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

Noytext: A Web platform to annotate social media documents on noise perception for their use in opinion mining research.

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

Boost of online social networks has demonstrated that some people are willing to share their views about everyday problems, including noise. With the advent of Natural Language Processing and Machine Learning technologies to the majority of the scientific fields, we have begun to analyze the textual content of social media, and more specifically online social networks, to extract insights about the noise attitude of the population that uses this channel to express their opinion in this matter.

Some of the state-of-the-art algorithms, such as deep neural networks, are supervised machine learning algorithms. This means that researchers have to provide a set of labelled training data to build new models. The annotation process is known as one of the most time-costly tasks in a data science pipeline, since researchers among other thigs have to test the agreement between annotators and to measure the quality of the categories they had previously defined. For that reason in this paper, we introduce Noytext which is a customizable web tool to annotate texts from your database, that can be deployed in your own webserver and you can use to request help from colleagues and collaborators in the annotation process in a friendly way.

Keywords: Text mining, Community engagement, Noise annoyance
I-INCE Classification of Subject Number: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 advent of the Digital Revolution has radically changed the way we communicate and use technology. Today 4200 million people have an Internet connection, there are more than 3000 million users who actively use social media, and it is estimated that each user has an average of 5.5 accounts on these social platforms [1]. The importance of these changes has even led to the "human being" being named person of the year by Times magazine in year 2006 [2]. This shows the possibilities opened up by the Internet for users' opinions to be heard and for them to have decision-making power as a whole on issues such as politics, commercial products and the environment, among others.

In parallel, innovative techniques have reached all branches of science, including environmental acoustics. On the one hand, the lower cost of instrumentation has made it possible to implement affordable monitoring networks, both in combination with traditional monitoring devices [3] and sensor networks composed entirely of low-cost equipment [4]. On the other hand, the ease with which citizens can be involved in projects has led to the creation of many crowdsourcing platforms that allow population to collaborate with research in noise monitoring and acoustic evaluation tasks, some of those projects were focused on measuring noise [5], evaluating it [6], or combining both approaches [7]. Additionally, the proof of good performance of Artificial Intelligence systems has led to the development of programming libraries that have democratized the use of Machine Learning techniques that are being applied in a multitude of analyses which were unthinkable a few years ago, such as the automatic identification of sound sources [8].

Despite all this progress, there is an entry barrier for the application of these technologies. It is a fact that we work in an interdisciplinary environment where acoustic engineers, urban planners, environmentalists, policy makers, and computer scientists work all together, but it is also true that many teams do not have easy-to-use tools to start working with novel data-analysis approaches. For that reason, in this paper we present Noytext, an open source web application that is relatively easy to install, configure and personalize with the aim of simplifying one of the most time-consuming tasks in supervised machine learning projects: the data annotation.

2. THE PROBLEM

Currently, people that are active on Online Social Networks (OSN) observe, analyze, create and disseminate information about their feelings and opinions by writing texts, and uploading photographs and videos. In environmental acoustics, OSN data was used for the first time to diagnose New York noise sources by doing the analysis together with open data [9, 10].Yahoo researchers carried out the Chatty Maps project, where they analyzed tags from pictures and they were able to extract the primary noise source in big cities around the world at street level just using social media content [11]. More recently, we developed a Machine Learning model to detect and classify noise complaints written on Twitter [12]. During this project, we had to deal with the lack of annotated datasets and the difficulty to find environmental acoustic professionals to help us with that task. It is known that data annotation is one of the most time-consuming tasks in the data analysis pipeline; hence, we developed a web application to ask for assistance from other professionals who wanted to collaborate by annotating some text documents. After obtaining successful results in the project, in which we demonstrated that it was possible to detect the noise

complaints that people share on OSN, we decided to improve the tool and make it open source so that other researchers who wanted to apply a similar methodology could do their research more efficiently.

3. NOYTEXT

Noytext was created with the aim to help environmental acoustic researchers to annotate text documents for acoustic perception opinion mining [13]. It offers several features to be suitable to the needs of every research project in this field, including the possibility of sharing the application as a web page to request annotations from the community.

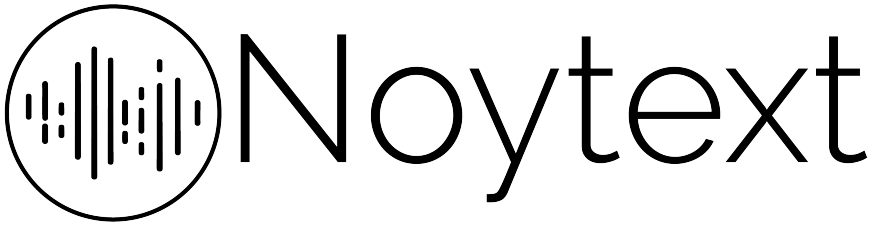

Figure 1: Noytext logotype

This web application offers some features that can allow researchers to obtain better training data, such as:

- *Simplifying the annotation task*: Since the annotating process is interactive, the user can tag a greater number of texts than with other plain text methods. In addition, since annotation is carried out interactively on a web server, rules can be set to maximize the performance of this task by displaying new texts when they reach a specific number of annotations.

- *Easy to install in your own server*: Noytext is a Shiny application that can be easily installed and used in a server with a Shiny server installed or a personal computer with an R installation. If you decide to install it on an empty server on the cloud, you could have your app working in 20 minutes following the steps given on the repository web-page

- *Standard back-end and front-end*: The app is written on R programming language, using the Shiny library. This library produces standard HTML, CSS, and JavaScript code, the standard front-end technologies on web development. The server side, also known as back-end, works using R language, a standard programming language in data science and research that will allow researchers to modify code according to their needs if required. On the other hand, MongoDB is the database system chosen for the app because of its easy adaptability to web-apps causality.

- *Cross platform*: In order to facilitate the annotation tasks to users, we have developed an application whose User Interface (UI) adapts to the device from which the annotation is carried out.

3.1. Structure and customizing options

Noytext is based on a 4-page schema that you can fully personalize based on your needs:

– *Project information page*: You can use your own HTML page to provide information about your project goals, or modify the example HTML file we provide you to better match the app appearance. This page can be disabled if you do not need it.

– *Help page*: This page gives the user a step-by-step guide on how to use the UI for data annotation. You can change the hints to your own language or needs, and disable the page as well.

– *Annotation page*: This is the main page of the app. It is composed by three main elements. The text box, where it will appear the texts store in your database; the radio button selectors, where the annotations categories are shown; and the buttons, which are used to save the selected option and to show another text. In order to avoid fake annotators, we have included a JavaScript code to enable and disable the save and next buttons until a category is selected, making it difficult for potential fake annotators to boycott the database labeling.

– *About page*: This is another HTML page that you can use to present information about your team, institution or whatever you consider. In the same way as in the information page, you can use your own HTML file, modify the example or disable it.

An example of the application's UI can be seen on Figure 2. Both the page titles and the project name can be defined by the researcher, being the Noytext logo the only element that is fixed by default. At the time of writing this manuscript, Noytext does not

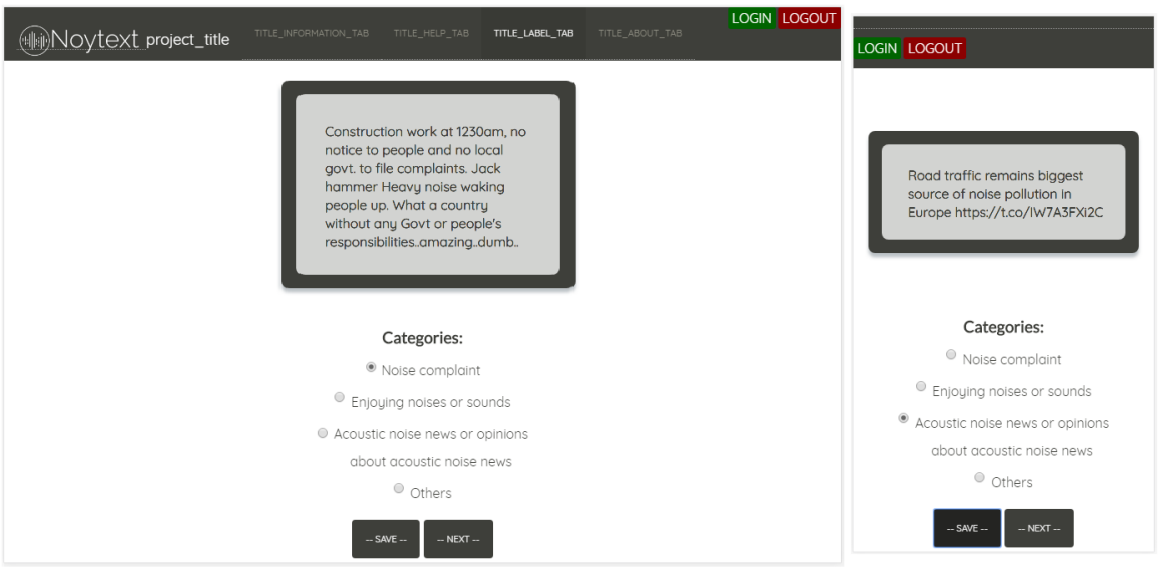

Figure 2: Graphic interface of the Noytext platform. On the left, the UI adapted for a large screen can be seen. On the right, the UI is shown adapting to small screen such as smartphones.

allow changing the categories of text classification, but it will be a customizable element in future versions of the app. It can also be seen that the UI elements adapt to the width of the device, hence it can be used both in standard personal computers as well as in tablets and smartphones. On the other hand, the figure also shows an example of the type of texts that can be annotated on the platform: The screenshot on the left shows a typical noise complaint that can be found on a social network like Twitter and that must be tagged before training a complaint detection model; the one on the right shows the typical tweet that mentions the word noise, but is just a post in which the user has shared a piece of news about noise instead of a complaint.

3.1.1. Decide your own application functionality

The application offers several options to personalize its functionality:

- A database running in a remote MongoDB server can be configured to be used with the application. This gives versatility and the option of using Noytext without changing your computer systems configuration.
- The number of times the same text can be annotated by different users can be defined by the researcher. This allows to get inter-annotator agreement statistics, used to measure consensus among them, and to check that annotation tasks are working correctly.
- When the project requires it, information about each annotator could be obtained. If this option is enabled, a login and registration button will appear. Then, the first time a user enters to Noytext, he will be asked to register and answer a small questionnaire defined by the researcher. This questionnaire has several types of survey questions, both free-text and numeric inputs, sliders, radio buttons, or multiple choice questions, which have been implemented in order to provide flexibility for this purpose. The variety of inputs can be seen in Figure 3.

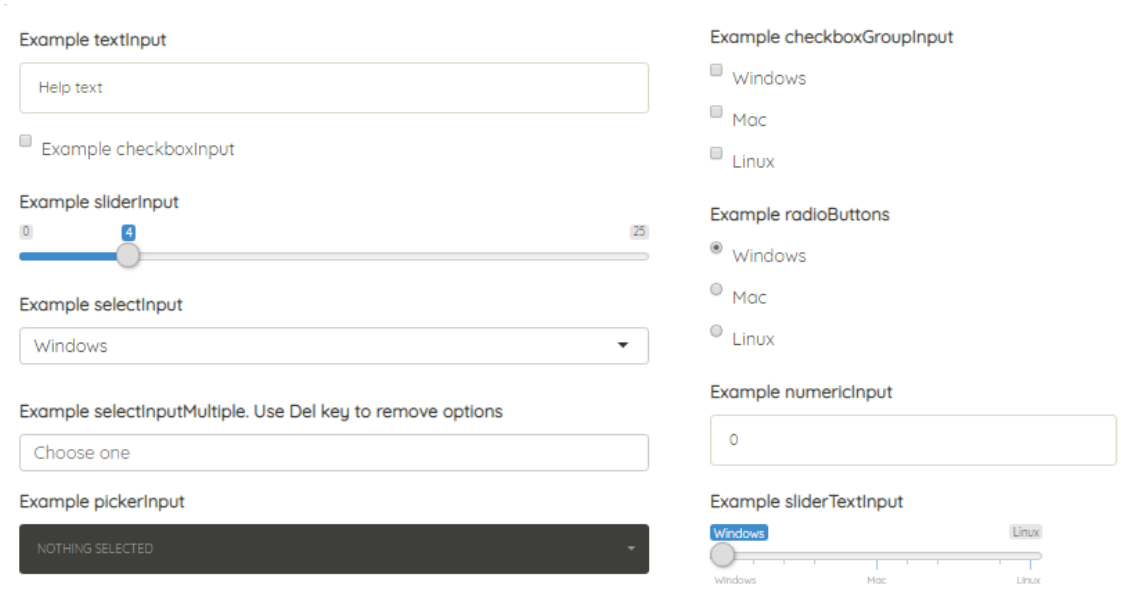

Figure 3: Questions types that can be used on Noytext questionnaires

4. EXPERIMENT

As we have previously mentioned, one of the problems we experienced during our project to detect and classify noise complaints from OSN was the lack of annotated datasets. In fact, this was the most time-consuming task of the project, standing for approximately 30% of the whole duration.

For that reason, and with the aim of testing the application functionality, we launched the experiment *Noise tweet lab*. In this experiment we seek collaboration from both environmental acoustics professionals and the general public to build a reliable labelled database that can be used by other researchers in the field. The creation of this database will allow researchers to focus on improving noise complaint detection algorithms and on the efficiency of the research projects of this nature by minimizing the data annotation process. If you want to know more information about this project, you can scan Figure 4 code or access to http://noisetweetlab.noytext.com

Figure 4: Access the website of the Noise Tweet Lab experiment built using Noytext app

5. CONCLUSIONS

As it has been introduced in this manuscript, text annotation is one of the problems that researchers who decide to apply machine learning techniques to detect textual noise complaints may face. For this reason, Noytext has been developed and presented. This open-source application allow you to obtain better training data by simplifying the annotation task to your annotators. It has shown several customization options, as well as the *Noise Tweet Lab* experiment, conceived to create the first public annotated dataset of noise complaints gathered from Social Media, that will be useful for future research projects.

6. REFERENCES

[1] Kit Smith. 122 amazing social media statistics and facts, Jan 2019. URL https://www.brandwatch.com/blog/amazing-social-media-statistics-and-facts/.

[2] A. Sheth. Citizen sensing, social signals, and enriching human experience. *IEEE Internet Computing*, 13(4):87–92, July 2009. doi: https://doi.org/10.1109/MIC.2009.77.

[3] Cesar Asensio, Luis Gasco, Guillermo De Arcas, Juan Manuel López, and Jesus Alonso. Assessment of residents' exposure to leisure noise in málaga (spain). *Environments*, 5(12), 2018. doi: https://doi.org/10.3390/environments5120134.

[4] Juan P. Bello, Claudio Silva, Oded Nov, R. Luke Dubois, Anish Arora, Justin Salamon, Charles Mydlarz, and Harish Doraiswamy. Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution. *Communications of the ACM*, 62(2):68–77, 2019. doi: http://doi.acm.org/10.1145/3224204.

[5] Ellie D'Hondt, Matthias Stevens, and An Jacobs. Participatory noise mapping works! an evaluation of participatory sensing as an alternative to standard techniques for environmental monitoring. *Pervasive and Mobile Computing*, 9(5):681 – 694, 2013. ISSN 1574-1192. doi: https://doi.org/10.1016/j.pmcj.2012.09.002. Special issue on Pervasive Urban Applications.

[6] Charlie Mydlarz, Ian Drumm, and Trevor Cox. Application of novel techniques for the investigation of human relationships with soundscapes. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 2011, pages 738–744. Institute of Noise Control Engineering, 2011.

[7] Antonella Radicchi, Dietrich Henckel, and Martin Memmel. Citizens as smart, active sensors for a quiet and just city. the case of the "open source soundscapes" approach to identify, assess and plan "everyday quiet areas" in cities. *Noise Mapping*, 4, 03 2018. doi: https://doi.org/10.1515/noise-2017-0008.

[8] J. Salamon and J. P. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, March 2017. doi: https://doi.org/10.1109/LSP.2017.2657381.

[9] 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.

[10] 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.

[11] 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.

[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] 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.

[14] 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.

[15] 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.