilient–low carbon–eco–knowledge cities; making sense of a multitude of concepts promoting sustainable urbanization. *Journal Of Cleaner Production*, 109, 25-38.

<https://doi.org/10.1016/j.jclepro.2015.02.004>

Kumar, P., Robins, A., Vardoulakis, S., & Britter, R. (2010). A review of the characteristics of nanoparticles in the urban atmosphere and the prospects for developing regulatory controls.

Atmospheric Environment, 44(39), 5035-5052. <https://doi.org/10.1016/j.atmosenv.2010.08.016>

Leetaru, K. (2019). *Is Twitter's Spritzer Stream Really A Nearly Perfect 1% Sample Of Its Firehose?*. Forbes. Retrieved 10 October 2020, from

<https://www.forbes.com/sites/kalevleetaru/2019/02/27/is-twitters-spritzer-stream-really-a-nearly-perfect-1-sample-of-its-firehose/#5aa45ef35401>.

Li, Q., Shah, S., Liu, X., Nourbakhsh, A., & Fang, R. (2016). TweetSift. *Proceedings Of The 25Th ACM International On Conference On Information And Knowledge Management*, 2429–2432. <https://doi.org/10.1145/2983323.2983325>

Li, Q., Shah, S., Liu, X., Nourbakhsh, A., & Fang, R. (2016). TweetSift. *Proceedings Of The 25Th ACM International On Conference On Information And Knowledge Management*, 2429–2432. <https://doi.org/10.1145/2983323.2983325>

Li, X., & Gar-On Yeh, A. (2004). Data mining of cellular automata's transition rules. *International Journal Of Geographical Information Science*, 18(8), 723-744. <https://doi.org/10.1080/13658810410001705325>

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (pp. 281--297). Berkeley, Calif.; University of California Press. <https://projecteuclid.org/euclid.bsmsp/1200512992>.

Max-Planck-Institut für Informatik: WebChild. Mpi-inf.mpg.de. (2018). Retrieved 10 October 2020, from <https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/commonsense/webchild>.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations

-
- of words and phrases and their compositionality. In *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (pp. 3111–3119). Curran Associates Inc.
- Miller, H., & Han, J. (2009). *Geographic data mining and knowledge discovery*. CRC Press.
- Morcol, G. (2012). Urban Sprawl And Public Policy: A Complexity Theory Perspective. *Emergence: Complexity And Organization* *goktug Morcol, 14(4)*, 1-16.
<https://doi.org/10.17357.38641212d93eb1f8554ed20c527c19a2>.
- Nagy, R., & Lockaby, B. (2010). Urbanization in the Southeastern United States: Socioeconomic forces and ecological responses along an urban-rural gradient. *Urban Ecosystems, 14(1)*, 71-86.
<https://doi.org/10.1007/s11252-010-0143-6>
- Nasukawa, T., & Yi, J. (2003). Sentiment analysis. *Proceedings Of The International Conference On Knowledge Capture - K-CAP '03*. <https://doi.org/10.1145/945645.945658>
- Nasukawa, T., & Yi, J. (2003). Sentiment analysis. *Proceedings Of The International Conference On Knowledge Capture - K-CAP '03*, 70-77. <https://doi.org/10.1145/945645.945658>
- New York City (NYC) Population Facts*. www1.nyc.gov. (2017). Retrieved from
<http://www1.nyc.gov/site/planning/data-maps/nyc-population/populationfacts.page>.
- Nuortimo, K. (2018). Measuring public acceptance with opinion mining: The case of the energy industry with long-term coal R&D investment projects. *Journal Of Intelligence Studies In Business, 8(2)*. <https://doi.org/10.37380/jisib.v8i2.319>

-
- Nuortimo, K., & Härkönen, J. (2018). Opinion mining approach to study media-image of energy production. Implications to public acceptance and market deployment. *Renewable And Sustainable Energy Reviews*, 96, 210-217. <https://doi.org/10.1016/j.rser.2018.07.018>
- NYC Smart City Trek. www.eventa.us. (2020). Retrieved from <https://www.eventa.us/events/new-york-ny/nyc-smart-city-trek>.
- O'Connell, C. (2020). *15 Ordinary People Who Changed History*. Reader's Digest. Retrieved from <https://www.rd.com/true-stories/inspiring/inspiring-stories-9-ordinary-people-who-changed-history/>.
- Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010* (pp. 17-23).
- Pampoore-Thampi, A., Varde, A., & Yu, D. (2014). Mining GIS Data to Predict Urban Sprawl. In *ACM KDD (Knowledge Discovery and Data Mining conference) Bloomberg Track* (pp. 118-125). New York, NY, United States; Association for Computing Machinery.
- Pandey, A., Puri, M., & Varde, A. (2018). Object Detection with Neural Models, Deep Learning and Common Sense to Aid Smart Mobility. *2018 IEEE 30th International Conference On Tools With Artificial Intelligence (ICTAI)*, 859–863. <https://doi.org/10.1109/ictai.2018.00134>
- Pang, B., & Lee, L. (2004). A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings Of The 42Nd Annual Meeting On*

Association For Computational Linguistics - ACL '04, 271-278.

<https://doi.org/10.3115/1218955.1218990>

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations And Trends®*

In Information Retrieval, 2(1–2), 1-135. <https://doi.org/10.1561/1500000011>

Pant, P., & Harrison, R. (2013). Estimation of the contribution of road traffic emissions to

particulate matter concentrations from field measurements: A review. *Atmospheric*

Environment, 77, 78-97. <https://doi.org/10.1016/j.atmosenv.2013.04.028>

Pawlish, M., & Varde, A. (2010). Free cooling: A paradigm shift in data centers. *2010 Fifth*

International Conference On Information And Automation For Sustainability, 347–352/347–

352. <https://doi.org/10.1109/iciafs.2010.5715685>

Pawlish, M., Varde, A., & Robila, S. (2015). The Greening of Data Centers with Cloud

Technology. *International Journal Of Cloud Applications And Computing*, 5(4), 1-23.

<https://doi.org/10.4018/ijcac.2015100101>

Pawlish, M., Varde, A., Robila, S., & Ranganathan, A. (2014). A call for energy efficiency in data

centers. *ACM SIGMOD Record*, 43(1), 45-51. <https://doi.org/10.1145/2627692.2627703>

Persaud, P., Varde, A., & Robila, S. (2017). Enhancing Autonomous Vehicles with

Commonsense: Smart Mobility in Smart Cities. *2017 IEEE 29Th International Conference On*

Tools With Artificial Intelligence (ICTAI). <https://doi.org/10.1109/ictai.2017.00155>

phpMyAdmin. phpMyAdmin. (2021). Retrieved from <https://www.phpmyadmin.net/>.

Policy Assessment for the Review of the National Ambient Air Quality Standards for Particulate

Matter. Epa.gov. (2020). Retrieved from https://www.epa.gov/sites/production/files/2020-01/documents/final_policy_assessment_for_the_review_of_the_pm_anaqs_01-2020.pdf.

Preece, J., Sharp, H., & Rogers, Y. (2015). *Interaction design - beyond human-computer interaction* (4th ed.). Wiley.

Publishing, P. (2019). *Artificial Intelligence for Smart Cities*. Medium. Retrieved from <https://becominghuman.ai/artificial-intelligence-for-smart-cities-64e6774808f8>.

Puri, M., Du, X., Varde, A., & De Melo, G. (2018). Mapping Ordinances and Tweets using Smart City Characteristics to Aid Opinion Mining. *Companion Of The The Web Conference 2018 On The Web Conference 2018 - WWW '18*. <https://doi.org/10.1145/3184558.3191632>

Puri, M., Varde, A., & Dong, B. (2018). Pragmatics and Semantics to Connect Specific Local Laws with Public Reactions. *2018 IEEE International Conference On Big Data (Big Data)*, 5433–5435. <https://doi.org/10.1109/bigdata.2018.8622162>

Puri, M., Varde, A., Du, X., & De Melo, G. (2018). Smart Governance Through Opinion Mining of Public Reactions on Ordinances. *2018 IEEE 30Th International Conference On Tools With Artificial Intelligence (ICTAI)*. <https://doi.org/10.1109/ictai.2018.00131>

Quinlan, J. (1993). *C 4.5: programs for machine learning*. Morgan Kaufmann.

Rajasekar, U., & Weng, Q. (2009). Application of Association Rule Mining for Exploring the Relationship between Urban Land Surface Temperature and Biophysical/Social Parameters.

Photogrammetric Engineering & Remote Sensing, 75(4), 385-396.

<https://doi.org/10.14358/pers.75.4.385>

Read, J. (2005). Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In *Proceedings of the ACL Student Research Workshop* (pp. 43–48). Association for Computational Linguistics. <https://www.aclweb.org/anthology/P05-2008.pdf>.

Revised Air Quality Standards for Particle Pollution and Updates to The Air Quality Index (AQI). Epa.gov. (2016). Retrieved from https://www.epa.gov/sites/production/files/2016-04/documents/2012_aqi_factsheet.pdf.

Rexha, A., Kröll, M., Dragoni, M., & Kern, R. (2016). Polarity Classification for Target Phrases in Tweets: A Word2Vec Approach. *The Semantic Web*, 217-223. https://doi.org/10.1007/978-3-319-47602-5_40

Rouces, J., De Melo, G., & Hose, K. (2016). Complex Schema Mapping and Linking Data: Beyond Binary Predicates. *Proceedings Of The Workshop On Linked Data On The Web : Co-Located With 25Th International World Wide Web Conference (WWW 2016)*, 1593.

Rouces, J., De Melo, G., & Hose, K. (2016). Heuristics for Connecting Heterogeneous Knowledge via FrameBase. *The Semantic Web. Latest Advances And New Domains*, 20-35. https://doi.org/10.1007/978-3-319-34129-3_2

Roues, J., De Melo, G., & Hose, K. (2015). Representing Specialized Events with FrameBase. *Representing Specialized Events with FrameBase. Proceedings Of The 4Th*

-
- International Workshop On Detection, Representation, And Exploitation Of Events In The Semantic Web (Derive 2015) Co-Located With The 12Th Extended Semantic Web Conference (ESWC 2015)*, 1363, 58-69. http://ceur-ws.org/Vol-1363/paper_7.pdf.
- Rozeva, A., & Zerkova, S. (2017). Assessing semantic similarity of texts – Methods and algorithms. *AIP Conference Proceedings*, 1910(1), 1-8. <https://doi.org/10.1063/1.5014006>
- Russell, S., Norvig, P., & Canny, J. (2003). *Artificial intelligence* (2nd ed.). Prentice Hall.
- Sachdeva, S., McCaffrey, S., & Locke, D. (2016). Social media approaches to modeling wildfire smoke dispersion: spatiotemporal and social scientific investigations. *Information, Communication & Society*, 20(8), 1146-1161. <https://doi.org/10.1080/1369118x.2016.1218528>
- Santé, I., García, A., Miranda, D., & Crecente, R. (2010). Cellular automata models for the simulation of real-world urban processes: A review and analysis. *Landscape And Urban Planning*, 96(2), 108-122. <https://doi.org/10.1016/j.landurbplan.2010.03.001>
- Sayce, D. (2020). *The Number of tweets per day in 2020* | David Sayce. David Sayce. Retrieved 18 October 2020, from <https://www.dsayce.com/social-media/tweets-day/#:~:text=Every%20second%2C%20on%20average%2C%20around%206%2C000%20tweets%20are%20tweeted%20on,200%20billion%20tweets%20per%20year>.
- Schneider, A., & Woodcock, C. (2008). Compact, Dispersed, Fragmented, Extensive? A Comparison of Urban Growth in Twenty-five Global Cities using Remotely Sensed Data, Pattern Metrics and Census Information. *Urban Studies*, 45(3), 659-692.

<https://doi.org/10.1177/0042098007087340>

Scott, J. (2012). *Archive Team: The Twitter Stream Grab*. Archive.org. Retrieved 10 October 2020, from <https://archive.org/details/twitterstream?tab=about>.

Smart Cities Council | Definitions and overviews. Smartcitiescouncil.com. (2015). Retrieved from <http://smartcitiescouncil.com/smart-cities-information-center/definitions-and-overviews>.

So, Y. (2017). *Designing for Mobile Apps: Overall Principles, Common Patterns, and Interface Guidelines*. Medium. Retrieved from <https://medium.com/intuit-engineering/native-mobile-app-design-overall-principles-and-common-patterns-26edee8ced10>.

Soni, S. (2019). *A Mobile App Designer's Guide on Google Material Design*. Appinventiv. Retrieved from <https://appinventiv.com/blog/mobile-app-designers-guide-on-material-design/>.

Tandon, N., & De Melo, G. (2010). Information Extraction from WebScale N-Gram Data. *Web N-Gram Workshop. Workshop Of The 33Rd Annual International ACM SIGIR Conference On Research And Development In Information Retrieval*, 5803, 8-5.

Tandon, N., De Melo, G., & Weikum, G. (2011). Deriving a Web-Scale Common Sense Fact Database. *Proceedings Of The 25Th AAI Conference On Artificial Intelligence (AAAI 2011)*, 152–157.

Tandon, N., De Melo, G., & Weikum, G. (2014). Acquiring comparative commonsense knowledge from the web. In *AAAI'14: Proceedings of the Twenty-Eighth AAI Conference on Artificial Intelligence* (pp. 166–172). AAAI Press.

-
- Tandon, N., De Melo, G., & Weikum, G. (2017). WebChild 2.0 : Fine-Grained Commonsense Knowledge Distillation. In *Proceedings of ACL 2017, System Demonstrations* (pp. 115–120). Association for Computational Linguistics. <https://www.aclweb.org/anthology/P17-4020.pdf>.
- Tandon, N., De Melo, G., Suchanek, F., & Weikum, G. (2014). WebChild: harvesting and organizing commonsense knowledge from the Web. *Proceedings Of The 7Th ACM International Conference On Web Search And Data Mining*, 523-532. <https://doi.org/10.1145/2556195.2556245>
- Tandon, N., Varde, A., & De Melo, G. (2018). Commonsense Knowledge in Machine Intelligence. *ACM SIGMOD Record*, 46(4), 49-52. <https://doi.org/10.1145/3186549.3186562>
- Taylor, R. (2012). Urbanization, Local Government, and Planning for Sustainability. *Sustainability Science*, 293-313. https://doi.org/10.1007/978-1-4614-3188-6_14
- Taylor, R., Carandang, J., Alexander, C., & Calleja, J. (2012). Making Global Cities Sustainable: Urban Rooftop Hydroponics for Diversified Agriculture in Emerging Economies. *OIDA International Journal Of Sustainable Development*, 5(7), 11-28. <https://ssrn.com/abstract=2192203>.
- The 2017 smart cities index*. Easyparkgroup.com. (2017). Retrieved 1 January 2017, from <https://www.easyparkgroup.com/smart-cities-index/>.
- The New York City Council - Committees*. Legistar.council.nyc.gov. (2017). Retrieved from <http://legistar.council.nyc.gov/Departments.aspx>.

The New York City Council - Legislation. Legistar.council.nyc.gov. (2020). Retrieved 10 October 2020, from <https://legistar.council.nyc.gov/Legislation.aspx>.

The World Bank, Data By Country. (2015). Retrieved 26 April 2015, from <http://data.worldbank.gov>.

Top 10 Features of Android Studio for Developers Not to Miss. ADMEC Multimedia. (2018). Retrieved from <https://www.admecindia.co.in/blog/top-10-features-android-studio-developers-not-miss>.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Proceedings Of The International AAAI Conference On Web And Social Media*, 4(1), 178-185.

Varde, A., & Du, X. (2015). *Multicity Simulation with Data Mining for Urban Sustainability*. Presentation, Bloomberg Data Science Labs, New York, NY.

Varde, A., Tandon, N., Chowdhury, S., & Weikum, G. (2015). *Commonsense Knowledge in Domain-Specific Knowledge Bases*. Saarbruecken, Germany: Max Planck Institute for Informatics (MPII).

Varghese, A., Varde, A., Peng, J., & Fitzpatrick, E. (2015). A Framework for Collocation Error Correction in Web Pages and Text Documents. *ACM SIGKDD Explorations Newsletter*, 17(1), 14-23. <https://doi.org/10.1145/2830544.2830548>

Varghese, C. (2019). *Disseminating Results of Mining Ordinances and their Tweets by Android*

-
- App Development*. Montclair, NJ, USA. Retrieved from <http://cs.montclair.edu/research/ProjectChristinaVarghese19Dec.pdf>
- Varghese, C., Varde, A., & Du, X. (2020). An Ordinance-Tweet Mining App to Disseminate Urban Policy Knowledge for Smart Governance. *Lecture Notes In Computer Science*, 389-401. https://doi.org/10.1007/978-3-030-45002-1_34
- Walden, M. (2005). *Smart Economics: Commonsense Answers to 50 Questions about Government, Taxes, Business, and Households*. Praeger.
- Wang, J., Varshney, K., & Mojsilović, A. (2012). Legislative Prediction via Random Walks over a Heterogeneous Graph. *Proceedings Of The 2012 SIAM International Conference On Data Mining*. <https://doi.org/10.1137/1.9781611972825.94>
- Wang, L., Liu, K., Cao, Z., Zhao, J., & De Melo, G. (2015). Sentiment-Aspect Extraction based on Restricted Boltzmann Machines. *Proceedings Of The 53Rd Annual Meeting Of The Association For Computational Linguistics And The 7Th International Joint Conference On Natural Language Processing (Volume 1: Long Papers)*. <https://doi.org/10.3115/v1/p15-1060>
- Wang, L., Wang, Y., Liu, B., He, L., Liu, S., Melo, G., & Xu, Z. (2017). Link prediction by exploiting network formation games in exchangeable graphs. *2017 International Joint Conference On Neural Networks (IJCNN)*. <https://doi.org/10.1109/ijcnn.2017.7965910>
- Wang, R., Mao, H., Wang, Y., Rae, C., & Shaw, W. (2018). Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. *Computers & Geosciences*, *111*, 139-147.

<https://doi.org/10.1016/j.cageo.2017.11.008>

Wang, S., Paul, M., & Dredze, M. (2015). Social Media as a Sensor of Air Quality and Public Response in China. *Journal Of Medical Internet Research*, 17(3), e22.

<https://doi.org/10.2196/jmir.3875>

Wang, S., Zhang, X., Cao, J., He, L., Stenneth, L., & Yu, P. et al. (2017). Computing Urban Traffic Congestions by Incorporating Sparse GPS Probe Data and Social Media Data. *ACM Transactions On Information Systems*, 35(4), 1-30. <https://doi.org/10.1145/3057281>

What is smart growth? | Smart Growth America. Smart Growth America. (2021). Retrieved from <http://www.smartgrowthamerica.org/what-is-smart-growth>.

Whitelaw, C., Garg, N., & Argamon, S. (2005). Using appraisal groups for sentiment analysis. *Proceedings Of The 14Th ACM International Conference On Information And Knowledge Management - CIKM '05*, 625–631. <https://doi.org/10.1145/1099554.1099714>

Witten, I., Frank, E., & Hall, M. (2011). *Data Mining : Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann Publishers.

World Health Organization, Data Repository,. Who.int. (2015). Retrieved 26 April 2015, from <http://www.who.int/gho/en/>.

Wu, Y., Xie, L., Huang, S., Li, P., Yuan, Z., & Liu, W. (2018). Using social media to strengthen public awareness of wildlife conservation. *Ocean & Coastal Management*, 153, 76-83.

<https://doi.org/10.1016/j.ocecoaman.2017.12.010>

XML Editor: XMLSpy. Altova.com. (2021). Retrieved from <https://www.altova.com/xmlspy.html>.

Yang, C., Lin, K., & Chen, H. (2007). Emotion Classification Using Web Blog Corpora.

IEEE/WIC/ACM International Conference On Web Intelligence (WI'07), 275-278.

<https://doi.org/10.1109/wi.2007.51>

Yu, Y., Lou, Q., Tang, J., Wang, J., & Yue, X. (2017). An exact decomposition method to save trips in cooperative pickup and delivery based on scheduled trips and profit distribution.

Computers & Operations Research, 87, 245-257. <https://doi.org/10.1016/j.cor.2017.02.015>

Zadeh, L., Abbasov, A., & Shahbazova, S. (2015). Analysis of Twitter hashtags: Fuzzy clustering approach. *2015 Annual Conference Of The North American Fuzzy Information Processing Society (NAFIPS) Held Jointly With 2015 5Th World Conference On Soft Computing (Wconsc)*.

<https://doi.org/10.1109/nafips-wconsc.2015.7284196>

Zauli Sajani, S., Ricciardelli, I., Trentini, A., Bacco, D., Maccone, C., & Castellazzi, S. et al.

(2015). Spatial and indoor/outdoor gradients in urban concentrations of ultrafine particles and PM 2.5 mass and chemical components. *Atmospheric Environment*, 103, 307-320.

<https://doi.org/10.1016/j.atmosenv.2014.12.064>

Zhou, J., Chen, A., Cao, Q., Yang, B., Chang, V., & Nazaroff, W. (2015). Particle exposure during the 2013 haze in Singapore: Importance of the built environment. *Building And Environment*, 93, 14-23. <https://doi.org/10.1016/j.buildenv.2015.04.029>

Zou, L., Lam, N., Cai, H., & Qiang, Y. (2018). Mining Twitter Data for Improved Understanding

of Disaster Resilience. *Annals Of The American Association Of Geographers*, 108(5), 1422-1441. <https://doi.org/10.1080/24694452.2017.1421897>