Monitoring Social Media to Identify Environmental Crimes through NLP A Preliminary Study

Raffaele Manna, Antonio Pascucci, Wanda Punzi Zarino, Vincenzo Simoniello, Johanna Monti
UNIOR NLP Research Group
University "L'Orientale"
Naples, Italy
[rmanna, apascucci, jmonti]@unior.it
[w.zarino, vincenzosimoniello]@gmail.com

Abstract

This paper presents the results of research carried out on the *UNIOR Eye* corpus, a corpus which has been built by downloading tweets related to environmental crimes. The corpus is made up of 228,412 tweets organized into four different subsections, each one concerning a specific environmental crime. For the current study we focused on the subsection of waste crimes, composed of 86,206 tweets which were tagged according to the two labels *alert* and *no alert*. The aim is to build a model able to detect which class a tweet belongs to.

1 Introduction

In the current era, social media represent the most common means of communication, especially thanks to the speed with which a post can go viral and reach in no time every corner of the globe. The speed with which information is produced creates an abundance of (linguistic) data, which can be monitored and handled with the use of hashtags (#). Hashtags are user-generated labels, which allow other users to track posts with a specific theme on Twitter. Moreover, social media such as Twitter can be powerful tools for identifying a variety of information sources related to people's actions, decisions and opinions before, during and after broad scope events, such as environmental disasters like earthquakes, typhoons, volcanic eruptions, floods, droughts, forest fires, landslides (Imran et al., 2015; Maldonado et al., 2016; Corvey et al., 2010). In light of the above, our aim is to monitor social media in order to detect environmental crimes.

Our research is guided by the following question: can Natural Language Processing (NLP) represent a valuable ally to identify these kinds of crimes through the monitoring of social media? For this purpose, we compiled a corpus of tweets starting from a list of 41 terms related to environmental crimes, e.g. *combustione illecita* (*illicit combustion*), *rifiuti radioattivi* (*radioactive waste*), *discarica abusiva* (*illegal dumping*), and we used the Twitter API to download all the tweets (specifically 228,412) related to these terms introduced by hashtag. In this research, a special focus is dedicated to the tweets related to *La terra dei fuochi* (literally *the Land of Fires*) (Peluso, 2015), a large area located between Naples and Caserta (in the South of Italy) victim of illegal toxic wastes dumped by organized crime for about fifty years and routinely burned to make space for new toxic wastes.

In order to achieve our purpose, we trained different machine learning algorithms to classify report emergency text and user-generated reports. The paper is organized as follows: in Section 2 we discuss Related Work, in Section 3 we present the *UNIOR Earth your Estate (UNIOR Eye)* corpus. The case study is described in Section 4 and Results are discussed in Section 5. Conclusions are in Section 6 along with directions for Future Work.

2 Related Work

As previously mentioned, hashtags are one of the most important resources - if not the most important - in text data such as those of Twitter. The possibility to aggregate data according to their content allows users to monitor all the discussion about a specific subject in real-time (an emblematic case is the hashtag *#Covid_19*).

Concerning the topic of our research, namely environmental issues, the most representative and

productive hashtags have proved to be *#terradeifuochi* and *#rifiuti* (respectively with a frequency of 92,322 and 62,750 occurrences), that directly refer to circumstances that have a strong impact on the environment and people's health. The use of hashtags proved to be useful in monitoring natural disasters, such as earthquakes, flood and hurricane.

For a survey on information processing and management of social media contents to study natural disasters, see (Imran et al., 2016). (Neubig et al., 2011) focused on the 2011 East Japan earthquake. The scholars built a system able to extract the status of people involved in the disaster (e.g. if they declared to be alive, they request for help, their information requests, information about missing people). About one hundred scholars participated spontaneously in the project *ANPI_NLP* (ANPI means Safety in Japanese) and the results show convincing performances by the classifier they built. (Maldonado et al., 2016) investigated natural disasters in Ecuador, monitoring Twitter to filter contents according to four different categories: volcanic, telluric, fires and climatological. The filtering process is based on keywords related to the four categories. The scholars released a web application that graphically shows the database evolution. The efficiency of the tweet filtering algorithm that they developed is expressed in terms of precision (%93.55). (Tarasconi et al., 2017) investigated tweets related to eight different event types (floods, wildfires, storms, extreme weather conditions, earthquakes, landslides, drought and snow) in Italian, English and Spanish. The corpus is composed of 9,695 tweets and can be extremely useful to perform information extraction in the aforementioned three languages. (Sit et al., 2019) used the Hurricane Irma, which devastated Caribbean Islands and Florida in September 2017, as a case-study: the scholars demonstrate that by monitoring tweets it is possible to detect potential areas with high density of affected individuals and infrastructure damage throughout the temporal progression of the disaster. By focusing on tweets generated before, during, and after Hurricane Sandy, a superstorm which severely impacted New York in 2012, (Stowe et al., 2016) proposed an annotation schema to identify relevant Twitter data (within a corpus of 22.2M unique tweets from 8M unique Twitter users), categorizing these tweets into fine-grained categories, such as *preparation* and *evacuation*.(Imran et al., 2016) presented Twitter corpora composed of over 52 million crisis-related tweets, collected during 19 different crises that took place from 2013 to 2015. These corpora were manually-annotated by volunteers and crowd-sourced workers providing two types of annotations, the first one related to a set of categories, the second one concerning out-of-vocabulary words (e.g. slangs, places names, abbreviations, misspellings). The scholars then built machine-learning classifiers in order to demonstrate the effectiveness of the annotated datasets, also publishing *word2vec* word embeddings trained on more than 52 million messages. The preliminary results of this study posit that a classification with a high precision of tweets relevant to the disaster is possible to assist crisis managers and first responders. Our study is not devoted to monitor natural disasters but to monitor natural human-caused disasters. More specifically, the aim is to exploit NLP techniques to contribute to the identification of intentional environmental crimes through social media analysis. To the best of our knowledge, this perspective of investigation is rather novel in the field.

3 The *UNIOR Eye* Corpus

This section outlines the way the *UNIOR Eye* corpus was created and how it is internally structured. The research has been carried out in the framework of the C4E - Crowd for the Environment (Progetto PON Ricerca e Innovazione 2014-2020) project[2].

The *UNIOR Eye* corpus is made up of 228,412 tweets related to environmental crimes downloaded through Twitter API, covering the period from 01 January 2013 to 06 August 2020. The compilation phase of the corpus was divided into two steps: the creation of a vocabulary containing keywords related to environmental crimes and the creation of the corpus. During this work phase, the data was structured and organized according to the different keywords, obtained from glossaries and documents specific to the topic.

Precisely, the following resources

- *Glossario di termini sull'ambiente* (FIMP, 2017) (a guide from A to Z concerning the complex issue of environmental pollution);

[2]http://www.unior.it/ateneo/20574/1/c4e-crowd-for-the-environment-progetto-pon-ricerca-e-innovazione-2014-2020.html

- *Glossario dinamico per l'Ambiente ed il Paesaggio* (ISPRA, 2012) (a glossary supplied by the Italian Institute for Environmental Protection and Research);
- *Glossario ambientale*[3] (a glossary supplied by the national agency for the environmental protection of Tuscany);
- *BeSafeNet*[4] (a glossary based on the *Glossary on Emergency Management*, which has been developed in 2001 by European Centre of Technological Safety (TESEC) of Euro-Mediterranean network of Centres EUR-OPA Major Hazard Agreement of Council of Europe in collaboration with other centres of network);
- *HERAmbiente*[5] (a glossary provided by Herambiente, the largest company in the waste management sector);
- *Enciclopediambiente*[6] (the first freely available online Encyclopedia on the Environment, designed by a group of four engineers with the aim of spreading "environmental knowledge")

and the following two web sources

- a dossier containing important provisions aimed at dealing with environmental and industrial emergencies and encouraging the development of the affected areas[7];
- a document on environmental crimes and environmental protection[8].

were consulted. All of these language resources contain information and definitions of the basic terms related to environmental disasters and crimes, e.g. *Rifiuti pericolosi* (hazardous waste): waste products which can generate potential/substantial risk to human health/the environment if handled improperly. Hazardous waste contains at least one of these characteristics: flammability, corrosivity, or toxicity,[9] and is included in special lists. Here are some examples.

- *HASHTAG HASHTAG Fiumicino: eternit e rifiuti pericolosi al Passo della Sentinella URL HASHTAG* (HASHTAG HASHTAG Fiumicino: eternit and hazardous waste in Passo della Sentinella URL HASHTAG);
- *Cani in gabbia in discarica abusiva: Due animali tra rifiuti pericolosi, amianto e bombole gas URL* (Caged dogs in illegal dump: two pets among hazardous waste, asbestos and gas cylinders URL)

After this phase it was possible to create the corpus by downloading from Twitter all the tweets containing these keywords preceded by the hashtag. These hashtags helped us to gather the information needed to detect crimes against the environment. More specifically, the corpus is internally divided into semantic areas, each one concerning a specific environmental crime: *rifiuti e terra dei fuochi* (waste and Terra dei fuochi); *reati contro le acque* (water-related crimes); *materiali e sostanze pericolose* (hazardous substances and materials); *incendi e roghi ambientali* (environmental fires). These sets are further divided into more specific subsets, e.g. the folder *reati contro le acque* (water-related crimes) contains the subsets *acque di scarico, acque reflue, fiumi inquinati, liquami* (sewage, wastewater, polluted rivers, slurry). The resulting corpus contains, therefore, a total of 228,412 tweets, 22,780,746 tokens, 569,905 types with a type/token ratio (TTR) of 0.025.

4 Case Study

This section describes the steps taken to perform the preliminary experiments on a selected part of the *UNIOR Eye* corpus. First, the dataset on which the experiments and data preparation operations were carried out is presented, then the preprocessing steps are listed and, finally, the different machine learning approach used are described.

4.1 Dataset

As described in Section 3, the *UNIOR Eye* corpus is divided into four semantic areas related to the most common crimes against the environment. Among the four semantic areas, we decided to use the waste crimes subsection to test a specific use case: whether an NLP system can understand and report emergency text and user-generated reports. Therefore, for the experiments described in this paper, we focus our investigation on a sub-section of the *UNIOR Eye* corpus, namely tweets about waste related crimes and tweets with the hashtag #terradeifuochi contained in the corresponding semantic area: *waste and Terra dei fuochi*. This subsection of the corpus contains 86,206 tweets. First, for the total number of tweets, hashtags, mentions and URLs are replaced with placeholder words. Then tweets were annotated by the paper authors on the basis of two labels: i) alert and ii) no alert,

[3] http://www.arpat.toscana.it/glossario-ambientale
[4] http://www.besafenet.net/it-it/glossary
[5] http://ha.gruppohera.it/glossario_ambiente/
[6] http://www.enciclopediambiente.com
[7] https://www.senato.it/japp/bgt/showdoc/17/DOSSIER/0/740667/index.html?part=dossier_dossier1-sezione_sezione12-h2_h28
[8] https://scuola21.fermi.mn.it/documenti/reati_ambientali.pdf
[9] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:C:2018:124:FULL&from=IT

i.e. if the tweet contains or not a message aimed at reporting and locating a waste related crime.
Below, we provide a sample of annotated tweets following our two labels, **alert - no alert**:

- *Ore 11:40 autostrada A1 altezza Afragola Acerra direzione Roma. Roghi Tossici indisturbati, la HASHTAG... URL HASHTAG HASHTAG* (11:40 am A1 motorway near Afragola Acerra towards Rome. Undisturbed toxic fires, the HASHTAG ... URL HASHTAG HASHTAG) — **ALERT**

- *MENTION ministro, piuttosto che pensare alla HASHTAG pensi ai continui roghi MENTION* (MENTION Minister, rather than thinking about the HASHTAG think about the continuous fires MENTION) — **NO ALERT**

During the annotation phase, we noted that the **no alert** class is the one which contains the majority of tweets and includes examples of hate speech, satirical texts, news about emergency actions as well as politically oriented texts. Consequently, our dataset built in this way is unbalanced for the two classes, counting 81,235 tweets for the **no alert** class and 4,970 **alert** tweets. In order to visualize alert tweets, we exploit Carto[10], a cloud computing platform that provides a geographic information system, web mapping, and spatial data science tools[11].

4.2 Inter-annotator Agreement

When different annotators label a corpus, it is important to calculate the inter-annotator agreement (IAA) with a twofold objective: i) make sure that annotators agree and ii) test the clarity of guidelines. As previously mentioned, the dataset (composed of 86,206 tweets) has been annotated by four of the paper authors on the basis of two labels: i) alert and ii) no alert. This implies that each author annotated about 21,000 tweets. Then, to calculate inter-annotator agreement we randomly selected 10% of the tweets (i.e. 8,620) which were tagged by all annotators.
The agreement among the four annotators is measured using Krippendorff's α coefficient; instead, to estimate the agreement between pairs of annotators, we use Cohen's κ coefficient (Artstein and Poesio, 2008). Taking into account the recommendations set out in (Artstein and Poesio, 2008; Krippendorff, 2004), we interpret the κ values obtained in IAA according to the strength of agreement criteria described in (Landis and Koch, 1977) for each pair of annotators; whereas, for agreement among four annotators, we follow the suggested standard in (Krippendorff, 2004). The calculated value of Krippendorff's α is ***0.706***. Considering the standard value in (Krippendorff, 2004), our value of α=*0.706* is considered as acceptable and expressing a good data reliability. In Table 1 we show the results for pairs of annotators.

Pair of annotators	**Value of κ**
a1 - a2	0.691
a1 - a3	0.742
a1 - a4	0.841
a2 - a3	0.676
a2 - a4	0.644
a3 - a4	0.641

Table 1: Cohen's κ values for pairs of annotators.

According to (Landis and Koch, 1977), five out of six Cohen's κ values show a "substantial" strength of agreement for each pair; while a pair (a1-a4) show a κ value considered "almost perfect" in the research cited.

4.3 Preprocessing

Before feeding the machine learning algorithms, some pre-processing steps are performed. Since the majority of mentions and hashtags are shared by both alert and no-alert samples, we focus on the tweet itself, by removing any reference to people, entities and organizations conveyed through hashtags and mentions. Therefore, the placeholder words related to hashtags, URLs and mentions are removed. Then, punctuation is removed from the tweets along with a custom list of function words such as determiners, prepositions and conjunctions. Finally, the tweets are lower-cased and the tokenization is performed.

4.4 Machine Learning Approaches

We set the problem of tweets related to waste crimes as a supervised binary classification problem between different textual content.
To tackle the problem as first task within the *C4E Project*, we select a machine learning approach using Support Vector Machines (SVM) with linear kernel and *C*=1 and Multinomial Naive Bayes (MNB) as classification algorithms (Imran et al., 2015). Since the task concerns the classification

[10] carto.com

[11] A map showing toxic fires alert tweets in the UNIOR Eye corpus is available at this link https://uniornlp.carto.com/builder/04f2cca9-08cd-4b9f-90cd-79fc0d93af42/embed

of tweets belonging to the **alert** class, to deal with the unbalanced dataset, we use the undersampling technique by automatically reducing the number of samples for the majority class (**no alert**) (Li et al., 2009), until they were balanced with the samples of the **alert** class. We used the tf-idf technique to extract the features used by both algorithms. To build algorithms and extract features, we used the Python scikit-learn library.

In addition to the MNB and SVM with tf-idf technique, we built two models with sentence embeddings as features and SVM with the tuning of *C* parameter as a classification algorithm. In the first model (FT-SVM), we used the Italian pre-trained word vectors from *fastText*[12](Bojanowski et al., 2017) to build our sentence embeddings by averaging word embeddings for all tokens for each tweet; then, *C*=10 is found as the best C parameter value using GridSearchCV[13] instance. In the second model (mDB-SVM), we generated sentence embeddings using the pretrained multilingual DistilBERT (Sanh et al., 2019) model from *Transformers*[14]. To accomplish this, each tweet is represented as a list of tokens and, then, each list is padded to the same size (*max_len* = 94). The attention mask is used. Before fitting the sentence embeddings thus constructed in the SVM classifier, it is searched for the best value of the *C* parameter set to *C*=0.1. For both models (FT-SVM and mDB-SVM) the pre-processing steps described above are performed.

5 Results

In this section, we show the results obtained by our models in terms of Precision, Recall, F-Measure and Accuracy. For all models, the results are obtained on 30% of the dataset set aside as a test set, keeping the samples balanced between the two classes. Furthermore our models were evaluated using a 10-Fold Cross-Validation[15].

As a baseline to compare with, we used Dummy classifier which achieves an accuracy of 0.501. On the test set, the SVM classifier achieves an accuracy of 0.870, while for the MNB classifier it is 0.839. Regarding the evaluation by 10-fold cross validation, our SVM reaches an accuracy of 0.868 with the mean and standard deviation of 0.008, instead the accuracy of the MNB is 0.841 with the mean and standard deviation of 0.010. In Table 2 we show the performances achieved by both models.

MNB	Precision	Recall	F-Measure
alert	**0.871**	0.816	**0.843**
no alert	0.807	**0.864**	0.835
SVM	Precision	Recall	F-Measure
alert	0.857	**0.878**	0.867
no alert	**0.883**	0.862	**0.873**

Table 2: Results in terms of Precision, Recall and F-Measure.

Both classifiers with tf-idf achieve good accuracy and seem to have a good ability to classify a considerable amount of tweets providing good results in terms of precision and recall. One of the reasons for these performances may be ascribed to a discriminating lexical composition regarding the samples belonging to the **alert** and **no alert** classes.

Regarding the accuracy of sentence embeddings models on the test set, FT-SVM reaches an accuracy of 0.822, while mDB 0.774. By evaluating the predictive performance of the two models with 10-fold cross-validation, FT-SVM achieves an accuracy of 0.825 with the mean and standard deviation of 0.011, while mDB-SVM reaches the accuracy of 0.773 with the mean and standard deviation of 0.013. In Table 3, the results in terms of Precision, Recall and F-Measure are shown.

FT-SVM	Precision	Recall	F-Measure
alert	**0.826**	0.817	0.821
no alert	0.818	**0.827**	**0.822**
mDB-SVM	Precision	Recall	F-Measure
alert	**0.785**	0.766	**0.775**
no alert	0.765	**0.783**	0.774

Table 3: Classification Reports for FT-SVM and mDB-SVM.

Both models fed with sentence embeddings constructed with different techniques, seem to perform well in this classification task. In particular, the FT-SVM model based on sentence embeddings built with *FastText* seems to have better scores in terms of Precision and F-measure than those achieved by the mDB-SVM model. One

[12]https://fasttext.cc/docs/en/pretrained-vectors.html

[13]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#

[14]https://huggingface.co/transformers/pretrained_models.html

[15]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

of the reasons could be that sentence embeddings built with *FastText* benefit from a resource tailored on the Italian language compared to a multilingual one used in DBert-SVM. Specifically, mDB-SVM achieved good results in terms of precision and f-measure for the alert class. Instead, in terms of Recall, both models have a high proportion of relevant instances for the no alert class.

5.1 Confusion Matrices

In this section we show the four confusion matrices in order to graphically display the performances achieved by the different models. In Figure 1 we show the confusion matrix of the MNB model, while in Figure 2 that of the SVM model.

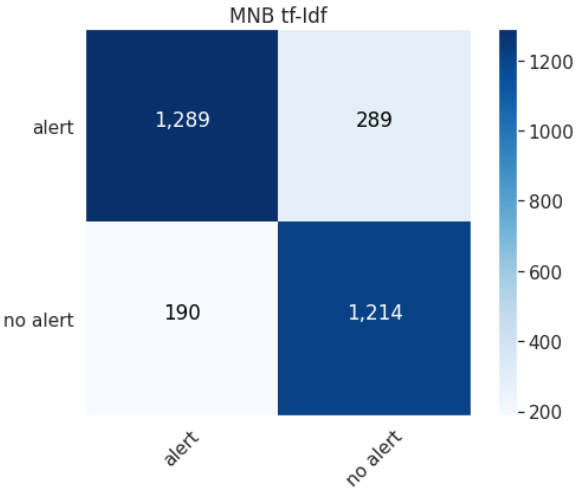

Figure 1: MNB model confusion matrix.

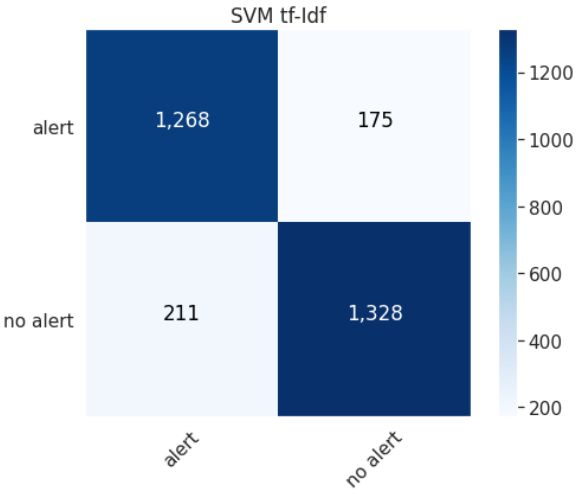

Figure 2: SVM model confusion matrix.

The confusion matrices of the FT-SVM and the mDB-SVM model are shown respectively in Figure 3 and Figure 4.

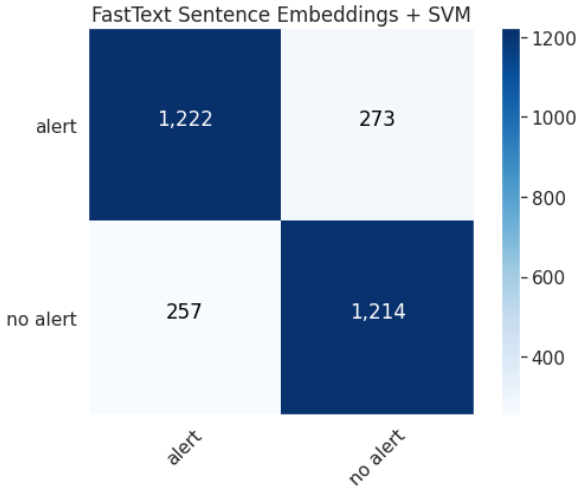

Figure 3: FT-SVM model confusion matrix.

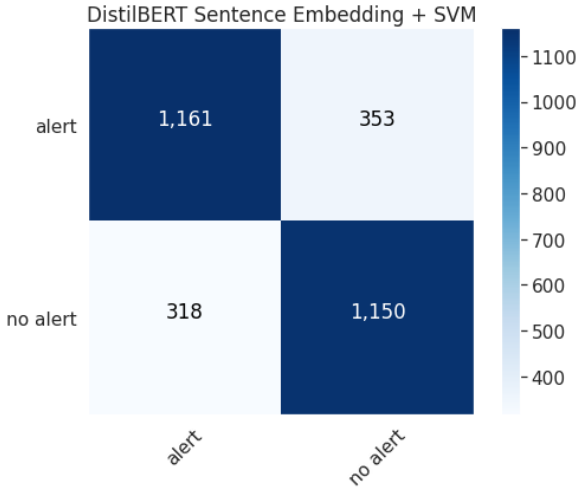

Figure 4: mDB-SVM model confusion matrix.

6 Conclusions and Future Work

We presented a case study within the C4E project aimed at monitoring social media to provide support against environmental crimes. In particular, we described the *UNIOR Eye* corpus, in some sections still in progress, on which we tested four models with three different features extraction and construction techniques on a part of the corpus. We proposed two classifiers, namely SVM and MNB, with tf-idf features as the first experiment; then, SVM with C parameter tuning fed with sentence embeddings. These embeddings were built both using Italian pre-trained *fastText* model and using pre-trained DistilBert multilingual model. Our purpose was to classify alert tweets related

to waste crimes vs no alert tweets. Future research will include the enlargement of the corpus, applications of NLP in the field of environmental protection as well as the analysis of contextual features related to environmental issues used as a medium to polarize public opinion (Karol, 2018).

Acknowledgements

This research has been carried out within the framework of two Innovative Industrial PhD projects supported by the PON Ricerca e Innovazione 2014/20 and the POR Campania FSE 2014/2020 funds and two research grants supported by the PON Ricerca e Innovazione 2014/20 in the context of the C4E project.
Authorship contribution is as follows: Raffaele Manna is author of section 4. Section 2 is by Antonio Pascucci. Section 5 is by Raffaele Manna and Antonio Pascucci. Sections 1, 3 and 6 are by Wanda Punzi Zarino and Vincenzo Simoniello. We are grateful to Prof. Johanna Monti for supervising the research.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

William J Corvey, Sarah Vieweg, Travis Rood, and Martha Palmer. 2010. Twitter in mass emergency: What nlp can contribute. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 23–24.

FIMP. 2017. Fimp ambiente - federazione italiana medici pediatriglossario di termini sull'ambiente. una guida dalla a alla z per orientarsi nel complesso tema dell'inquinamento ambientale.

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):1–38.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894*.

ISPRA. 2012. Ispra – l'istituto superiore per la protezione e la ricerca ambientale, glossario dinamico per l'ambiente ed il paesaggio.

David Karol. 2018. Party polarization on environmental issues: toward prospects for change. *Research Paper. Niskanen Center, Washington, DC*.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Yaoyong Li, Kalina Bontcheva, and Hamish Cunningham. 2009. Adapting svm for data sparseness and imbalance: a case study in information extraction. *Natural Language Engineering*, 15(2):241–271.

Miguel Maldonado, Darwin Alulema, Derlin Morocho, and Marida Proaño. 2016. System for monitoring natural disasters using natural language processing in the social network twitter. In *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*, pages 1–6. IEEE.

Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. 2011. Safety information mining—what can nlp do in a disaster—. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 965–973.

Pasquale Peluso. 2015. Dalla terra dei fuochi alle terre avvelenate: lo smaltimento illecito dei rifiuti in italia. *Rivista di Criminologia, Vittimologia e Sicurezza*, 9(2):13–30.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Muhammed Ali Sit, Caglar Koylu, and Ibrahim Demir. 2019. Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of hurricane irma. *International Journal of Digital Earth*, 12(11):1205–1229.

Kevin Stowe, Michael Paul, Martha Palmer, Leysia Palen, and Kenneth M Anderson. 2016. Identifying and categorizing disaster-related tweets. In *Proceedings of The fourth international workshop on natural language processing for social media*, pages 1–6.

Francesco Tarasconi, Michela Farina, Antonio Mazzei, and Alessio Bosca. 2017. The role of unstructured data in real-time disaster-related social media monitoring. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3769–3778. IEEE.