5-2019

Commonsense Knowledge in Sentiment Analysis of Ordinance Reactions for Smart Governance

Manish Puri

Montclair State University

Commonsense Knowledge in Sentiment Analysis of Ordinance Reactions for Smart Governance

Abstract

Smart Governance is an emerging research area which has attracted scientific as well as policy interests, and aims to improve collaboration between government and citizens, as well as other stakeholders. Our project aims to enable lawmakers to incorporate data driven decision making in enacting ordinances. Our first objective is to create a mechanism for mapping ordinances (local laws) and tweets to Smart City Characteristics (SCC). The use of SCC has allowed us to create a mapping between a huge number of ordinances and tweets, and the use of Commonsense Knowledge (CSK) has allowed us to utilize human judgment in mapping.

We have then enhanced the mapping technique to link multiple tweets to SCC. In order to promote transparency in government through increased public participation, we have conducted sentiment analysis of tweets in order to evaluate the opinion of the public with respect to ordinances passed in a particular region.

Our final objective is to develop a mapping algorithm in order to directly relate ordinances to tweets. In order to fulfill this objective, we have developed a mapping technique known as TOLCS (Tweets Ordinance Linkage by Commonsense and Semantics). This technique uses pragmatic aspects in Commonsense Knowledge as well as semantic aspects by domain knowledge. By reducing the sample space of big data to be processed, this method represents an efficient way to accomplish this task.

The ultimate goal of the project is to see how closely a given region is heading towards the concept of Smart City.

Keywords: *Artificial Intelligence, Big Data, Commonsense Knowledge, Complexity, Data Mining, Machine Learning, Natural Language Processing, Ordinances, Pragmatics, Semantics, Sentiment Analysis, Smart Cities, Social Media, Text Mining, Twitter, Urban Policy*

MONTCLAIR STATE UNIVERSITY

Commonsense Knowledge in Sentiment Analysis of Ordinance Reactions for Smart Governance

by

Manish Puri

A Master's Thesis Submitted to the Faculty of

Montclair State University

In Partial Fulfillment of the Requirements

For the Degree of

Master of Science

May 2019

School: College of Science and Mathematics

Department: Computer Science

Thesis Committee:

[redacted]

Dr. Aparna Varde
Thesis Sponsor

[redacted]

Dr. Anna Feldman
Committee Member

[redacted]

Dr. Boxiang Dong
Committee Member

Commonsense Knowledge in Sentiment Analysis of Ordinance Reactions for Smart Governance

A THESIS

Submitted in partial fulfillment of the requirements for the degree of Master of Science

By

Manish Puri

Montclair State University

May 2019

Acknowledgements

I would like to thank Dr. Aparna Varde for her guidance while writing this thesis and my research endeavors during my graduate studies at Montclair State University. I am especially grateful for her patience during challenging times and sharing her knowledge and experience with me. Working with Dr. Varde on this project has helped me increase my technical and research skills, and my work would not have been possible without her mentoring.

I would like to thank Dr. Boxiang Dong for his inputs, especially in the final phase of my research work. He has been instrumental in helping me refine the content in my papers and giving me direction when needed. I am grateful for the initiative and support he has shown in helping me complete my project.

I would like to thank Dr. Anna Feldman for her inputs in this research work. My interest in the area of Natural Language Processing began after I took her class in Computational Linguistics and she has been very generous with her time in reading my research papers and giving me feedback on my work.

I would like to thank Dr. Niket Tandon from the Allen Institute of Artificial Intelligence (Seattle, WA), Dr. Gerard de Melo from Rutgers University (New Brunswick, NJ) and Xu Du from the Environmental Science and Management PhD program at Montclair State University for their inputs during different phases of my project.

Finally, I would like to thank my family for being very supportive of my endeavors during graduate studies.

Table of Contents

1. Introduction
   1.1 Background and Motivation……………………………………..…………..… 7
   1.2 Problem Definition………………………………………………….……... 8
   1.3 Layout and Organization …………………………………………………… 9

2. Proposed Methodology:
   2.1 General Description ………………………………………………………... 11
   2.2 Commonsense Knowledge ...……………………………………………… 14
   2.3 Smart City Characteristics……………………………………………..…… 20
   2.4 Methodology………………………………………………………………..… 22
   2.5 Mapping Tweets and Ordinances to SCCs……………………………………. 24
   2.6 The TOLCS approach ……………………………………………………... 27
   2.7 Sentiment Analysis………………………………………………………………. 28

3. Implementation details:
   3.1 Smart City Classifier Tool ………………….……………………………... 31
   3.2 TOLCS Approach Description…………………………...…………………. 32
   3.3 TOLCS Approach Implementation …………………………………………… 34

4. Experimental Evaluation
   4.1 Experiments and Observations on Mapping ordinances and tweets to SCC…...35
   4.2 Experiments and Observations on Mapping ordinances to tweets ……….......39
   4.3 Experiments and Observations on sentiment analysis of ordinances and tweets.40
   4.4 Challenges ………………………...…………………………………………43

5. Related Work
   5.1 Smart Cities …………………………..………………………….…..………45
   5.2 Enabling Data Driven Research for Smart Governance ……………………..…48
   5.3 Use of Commonsense Knowledge ……………………………………………..49

6. Conclusions
   6.1 Summary ……………………………………………………………………… 54
   6.2 Technical Contributions ………………………………………………………. 54
   6.3 Future Work ……………………………………………………………...... 56

7. References ……………………………………………………………………… 57

List of Figures

1. Figure 1 ……………………………………………………………….………..... 8
2. Figure 2 ……………………………………………………………….…………13
3. Figure 3 ……………………………………………………………….………… 16
4. Figure 4 ……………………………………………………………………………. 17
5. Figure 5 …………………………………………………………………….…. 19
6. Figure 6 ……………………………………………………………………………. 24
7. Figure 7 ……………………………………………………………………………..26
8. Figure 8 ……………………………..……………………………………….. 27
9. Figure 9 ………………………………...………………………………….. 28
10. Figure 10 ………………………………...…………………………………… 29
11. Figure 11 ………………………………………..………………………….31
12. Figure 12 …………………………………………………..……………….33
13. Figure 13 ………………………………………………………………..….. 36
14. Figure 14 ………………………………………………………………..….. 37
15. Figure 15 ………………………………………………………………..….. 38
16. Figure 16 ………………………………………………………………..….. 40
17. Figure 17 ………………………………………………………………..….. 42
18. Figure 18 ………………………………………………………………..….. 42
19. Figure 19 ………………………………………………………………..….. 52

1. Introduction

1.1 Background and Motivation

This work focuses on the area of "Smart Governance", which is part of a broad concept in Artificial Intelligence known as "Smart City". Smart Governance has been described as "the capacity of employing intelligent and adaptive acts and activities of looking after and making decisions about something" [55]. The use of technology in this realm enables governments and lawmakers to achieve better decision-making capabilities. This also helps in making democracy stronger, increasing participation among citizens and having a greater influence on public welfare.

Legislative bodies issue ordinances, which are local laws passed in municipalities or counties. It is important for lawmakers to gauge public sentiment towards ordinances. With the rise in the number of social media sites and blogs, more people have a platform to engage in discussions about ordinances. The social media site Twitter has been a prominent medium for people to express their opinions on news events across the globe. Hence, conducting sentiment analysis on tweets is an effective method for determining public reaction.

The method we have chosen in order to gauge public satisfaction is by linking ordinances to the tweets to which they are related. However, linking tweets to ordinances is not a trivial operation. Both ordinances and tweets have an intricate language structure, and a simple matching mechanism would not work. However, by making use of semantics as well as pragmatics, we have been able to develop an effective method for mapping tweets to ordinances.

1.2 Problem Definition

This work is primarily concerned with the area of Smart Governance, with an emphasis on urban policy. Our objective is to analyze ordinances (local laws) by taking into account various reactions to ordinances expressed on social media platforms. Since our objective is to determine to what extent ordinances contribute to establishing Smart City Characteristics, our aim is to group ordinances into one of six different Smart City Characteristics (SCC) [1] based on their relevant characteristics. A snapshot of the SCCs is shown in Figure 1 (source [2]).

Figure 1: Smart City Characteristics [2]

The key objective that drives our research is to connect ordinances to relevant tweets by taking into account their semantic relatedness. The process of matching is a non-trivial operation and cannot be done my means of simple one-to-one keyword matching techniques. Traditional techniques used in Machine Learning have not been found to be suitable for this project as implementing such techniques would require huge datasets for training [3]. To the best of our knowledge, no prior research work has been done in the area of ordinance mining, hence, there is an absence of datasets and training data in order to implement machine learning techniques. Ours is thus pioneering work in the area. For the purpose of matching ordinances with tweets, we aim to find a matching pair (*O*,*T*) between every ordinance *O* and tweet *T* [53]. However, as there are millions of tweets and thousands of ordinances, a simple mapping approach would result in a complexity that would be computationally infeasible due to the extremely large mapping space. We therefore propose to develop an efficient linkage approach which will reduce the number of similarity computations while still being able to capture the semantic and pragmatic aspects of ordinances and tweets.

1.3 Layout and Organization

The content in this thesis is organized as follows: in section 2, we discuss Commonsense Knowledge (CSK) sources and describe in detail how we incorporated them into our approach for linking ordinances to tweets. We also discuss the role of CSK in building knowledge bases for tweets as well as ordinances. We then introduce the TOLCS approach that was proposed by us and implemented to map ordinances to tweets. We also discuss the field of sentiment analysis and how it applies to our work. In section 3, we look at the classification tool developed for mapping tweets and ordinances to SCC followed by a discussion of the TOLCS approach. Section 4 is concerned with evaluation of our experiments and observations as well as discussions on experiments, the challenges in this project as well as future contributions to this work. We explore a literature on related work in section 5. Finally, we present a summary of our project and our contributions in section 6, which includes a section on technical contributions and future work.

2. Proposed Methodology

2.1 General Description

Twitter is a major source of data today with over 250 million users [4] and an estimated 500 million tweets sent each day [4]. Due to the vast number of tweets sent out each day, it is a valuable source of data on public reactions to ordinances.

Our work involves mapping tweets to ordinances by taking into account mutual connections with different Smart City Characteristics (SCC). To the best of our knowledge, our work in this area represents a pioneering work in the field of mining ordinances. In order to accomplish our objective, a two-step approach for mapping was followed which takes into account the connection between ordinances and tweets. [53] The connection is transitive in nature and the specific transitive property we invoke is: i*f an ordinance relates to a tweet and a tweet relates to the same SCC, it can be inferred that the ordinance bears some relation to the tweet.* The existing sources of SCC data [1][5][6] are limited to a set of identifying features which can be used for mapping. Thus, this transitive method is more practical than a method which involves mapping a large number of tweets and ordinances from different websites.

The first step in the mapping process was to connect ordinances to relevant tweets by drawing on their semantic relatedness. Since ordinance to tweet matching is not a trivial operation, a simple keyword matching algorithm is not a good approach for this. Also, we did not have sufficient prior training data in order to address various issues. [53] This approach made sense because classical sources of SCC data are finite and are restricted to only a limited set of identifying features which can be relied upon for mapping as shown in Figure 1.

Developing the technique for mapping ordinances to SCC involved discovering connections between SCCs and ordinances by making use of classical SCC sources guided by Commonsense Knowledge. We also considered the mapping of tweets to SCCs by drawing on such CSK. This approach enabled us to directly relate ordinances and tweets to pertinent aspects of Smart Cities and also set the stage for sentiment polarity classification as well as sentiment aspect analysis [9] of related tweets by using appropriate methods to assess public opinion. This work aimed for a broader impact by contributing to development towards smart cities. The task is to identify which SCCs addressed by local laws or ordinances are passed by urban agencies in order to provide feedback on how well urban policies head towards Smart City development in different categories. The Figure 2 below shows the mapping approach we used in the first part of the project.

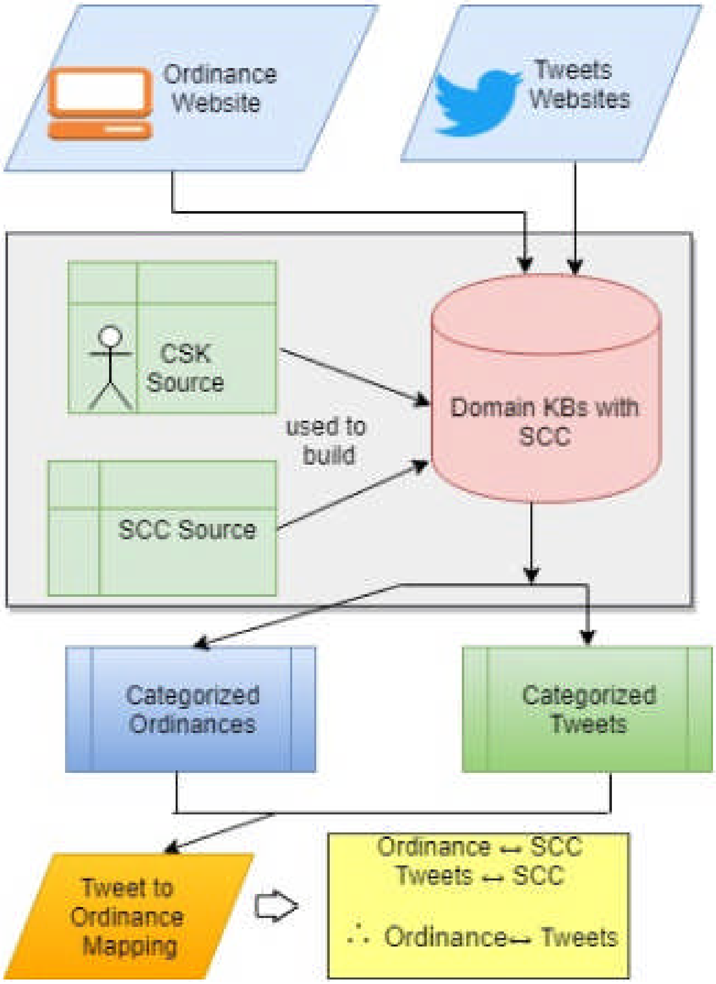

Figure 2: Mapping approach for mapping tweets and ordinances to SCC [53]

2.2 Commonsense Knowledge

Commonsense Knowledge (CSK) has been an integral component of this work. CSK can be best described as knowledge which is inherent in human beings but is not obvious to machines. Encyclopedic knowledge tends to be factual and explicit, while CSK does not have those traits. With the advent of the Internet, machines have been able to have greater encyclopedic knowledge than humans. [44] Today, a gadget such as Amazon's Alexa would be able to provide more details on yesterday's weather or the 10 highest rated movies on IMDB (Internet Movie Database) than a human. However, machines with encyclopedic knowledge do not possess commonsense reasoning, such as not littering on an open street. Machines today fail to distinguish between seemingly similar but unrelated objects, such as falsely classifying a bridge as a bench. [45]. While this trivial example may not have much repercussion in the real world, another example where a truck was misclassified as an overpass [45] can present risk to someone driving a car with such a feature. Hence, the use of CSK helped us to capture intuitive human judgment and make matches to the closest matching SCCs. The context-dependent nature of Commonsense Knowledge can be affected by reporting bias [46] and requires us to consider multiple modalities. Common sense can be partitioned into three dimensions:

i) Common sense of objects in the environment.

ii) Common sense of object relationships such as taxonomic or spatial

iii) Common sense of object interactions, such as actions, processes and procedural knowledge.

Based on this discussion, the various sources of CSK are explained below.

2.2.1 WebChild

WebChild repository [10] [11] comprises of Commonsense concepts from large amounts of data along with their properties and relationships. Today, there has been a rise in automatically constructed knowledge bases (KBs) online such as Google's knowledge graph [47] and IBM Watson project [48]. Today, dbpedia.com and freebase.com represent the largest publicly available KBs, however, the strength of these KBs is taxonomic and factual in nature, and not having knowledge which Machines would benefit from having such knowledge. The WebChild repository seeks to contribute in this area as a method for automatically extracting and cleaning commonsense properties from the web [10]. WebChild has more than 4 million triplets for fine grained relations such as *hasSubstance*, *hasProperty*, etc. The KB makes use of a Label Propagation graph for learning domain and range set, as well as extension of such relations to large scale. Graphs which connect nouns, adjectives and WordNet senses by weighted edges are built and sense relatedness and pattern statistics are used to get edge weights. WordNet and Web data are used to get seeds in order to start the LP graphs and LP algorithms are used to get assertions of good quality for refined connections between noun senses and adjective senses [11]. A partial screenshot of the WebChild repository can be seen in Figure 3 herewith.

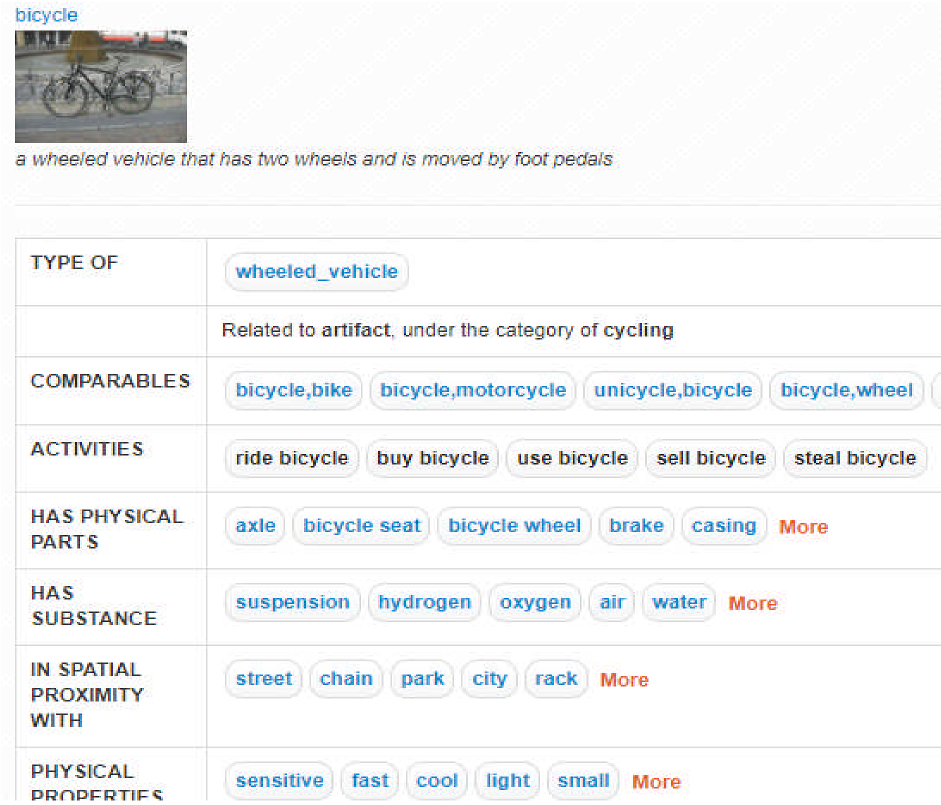

Figure 3: Snapshot of WebChild repository [63]

2.2.2 WordNet

WordNet is an electronic lexical database for the English language which was developed at Princeton University [14]. WordNet represents a semantic network linking words and groups of words by using lexical and conceptual relations which are represented by label arcs. Synsets represent building blocks of unordered sets of synonymous words and phrases. Synsets can include one or more sentences which illustrate the synonyms' usage. For example, the word "base" appears in different synsets such as *floor*, *basis, root,* and *infrastructure.* WordNet comprises of four components – each having synsets with words from 4 categories: nouns, verbs, adjectives and adverbs. WordNet has more than

100,000 synsets , which consists of 81,000 noun synsets, 13,600 verb synsets, 19,000 adjective synsets and 3,600 adverb synsets. In WordNet, noun concepts are linked as specific concepts to general ones and are derived from entity types vs instances. Another relation among noun synsets is metonymy, where synsets denoting parts, components and members are linked to those denoting a whole part [59].

Inheritance is an important component of WordNet. Inheritance relates to relations which are based on hierarchy building. For example, if a car has four wheels then a racecar would have four wheels as well. However, upward inheritance does not apply here i.e., a car has four wheels, that does not mean that every vehicle has four wheels.

WordNet has been widely used in the area of Natural Language Processing (NLP), especially in the area where a particular word can have multiple senses. However, with it being a lexical resource, does not provide syntactic information.

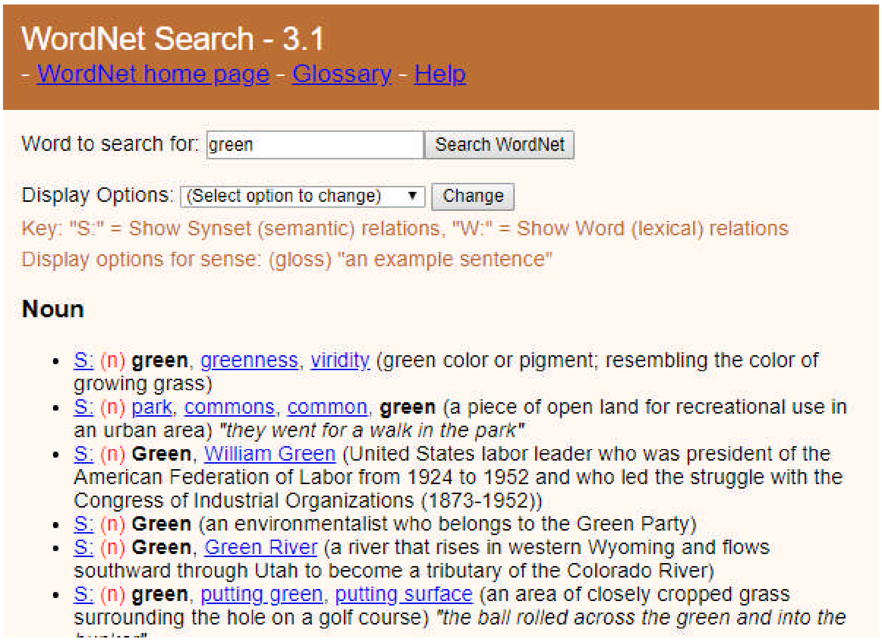

Figure 4: Sample search on WordNet [60]

2.2.3 Cyc

The Cyc project is one of the oldest projects in AI. The project is based on the idea that a core of knowledge which provides context for novel learned information results in effective machine learning [61]. The project's long-term goal is to automate the building of a formal representation of the world in the Cyc knowledge base using machine learning and build a repository of commonsense knowledge in machines. Cyc comprises of three components – a knowledge base (KB), an inference engine and a natural language system. The Cyc KB has more than 3.2 million fact and rules which describe approximately 300,000 concepts. The natural language system has a lexicon which is responsible for mapping words as well as phrases to Cyc concepts, as well as a parsing and generation subsystem, which translates English to CycL [61]. Cyc has been effective in both driving a learning process as well as showing results.

2.2.4 DBpedia

DBpedia is a web project designed to get structured content from Wikipedia pages and make the information available on the Web. It allows users to perform complex queries for Wikipedia datasets and find links between Wikipedia data and other datasets [62]. The project aims to address one of the key challenges in the area of Computer Science – combining structured information in the world in order to answer complex queries dealing with semantics. The project is concerned with deriving a diverse data corpus from Wikipedia. The project consists of an information extraction framework from which information extraction , clustering and querying can be done. Its dataset consists of 103 million RDF triplets, which can be applied to a variety of semantic web applications. The use of various interfaces and access modules allows datasets to be accessed from Web services and can also be imported using other applications. The DBpedia dataset holds information on approximately 2 million objects. Datasets may contain specific information or meta-data without specific semantics. We used the *Search DBPedia.org* application to explore datasets on DBpedia. A snapshot of various components related to DBpedia is shown below in Figure 5.

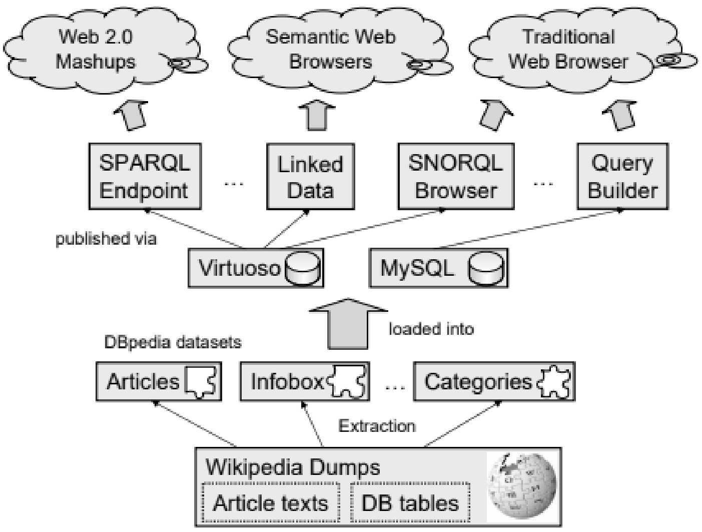

Figure 5: Snapshot of DBpedia [62]

2.3 Smart City Characteristics

"Smart City" refers to a concept in which cities around the globe are more technologically advanced and more efficient with the ultimate aim of improving the quality of life among its residents. The residents of a Smart City work together to achieve common values such as energy efficiency, transparency in government, and conservation of environment. Smart City Characteristics represent closely related building blocks which contribute to a Smart City system. The rapid urbanization in the world today has given rise to several challenges in improving urban living in different parts of the world. There are six characteristics, known as " Smart City Characteristics", which serve as building blocks for a Smart City system.

Some of the global efforts of promoting engagement in the area of Smart Cities can be seen on the IEEE Smart Cities website [49][50]. Via this initiative, IEEE aims to help different towns and cities to manage their transition to urbanization. It also hopes to promote awareness about the advantages as well as disadvantages of technology and how it can be used appropriately. With the world population expected to double by 2050 [49], there is an increased need to have intelligent, sustainable environments while having minimal impact on it. By bringing together different aspects of technology, government and society, the organization aims to make the transition towards Smart Cities efficient.

The source of SCC in our work is derived form a widely accepted tech report from TU Wien [1]. This report offers a perspective on medium-sized cities by identifying their strengths and weaknesses by taking into account a range of factors related to the concept of "Smart Cities". The rise in globalization has brought about rapid changes in production and consumption, which carries significant effects on the development of cities. Cities in Europe face a huge challenge of integrating competition with sustainable development, and this will likely have an impact on urban quality and housing [38]. Most research on globalization today tends to focus on “alpha cities” or metropolises. Hence, medium-sized cities have not received much attention. Assigning rankings to different cities has been a popular way for gauging potential for different urban region for the past few decades, as a higher ranked city will have better marketing capabilities. However, this approach has several drawbacks, namely, that it fails to consider complex relationships in regional development and that it ignores long-term development strategies while strengthening existing stereotypes. To use a better approach, the report describes six different Smart City Characteristics (SCC):

i) Smart Governance

ii) Smart Economy

iii) Smart Mobility

iv) Smart Environment

v) Smart People

vi) Smart Living

Taking the example of Smart Living, this SCC comprises of various aspects of quality of life such as:

i) Cultural Facilities

ii) Health Conditions

iii) Housing safety

iv) Education facilities

The detailed explanation about Smart Cities in the TU Wien report and is being used as a basis for this work.

If a tweet or ordinance is found to reference any of the above aspects, it can be inferred that it is likely to relate to Smart Living. However, these expressions are not likely to be inspected in ordinances, which can be subtle. In order to automate the process, we rely on CSK web sources such as WebChild and WordNet.

WebChild and WordNet have served as the two main sources of CSK to build the Knowledge Bases (KB) in this work. The KBs are text based and contain terms for specific Smart City Characteristics which are derived from CSK repositories, as well as SCC sources, with the aid of Natural Language Processing and semantic linking. The size of these domain KBs can further be increased by applying techniques such as knowledge-based extraction of text [5] [12] as well as association rule mining.

2.4 Methodology:

Our process starts by identifying connections between SCCs and ordinances by using classical sources of SCC which are guided by Commonsense Knowledge (CSK) from web-based data repositories. We then consider mapping tweets with SCCs by drawing on the relevant CSK. This allows us to link ordinances and tweets to relevant aspects of Smart Cities which also forms a basis for sentiment polarity classification [5] [7] [8] as well as sentiment analysis of relevant tweets which allows us for public opinion assessment [9]. The mapping technique is then enhanced to allow for mapping of tweets and ordinances to multiple SCC. Finally, we develop a technique to link individual ordinances to tweets by drawing on their semantic relatedness.

This work serves to make a broad impact by indirectly making a contribution towards the development of Smart Cities. Once we identify which SCCs are being referred to by local laws, feedback can be provided on how urban policies can relate to Smart City development throughout different categories. This work also relates to the area of Social Sensing. Public reactions derived from opinion mining can enable urban councils and management agencies to judge public reaction. This can help us assess the appeal of SCC among the public which provides useful feedback to agencies which enables them to provide better policies for Smart City development. The analysis involves invoking different concepts from Natural Language Processing (NLP), Commonsense Knowledge (CSK) and text mining.

2.5 Mapping Tweets and Ordinances to SCCs

The first step of this project involved building an algorithm to map tweets and ordinances to SCCs. A single ordinance or tweet can map to one or more SCCs [53]. The ordinances and tweets are linked to the SCCs having the most relevant number of features. By using WebChild and WordNet as the main CSK sources, domain-specific knowledge bases on Smart City Characteristics have been built. The knowledge bases are text based and have terms relevant to Smart City Characteristics derived from CSK repositories and SCC sources using NLP and semantic matching. The linking algorithm used for this process is shown below in Figure 6:

Algorithm 1 Linking algorithm

for each SCC S_j **do**
  Build domain knowledge base K_j
for each ordinance O_i **do** ▹ ordinance linking
  for each SCC S_j **do**
    $L_{i,j} \leftarrow \sum_{x \in K_j} \mathrm{C}(O_i, x)$
  Assign O_i to the S_j with $j = \mathrm{argmax}_j \, L_{i,j}$
for each social media posting T_i **do** ▹ social media linking
  for each SCC S_j **do**
    $M_{i,j} \leftarrow \sum_{x \in K_j} \mathrm{C}(T_i, x)$
  Assign T_i to the S_j with $j = \mathrm{argmax}_j \, M_{i,j}$
$O \leftarrow \{(O_i, T_k) \mid \exists S_j : (O_i \text{ assigned to } S_j) \wedge (T_k \text{ assigned to } S_j)\}$
return O ▹ Links between ordinances and social media

Figure 6: Algorithm used for tweet to SCC mapping [53]

The CSK concepts are deployed to semantically relate to terms *x* in ordinance text *T* to SCCs. This is denoted as *C(T,x).* If a tweet or ordinance text contains the term *smoke detector,* CSK concepts will help us to semantically relate this term with the SCC Smart Environment through the CSK properties of smoke detector which are relevant to the SCC. We find this information in domain KBs derived from SCC and CSK sources. It is possible that a single ordinance may map to multiple SCCs. In that case, their occurrences would be counted towards multiple categories. For example, if a term sustainability occurs in an ordinance, the ordinance would be counted under the characteristics of *Smart Mobility* as well as *Smart Environment.* The aggregate SCC counts are examined, and each ordinance is accordingly linked to SCC with the maximum number of relevant features. [53] CSK plays an important role in finding semantic relatedness for this type of mapping via concepts, properties, etc. Tweets are mapped to SCCs using a similar CSK guided procedure. Then, we output the linkages between ordinances and tweets via mutual SCC connection. This approach is summarized in the algorithm above in Figure 6.

The next step of the project was to enhance the original technique to map multiple ordinances and tweets to SCC. The mapping to different SCC depends on weights of relevant terms assigned to them. For example, if an ordinance has three terms – two related to Smart Mobility and the third related to Smart Governance, the ratio of Smart Mobility: Smart Governance is 2:1. [52] An ordinance or tweet could contain a significant number of terms related to a single SCC, and other terms may be ignored if they are not relevant to the core message of the tweet. This feature can be determined by a threshold, which can be adjusted as needed. Below is the algorithm we used for mapping ordinances and tweets to SCCs. For each word mapping, we use two techniques for word normalization – stemming and lemmatization, which returns the words in tweets to their root forms when determining a match with a term in the SCC knowledge base. Both stemming and lemmatization operations have been performed in this work using

NLP libraries in Python. Figure 7 present our algorithm for mapping tweets and ordinances to Smart City Characteristics (SCCs).

for each SCC S_i **do**:
    build domain KB K_i
$A \leftarrow \emptyset$
for each ordinance O_i **do**:
    for each SCC S_j **do**:
        $L_{i,j} \leftarrow \sum_{x \in K_j} C(O_i, x)$
    $A \leftarrow A \cup \{(O_i, S_j) \mid j = \text{argmax}_j L_{i,j}\}$
for each tweet T_i **do**:
    for each SCC S_j do:
        $M_{i,j} \leftarrow \sum_{x \in K_j} C(T_i, x)$
    $A \leftarrow A \cup \{(T_i, S_j) \mid j = \text{argmax}_j L_{i,j}\}$
$\theta \leftarrow \{(O_i, T_k) \mid \exists S_j : (O_i, S_j) \in A \wedge (T_k, S_j) \in A\}$
return θ

Figure 7: Algorithm for mapping tweets and ordinances to SCCs [52]

This algorithm invokes a transitive property that if an ordinance maps to one or more SCCs and if the tweets map to the same SCCs, the ordinances broadly map to the tweets. This property can be used to determine actual ordinance to tweet mapping on a broad scale. This ordinance-tweet mapping does not deal with the finest levels of granularity and maintains a connection at the level of relevance to SCCs. A sample of various SCC domains that the ordinance / tweet is mapped to is presented in the Figure 8 below.

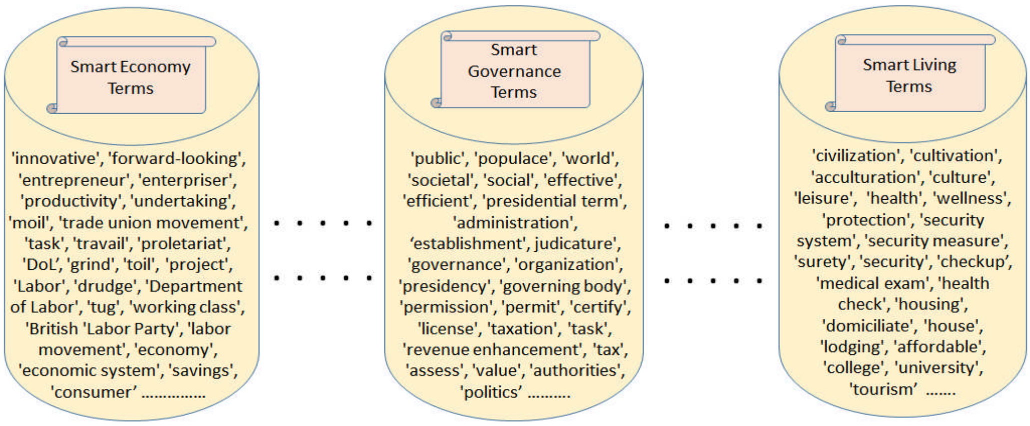

Figure 8: Sample SCC Domains [52]

2.6 The TOLCS Approach

An approach called TOLCS (Tweet Ordinance Linkage by Commonsense and Semantics) is proposed in order to link ordinances to relevant tweets [51]. This method uses Commonsense Knowledge (CSK) and text similarity methods in order to capture pragmatics and semantics. Rather than figuring out the similarity between every tweet and ordinance, we use external knowledge sources to design blocking steps which result in candidate matching pairs. A reduction of the linkage complexity is done to improve efficiency in mapping. The contributions made by this approach are as follows:

i) A knowledge base (KB) for Smart Cities and Ordinances is constructed using sources of SCCs as well as ordinances websites. We harness CSK from different repositories worldwide in order to build the KB.

ii) A blocking mechanism was designed to reduce the mapping space between the tweets and ordinances using the KBs.

iii) Extensive experiments were conducted to evaluate the performance of TOLCS.

The TOLCS methodology is based on the previous mapping methodology used for mapping tweets and ordinances to SCC and is shown in figure 9.

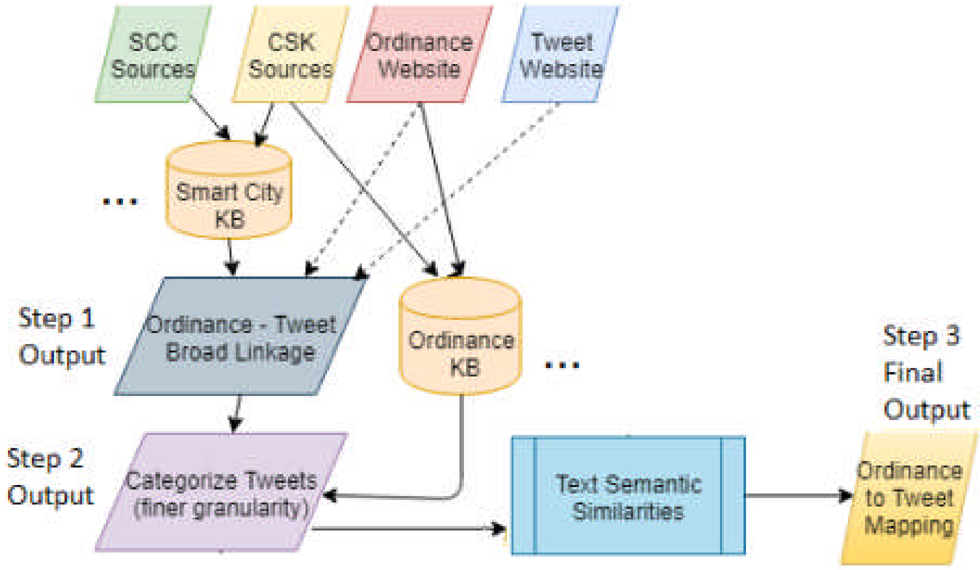

Figure 9: Proposed ordinance to tweet mapping technique [51]

2.7 Sentiment Analysis

In our research, sentiment analysis is conducted in order to discover knowledge with respect to opinion mining of ordinances to tweets. The task focuses on assessing whether a sentiment expressed in a piece of writing is “positive”, “negative” or “neutral”, and also whether the writing heads in a particular direction of polarity, such as “strongly positive”, or “strongly negative.” Sentiment analysis serves as the basis for opinion mining, which involves discovering patterns or knowledge from people’s opinions. The area of sentiment analysis has seen an increase in the number of research activities as it has applications in a wide variety of areas such as business, politics and law.

We use SentiWordNet [13] as the primary database for Sentiment Analysis in this work. This has been built for guiding sentiment classification and opinion mining. SentiWordNet makes groups of words as synonym sets and annotates them by how positive or negative the words are. Hence, terms are classified into either positive, negative or neutral based on their polarity. If a word has multiple meanings, the meanings have different scores assigned to them.

For example, the word *estimable* when it pertains to computation has a neutral score, while the same word in the sense of “deserving respect” has a positive score of 0.75. We employ a semi-supervised learning method using SentiWordNet. This helps to incorporate subtle human judgment into the mining processes. The algorithm for implementation of sentiment analysis is given below in Figure 10.

for each tweet t_i **do**:
    if not (t_1 relevant according to SCC KB):
        continue (with next tweet)
    map t_i to ordinances using Algorithm 1
    $W_i \leftarrow$ set of words in t_i
    for each $w \in W_i$ **do**:
        $s_w \leftarrow$ polarity score of w in SentiWordNet
    $s_i \leftarrow \sum_{w \in W_i} s_w$
return final polarity scores s_i for relevant t_i

Figure 10: Implementation of sentiment analysis [52]

This algorithm helps us in polarity classification of tweets obtained from twitter. The selection of relevant tweets and mapping them to relevant ordinances is guided by CSK.

The domain KBs from WebChild and WordNet help to distinguish the relevant tweets in the first step and is followed by mapping tweets to ordinances.

3. Implementation Details:

3.1 Smart City Classifier Tool

A Smart City Classifier Tool was built in Python using the methodology discussed in the earlier section. The application works by semantically relating *n* terms in an ordinance text to Smart City Characteristics (SCC). It can be possible for terms in ordinances to overlap with multiple SCCs. In that case, the term would result in a count increase for both SCCs. The aggregate SCC counts are linked to SCC with the maximum number of relevant features. CSK helps to find semantic relatedness for mapping via different concepts and properties. Finally, we find the linkages between ordinances and tweets via mutual SCC connections. Our mapping principle can be summarized as: if an ordinance maps to a particular SCC and the same SCC maps to a particular tweet, the ordinance can be mapped to the tweet. The Figure 11 below shows a snapshot of the GUI (Graphical User Interface) for this tool.

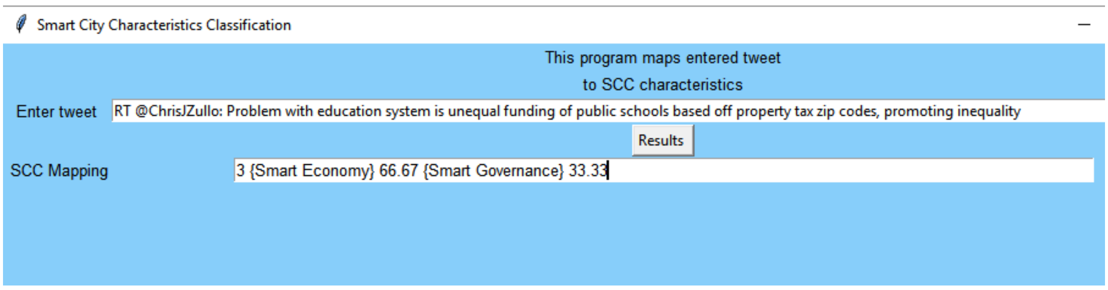

Figure 11: Snapshot of Smart City Classifier Tool

3.2 TOLCS Approach Description

TOLCS comprises of three steps when mapping ordinances with their respective tweets. Smart City Characteristics such as Smart Governance and Living are used as a nexus between ordinances and tweets, since these SCCs are finite compared to an overwhelming number of ordinances and tweets. CSK from different repositories worldwide is used, as well as SCC sources to build Knowledge Bases (KB) relating to each SCC. CSK and ordinance websites can be further used to build KBs with relevant ordinances terms which incorporate domain knowledge on urban policy. So, for each ordinance department, an ordinance KB was built. Our implementation of TOLCS has the following three steps.

1. The mapping space was reduced by broadly relating large groups of ordinances and tweets, with SCC as a nexus.
2. For each SCC group, the mapping space was trimmed by relating smaller groups of ordinances and tweets by mutual relevance with each department's ordinance KB.
3. With a finer mapping space being obtained, specific ordinances were linked to tweets using properties of text similarity and semantic relatedness.

Based on this discussion, the TOLCS algorithm based on this approach is shown below in Figure 12.

Input: θ: similarity threshold to link ordinances and tweets
Output: $\mathcal{R}$: the set of correlated ordinances and tweets
1: let $\mathcal{R} = \emptyset$
2: **for all** ordinances O_i **do**
3: find the most related SCC and ordinance department
4: **end for**
5: **for all** tweets t_j **do**
6: find the most related SCC and ordinance department
7: **end for**
8: **for all** ordinances O_i and tweets t_j that are related to the same SCC and ordinance department **do**
9: **if** $sim(O_i, t_j) \geq \theta$ **then**
10: $\mathcal{R} = \mathcal{R} \cup \{(O_i, t_j)\}$
11: **end if**
12: **end for**
13: **return** $\mathcal{R}$

Figure 12: TOLCS Algorithm [51]

Considering Smart city KBs and ordinance KBs, a semantic similarity between an ordinance and tweet is calculated if they are found to be matching with the same SCC and ordinance department. In the algorithm above, lines 2-7 result in a complexity of *O((m+n)(p+q))*, where *m* and *n* represent ordinances and tweets and *p* and *q* represent the number of SCCs in Smart City KBs and departments in Ordinance KBs. Since the number of SCCs is less than 12 and ordinance departments is less than 25, p and q are relatively small constants. The time complexity is then reduced from O((*m+n)(p+q*)) to O(*m+n*). As we are dealing with millions of tweets, the decrease in time complexity is significant. If 10 million tweets are found initially, the number is reduced to 1000 for similarity computation, which represents a decrease of 4 orders of magnitude. Building the KB represents a one-time effort but can be used repeatedly once it is built.

3.3 TOLCS Approach Implementation

For the implementation of TOLCS, we build Smart City KBs using sources of SCC such as [1] as well as sources of CSK such as WebChild [10] and WordNet [14]. The ordinance KBs are built using the NYC legislative websites [15] as well as CSK sources used to build Smart City KBs. For the experiments here, over 5000 tweets from the NYC area were used.

When measuring the similarity for linking ordinances to tweets, the skip-gram model [16] is applied to embed root words in ordinances and tweets in a manner which preserves semantics. For an ordinance and tweet, the similarity factor is measured as an average pairwise cosine similarity between the enclosed word embeddings. These capture the pragmatic and semantic aspects when mapping ordinances to tweets.

Cosine similarity works by ignoring the magnitude and focusing on the orientation of texts. This allows for detection of similarities between a larger document with the same theme as a shorter document.

4. Experimental Evaluation

4.1 Experiments and Observations on Mapping Ordinances and tweets to SCC

For the first part of the project, we conducted 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 the experimental evaluation is as follows.

We had gathered large amounts of historical data on ordinances from the NYC council website [15], which is openly available to the public. A portion of the website is seen in Figure 14. The ordinances are first extracted into a machine- readable form and subjected to preprocessing step such that only their textual content is retained. Other attributes such as “Prime Sponsor”, “Council Member Sponsor” are filtered out during the preprocessing phase. The textual content of the ordinances serves as input to our algorithm which conducts the ordinance to SCC mapping. A sample ordinance and its SCC mapping is shown below in figure 13.

Sample Ordinance: *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...*

Smart City Characteristic	Count of Terms
Smart Economy	1
Smart Mobility	0
Smart Environment	0
Smart Governance	2
Smart People	0
Smart Living	0

Figure 13: SCC Mapping for sample ordinance [52]

With reference to this ordinance excerpt, our algorithm relies on SCC Domain KBs and comes to the conclusion that only term relevant to the *Smart Economy* characteristic is *business*, while many terms are relevant to *Smart Governance* such as *law* and *district plan* as summarized in the table above.

We use the NYC legislative council as the source of the ordinances, shown in Figure 14 below. The tweets are mapped using the mapping algorithm described in Figure 5. We consider different terms from input tweets and ordinances and assign weights to them. This helps us find how relevant the tweets and ordinances are to SCCs based on different threshold levels. The tweets and ordinances are then entered as input in our GUI as shown in Figure 11, and we receive a mapping of the relevant SCC or “No Mapping” if a relevant mapping is not found.

Figure 14: The New York City Council Website [15]

Let us take the example of another ordinance obtained from this NYC source:

"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."

For this ordinance, we get the following mapping results: Smart Governance (50%) and Smart Economy (50%). Hence, the ordinance relates to both SCCs at roughly an equal level.

To illustrate the example of tweet to SCC mapping, we take the examples of two tweets:

i) *Problem with education system is unequal funding of public schools based off property tax zip codes, promoting inequality*

Here, the number of SCC terms equals 3. The weight of terms for Smart Economy versus Smart Governance is 2:1

ii) *Adding plants to your home or office has so many benefits such as reducing stress and increasing productivity*

Here, the number of SCC terms equals 1. The tweet maps to Smart Environment.

The outcomes of the mapping are submitted for assessment to domain experts from the Department of Earth and Environmental studies at Montclair State University. The experts define ground truth such that they identify a mapping to be correct if it agrees with their judgment. This is considered as True Mapping (TM). If the experts disagree with the results, this is known as false mapping (FM). Hence, precision is given by:

$\text{Precision} = \frac{TM}{TM + FM}$ and this metric is used in accuracy evaluation.

In this research, we do not consider true negatives and false negatives since their definition requires deeper insights and more complex discussions with domain experts.

The results of the evaluation are as follows:

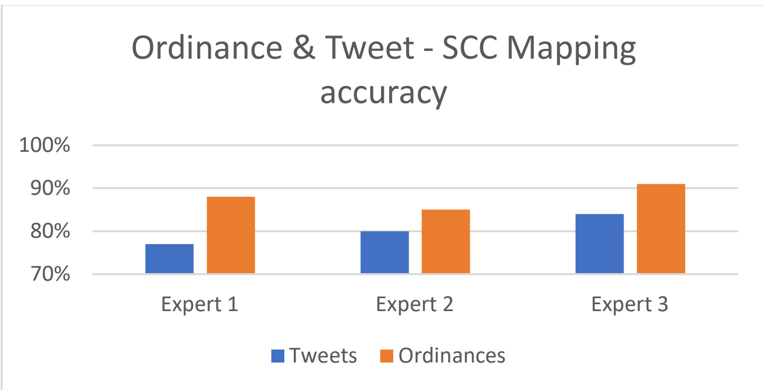

Figure 15: Summary of mapping assessment

From the evaluations, we found that 88% of the ordinances were correctly mapped while 78% of tweets were correctly mapped.

4.2 Experiments and Observations on Mapping Ordinances to tweets.

For evaluating TOLCS, real data from two NYC ordinance sessions was used: Session 1 (2010-2013) and Session 2 (2014-2017), as seen in [15]. Tweets from the NYC area were chosen.

These are examples of linking ordinances to relevant tweets:

"A Local Law to amend the New York city charter, in relation to planning for resiliency to climate change as a responsibility of the office of long-term planning and sustainability."

RT @NRDC: Climate change poses a threat to national security-and the Pentagon doesn’t seem inclined to keep quiet about

"Red pandas, climate change, and the fight to save forests | Stories | WWF change believers”

Remember the very warm winter in NYC that liberals said was proof of climate change? Now that we're having the coldest winter ever here...”

A Local Law to amend the administrative code of the city of New York, in relation to tax exemption and abatement for certain rehabilitated buildings as authorized by section 488-a of the real property tax law

@NYCSpeakerCoJo @MarkGjonajNY\\n\\"the rent is too damn high, property taxes are too damn high, the subways are in a ...

As property taxes go up, so does the rent. 44% increase since 2013 completely unsustainable & a big reason NYC has become so unaffordable."

In order to evaluate the effectiveness of the mapping, we looked at the evaluation of three different domain experts. If an expert agreed with the linkage, the linkage was considered to be accurate. We measure precision as a percentage of accurate linkages.

The performance evaluation of TOLCS can be seen in the figure below. We observe that about 77% of the ordinances and tweets linked by TOLCS are accurate.

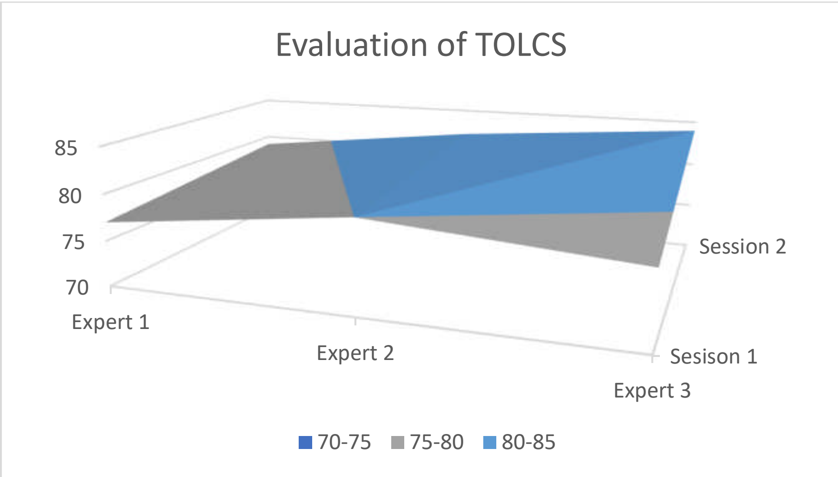

Figure 16: Summary of TOLCS evaluation

4.3 Experiments and Observations on sentiment analysis of ordinances and tweets

We use tweets from the NYC region for purposes of sentiment analysis. For sentiment analysis, we use a dictionary-based sentiment scoring mechanism based on SentiWordNet. Sentiment scores between -1 and 1 are assigned. The method we choose is domain-independent as it is based on prior knowledge of the sentiment values assigned to different words. The approach works by weighting each sense with a geometric series ratio of ½. This is done so that the senses which occur most frequently have a more affective weight than senses which rarely occur [17].

We use the following examples to conduct sentiment analysis of tweets:

Tweet: *and now... time to get ready for my exercise class... what a fun morning this has been*

Sentiment score: 0.43 (positive)

Tweet: *"Ok this sucks, I'm still feeling under the weather, but, i'm soldering on. I'm in the office today. Sunday night I had a 103 deg temp. "*

Sentiment score – 0.29 (negative)

We send the polarity classification for evaluation from experts, who assess it as correct or incorrect depending on whether or not they agree with the sentiment expressed in the tweet. Accuracy is then calculated as the percentage of correct assessments of all polarity classifications. On the basis of this, the overall evaluation receives an average accuracy of 84%. The results from three domain experts are given below.

- Domain Expert 1 – Accuracy = 83%
- Domain Expert 2 – Accuracy = 81%
- Domain Expert 3 – Accuracy = 88%

We then summarize the public reactions expressed on all tweets and express the results in the pie chart below in Figure 17.

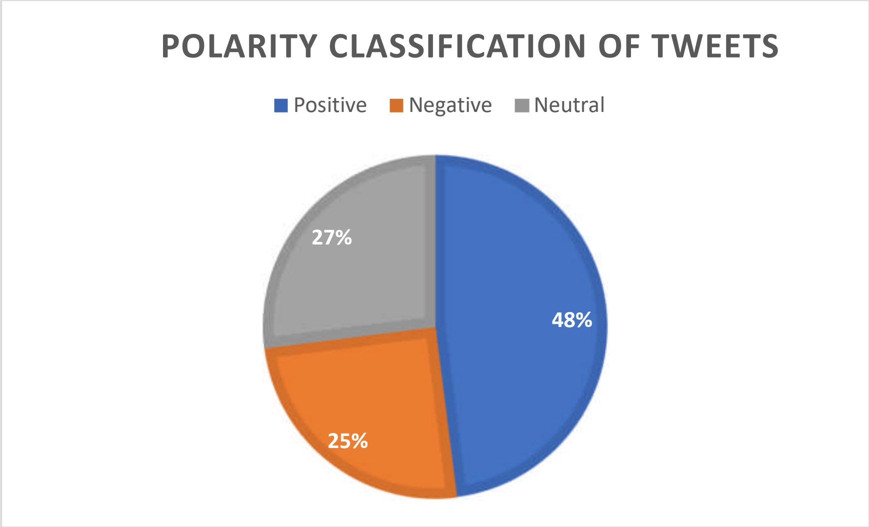

Figure 17: Polarity Classification of Tweets

This figure indicates that on the whole, the public is satisfied with the ordinances in NYC, since there are more positive tweets than negative and neutral ones. However, since there are still some negative and neutral tweets, it implies that there is scope for improvement in urban policies, especially with respect to Smart City initiatives.

Furthermore, we also analyzed the public satisfaction with respect to each SCC, the results of which are tabulated in Figure 18.

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

Figure 18: Public contentment for each SCC

From the results above, we see that the public is most positive about tweets related to Smart Living and is the least positive about those related to Smart Environment. On the whole, the public seems content with many of the urban policies pertaining to different Smart City Characteristics. From our observations, we can conclude that NYC exhibits characteristics which shows that it is leaning towards a Smart City. These results can be provided as feedback to different agencies in order to help them make decisions regarding Smart Cities in general and NYC in particular.

4.4 Challenges

To the best of our knowledge, our work is the first of its kind in the field of Smart Governance. We encountered several challenges when mapping tweets and ordinances to SCC:

a) Extraction and analysis of tweets is very difficult due to their informal language.
b) Excessive use of acronyms is rampant in tweets due to the limited number of characters in tweets, which further increases the difficulty in analysis.
c) Due to the limited coverage of tweets, 1/3 of the mentions cannot be linked to Wikipedia mentions.
d) There are several degrees of uncertainty associated with Named Entity Extraction (NEE) and Named Entity Disambiguation (NED).

We have been able to map ordinances and tweets to SCC on a broad level of semantic relatedness. Due to this, a few of the ordinance and tweet mappings have been observed to have lack of precision with respect to their SCC identification.

In the case of sentiment analysis, the domain experts have been asked to assess our classification of sentiments evoked by tweets. The polarity classification of tweets has been found to be mostly accurate. This means that if a tweet was classified as “positive” by our method, the experts have also determined that this is “positive” on the basis of ground truth defined by them. However, we have also found instances of incorrect polarity classification since tweets do not follow a systematic grammar structure, which makes it difficult to get semantic patterns in a few instances.

While mapping ordinances to tweets, we have been able to reduce the complexity from polynomial time to linear time as shown in our calculations. There are several challenges faced when establishing a mapping technique between ordinances and tweets. The primary obstacle we have faced is that tweets do not follow a standard language structure, and they include different acronyms and hashtags. This makes establishing relationships between ordinances and tweets very difficult. Through our experiments, we have achieved the best possible results by using our proposed methods in this work. However, additional research needs to be done if the accuracy can be improved using other text similarity / machine learning techniques. This calls for further research and presents the scope for future work by others working in the area.

5. Related Work

The dramatic rise in the number of social media users in the past decade has brought about different opportunities to analyze data from users which has applications in various fields such as business, economics and politics. While there has been tremendous activity in the area of social mining, previous work in this area is very different from our work.

The activity of predicting links on social networks [18] has a long history. However, existing methods have been focused towards creating links between homogenous nodes, for example finding connections among users of a social network. There has also been some interest in the area of entity resolution [19]. There have only been a few approaches which focus on open-domain linking between arbitrary entities and concepts [20] [21] [22]. However, these approaches involve dealing with structured data. Data in this form have different attributes. Our work is different in that we are trying to connect two forms of unstructured text for a given natural language. While terms in ordinances are expressed using formal language, which includes various legal terminologies, social media posts use language which is very informal, including hashtags, URLs, etc.

5.1 Smart Cities

With the advent of increased social media use in the past decade, we have seen a lot of work in the area of social media mining. However, our work involving tweets and Smart Governance is novel. Link prediction is an active research in the area of social networks [23] [24]. However, most research in the area is in the area of predictions related to connections between users. Various attempts have also been made in order to solve problems related to entity resolution as well as resource allocation [19]. While ample work has been done in other areas, there has been only a handful of research in the area of finding connections between arbitrary entities [5] [25] [26]. Moreover, the research in these areas involve looking at structured data, which have objects with different attributes. In contrast, our work involves finding connections between unstructured language texts. While the ordinances we work with are written in formal language, tweets have a very informal language with the use of different hashtags, URLs and emojis.

An important area of research in the area of social media mining is unsupervised topic modeling and trend detection. [27]. An example can be seen in [28], where hashtags were analyzed using a fuzzy-based approach in an attempt to investigate trends in hashtag popularity. However, this cannot be easily applied for mapping tweets to a set of ordinances, which we have taken into account in our research. Vector based representations of elements [29] are insufficient when we have two heterogenous items as in our case. Recent approaches involving linking social media text have relied on supervised classification. While standard methods can be used for predicting links between heterogenous items [36], the relatively short length and variability of tweets makes it difficult for large training sets to accurately cope with them. It also leads to data sparsity. The TweetSift system [30] classifies tweets by topic while using external entity knowledge and topic enhanced word embeddings. The embeddings result in topic-specific word embeddings where different senses of ambiguous words can result in different representations. This model makes an assumption that entities such as signals for specific twitter users can be determined from the knowledge base. In contrast, our methodology uses generic Commonsense Knowledge without the use of a detailed labeled training set.

Based on a literature survey of various such works, we claim that our research on ordinance and tweet analysis with respect to Smart City Characteristics is pioneering in this area, as previous work has not considered a setting involving the aforementioned topics. This claim is to the best of our knowledge. Our work herein certainly propels more joint work on topics in urban studies, data mining and related fields.

There has been much attention given to Smart Cities in recent years. We have seen that buses in Barcelona are run on routes which achieve optimal power consumption. [1] Also, canal lights in Amsterdam are automatically able to adjust their brightness based on pedestrian usage [1]. A previous work [31] describes how automated vehicles can be enhanced by using Commonsense Knowledge. These advancements can be classified as the “Smart Mobility” characteristic. There has also been significant research done in the use of technology for fighting crime, [32] which was developed as part of the EU ePOOLICE project. This work falls under “Smart Living.” In another work, [33], we see the Smart Environment characteristic being targeted via cloud computing solutions for data centers. Scenarios were analyzed where greater energy efficiency was achieved using cloud models while meeting productivity targets. Security, privacy and availability were key issues discussed for using cloud computing in greening of data centers. Data centers which have free cooling, as addressed in [34] considering temperature, humidity and other factors fall under the concept of “Smart Environment.” In [35], an algorithm was shown to save trips on cooperative pickup and delivery, which loosely falls under “Smart Economy”. The research conducted on [36] is seen to fall under the category of

"Smart People". This described a process of addressing an aspect of 21st century education suing collocation-based writing aids. In another piece of research seen in [26], we see a contribution towards "Smart Environment" by estimating air quality via analysis of pollutant data. Hence, we observe various research activities which contribute to each of the Smart City Characteristics. The research that is presented in our work seeks to make an impact by advocating for deployment of Commonsense Knowledge in the field of Smart Cities.

5.2 Enabling Data Driven Research for Smart Governance

The first mention of the words "Smart" and "Governance" can be traced back to a World Bank report which dealt with reforming civil service [37]. Today, governments around the world face numerous challenges, most of which are beyond the reach of traditional institutions to tackle. From tacking policies related to pollution to government spending, we observe an increased lag between The field of Smart Governance has been getting traction in recent years due to widespread awareness in providing transparency in governance and involving public in decision making. While maintaining western democratic principles, government bodies need to implement institutional formats and mechanisms in order to keep up with ever changing dynamics in an interconnected society. We believe that governments can use data driven approaches in enacting policies for fostering participation and collaboration with the public at various levels. Active involvement with public as well as stakeholders would help cities reach a level of "Smart Governance" [39][40] which will create a more sustainable economic environment. [41][42]. Hence, different aspects of Smart Governance include feasibility, contribution from stakeholders and increased transparency and coordination. Various empirical studies done have also supported this idea [42][43].

5.3 Use of Commonsense Knowledge

A recent work in the area of Smart Environment was to simulate human judgment in mining social media data for a smarter environment. [26] The work involved mining social media data and structured data by taking into account their domain-specific context, incorporating commonsense knowledge in mining media opinions and focusing on urban planning domain in a multicity environment. Methods such as classification and clustering were used to get structured data from global sources and Twitter was used to incorporate CSK and build domain KBs. The outcome of this work has an application for predictive analysis in urbanization.

Another work in on enhancing autonomous vehicles with Commonsense [31] dealt with advancing CSK in the area of Smart Mobility. The project aimed to utilize commonsense knowledge in order to enable vehicles make humanlike decisions. The first step of the project involved investigating existing literature related to CSK and domain-specific knowledge bases. Then, a transportation domain KB was created by using the transportation sector of WebChild. It is then modified by including images of obstacles on the road when driving. The third step involved programming the domain KB for transportation into a small robot which served as a simulated autonomous vehicle. The robot was able to navigate well at a basic level using these instructions, however, was not able to make complex decisions due to limitations in its computing power. Hence, this work proposed an approach using CSK to enhance decision making in automated driving and built a transportation KB using filtering, ranking and augmentation. By enabling image detection capabilities on a small robot, the system was able to make basic recommendations based on its environment. It set the stage for more large-scale research in this problem, useful from a Smart Mobility perspective.

Another related work which made use of CSK is a framework known as COTIR (Commonsense Knowledge, Ontology and Text Mining for Implicit Requirements) developed for identification and management of IMplicit Requirements (IMR) in the requirements specification phase of the software development life cycle. By combining CSK, text mining and ontology, the work showed that discovering and handling of unknown as well as non-elicited requirements would reduce risks as well as costs for software development [56]. The framework has great potential for AI applications such as smart city tools.

In another study, concepts related to smart cities were discussed, where categories such as “knowledge city and “resilient city” was used interchangeably by urban policy makers. Issues of categories having distinct conceptual perspectives were addressed [57]. Their design involved using articles in databases and counting co-occurring keywords, presenting how keywords have evolved and building a network of keywords and categories. The implicit use of Commonsense knowledge in this work made use of CSK repositories and relevant domain KBs in the area of urbanization.

The use of routes which minimize the use of signals to maximize energy efficiency in cities like Barcelona [1] as well as the dimming of street and canal lights based on pedestrian usage in places like Amsterdam [1] can further benefit from the use of CSK.

In another research study in the area of image processing, in order to improve the accuracy of an object detection framework, contextual information such as global scene context and geometric context was employed in the form of 3D surface orientations, relative location and geographic information. [58] CSK provided the basis for contexts related to spatiality and part-whole knowledge.

Transportation is one of the areas receiving the most attention in the field of AI today. There has been much focus in the area of self-driving cars in order to minimize human involvement in driving and make driving more efficient. Autonomous vehicles would also would be of tremendous help for people without driving licenses or physical handicaps which prevent them from driving. Hence, this is one of the most challenging areas in Smart Mobility, as a reasonably effective autonomous vehicle should have the capability of distinguishing different objects and make split second decisions at times of danger. Existing object detection models may have some identification capabilities but do not possess abilities for commonsense reasoning. [44] In 2016, a Tesla vehicle collided with a truck as it mistook a truck for an overpass, due to the object detection system finding similar attributes between a truck and an overpass. A machine with Commonsense reasoning capabilities would have been able to distinguish between these two objects and prevent the accident. An example of this type of knowledge is that we know that an overpass is immobile while trucks are mobile.

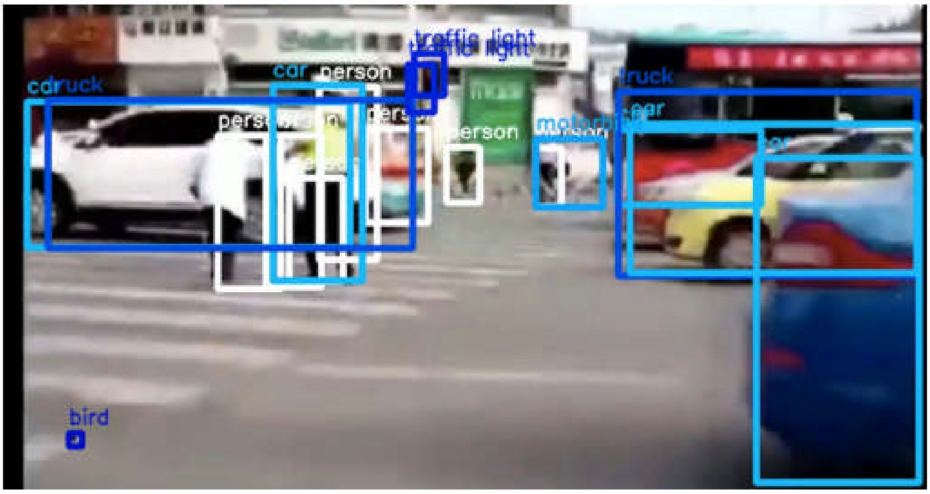

Figure 19: Object detection in Real Time Traffic [45]

Figure 19 shows a real time view of traffic in a busy street, where a slight miscalculation resulting in an incorrect object classification can be hazardous.

In another piece of research [45], a method to tackle such a challenge was proposed where VISIR (VIsual and Semantic image label Refinement) labels [54] were used. The technique of VISIR had three steps:

i) remove incoherent tags among visual labels

ii) extend tag spaces by adding visually similar tags missed by object detectors

iii) joint inference on enriched tag spaces, by integer linear programming.

VISIR labels serve to aid autonomous driving with respect to identifying vehicles and abstraction, which refers to situations such as transportation, traffic jam and rush hour. The use of object existence property of WebChild also helped to integrate this methodology into autonomous vehicles.

Hence, it is important for transportation systems to be able to make well-informed decisions in order to promote safety, energy efficiency and cost effectiveness. Combining CSK with neural models as well as deep learning can help address different problems related to object detection. Another example of enhancing Smart Mobility is increasing efficiency of street lights in places such as Amsterdam.

6. Conclusions

6.1 Summary

In this thesis, we have presented a contribution to the area of Smart Governance by providing ways to use data driven decisions in this area. We first proposed a mapping technique which maps ordinance and tweets to Smart City Characteristics (SCCs). We then enhanced the original technique to allow for mapping of ordinances and tweets to multiple SCCs. We also conducted a sentiment analysis of tweets and observed people's reactions towards different SCCs. Finally, we developed a three-step mapping approach called TOLCS to link individual ordinances to tweets while reducing the complexity of the mapping significantly. After submitting the results of our experiments to external evaluators, we saw that our approaches satisfactorily accomplish their respective goals. We believe that our work lays the groundwork for further research in the area of Data Mining and Machine Learning, stimulating more AI tools in general.

6.2 Technical Contributions

The research involved in this work has been selected for publication at three different conferences:

i) Mapping Ordinances and Tweets using Smart City Characteristics to Aid Opinion Mining -The World Wide Web (WWW) Conference in April 2018 [53]

ii) Smart Governance Through Opinion Mining of Public Reactions- The IEEE International Conference on Tools with Artificial Intelligence (ICTAI) 2018 in November 2018.[52]

iii) Pragmatics and Semantics to Connect Specific Local Laws with Public Reactions- The IEEE Big Data Conference in Seattle, WA in December 2018.[51]

With the rise in interest in mining data from social media websites, the area of Smart Governance has not been given as much attention. Our work in this area aims to help law makers promote transparency and data driven decision making in enacting ordinances. The technical contributions from this work are as follows:

1. Introducing a mapping technique to map ordinance and tweets to multiple SCC.

2. Conducting sentiment analysis to gauge people's reactions towards tweets related to different SCC and presenting a summary.

3. Building a Knowledge Base (KB) for SCC and ordinances using different sources of Commonsense Knowledge to be used for semantic mapping.

4. Developing the TOLCS approach for mapping individual ordinances to tweets. This involved a nontrivial task of significantly reducing the complexity of mapping from polynomial time to linear time.

5. Evaluating the performance of the mapping techniques in order to depict its effectiveness in aiding lawmakers in data driven decision making.

6.3 Future Work

Our work represents an interesting solution for mapping ordinance and tweets to SCCs as well as linking ordinances to tweets for gauging their public reactions, especially from a Smart City angle. Further work in this study includes performance improvements to handle larger quantities of data as well as improving accuracy of different mapping techniques. This also includes achieving a finer level of granularity for analysis of ordinance reactions within each Smart City Characteristic (for example, Smart Environment could imply ordinances related to features such as pollution control, wildlife preservation etc. and these could be analyzed in terms of their relevance with that respective feature, along with public reactions). This would involve exhaustive study in the area of Data Mining as well as Natural Language Processing. It would foster collaborations among different researchers, including those in Urban Studies, Environmental Science and related areas in order to optimize the existing techniques and develop new approaches for solving various sub-problems as needed. The work can also be extended to other applications such mapping news to tweets as well as conducting historical analysis of tweets analogous to our current work.

This research activity on the whole encourages further research in the area of Big Data, NLP and Sentiment Analysis. We believe aiding lawmakers in making data driven decisions will help in improving transparency in governance and will help their cities grow closer to becoming a Smart City. In general, this work also makes a positive impact on sustainability.

7. References

1. TU Wien (Vienna University of Technology). 2015. European Smart Cities, Technical Report. Technical Report. Vienna, Austria.

2. Smart City Consortium. Background.
https://smartcity.org.hk/index.php/aboutus/background

3. 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

4. https://www.statista.com/statistics/303681/twitter-users-worldwide/

5 .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

6. IEEE. 2018. IEEE Smart Cities Technical Community. (2018). Retrieved March 4,2018 from https://smartcities.ieee.org/

7. 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

8. 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

9. Linlin Wang, Kang Liu, Zhu Cao, Jun Zhao, and Gerard de Melo. 2015. Sentiment Aspect Extraction based on Restricted Boltzmann Machines. In Proceedings of ACL 2015. 616–625

10. Niket Tandon, Gerard de Melo, Fabian Suchanek, and Gerhard Weikum. 2014. WebChild: Harvesting and Organizing Commonsense Knowledge from the Web. In Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14).

11. 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, 115–120. http://aclweb.org/anthology/P17-4020

12. 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

13. 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

14. G. A. Miller, "Wordnet: A lexical database for English," Communications of the ACM, vol. 38, no. 11, pp. 39–48, 1995

15. N. Y. C. Council, "Legislative research center web page." http://legistar.council.nyc.gov/, 2018

16. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems, 2013

17. Guerini M, Gatti L, Turchi M. Sentiment analysis: how to derive prior polarities from SentiWordNet. In: Proceedings of empirical methods in natural language processing (EMNLP); 2013. p. 1259–69.. 2013. http://arxiv.org/abs/1309.5843

18. 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.

19. 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

20. 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

21. 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.

22. 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 Web6 (LDOW 2016).

23. 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.

24. 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

25. 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

26. 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.

27. 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

28. 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

29. Gerard de Melo. 2017. Multilingual Vector Representations of Words, Sentences, and Documents. In Proceedings of IJCNLP 2017. http://www.aclweb.org/anthology/I17-5002

30. 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

31. 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

32. 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

33. 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

34. 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.

35. 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 proft distribution. Computers Operations Research 87 (2017), 245 – 257. https://doi.org/10.1016/j.cor.2017.02.015

36. 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.

37. B. Nunberg, Re-thinking civil service reform: An agenda for smart government: World Bank, 1997.

38. R. Giffinger, C. Fertner, H. Kramar, R. Kalasek, N. Pichler-Milanović, and E. Meijers, "Smart Cities: Ranking of European Medium-Sized Cities," Centre of Regional Science (SRF), Vienna University of Technology, Vienna, Austria2007.

39. R. G. Hollands, "Will the real smart city please stand up?," City, vol. 12, pp. 303-20, 2008.

40. R. Giffinger, C. Fertner, H. Kramar, R. Kalasek, N. Pichler-Milanović, and E. Meijers, "Smart Cities: Ranking of European Medium-Sized Cities," Centre of Regional Science (SRF), Vienna University of Technology, Vienna, Austria2007

41. S. Alawadhi, A. Aldama-Nalda, H. Chourabi, R. J. Gil-Garcia, S. Leung, S. Mellouli, T. Nam, T. Pardo, H. J. Scholl, and S. Walker, "Building Understanding of Smart City Initiatives," in Electronic Government. vol. 7443, H. J. Scholl, M. Janssen, M. Wimmer, C. Moe, and L. Flak, Eds.: Springer Berlin / Heidelberg, 2012, pp. 40-53.

42. H. Chourabi, T. Nam, S. Walker, J. R. Gil-Garcia, S. Mellouli, K. Nahon, T. A. Pardo, and H. J. Scholl, "Understanding smart city initiatives: An integrative and comprehensive theoretical framework," in 45th Hawaii International Conference on System Sciences, Maui, Hawaii, 2012, pp. 2289-2297.

43. T. Nam and T. A. Pardo, "Conceptualizing smart city with dimensions of technology, people, and institutions," in 12th Annual International Conference on Digital Government Research (dg.o 2011), College Park, MD, USA, 2011, pp. 282–291

44. Commonsense for Machine Intelligence: Text to Knowledge and Knowledge to Text. Gerard de Melo, Niket Tandon, Aparna S. Varde. CIKM, 53

45. Abidha Pandey, Manish Puri, Aparna S. Varde: Object Detection with Neural Models, Deep Learning and Common Sense to Aid Smart Mobility. ICTAI 2018: 859-863

46. Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In AKBC, 2013

47. A. Singhal: Introducing the Knowledge Graph: Things, Not Strings. googleblog.blogspot.co.uk, 16 May 2012.

48. IBM Journal of Research and Development 56(3/4), Special Issue on “This is Watson”, 2012.

49. IEEE Smart Cities https://smartcities.ieee.org/home/sitemap

50. IEEE Smart Cities. https://www.ieee-pes.org/pes-communities/ieee-smart-cities

51. Manish Puri, Aparna Varde and Boxiang Dong --- Pragmatics and Semantics to Connect Specific Local Laws with Public Reactions -- In IEEE Big Data Conference, Seattle, WA, Dec 2018, pp. 5433-5435.

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

53. 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), Apr 2018, Lyon, France, pp. 1721 - 1728.

54. S. Nag Chowdhury, N. Tandon, H. Ferhatosmanoglu, and G. Weikum, “VISIR: Visual and semantic image label refinement,” in ACM WSDM conference, 2018, pp. 117–125.

55. Alawadhi, S., & Scholl, H. J. (2016). Smart Governance: A Cross-Case Analysis of Smart City Initiatives. In 2016 49th Hawaii International Conference on System Sciences (HICSS) (pp. 2953- 2963). IEEE.

56. Emebo O., Varde A., Olawande D., “Common sense knowledge, text mining and ontology for implicit requirements”, DMIN, Las Vegas, NV, Jul 2016, pp. 146 – 152.

57. de Jong M., Joss S., Schraven D., Zhan C., Weijnen M. ‘‘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.

58. Torralba, A., Murphy, K., Freeman, W.T., ‘‘Using the Forest to see the Trees: Object Recognition in Context’’, Communications of the ACM, Mar 2010, 53(3):107-114.

59. Introduction to WordNet: An On-line Lexical Database. August 1993. http://wordnetcode.princeton.edu/5papers.pdf

60. WordNet Search 3.1 http://wordnetweb.princeton.edu/perl/webwn

61. Douglas B. Lenat, R. V. Guha, Karen Pittman, Dexter Pratt, and Mary Shepherd. 1990. Cyc: toward programs with common sense. Commun. ACM33, 8 (August 1990), 30-49. DOI=10.1145/79173.79176 http://doi.acm.org/10.1145/79173.79176

62. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: a nucleus for a web of open data. In Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference (ISWC'07/ASWC'07),. Springer-Verlag, Berlin, Heidelberg, 722-735

63. WebChild CommonSense Browser. https://gate.d5.mpi-inf.mpg.de/webchild/