On Classifying Urban Issues from TV News Using Natural Language Processing and Geoprocessing

Rich Elton Carvalho Ramalho
Information Systems Laboratory
Federal University of Campina Grande
Campina Grande - PB, Brazil
rich.ramalho@ccc.ufcg.edu.br

Anderson Almeida Firmino
Information Systems Laboratory
Federal University of Campina Grande
Campina Grande - PB, Brazil
andersonalmeida@copin.ufcg.edu.br

Cláudio de Souza Baptista
Information Systems Laboratory
Federal University of Campina Grande
Campina Grande - PB, Brazil
baptista@computacao.ufcg.edu.br

Ana Gabrielle Ramos Falcão
Information Systems Laboratory
Federal University of Campina Grande
Campina Grande - PB, Brazil
anagabriellee@gmail.com

Maxwell Guimarães de Oliveira
Information Systems Laboratory
Federal University of Campina Grande
Campina Grande - PB, Brazil
maxwell@computacao.ufcg.edu.br

Fabio Gomes de Andrade
Federal Institute of Paraíba
Cajazeiras - PB, Brazil
fabio@ifpb.edu.br

***Abstract*—Citizens as sensors enable society to discuss urban issues. Although some geosocial networks have been developed in recent years to enable citizens to report many types of urban problems, the engagement of the users of these networks usually decreases in time. Hence, many relevant issues are not posted reducing the effectiveness of these networks. Aiming to overcome this limitation, this article proposes an approach in which urban issues are automatically detected from a television news program. The proposed solution uses geoparsing and Natural Language Processing techniques to geocode and classify the identified complaints. The results are published in the Crowd4City geosocial network that deals specifically with urban issues. Finally, our method was evaluated using data from a real news TV program in Brazil. Our results indicated 59.8% of success in extracting text and location from the video news.**

***Index Terms*—Geosocial network; NLP; Urban Issues; Crowdsourcing.**

I. Introduction

The high concentration of population in urban areas has imposed on local authorities several challenges to address issues concerning mobility, security, infrastructure, education, health, etc. These are what we call urban issues. One important challenge for these authorities consists of identifying the problems that have been faced by citizens.

Aiming to solve this limitation, in the context of Smart Cities, some authors have developed geosocial networks that deal specifically with urban issues. These networks enable the use of context-aware services to locate users and their complaints.

In a previous work [1] we provided a discussion towards the proposal for an approach for automated detection of urban issues from TV news programs. In this work, we improve our previous proposal defining domain ontology named UIDO (Urban Issues Domain Ontology), which aims to model the semantics of urban issues. We also extend the case study with a performance analysis in order to compare our classifier with XGBoost and Bi-LSTM classifiers. Finally, this work also provides a state-of-the-art comprehensive review in the field.

The authors would like to thank the Brazilian Research Council - CNPq for funding this research.

In the context of Smart Cities, geosocial networks enable the use of context-aware services to locate users and their complaints about urban issues. Several tools, such as Crowd4City [2], Wegov [3], and FixMyStreet [4] provide urban complaint environments. However, people's motivation to use such geosocial networks decreases throughout time. Hence, to ensure a high engagement of society, different approaches to gathering information are required.

Several local TV stations in Brazil portray urban issues addressed by the community. An example is the 'Calendar' report in a daily open TV channel news program in the State of Paraíba, Brazil. That news broadcast exhibits several urban issues from the main cities of that particular state. Then, we decided to use the information presented in this broadcast to automatically input new urban issues into geosocial networks, improving citizenship and increasing awareness. To accomplish this task, we initially convert the audio descriptions in the news channel into text. After that, we use geoparsing tools from Geographic Information Systems (GIS) and Natural Language Processing (NLP) techniques to extract the correct location of the respective urban issues.

In this article, we propose a framework to extract audio files from TV news, convert them into text documents, then extract location using a gazetteer and urban issues from text using NLP techniques in order to feed the Crowd4City geosocial network. It is important to mention that the news is up-to-date and extracted from a real context. Our main contribution consists in the integration of GIS and NLP.

The remainder of this article is structured as follows. Section II discusses related work. Section III presents an overview of the Crowd4City geosocial network. Section IV focuses on our proposed method for extracting and structuring urban issues reported in TV news. Section V presents a case study and

discusses the results. Finally, Section VI concludes the paper and points out further research to be undertaken.

II. Related Work

NLP techniques have been broadly used in several application domains including machine translation, speech recognition, chatbots/question answering, text summarization, text classification, text generation, sentiment analysis, recommendation systems and information retrieval. Britz et al. [5] discuss machine translation using a seq2seq model. Reddy et al. [6] present a question answering approach. Schwenk et al. [7] focus on text classification. Radford et al. [8] propose a language model using unsupervised learners. NLP is difficult to accomplish since text differs from language to language.

Upon developing our proposed approach, we first performed an extensive study on the already existing models within scenarios similar to ours. Given that our method is based on NLP and geoparsing, we discovered some useful corpora. Oliveira et al. [9], for instance, contributed to the creation of a gold-standard english corpus of urban issues identified from geo-located tweets. Such information can be very useful for improving geoparsers and for developing classifiers for the detection of urban issues. Focusing on our TV news domain, Camelin et al. [10] composed a corpus of different TV Broadcast News from French channels and online press articles. These articles were manually annotated in order to obtain topic segmentation and linking annotations between topic segments and press articles. The FrNewsLink is freely available online. Although both corpora are based on different languages than the one used in our study, they proved to be useful in such domains.

Aiming at obtaining information from the TV news videos, Kannao and Guha [11] focused their study on extracting text from the overlay banner presented in broadcasts. Such text usually contains brief descriptions of news events. They performed experiments using Tesseract Optical Character Recognition (OCR) for overlay text recognition trained using Web news articles. The authors validated their approach using data from there Indian English television shows and obtained significant results. However, their domain is limited. In a more recent study [12], the authors proposed a system architecture that enables the semantic segmentation of TV news broadcast videos. In this architecture, the collected videos were segmented into shots from which advertisements were detected and removed. Then, the remaining non-commercial content was classified as news bulletins, debates, or interviews. Finally, contents categorized as news bulletins were processed aiming at obtaining news stories. To validate the proposed architecture they used videos from three Indian English news channels. Nonethetess, they do not address urban issues.

Chifu and Fournier [13] developed the SegChainW2V framework. In their work, video segmentation was accomplished from lexical chains obtained by transforming the results of video transcriptions into vectors. Then, they used cosine-based similarity measures to detect topic variations along with a video and applied the word2vec word embeddings model for computing similarities between the videos. They used data from a French TV news broadcast and an English MOOC video, and the preliminary results showed the viability of the proposed solution. Nevertheless, unlike the solution described in this article, this work did not deal with urban complaints.

Iwata et al. [14] enabled users to retrieve news videos from keywords. In their solution, the author used OCR for extracting captions from video images in Arabic. Then, they applied of text processing techniques to perform character recognition and automatic language translation. Their approach aimed at evaluating the use of frame images extracted from the AlJazeera broadcasting programs. Despite the relevance of this work, it does not evaluate the semantics of the videos.

Similarly, Pala et al. [15] developed a system for the transcription, keyword spotting and alerting, archival and retrieval for broadcasted Telugu TV news. Their main goal was to aid viewers in easily detecting where and when topics of their interest were being presented on TV news in real-time and they were also hoping to assist anyone (including editorial teams at TV studios) in discovering videos of TV news reports about specific topics, defined by the user with keywords. Their system was the first that enabled the simultaneous execution of the broadcasted audio (speech), video and transcription of the audio in real-time with the Indian Language, with keyword spotting and user alerts. Although it can detect topics of interest based on keyword, the system cannot identify the theme to extract the theme or domain being discussed in the video.

Bansal and Chakraborty [16] proposed an approach for content-based video retrieval by combining several state-of-the-art learning and video/sentence representation techniques given a natural language query. They aimed at overcoming the robustness and efficiency problems found in the existing solutions using deep-learning based approaches, combining multiple learning models. Their results showed that their solution was able to capture the videos' and sentences' semantics when compared to other existing approaches. However the authors lack retrieving any geographic information.

Dong et al. [17] focused on developing a method for subject words extraction of urban complaint data posted on the Internet. Their approach consisted in the segmentation of the complaint information, extraction and filtering of candidate subject words, and was validated using 8289 complaints posted on a Beijing website. The proposed method showed that better results can be obtained than the Term Frequency–Inverse Document Frequency (TF-IDF) and TextRank methods in the context of written informal content made by Internet users. Nonetheless, such an approach would need to be validated in other scenarios.

Mocanu et al. [18] proposed a method for news extraction by using temporal segmentation of the multimedia information, allowing it to be indexed and thus be more easily found the users interested in specific topics. Their approach was based on anchorperson identification, where the TV news program presenter was featured on the video. They performed

some tests with a limited database of French TV programs. Nevertheless, the topic detection method implemented in this work is not very robust, since it is based only on the video subtitles.

Zlitni et al. [19] addressed the problem of automatic topic segmentation in order to analyze the structure and automatically index digital TV streams using operational and contextual characteristics of TV channel production rules as prior knowledge. They used a two-level segmentation approach, where initially the program was identified in a TV stream and then the segmentation was accomplished, thus dividing the news programs into different topics. They obtained reasonable results in their experiments, but their approach is completely dependent on the production rules of TV channels. Also aiming at achieving news story segmentation, Liu and Wang [20] focused their efforts on using a convolutional neural network in order to partition the programs into semantically meaningful parts. They based their input on the closed caption content of the news and trained and tested their model using the TDT2 dataset, from Topic Detection and Tracking (TDT). Although they obtained significant results, their approach is limited to the linguistic information extracted from the closed caption and thus it is not applicable to programs without such resource.

Even though several studies could be found, none of them comprises the same aspects and goals we aim at achieving with our study, which is to perform NLP and GIS extraction and structuring of stories depicted in TV news reports, focusing especially on urban complaints.

III. Ontologies versus Topic Modeling for urban issues automated extraction

The classical definition of ontology is provided by Gruber [21]: "An ontology is a formal explicit specification of a shared conceptualization". In other words, ontologies provide a shared vocabulary that can be used to model a domain. Such a shared vocabulary can be represented by objects and/or concepts that may contain properties and relationships [22].

In computer science, ontologies have been used since the last decades for accomplishing several tasks such as improving communication between agents (human or software); enabling computational inference; and reusing data models or knowledge schemas [23]. In this context, ontologies can be classified according to the language expressivity and scope. Language expressivity focuses on knowledge representation. Such a classification includes information ontologies, linguistic ontologies, software ontologies and formal ontologies. On the other hand, the scope addresses the level of specificity of the knowledge represented by the ontology and includes domain ontologies and general ontologies.

Concerning language expressivity, information ontologies are a clarified organization of ideas useful solely by humans. Linguistic ontologies can be dictionaries, folksonomies [24], lexical databases, etc. One known information ontology example is the Resource Description Framework (RDF). Software ontologies focus on data storage and data manipulation for data consistency in software development activities. The Unified Modeling Language (UML) is one example of a software ontology. Formal ontologies require some clear semantics and involve formal logic and formal semantics, with strict rules about how to define concepts and relationships. The Web Ontology Language (OWL) a widely used language that enables people to build formal ontologies. Finally, concerning the scope of this article, domain ontologies stand out. Different from general ontologies, domain ontologies focus on specific domains, such as urban issues, with specific viewpoints and characteristics, to better represent the domain semantics.

A way of extracting valuable information from texts is using Ontology-based Information Extraction (IE), as performed by Yang et al. [25]. Ontology-based IE processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information [26]. Two main categories split ontology-based IE: document-driven and ontology-driven. Document-driven [27] is known as semantic annotation, where the discovered knowledge is structured in domain ontologies, while ontology-driven [28] extracts information from unstructured documents based on a domain ontology constructed under the help of domain experts or even by combining domain ontologies with core reference ontologies and folksonomies.

While domain ontologies represent structured and specialized semantic knowledge, topic modeling is the discovery of hidden semantic structures in a text [29]. In this context, unsupervised topic modeling is an unsupervised machine learning task that can scan a set of documents to detect word and phrase patterns a set of documents, detecting word and phrase patterns within them, being suitable for dealing with a large-scale dataset [30]. Topic modeling also has the capability of automatically clustering word groups and similar expressions that best characterize a set of documents. Thus, topic modeling can be seen as a baseline to build domain ontologies in order to enable performing ontology-based IE in specific domains such as urban issues.

In the context of the urban issues domain, this article applies topic modeling combined with location extraction in order to identify not solely the main keywords for this domain, but also the hidden semantics of urban issues. Such hidden semantics may include the way of people complain about urban issues such as potholes, broken lights, litter and others. This is a key step towards building a domain ontology in the urban issues domain that can be applied to perform ontology-based IE in unstructured texts from social media. Moreover, once a domain ontology in urban issues is built, it can be useful to many other tasks. One of the main advantages in using ontologies is the interoperability by humans and the ability to perform inferences, which makes the domain modeling easier and readable by machines. In order to illustrate how ontologies can be used to model semantically urban issues, in the following we present a domain ontology for urban issues.

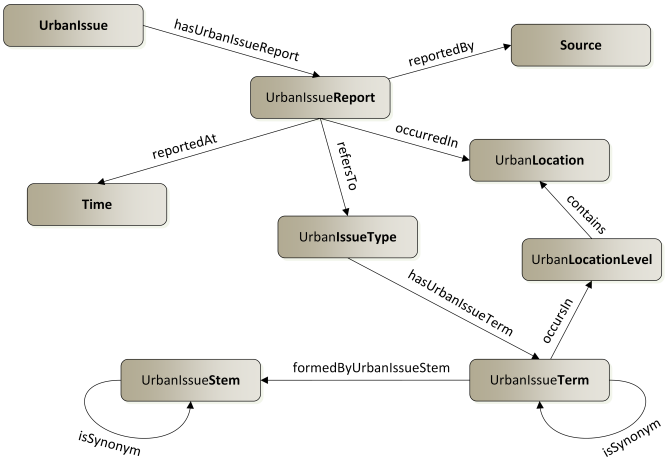

Fig. 1. Main concepts and relationships from the UIDO ontology

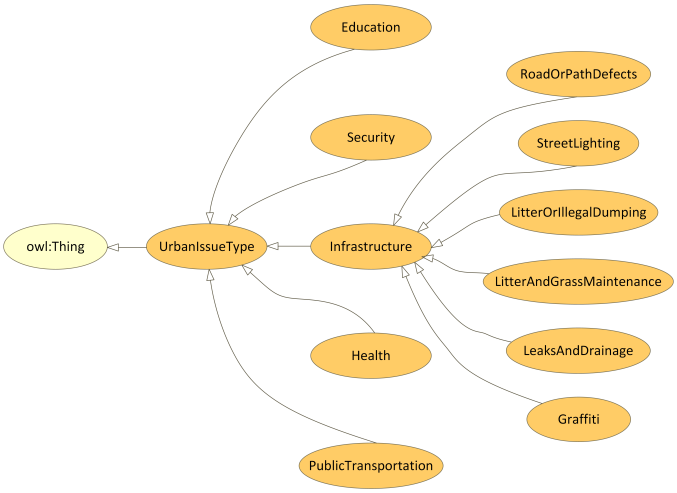

Fig. 2. The urban issue type hierarchy

A. The UIDO ontology

We designed the UIDO - Urban Issues Domain Ontology for semantic modeling of urban issues [9], [31]. The design phase for the UIDO ontology started by analyzing the CContology [32], a domain ontology developed to support online customer complaint management. Such ontology contains seven main classes: Complaint Problems, Complaint Resolutions, Complaint, Complainant, Complaint-Recipient, Address, and Contract. From these, only four classes would be relevant to create the UIDO ontology: Complaint, Complainant, Complaint Problems and Address. However, these four concepts could not be directly reused because they focus on specific customer complaints that are quite different from the complaints in the urban issues domain. Therefore, only the CContology design ideas were used for defining the concepts and relationships for the UIDO ontology.

The UIDO ontology was developed using the Web Ontology Language (OWL) [33] using the Protégé 5.0 system. The core of the UIDO ontology is the classes UrbanIssueReport. this class has two main goals: to act as a top concept during the reasoning process on discovering urban issues from a text; and to be a metamodel for the Linked Data entities from the social media messages related to urban issues. Figure 1 shows the main concepts and relationships of the UIDO ontology. These concepts and relationships make the basic skeleton, but there are others in the sublevels of the hierarchy.

The classes UrbanIssueType and UrbanLocationLevel (shown in Figure 1) are explained in more detail because they are the most important concepts beyond the core. The core class *UrbanIssueReport* represents the reports regarding urban issues. The "user" attribute is encapsulated into the *Source* class as well as the "description" attribute. The *UrbanIssue* class groups urban issue reports of the same type, location and time by the *hasUrbanIssueReport* relationship.

The *UrbanIssueType* class models the classification of issues related to the urban environment. There are initially five subclasses for urban issue types: education, security, infrastructure, health and transportation. Figure 2 shows the UrbanIssueType hierarchy. The six issue types learned from the FixMyStreet corpora fits in the infrastructure concept. Thus, they are included as *Infrastructure* subclasses. Other corpora and specialist knowledge would be necessary in order to expand the ontology within other classes and subclasses that may be further included.

Figure 3 shows an example of two instances ("broken bollard" and "pavement damage") from the classes *UrbanIssueTerm* and *UrbanIssueStem* in a relationship with the issue type "Road or Path Defects" from the *UrbanIssueType* class.

The *UrbanLocationLevel* class models the level of detail for geographical locations. Figure 4 shows the urban location level hierarchy. Such levels are used in constraints for urban issues locations according to the urban issue term. These constraints are modeled through the relationship *occursIn*. For example, a detected urban issue report about a pothole is relevant if it occurred at street or POI (Point of Interest) levels. However, at higher levels such as city or county, such a report tends to be irrelevant due to the geographic vagueness. There are initially three subclasses: Street/Road, District and POI. Nevertheless, other levels may be further included according to new studies. The semantics of such instances are explained in details in the following section.

The *UrbanLocation* class models the geographical information of urban issue reports. Such class reuses concepts from the LinkedGeoData ontology (LGDO) [34]. The LGDO ontology was chosen for the spatial locations related to urban issues for the following reasons: 1) It is derived from concepts defined by the OpenStreetMap [35]; 2) OpenStreetMap is currently a huge up-to-date and open spatial database that contains spatial features, mostly inside urban areas, such as street networks, which are commonly found in urban issue reports; 3) A preliminary work (presented in the following section) that focused on the geographical facet of urban issues enables to store spatial features from OpenStreetMap and provides the

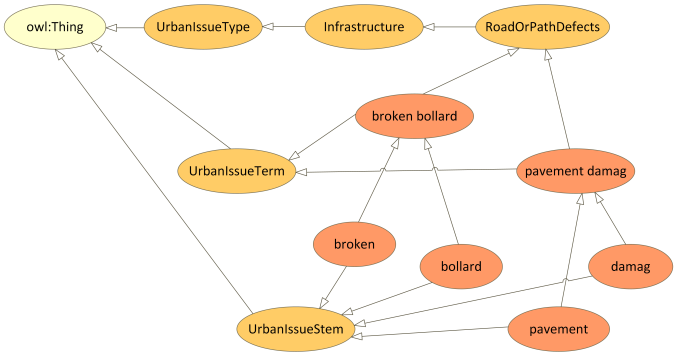

Fig. 3. An example of urban issue stems and terms related to an urban issue type

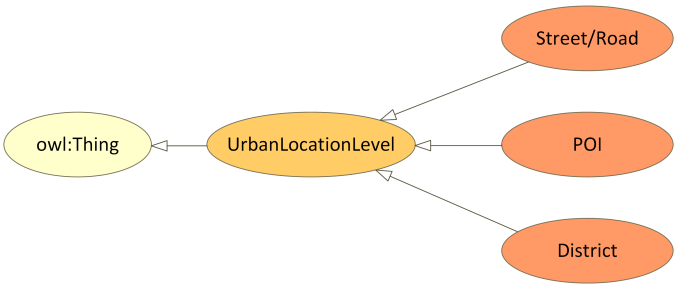

Fig. 4. The urban location level hierarchy

location levels for the UIDO ontology.

The *Time* class models the temporal information of an urban issue. The temporal information can be assigned to the time which the social media message was posted or it can be extracted from the social media message using temporal parsers. Such class reuses concepts from the OWL-Time ontology [36]. The OWL-Time ontology models temporal concepts describing the temporal properties of resources in the world or described in Web pages such as urban issues reported in social media. Thus, the class *Time* in the UIDO ontology is equivalent to the *TemporalEntity* in the OWL-Time.

Finally, the classes *UrbanIssueTerm* and *UrbanIssueStem* model the key terms learned during the knowledge acquisition stage. These concepts are in charge of connecting the top concepts modeled with the terms learned from the analyzed corpus. A set of relevant stems are instances of the urban issue stem concept instead of words (elements from W_{ui}) as a stem is the root of a word and thus it better represents a group of words with similar semantics. The combined stems produce a set of relevant terms T_{ui} through the relationship *formedByUrbanIssueStem*. These terms are instances of the urban issue term concept. Then, subsets of these relevant terms compose the vocabulary of urban issue types through the relationship *hasUrbanIssueTerm*. Both term and stems can be synonyms for each other. For those cases, the relationship *isSynonym* applies.

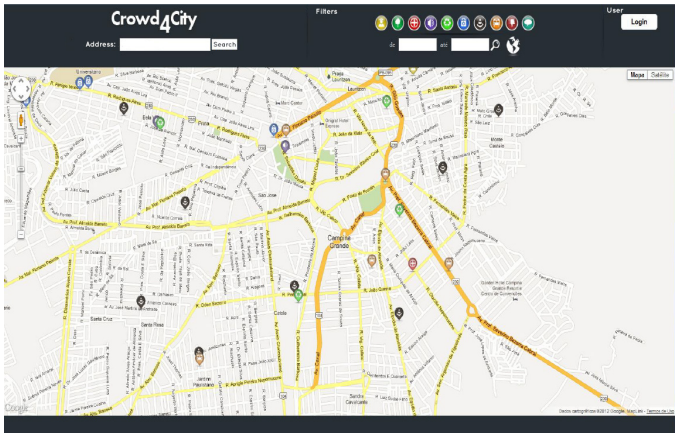

Fig. 5. Crowd4City's main user interface.

IV. The Crowd4City Geosocial Network

The Crowd4City system is a geosocial network aiming at providing e-participation to citizens, which enables them to take part more actively in their city's management, acting as sensors. The Crowd4City users can share and comment on many kinds of geolocated urban issues including traffic jam, criminality, potholes, broken pole lights and so on. Citizen's complaints about urban issues are shared publicly in the Crowd4City aiming to draw the attention of the authorities and the society as a whole. Hence, Crowd4City enables humans as sensors in a smart city environment. Figure 5 depicts the Crowd4City interface in which users can see the spatial distribution and pattern of different topics related to urban issues.

Regarding Crowd4City's use (Figure 5), the citizens can create complaint posts using their personal information or even anonymously, and they can input their dissatisfactions making use of the geographical tools. They can mark a single point on the map where the problem took place (for instance, if the user is reporting a pothole on a street); they can draw lines, perhaps to show routes where there are lighting issues; or they can even draw polygons on the map, thus being able to report regions that can be considered insecure.

Crowd4City presents some predefined categories for the problems reported including: Education, Sanitation, Transportation, Work Under Construction, Security and Others (Noise Pollution, Rubbish, Lighting, Potholes, etc.). However, if the user wishes to report something else, there is a category named "Other", which can be used for such uncategorized complaints.

Crowd4City's posts consist mainly of a location (geographic information), a title, a brief description and optionally multimedia attachments if the user has pictures or videos of the problem being reported. Additionally, the system provides a section for the other users's feedback with like/dislike buttons and a comment section, as seen in Figure 6.

Crowd4City enables operations such as pan and zoom. Also, the system provides several filters so that the users may

perform more specialized searches for their information of interest. There are the basic filters, where the posts may be refined by the selected categories or creation dates; and the advanced filters, which may consider the complaints' contents and their geographic information. With the advanced filters, the users may perform searches using the buffer and contains operations, select some Points Of Interest (POI) categories (such as schools, hospitals, squares, airports and so on) and combine all the available filters.

A serious problem when dealing with information submitted directly from the internet users is the reliability of such data. Malicious users and spammers may use this tool to post irrelevant, erroneous or misleading information or even harmful URLs that can affect the users' devices. In order to bypass this problem, we elaborated a reputation model to be implemented in the Crowd4City geosocial network that evaluates every information submitted to the system and allows the users to better understand if what they are reading could be trustworthy or not. That way, we aim to provide a spam-free environment for Crowd4City, minimizing misinformation.

Our reputation model evaluates every action the users make, such as post creation and comments. That way, we assign a score to each user, which represents how trustworthy they are. When analyzing the content submitted in a new post or comment, the reputation model first checks if there are any of the flagged expressions (commonly used words associated with spam or malicious content, such as "make money" or "lose weight", which were not expected to be used in the context of this geosocial network). According to the number of flagged expressions used, a spam score is attributed to it and that will have on user's reputation.

The second step adopted was the prohibition of uploading possibly harmful files, such as executables (.exe files). However, this is just a precautionary step and does not affect the reputation, since the file to be submitted could be.

We added in Crowd4City the possibility for the users to evaluate every post with "I agree" or "I disagree". With the information retrieved from such feedback, we wish to determine if the content evaluated could be inappropriate, false or incorrect. Posts with a higher rate of "I agree" feedback can indicate true and impartial content. The result of the feedback in a post created by the user affecteds his reputation score. However, the reputation of the user who sends a feedback is also taken into account, since a malicious user may try to ruin other users' reputation.

A more direct way for the users to express their discontentment with users they see that are not using Crowd4City properly is to report them as spammers on their profile pages.

Finally, one last feature is considered, which is the user participation in the system. Users that participate more actively creating post, commenting and evaluating other posts are more likely to have the system's best interests in mind, wishing for Crowd4City to achieve its full potential and be filled with truthful information that could be used to improve the city's conditions and its population's quality of life.

This way, users' reputation score is calculated considering: the reputation score associated with their posts and comments; the number of times they have been flagged as spammers; and how active they are. To each of these points, an expiration date is associated, in case of users change their behavior over time (once malicious users may change their mind and start using the geosocial network correctly).

As a result, this reputation score is shown graphically on the user's profile page with a red stamp (for users with low reputation scores), a white stamp (for neutral users), a bronze stamp, a silver stamp or a gold stamp (the last three for users with a good reputation, according to how good their scores are).

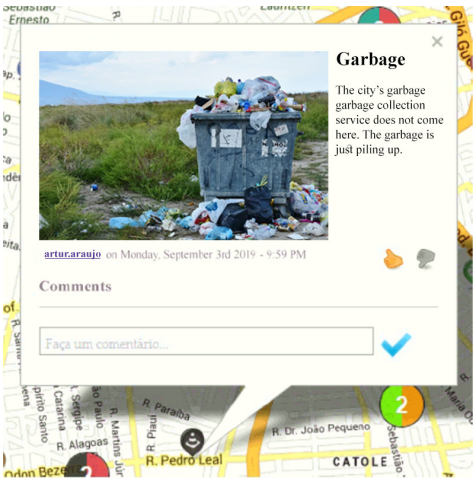

Fig. 6. A rubbish complaint.

V. Automated method for extracting and structuring Urban Issues reported in TV news

The main problem addressed in this research deals with obtaining urban issues complaints from TV news, georeferencing them and automatically classifying them into one of the defined categories. The categories include sanitation, transportation, work under construction, among others. The urban issues context considered in this work is based on a corpus built in a previous work [9].

Our methodology comprises the following steps, according to Figure 7. First, we implemented a Web scraping method to extract the audio from video news. Second, we convert the audio into text using a speech recognition tool. Third, we use a gazetteer to perform geoparsing on the mentioned addresses and locations obtained from the Named Entity Recognition (NER) process, without preprocessing. Then, we implemented a preprocessing step comprising word capitalization, stopwords removal and lemmatization. Fourth, we use NER to obtain the named entities from the text. Then, we perform topic modeling to obtain the class of urban issues related to the text. Finally, the urban issues put into the Crowd4City geosocial network. We detail each step of our methodology next.

A. Web scraping - Video 2 Txt

Initially, we developed a Web scraping tool for obtaining the videos from a TV news website. The data comes from a Brazilian TV broadcast website in Portuguese. We used the Selenium library [37] and YouTubeDL [38] to download the audio files from the video URLs that were stored in a JavaScript Object Notation (JSON) file. Then, we used the SpeechRecognition library [39] with the Google Speech Recognition API to convert audio into text. In order to decode the speech into text, groups of vectors are matched to one or more phonemes, which is a fundamental unit of speech. The SpeechRecognition library relies on modern speech recognition systems based on neural networks and Voice Activity Detectors (VADs). In addition, Google Speech Recognition API is free and supports Brazilian Portuguese language with good results.

B. Preprocessing

In the preprocessing step, we converted the text into lower case, removed stopwords and performed lemmatization. We used the Spacy library [40] to perform entity recognition of locations. The Natural Language Toolkit (NLTK) Python library [41] was also used for the lemmatization process.

The strategy defined for the preprocessing was to extract words that are entities from locations using the Spacy NER, for which the library performs very well when aided by the SpeechRecognition tool. Spacy also offers support for the Portuguese language, which avoids the translation of all texts into English, as it may reduce performance.

Spacy recognizes the location entities of the text and their title, so we combine all the location entities found to then search for those addresses and choose the one with the highest reliability.

We also have guaranteed anonymization by removing the names of people that took part in the audio extracted from video URLs. Hence, privacy was preserved, although it is important to note that all the videos processed in this work are publicly accessible from the sources.

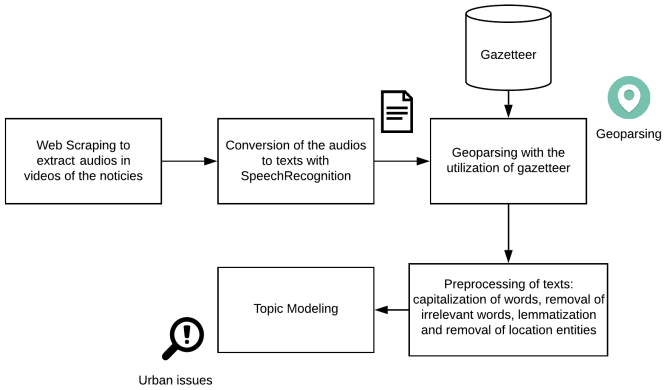

Fig. 7. Our proposed methodology.

C. Geoparsing

We used the Geocoder library [42], which offers an API that enables the use of geocoding services such as Google, ArcGIS, and Bing. The chosen API service was ArcGIS, which provides simple and efficient tools for vector operations, geocoding, map creation and so on. Brazil is ranked level 1 in the library, which means that an address lookup will usually result in accurate matches to the “PointAddress” and “StreetAddress” levels, which fulfills our requirements. After having the entities properly combined, we iterate through this structure by checking which one is the most accurate address (among the ones resulted by the Geocoder using ArcGIS). The accuracy is increased by filtering the addresses found, so the user may perform filtering by state, city or even geographic coordinate.

We used the Open Street Map (OSM) to obtain spatial data from some cities of the State of Paraiba in Brazil, and a gazetteer to improve the geoparsing accuracy. The gazetteer contains streets, neighborhoods, roads, schools, hospitals, supermarkets, pharmacies, etc. Notice that we do not deal with place name pronunciation, as the audio files do, because the names of the places were converted into text. We performed a cleaning of this data to keep only the information of interest to us: name, type and coordinates. Such cleaned data was stored in a PostgreSQL/PostGIS database system. Figure 8 presents the geoparsing step.

D. Topic Modelling

Concerning the topic modeling, we used Gensim [43], an open-source library for unsupervised topic modeling and NLP, which provides statistical machine learning tools. We used LDA from Gensim (LDAMulticore and LDAModel) to implement topic modeling, which considers each document as a collection of topics, and each topic as a collection of keywords.

In order to implement a topic classifier in Gensim, we needed to follow a few steps: creating both a word dictionary and a corpus (bag of words), then providing the desired number of topics and some algorithm tuning parameters. The word dictionary chooses an identifier for each words contained in documents. The corpus (bag of words) was a dictionary containing a set of the word identifiers and the number of occurrences of each word along the document. TF-IDF was also used, transforming the corpus co-occurrence matrix into a local TF-IDF co-occurrence matrix.

Concerning topic modeling, we removed all the words that were location entities, as they were not useful for the class classification process, aiming at increasing the accuracy of the model. Thus, our classifier focused on words of a given class, without worrying about locations.

To find out the best number of topics, some tests were performed and then we identified the best model comparing the evaluated models using the measure coherence score, which evaluates the quality of the obtained topics. After these tests, we concluded that the best number of data topics would be four, as shown in Figure 9.

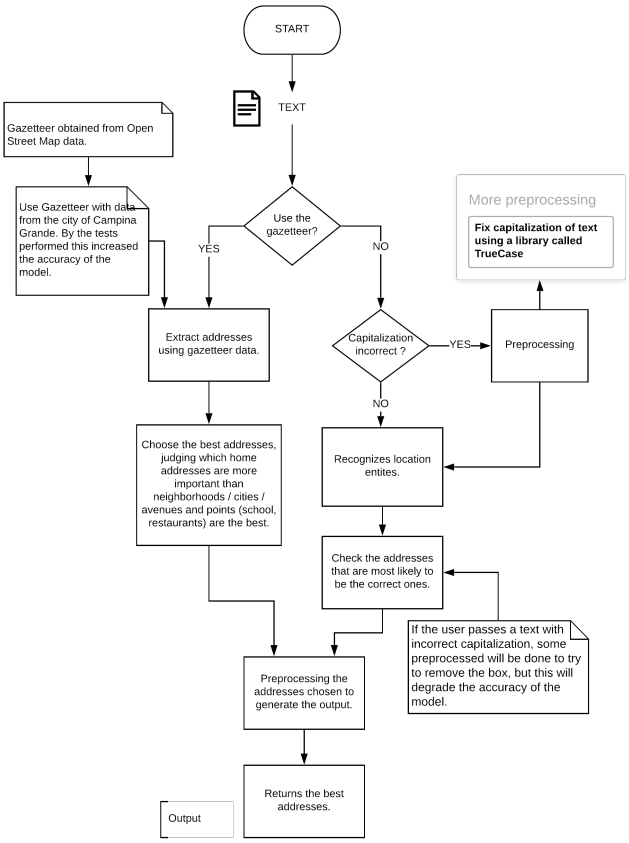

Fig. 8. The geoparsing process.

With four topics, the algorithm achieve a coherence score of 0.527613, the best result in the used dataset. Another improvement was the generation of the 15 most in the topics generated by the algorithm. After that, we manually selected the words that should not be considered and we added them to the list of stop words. Then, we repeated the process until the 10 words in each topic were strongly related to the topic.

In topic modeling, we can analyze which topics represent all documents and also the keywords of each topic. Figure 10 shows the thirty most frequent words in the first topic and also presents the words of the first topic sorted according to their relevance. The most important words were water, sewage, pavement, and home, thus indicating that the topic addresses sanitation problems.

VI. A CASE STUDY IN CAMPINA GRANDE NEIGHBOURHOOD

Usually, urban issues raises the attention of the local press, in order to establish a connection between population and city councils. In Campina Grande, a 400,000 inhabitants Brazilian city, citizens report their complaints to local TVs through messaging service platforms.

Thus, this research aims to fill the gaps mentioned above, helping to share complaints and providing a centralized means

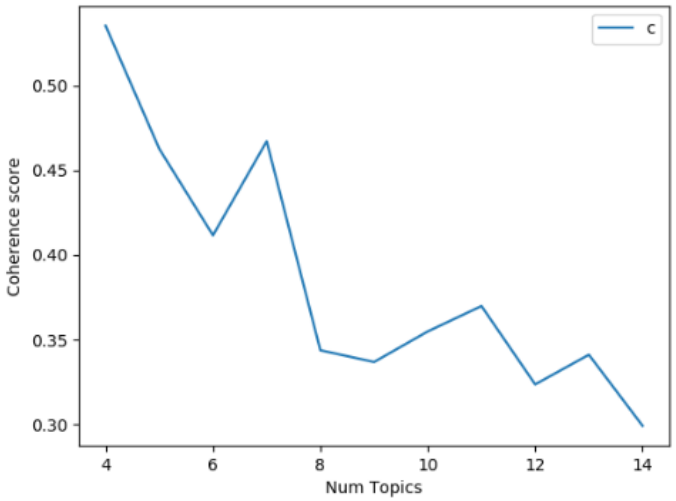

Fig. 9. Coherence score per number of topics.

with this information, it makes it easier for both the inhabitants to make complaints and for the local authorities to solve the reported problems.

A. Setup

In this research, we collected 1,007 videos of the news story "My Neighborhood on TV", covering the years 2016 to 2019, with an average duration of five minutes per video. We took all videos from the Paraiba TV news program website [44].

From all the videos obtained, in 602 of them (59.8 %) it was possible to obtain the text and the locations. Unfortunately,

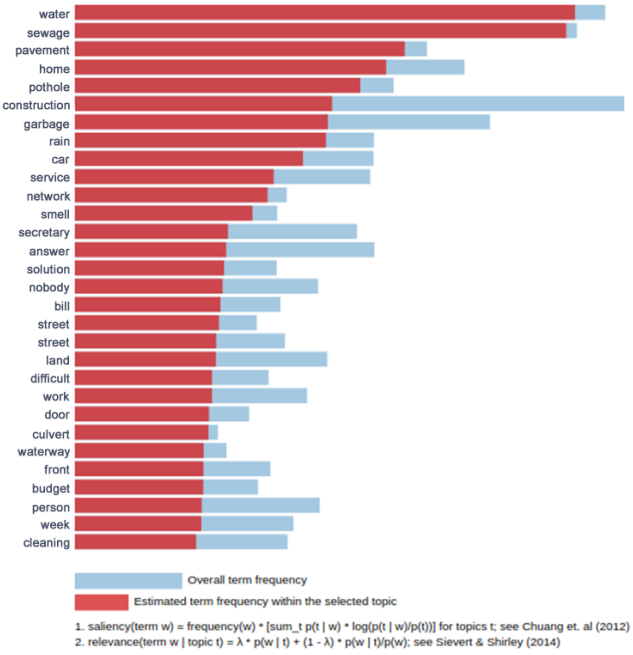

Fig. 10. Most frequent words in the first topic.

TABLE I
NUMBER OF TEXTS PER TOPIC

Class	Number of instances
Streets	308
Other (construction works)	236
Other (sanitation)	108
Education	91
Garbage	79
Other (traffic)	64
Complaints	26
Health	24
Security	17
Other (water shortage)	17
Public transportation	14
Forestation	5
Other	1

some videos did not specify the location. At the same time, some videos did not specify the urban issue, as the word "obra", which means "something not concluded that is being built or repaired" in Portuguese, is applied as a problem generalization.

We extracted the problem classes reported in the videos, enabling various applications to use them in an attempt to improve city management. Municipality authorities may be notified to provide solutions for urban issues.

When analyzing the data, we could see that there was a clear imbalance. Table I shows the arrangement of classes and the respective number of elements in each one.

As we can see in Table I, the majoritarian class had 308 elements, while there were many classes with less than 50 elements. However, as pointed by Branco et al. [45], when using unbalanced datasets, problems such as the use of biased assessment metrics to enrich the performance of the models in less frequent cases will occur. Also, it is necessary to force machine learning algorithms to focus on these less frequent cases.

Therefore, we carry out tests with all classes and decided to with performing use only the four classes with over 90 elements: Education, Other (sanitation), Other (construction works), and Streets.

B. Results

We performed several tests in the Gensim library, using different pre-processing functions and tests different values for their parameters. We set the number of steps equals to 10 because we saw that with this number we obtained better results without losing performance (see Figure 11). When we tried to use values below or above 10, the accuracy decreased.

We used Coherence Score to test the topics, which measures the relative distance between the words within a topic. The number of parameters used was 4, with a score of 0.52. Such a score is acceptable in this preliminary study due to our dataset. A problem using geoparsing is entity recognition. This is because the tools used for NLP cannot recognize entities that are misspelled (for example, if someone wrote the Campina Grande entity lowercase). However, this problem was mitigated with the use of the TrueCase library [46], which corrected the capitalization of words, so that the geoparsing could recognize the entities, obtaining good accuracy. It is important to notice that NLP is by definition not capable of getting the real meaning of any term or context, as text is something by nature completely different than language. In order to deal with context, we needed to combine NLP with other resources such as Part-Of-Speech tagging and supervised machine learning algorithms for instance.

The TrueCase library supports the English language only. As our data was in Portuguese, we needed to use a library to translate the words from Portuguese to English - GoogleTrans [47] - use the TrueCase and then do the reverse process, resulting the words in Portuguese with the correct capitalization.

As an additional process to improve the performance of geoparsing, we used a gazetteer, to improve in the geolocation process for texts of the city of Campina Grande - which is the object of this research.

Regarding topic classification, we analyzed the F-Score: 1) With all 13 classes and 993 complaints, and 2) Removing the classes with less than 90 instances, with 743 complaints remaining. In both cases, we carried out the same pre-processing (stemming, TF-IDF, etc.). Besides, we used two classification algorithms and compared their results. The classification algorithms used in this study were XGBoost [48] and a Bi-LSTM neural network (Bidirectional LSTM). [49].

Using XGBoost with all classes we obtained an a F-Score of 68.55%. After removing classes with few instances, we improved the results and obtained an F-Score of 83.34%. In Figure 12, we can see the F-Score with full data and only with the major classes. On the other hand, using the Bi-LSTM with all classes, we obtained an F-Score of 32%. With the removal of minority classes, the F-Score increased to 75%, as we can see in Figure 13.

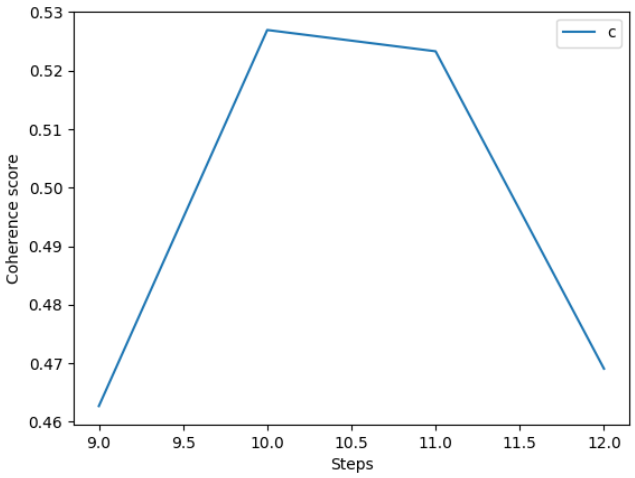

Fig. 11. Comparative chart for influence of the steps value in the coherence score.

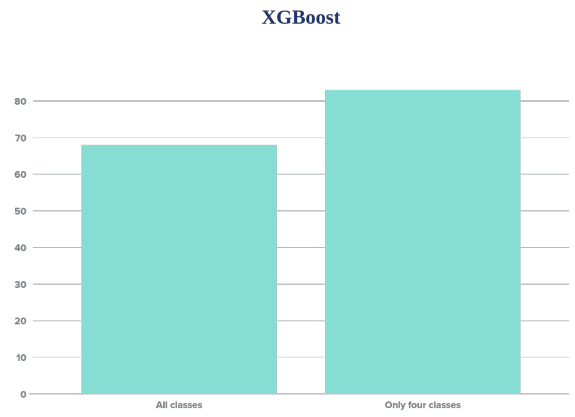

Fig. 12. Comparison between the results with full data and with four classes using XGBoost.

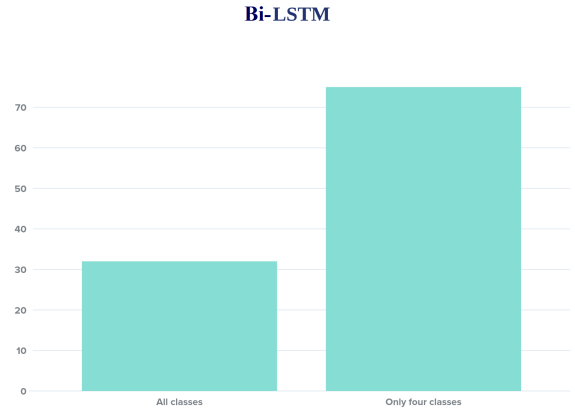

Fig. 13. Comparison between the results with full data and with four classes using Bi-LSTM.

VII. Conclusion

Citizens as sensors enable the engagement of society through technology to complain about urban issues. Smart cities demand tools for such engagement promoting e-citizenship and e-participation. Nonetheless, although some of such tools have already been proposed, it turns out that people engagement decreases in time. Hence, the gathering of urban issues from any media is very important to keep people engaged. As such, this article proposes an approach to gather urban issues data from a TV news program using geoparsing and NLP techniques to locate and classify the urban issues in order to input it in the Crowd4City geosocial network.

The results showed that our approach is feasible and that we manage to classify urban issues into four topics: mobility, sanitation, buildings and others. As future work, we plan to perform an in-depth performance analysis of geoparsing, as well as topic modeling, by manually identifying the topics of the videos as ground truth and comparing them with the topic modeling results. Another plan consists of performing a comparative study between topic modeling and supervised machine learning.

References

[1] R. Rich Elton, F. Anderson, B. Cláudio, F. Ana Gabrielle, O. Maxwell, and A. Fabio, "Using Natural Language Processing for Extracting GeoSpatial Urban Issues Complaints from TV News." Valencia, Spain: IARIA, Mar. 2020, pp. 55–60.

[2] A. G. R. Falcão et al., "Towards a reputation model applied to geosocial networks: a case study on crowd4city," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing, SAC*, Pau, France, 2018, pp. 1756–1763.

[3] T. Wandhofer, C. van Eeckhaute, S. Taylor, and M. Fernandez, "We-Gov analysis tools to connect policy makers with citizens online," in *Proceedings of the tGovernment Workshop*, 2012, pp. 1–7.

[4] N. Walravens, "Validating a Business Model Framework for Smart City Services: The Case of FixMyStreet," in *Proceedings of the 27th International Conference on Advanced Information Networking and Applications Workshops*, 2013, pp. 1355–1360.

[5] D. Britz, A. Goldie, M.-T. Luong, and Q. V. Le, "Massive exploration of neural machine translation architectures," *ArXiv*, vol. abs/1703.03906, 2017.

[6] S. Reddy, D. Chen, and C. D. Manning, "Coqa: A conversational question answering challenge," vol. Transactions of the Association for Computational Linguistics, Volume 7, March 2019, pp. 249–266. [Online]. Available: https://www.aclweb.org/anthology/Q19-1016 [accessed: 2020-03-02]

[7] H. Schwenk, L. Barrault, A. Conneau, and Y. LeCun, "Very deep convolutional networks for text classification." in *EACL (1)*, M. Lapata, P. Blunsom, and A. Koller, Eds. Association for Computational Linguistics, 2017, pp. 1107–1116. [Online]. Available: https://www.aclweb.org/anthology/E17-1104/ [accessed: 2020-03-02]

[8] A. Radford et al., "Language models are unsupervised multitask learners," 2018. [Online]. Available: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf [accessed: 2020-03-02]

[9] M. G. de Oliveira, C. de Souza Baptista, C. E. C. Campelo, and M. Bertolotto, "A Gold-standard Social Media Corpus for Urban Issues," in *Proceedings of the Symposium on Applied Computing (SAC)*, ser. SAC '17. New York, NY, USA: ACM, 2017, pp. 1011–1016.

[10] N. Camelin et al., "FrNewsLink : a corpus linking TV Broadcast News Segments and Press Articles," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. [Online]. Available: https://www.aclweb.org/anthology/L18-1329 [accessed: 2020-03-02]

[11] R. Kannao and P. Guha, "Overlay Text Extraction From TV News Broadcast," *CoRR*, vol. abs/1604.00470, 2016. [Online]. Available: http://arxiv.org/abs/1604.00470 [accessed: 2020-03-02]

[12] R. Kannao and P. Guha, "A system for semantic segmentation of tv news broadcast videos," *Multimedia Tools and Applications*, vol. 79, no. 9, pp. 6191–6225, 2020.

[13] A.-G. Chifu and S. Fournier, "Segchain: Towards a generic automatic video segmentation framework, based on lexical chains of audio transcriptions," in *Proceedings of the 6th international conference on web intelligence, mining and semantics*, 2016, pp. 1–8.

[14] S. Iwata, W. Ohyama, T. Wakabayashi, and F. Kimura, "Recognition and connection of moving captions in arabic tv news," in *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*. IEEE, 2017, pp. 163–167.

[15] M. Pala, L. Parayitam, and V. Appala, "Real-time transcription, keyword spotting, archival and retrieval for telugu TV news using ASR," *International Journal of Speech Technology*, vol. 22, no. 2, pp. 433–439, 2019.

[16] R. Bansal and S. Chakraborty, "Visual Content Based Video Retrieval on Natural Language Queries," in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, ser. SAC '19. New York, NY, USA: ACM, 2019, pp. 212–219.

[17] Z. Dong and X. Lv, "Subject extraction method of urban complaint data," in *Proceedings of the IEEE International Conference on Big Knowledge (ICBK)*, 2017, pp. 179–182.

[18] B. Mocanu, R. Tapu, and T. Zaharia, "Automatic extraction of story units from TV news," in *Proceedings of the IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2017, pp. 414–415.

[19] T. Zlitni, B. Bouaziz, and W. Mahdi, "Automatic topics segmentation for TV news video using prior knowledge," *Multimedia Tools and Applications*, vol. 75, no. 10, pp. 5645–5672, 2016.

[20] Z. Liu and Y. Wang, "TV News Story Segmentation Using Deep Neural Network," in *Proceedings of the IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2018, pp. 1–4.

[21] T. R. Gruber *et al.*, "A translation approach to portable ontology specifications," *Knowledge acquisition*, vol. 5, no. 2, pp. 199–221, 1993.

[22] M. Gruninger, "Ontology: applications and design," *Commun. ACM*, vol. 45, no. 2, 2002.

[23] C. Roussey, F. Pinet, M. A. Kang, and O. Corcho, "An introduction to ontologies and ontology engineering," in *Ontologies in Urban development projects*. Springer, 2011, pp. 9–38.

[24] I. Peters, *Folksonomies. Indexing and retrieval in Web 2.0*. Walter de Gruyter, 2009.

[25] X. Yang, R. Gao, Z. Han, and X. Sui, "Ontology-based hazard information extraction from chinese food complaint documents," in *International Conference in Swarm Intelligence*. Springer, 2012, pp. 155–163.

[26] D. C. Wimalasuriya and D. Dou, "Ontology-based information extraction: An introduction and a survey of current approaches," 2010.

[27] O. Corcho, "Ontology based document annotation: trends and open research problems," *International Journal of Metadata, Semantics and Ontologies*, vol. 1, no. 1, pp. 47–57, 2006.

[28] D. W. Embley, D. M. Campbell, R. D. Smith, and S. W. Liddle, "Ontology-based extraction and structuring of information from data-rich unstructured documents," in *Proceedings of the seventh international conference on Information and knowledge management*, 1998, pp. 52–59.

[29] D. M. Blei, "Introduction to probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2011.

[30] A. Lesnikowski, E. Belfer, E. Rodman, J. Smith, R. Biesbroek, J. D. Wilkerson, J. D. Ford, and L. Berrang-Ford, "Frontiers in data analytics for adaptation research: Topic modeling," *Wiley Interdisciplinary Reviews: Climate Change*, vol. 10, no. 3, p. e576, 2019.

[31] M. G. d. Oliveira, "Ontology-driven urban issues identification from social media." Ph.D. dissertation, Federal University of Campina Grande (UFCG), 2016.

[32] M. Jarrar, "Towards effectiveness and transparency in e-business transactions: an ontology for customer complaint management," in *Semantic Web for Business: Cases and Applications*. IGI Global, 2009, pp. 127–149.

[33] D. L. McGuinness, F. Van Harmelen *et al.*, "Owl web ontology language overview," *W3C recommendation*, vol. 10, no. 10, p. 2004, 2004.

[34] C. Stadler, J. Lehmann, K. Höffner, and S. Auer, "Linkedgeodata: A core for a web of spatial open data," *Semantic Web*, vol. 3, no. 4, pp. 333–354, 2012.

[35] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008.

[36] S. Cox, C. Little, J. Hobbs, and F. Pan, "Time ontology in owl," *W3C recommendation*, vol. 19, 2017.

[37] Selenium, "Selenium Library." [Online]. Available: https://www.selenium.dev/ [accessed: 2020-03-02]

[38] R. Gonzalez et al., "YouTubeDL." [Online]. Available: https://github.com/ytdl-org/youtube-dl [accessed: 2020-03-02]

[39] A. Zhang, "Selenium." [Online]. Available: https://github.com/Uberi/speechrecognition [accessed: 2020-03-02]

[40] Explosion AI, "Spacy." [Online]. Available: https://spacy.io/ [accessed: 2020-03-02]

[41] NLTK Project, "The Natural Language Toolkit." [Online]. Available: https://radimrehurek.com/gensim/ [accessed: 2020-03-02]

[42] D. Carriere et al., "Geocoder." [Online]. Available: https://geocoder.readthedocs.io/ [accessed: 2020-03-02]

[43] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora." [Online]. Available: https://radimrehurek.com/gensim/ [accessed: 2020-03-02]

[44] G1 Paraiba, "JPB1 TV News program official website." [Online]. Available: http://g1.globo.com/pb/paraiba/jpb-1edicao/videos/ [accessed: 2020-03-02]

[45] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 31:1–31:50, 2016.

[46] D. Fury, "TrueCase." [Online]. Available: https://github.com/daltonfury42/truecase [accessed: 2020-03-02]

[47] S. Han, "Googletrans." [Online]. Available: https://github.com/ssut/py-googletrans [accessed: 2020-03-02]

[48] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 785–794.

[49] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602 – 610, 2005, iJCNN 2005.