NLP for Social Good: Automated Detection of Disaster Tweets

Athiya Deviyani
Carnegie Mellon University
adeviyan@cs.cmu.edu

Abstract

The colossal scale of data generated by social media presents a unique opportunity for disaster analysis. With the availability of smartphones and widespread use of social media platforms, people can instantly announce events such as natural disasters in real-time. Because of this, more and more disaster relief agencies and governments are monitoring Twitter to track disaster events to gain more information and provide a swift response and recovery plan. This provides the opportunity to build an automated tool filter disaster tweets to be routed directly to disaster response teams and news agencies to raise awareness.

1 Goals

The main goal of this assignment is to build a language model using machine learning-based methods to identify tweets that give information about disasters. This will allow students to learn that the field of NLP is applicable to real life time-sensitive scenarios with the potential of saving lives. From this, students will reason about the appropriate metrics to be used as the response from a prediction may have disproportionate outcomes.

Additionally, unlike most news articles, tweets are often tainted with noise in the form of colloquial language, emojis, and sarcasm, which makes tasks such as disaster tweet prediction more challenging. Therefore, students will be required to carefully think about the appropriate preprocessing techniques for a dataset containing tweets. For example, formal corpora such as Wikipedia articles will contain subscripts, links, and possibly mathematical symbols, while Twitter data may contain attributes such as usernames, special tokens such as “RT”, and more. We would want the students to discuss the ethical implications of using a dataset from Twitter and how the appropriate preprocessing methods can help with preserving the privacy of the users, while also maintaining the relevant information within the tweet.

The main learning objective of this assignment is not only to obtain the best performing model in terms of accuracy with respect to the disaster tweet classification task. We want the students to be able to discuss why the chosen model is most appropriate for the task, as well as other pipeline design decisions and ethical considerations that they have made along the way, such as how they dealt with missing data or how they preprocessed and featurized the Tweets. Finally, we would like the students to explore the various challenges and ethical implications that may come with the deployment of an automated disaster tweet detection system.

2 Overview

The primary dataset that will be used is the “Natural Language Processing with Disaster Tweets” which comes from an ongoing Kaggle competition. The competition has over 3,000 entries and 800 submissions utilizing various NLP techniques to solve the binary classification task of identifying whether or not a tweet indicates a real disaster or not. The popularity of this dataset has also extended to academic publications such as Chanda (2021), Plakhtiy et al. (2020), and Gupta et al. (2022), where the authors attempt to use various state-of-the-art machine learning-based text classification algorithms to improve the existing disaster tweet detection models. In addition to its popularity, this dataset is part of the “Getting Started with Kaggle” collection, which means that this dataset contains real information which is presented in an intuitive way, such that anyone without a data science background would be able to use it without much hassle. This makes it appropriate for the Computational Ethics for NLP course, as students may come from non-technical backgrounds.

The `train.csv` and `test.csv` files contain 10,876 tweets annotated for the indication of a disaster-level event. For the purpose of the homework, we will split the training set further, where a random 20% will go to the validation set to be used by the students to test their model. Each sample in the train (`train.csv`), validation (`val.csv`), and test (`test.csv`) set has the following columns:

- TEXT: the tweet content
- KEYWORD: the keyword of the tweet (may be blank)
- LOCATION: the location where the tweet was sent from (may be blank)

The training and validation set has the additional TARGET column which denotes whether the tweet is about a real disaster (1) or not (0). Below are example tweets taken from the dataset:

> DISASTER: '#flood #disaster Heavy rain causes flash flooding of streets in Manitou, Colorado Springs areas'
> NON-DISASTER: 'I liked a @YouTube video http://t.co/**** Kalle Mattson - 'Avalanche' (Official Video)'

In this assignment, the students will first perform a preliminary analysis on the dataset containing tweets that may or may not indicate a disaster-level event. They can perform a quantitative analysis or visual analysis to help give them a "bigger picture" of the dataset. We do not expect complicated preprocessing methods to be done on the data, however we will be looking for justifications such as the username tokens being removed to protect the privacy of the usernames, the effect of keeping or removing emojis, methods of separating hashtags, and so on.

After preprocessing the data using the method of their choice, the students will be able to perform meaningful qualitative analysis by examining a random sample of tweets or by plotting a graph to show which words occur most frequently in disaster tweets. The students should evaluate the model based on accuracy, as well as the F1-score (where 1 or real disaster is considered the positive label). Based on the distribution of classes, the students should make a choice on which metric is most appropriate to use in this scenario, and what the final performance entails. A poor disaster tweet classification performance will yield a model which won't much use to disaster response teams and news agencies, and can be potentially misleading. Therefore, students should aim to have a model which achieves a good performance over the train and validation set. A good performance on the training set and a low performance on a validation set might indicate overfitting. At the end of the assignment, we would like the students to reflect and provide a discussion on their model choices, the outcome of deploying their final model and other ethical implications and design decisions that they have thought about during the development of the model.

3 Basic Requirements

3.1 Preliminary dataset analysis

The students should first do a preliminary analysis of the tweets in the given dataset. An example analysis would be finding words that occur most frequently in tweets labeled as 1 (containing real disaster). The contents of the tweet might contain noise in the form of punctuations and stop words, so we would like the students to perform the appropriate preprocessing to obtain a meaningful visual representation of the dataset. An example preprocessing method would be to lowercase the tweets, tokenize them, and remove stopwords and additional characters such as symbols or Twitter-specific tokens such as "RT" (which means ReTweet). We would like the students to include the diagrams or tables (if any) from their analysis in the final report. Make sure that they are properly labeled and clearly visible. We would also like them to provide a brief comment on any observations that they have found after performing the preliminary analysis on the dataset. The students could use metrics such as Pointwise Mutual Information (PMI) as described by Wikipedia (2022), to more appropriately measure which words are highly indicative of a disaster happening. From this, they will be able to make an observation and discuss which words are appropriately indicative of a disaster and which words are most misleading (words that they think are not disaster-related yet have a high PMI with the DISASTER label).

3.2 Build and run a basic classification model

Use a basic machine learning-based model available in the Python `scikit-learn` library to distinguish tweets which contain disaster information and which do not. The students should train their model on the TEXT column (and ignore the other columns) on `train.csv` using the labels available in the TARGET column, and report the accuracy and F1-score on the TEXT column on `val.csv`. They should provide a brief discussion on their results, such as by connecting trends observed in the predictions with the ones that they have observed in their prior analysis. They should obtain an accuracy of at least 60% and an F1-score of at least 70%. The students can also perform a comparison on various models and report their corresponding metrics. The students might find the Scikit Learn tutorial on building a text classification pipeline helpful. The students should then use their model to generate predictions on `test.csv` as and submit their predictions along with their final submission files.

4 Advanced Analysis

4.1 Error analysis

Great data scientists go beyond quantitative methods such as the accuracy and F1-score metric to evaluate the performance of their model. Students who choose to do detailed error analysis should sample misclassified tweets and try to observe a pattern between them. For example, the word "fire" might indicate that there is a fire ongoing in some location, however since Twitter contains a lot of colloquial text, the word "fire" can be used in contexts such as "your new music is fire!" which is not disaster-related. Other forms of analysis not mentioned in the basic analysis will also be accepted.

4.2 Sentiment analysis

The students can use an off-the-shelf sentiment analyzer on top of their disaster tweet classification model to aid their predictions. For further information, the students can refer to Ragini et al. (2018). They should report their classification results after using a sentiment analyzer in conjunction with their original classifier, and state whether or not adding the sentiment analyzer improves or worsen their accuracy and F1-score. They should provide a brief description of which sentiment analyzer they have used and a short discussion on why the results are better or worse.

4.3 Model improvement

In the basic requirements section, we have asked the students to use classifier models that are available in the Python `scikit-learn` library as well as build a basic text classification pipeline, hence there is plenty of room for experimentation and improvement. As a suggestion, they can use word embeddings to better represent the textual features of the tweet. The students should justify why they chose a specific word embedding. For example, the Glove-Twitter (Pennington et al., 2014) word embeddings might be more suitable for this task than Word2Vec (Mikolov et al., 2013) as the former is trained on tweets while the latter is trained on a more formal corpora. Other than that, the user can employ a more complex machine learning method such as using deep neural networks, similar to what Plakhtiy et al. (2020) has done. The students can use any of the relevant datasets listed on CrisisNLP (2019) to train the model if they believe it would improve the accuracy of the model on the validation and test sets, however the provided dataset should suffice. The students will be required to justify their chosen improvement method and their corresponding results.

5 Bonus Points

Like data science and machine learning, the NLP field has a large online community. Kaggle is a platform which allows users to find and publish datasets, explore and build models in a web-based data-science environment. The dataset that we have provided in this assignment is part of a Kaggle competition, where users all over the world work together and compete to obtain the best model that obtains the best F1-score in disaster tweet classification. For bonus points (an additional 5% in the total score) in this assignment, students can try and run their best model on `test.csv` and submit it to Kaggle. This will introduce and encourage students, particularly the ones who do not come from a technical background, to engage in a community filled with people with similar interests and explore the Kaggle platform.

6 Write up

The students should submit a report which should include the information described in the previous sections. Additionally, the report should include a brief discussion of the challenges they have encountered during the completion of the assignment as well as specific design decisions they have chosen, and the potential ethical implications of deploying an automated disaster tweet prediction system in a large-scale social media platform such as Twitter, such as a brief discussion on the cost of misclassifying a tweet as DISASTER and vice versa.

References

Ashis Kumar Chanda. 2021. Efficacy of bert embeddings on predicting disaster from twitter data. *arXiv preprint arXiv:2108.10698*.

CrisisNLP. 2019. Resources for research on crisis informatics topics.

Deepak Gupta, Nakul Narang, Mihir Sood, et al. 2022. Disaster tweets classification.

Kaggle. Natural language processing with disaster tweets.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Maryan Plakhtiy, Maria Ganzha, and Marcin Paprzycki. 2020. Comparing performance of classifiers applied to disaster detection in twitter tweets – preliminary considerations. In *Big Data Analytics*, pages 236–254, Cham. Springer International Publishing.

J Rexiline Ragini, PM Rubesh Anand, and Vidhyacharan Bhaskar. 2018. Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management*, 42:13–24.

Scikit Learn. Working with text data.

Wikipedia. 2022. Pointwise mutual information — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Pointwise%20mutual%20information&oldid=1073278253`. [Online; accessed 19-April-2022].