POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

Analyzing drought impact through social media and geophysical parameters

TESI DI LAUREA MAGISTRALE IN COMPUTER SCIENCE AND ENGINEERING - INGEGNERIA INFORMATICA

Author: **Violeta Marić**

Advisor**: Prof. Barbara Pernici**

Co- Advisor: **Carlo Bono**

Academic year**: 2022-2023**

Introduction

Each year, millions of people are affected by natural disasters. These effects are seen on infrastructure, economy, environment, and the most valuable resource - human lives. The crucial point is to understand the causes and impacts to be able to build resilient societies accordingly. Traditional sources that include official reports, sensor networks or meteorological agencies are often not enough. To fill the gaps that missing information during disaster is carrying along, citizen observations spread through social media are widely considered to be a promising source of relevant information, and many studies propose new methods to tap this source. It has been anticipated that these methodologies can be helpful for policymakers, disaster relief organizations, researchers and members of civil society who wish to leverage an additional tool to better understand the impacts of extreme weather events and how to focus their efforts.

Drought is one type of natural disaster that faces the challenges in having its impact assessed, due to the gradual development of its effects over an extended period.

This work will investigate whether and to what extent social media can truly be beneficial in drought case to help overcome the lack of information that the researchers and emergency responders are facing. By cross-referencing information from multiple sources, the study will try to evaluate user credibility during emergency events and confirm its perimeter. Furthermore, the study addresses some of the questions that were raised due to the lack of existing connections between an event and its effects. The contribution is also in helping the assessment of situational awareness reached during the detected natural disaster and guiding the response to perform the actions in categories where it left damage. To capture existing drought conditions, the work will consider assessing drought impacts, response, and its recovery, from Twitter user's perspective. In this way, it is easier to note viral trends and

dynamics, as it provides more insights through user engagement metrics and other metadata.
The innovative approach adopted in this study is the leverage of multilingual BERT model for the classification of tweets into multiple drought impact categories. The study focused on the drought situation in Italy, mainly during 2021 and 2022, a period characterized by the anticipation of reaching peak values of drought severity.
The analysis will encompass multiple data sources, including official reports, social media data, and other relevant information, to provide a comprehensive understanding of the drought situation and its impacts in Italy during these critical years.

1. State of the art

Getting information about areas affected by natural disasters and assessing the severity of impacts is possible by applying some data analysis techniques to automatically extract insights and patterns from the data. User-generated data can fill information gaps by combining human intelligence together with artificial intelligence improving disaster management capabilities for labor-intensive manual tasks through employing digital volunteers [1].
Recent advances in deep learning technologies have been applied to design disaster classification models [8]. Transformers-based language models are widely used in natural disasters domain where they demonstrate outstanding performance.
Even though the leverage of these models is examined in various studies on the topic of natural disasters, in drought domain the research is more restricted in providing impact assessment. The demanding topic of drought is studied broadly, but there were only a few research studies that dealt with social media generated data [2], [3].
The study [2] performed research on drought relevant data using news media articles and classification with the base version of BERT. The validation of the model built is done on drought relevant tweets from United States having as a validation parameter category-specific keyword label.
The innovative approach adopted in this study is the combination of multilingual BERT model trained on social media data and traditional sources for assessment of drought impact.

2. Drought in Italy 2021-2022

In Italy, 2022 turned out to be a year of climatic extremes, as one of the hottest years. From the Drought Observatory[1] report, severe drought has affected areas of northern Italy and the Po River since December 2021.
By analysing Standard Precipitation Index (SPI), the results revealed a progressive worsening from the late 2021. The lack of precipitation lasted during the whole year, with several regions affected at the end of December 2022. As far as Italy is concerned, it has been detected in the provincial capitals an increase of temperature by 1.2C° compared to average 1971-2000. During the period preceding the core part of the growing season, plants have lower photosynthetic activity and lower water needs, so temperature becomes the main constraint for photosynthetic activity. The significantly lower indicator's values are visible in July and August, especially in the north area, but also in the central and south parts.
The strong instability of the international markets of agricultural raw materials and energy products was amplified partially by drought, that characterized the entire year affecting the volumes and quality of many crops. In 2022, agricultural production is reduced by 0.7%. There was a significant decrement of oil (-17%) and cereal (-10,4%) production. The drought influenced the strong reduction in produced hydroelectric energy by about -40%. The impact report showed energy storage being affected from the middle of 2021, but the exceptionality of 2022 in comparison to the previous 6 years is evident.
With the start of the new summer season yet another drought alarm returns and forcefully to flip. This time the data appears particularly alarming with some areas of the Peninsula, especially the North, who are living through the worst crisis of the last 70 years.

3. Analysis of drought-related tweets

Twitter as a free social networking site is easily accessible by a vast majority of the population.

[1] https://drought.climateservices.it/

Users broadcast on it short posts – tweets that can contain text, videos, photos, or links. They are permanent, searchable, and public. Keyword-based extraction of the tweets is performed using Twitter API with words '*siccita*' or '*siccità*', Italian translation of drought, for 2020, 2021, and 2022.
This kind of content often poses challenges for analysis since data contain a lot of noise, which makes more difficult to isolate signals from it. To remove mentioned noise, it is proceeded with pre-processing task, to remove irrelevant parts of the tweets. In this case, the cleaning part consisted of removing URLs, HTML tags, user mentions, special characters and stop words for Italian language. Additionally, all duplicate values are removed, and the final number of samples in datasets is presented in *Figure 3.1*.

Year	Samples after cleaning
2020	6949
2021	8313
2022	85478

Figure 3.1: Number of samples for each year after pre-processing task

3.1. Labelling task

Due to lack of publicly available tweet datasets that contain specific annotations in drought assessment field, it has been decided to perform manual labelling of drought-related tweets, to have dataset tailored to this specific case study.
The categories of drought impacts considered were the same as in Drought Impact Reporter (DIR) dataset. This dataset consists mainly of news media data that has been manually annotated by experts from NDMC[2].
To each tweet, 7 values have been assigned. The values considered were 0 (a tweet does not belong to a certain category) or 1 (a tweet belongs to a certain category) and were associated to each of 7 categories. Alternatively, if all 7 values are 0, this tweet sample is further considered as irrelevant.
The final distribution of labelled categories is represented in *Figure 3.2*.

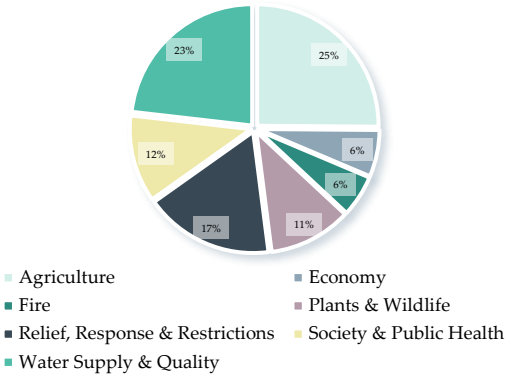

Figure 3.2: Labelled tweets distribution of categories they belong to

4. Applying multilingual BERT model on drought-related tweets

Encoder-based models are best suited for tasks that require an understanding of the entirety of the provided context. BERT is considered having state-of-the-art performance on various NLP tasks [4].
The study [2] performed research on drought relevant data using news media articles and classification with the base version of BERT. The validation of the model built is done on drought relevant tweets from United States having as a validation parameter category-specific keyword label.
In this chapter, multilingual BERT model is applied instead and further fine-tuned for multilabel classification problem on tweet content. This phase of the study was mainly experimental, due to faced limitations of more specific documentation on the methodology and common practices used for multilabel classification with BERT model on tweet content.

4.1. Fine-tuning BERT model

To fine-tune pre-trained multilingual BERT on drought-related data, two techniques were evaluated:

1. Freezing pre-trained weights
2. Updating pre-trained weights

Dataset considered was DIR news media dataset, due to its larger number of samples and higher

[2] National Drought Mitigation Center https://drought.unl.edu/

accuracy in annotation, since it has been labelled by domain experts. These approaches were assessed with two different architectures:

1. Dense layer followed by Sigmoid activation function
2. Dense layer and ReLU function are added to 1.

When freezing weights, best performance showed more complex architecture with batch=32 and learning rate=1e-3. Model was trained for 5 epochs and achieved F1 score of 0.7266.
In the other case, model performed better overall when updating pre-trained parameters, with F1 score variating between 0.8345 and 0.8393.

4.2. Multilabel tweet classification

In previous section, the model has been trained on news media data to find the architecture that shows the best results. However, news media data can fail to capture the characteristics that tweet posts contain and primarily informal language of its users. As the focus of the study is social media content, and the assessment of the impacts recognized by Twitter population, the ultimate goal is to classify tweet datasets. The architecture that yielded the best performances will be employed in two different strategies to evaluate tweet datasets:

1. applying the model trained on news media data on tweet dataset
2. using tweet datasets for both training and evaluation

In this phase, since all architectures gave analogous results, the selected model, to move forward with is one dense layer and batch size=16, learning rate=3e-5 and with 4 epochs of training.
The results of the model are reported in the next chapter.

5. Results

A comprehensive analysis will be reported in order to assess the results of multilabel classifier model built on drought-related tweets in Italian. Additionally, the patterns and trends observed from model results and tweet datasets retrieved for three consecutive years, will be compared with the information obtained from traditional sources.

When examining the content of the tweet, in last 2020, 2021 and 2022, mainly the same topics were dominate in drought datasets. One example for 2022 is presented in *Figure 5.1.*

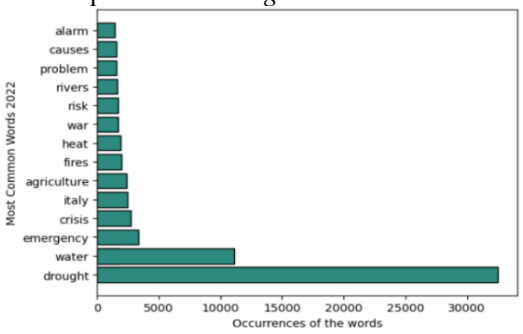

Figure 5.1: Distribution of top 10 most common words in tweets for 2022

Words *crisis* and *emergency* are indicating severity and urgency of an event, and the fact that on *Figure 5.1* the words reached top 4 most common is suggesting that drought situation worsened.

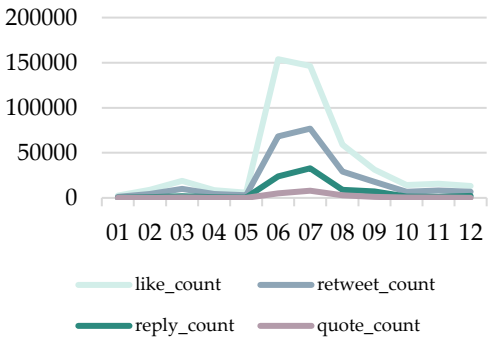

Figure 5.2: Evaluation of user engagement metrics for tweets from 2022

Instead, on *Figure 5.2* it is visible how the user engagement metrics reached their peak in summer period of 2022, the most important season for drought conditions.

5.1. Model results

In the first part, in case of applying transfer learning, the model is trained with news media and then applied on tweet dataset.
The results obtained indicate *Society & Public Health* as the most impacted from drought followed by *Water Supply & Quality*. Although this scenario could be explained with other impacts having effect on population, directly or indirectly, it is very unlikely to happen. Furthermore, it would be expected to see society related topics into the trends and patterns, but it is not the case.

In the second case, multilingual BERT pre-trained model with added classification layer on top was fine-tuned on drought-related, manually labelled dataset and then assessed on tweet datasets for different years.
The tweets signed as relevant have distribution for each category as showed in *Figure 5.3*

Figure 5.3: Distribution of drought impacts contained in tweets from 2022

Classification results found water-related topics dominating in Twitter discussions among Italian users. *Water Supply & Quality* gained the most attention, indicating significant concerns regarding water scarcity and degradation. Agriculture emerged as the second most prominent drought topic, showing impacts on crops, soil, and the economy. Water management plays a crucial role in this field too, influencing crop quality, yield, and overall sustainability.

5.2. Comparing model results with traditional data sources

When it comes to assessing the results obtained, one of the best practices is to confront them with the ones coming from the official sources, taking into consideration their reliability and credibility. The correlation and combination of the two sources can help in defining the accuracy and validity of the findings.
An important conclusion drawn from the posts and also confirmed by other sources was that the drought situation is significantly worsening considering the three-year time frame. From the analysis performed on tweets, there is a momentous increase in tweets posted in 2022, 10 times more than 2021 and even 12 times more than 2020. Even though the drought has been detected according to the reports since the end of 2021 at the north of the country, some time needed to pass for users to start to discuss this topic since the peaks are visible in dry season months.
The official reports from EDO[3] can be supported by the significant attention that was brought by people on Twitter, where it can be also seen a serious decrease in SPI measurements. As a matter of fact, one of the most considered topics on Twitter was water, in all forms, whether when talking about the water scarcity in rivers and households or water necessary for soil irrigation. *Water Scarcity & Quality* impact category certifies the substantial role that water plays in the drought effects, whether as a direct, or indirect impact on overall water cycle and ecosystem dynamics as well as its impact in other categories. Due to all mentioned statistics, it is more likely that water resource is the category that has priority. Hence, it can be deducted that the model trained and evaluated on tweets has a higher possibility of depicting the real scenario on drought impacts.
Moreover, there are some other connections that we can make between Twitter analysis and traditional reports. The temperature deviations registered in Maximum Temperatures Anomaly can be justified by heat reaching to one of the most common words discussed about, for all three years. On the other side, the decrement in energy storage did not pass unnoticed on Twitter either. Energy crises as a term can be seen in the top 10 most common biagrams in 2022. The percentage of the population exposed to severe-extreme drought reached the highest value in July 2022 with the long-term overall percentage of 47%. The analysis of user engagement metrics in 2022 is indeed analogous to statistics since it shows significant spike in the months of June and July.

6. Conclusion and future development

Acquiring a better understanding of drought impacts becomes increasingly vital under a warming climate. Traditional drought indices describe mainly biophysical variables and not impacts on social, economic, and environmental systems [2].
The Twitter analysis performed in this study can help to understand better its role and in general social media role in disaster management and the ways in which can be beneficial for the process.

[3] European Drought Observatory

The secondary goal of deep tweet analysis was to detect the baseline rate of tweets in Italy over the past two years (2020-2021) and then use this information to detect whether the rate was within the boundaries for the third year (2022). Indeed, the observed tweet volumes were analogous to the surges or spikes in attention that emergency managers use to identify events of interest.

The study was also complemented by the qualitative part about some most common experiences. From all the categories that may have been impacted by the drought conditions, the results showed that the most affected was water resources. Approximately 46% of tweets were related to this category, highlighting its central role during drought events, and even after leaving impacts for the future. Moreover, for other categories where water management plays a crucial role, the impacts are escalating, which puts the category of agriculture in the second place of the most impacted and damaged fields.

The awareness of the drought problem, raised by tweets analysis can positively impact the decision-making process in the impacted fields. Even basic content retrieved from Twitter by a single keyword search, like in this case, can provide insight on what type of impacts people are experiencing and eventually identify new types of categories to consider.

By looking at the tweets extracted, it is noticed that they could also be grouped based on whether they are reporting information on the current situation and assessing the damage or providing possible suggestions for improvements instead. The division in this way could help distinguish pure drought assessment from assistance demand and provide relevant information to each field separately. Another possible improvement can be exploring other search terms more specific to each category. Many tweets may relate to dry conditions even without explicitly using the word "*drought*".

This topic is still under discussion, and it continues to attract the needed attention, as by considering moderate emission scenario (RCP4.5) drought events are calculated to become considerably more severe in the period 2071–2100 than 1981–2010 for Mediterranean region or even larger [5].

7. Bibliography

[1] C. Havas *et al.*, "E2mC: Improving emergency management service practice through social media and crowdsourcing analysis in near real time," *Sensors (Switzerland)*, vol. 17, no. 12, Dec. 2017, doi: 10.3390/s17122766.

[2] B. Zhang, F. Schilder, K. H. Smith, M. J. Hayes, S. Harms, and T. Tadesse, "TweetDrought: A Deep-Learning Drought Impacts Recognizer based on Twitter Data."

[3] S. Mukherjee, S. Wang, D. Hirschfeld, J. Lisonbee, and R. Gillies, "Feasibility of Adding Twitter Data to Aid Drought Depiction: Case Study in Colorado," *Water (Switzerland)*, vol. 14, no. 18, Sep. 2022, doi: 10.3390/w14182773.

[4] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." [Online]. Available: https://github.com/tensorflow/tensor2tensor

[5] J. Spinoni, J. V. Vogt, G. Naumann, P. Barbosa, and A. Dosio, "Will drought events become more frequent and severe in Europe?," *International Journal of Climatology*, vol. 38, no. 4, pp. 1718–1736, Mar. 2018, doi: 10.1002/joc.5291.