Versione PROVVISORIA del contributo presentato al Convegno Annuale

Hashtags as an information source. Analyzing tweets to map *La Terra dei Fuochi*

Raffaele Manna, Antonio Pascucci, Wanda Punzi Zarino, Vincenzo Simoniello, Johanna Monti

L'Orientale University of Naples - UNIOR NLP Research Group, Italy

{rmanna,apascucci,wzarino,vsimoniello,jmonti}(«»)unior.it

ABSTRACT

In this paper, we present the *UNIOR Eye* (*Earth your estate*) corpus made up of 228,412 tweets dealing with environmental crimes. The core of this research is the extraction of information from tweets according to hashtags and their frequency. Then, in an unsupervised scenario, the K-means clustering algorithm is used to automatically cluster the different contents of the tweets by keeping the hashtags hidden. The analysis of each cluster at the lexical level allows us to identify i) alert tweets, ii) political background tweets and iii) personal opinion tweets. Once the alerts tweets about *La Terra dei Fuochi* are identified (precisely 36,207), we first extract the exact location (preceded by a hashtag) where the crime took place and then establish the type of crime. This information is displayed on a map with the use of Carto.

KEYWORDS

Natural Language Processing, Corpus Linguistics, Environmental Crimes

1. INTRODUCTION

The aim of our research is to study the frequency and co-occurrence of hashtags in order to gather information related to the contents of tweets and then to monitor natural human-caused disasters. More specifically, we focus on *La Terra dei Fuochi* (literally The Land of Fires), a large area between Naples and Caserta (in the north of the Campania region, Italy) which has been plagued for about fifty years by illegal toxic wastes routinely dumped by criminal organizations to make space for new ones [Peluso, 2015]. Over the years, the public awareness about what happens in *La Terra dei Fuochi* has increased and common people have started playing a role in the fight against illegal and toxic waste disposal. In particular, they have started using social media to share real-time information [Ampofo, 2011] and report the presence of illegally disposed toxic wastes. Twitter, especially, is one of the most popular microblogging and represents a fundamental channel for sharing news and opinions related to a specific topic thanks to hashtags, namely words or unspaced phrases preceded by #, that are a valuable resource for identifying and aggregating messages and posts related to a specific subject [Austin and Jin, 2018]. Moreover, the fact that tweets contain information shared by ordinary people helps to assess the real situation of events [Cobo, Parra and Navón, 2015]. Therefore, by monitoring tweets about criminal events in a specific place, some crimes could be detected immediately [Crowe, 2012]. For this purpose, we present an unsupervised approach that hides hashtags from the tweets, in order to classify the textual contents of tweets thanks to the K-means clustering algorithm and so to discriminate the clusters related to i) alert tweets, ii) political background tweets and iii) personal opinion tweets. This research is carried out in the framework of the C4E (Crowd for the Environment) project. The paper is organized as follows: in Section 2 we discuss Related Work, in Section 3 we present the *UNIOR Eye* corpus, in Section 4 we show the case study and the map of environmental crime places detected in the corpus. Conclusions are in Section 5 along with Future Work.

2. RELATED WORK

Many researchers propose event detection approaches that monitor Twitter data and determine whether special events, such as accidents, extreme weather conditions, earthquakes or crimes, occur by analyzing hashtags. [Yang and Rayz, 2018] propose an event detection approach based on hashtags in tweets to collect information related to the Paris attack from 13 to 17 November 2015. [Wang, Liu and Gao, 2016] examine how the co-occurrence of specific hashtags leads to the virality of information during the Occupy Wall Street movement in 2011. Through network analysis, the authors identify popular hashtag types and examine co-occurrence patterns during the two days of the movement. This study shows that participants in the self-organised movement make a strategic use of hashtags to better spread their message and drive the viral movement. [Murzintcev and Cheng, 2017] propose a method for data mining on Twitter to retrieve messages about floods and tropical cyclones. The scholars describe an automated process for collecting hashtags which are strongly related to a specific event and demonstrate that hashtags are good markers for separating similar and

simultaneous events. The method uses disaster databases to find the location of an event and to estimate the area of impact.
[Chowdhury, Caragea and Caragea, 2020] construct a dataset of more than 67,000 tweets related to natural disasters of various types. The corpus was then annotated with the hashtags from the tweets, where present. By using this large dataset, the scholars further investigate a deep learning model for keyphrase extraction from general tweets, namely a joint-layer Long-Short Term Memory network trained using Multi-Task Learning (LSTM-MTL) and its variants that incorporate informal writing styles, in order to evaluate its performance capability for hashtag extraction from tweets related to different natural and environmental disasters.

3. THE *UNIOR EYE* CORPUS

The *C4E - Crowd for the Environment* project involves several partners dealing with artificial intelligence and is partly supported by the PON 2014-20 Ricerca e Innovazione fund. The C4E project aims at developing an innovative framework for the detection and monitoring of illegal spills, such as illegal landfills, micro-dumps or illegal releases in surface waters, and the organization of subsequent monitoring actions on site. For this purpose, we compiled the *UNIOR Eye* corpus [Manna et al., 2020], a corpus composed of tweets related to such environmental crimes posted over the time span 1 January 2013 - 6 August 2020. The corpus is made up of 228,412 tweets, 22,780,746 tokens, and 569,905 types and with a type/token ratio (TTR) of 0.025. The creation of the corpus was preceded by the compilation of a glossary containing 43 terms often used in relation to environmental crimes. The glossary has allowed us to create the corpus as, with the use of Twitter API, it was possible to download all tweets where these words, preceded by hashtags, were present. Therefore, hashtags were essential elements to gather the information needed to detect crimes against the environment. As regards the corpus configuration, it is internally divided into four semantic areas, each one concerning a specific environmental crime: *rifiuti e Terra dei fuochi* (waste and Terra dei fuochi) composed of 142,174 tweets; *reati contro le acque* (water-related crimes) composed of 37,473 tweets; *materiali e sostanze pericolose* (hazardous substances and materials) composed of 13,536 tweets; *incendi e roghi ambientali* (environmental fires) composed of 35,229 tweets. These sets are further divided into more specific subsets, e.g. the folder *incendi e roghi ambientali* (environmental fires) contains the subset *incendi boschivi* (forest fires) consisting of 1,434 tweets and *incendi dolosi* (arsons) consisting of 23,130 tweets.

4. CASE STUDY

This section outlines the methodology adopted to analyze the information conveyed by hashtags. In particular, our aim is to show that:

- the information conveyed by pairs (e.g. #terradeifuochi - #stopbiocidio) and trios (e.g. #rifiuti - #terradeifuochi - #m5s) of hashtags is shared by the corresponding tweets [Hong and Davidson, 2010].
- different textual contents can be discerned through the information affixed by the user via hashtags [Lim and Buntine, 2014]. More specifically, the main purpose is to separate alert tweets from tweets containing political and personal opinions in a dataset of tweets concerning *La Terra dei Fuochi*.

We start selecting from the semantic area *Terra dei Fuochi* of our corpus 77,205 tweets, previously extracted using the hashtag *#terradeifuochi* and its spelling variants (e.g. *#Terradeifuochi, #TerradeiFuochi, #terredeifuochi*). Then, we perform the following preprocessing steps: lowercasing the textual content of the tweets and the hashtags, thus avoiding the spelling variations of the same information; removing punctuation and special characters from the tweet; removing a custom set of Italian stopwords (i.e. determiners, conjunctions and prepositions) from tweets. All preprocessing steps are performed using the *NLTK* package in Python. Once these steps are completed, we focus on analyzing the co-occurrences (n-grams) of hashtags. The calculated average of token-hashtags in our corpus corresponds to 2,425 tokens-hashtags per tweet. Therefore, we extract, visualize and analyze the textual content of bigrams and trigrams, both shown in Figure 1, from the hashtags following each tweet in the dataset. Considering the hashtags bigrams and trigrams shown below, we qualitatively identify three types of textual information related to *Terra dei Fuochi*: political parties, criminal organizations and alerts. In particular, the textual link between Terra dei Fuochi and politics is reported in bigrams (#terradeifuochi and # m5s, #renzi, #napolitano and so on), while texts related to opinions about criminal organizations are more present in trigrams (e.g. #terradeifuochi #caserta #lupi). Then, keeping the hashtags hidden, we cluster the tweets using the K-means algorithm in an unsupervised way. We evaluate the number of clusters using the Elbow method, shown in Figure 2 on the left side, and silhouette score with a peak at k = 6 with value of 0.78. Therefore, six appropriate clusters are identified to describe the dataset. These clusters are presented in Figure 2 on the right. By

inspecting the words contained in each cluster, we identify two alert clusters containing words such as: *rifiuti, tossici, vicino, località, etc.* Then the cluster labels are assigned to the alert tweets. Therefore, we are able to identify the most relevant hashtags for alert tweets shown in Figure 3.

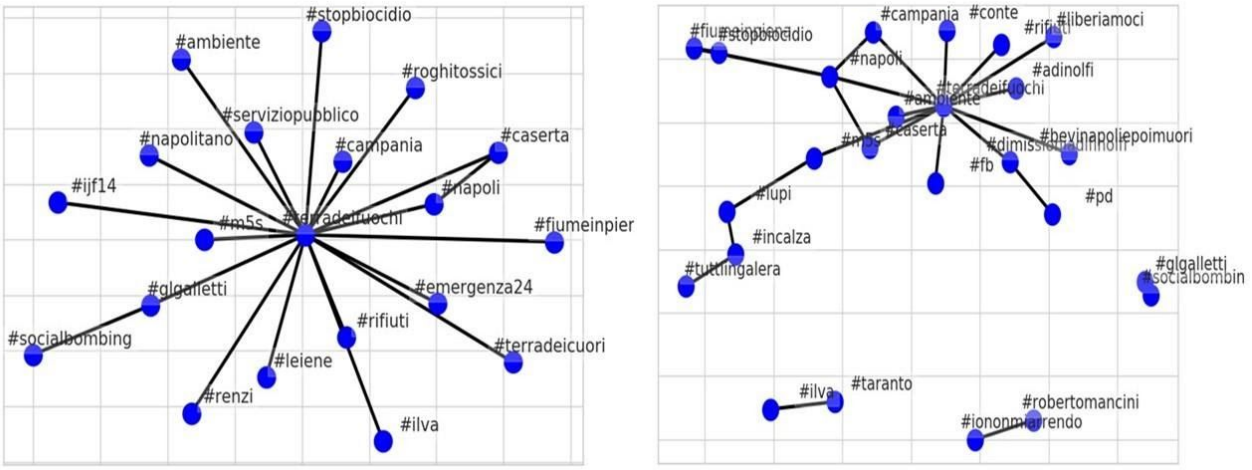

Figure 1: Hashtags bigrams and trigrams

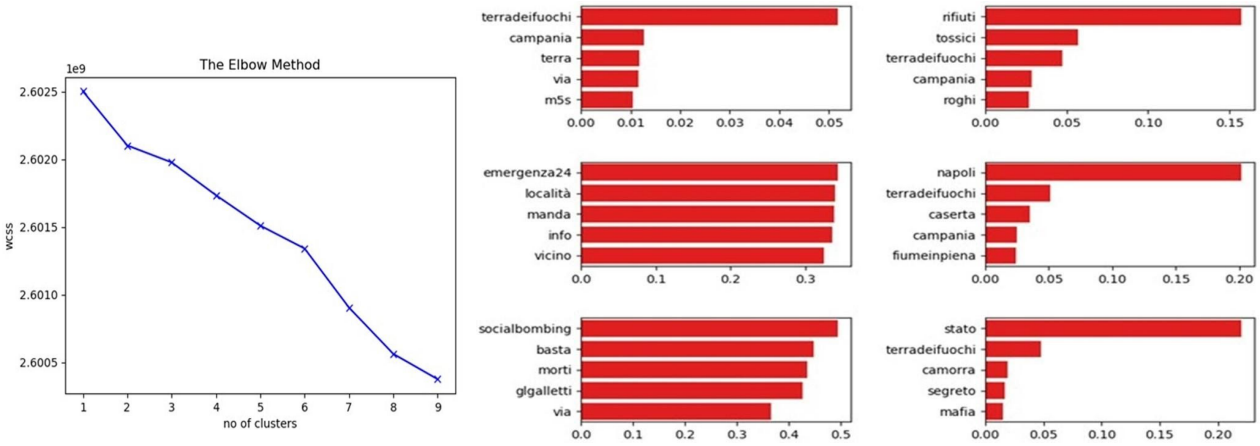

Figure 2: Elbow method on the left and six clusters on the right side

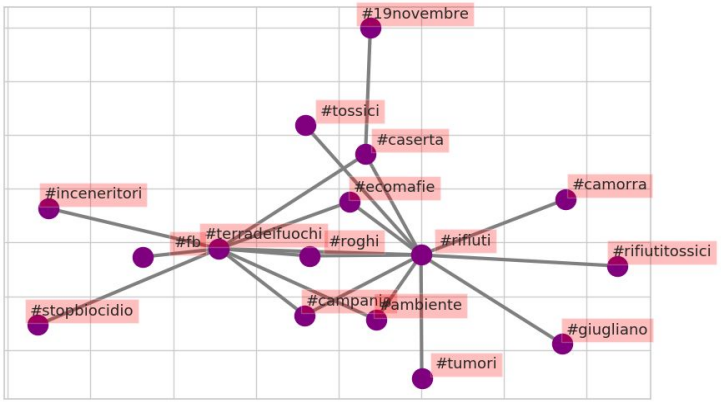

Figure 3: Relevant hashtags for the alert tweets

Once the alerts tweets about *La Terra dei Fuochi* are identified (precisely 36,207), we extract the exact place (preceded by a hashtag) where the crime took place and establish the type of crime, with the view to representing this information

on a map. Using this link it is possible to visualize the interactive Hashtag frequency map, namely the map of places preceded by a hashtag reported in the alert tweets. It is also possible to verify the instances of the alert tweets preceded by a hashtag for the single place by placing the mouse cursor on them. In Figure 4 we show a screenshot of the map.

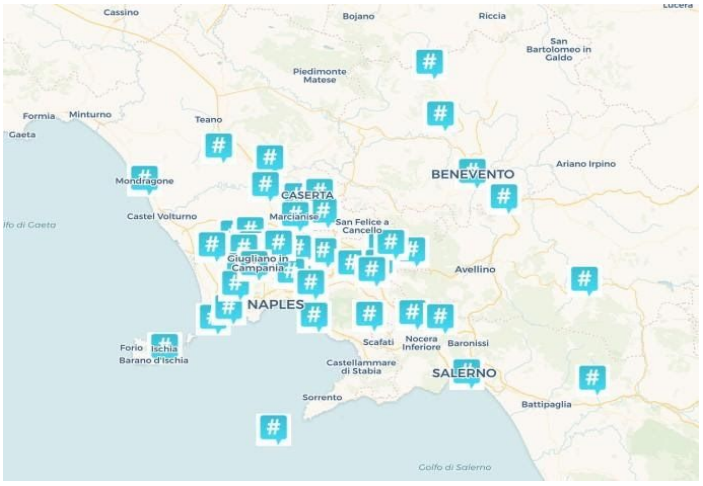

Figure 4: A screenshot of the Hashtag frequency map

5. CONCLUSIONS AND FUTURE WORK

We presented a case study concerning the *UNIOR Eye corpus,* a corpus made up of 228,412 tweets related to environmental crimes, aimed at investigating the frequency of co-occurrence of hashtags to monitor environmental crimes, particularly related to *La Terra dei Fuochi*. The research is carried out in the framework of the C4E - Crowd for the Environment project, partly supported by the PON Ricerca e Innovazione fund, which aims at developing an innovative framework for the detection and monitoring of illegal spills. Our method is based on the extraction of hashtags contained in the tweets gathered in the dataset, and then on clustering these tweets in order to identify the most relevant hashtags for each labeled alert tweet. Future work will focus on extending our hashtag-based analysis to cluster and map the entire corpus in order to provide computational methods to monitor the environment through social media.

ACKNOWLEDGEMENTS

This research has been conducted 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. Special thanks to Annarita Magliacane for helping us during this research. We are grateful to Prof. Johanna Monti for supervising the research.

BIBLIOGRAPHY

[1] Ampofo, Lawrence. (2011) "The Social life of real-time social media monitoring." *Participations. Journal of Audience and Reception Studies* 8, no. 1, 21-47.

[2] Austin, Lucinda, and Jin, Yan. (2018) "Social Media and Crisis Communication." *New York: Routledge*.

[3] Chowdhury, Jishnu R., Caragea, Cornelia and Caragea, Doina. (2020) "On Identifying Hashtags in Disaster Twitter Data." *Proceedings of the AAAI Conference on Artificial Intelligence* 34, no. 01, 498-506.

[4] Cobo, Alfredo, Parra, Denis, and Navón, Jaime. (2015) "Identifying Relevant Messages in a Twitter-based Citizen Channel for Natural Disaster Situations." *Proceedings of the 24th International Conference on World Wide Web*, 1189-1194.

[5] Crowe, Adam. (2012) "Disasters 2.0: The Application of Social Media Systems for Modern Emergency Management." *Boca Raton (FL): CRC Press*.

[6] Hong, Lianjie, and Davison, Brian D. (2010) "Empirical study of topic modeling in twitter." *Proceedings of the first workshop on social media analytics*.

[7] Lim, Kar Wai, and Buntine, Wray. (2014) "Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon." *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*.

[8] Manna, Raffaele, Pascucci, Antonio, Punzi Zarino, Wanda, Simoniello, Vincenzo and Monti, Johanna. (2020) "Monitoring Social Media to Identify Environmental Crimes through NLP. A Preliminary Study." *Proceedings of the Seventh Italian Conference on Computational Linguistics (Clic-IT)*.

[9] Murzintcev, Nikita, and Cheng, Changxiu. (2017) "Disaster hashtags in social media." *ISPRS International Journal of Geo-Information 6.7*, 204.

[10] Peluso, Pasquale. (2015) "Dalla terra dei fuochi alle terre avvelenate: lo smaltimento illecito dei rifiuti in Italia." *Rivista di Criminologia, Vittimologia e Sicurezza* 9, no. 2, 13-30.

[11] Wang, Rong, Liu, Wenlin, and Gao, Shuyang. (2016) "Hashtags and information virality in networked social movement." *Online Information Review* 40, no. 7, 850-866.

[12] Yang, Shih-Feng, and Rayz, Julia T. (2018) "An event detection approach based on Twitter hashtags." *arXiv preprint arXiv:1804.11243*.