Lang Resources & Evaluation
https://doi.org/10.1007/s10579-022-09584-6

ORIGINAL PAPER

Commonsense based text mining on urban policy

Manish Puri[1,2] · **Aparna S. Varde**[3,4] · **Gerard de Melo**[5,6]

Accepted: 26 January 2022

Abstract Local laws on urban policy, i.e., *ordinances* directly affect our daily life in various ways (health, business etc.), yet in practice, for many citizens they remain impervious and complex. This article focuses on an approach to make urban policy more accessible and comprehensible to the general public and to government officials, while also addressing pertinent social media postings. Due to the intricacies of the natural language, ranging from complex legalese in ordinances to informal lingo in tweets, it is practical to harness human judgment here. To this end, we mine ordinances and tweets via reasoning based on commonsense knowledge so as to better account for pragmatics and semantics in the text. Ours is pioneering work in ordinance mining, and thus there is no prior labeled training data available for learning. This gap is filled by commonsense knowledge, a prudent choice in situations involving a lack of adequate training data. The ordinance mining can be beneficial to the public in fathoming policies and to officials in assessing policy effectiveness based on public reactions. This work contributes to smart governance,

✉ Aparna S. Varde
vardea@montclair.edu

Manish Puri
manishpuri319@gmail.com

Gerard de Melo
gdm@demelo.org

[1] Allstate Insurance Company, Northfield Township, AZ, USA

[2] Department of Computer Science, Montclair State University, Montclair, NJ, USA

[3] Department of Computer Science, and Environmental Science & Management, PhD Program, Montclair State University, Montclair, NJ, USA

[4] Visiting Researcher at Max Planck Institute for Informatics, Saarbrücken, Germany

[5] Artificial Intelligence & Intelligent Systems, Hasso Plattner Institute, Potsdam, Germany

[6] Rutgers University, New Brunswick, NJ, USA

Published online: 25 February 2022

leveraging transparency in governing processes via public involvement. We focus significantly on ordinances contributing to smart cities, hence an important goal is to assess how well an urban region heads towards a smart city as per its policies mapping with smart city characteristics, and the corresponding public satisfaction.

Keywords Commonsense reasoning · Opinion mining · Ordinances · Smart cities · Social media · Text mining

1 Introduction

1.1 Background and motivation

While there are ample resources that aid in interpreting national legislation, there are also ever-increasing amounts of local laws or ordinances affecting the lives of people in various direct ways. These include urban policy relating to issues such as health, business etc. as depicted in the excerpt in Fig. 1.

Although such public policy has become more easily retrievable via the web, often only experts possess the ability to locate pertinent documents and make sense of their contents. To most people, these repositories tend to seem impervious and complex. Even the experts, however, typically lack the ability to monitor more macroscopic trends, both in the actual policy and its public reaction. To address these important challenges, it is important to provide novel ways for both the public and for experts to peruse and analyze local ordinance data. In order to achieve this, it is beneficial to discover knowledge of local laws (ordinances) by mining relevant ordinance data collected from public websites. At the same time, public reactions to such policies may be obtained by analyzing social media data, such as tweets, as illustrated in Fig. 2. This brings us to the goals of our work.

1.2 Goals

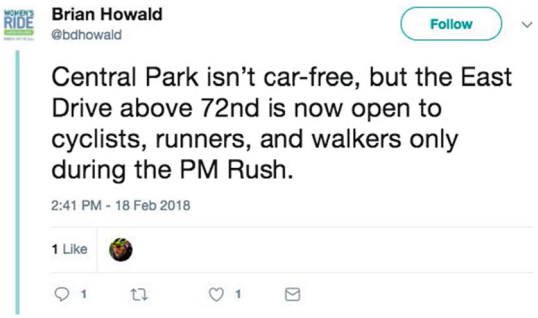
Our work focuses on the area of urban policy with broader implications in the field of smart governance and smart cities in general. A principal goal of this project is to conduct an analysis of ordinances by analyzing the reactions to them observed on social media platforms. To enable this, a key challenge is to link public ordinances to tweets by taking into account their semantic connections. This is a non-trivial endeavor, since traditional machine learning techniques are not applicable due to a lack of training data, and using simple keyword matching techniques is not feasible due to the diversity of natural language in a vast number of ordinances and tweets, wherein the former involve complex legalese, while the latter encompass informal lingo, abbreviations etc. Our problem addressed in this article therefore involves defining an approach to conduct ordinance and tweet mining, which entails first mapping the ordinances and tweets, so as to ultimately assess to what extent the concerned public is satisfied with urban policy as well as to assess to what extent the given region heads towards becoming a Smart City. Important desiderata include a low computational complexity, and capturing semantics and pragmatics.

§ 2. Chapter one of title 19 of the administrative code of the city of New York is amended by adding a new section 19-132.1 to read as follows:

§ 19-132.1 Restrictions on the central park loop drive. a. The central park loop drive shall be closed to motor vehicle traffic between June 24, 2018 and September 25, 2018.

Fig. 1 Excerpt of NYC Int. 0363-2018

Fig. 2 Tweet example with reference to NYC policy



1.3 Proposed approach

We propose the use of Commonsense Knowledge (CSK) (Tandon et al., 2017) to capture human judgment in mining ordinances and tweets. It is to be noted that ours is among the pioneering works in ordinance mining to the best of our knowledge, hence huge training data sets are not available for suitably deploying machine learning techniques. In order to bridge this gap, we aim to conduct reasoning guided by commonsense knowledge, thus harnessing (in a manner analogous to humans) the semantics and pragmatics of the text, including complex legalese of ordinances and informal lingo in tweets. CSK is deemed a good choice in situations involving a dearth of training data (Razniewski et al., 2021). We present the details of this approach in Sect. 3. Much of the urban policy we address in this work targets *smart cities*, a term that is no longer just a popular buzzword today in many urban regions. Smart cities gear towards *sustainability*, which is an increasingly important paradigm in the 21st century, helping to combat issues such as climate change and global warming. Insights gathered from the mining conducted in this work can thus be used to assess the readiness of an urban area with respect to several relevant criteria (Wien, 2015) for smart cities. Based on the overall results of this analysis, we aim to provide feedback to urban agencies on how well they are doing in urban policy management, where they need improvement, how closely they head towards smart city goals and how their policies are perceived by the public.

1.4 Contributions

In general, this research makes the following contributions.

1. It is among the first of its kind to conduct in-depth data mining on urban policy, particularly in conjunction with smart cities.
2. It mines the public reaction to diverse urban policy decisions through social media postings.
3. It deploys reasoning guided by commonsense knowledge to simulate human judgment in the mining processes.
4. It creates knowledge bases pertaining to smart city characteristics as well as ordinance departments guided by commonsense knowledge, the process and products of which can be reusable in other suitable applications.

1.5 Article layout

The rest of this article is organized as follows. Section 2 provides a literature survey of related work in the area. Section 3 describes the details of our proposed methodology for urban policy mining guided by commonsense knowledge. Section 4 provides excerpts from our experimental evaluation with discussion. Section 6 states the conclusions along with potential future work.

2 Related work

2.1 Commonsense knowledge and reasoning

There is much research on commonsense knowledge and reasoning from classical to state-of-the-art. Cyc (Lenat et al., 1990) represents one of the earliest projects in this area. Cyc includes a large-scale axiomatic knowledge base consisting of millions of assertions with 300,000+ concepts, many of which relate to CSK. Apart from the KB, it also includes a reasoning mechanism and a system for NLP. ConceptNet is a large-scale semantic network to aid AI systems in comprehending the meanings of words and relationships between concepts. The project was initiated at MIT Media Lab in 1999 within the crowdsourcing project Open Mind Common Sense and later was extended to encompass knowledge from other crowdsourced sites, expert-created resources, and games with a purpose (Speer et al., 2016). While these projects represent some of the earliest attempts at building commonsense knowledge bases, with further enhancements following later, there are notable challenges in terms of scaling up such knowledge acquisition.

Aristo TupleKB is among the more recent efforts dealing with knowledge extraction on simple concepts related to 4th grade science (Mishra et al., 2017). It includes 290,000+ relevant knowledge tuples and provides reasonably high precision and coverage. A system called Quasimodo extracts commonsense concepts from non-standard web sources such as query logs in search engines

and Question-Answering (QA) forums (Romero et al., 2019). It addresses typicality, salience and meaningfulness of concepts in addition to precision and coverage. In the project Distribution over Quantities (DoQ) (Elazar et al., 2019), an unsupervised collection technique gathers large amounts of quantitative data from the web to create a repository with quantitative information on nouns, adjectives and verbs. Recent research presents DICE (Chalier et al., 2020) a multifaceted model of CSK that emphasizes concept properties of plausibility, typicality, remarkability and salience in everyday concepts. There is further research on the connection between commonsense knowledge and concept-level sentiment analysis, most notably the SenticNet resource (Cambria et al., 2020).

In the past few years, efforts at commonsense knowledge acquisition have often drawn on language models to better generalize from the specific observations encountered on the web (Davison et al., 2019; Tandon & de Melo, 2010). However, it has been found that even modern large language models still have notable shortcomings and are best consulted after fine-tuning them using manually constructed commonsense knowledge. In particular, COMET (Bosselut et al., 2019) fine-tunes Transformers based on triples taken from ConceptNet, while a more recent version (Hwang et al., 2020) is fine-tuned on a carefully curated set of triples obtained using crowdsourcing as well as by filtering ConceptNet.

Apart from these advances in knowledge acquisition, there have also been substantial advances in commonsense reasoning, often drawing on subsymbolic deep neural approaches or neuro-symbolic approaches. Neuro-symbolic systems that are capable of approximate deductive reasoning can open possible investigations into combining deductive and inductive reasoning, as well as commonsense reasoning within a single system (Hitzler et al., 2019).

Commonsense reasoning challenges include the Winograd Schema Challenge (WSC), originally proposed as an alternative to the Turing Test using challenge sentences such as "The guitar does not fit in the bag since *it* is too large" *(the guitar)* vs. "The guitar does not fit in the bag since *it* is too small" *(the bag)*. Yet, advances in neural models give over 90% accuracy on WSC variants. WinoGrande (Sakaguchi et al., 2020) is a benchmark dataset extending this idea to around 44,000 problems, enhancing both the scale and difficulty. The ARC (Clark et al., 2018) dataset provides questions on grade-school science and is the largest public-domain set of its kind (7787 questions). In recent years, a number of benchmark datasets have been proposed specifically to assess to what extent deep neural models are able to capture commonsense knowledge. Examples include the SemEval 2020 Commonsense Validation and Explanation (ComVE) (Wang et al., 2020) task, the CommonsenseQA benchmark (Talmor et al., 2021), and the COM2SENSE benchmark (Singh et al., 2021).

The above lines of work seek to mine commonsense knowledge and address particular kinds of benchmark tasks that assess to what extent commonsense knowledge has been acquired. The focus of our work is instead to exploit sources of machine-readable commonsense knowledge to better relate social media postings to urban policy. To this end, our work draws on two pre-existing databases of commonsense knowledge, WebChild (Tandon et al., 2014) and WordNet (Miller & Fellbaum, 1998), which we introduce later in Sect. 3.2.

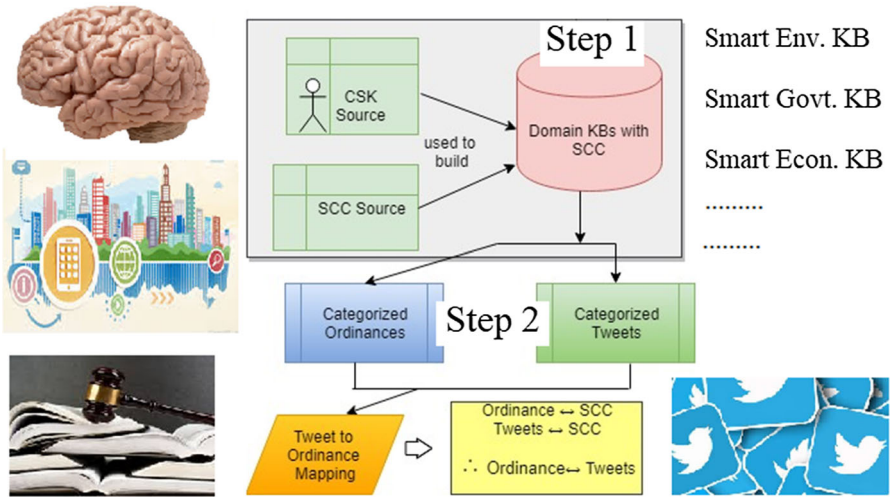


Fig. 3 Overview of the CSK-SCC approach (high-level mapping of ordinances and tweets)

2.2 Social media mining

The increased accessibility of social media has led to a proliferation of studies on social media analytics. Unsupervised topic modeling and trend detection (Jayadharshini et al., 2018) are among the most well-known subareas within social media mining. For example, one study (Shams et al., 2020) uses hashtags to analyze gaming trends on Twitter via a Long Short-Term Memory (LSTM) network trained on Twitter data. Some studies also touch on the area of urban policy, e.g. efforts to draw on social media for air quality assessment (Du et al., 2016) and environmental management (Du et al., 2019). However, techniques such as the above cannot be applied to the task of mapping tweets to ordinances, which are two sets of heterogeneous items.

Another notable stream of research focuses on link prediction in social networks (Cao et al., 2018; Wang et al., 2017). However, most such efforts involve structured data such as connections between users. In contrast, our work on mapping ordinances to tweets entails connecting two forms of unstructured data. Ordinances generally use formal language, often with intricate legal terms, while tweets tend to be very informal, with hashtags, URLs and emojis (Shoeb et al., 2019).

Tweet classification with external entity knowledge is provided by TweetSift (Li et al., 2016), which categorizes tweets into different topics so as to give more context to tweets and combine them with word embeddings from different topics. The system is supervised, however, and thus the technique cannot be readily adopted for our project on mapping ordinances to tweets due to a lack of training data.

Relevant related work includes a study on a semi-automatic annotation of Twitter content for the promotion of citizen engagement (Alkhammash et al., 2019). Another study (Rose & Willis, 2019) involving Twitter and smart cities draws on

the visualization software ImagePlot to create a visualization of 9030 tweeted images related to smart cities. To the best of our knowledge, no prior work has focused on mining ordinances and their social media reactions on Twitter. This is precisely the problem addressed in our work on discovering knowledge from the complex legalese in urban policy and informally expressed tweets about them, both constituting different aspects of natural language.

3 Methodology for urban policy mining with CSK

We propose an approach for mapping ordinances and tweets based on SCCs guided by commonsense knowledge (CSK), the big picture of which is illustrated in Fig. 3. We refer to this as the CSK-SCC approach for high-level mapping of ordinances and tweets. We build this high-level approach, described in more detail in Sect. 3.3, to gauge the effectiveness of such a CSK-guided reasoning leveraging SCCs. This is significant in order to proceed further with the fine-grained linkage of the individual ordinances and tweets, which is described later in Sect. 3.4.

The inventory of SCCs $\mathcal{C}$ used for reference here are widely accepted ones from sources in the literature Wien (2015). These include *smart environment*, *smart mobility*, *smart governance* etc. This high-level mapping approach for urban policy mining is motivated by the fact that both ordinances and tweets can be unrestricted (infinite) sets with intricate natural language, comprising complex legalese and informal lingo, respectively. Mapping them to each other directly appears challenging without relevant prelabeled training data, which unfortunately is not available for this task. SCC sources are limited and hence can be considered appropriate to establish a nexus between ordinances and tweets. Additionally, an important goal is to assess how well the concerned urban region heads towards a smart city, by discovering how ordinances address SCCs and to what extent the public is satisfied (based on tweets).

In light of the above reasoning, we invoke a simple transitive property $P1$ as follows.

Property $P1$: If an ordinance O_i semantically relates to the SCC C_k and a tweet T_j relates to the same SCC C_k then O_i bears some connection to T_j. Hence, $(O_i \mapsto C_k) \wedge (T_j \mapsto C_k) \Rightarrow (O_i \mapsto T_j)$.

In order to find this semantic relatedness, we harness commonsense knowledge (CSK) from sources such as WebChild (Tandon et al., 2014) (a large repository of CSK facts, properties, and relationships derived from the web) as well as WordNet (Miller & Fellbaum, 1998) (a large lexical source of nouns, verbs, adjectives and adverbs grouped as synonym sets, expressing distinct CSK concepts with semantics). It is crucial to harness such commonsense sources in this work because terms from ordinances and tweets may not directly exist in descriptions of SCCs, and hence CSK is beneficial for establishing the connection. For example, consider the following ordinance snippet: *"... carbon monoxide detectors must be installed in every apartment as per NY state law"*. While a human would intuitively be able to connect this with the SCC "smart environment", this is not obvious in an

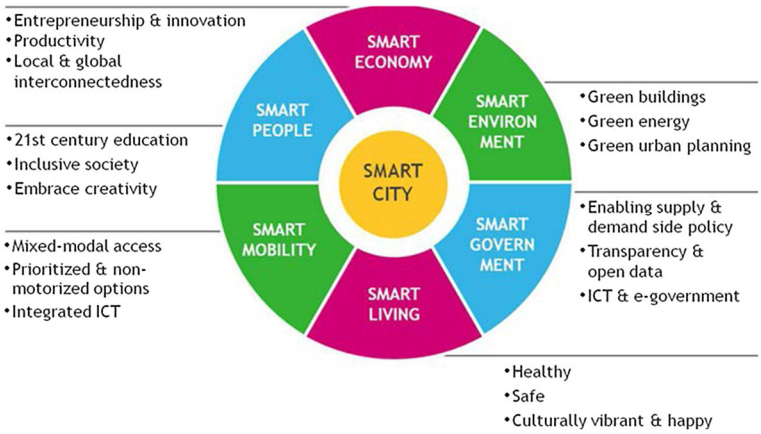


Fig. 4 Overview of some widely accepted SCCs, adapted from TU Vienna (Wien, 2015)

unsupervised automated mapping, since the terms in this ordinance snippet may not directly appear in the descriptions of the corresponding SCC in sources from the literature. Hence, CSK is useful here. Our CSK-guided reasoning is thus based on two steps:

1. Construct SCC Domain KBs using CSK sources and SCC sources
2. Use these SCC KBs for reasoning to map ordinances with tweets using transitive property $P1$ and its implications

Before proceeding with further details on our proposed approach, we first provide an overview of urban policy and smart cities along with a brief description of WebChild and WordNet.

3.1 Urban policy and smart cities

The term "smart cities" refers to urban regions (not just metropolitan areas) around the world using technological advancements to improve the quality of life among their residents. People in smart cities have a common objective of promoting advanced technology and efficient use of energy, while maintaining transparency in the government. Over the past few decades, rampant urbanization in different parts of the world has given rise to several challenges for improving the quality of life for citizens (Shahidehpour et al., 2018). The population growth in urban areas may exacerbate challenges such as increased pollution, increase in crime and traffic congestion. Accordingly, there has been increasing global interest in promoting involvement in smart cities. For instance, the IEEE organization has several Smart Cities initiatives (The IEEE Smart Cities Technical Community, 2018) to help cities around the globe address challenges when migrating towards different aspects related to urbanization. This is especially important due to the growing population of the world, which according to a recent report by the United Nations, is expected

to reach 9.3 Billion by 2050 (United Nations: Department of Economic and Social Affairs: Population Division, 2019). In a smart city, data stemming from various sources is used to create a better approach to deal with urban challenges of the aforementioned sorts and manage efficient allocation of resources. An example of this is the deployment of the ATAK (Full Adaptive Traffic Management System) Signal Control system in Istanbul, Turkey, which relies on data-driven algorithms to optimize traffic signal timings (Gundogan, 2015). Smart city characteristics (SCCs) represent the building blocks of such endeavors (Wien, 2015). A visual representation of a sample of smart city characteristics is given in Fig. 4. Note that the illustrated list is not exhaustive, for example some systems may consider "smart health", "smart manufacturing" and so on as further SCCs. Regardless of the particular inventory $\mathcal{C}$ of SCCs invoked, our overall methodology remains the same.

3.2 Main CSK sources harnessed in this research

3.2.1 WordNet

One of the most well-known lexical knowledge bases available is WordNet. (Miller & Fellbaum, 1998), originally created at Princeton University. WordNet can be viewed as a semantic network of groups of near-synonymous words or phrases, known as synonym sets or synsets for short, which are linked by lexical and conceptual relationships. The inventory of synonym sets distinguishes not only different parts of speech but also different senses of a word. For example, the word *fast* can denote the act of abstaining from food (a verbal synset) or the property of moving quickly (an adjective synset). The repository consists of over 100,000 synsets, including over 80,000 noun synsets, approximately 20,000 adjective synsets, approximately 14,000 verb synsets and about 3600 adverb synsets. The relationships include hypernymic links that connect specific fine-granular concepts to more general concepts. Another CSK-related relationship is that of meronymy, i.e., part–whole relationships (Miller & Fellbaum, 1998). However, WordNet was not designed to serve as an extensive compilation of CSK, so incorporating further sources is useful to obtain more diverse kinds of CSK.

3.2.2 WebChild

The WebChild repository (Tandon et al., 2014, 2017) provides commonsense related knowledge pertaining to the properties of objects and relationships to other objects. In recent years, there have been a number of large-scale knowledge bases, such as Google's Knowledge Graph (Singhai) and DBpedia, but the information they provide is largely encyclopedic and factual. In contrast, WebChild serves to fill the gap by focusing on less crisp commonsense knowledge.

Methodologically, it is based on text extraction along with graph algorithms (Tandon et al., 2014). One of the sources it uses is the Google Web search 5-g corpus, an n-gram corpus sourced from approximately 1 trillion words with manual design patterns involving approximately 20 copula verbs. The repository has over 4

million triplets which define properties of relational mapping between nouns and adjectives such as *hasProperty* and *hasSubstance* for objects at a granular level. The repository consolidates information by making use of label propagation on graphs when defining a domain set and range set. The graphs link nouns with WordNet senses and adjectives by using weighted edges, and edge weights are computed based on statistics on how they are interrelated. The learning propagation algorithms are invoked using WordNet as well as Web data, and are used to determine connections between noun senses and adjective senses (Tandon et al., 2017).

In addition to the above two knowledge sources, we also use SentiWordNet and VADER for sentiment analysis, which we describe later in Sect. 3.5.

3.3 Details of CSK-SCC approach: high-level mapping

3.3.1 SCC KB development guided by CSK

The first step in our proposed approach involves using sources of CSK and SCC in the literature to build knowledge bases specific to the different domains pertaining to the smart city characteristics. These are referred to as *Domain KBs with SCC* (or *SCC KBs*). The SCC sources include technical reports, websites and other documents, e.g., the IEEE Smart Cities website, and the TU Wien Technical Report (Wien, 2015). Using CSK sources such as WebChild, we are able to derive common sense concepts based on Web data. These concepts are in the form of objects with properties and relationships. The CSK sources are used to create text-based SCC KBs that harvest words related to various SCCs derived from these CSK repositories with the use of NLP techniques such as stemming and lemmatization (Stanford University, 2021), as well as text matching based on semantics. We further expand the content of the KBs via knowledge based text extraction (Tandon et al., 2011) as well as finding relationships between terms in large datasets based on association rules, i.e., association rule mining (Solanki & Patel, 2015).

While conducting the association rule mining, we gather texts from sources such as ordinance websites, where each line of text acts as a transaction. After cleaning each line of text to retain only nouns, we apply the Apriori algorithm for frequent item set mining to induce a set of rules, which we sort by confidence in decreasing order. Below are some examples of rules derived with reference to the word "economy" extracted from WebChild, and the web-based sources pertaining to the SCC "smart economy".

1. stocks ⇒ economy
2. taxes ⇒ income
3. insurance ⇒ health
4. startup ⇒ entrepreneur
5. cost of living ⇒ salary
6. jobs ⇒ economy
7. profits ⇒ revenues

Likewise, we extract dozens of association rules for each SCC. The overall process is outlined in Algorithm 1.

ALGORITHM 1: CSK-SCC Rule Mining

1. **for each** SCC $C_i \in \mathcal{C}$ **do**:
1. $L_i = \text{preprocess}(\text{texts}(C_i))$
2. $R_i = \text{apriori}(L_i, \text{support, confidence, min\_len=2, max\_len=6})$
3. $R_i^{\text{S}} = \text{sort}(R_i, \text{by} = \textsc{Support}, \text{descending} = \textsc{True})$
4. $O_i = \text{topk}(R_i^{\text{S}}, \text{k})$
5. **return** $(O_1, \ldots, O_n)$

Figure 5 illustrates our process for creating SCC KBs. We provide an example of populating a subset of the *smart economy KB*. We start with text from a smart city source, i.e., the TU Wien Technical Report on *European Smart Cities* (Wien, 2015). In the description for the SCC *smart economy*, the following concepts are enumerated.

Smart economy
– Innovative spirit
– Entrepreneurship
– Economic image and trademarks
– Productivity
– Flexibility of labour market
– International embeddedness
– Ability to transform

Given such information, we search the CSK sources to find matching descriptions of concepts. For instance, looking up the word "economy" in WebChild, we obtain *comparables* (Tandon et al., 2017) such as "Wall Street" , "market" and "indicator" as well as *activities* (Tandon et al., 2017) such as "create economy" and "practise economy". We include these in our intermediate KB for *smart economy*. We further apply association rule mining to obtain related terms such as "enterprise", "commercialize", "monetary value", and "inflation". These are then compiled into the final *smart economy KB*. Likewise, we proceed with other terms in the description of *smart economy* above, such as "innovative spirit", "entrepreneurship" etc. to search WebChild and WordNet, thereby populating the intermediate KB, and further apply association mining to populate the final KB. As per this process, we derive numerous relevant terms to create the *smart economy KB* such as the obvious "innovative", "entrepreneur", "productivity" and less obvious yet pertinent ones, e.g. "proletariat", "trade union movement", "enterpriser" etc. A similar process is applied for each SCC in the SCC inventory to build a corresponding SCC KB. Due to the large size of the KBs, we observe a few

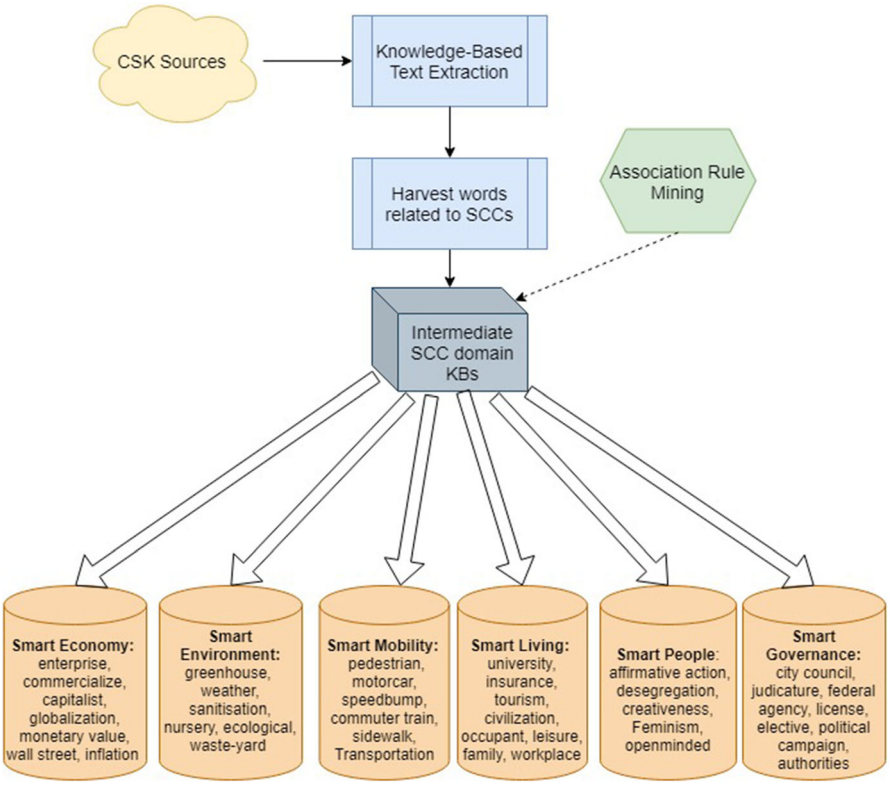


Fig. 5 Creation of SCC KBs based on CSK (showing a subset of concepts)

overlapping concepts between them, e.g. ”energy efficiency” can be attributed to both “smart environment” as well as “smart economy”.

The SCC KB creation is a one-time process, while the SCC KBs can then be used recurrently for ordinance and tweet mining in any contexts. They can also be used for other research (outside our own work) relevant to smart cities. Hence, this part of our effort makes contributions to language resources as a whole. We create 6 SCC KBs here since we consider 6 commonly used SCCs from sources in the literature (Wien, 2015), in particular *smart economy*, *smart environment*, *smart living*, *smart mobility*, *smart people* and *smart governance*. In practice, the same process can be applied to further SCCs. Thus, if there is the need to include others such as smart energy, smart health, smart lighting, smart manufacturing etc., the process can be repeated to obtain further KBs. Some SCCs may exhibit topical overlap, e.g. smart energy and smart environment could both cater to greenness, while smart living and smart mobility could both entail mobile app development. Thus, as explained later, our mapping procedure explicitly accounts for this phenomenon.

3.3.2 Mapping using KBs along with transitive properties

Once the SCC KBs are built, they serve as the basis for reasoning guided by CSK in order to map ordinances with tweets entailing the smart city angle. Note that an ordinance/tweet can potentially have a many-to-many mapping relationship with SCCs, i.e., one ordinance/tweet can map to multiple SCCs (and obviously one SCC can map to multiple ordinances and tweets) (Puri et al., 2018). Our mapping thus provides a multiple linkage process.

Leveraging subjective human judgment in mapping ordinances and tweets deserves some consideration when they contain terms relevant to multiple SCCs. In order to address this issue, we translate the subjective judgment of humans into numeric scores as follows. We assign weights for different terms in ordinance/tweets pertaining to SCCs in proportion to their frequency of occurrence. For example, an ordinance/tweet may have 1 term related to SCC C_1 and 2 terms related to SCC C_2. In that case, it would be given a 33% weighting towards C_1 and a 67% weighting towards C_2. The same logic applies to mapping with more than two SCCs. We ignore terms in tweets that are not relevant to any SCCs. During the mapping process, we preprocess the tweet and ordinance texts by eliminating hashtags, punctuation marks, as well as stopwords such as *is*, *are*, *the*, *this*. We then perform stemming (Han et al., 2012) using NLTK, in order to remove morphological affixes from words and map them to a canonical base form, along with lemmatization using the spaCy library in order to map various forms of a word to the common base form of the word. For example, the words *bicycle*, *bicycles*, *bicycle's* and *bicycles'* map to *bicycle*. Since Twitter data contains numerous misspelled words, we additionally invoke a text processing framework (Baziotis et al., 2017) based on word statistics from two big corpora—Wikipedia and Twitter—to allow replacing mispelled words with the most probable candidate words.

Using the results of this preprocessing, we proceed further in order to map ordinances and tweets to each other via multiple SCCs, We make use of the transitive property $P2$ as follows.

Property $P2$: If an ordinance O_i maps to any subset S_a of SCCs $(C_1 \ldots C_m)$ and if a tweet T_j maps to another subset S_b thereof, then the extent of semantic relatedness R between O_i and T_j is directly proportional to the size of the intersection of S_a and S_b. Formally: If $(O_i \mapsto S_a) \wedge (T_j \mapsto S_b)$ then $R(O_i, T_j) \propto |S_a \cap S_b|$, where $S_a, S_b \subset \{C_1 \ldots C_m\}$.

By using this generic concept of semantic relatedness via SCCs guided by CSK, our objective is to create an efficient high-level mapping process initially without the complexities associated with fine-grained connections between tweets and ordinances that are vast in number, entailing intricate natural language. By using this transitive property described above, our mapping algorithm for this high-level SCC classification is shown in Algorithm 2, i.e., the CSK-SCC Mapping Algorithm. The function σ assesses the semantic relevance of KB terms in a given input text.

Therefore, in the process of harnessing human judgment in ordinance–tweet mapping, for every relevant term in an ordinance text or in a tweet text, our SCC

KBs are used to assign the respective smart city characteristic(s) C_j. Hence, the SCC KBs we construct guided by CSK are very important in driving the process of mapping.

ALGORITHM 2: CSK-SCC Mapping Algorithm

1. **for each** SCC $C_i \in \mathcal{C}$ **do**:
2. Establish KB K_i
3. $D \leftarrow \emptyset$
4. **for each** ordinance O_i **do**:
5. **for each** SCC $C_j \in \mathcal{C}$ do:
6. $S_{i,j} \leftarrow \sum_{x \in K_j} \sigma(O_i, x)$
7. $D \leftarrow D \cup \{(O_i, C_j) \mid j = \text{argmax}_{j'} S_{i,j'}\}$
8. **for each** mined tweet T_i:
9. **for each** SCC $C_j \in \mathcal{C}$ do:
10. $M_{i,j} \leftarrow \sum_{x \in K_j} \sigma(T_i, x)$
11. $D \leftarrow D \cup \{(T_i, C_j) \mid j = \text{argmax}_{j'} M_{i,j'}\}$
12. $\theta \leftarrow \{(O_i, T_k) \mid \exists C_j : (O_i, C_j) \in D \wedge (T_k, C_j) \in D\}$
13. **return** θ

3.4 The TOLCS technique for fine-grained CSK-based mapping

After linking both ordinances and tweets to SCCs in a high-level manner as described above, we introduce a technique called TOLCS (Tweet Ordinance Linkage by Commonsense and Semantics) (Puri et al., 2018) for linking individual ordinances to tweets at a more fine-grained level. This is because the high-level mapping provides a good starting point to understand the semantic relatedness between ordinances and tweets via their mutual relevance with SCCs. However, in order to proceed with details of sentiment analysis on tweets, it is important to delve into a deeper linkage. The TOLCS technique is illustrated in Fig. 6.

In TOLCS, we construct ordinance KBs analogous to SCC KBs, guided by CSK. The same CSK sources of WebChild and WordNet are used to guide this KB construction. Ordinance sources can be found online from public websites on local laws, e.g. the New York City public ordinance website. We build a KB per ordinance department, and it is to be noted that some ordinance departments are relevant to the same SCC, e.g. "NYC Department of Finance" and "NYC Department of Economic Development" both refer to "smart economy". Even then, we build a separate KB for each department, so the number of ordinance KBs can be greater than the number of SCC KBs, hence allowing for a more fine-grained mapping at this level. Such ordinance KBs are created similar to SCC KB creation as described earlier, and serve as the basis for deeper level of linkage between ordinances and tweets. We use CSK and text similarity methods to harness pragmatics and semantics from tweets. The ordinance KBs, though larger in number, are smaller in size compared to the SCC KBs, ranging from 70 to 150 keywords, as ordinances have highly structured language, which results in a smaller KB size. The SCC KBs, in comparison, have a size ranging from 200 to 400 keywords. Since finding direct linkage between ordinances and tweets would

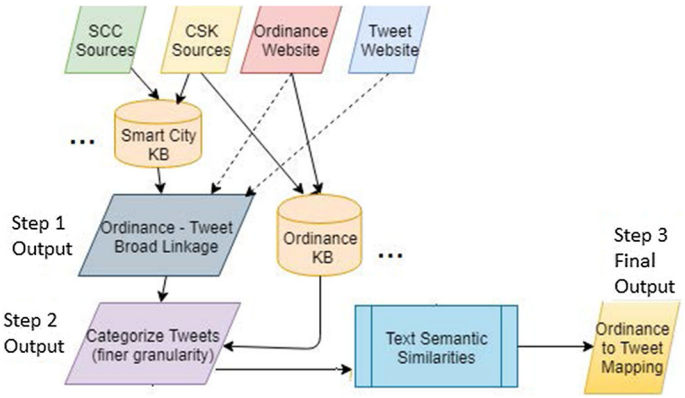


Fig. 6 Steps of TOLCS for fine-grained tweet ordinance linkage

require large datasets with a process of polynomial time complexity, we use blocking steps to reduce linkage complexity to linear time for a more efficient mapping. TOLCS thus entails the following steps.

1. Reduce mapping space by using SCC KBs to find high-level relationships between large groups of ordinances and tweets.
2. Use ordinance KBs to find subsets of ordinances and tweets that relate to each other on a deeper level.
3. Link ordinances to tweets directly in a smaller mapping space with properties of text semantic similarity.

ALGORITHM 3: The TOLCS Linkage Algorithm

Input: θ: threshold for similarity links between ordinances and tweets
Output: $\mathcal{R}$: the set of correlated ordinances and tweets
1. **for each** ordinance O_i:
2. find relevant SCC and ordinance department
3. **for each** tweet T_i:
4. find relevant SCC and ordinance department
5. **for each** Ordinance O_i and Tweet T_j:
6. **if** $\sigma(O_i, T_j) \geq \theta$:
7. $\mathcal{R} = \mathcal{R} \cup \{(O_i, T_j)\}$
8. **return** $\mathcal{R}$

Based on the above explanation, the TOLCS technique is described in Algorithm 3. In lines 5–7, the algorithm considers a semantic similarity $\sigma(O_i, T_j)$ only if both the tweet and the SCC are related to the same SCC and ordinance department. In a real-world scenario, since the number of ordinance departments is typically less than 40 and the number of SCCs is typically less than 15, the complexity of TOLCS is reduced from the quadratic $O(N_O \times N_T)$ (i.e., the case of investigating all pairwise combinations of $O_i \mapsto T_j$ in the original space to near-linear in practice, because only a limited number of pairings need to be compared for each SCC. Here, N_O is the number of ordinances and N_T the number of tweets.

Given millions of available tweets, this efficiency boost due to the reduction in complexity is significant. For example, we find that if we have 1 million tweets initially, they reduce to around 1000 tweets finally for similarity computation, a decrease by 3 orders of magnitude.

In the implementation of TOLCS, it is thus important to first ascertain whether ordinances and tweets have semantic similarities between them, which is done by checking if they are related to the same SCC and the corresponding ordinance department. We then apply the word2vec skip-gram model (Mikolov et al., 2013) for embedding words in root form for ordinance and tweets while maintaining their semantic properties. For measuring the similarity, we consider two techniques—Jaccard's similarity coefficient and cosine similarity. Jaccard's similarity is measured by taking the size of the intersection divided by the size of the union of words between two sentences (Leskovec, 2020). However, the former takes into account unique sets of words for each sentence. In contrast, the latter involves converting sentences to vectors. The variable that measures similarity is computed by using the average pairwise cosine similarity between the terms (Singhal, 2001). This use of cosine similarity allows for detection of similarities between two texts of different sizes.

Once the ordinances and tweets have been mapped to each other at this fine level of granularity, it sets the stage for sentiment analysis of the tweets guided by human judgment via CSK. This is described next.

3.5 Sentiment analysis of tweets guided by CSK

Given our interest in public ordinances, our ultimate goal is not just to connect tweets to ordinances but to monitor public opinion with regard to matters covered by such ordinances. Twitter is one of the most prominent social networking sites and its users often consider it a suitable medium to convey their emotions such as happiness or anger. Thus, we further perform sentiment analysis of linked tweets.

Tweets can evoke emotions that are positive, neutral or negative to varying degrees. By measuring how people's reaction to various tweets, we can capture their overall level of satisfaction with tweets related to SCC. We harness three different methods for sentiment analysis and later submit our findings to domain knowledge experts to evaluate our results.

3.5.1 SentiWordNet for sentiment analysis

SentiWordNet (Baccianella et al., 2010) is a knowledge base that enables classification of sentiments and opinion mining applications. It is derived by annotating synsets of WordNet (Baccianella et al., 2010) based on whether they are positive, negative or neutral. Every synset has three different scores, i.e., Pos(s), Obj(s) and Neg(s), which express how positive, objective or negative the terms contained in the synset are. The automatic annotation process used in SentiWordNet consists of two steps: a semi-supervised learning step and a random-walk step. The semi-supervised learning step trains the classifier used in the annotation, while the

random walk step looks at a graph of the terms and seeks to determine whether most of the terms are either positive, negative or neutral. Hence, CSK is captured here via the original WordNet synsets as well as the scores assigned to emotions being positive, negative or neutral based on intuitive human judgment.

In our system, given a tweet t_i, we first construct the set of relevant terms W_i and the polarity score S_i for t_i is computed as

$$S_i = \sum_{w \in W_i} s_w$$

where s_w is the overall word-level polarity score in SentiWordNet for word w.

3.5.2 VADER for sentiment analysis

VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto & Gilbert, 2014) is a rule-driven approach for text sentiment analysis with the advantage that it does not require training data and thus is less susceptible to domain-specific overfitting. It supports efficient real-time sentiment analysis and is built on a sentiment lexicon that shows sensitivity with respect to the content as well as polarity of text. It then assesses the semantic intensity by applying rules for various syntax-related traits. VADER relies on data structures that link lexical aspects to intensity scores, which are computed by combining the intensity scores of terms in the text. We use the same approach as above for SentiWordNet. Hence, it captures CSK via aspects such as sensitivity, traits and intensity scores that take into account the subjective aspects of subtle human reasoning and help to embed them within the sentiment analysis.

3.5.3 Multi-class supervised machine learning for sentiment analysis

In addition to SentiWordNet and VADER, we apply a supervised machine learning technique using Support Vector Machines (SVM) in order to train a model that extracts sentiment information from tweets. SVMs are a widely known machine learning algorithm that can be used for classification as well as regression and performs classification using separation hyperplanes chosen to maximize the margin between data points in different classes. We want to classify the tweets as either "positive", "negative" or "neutral". Hence, we use a multi-class classifier using SVM, following the one-vs-rest (OVR) approach. The multi-class classification is achieved using SVM by breaking down the problem into multiple binary classification problems. Building a model involves three steps: establishing training and test data sets, vectorizing them, and finally training an SVM model for prediction. We use a training dataset for this purpose that consists of approximately 1.6 million tweets along with their polarity scores (Kaggle, 2021). In our experiments, we typically assign a 80/20 split for training and testing.

Once we conduct our experiments using our proposed techniques described here, we submit our results to domain experts for evaluation. These results are presented in the following section on experimental evaluation.

4 Experimental evaluation

In the following, we evaluate the high-level and fine-grained mapping processes using annotations provided by domain experts and assess the potential of our approach for urban policy mining via sentiment analysis. The relevant code base for our experiments is available online.[1]

4.1 Ordinance mining

In our work on ordinance mining, the official website of the NYC legislative council (2018) serves as the source of urban policy data. This council is a body that represents 51 members from 51 council districts. A partial snapshot of the website is given in Fig. 7, showing its user interface with diverse examples of legislation along with their details. We use ordinances from different sessions in our experiments. A sample instance of such ordinance data is as follows:

A Local Law in relation to expanding the city's temporary outdoor dining program to include food service establishments not located adjacent to sidewalks or roadways, and to allow limited food preparation activities on such sidewalks or roadways.

In our experiments, we observe that the ordinance above maps to the Smart City Characteristics (SCCs) as shown in Table 1. This ordinance has been introduced by the Committee on Consumer Affairs and Business Licensing. The mapping we obtain for this ordinance based on the implementation of our CSK-guided mapping approach is 60% for smart governance and 40% for smart economy. Hence, this ordinance is focused on both smart economy and governance, with a tilt towards smart governance. Our domain experts on urban policy (researchers from the department of Earth and Environmental Studies at Montclair State University) confirm that this mapping is indeed appropriate as per their human judgment and expertise.

4.2 Social media text mining

We use Twitter as our source for social media mining. We conduct mining on approximately 30,000 tweets geo-tagged to the NYC area. In Table 2, we present a sample of the tweets we mine for the experiments in this paper.

We conduct a tweet to SCC mapping guided by CSK based on the steps described in Algorithm 2. If we show that ordinances have a mapping to one or more SCCs and that tweets show a mapping to overlapping SCCs, then we can establish a high-level mapping from the ordinances to the tweets. At this point, the process does not involve diving into mapping ordinances to tweets at finer levels of granularity.

An ordinance or tweet can be mapped to multiple SCCs. In this case, we consider the proportion of each SCC therein. In order to facilitate the mapping, we have developed an application called a *CSK-SCC classifier* with a GUI (graphical user interface) that takes tweets or ordinances as the input and maps them to SCCs. This

[1] https://github.com/mpuri14/urbanpolicymining.

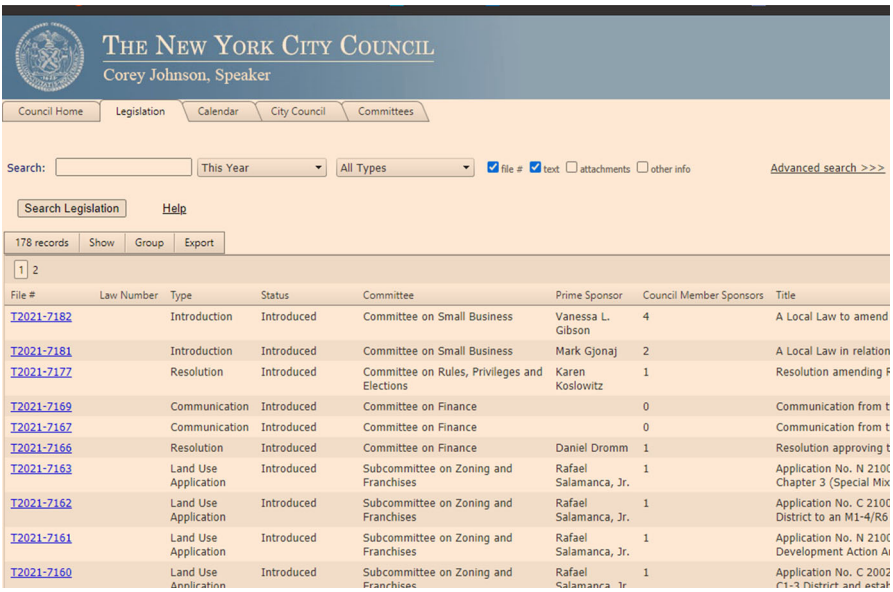


Fig. 7 Snapshot of ordinances from the NYC Council

CSK-SCC classification works by taking ordinances or tweets and extracting relevant information from them guided by CSK (as described earlier). It then maps them to appropriate SCCs. Once the text is propagated and analyzed (using Algorithm 2), we obtain the results of the SCC classification, as shown in Fig. 8.

4.3 Comparative evaluation of high-level mapping

We compare three different classifiers for evaluating the high-level mapping of ordinances and tweets:

1. Our CSK-SCC classifier
2. Non-CSK classifier
3. Multi-label supervised ML (machine learning) classifier using classifier chains

Table 1 Sample ordinance to SCC mapping

SCC	Frequency
Smart economy	2
Smart environment	0
Smart governance	3
Smart people	0
Smart mobility	0
Smart living	0

Table 2 Snapshot of tweets mined from Twitter

Tweets
1. "Construction on #CLine Both directions from 59th Street-Columbus Cir Station to Canal Street Station"
2. "RT @TheBradCurrie: Q2: What do effective digital learning environments look like? #satchat"
3. "Adding plants to your home or office has so many benefits such as reducing stress and increasing productivity!"
4. "RT @mentalhealth_ph: We should push more efforts to give awareness to most people and open their minds to mental illness. Many people still thinks that it is only in the mind of the person suffering but that is not the case"
5. "I really had a strong interest in going to see the black panther movie, but that has waned since social justice warriors are politicizing it."
6. "How do electric vehicles help the environment? In part through green electricity. How do renewables provide green electricity? In part through green storage. Tesla attempts to put together the pieces."

The non-CSK classifier baseline does not utilize any lexical databases, and instead relies on manually creating the KBs using words, their synonyms, and related words derived from various online sources. Hence, the non-CSK classifier does not use any CSK sources, but instead proceeds with mapping directly. For the machine learning based classifier, we would like to note that since we do not have prior training data large enough for this evaluation, we generate a sample training set manually.

No prior datasets exist for mapping tweets to SCCs, hence we use our NYC tweet dataset, which consists of 30,000 tweets obtained using the Twitter API over a

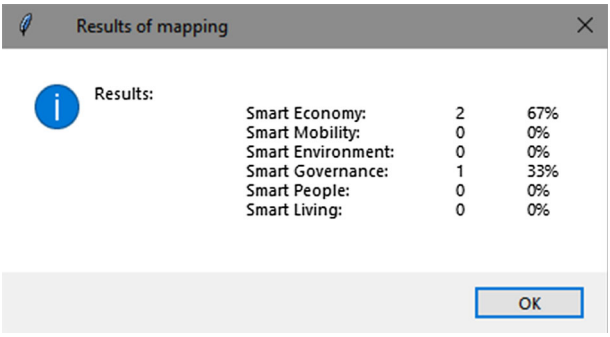


Fig. 8 Sample results from CSK-SCC classifier

Table 3 Comparison of sample tweet across three classifiers

SCC	CSK-SCC classifier	Non-CSK classifier	Multi-label supervised ML classifier
Smart economy	0	2	1
Smart environment	0	0	0
Smart governance	2	0	2
Smart people	1	0	0
Smart mobility	0	0	0
Smart living	1	1	0

period of 4 months from January 2018 to April 2018, using a 80/20 training/test split. We use a multi-label classification technique in order to map ordinances and tweets to multiple potential smart city characteristics. In this classification type, the training set comprises ordinances and tweets mapped to various SCCs. This training set is then used to predict those that do not have this mapping. The classes are not mutually exclusive and may be related. Tweets and ordinances can map to one or more SCC, which makes this an appropriate mapping technique. We preprocess the data so that only relevant parts of the tweet/ordinance are used, eliminating hashtags, punctuation marks, as well as stopwords. We then employ lemmatization in order to map various forms of a word to a common base form of the word. For this problem of multi-label classification, we considered 3 popular classification techniques: Binary Relevance, Label Powerset, and Classifier Chains. While the Label Powerset method takes into account correlations between labels, the computational complexity is significant and can grow exponentially as the number of target labels increases. The trivial Binary Relevance approach has the drawback that correlations between labels are disregarded. Hence, we ultimately adopted the Classifier Chains method, which uses a chain of binary classifiers $B_0, B_1, B_2, ..., B_n$.

Table 3 shows a sample classification using the three techniques above for the following tweet. Here, a single tweet with a frequency of 2 for Smart Governance means that 2 terms relevant to Smart Governance are present in the tweet.

> "Amazing work kicking off our @TexasAFT advocacy for public school funding RGV! Tons of members participated in a socially distanced event to bring attention to the kids and educators around Texas that need our continued investment in public schools".

As per the assessment provided by our domain experts on urban policy, the mapping provided by the CSK classifier is the most appropriate in this example. Likewise, for all our experiments, we assess the effectiveness of our results via evaluation by domain experts. We describe this next.

We submit the results of our CSK-based mapping approach to be assessed by domain experts on urban policy. These experts are faculty and other researchers from Montclair State University's Earth and Environmental Studies (EAES)

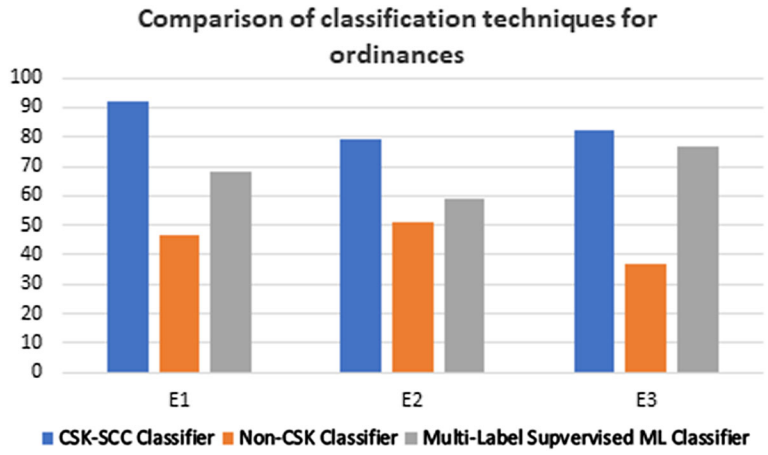


Fig. 9 Evaluation of mapping methods for ordinances

department. The experts define the concept of ground truth, where their individual human judgment based on their expertise on urban policy is used to determine whether a mapping is considered appropriate. Here a "true mapping" *(TM)* is defined as one for which the ground truth provided by experts matches the classification provided by our approach within a given error threshold *(E)*. Conversely, a "false mapping" *(FM)* is one which does not match with the ground truth offered by domain experts within threshold *E*. For example, if experts indicate that an ordinance/tweet matches "smart environment" and "smart mobility" to an equal extent and our approach indicates the same, it is clearly a *TM*. Also, if our approach indicates mapping with the same SCCs but with the extent of a greater tilt towards one than the other, it is can still be a *TM* if it is within the limits of *E*. The experiments shown in this paper are with a threshold of *E=10%*, which is a fairly tight limit in such settings. Hence, if the mapping of the approach matches ground truth given by experts to the extent of at least 90%, it is a *TM*, else *FM*. We thus use the following formula to measure P_{map}, i.e., precision of mappings.

$$P_{\text{map}} = \frac{TM}{TM + FM}$$

In our experiments, we do not evaluate recall, since the data to calculate it is not feasible to obtain. Hence, we present our evaluation based on precision, using the ground truth provided by domain experts. We use the domain experts' judgment for determining correctness, hence their response is considered authentic enough to serve as the ground truth. In some cases, where the expert's response to the mapping of an ordinance/tweet to the SCCs is "none" and the system assigns a label or vice versa, the expert's response is considered correct, which means that such a case would be counted as an "incorrect classification" or a "false mapping (FM)" and this is taken into account in computing the precision of the mappings.

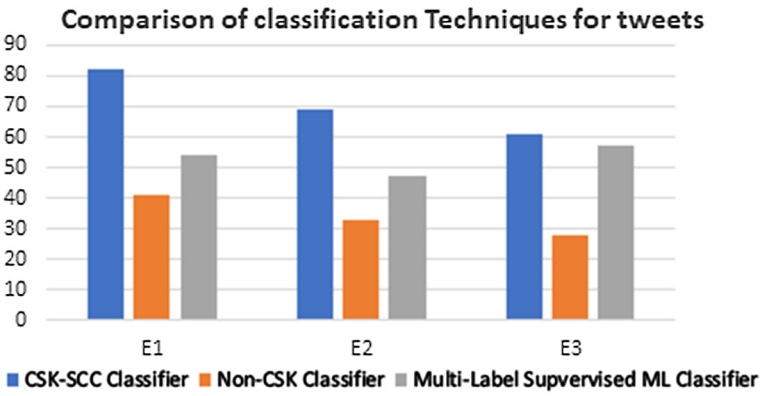


Fig. 10 Evaluation of mapping methods for tweets

The precision metric yields the following observations in our high-level mapping approach (prior to deeper mapping with TOLCS). For ordinances, we find that the CSK-SCC classifier obtains a precision of 84%, the non-CSK classifier has 45%, while the Multi-Label Supervised ML classifier has 68% precision. This is shown in Fig. 9 based on the ground truth provided by 3 experts. For tweets, we observe that the CSK-SCC classifier gives 71% precision, the non-CSK classifier gives 34% while the Multi-Label Supervised ML classifier has 53% precision. This is shown in Fig. 10 as per the ground truth provided by 3 different experts. Note that although larger training datasets would yield higher accuracy for the Multi-Label Supervised ML classifier, such datasets do not exist.

4.4 TOLCS evaluation based on ordinance sessions

Once we have established high-level mapping between ordinances and SCCs, as well as tweets and SCCs, we have proved that our CSK-SCC classifier is suitable for the mapping required in our work. In fact this is a significant aim behind conducting our mapping in two different stages: a high-level mapping to ascertain the effectiveness of the CSK-guided approach, and a fine-grained mapping for the actual linkage of the ordinances and tweets.

Therefore, we proceed to use this CSK-SCC classifier within the TOLCS approach. Having accomplished this, we now consider data from two different ordinance sessions for the TOLCS mapping process to in order link the individual ordinances to their pertinent tweets.

1. Session 1: NYC ordinances for the years 2010–2013
2. Session 2: NYC ordinances for the years 2014–2017

An example of the mapping process is illustrated in Fig 11 where sample ordinances are linked to their relevant tweets. We give the results of our experiments to the domain experts for evaluation. They find that the precision for the TOLCS mapping

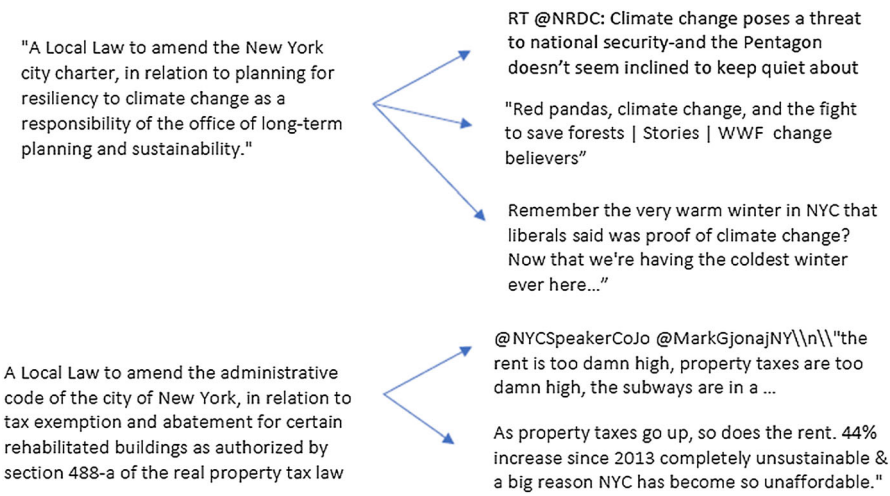


Fig. 11 Sample mapping using TOLCS

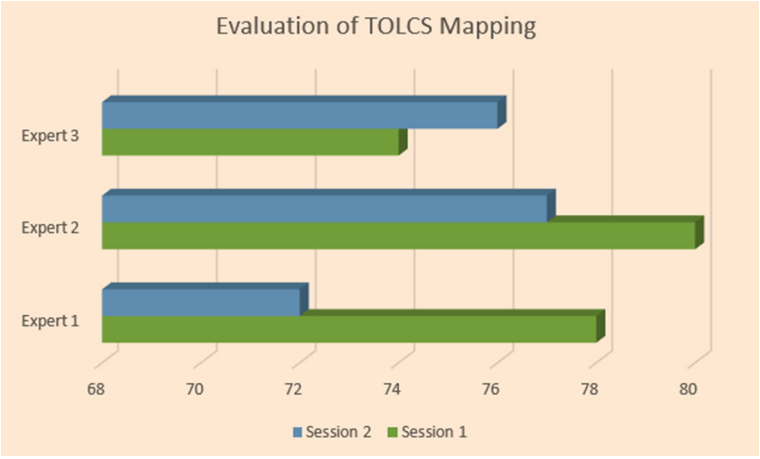


Fig. 12 Evaluation of TOLCS

for the 2 sessions is approximately 77%, as shown in Fig. 12 based on the judgment of 3 domain experts.

4.5 Evaluation of polarity analysis

We conduct sentiment analysis using polarity classification by deploying three different methods as follows.

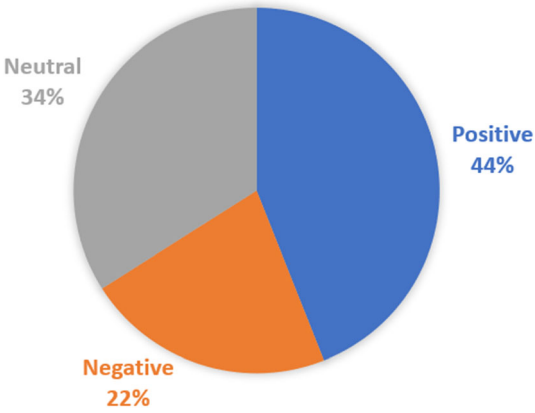


Fig. 13 Distribution of polarity classification of tweets

1. SentiWordNet for Sentiment Analysis (CSK based on WordNet synsets with scores for emotions)
2. VADER for Sentiment Analysis (CSK based on capturing sensitivity, traits and intentsity)
3. Multi-Class Supervised ML classifier for Sentiment Analysis using SVM

These three methods are described earlier in the paper. In our experiments on polarity classification, we use our SCC KBs leveraging CSK as a guide for selecting relevant tweets. After conducting the polarity classification with the three methods mentioned above, we submit the results of this sentiment analysis to domain experts on urban policy (Montclair's EAES faculty and researchers) who evaluate the mapping process described earlier. The precision metric described earlier is used here as well, this time with true or false "polarity classifications" instead of "mappings" based on how well they match the judgment of the domain experts. Based upon this evaluation, we find that VADER has a precision 72% compared to 64% for SentiWordNet and 57% for the Multi-Class ML classifier. The aggregate results from the sentiment analysis of tweets is visualized in Fig. 13. While most of the tweets on the NYC ordinances are positive, the significant numbers of negative and neutral tweets may be investigated to identify potential points of improvement.

4.6 Assessment of public satisfaction on urban policy

An important aspect of evaluation in this research on the whole involves gauging the public level of contentment on policies related to various smart city characteristics (based on the opinions they express on Twitter) for the data analyzed in this work. An aggregate representation of their contentment for NYC legislation data is shown in Table 4. We see that people are in general satisfied with most SCCs, and that NYC has shown much progress towards becoming a smart city. Our findings in this paper seem to be compatible with the worldwide Smart City Index 2020, where NYC ranks at position 10, the top few being Singapore, Helsinki and Zurich, with many European cities being quite high in the rankings (IMD Business School,

Table 4 Gauging public satisfaction on urban policy

SCC	Polarity (%)
Smart economy	51
Smart environment	34
Smart governance	42
Smart people	47
Smart mobility	49
Smart living	56

2020). Providing urban agencies with various types of feedback based on the public attitude towards urban policy with respect to SCCs may help such agencies to better capture the public perception and enable them to incorporate data-driven decision-making into their process of enacting ordinances. This is in line with the concept of smart governance that entails transparency via public involvement.

In general we observe that our evaluation has an important impact of gauging the urban region analyzed in this work from a smart city perspective, in addition to assessing the effectiveness of our research methods. This contributes towards smart cities and sustainability as a whole. These paradigms are highly significant today in the light of issues such as climate change and global warming; in order to combat these issues, sustainable living is highly encouraged and smart cities gear towards that notion.

5 Discussion regarding the role of CSK

Commonsense knowledge is described as knowledge that humans inherently possess but machines do not. Both humans and machines are capable of factual knowledge. The growth in Internet related technologies has enabled machines to exceed humans in terms of encyclopedic knowledge (Tandon et al., 2017). Gadgets such as virtual assistants developed by Amazon and Google can provide more details on events or news than humans. Despite this, only humans have advanced commonsense enabled decision-making abilities, such as understanding that although a bench and a bridge have a similar structure. For example, consider a real-world scenario where a truck was misclassified as an overpass by a semi-autonomous Tesla vehicle, causing a fatal accident in 2016. The use of CSK in such scenarios could capture human judgement so that the vehicle would avoid such misclassifications. As per such examples and other contexts including our work in this paper, we can observe the use of commonsense knowledge in the following aspects:

1. Knowledge pertaining to objects present in our surroundings
2. Knowledge about how different objects are related to one another
3. Knowledge describing interactions of different types among various objects

Given the substantial body of research on deep learning methods and the increasing availability of large-scale training datasets, artificial intelligence has made notable strides in the past decade. However, important challenges remain with regard to CSK. For example, even with the latest techniques in object detection, the addition of random overlays in images can lead to incorrect classification of images (Pandey et al., 2018; Szegedy et al., 2014). We have also observed some evasive strategies by natural language processing models when trying to build conversations with humans (Holtzman et al., 2019). In particular, CSK is very useful in situations with limited training data. On the whole, it has been observed that CSK and deep learning benefit from each other (Hwang et al., 2020; Razniewski et al., 2021).

Commonsense knowledge is reproducible and has a multitude of applications spanning the AI realms of Knowledge Representation (KR) and Machine Learning (ML). In our work in this paper, CSK plays a crucial role from the following perspectives:

1. It helps in harnessing human judgment in the high-level mapping of ordinances and tweets through SCCs as a nexus.
2. It assists in significantly reducing complexity for fine-grained ordinance-tweet mapping via SCC KBs and ordinance KBs.
3. It guides sentiment analysis of tweets to discover knowledge from public opinions on ordinances, capturing nuances.
4. It plays a subtle role in assessing how well a given urban region heads towards a smart city.

6 Conclusions and roadmap

This research is focused on capturing human judgment via commonsense knowledge for text mining of urban policy, with impact on smart cities, specifically with respect to smart governance. It has the following highlights.

- It proposes mapping ordinances to tweets via CSK, incorporating SCCs as a nexus, helping to link policies to public reactions.
- It analyzes public satisfaction due to which urban agencies can evaluate policy effectiveness, and include data-driven methodologies in governance.
- It deploys ground truth from domain experts on urban policy to corroborate that human judgment is well-harnessed in mining.
- It builds SCC KBs and ordinance KBs reusable in other research, in terms of the KBs themselves and the corresponding processes.

Some challenges encountered in this research, which could be considered as the potential limitations of this study, include the following.

- Accuracy of mapping techniques with highly subtle nuances still poses a challenge that can be further addressed as ongoing work by drawing on recent

advances in neural language modeling, while achieving a finer level of granularity for the analysis of ordinance reactions.
- Research on analyzing emojis is a fairly challenging task when conducting sentiment analysis, and though we have not dwelt upon this in our current research, it can be considered further via several developments in the literature.
- Addressing the issue of Named Entity Disambiguation (NED) in tweets offers yet another challenge that has not been an item of focus in our present study.

Notable takeaways from this research that provide the scope for potential future work for us as well as other researchers include the following aspects.

- Handling iterative changes in larger quantities of data, dealing with reporting bias of different words and phrases, and incorporating idioms and noise in sentiment analysis from social media platforms.
- Mapping news to tweets, since news represents formal text analogous to ordinances, and perform mining based on this data.
- Conducting a historical analysis of tweets before and after a given ordinance or piece of news, and thereby addressing time-series perspectives.

Our cross-disciplinary exploration lays the foundation for further investigation with text mining from domain-specific perspectives. This could result in collaborations between departments such as Earth Science, Environmental Engineering, Urban Studies on the one hand and Computer Science, Information Technology, Data Science on the other, especially with a focus on AI areas such as Knowledge Representation, Data Mining and Machine Learning. Substantial research is available on mining social media. However, to the best of our knowledge, ours is among the first academic studies on ordinance mining, particularly from a smart city angle. Ultimately, our work encourages greater scrutiny in the area of smart cities as well as social media. This would result in governing bodies taking a stronger initiative in incorporating data-driven methodologies within their functioning. Hence, our research makes a modest contribution to the overall paradigm of Data Science for Social Good.

Acknowledgements Manish Puri was supported by a Graduate Teaching and Research Assistantship from the Computer Science (CS) department at Montclair State University (MSU) as an MS student in CS. Aparna Varde's research has support via grants from NSF (USA), Award Number 2018575 on MRI: Acquisition of a High-Performance GPU Cluster for Research and Education, and Award Number 2117308 on MRI: Acquisition of a Multimodal Collaborative Robot System (MCROS) to Support Cross-Disciplinary Human-Centered Research and Education at Montclair State University, She is a visiting researcher at Max Planck Institute for Informatics, Saarbrücken, Germany, in the research group of Dr. Gerhard Weikum, during her sabbatical. Additionally, we thank Xu Du, Boxiang Dong, Anna Feldman and Matthew Kowalski from MSU for some early inputs on this work.

References

Alkhammash, E. H., Jussila, J., Lytras, M. D., & Visvizi, A. (2019). Annotation of smart cities twitter micro-contents for enhanced citizen's engagement. *IEEE Access, 7,* 116267–116276.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*.

Baziotis, C., Pelekis, N., & Doulkeridis, C. (2017). DataStories at SemEval-2017 Task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 747–754). Association for Computational Linguistics.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., & Choi, Y. (2019). COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4762–4779). Association for Computational Linguistics.

Cambria, E., Li, Y., Xing, F. Z., Poria, S., & Kwok, K. (2020). Senticnet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. Association for Computing Machinery.

Cao, Z., Wang, L., & de Melo, G. (2018). Link prediction via subgraph embedding-based convex matrix completion. In *AAAI*.

Chalier, Y., Razniewski, S., & Weikum, G. (2020). Joint reasoning for multi-faceted commonsense knowledge. In *AKBC conf*.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018) Think you have solved question answering? Try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Davison, J., Feldman, J., & Rush, A. (2019). Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 1173–1178). Association for Computational Linguistics.

DBPedia: Global and unified access to knowledge graphs. DBPedia.

Du, X., Emebo, O., Varde, A., Tandon, N., Chowdhury, S. N., & Weikum, G. (2016) Air quality assessment from social media and structured data: Pollutants and health impacts in urban planning. In *IEEE ICDE workshops* (pp. 54–59).

Du, X., Kowalski, M., Varde, A. S., de Melo, G., & Taylor, R. W. (2019). Public opinion matters: Mining social media text for environmental management. In *ACM SIGWEB*, *5,* 1–5:15.

Elazar, Y., Mahabal, A., Ramachandran, D., Bedrax-Weiss, T., & Roth, D. (2019) How large are lions? Inducing distributions over quantitative attributes. *CoRR*, abs/1906.01327.

Gundogan, F. (2015). Real-time signal control in developing cities: Challenges and opportunities. In *IEEE international conference on intelligent transportation systems* (pp. 38–41).

Han, P., Shen, S., Wang, D., & Liu, Y. (2012). The influence of word normalization in english document clustering. *IEEE CSAE, 2,* 116–120.

Hitzler, P., Bianchi, F., Ebrahimi, M., & Sarker, Md.K. (2019). Neural-symbolic integration and the semantic web. *Semantic Web*, *11,* 1–9.

Holtzman, A., Buys, J., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text.

Hwang, J.D., Bhagavatula, C., Le Bras, R., Da, J., Sakaguchi, K., Bosselut, A., & Choi, Y. (2020). Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs.

IMD Business School – Lausanne Switzerland. (2020). Smart City Index 2020: Singapore, Helsinki and Zurich triumph in global smart city index. https://www.imd.org/smart-city-observatory/smart-city-index

Jayadharshini, J., Sivapriya, R., & Abirami, S. (2018) Trend square: An android application for extracting twitter trends based on location. In *2018 international conference on current trends towards converging technologies (ICCTCT)* (pp. 1–5).

Kaggle. (2021). Sentiment140 dataset with 1.6 million tweets. https://www.kaggle.com/kazanova/sentiment140

Lenat, D. B., Guha, R. V., Pittman, K., Pratt, D., & Shepherd, M. (1990). Cyc: Toward programs with common sense. *Communications of the ACM, 33*(8), 30–49.

Leskovec, J. (2020). *Mining of massive datasets*. Cambridge University Press.

Li, Q., Shah, S., Liu, X., Nourbakhsh, A., & Fang, R. (2016) Tweetsift: Tweet topic classification based on entity knowledge base and topic enhanced word embedding. In *ACM CIKM* (pp. 2429–2432).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *NIPS - Volume 2* (pp. 3111–3119).

Miller, G., & Fellbaum, C. (1998). *WordNet: An electronic lexical database*. The MIT Press.
Mishra, B. D., Tandon, N., & Clark, P. (2017). Domain-targeted, high precision knowledge extraction. *TACL Journal, 5,* 233–246.
Pandey, A., Puri, M., & Varde, A. (2018). Object detection with neural models, deep learning and common sense to aid smart mobility. *IEEE ICTAI* (pp. 859–863)
Puri, M., Varde, A. S., Du, X., & de Melo, G. (2018a). Smart governance through opinion mining of public reactions on ordinances. In *IEEE ICTAI* (pp. 838–845) IEEE.
Puri, M., Varde, A. S. & Dong, B. (2018b). Pragmatics and semantics to connect specific local laws with public reactions. In *IEEE Big Data* (pp. 5433–5435).
Razniewski, S., Tandon, N., & Varde, A. (2021). Information to wisdom: Commonsense knowledge extraction and compilation. In *ACM WSDM* (pp. 1443–1446).
Romero, J., Razniewski, S., Pal, K., Pan, J. Z., Sakhadeo, A., & Weikum, G. (2019). Commonsense properties from query logs and question answering forums. *CoRR*, abs/1905.10989.
Rose, G., & Willis, A. (2019). Seeing the smart city on twitter: Colour and the affective territories of becoming smart. *Environment and Planning D: Society and Space, 37*(3), 411–427.
Sakaguchi, K., Le Bras, R., Bhagavatula, C., & Choi, Y. (2020). WinoGrande: An adversarial winograd schema challenge at scale. In *AAAI conference* (pp. 8732–8740).
Shahidehpour, M., Li, Z., & Ganji, M. (2018). Smart cities for a sustainable urbanization: Illuminating the need for establishing smart urban infrastructures. *IEEE Electrification Magazine, 6*(2), 16–33.
Shams, M. B., Hossain, M. J. & Noori. S. R. H. (2020). A time series analysis of trends with twitter hashtags using lstm. In *2020 11th international conference on computing, communication and networking technologies (ICCCNT)* (pp 1–6).
Shoeb, A. A. Md., Raji, S., & de Melo, G. (2019). EmoTag: Towards an emotion-based analysis of emojis. In *Proceedings of RANLP 2019* (pp. 1094–1103).
Singh, S., Wen, N., Hou, Y., Alipoormolabashi, P., Wu, T., Ma, X., & Peng, N. (2021) COM2SENSE: A commonsense reasoning benchmark with complementary sentences. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021* (pp. 883–898). Association for Computational Linguistics.
Singhai, A. Introducing the knowledge graph: Things, not strings. googleblog.blogspot.co.uk
Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin, 24,* 35–43.
Solanki, S. K., & Patel, J. T. (2015). A survey on association rule mining. In *Internaional conference on advanced computing communication technologies* (pp. 212–216).
spaCy. (2021). Spacy: Industrial strength natural language processing. https://spacy.io/api
Speer, R., Chin, J., & Havasi, C. (2016) ConceptNet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975.
Stanford University. (2021). Stemming and lemmatization. https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going deeper with convolutions. *CoRR*, abs/1409.4842.
Talmor, A., Yoran, O., Le Bras, R., Bhagavatula, C., Goldberg, Y., Choi, Y., & Berant, J. (2021). Commonsenseqa 2.0: Exposing the limits of ai through gamification. In *Proceedings of the neural information processing systems track on datasets and benchmarks 2021*.
Tandon, N., & de Melo, G. (2010). Information extraction from web-scale n-gram data. In Zhai, C., Yarowsky, D., Viegas, E., Wang, K., & Vogel, S. (Eds.) Web N-gram Workshop ACM SIGIR (Vol. 5803, pp. 8–15).
Tandon, N., de Melo, G., Suchanek, F., & Weikum, G. (2014) WebChild: Harvesting and organizing commonsense knowledge from the web. In *ACM WSDM* (pp. 523–532).
Tandon, N., de Melo, G., & Weikum, G. (2011) Deriving a Web-scale common sense fact database. In *AAAI* (pp. 152–157).
Tandon, N., de Melo, G., & Weikum, G. (2017) WebChild 2.0: Fine-grained commonsense knowledge distillation. In *ACL system demo* (pp. 115–120)
Tandon, N., Varde, A. S., & de Melo, G. (2017). Commonsense knowledge in machine intelligence. *ACM SIGMOD Record, 46*(4), 49–52.
The IEEE Smart Cities Technical Community. (2018). https://smartcities.ieee.org/
The New York City Council. Legislative research center web page. http://legistar.council.nyc.gov/, 2018.

United Nations. (2019). *Department of Economic and Social Affairs: Population Division*. World population prospects: Highlights, Key findings and advance tables. United Nations.
Wang, C., Liang, S., Jin, Y., Wang, Y., Zhu, X., & Zhang, Y. (2020). SemEval-2020 Task 4: Commonsense validation and explanation.
Wang, L., Wang, Y., Liu, B., He, L., Liu, S., de Melo, G., & Xu, Z. (2017). Link prediction by exploiting network formation games in exchangeable graphs. In *IJCNN*
Wien, T. U. (Vienna University of Technology). (2015). European smart cities. *Technical report*.