MADRID
inter.noise 2019
June 16 - 19
NOISE CONTROL FOR A BETTER ENVIRONMENT

Evaluating noise perception through online social networks: A text mining approach to designing a noise-event alarm system based on social media content.

Gascó, Luis[1]
Universidad Politécnica de Madrid (Spain) - EIT Digital Doctoral School Madrid

Asensio, César [2]**; De Arcas, Guillermo**[3]
Universidad Politécnica de Madrid (Spain)

Clavel, Chloé[4]
Télécom ParisTech (France)

ABSTRACT

With the rapid rise of the use of Online Social Networks, people have been sharing their opinions and feelings on the Internet: they write about their personal interests and political opinions, but also about their feelings about noisy activities and sounds they hear during their daily life. This textual information could provide policy makers and city managers with insights about the community response towards specific noisy events in cities that may be useful for improving the management of these activities.

In this paper, we present a methodology to analyze automatically these Internet opinions by using machine learning and Natural Language processing Technologies. This approach has allowed us to build a system that automatically detects and classifies noise complaints by source, using texts written on online social networks as input. We also present a noise-event alarm system based on statistical process control theory that uses the power of our methodology to detect problematic noise events, as well as the reason why those events caused annoyance to population.

Keywords: Text mining, Community engagement, Noise annoyance, Machine Learning, NLP
I-INCE Classification of Subject Number:52, 56, 61, 66, 69

[1]luis.gasco@i2a2.upm.es
[2]casensio@i2a2.upm.es
[3]g.dearcas@upm.es
[4]chloe.clavel@telecom-paristech.fr

1. INTRODUCTION

The Digital Revolution is a phenomenon that we have been living since the 70s, we have experienced a fast evolution of electronics and an adoption of technologies that have resulted in changes in the way we interact with machines and between us [1]. This circumstance, especially exacerbated since the 2000s when the Internet became a global event, has led to the creation of virtual networks in which humans relate to each other digitally. These platforms, commonly known as Online Social Networks (OSN), are now used by the population to express their opinions and feelings on various topics, including politics, sports or the environment.

Parallel to this social phenomenon, we have experienced an increase in the importance of environmental policies in the European Union, and many regulations have appeared for managing of pollutants, including environmental noise. The implementation of Community Directives such as 2002/49/CE have led to greater noise awareness in citizenship and policy makers, who have developed new local laws to control this issue [2].

Technological progress has also changed the manners of policy making. For years now, new policy-making methodologies have been proposed in which citizenship has a greater importance in the decision making, an example of those new methodologies is Policy Making 3.0. [3]. This methodology has been previously adapted to environmental acoustics in the past [4], and the content of this manuscript is somehow included in the development of this methodology, since it considers the use of social networks as a tool to gather information from the population.

In this document we propose a novel approach for assessing subjective response to noise. Our methodology aims to gather, detect, and analyze people's opinion from OSN by using text mining and sentiment analysis techniques. This approach has been explored in other areas; i.e. abuse of drug prescriptions using Twitter data, monitor adverse drug reactions using social media content, and forecasting of the daily Air Quality Index only based on social media content [5–7]. In environmental acoustics, OSN data was used for the first time to diagnose New York noise sources [8, 9]. Researchers were also able to extract the primary noise source in big cities just using photographs metadata from Flickr [10]. More recently, we proposed a new approach to analyze texts from the Internet in environmental acoustic perception [11]: the hypothesis were validated and published last year [12], and now we present them in this paper.

The first section of this paper shows the technological background and the future importance of new text mining technologies in both research, management, and commercial applications. Then, we present a novel methodology, drawn from the areas of machine learning and Natural Language Processing (NLP), to detect noise complaints from OSN automatically. The third section illustrates an example of how city managers could use these noise complaints to detect the attitude towards noisy events in a city. Finally, the last section focuses on discussing future developments in this field.

2. TECHNOLOGICAL BACKGROUND

As stated above, new technologies, especially those related to Artificial Intelligence, such as Machine Learning, NLP and text mining, have been widely studied and developed at a technical level. It is known that these tools are currently used in production by big technology leaders such as Google, Microsoft, or Apple, but also by smaller

companies to get customers' opinion about their products. For example, food franchise enterprises automatically analyze OSN data to know what the people's opinions about new sandwiches are, what is their position on competing products, and to identify new audience segments [13]. Other sectors such as car brands use these technologies to reveal brand perceptions in each market they operate and to build the next marketing campaigns in a more effective way [14].

In fact, the consulting firm Gartner annually develops the Hype cycle on some technologies applied to different branches. These graphs represent the maturity, adoption, and commercial application of specific technologies, which go through five phases during its cycle of life: Innovation Trigger, which is the period when the technology is investigated and developed but there are not products in the market; Peak of inflated Expectation, when some companies decide to use them in real environments; Trough of Disillusionment, when after some fails in applying a technology a disillusionment feeling appears in the industry; Slope of Enlightenment, when the real usability of the technology is found and most of the companies use it; and Plateau of Productivity, when the mainstream adoption starts to take off and the usefulness of the technology is clear [15].

The Figure 1 shows the last available hype cycle for Data Science. When we take a look at the graph, we can see that technologies such as Deep Neural nets or Machine Learning are at the point of "Peak of inflated Expectations". As previously said, these technologies could be potentially be applied in many branches, hence the industry is very interested in them. But only big technological giants like Google and Amazon, or small technological startups are applying them in 100% in their processes.

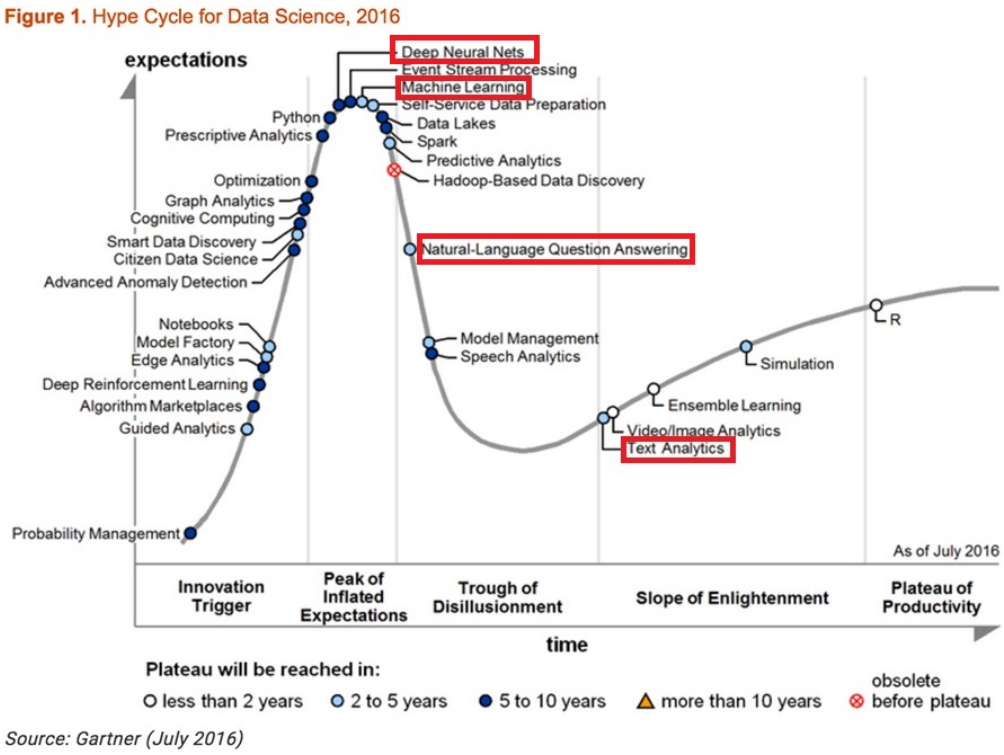

Figure 1: Gartner Hype Cycles for emerging technologies and Data Science. Source: Gartner

However, when we go to technologies such as Text Analytics, which is based on Machine Learning and NLP, we observe that it has already reached a technological maturity that allows its commercial application by many companies. Furthermore, Gartner forecasts that it will be applied by the majority of companies in the next 2 years. But, why is it that text mining technologies have not been applied to environmental acoustics yet? Its application could be very useful in topics ranging from automatic analysis of noise complaints, to measurement of public acceptance of noisy activities analyzing alternative communication channels.

In order to test that those technologies would be able to work in this branch, we developed a methodology that uses them and that is shown in this paper. Our methodology is focused on detecting noise complaints that can be present in online communication channels such as OSN, but that could applied to every textual content with an opinion about noisy activities.

3. METHODOLOGY

The nature of OSN, which gives users a feeling of anonymity, makes people write their opinions on various topics without constraints, including useful information about noisy events in cities. For this reason, in this paper we propose a baseline methodology for acquiring and processing textual data from social networks, detecting present noise complaints, and classifying them according to the noise source to which they refer.

We had three different tasks to overcome when designing the method. The first one was to gather texts of an OSN, then detecting whether those texts were a were a noise complaint or not, and if they were, classifying them by sound source considering a standard classification. This process is shown on Figure 2

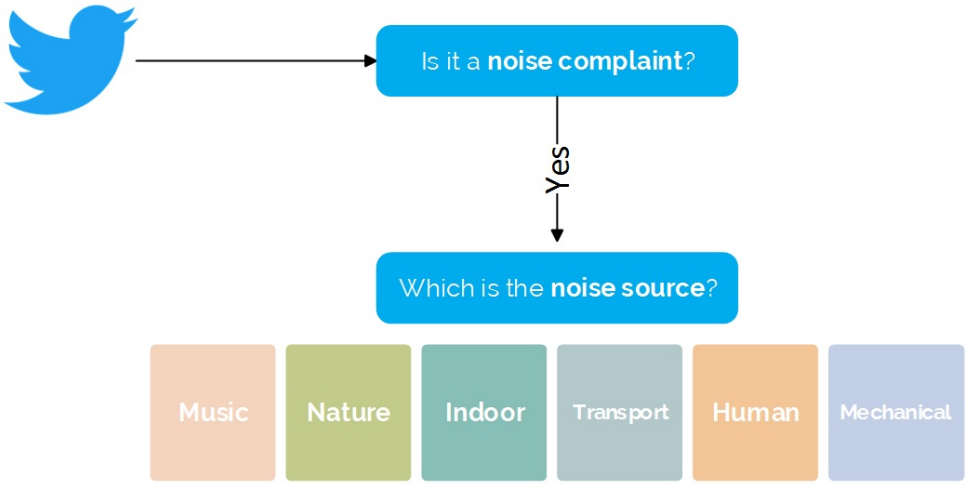

Figure 2: Goals of the proposed methodology.

A more complex schema of our methodology is shown in Figure 3. The first step is to acquire text data from the Internet. In data preparation, we normalize the text data correcting spelling mistakes and Internet slang, and we carry out a manual annotation assigning labels according to whether the text is a complaint or not. In the third step we train an algorithm to automatically identify complaints. In the last step, we build a system to arrange previously detected noise complaint by source, using a taxonomy we have built with that aim.

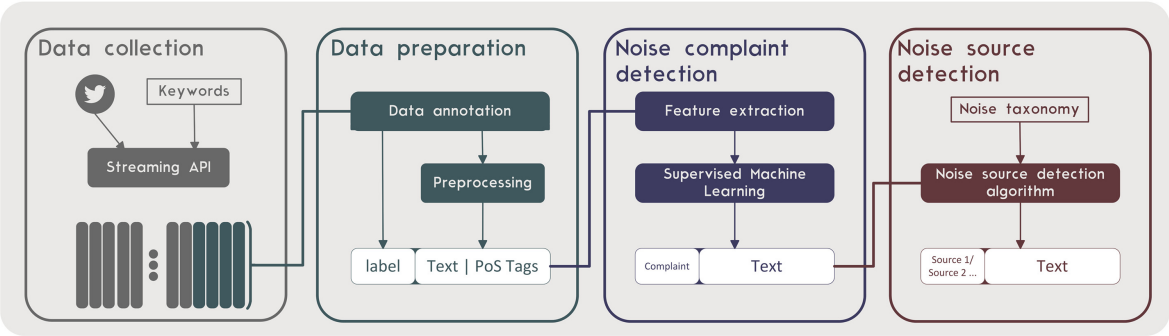

Figure 3: Methodology of OSN analysis for noise complaint detection and classification.

3.1. Data collection

Text data is our raw material and a proper data collection is essential for our study. We decided to focus on analyzing microblogging OSN because users usually post there short messages giving their opinion about almost any topic. More specifically, we have selected Twitter because of the number of active users,the extensive bibliography in the analysis of these type of texts, and the fact that people are straightforward in expressing their ideas because of the character limitation in each post [16, 17].

We have used Twitter's Streaming Application Programming Interface (API) to collect and save in a database the tweets where the word "noise" appeared. The data collection was performed between the 1st June and 1st September 2017. As a result, we got a collection of more than 5.6 million tweets, both original tweets, response to tweets, and retweets. For the analysis we established the requirement to build our model based only on original tweets detected in English by Twitter's API. We decided to remove retweets because they are used to support other people's opinions. We also removed tweet responses because when a person interacts with the content of other users, the context on the topic can be lost and the performance of the model could decrease. In addition, we eliminated some tweets that used keywords that were not related to environmental acoustics such as advertisements about noise reduction headphones or popular proverbs such as "Work hard in silence, let success make the noise".

After applying those constrains, we got the database shown in Table 1, which was the dataset we used during the rest of the experiment.

Database of tweets	
No. of tweets	843,300
No.of different hashtags	90,707
Statistics per tweet	
Average no. of mentions	0.1
Average no. of hashtags	0.4
Average no. of URLs	0.4

Table 1: Statistics of tweet database after applying filters

3.2. Data preparation

Because the acquired texts come from the Internet, they have many misspellings, slang, and acronyms commonly used in OSN. We applied a pre-processing pipeline represented by Figure 4 and based on the Sarker's proposal [18].This pre-processing pipeline is divided into three main components: A tokenizer, in which texts are split into smaller elements known as tokens using the Carnegie Mellon tokenizer [19]; a normalizer, which first transforms words to lowercase, then normalizes American terms to their British English form, removes repeated characters, and then goes through a process of slang corrector, based on online slang databases [20–22], and finally replacing the English contraction to its original form as this improves sentiment detection performance [23]. Finally, the Part Of Speech (PoS) is extracted from each token with the intention of knowing its word category, since it is a useful feature for generating the model.

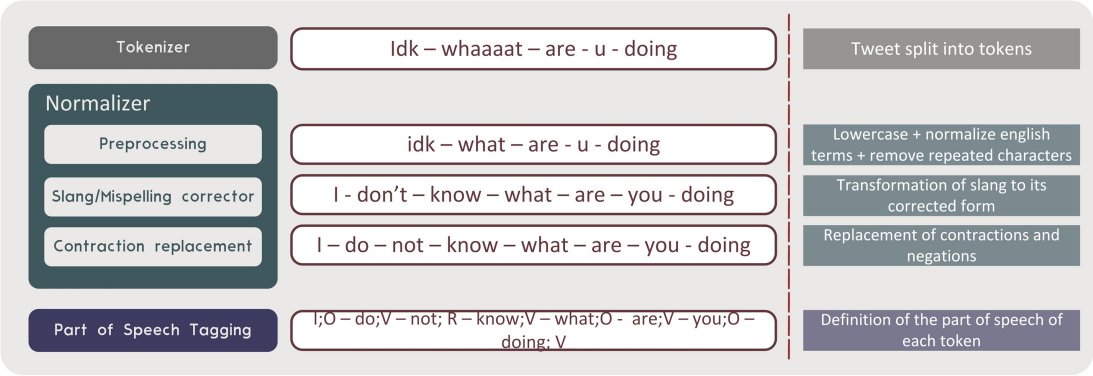

Figure 4: Pre-processing pipeline to be used with tweets

Additionally, because we implemented a Supervised Machine Learning algorithm, we had to manually classify some of the tweets to train our system. We used a previous version of the Noytext platform [24] to annotate more than 10,000 tweets, differentiating between two classes: class 1 tweets, texts containing a noise complaint; and class 0 tweets, other texts. We got a total of 9427 class 0 tweets, and 580 class 1 tweets.

3.3. Noise complaint detection

The third step of our methodology is the detection of tweets in which people are complaining about noise. We trained a Maximum Entropy Classifier, which is a Supervised Machine Learning algorithm based on class probabilities. We used a diverse set of features extracted from our texts, and that usually works well to detect negative sentiments about topics such as N-grams, which are a sequence of n-tokens from a text; sentiment features, which are sentiment scores given to texts based on its lexical content ; embeddings, which are n-dimensional representation of words where better analysis can be obtained; and PoS features, which are the number and percentage of PoS for each of the tweets.

To evaluate our model, and because our dataset had imbalanced classes, we decided to divide it into 8 splits using a stratified 8-fold cross validation algorithm. Thus, we used 7 folds for training and the last one for testing purposes in each iteration. We used stratification to maintain the original percentage of tweets of each class in each split, but we applied a weight to decrease the effect of class imbalance in the model performance.

Figure 5 presents the performance of the classifier in terms of Receiver Operating Characteristic curve (ROC curve). The x-axis represents the False Positive Ratio (FPR) and the y-axis the True Positive Ratio (TPR). At the optimum point of the curve a TPR of 0.85 is obtained, meaning that 85% of noise complaints are correctly identified, and a FPR of 0.16, indicating that 16% of tweets that are not noise complaints are categorized as such.

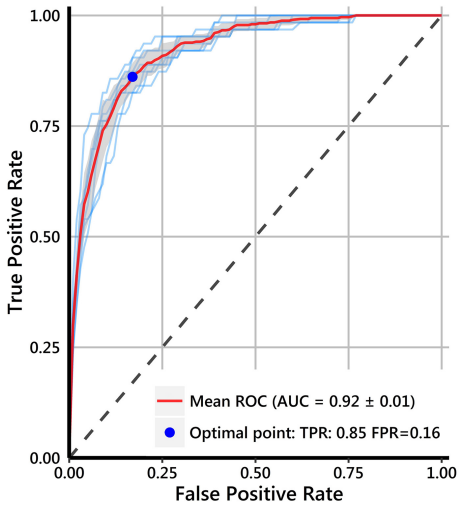

Figure 5: ROC Curve of noise complaint classifier.

3.4. Noise source detection

Finally, we wanted to classify those complaints based on their sound origin. To achieve that goal, we used an approach that combines the use of lexicons and taxonomies. We built a taxonomy based on the one used in Chatty maps project, that was created only using Social Media data and offers similar categories to the ones given by Schafer [25]. Each category was accompanied by a list of words extracted from Freesounds platform, in this way, a 228-word lexicon is arranged into 6-noise sources: transport, nature, human, music, indoor, and mechanical [10]. Since twitter vocabulary sparsity is higher than the Freesounds' one, we expanded that lexicon looking for synonyms and similar words in WordNet 3.0., that is a lexical database for the English language that arranges words by hierarchies based on their relationship [26].

We used that lexicon and the taxonomy to calculate the noise source mentioned in each tweet by looking each tweet token in the lexicon. If the token was found in the lexicon, we added that category to the tweet. After that process, we obtained a list of likely categories for each tweet.

Since a tweet can have several sound categories, our noise source detector could be considered as a multi label classifier. In the first place, we annotated a noise complaint dataset comprised of 510 tweets, assigning each tweet one or more sound categories. Then, we used the Hamming Loss score (HL), which represents the fraction of misclassified labels, and the Subset 0/1 Loss function, a more restrictive metric that only considers a correct classification when all the labels of a sample match with the correct ones.

We obtained a value Subset 0/1 loss score of 0.58, which means that 42% of tweets had all their labels well predicted. That metric was highly restrictive, since if a label is not

correctly predicted, it considers a wrong classification even if the the other labels could provide useful information. For that reason, we calculated the HL score, in which we got a value of 0.14, which means that 86% of the dataset labels were correctly predicted.

4. EXAMPLE OF USE - NOISE EVENT ALARM SYSTEM

We developed a model with acceptable performance in detecting complaints and classifying them by noise source but, what kind of analysis could we carry out to apply the model in a useful way? We have implemented an example that uses the power of our methodology, together with statistical techniques, to detect events that could raise noise concerns and complaints in the population.

After applying our model to the dataset shown in Table 1, we got more than 32,000 tweets classified as noise complaints. And almost 19,000 of them were identified with at least one noise source. Then, we computed the number of complaints per day for each sound source and applied an AutoRegressive Integrated Moving Average (ARIMA) model to detect days with anomalous number of complaints about noise. The result is shown in Figure 6.A, where we can see that an anomaly in human, mechanical, and nature noise complaints appeared during the U.S. Independence Day festivities. With our model and ARIMA models, it is easy to infer that the noise problem is related to an increase in annoyance due to those noise sources, since the number of complaints on those days increased significantly.

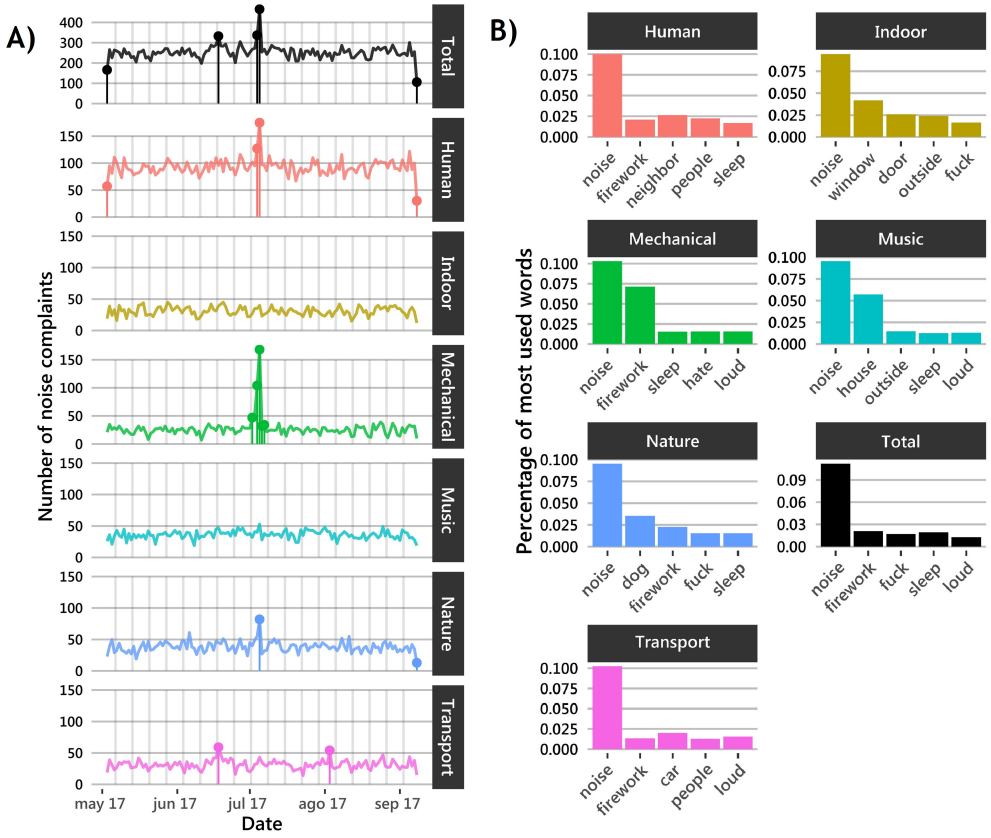

Figure 6: A- Time series of the number of noise complaints detected each day by noise source. B- Top-6 most used words in complaints written between the 3rd and the 5th of July on each noise source.

But how could we know what the real origin of those complaints is? We have taken the complaints of those anomalous days and plotted the histogram with the most commonly used words for each category. The Figure 6.B shows that the word *firework* is present as one of the most used words in many categories, which means that the noisy event was probably produced by fireworks. If we take a look at the human category, *firework* is often used near the word *neighbour*. Therefore, people complain about neighbours who use fireworks.

We have also implemented a system to trigger alerts related to fireworks problems in the future through the use of Statistical Process Control theory [27]. We have analyzed the 5 most used words in the category "total". We have defined our statistical control limits by computing the average use of those words plus three standard deviations. This limit, which covers 99.7% of normal days, is represented by a dashed line in the Figure 7. If a day is over that limit, we might conclude that a noise event related to fireworks is occurring.

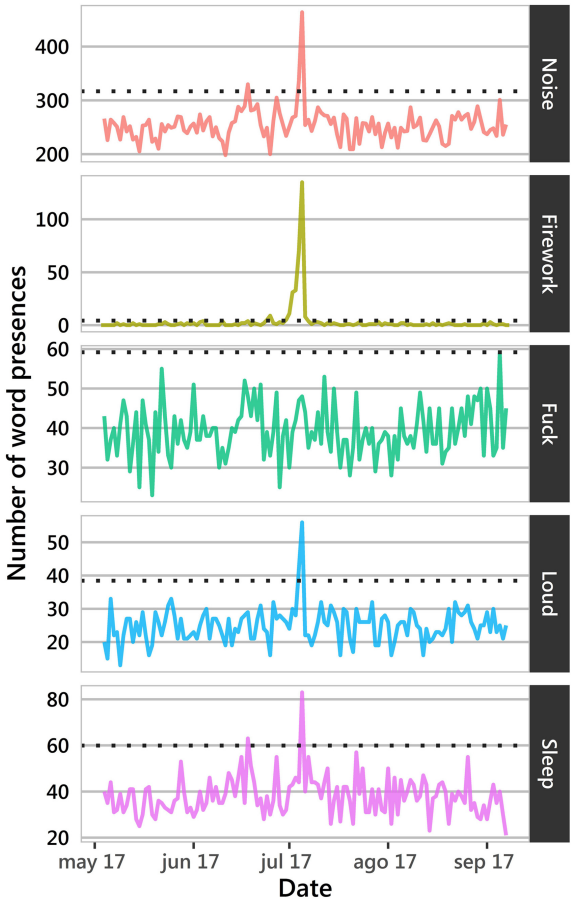

Figure 7: Evolution of daily tweets that contain the most used words in the category "total".

When we plot the evolution of tweets using these words, we can see that the limit was exceeded on the 4th of July for the words *noise*, *firework*, *loud*, and *sleep*. We could conclude that analyzing the number of appearances of those words could be used like an alarm system that warns of an annoying sound event occurring related to this type of celebration in the future

5. CONCLUSION

This paper has described a methodology for identifying and classifying complaints about noise posted by OSN users. The use of Artificial Intelligence technologies such as Machine Learning and NLP has been shown to be an appropriate way of detecting noise complaints. After testing the performance of our model, we have shown a case study in which, applying our methodology to the database, we were able to obtain temporal series with the number of complaints over time and organising them by noise source. After detecting anomalies in those temporal series, we were able to see what the most annoying noise sources were during the U.S. Independence Day event. Applying NLP techniques, we were able to more explicitly observe that the main annoyance noise source came from fireworks, and we proposed a noise annoyance alarm system for this kind of event based on Statistical Process Control, which could be extended to other noise events and could be implemented by city managers to measure the effectiveness of actions against noise. We have also found that the system could have been used to trigger a-priori warnings on the same day, if it worked on real-time basis.

This type of analysis, or an alarm-system, based on the technology and the methodology presented in this paper could be extrapolated to other noisy events such as monitoring if nightclubs respect their closing time, or the annoyance produced by bar terraces. With this approach, city managers could measure the success of their actions against noise in a convenient manner.

We have focused this paper on showing the possibilities of the use of NLP technologies together with machine learning models in the environmental acoustics branch, as they are already used commercially in other fields and could be successfully implemented in current urban management systems. An extended version going deeper into the technical aspects of the research has been published in the paper *Beyond sound level monitoring: Exploitation of social media to gather citizens subjective response to noise*, published by Science of The Total Environment journal [12].

6. REFERENCES

[1] Irena Bojanova. The digital revolution: What's on the horizon? *IT Professional*, 16 (1):8–12, 2014.

[2] EU Directive. Directive 2002/49/ec of the european parliament and the council of 25 june 2002 relating to the assessment and management of environmental noise. *Official Journal of the European Communities, L*, 189(18.07):2002, 2002.

[3] F. Accordino. The futurium - a foresight platform for evidence-based and participatory policymaking. *Philosophy and Technology*, 26(3):321–332, 2013. doi: https://doi.org/10.1007/s13347-013-0108-9.

[4] C. Asensio, G. De Arcas, J.M. López, I. Pavón, and L. Gascó. Awareness: A parallel approach against noise. In *Proceedings of the 22nd International Congress on Sound and Vibration (ICSV 22), Florence, Italy*, pages 12–16, 2015.

[5] Abeed Sarker, Karen O' Connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. Social media mining for toxicovigilance : Automatic monitoring of prescription medication abuse from twitter. *Drug Safety*, 39(3):231–240, 2016. doi: https://doi.org/10.1007/s40264-015-0379-4.

[6] Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O' Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202 – 212, 2015. doi: https://doi.org/10.1016/j.jbi.2015.02.004.

[7] W. Jiang, Y. Wang, M.-H. Tsou, and X. Fu. Using social media to detect outdoor air pollution and monitor air quality index (aqi): A geo-targeted spatiotemporal analysis framework with sina weibo (chinese twitter). *PLoS ONE*, 10(10), 2015. doi: https://doi.org/10.1371/journal.pone.0141185.

[8] Hsun-Ping Hsieh, Rui Yan, and Cheng-Te Li. Dissecting urban noises from heterogeneous geo-social media and sensor data. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1103–1106, New York, USA, 2015. ACM. doi: http://doi.acm.org/10.1145/2733373.2806292.

[9] Yu Zheng, Tong Liu, Yilun Wang, Yanmin Zhu, Yanchi Liu, and Eric Chang. Diagnosing new york city's noises with ubiquitous data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 715–725, New York, USA, 2014. ACM. ISBN 978-1-4503-2968-2. doi: http://doi.acm.org/10.1145/2632048.2632102.

[10] L.M. Aiello, R. Schifanella, D. Quercia, and F. Aletta. Chatty maps: Constructing sound maps of urban areas from social media data. *Royal Society Open Science*, 3 (3), 2016. doi: https://doi.org/10.1098/rsos.150690.

[11] L. Gasco, C. Asensio, and G. De Arcas. Towards the assessment of community response to noise through social media. In *INTER-NOISE 2017 - 46th International Congress and Exposition on Noise Control Engineering: Taming Noise and Moving Quiet*, 2017.

[12] L. Gasco, C. Clavel, C. Asensio, and G. de Arcas. Beyond sound level monitoring: Exploitation of social media to gather citizens subjective response to noise. *Science of The Total Environment*, 658:69 – 79, 2019. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2018.12.071.

[13] NETBASE. Netbase Arbys Onlinse Social Networks sentiment analysis case study. Technical report, 2018. URL https://www.netbase.com/wp-content/uploads/Casestudy_Arby_03_2018.pdf.

[14] NETBASE. Netbase Chevrolet customer insights case study. Technical report, 2016. URL https://www.netbase.com/wp-content/uploads/NetBase_CS_Chevrolet_2016.pdf.

[15] J. Fenn and M. Raskino. *Mastering the Hype Cycle: How to Choose the Right Innovation at the Right Time*. Gartner Series. Harvard Business Press, 2008. ISBN 9781422121108.

[16] Twitter. Selected twitter company metrics and financials. Technical report, 2016. URL http://files.shareholder.com/downloads/AMDA-2F526X/5887909887x0x961126/1C3B5760-08BC-4637-ABA1-A9423C80F1F4/Q317_Selected_Company_Metrics_and_Financials.pdf.

[17] David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation. *ACM Trans. Manage. Inf. Syst.*, 9(2), 2018. doi: http://doi.acm.org/10.1145/3185045.

[18] A. Sarker. A customizable pipeline for social media text normalization. *Social Network Analysis and Mining*, 7(1), 2017. doi: https://doi.org/10.1007/s13278-017-0464-z.

[19] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390. Association for Computational Linguistics, 2013. URL http://aclweb.org/anthology/N13-1039.

[20] NoSlang. Noslang - internet text slang dictionary translator, 2017. URL www.noslang.com.

[21] Fei Liu, Fuliang Weng, and Xiao Jiang. A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pages 1035–1044. Association for Computational Linguistics, 2012. URL http://dl.acm.org/citation.cfm?id=2390524.2390662.

[22] Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. Insertion, deletion, or substitution? normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 71–76. Association for Computational Linguistics, 2011. URL http://aclweb.org/anthology/P11-2013.

[23] Zhao Jianqiang. Pre-processing boosting twitter sentiment analysis? In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pages 748–753. IEEE, 2015.

[24] Luis Gasco. Noytext: A web-based platform for annotating short-text documents to be used in applied text-mining based research. February 2019. doi: https://doi.org/10.5281/zenodo.2566448. URL https://github.com/luisgasco/noytext.

[25] R Murray Schafer. *The soundscape: Our sonic environment and the tuning of the world*. Simon and Schuster, 1993.

[26] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. doi: http://doi.acm.org/10.1145/219717.219748.

[27] P. Winkel and N. Fan Zhang. Theory of statistical process control. *Statistical Development of Quality in Medicine*, pages 9–36, 2007. doi: https://doi.org/10.1002/9780470515884.ch1.