Disaster Tweets Classification using NLP

Name: Narva Siddhartha, SR University, Narvasiddhartha@gmail.com

CSE{AI & ML}, Warangal,India,

Abstract*— *Social media platforms like Twitter brings many details during disasters! Currently data is not just words, but also special characters, sounds, symbols and more on Twitter! Time is pricelessness in sharing emergency information and responses according to studies. Twitter was researched and discovered as utilized effectively for spreading valuable information about different disasters such as volcano eruption, floods, geopolitical war devastation to help authorities (like NDMA) respond quickly and effectively to disasters! The main objective of the research is attained by using supervised machine learning and natural language processing to classify tweets as disaster-related or non-disaster related. The suggested approach is based on Word2Vec & Machine Learning Algorithm, where Word2Vec provides feature vectors and Classification Algorithms is used to categorize tweets.

Keywords— Disaster Responses, Twitter data, Natural language processing, Word2Vec, NDMA (National Disaster Management Authority).

I. Introduction

Social media helps the scientists understand the people's conduct, their reaction to and source of information in the events that affect them, and natural disasters including. Such things are useful for understanding the type of information which is disseminated through which channels, as well as what techniques and decisions for choices do people undertake. On the one hand, a significant amount of information is being provided by Twitter and other social media that filtering of noise or paying for the expenses of an appropriate platform is obligatory. Such a system has several processing phases, including the classifier that picks out disaster-related-tweets and the role of the classifier that categorizes the relevant tweets according to fine-grained types such as preparation and evacuation. Fundamentally, there is a lot of research on both which social media data makes sense and gets retained in data warehouses, and also is an analysis of which of the most important information gets extracted. In this system looking at finding out which of all tweets are useful and consequently during natural catastrophes when people need to be cautious are considered, the identification of informative tweets from social media data will be the focus. Twitter, microblogging sites, and these kinds of media sites have made it possible for everything to be in real time; in these places online, people are digitally coming together in a fast and frantic kind of way. It would be of paramount importance if there would be a system which would not filter tweets on the basis of connotational & dialogical content because relief efforts might obtain a remarkable preponderance in learning where to send aid.

By teaming up the technology of Geo-location, Sentiment Analysis and other social media data mining techniques, an effective social media filtering research can result to well-informed and correct decisions, thereby, leading to reduction in casualties or harmed bystanders. The workflow stage being in this system, we want to develop original features that can be remitted to machine learning classifiers so that they can automatically and correctly determine which tweets are informational and which are not in short time. In the project, their design would classify the tweets into categories that are related to the disaster, Filter tweets with narrow granularities, and evacuation preparations are the system as well. This automatic tweeting content tidying lock hold will not only be useful in the light of disaster strikes and also in there after. In crisis times, tweets can have the capability to govern situations and responders by their ability to act effectively! The reason might be that results speaker's model may be caused by social information data during the instance. With this awareness, communication a response measure to be more effective in risk management and minimizing of the potential disaster driving the eventual damages. In this proposed system besides the safety information mining task which we described in the previous sections, we observed many other efforts to help the earthquake victims via NLP technologies as sub-projects of ANPI NLP.Association of tweets to evacuation shelter locations: This includes geo-coding shelter lists which were taken from tweets, and assigning them locations as we geo-code the evaluation shelter lists. In addition, we consider automatic ways to extract knowledge from twitters under the proposed system. Here, we would like to emphasize on extracting the key "information nuggets", i.e., brief and complete data bits which are important for response in disaster. Our approach employs classifiers based on machine learning algorithms and information extraction technologies for this purpose. The outputs of our research, confirmed by working through one large data set for disaster relief, demonstrate that the well-thought-out design of the system can produce a usable system, which can be the base for the further development of more sophisticated data analysis and visualization systems. System must be able to find the mentions that add more situational awareness information - that is tweets which give the following types of information: tactical, actionable information, informing individuals on where to obtain it, or providing immediate after-impact help to people experiencing the great emergency. Mainly, it's the use of supervised classification algorithm.

This proposed system is divided into four categories which are: volcanic, telluric, lightning, and shots as it lies in a seismic zone and that rainfall cannot be anticipated and all the tweets are kept in the database for systematic investigation. Filtering through the NLTK tool to determine the frequency of a word in a tweet is an example of which the word is eventually classified into one of the categories. The results of the statistical data collected through these three categories get represented in a public website which shows the graphs on real time webpages. Conversely with the aid of people this work not only help this information supply for the class disasters, both the natural and private.

II. LITERATURE SURVEY

- This paper is concerned with the analysis of twitter data, produced by thousands of posts shared by citizens worldwide. In this case mention the information is unstructured and untabulated mean no data belongs to natural formatted data format. To discover that we should use Natural Language Processing (NLP) to recognize and analyze the fields we can extract information for our research. NLP can be considered as a field of research and development that studies linguistics, natural language understanding, designing and implementing information technology, text mining, opinion mining and artificial intelligence by

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

focusing on interactions between and machines and users natural spoken or written languages. It pertains to a subject of study on computers that applies in space for NLP extraction, processing and analysis of vast data. The language that humans speak is utilized for communication everyday. Natural language data is what it is called. Sources of the natural language processing problems are text extraction from speeches, speech recognition, natural language analysis, to figure it out and natural language generation. NLP is also considered as the humanization or the automation of the natural language such as speech and text but with the aid of artificial intelligence which is usually based on a programming system. The subject of natural language processing has evolved so much and it has spread its net to a lot of different areas even before the 50th year. Linguistic automated tools and their tremendous growth in processing capacity has made NLP as one of the hottest areas of technology discipline used today. Therefore, NLP can basically be defined as the machines' understanding of human speech manifestations and ability to make sense of human speech so as to unearth the underlying message. NLP is a widely used applications today, which range from similarity of text in E-commerce sites to the drug interaction risk assessment in health systems. Today, NLU is embedded in a number of translation software systems, word processors, spelling and grammar checkers such as Grammarly, Interactive Voice Response (IVR) and many more.

- A key part of the paper is an area of twitter data analysis. There is a great deal of research concerning Sentiment Analysis of data tweeter. Basically tweets are a kind of publicity created by individuals. Heaven knows there are many institutions that try to find out how the public in general or with regard to any particular issue feels. The main purpose of sentiment analysis of tweets is to calculate through computation, if the data in a tweet is positive, negative or neutral. The activity is the extraction of twitter data from consumer key, access key and access token, using the Twitter API. For this, the data is explored using advanced analytical tools like data exploratory analysis in order to understand whether there is a class imbalance or not. Class imbalance is a problem where there is one class that is hugely under-represented compared to the others. In the case of class imbalance, up-scaling or down-scaling is a must. Up-scaling covers increasing the class occurrence of the class feature which has a low relative number and comes down to reducing the class occurrence of the class characteristic which is a striking big number. It is only after class imbalance is removed, that a training method for generating a classifier can be yielded. The tweets that we have taken from the twitter repository need pre-processing and data cleaning. Thus, data pre-processing processes are corpus creation, lower-case conversion, stop-words removal, punctuation removal, URL removal, username removal, leading and ending blank spaces removal and text stemming.

- ***DATA COLLECTION :***

- The data collection process for the project was the obtaining of tweets from different platforms and repositories to create a dataset the classification model could be trained and evaluated. The Twitter API act as a national source facilitating real-time tweets with the use of specified keywords. hashtags and the location of geographic disasters. Apart from the original stakeholder datasets, the data from the disaster response organizations and research repositories were also used to supplement the data set. The preprocessing steps were applied to clean the collected data among which were the removal of any noise, setting the text to a specific format and handling missing values. One crucial aspect of data collection was labeling of tweets to determine what kind of tweets referred to disasters and what did not. The annotation process requiring manual labeling by human annotators or, if not the case, automated labeling by already-existing data labels. During data collection, the constant ethical considerations were our main concern and so we put all the necessary measures in place to protect user privacy and adhere to usage policies of the data. Dataset that was summarized showed different features including distribution of natural disasters and other non-disasters tweets, also biases and limitations of social media information were considered. These important aspects were documented in a precise way to ensure the clarity and repeatability during the stages of model development and evaluation.

- **Source and Methods of Collecting Data:**

- The data collection process in this project for tweets classification related to the disaster was to obtain data mostly from Twitter through a Twitter Application Programming Interface and archived datasets obtained from repositories and disaster response agencies. The Twitter API was used as a tool to scrape tweets in real-time in which they were filtered by specific keywords, hashtags and locations that were relevant to disasters enabling more up-to-date and relevant information to be obtained. Besides, labelled tweets were acquired for supplementary purposes from archived datasets, which cover different disasters or event.
- The keywords/hashtags were meticulously chosen to cover a wide range of potential disasters like the earthquakes, hurricanes and any kind of human-made incidents like accidents and terrorist attacks. Twitter data were collected using Twitter API, which was operated following a systematic approach and kept within the rate limits and API restrictions to comply with the Twitter's terms of service. At the same time, the aim was to obtain open datasets from reliable sources, which after preprocessing steps were used for the quality control purposes.
- Human annotators were allocated to datasets that required labeling and were responsible for classifying tweets with regards to predetermined criteria's. Besides, automated approaches like machine learning models were investigated which were trained on labeled data in order to accelerate annotation by methods and increase the data collection process. Ethical considerations which related to use of data and privacy were observed during the process of data collection.

- Summarizing, the use of the Twitter API and seizing archived datasets proved to be well-rounded dataset for training and evaluation of the disaster tweets classification model. This integrated methodology comprising both recent and historical data, increased the rigidity of the model ensures the reproducibility of the model even assuming 300 words as a description.

- **PREPROCESSING AND FEATURE SELECTION:**

- The data refinement process during the project of classifying disaster tweets was done through systematic way to modify raw data text for meaningful analysis. Conversion of all characters to lowercase in the beginning was to achieve uniformity in feature representation as all the characters were represented. The last step in the noise reduction techniques was to remove special characters, punctuation, numbers, hashtags and user mentions, and it also involved the process of cleaning up and focusing on trending subject matters.

- Tokenization which is the act of separating text into individual words and mean that the text can be analyzed by breaking it into adequate parts. Lemmatization achieved reduction of words to their core forms that generalized vocabulary and respectively contributed to the semantic copulativeness. We eliminated the insignificant words, which only have no particular semantic sense, to thrown away the noise and increase the computers' productivity.

- TF-IDF technique emerged as a strong candidate to numericize the textual features, the further analyzed tokens. Each word in the given sentence carries the relative frequency of the same word in all documents of a given corpus, with less frequent words being more informative than common words.

- What N-gram analysis did was to enrich feature representation by simultaneously capturing sequences of words, a fact that the model could use when trying to understand the subtle relationships contexts. The use of both unigram and n-grams was successful in improving the model's ability to capture the context of individual words within sentences.

- In sum, these steps guaranteed the details of the text data to be rightly processed, standardized and turned into a feature rich format, which is in turn suitable for classification. The database was shaped through gradual reduction of noise, standardization of text, and extraction of distinguishing features so that it could serve a sound basis for classifiers to work accurately upon. This skillful preprocessing formed the cornerstone of the subsequent modeling stage and made it possible for the classifier to separate the disaster-related tweets from the noise of social media hangouts with high precision.

• Overview of Feature Selection Methods:

- The methods of feature selection not only help in paring dataset down but also aim at identifying important features and reducing the number of features of the dataset so it can be easy to model. Classification task with respect to disasters, was made to be improved by using advanced technique as an approach in the project.

- **TF/TF-IDF Transformation** calculates the term frequencies for terms that are significant to differentiate documents. Information Gain seeks to compute the reduction in entropy used during the addition of a new feature. Those features are considered higher due to their information gain. Chi-squared Test reviews independence of not from class labels, keep those of having high scores. It is mutual information that pinpoints mutual dependency between features and class labels meanwhile concentrating on the features which have a higher information measure.

- RFE is an iterative method which eliminates features sequentially by importance feature selection the most appropriate ones. Lasso (L1 regularization) induces sparsity to the weighting of the features by preserving only those that feature is non-zero values. Trees-based approach type itself implies feature selection since tree segments the features by dividing the tree at each split to those which are most informative. PCA is the one that allows for a transformation of features into a lower-dimension space but preserving a key information about those lines.

- These methods of data refining are utilized to perform selection on informative and relevant features, and thus improving the result achieved as well as the model interpretability. Each method has its own aspects that can be usefully in different datasets and applications. Some methods will be better in some situations while in others the other methods will be preferred.

• Preprocessing and Feature Selection Steps:

- Data cleaning and feature selection were integral parts in data refinement process in disaster tweets classification project that is crucial for efficient tweet classification. The pre-processing primarily involved standardizing the text which became removing all the capitalized letters and keeping all the texts in lowercases for a consistent pattern. The noise removal process followed, which expunged special characters, punctuation, numeric figures, hashtags, and user mentions to concentrate more on the matters that brought out the importance. Tokenization gave words so as to enable analysis, as well as lemmatization (words to their base form) making words of same meaning be the same. From stop words to other types of noise, removal has also been a point to note. TF-IDF conversion of text into numeric features led to weight assignment of the words based on the the words being important.

- Feature selection includes identifying the most informative features which consider the dimensionality reduction measure. Usually, information gain, chi-squared test and mutual information algorithms check for feature independence from class labels and value, calculating they relevant ones. Instead of RFE, we utilized the feature selection technique which is Recursive Feature Elimination that removed least essential features each time. Shrinkage in L1 regularization (Lasso), on the other side, allowed a feature to be weighted significantly, discarding other features. Select feature at each split is under tree-based methods' control where it inherently performs feature selection. PCA - principal component analysis - can do several things at once. They combined the components together by selecting only the most essential ones which they then represent in a lower dimensional space.

- Therefore, by following these paths, the dataset that consists of just precisely related and informative characteristics is generated which, in turn, boosts the classifier performance and makes it predictive. Different from spoken dialogues, standardization and noise cancellation brought a stability and persistent focus to the text data composition. We used digital processing techniques for qualitative analysis enabling to tokenize and lemmatize text so that it was divided into smaller meaningful units, and the vocabulary normalized. The process of TF-IDF transformation converted the text into numerical format and thereby makes the word more important. Attribute selection methods were an efficient tool that distinguished important attributes and managing the feature dimensionality and complexity. Every move exactly added to the data set smoothly burning and including the features in an efficient manner of classification, which formed an essential part of the model improvement.

MODEL DEVELOPMENT:

- A systematic technique was employed during the model development stage of the disaster tweets categorization project, and it was facilitated the creation of reliable and accurate classification model. Initially, **exploratory data analysis (EDA)** had to be tackled first in order to discover the distribution of classes, characteristic of the dataset, and patterns behind it. After EDA, the classification methods employed were **sklearn's logistic regression, SVM** which stands for support vector machines, decision trees and random forests and MB-Classifier which is an algorithm that uses gradient boosting such as **XGBoost**. The performance of each algorithm was assessed by taking into account metrics such as accuracy, speed, and interpretability.

- Hyperparameters of the algorithms were the next step in the process, and they were tuned to the max for better performance. We did things like grid search or random search to try and identify hyperparameters that could ensure the highest accuracy in the model. Thereafter, the optimal hyperparameters were selected and the algorithms were trained on the further refined training data that had been cleaned and feature engineered. During training which includes the process of pattern and relationship discovery in data, our models learned to make true predictions.

- Then evaluation phase was conducted using the newly made validation dataset to check the models capabilities. Evaluation tools like accuracy, precision, recall, F1-score and area under the **ROC** curve **(AUC-ROC)** were applied to judge how good the models were in solving the problem of the right classification of tweets as a disaster and a non-disaster one. The evaluation results were the basis for the modifications and extensions to the models which were still necessary.

- We did comprehensive assessment and tuning in order to figure out the finest model. So, as a result, we chose the most performing model from all of them as the classification model for our scenario of the disastrous event. Afterwards, the deployment of the final model in a production environment, a continuous action where a real-time data processing and real-time predictions were performed. Through a systematic sequencing of processes, a robust model classifier with high accuracy has been developed, and it can be used for disaster related tweets classification to give valuable insights for disaster management and response. In this way, it may also contribute to mitigation efforts.

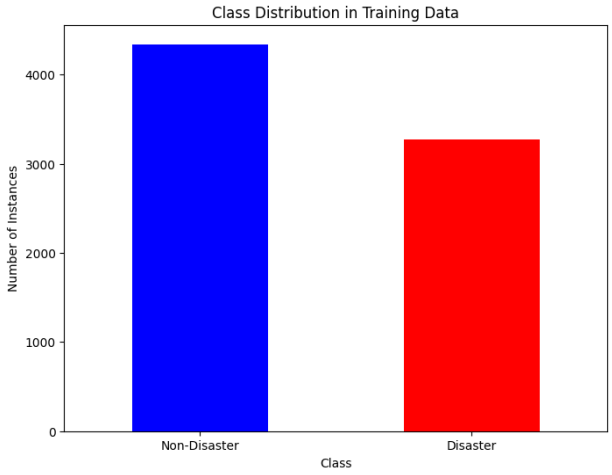
Experimental Results:

In this part, we will describe the experimental results in detail. We start with a sub-section that outlines the datasets, after which we illustrate the results of our experiments:

Class Distribution in Training Data:

The graph illustrates how tweets in the training data are divided into two classes: "Disaster" and "Non-Disaster". On the x-axis, these classes are represented individually. Each bar in the graph corresponds to one of these classes, with "non-disaster" tweets depicted in blue and "Disaster" tweets in red. The x-axis quantifies the quantity of instances, or tweets, associated with every class. Therefore, the tallness of every bar indicates the number of tweets credited to the particular class. For example, a green bar stretching to a height of 200 signifies the existence of 200 "Disaster" tweets in the testing data.

This representation serves to offer a swift glance at the distribution of categories within the testing data. It permits the evaluation of class balance or imbalance, which is crucial for model evolution and estimation. Additionally, it eases the recognition of potential prejudices in the dataset, supporting in the formulation of enlightened decisions throughout the modeling method.



Word Cloud:

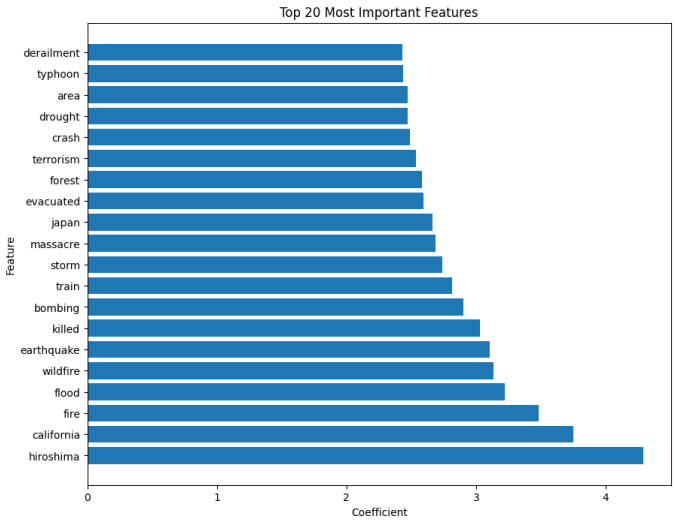
The visual representation showcases word clouds derived from tweets classified into two distinct categories: In particular, every type of "catastrophe" or "non-disaster" is teasing for us. These clouds represent the consciousness of speech that every area of language has brought into the language use. With the help of "Disaster" tweets, the cloud app comes to the judgment of what stands disaster in a widespread important with regard to the lively most outspoken words and phrases. These words and phrases manifest the overall perception of the community about disaster. The "Non-Disaster" tweets are, however, in stark contrast to the vocabulary used within them. This is because these ones portray

semantics of a realm which is not crisis-affected. Disposing of the word clouds in this way, reveals on the one hand a sharp visual difference between the now him motifs of language for normalized communication., compared to him the “unusual” language patterns of disaster-related parole. People can discover those critical insights highlighted by the word clouds graphical representations for them to comprehend the specific language trends that tends to be uniquely found in different contexts.

In the above picture we have, Disaster Tweets Word Cloud and non-Disaster Tweets Word Cloud.

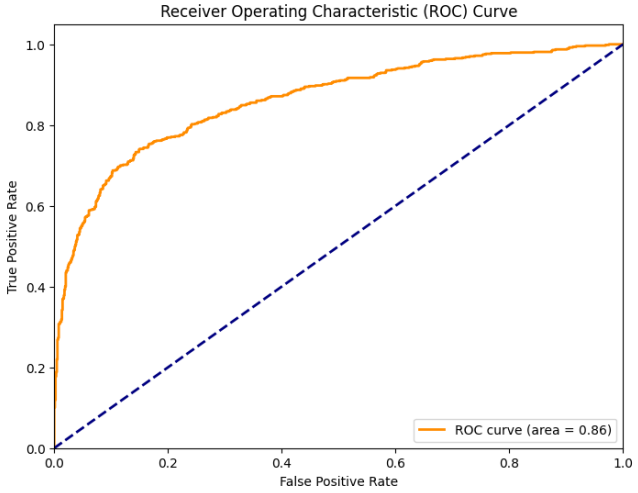
Top 20 Most Important Features:

Feature ranks top 20 from the text data in a graph along with transferred coefficients inferred with a logistic regression model are shown. Besides, these side factors act most importantly as a base of classification outcomes while their coefficients determine how important they are for target variable prediction. Every bar on the graph corresponds to a feature, with the length of its blocks signifying the absolute value of the coefficients. The longer verbalize suggest the elements are with a higher coefficient, thus a stronger classifier decision effect. On a contrary way, the thickness of bars stands for the bigger coefficients of the features that have more impact on the occurrence of the target variable. It is possible to divide these essential features of the model into groups by importance and to highlight the key terms or phrases that play the biggest role in the model's decision-making process. Consequently, the understanding of such linguistic cues enables us to contrast and discriminate the speech of different social classes with a richness in construction and form, that guides the process of analysis further.



Receiver Operating Characteristic (ROC) Curve:

ROC (Receiver Operating Characteristic) curve, presented in the picture below, will be employed for the purpose to illustrate how good classifier is doing in the current study which is logistic regression model. The curve shows the trajectory of the scenario, for different watchdog values: true positive rate (sensitivity) and false positive rate (1 - specificity). The orange curve provides the ROC curve. The sign of its shape is a simile depicting the success of the model in separating the positive and negative cases. A higher curve shows better discrimination ability while a straight line (which is cleverly shown as a let’s say dashed blue line) represents guessing randomly. The AUC of the ROC curve (AUC) characterizes how good a certain model is, the larger AUC corresponds to a better discriminating neural network. Here, AUC has the meaning which is combined with the graph label usage. Overall, the receiver operating characteristic curve with its underlying AUC represents a holistic measure of the model prediction power, which being beneficial in precising the threshold classification process for it.



Model Evaluation on Validation & Test Sets:

The evaluation results of the trained logistic regression model are summarized below: The short summary of model’s that has been trained is presented here:

Validation Accuracy: Specially over the validation period of **79.58%** got reported leading to the conviction that the model can use the data that is just generated.

Test Accuracy: The evaluation technique applied to the model proves that it provides **94.02%** accuracy set on the test which might be a demonstration of the ability to generalize the classifier well enough for a new dataset.

In conclusion, the mentioned algorithms reinforce the superiority of logistic regression in a well-crafted categorization of tweets into two classes; the disaster and the non-disaster. The models trained on the data set

shows little error on a test set this might mean that the model has grasped the subtle meaning of data and is trying to analogy form words the model hasn't seen in the training process.

Conclusion:

Here, we sailed through an exploratory study that looks at building a powerful text classification model, for the purpose of finding tweets related to disasters. Under data processing stage we carefully cleaned and standardized textual data of textual data to make it suitable for analysis. Utilizing techniques like lemmatization and removing of stop words, we improved our dataset which translated to better model performances.

Feature generation turned out to be that key factor in our undertaking given that we have exploited the TFIDF vectorization method to transform text into numerical. Therefore, this process made the semantic meaning of each tweet be in focus while coping with the contextual data targeted for classification. Following on is a step where we partition the dataset into the conducting of training and the evaluation of validation sets.

It was the discriminative power of a logistic regression algorithm that made the model evaluation successful, supplying promising outcome. An accuracy score of approximately 80.76% for our model on the validation set as evident was a confirmation of the suggested approach. The validation model was then deployed which calculated the Twitter classifications on the test dataset's samples, generating predictions for real-world utilization.

The actual integration of our model into the disaster response and crisis management to an assessment pipeline represents the primary goal of our study and opens opportunities to use the model to improve disaster response and address the consequences of crises and natural disasters. Through the automatic tagging of disaster-related tweets, decision-makers can immediately size up the level and spread of ongoing occurrences that sprung out. That way, quick and precise responses to the disaster are possible. Also, another positive aspect is that the product meets the needs of any current setup, which implies its wide availability and usability for different units.

What comes to an end in this journey, we can attempt to witness the change that data-driven methods do in society for all scales. Through the representation of machine learning and natural language processing methods we have been able to work out the major points of unstructured textual data, posing decision-makers with the opportunities for an informed look at the prospects of the complex scenarios they face. Our work here shows how data science can be used in building the ability to cope and taking different informed decisions during and after the disaster. Beyond this, data science could used in other fields e.g. economics and politics.

References:

This section should include : (1) papers mentioned in the related work section. (2) Describing algorithms(3) Code or libraries you downloaded and used. This includes libraries such as scikit-learn, Matlab toolboxes, Tensorflow, etc. Each reference entry must include the following (preferably in this order): author(s), title, conference/journal, publisher, year. Main body text, figures, and any discussions are strictly forbidden from this section.

[1]. G Neubig, Y Matsubayashi, M Hagiwara… - … Language Processing, 2011 - aclweb.org
Proceedings of the 5th International Joint Conference on Natural Language Processing

[2]. M Imran, S Elbassuoni, C Castillo, F Diaz, P Meier - Iscram, 2013 - idl.iscram.org the Web, especially during time-critical events such as "natural" and man-made disasters This plan to develop more sophisticated extractors that use complex Natural Language Processing(NLP) techniques in Social media: major tool in disaster response …

[3]. M Maldonado, D Alulema, D Morocho… - … conference on security …, 2016 - ieeexplore.ieee.org Natural Language Processing (NLP) is considered as a subarea of Artificial Intelligence, accordingto a storage and filtering criteria, being processed and analyzed with a tool NLP Twitter givegreater coverage to showbiz events than events related to Natural Disaster

[4]. Paul S. Earle, Daniel C. Bowden, and Michelle Guy. "Twitter earthquake detection: Earthquake monitoring in a social world". In: Annals of Geophysics 54.6 (2011), pp. 708–
715. ISSN: 15935213. DOI: 10.4401/ag-5364.

[5]. A Devaraj, D Murthy, A Dontula - International Journal of Disaster Risk …, 2020 - Elsevier
The build models using both traditional text features used in natural language processing (NLP), including n … ordering of words to make classification decisions, something that traditional NLP features like n … words, while useful for separating tweets relevant to the disaster from the social media
[1]