Using Natural Language Processing for Extracting GeoSpatial Urban Issues Complaints from TV News

Rich Elton Carvalho Ramalho, Anderson Almeida Firmino,
Cláudio de Souza Baptista, Ana Gabrielle Ramos Falcão
and Maxwell Guimarães de Oliveira

Information System Laboratory
Computer Science Department
Federal University of Campina Grande (UFCG)
Campina Grande - PB, Brazil
Email: rich.ramalho@ccc.ufcg.edu.br,
andersonalmeida@copin.ufcg.edu.br,
anagabriellee@gmail.com,
baptista@computacao.ufcg.edu.br,
maxwell@computacao.ufcg.edu.br

Fábio Gomes de Andrade

Federal Institute of Paraíba (IFPB)
Cajazeiras - PB, Brazil
Email: fabio@ifpb.edu.br

***Abstract*—Citizens as sensors enable the engagement of society through technology to complain on urban issues. Despite the fact that some geosocial networks have been developed in recent years to enable citizens to report many types of urban problems, it is possible to notice that the engagement of the users of these networks usually decreases in time. Hence, many relevant issues are not identified or published, which reduces the effectiveness of these networks. Aiming to overcome this limitation, this paper proposes an approach in which urban issues are automatically detected from a TV news program. The proposed solution uses geoparsing and Natural Language Processing (NLP) techniques to geocode and classify the identified complaints and publishes the results in Crowd4City, a geosocial Network that deals specifically with urban issues. Finally, our method was evaluated using data of a real news TV program in Brazil. Our results indicate 59.8% of success on extracting text and location from the video news.**

Keywords–Geosocial network; NLP; Urban Issues; Crowdsourcing.

I. Introduction

The high concentration of population in urban areas has imposed to 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 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 the context of Smart Cities, geosocial networks enable the use of context aware services to locate users and their complaints on urban issues. Several tools, such as Crowd4City [1], Wegov [2] and FixMyStreet [3] have been proposed in order to provide the citizen an opportunity to complain on urban issues. However, people's motivation in using such geosocial networks decrease in time. Hence, to ensure a high engagement of society, different approaches to gather information are required.

Several local TV stations in Brazil portray urban issues reported by the community. An example is the 'Calendar' board in a daily open TV channel news program in the State of Paraíba, Brazil. That news broadcast exhibits several urban issues faced by the main cities from that particular state. Hence, it is important to gather those urban issues and input them into geosocial networks in order to improve citizenship and increase awareness. The audio descriptions in the news channel need to be converted into text, then geoparsing tools from Geographic Information Systems (GIS) and Natural Language Processing (NLP) techniques need to be used to automatically extract the correct location of the respective urban issues.

In this paper, 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 are up-to-date and extracted from a real context. Our main contribution consists in the integration of GIS and NLP.

The remainder of this paper 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 has 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. [4] discuss machine translation using a seq2seq model. Reddy et al. [5] present a question answering approach. Schwenk et al. [6] focus on text classification. Radford et al. [7] propose a language model using unsupervised learners. NLP is difficult to accomplish as 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 corpus. Oliveira et al. [8], for instance, contributed with the creation of a gold-standard corpus of urban issues related tweets in the English language, including geographical information. 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. [9] composed a corpus of different TV Broadcast News from French channels and online press articles, which were manually annotated in order to obtain topic segmentation annotations and linking annotations between topic segments and press articles. They made the FrNewsLink available online for anyone who wishes to use it in their studies. 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 [10] focused their study on extracting text from the overlay banner presented in such broadcasts. Such text usually contains brief descriptions of news events and, since they may be in various formats. Therefore, they proposed a contrast enhancement preprocessing stage and a parameter free edge density based scheme for better text band and text extraction. They also performed experiments using Tesseract Optical Character Recognition (OCR) for overlay text recognition trained using Web news articles. The authors confirmed the validity of their approaches using three Indian English television shows and obtained significant results. However, their domain is limited, since only the overlay bands' contents are analyzed.

Similarly, Pala et al. [11] 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 with the keyword, the system does not have the ability to extract the theme or domain being discussed in the video.

Bansal and Chakraborty [12] 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 show they were able to capture the videos' and sentences' semantics when compared to other already existing approaches, however the authors lack retrieving any geographic information.

Dong et al. [13] focused on developing a method for subject words extraction of urban complaint data posted on the Internet. Their approach consisted on 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 approach would need to be validated in other scenarios.

Mocanu et al. [14] proposed a method for such extraction by using temporal segmentation of the multimedia information, allowing it to be indexed and thus be more easily found by the users interested in specific topics. Their approach was based on anchor person identification, where the TV news program presenter would be featured on the video. They performed a few tests with a limited database of French TV programs and obtained good results, however their topic detection is not very robust, since it is based only on the video subtitles.

Zlitni et al. [15] 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, however their approach is completely dependent on the production rules of TV channels. Also aiming at achieving news story segmentation, Liu and Wang [16] 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 on TDT2 dataset, from Topic Detection and Tracking (TDT). Although they obtained significant results, their approach is limited to the linguistic information extracted from closed caption and thus not applicable to programs without such resource.

Even though several studies could be found, none 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 specially on urban issues complaints.

III. 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 on urban issues are shared publicly in the Crowd4City aiming to draw the attention from the authorities and the society as a whole. Hence, Crowd4City enables humans as sensors in a smart city environment. Figure 1 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 1), 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

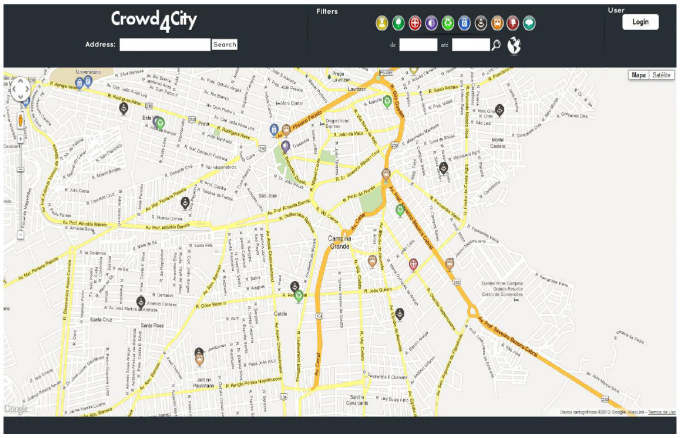


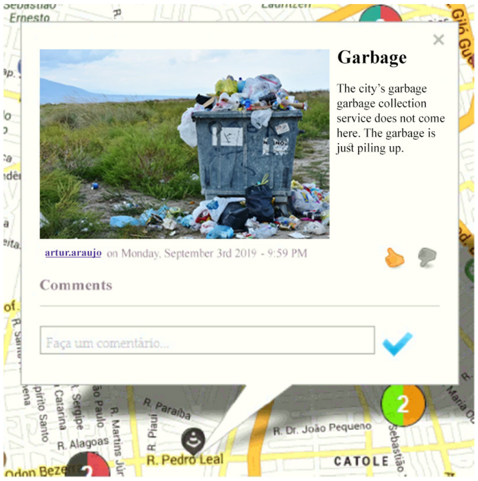
Figure 1. Crowd4City's main user interface.

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: 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 on Figure 2.

Crowd4City enables operations such as pan and zoom. Also, the system made available 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, may select some Points Of Interest (POI) categories (such as schools, hospitals, squares, airports and so on) and all the available filters may be used combined.



Garbage

The city's garbage garbage collection service does not come here. The garbage is just piling up.

artur.araujo on Monday, September 3rd 2019 - 9:59 PM

Comments

Figure 2. A rubbish complaint.

IV. 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 [8].

Our methodology comprises the following steps, according to Figure 3. 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 are located 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 TV news website. The data comes from a Brazilian TV broadcast website in Portuguese. We used the Selenium library [17] and YouTubeDL [18] 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 [19] 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.

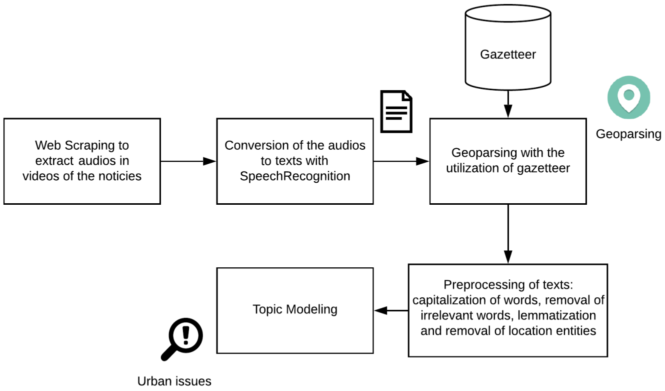

Gazetteer
Geoparsing
Web Scraping to extract audios in videos of the noticies
Conversion of the audios to texts with SpeechRecognition
Geoparsing with the utilization of gazetteer
Preprocessing of texts: capitalization of words, removal of irrelevant words, lemmatization and removal of location entities
Topic Modeling
Urban issues

Figure 3. Our proposed methodology.

Copyright (c) IARIA, 2020. ISBN: 978-1-61208-762-7

B. Preprocessing

In the preprocessing step, we converted the text into lower case, removed stopwords and performed lemmatization. We used the Spacy library [20] to perform entity recognition of locations. The Natural Language Toolkit (NLTK) Python library [21] was also used for the lemmatization process.

The strategy defined for the preprocessing is to extract words that are entities from locations using Spacy NER, for which the library works 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.

C. Geoparsing

We used the Geocoder library [22], that 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 form the addresses the Geocoder returns 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 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, pharmacy, etc. Notice that we do not deal with place names pronunciation, as the audio files do, because the names of the places are 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 4 presents the geoparsing step.

D. Topic Modelling

Concerning the topic modeling, we used Gensim [23], 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 need 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 ID for all the words contained in documents, the corpus (bag of words) is a dictionary with word IDs and how many times that word repeats in the document. TF-IDF was also used, transforming the corpus co-occurrence matrix into a local TF-IDF co-occurrence matrix.

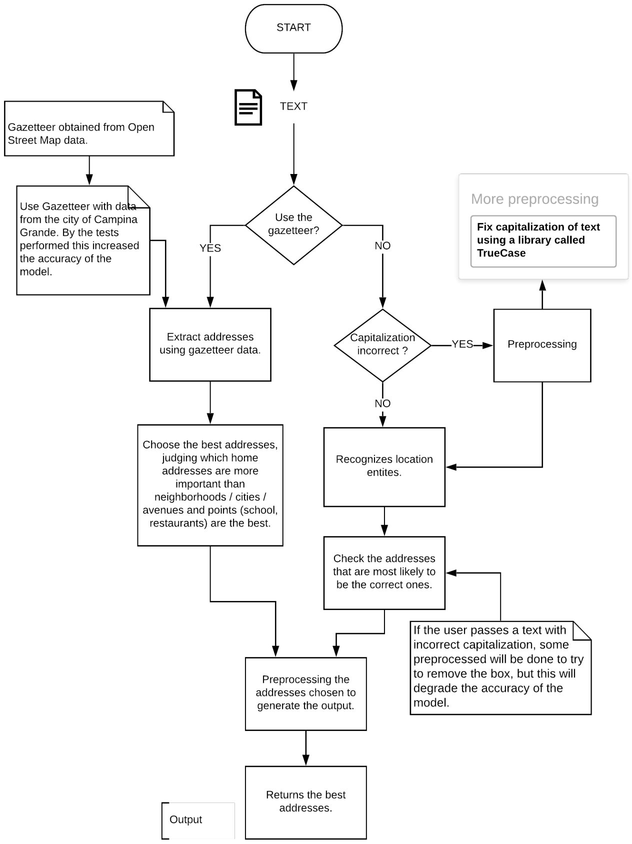

START
TEXT
Gazetteer obtained from Open Street Map data.
Use Gazetteer with data from the city of Campina Grande. By the tests performed this increased the accuracy of the model.
Use the gazetteer?
YES
NO
More preprocessing
Fix capitalization of text using a library called TrueCase
Extract addresses using gazetteer data.
Capitalization incorrect ?
YES
Preprocessing
NO
Choose the best addresses, judging which home addresses are more important than neighborhoods / cities / avenues and points (school, restaurants) are the best.
Recognizes location entites.
Check the addresses that are most likely to be the correct ones.
If the user passes a text with incorrect capitalization, some preprocessed will be done to try to remove the box, but this will degrade the accuracy of the model.
Preprocessing the addresses chosen to generate the output.
Returns the best addresses.
Output

Figure 4. The geoparsing process.

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

To find out the best number of topics, some tests were performed and then it was verified which was the best model, comparing them with the measure coherence score, which evaluates the quality of the obtained topics. After these tests, we came to the conclusion that the best number of data topics would be four, as shown in Figure 5.

With four topics, the algorithm achieves a coherence score of 0.527613, the best result in the used dataset. Another improvement was to generate the 15 most repeated words 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 6 shows the thirty most frequent words in the first topic and also presents the words of the first topic sorted according to their importance. The most important words are water, sewage,

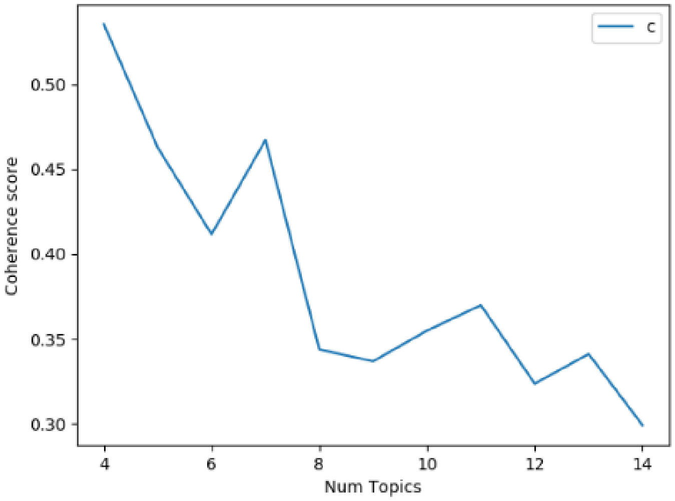


Figure 5. Coherence score per number of topics.

pavement, and home, thus indicating that the topic addresses sanitation problems.

V. A CASE STUDY IN CAMPINA GRANDE NEIGHBOURHOOD

Usually, urban issues raised attention from local press, in order to establish a connection between population and city councils. In Campina Grande, a 400,000 inhabitants Brazilian city, there is a story in a local newspaper called “My Neighborhood on TV” that weekly shows existing urban issues and proposes to notify the authorities to solve a problem, defining a deadline to solve it. In general, 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 with this information, so it is easier for both the inhabitants to make complaints and for the local authorities to solve the reported problems.

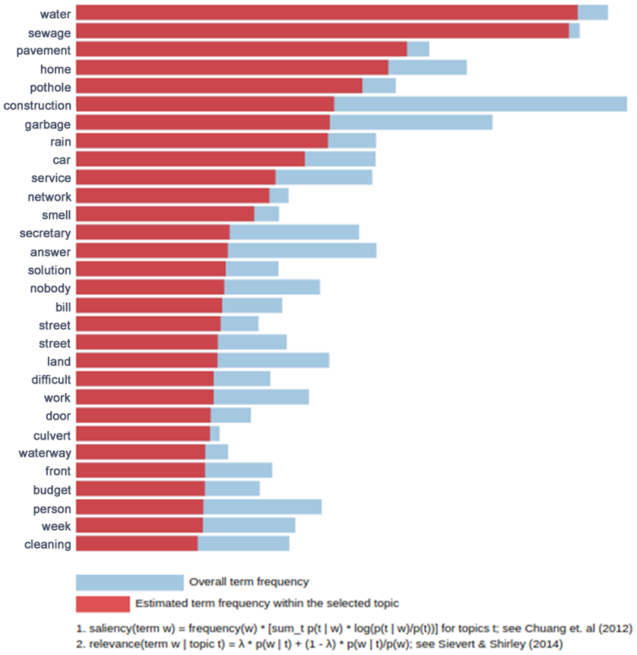


Figure 6. Most frequent words in the first topic.

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’s TV news program website [24].

From all the videos obtained, in 602 of them (59.8 %) it was possible to get the text and locations. Unfortunately, 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, as the authorities can be notified and then provide solutions for urban issues in this easy manner.

B. Results

We have performed several tests in the Gensim library, from changing pre-processing functions to changing parameters of the functions used. We used the number of steps equal to 10 because we saw that with this number we get good results without losing performance (see Figure 7). When trying to use values below or above 10, we saw that the accuracy began to decrease.

The metric we used to test the topics generated was the Coherence Score, which measures the relative distance between 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. We performed tests to verify how the generated model behaved with data not yet seen. One problem when using some geoparsing is in 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 with all lowercase characters). However, this problem was mitigated with the use of the TrueCase library [25], which corrected the capitalization

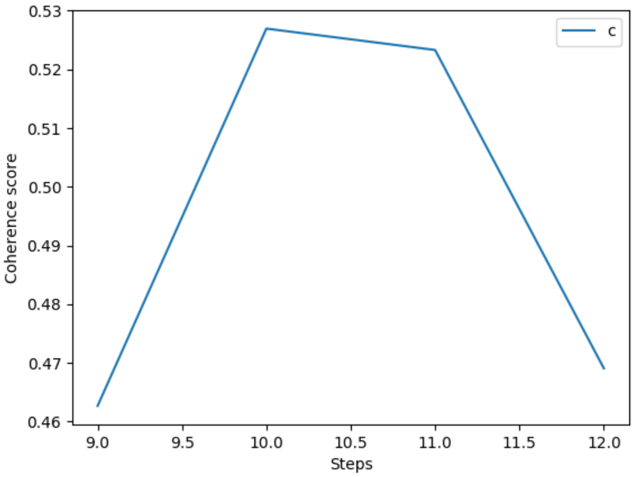


Figure 7. Comparative chart for influence of the steps value in the coherence score.

of words, so that the geoparsing used 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 need to combine NLP with other resources such as Part-Of-Speech tagging and supervised machine learning, for instance.

The TrueCase library supports the English language only. As in our case, the data was in Portuguese, hence we needed to use a library to translate the words from Portuguese to English - GoogleTrans [26] - use TrueCase and then do the reverse process, resulting in the words in Portuguese with the correct capitalization. However, sometimes, this procedure was unsuccessful due to problems in translations or problems in capitalized words.

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

VI. Conclusion

Citizens as sensors enable the engagement of society through technology to complain on 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 decrease in time. Hence, the obtention of urban issues from any media is very important to maintain people's engagement. As such, this paper proposes an approach to gather urban issues data from a TV news program and, using geoparsing and NLP techniques, to locate and classify the urban issues in order to input it in the Crowd4City geosocial network.

The results show 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.

Acknowledgment

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

References

[1] 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.

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

[3] 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.

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

[5] 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]

[6] 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]

[7] 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]

[8] 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.

[9] 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]

[10] 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]

[11] 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, 2019, pp. 433–439.

[12] 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.

[13] 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.

[14] 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.

[15] 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, 2016, pp. 5645–5672.

[16] 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.

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

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

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

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

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

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

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

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

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

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