Disaster Tweets NLP Classification Benchmarks

Franklin Chen [1] **Andrew Chun** [1] **Nick Nguyen** [1] **Shivram Ramkumar** [1]

Abstract

Twitter has become an important communication channel in times of emergency. The ubiquity of smartphones enables people to readily announce an emergency they're observing and because of this, more agencies are interested in programatically monitoring Twitter (i.e. disaster relief organizations and news agencies). However, automation faces the problem of screening whether a person's words are actually announcing a disaster, or just talking about their own lives perhaps using certain vocabulary or sentence structure. To examine this situation, we used used a Kaggle competition disaster Tweet data set and performed a binary classification test and benchmarked four machine learning methods. Data and code are available at: https://github.com/chun-andrew/Disaster-Tweets-NLP-Classification

1. Introduction

As of the 2nd quarter of 2021, Twitter had 206 million daily active users worldwide (Statista 2021). With such a large active user base, there are millions of daily Twitter posts (Tweets) globally. These publicly available Tweets can often contain breaking information on natural disasters such as forest fires, earthquakes, or hurricanes. Due to constant user interaction with Twitter, monitoring Tweets can be extremely valuable for both news agencies and disaster relief organizations. For example, in 2010 during the earthquake in Haiti, the use of instant messaging via civilian phones was instrumental in finding and reporting trapped individuals. Coordination for medical assistance, distribution of relief resources, and other key communications were also done over such mediums. It is undeniable that civilian organization and communication alongside government or international disaster response now has a part to play in saving lives (Heinzelman, Waters 2010). However, as the multitude of Tweets made daily continues to grow, manual classification grows more and more expensive and infeasible.

For this paper, we examine the use of Tweets in disaster information collection. We will also examine if natural language processing is capable of automating with high degree of confidence and accuracy whether or not a Tweet is referencing an actual disaster, its scale, and methods of distributing aid and resources to areas or even individuals who need it most becomes easier. To demonstrate the applicability of various machine learning techniques, we use the prelabeled Tweet data set from Kaggle's GettingStarted Prediction Competition on this precise topic. (Kaggle 2020) We will attempt binary classification on whether a Tweet is affiliated with a disaster or not using four different machine learning methods: logistic regression, k-nearest neighbors, convolutional neural networks (CNN), and long short-term memory networks (LSTM).

2. Related Work

In a 2016 paper, Tyler Robinson of KSU employed a word vector and part of speech tagging approach to disaster Tweet classification (Robinson). In his experiment, he used a disaster Tweet data set that went through part of speech processing, meaning each word had a part-of-speech tag such as noun, verb, or adjective assigned to it. He then applied Naive Bayes supervised learning to multiple sets of data. These data sets include word vectors alone, part-of-speech tags alone, word vectors + part-of-speech tags, and word vector bigrams + part-of-speech bigrams (bigrams being word or part-of-speech tag pairs). In his experiment, he found that the word vector dataset trained algorithms worked most effectively, with word vector + part of speech datasets being on average 2% less effective, and pure part-of-speech training to perform the worst.

In an ISCRAM 2015 conference paper, a group of researchers found that machine learning relies too heavily on labeled data to learn accurate classifiers to be effective in different, new disasters for which there was no labeled data. (Li 2015) They proposed using domain adaptation and training Naive Bayes classifiers on a dataset that contained only source labelled data and unlabelled target data (under the assumption that for an emergent disaster there is no available labeled data). Domain adaption is the process of adapting different models across domains - for example, using a model trained to detect fires to detect earthquakes. The data was represented through a bag-of-words representation that consists of vectorizing the words in an 0/1 representation. Data was sourced from Twitter, hand-annotated, and belonged to either a Hurricane Sandy or Boston Marathon

Bombing collection. The authors then trained and compared models based on target unlabeled data using domain adaptation and supervised sources. The results showed that unlabeled data was more effective for tasks related to a specific disaster as opposed to tasks similar between disasters as the data could act as noise during training. They suggested that domain adaption classifiers that use some data from an specific yet unlabeled disaster seemed to be superior to classifiers learned only from source data except in similar disaster related tasks. In conclusion, they called for more research to be done on larger datasets along with more classifiers such as SVM or random forests before being able to draw a general conclusion (Pennington 2015).

3. Methodologies

3.1. Data Preprocessing and Cleaning

Since Tweets often contain links, hashtags, and Retweets of other users (via the @ symbol), the data was processed to remove special characters, Unicode symbols like emojis, and entire Retweet usernames. This prevents parameter weights from learning to associate such noise with actual parameters. The dataset used for this paper was organized as a 5 column feature matrix, with columns corresponding to ID, keyword, location, text, and target (see Table 1). As part of preprocessing, the ID, keyword, and location were stripped from each Tweet since those often do not and did not have those fields (ID being the index in the dataset).

3.2. Feature selection

The baseline standard was a linear model consisting of binary logistic regression. The feature selection for the linear model and the k-nearest-neighbors entailed the use of Sci-kit Learn's TF-IDF Vectorizer; (Pedregosa 2011) because the dataset is presented as a collection of Tweets, vectorization of entire Tweets resulted in more meaningful data vectors and lower dimensionality as opposed to one-hot encoding of tokenized words and joining individual word vectors to form sentence vectors. In particular, the TF-IDF is a terms weighting algorithm that links the term frequency and inverse document frequency for a particular word among a corpus of text. In any larger amount of text, some "stop" words are heavily present (such as "the", "a", "is" in English) and carries little contextual meaning. The counts of these stopwords would overshadow the less frequent yet more important words. The corresponding outputs are calculated using a fit transform method that calculates vectors for each Tweet given the entire training data and the respective classifications.

3.3. Model baselines

Logistic regression has been chosen as the baseline linear model to compare the relative performance of different machine learning models owing to its simplicity, minimal training time, and the word-vector featurization. Such baselines are especially significant in considering the context of the algorithms' applications - disaster recognition from real time Tweets focuses on ease of training updates and processing, and simpler yet well-performing models are consequently preferable. The model relies on the TF-IDF vectors of Tweets as features and uses those to train a Sci-kit learn logistic regression model with a SAGA solver, a variant of stochastic average gradient.

The k-nearest neighbors algorithm has been chosen as a baseline non-linear model to compare against the performance of logistic regression and the neural networks. The model also relies on TF-IDF vectors of Tweets as features and uses a Sci-kit learn K nearest neighbors model with a neighbor consideration parameter of 5 (K = 5).

3.4. Neural networks

A convolutional neural network (CNN) model was chosen for investigation. In our model, N-grams of kernel sizes 1 - 5 were applied on the Tweets and then passed through a convolutional and a max pooling layer. The output from these layers is subsequently passed into a flattened linear layer. This method was chosen particularly to investigate whether using n-grams to embed words would produce better results over the baseline models in classifying disaster Tweets.

Long short-term memory (LSTM) networks are a form of recurrent neural networks that are able to keep track of long term dependencies in the input. This makes them particularly well suited for applications such as time-series predictions and natural language processing owing to the semantic nature of grammar and word ordering. The model was chosen to examine whether a relatively small network (with fewer neurons) would be able to derive substantial benefits over the baseline models in predicting the appropriate classification. The model was implemented in TensorFlow's Keras using word tokenization and padded sequences, and a sequential model was constructed, consisting of an LSTM layer followed by Dense and Dropout layers with a ReLU activation. The number of neurons within each layer was kept minimal without losing performance since the actual medium for text, Tweets, are significantly constrained in their complexity and width due to a character limit.

4. Dataset Results

Models were created using K-nearest neighbors (KNN), long short-term memory (LSTM), logistic regression, and

Classification	Keyword	Location	Text
False	accident	Anime World	@sakuma_en If you pretend to feel a certain way the feeling can become genuine all by accident. -Hei (Darker than Black) manga anime
True	ablaze	”GREENSBORO, NORTH CAROLINA”	How the West was burned: Thousands of wildfires ablaze in California alone http://t.co/vl5TBR3wbr
True	ablaze	”Concord, CA”	@Navista7 Steve these fires out here are something else! California is a tinderbox - and this clown was setting my ’hood ablaze @News24680
False	aftershock	304	’Remembering that you are going to die is the best way I know to avoid the trap of thinking you have something to lose.’ ÛÒ Steve Jobs

Table 1. Sample of data entries to show format of our initial data from Kaggle

convolutional neural networks (CNN). From these models, confusion matrices were generated and are shown below. The best accuracy and performance results logistic regression with the greatest accuracy and best distribution of the confusion matrix as displayed in Figure 3. It is closely followed by LSTM in terms of accuracy, then KNN, and lastly, convolutional neural network. The strong prediction of true negatives is consistent regardless of the network type chosen, as observed from the results of all four models. But there is some variability between the skew toward false positives and false negatives - both the logistic regression and the LSTM models have greater false positives than false negatives, whereas the CNN and KNN algorithms have a greater number of false negatives. The variability of the linear and KNN baselines suggest that the differences are an effect of the models and the model parameters themselves, and not necessarily a deeper correlation with the data. One may be able to assume that the semantic ordering for disaster Tweets have less variability than the broader category of non-disasters, which might indicate the LSTM’s skew.

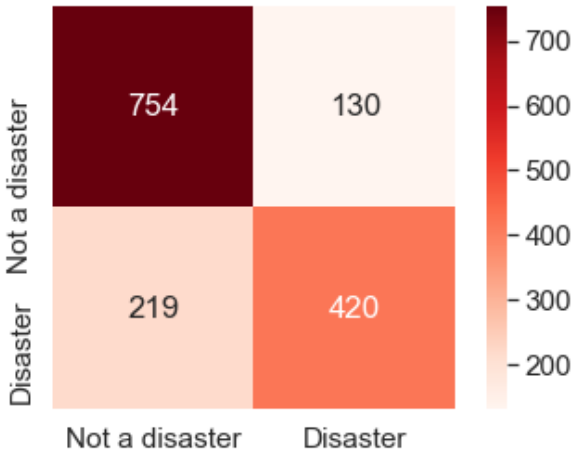

Figure 1. The model for KNN has correct classification 77.08% of the time.

5. Limitations

5.1. Problems with Twitter

There are many disadvantages inherent to using Tweets, and the format of a Tweet in attempting to classify disaster related Tweets. A Tweet often does not have grammatical structure or linguistic logic; they often are not in a logical sentence format which makes them difficult to interpret. Our models were unable to place special emphasis on words or substrings associated with hashtags or tags which often convey additional meaning which is then left unaccounted for. It is also difficult for NLP to ascertain whether or not a Tweet is talking about an actual disaster, or simply using metaphors or other euphemistic language. Very short Tweets are vulnerable to simply not providing enough data for classification, this problem is further compounded by very short Tweets having a much wider variety of applicable contexts. In these ways, the features we have chosen in order to represent the Tweets may not encompass the total meaning of the Tweets as well.

It is also notable that these models were trained on data which only took into account Tweets in English. Further verification is required in order to ascertain whether similar results can be generated for disaster Tweets in other languages.

Furthermore, disinformation and misinformation can be hard to account for. This problem can affect our results as these types of Tweets can and do seem like valid disaster Tweets. This problem can be especially prevalent on social media, where there is little to no information regulation.

5.2. Problems with Binary Classification

While simple binary classification is useful in determining whether Tweets are associated with disasters or not, we have no way of determining the relative urgency and relevancy

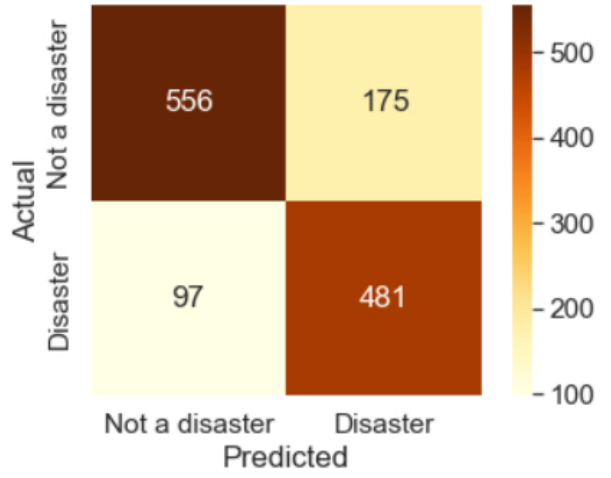

Figure 2. The model for LSTM has correct classification 79.22% of the time.

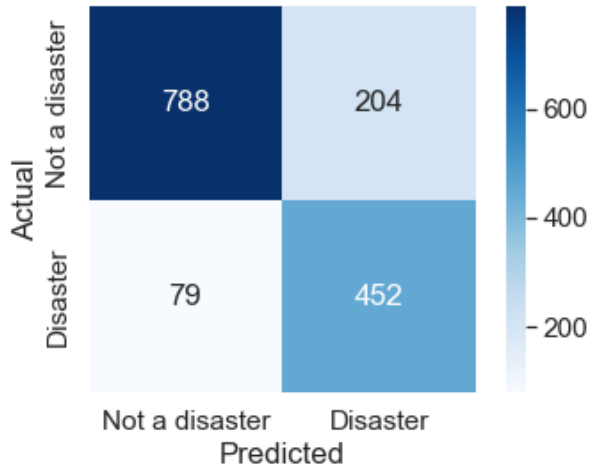

Figure 3. The model for Logistic Regression has correct classification 81.42% of the time on the test set

of Tweets. For instance some Tweets may be made by people giving live reporting on the ground which is useful, but ultimately not necessarily as urgent as reports of individuals trapped by debris or requests for emergency equipment. If further machine learning efforts would like to assign an evaluation for urgency then a departure from binary classification would likely be necessary to do so.

5.3. Discrepancy in the CNN model

With respect to our models in particular, it is also important to note here that our model for convolutional neural

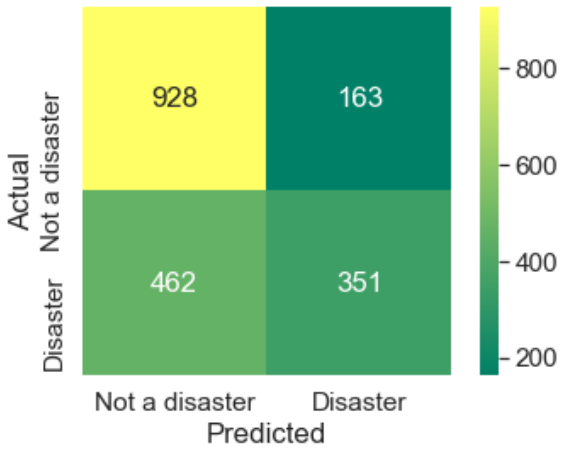

Figure 4. The model for CNN has correct classification 67.17% of the time

networks performed noticeably worse than the others. We attribute this primarily to the fact that the structure of convolutional neural networks was designed in order to facilitate the problem of image processing. This particularity is baked into the structure so to speak, and therefore most likely makes the model less generalizable to the problem of NLP. It may be possible to account for and adjust this error by additional manual hyper-parameter tuning. However, as Hosseini notes in one of our references, because disasters require real time feedback often with a high degree of urgency, there will not be time for manual hyper-parameter tuning in real time deployments of these models. Thus, we can, to an extent, discard the possibility of improved performance via manual tuning for this model. (Hosseini 2020)

5.4. Related Accuracy of other Models

A simple analysis of the most common words found in the Tweets categorized as disaster related yields words like ”fire, California, people, News, killed” etc. The most common words found in the Tweets categorized as non-disaster related are more mundane: ”like, Im, amp, one” etc. From this eye-scan alone we can tell that simply detecting a specific subset of disaster related words, which even can be tailored to the context of whatever specific disaster is being tracked in question, is probably sufficient for these models to make a prediction. This is perhaps the explanatory power behind the efficacy of binary classification for the other three models. The contexts of the words is less important for these models than their simple presence, and so our predictions ignore a significant part of the meaning in the Tweets. A more comprehensive analysis of these Tweets would hopefully attempt to address this reduction and reincorporate some of

the lost information here.

5.5. Model Abuse

Unfortunately, if we accept that the binary classification for these Tweets is mostly dependant on the presence of specific disaster related words, then these models are extremely susceptible to being abused and manipulated by malicious Tweets. Scammers or individuals simply seeking to amplify chaos by generating false reports would easily be able to provide false positives or true negatives to the model and disrupt the diagnostic application of any results. While the number of individuals who would engage in such behavior is likely very low, it is still important to note that these models are very easily susceptible to manipulation by adversaries.

6. Conclusion

Even with these relatively simple models and naive feature interpretations our confusion matrices correctly classified almost 80% of the data. The short disjointed nature of Tweets poses a problem in classification and feature processing, but using signal word frequencies is relatively successful. This is a promising result for the base ability for machine learning techniques to perform well in this problem. As discussed extensively in the limitations, there are various avenues of improvement ranging from moving beyond binary classification to the introduction of features or model considerations in an attempt to process more than just word frequencies which seem to make up most of the predictive power on the presented models.

7. Citations and References

Heinzelman, J., Waters, C. (2010). Crowdsourcing Crisis Information in Disaster-Affected Haiti. US Institute of Peace. http://www.jstor.org/stable/resrep12220

Hosseini, S. (2020, June 28). Binary classification of Disaster Tweets. Medium. Retrieved December 8, 2021, from https://towardsdatascience.com/binary-classification-of-disaster-tweets-73efc6744712.

Kaggle. (2020, March 24). Natural language processing with disaster Tweets. Kaggle. Retrieved December 8, 2021, from https://www.kaggle.com/c/nlp-getting-started/overview.

Li, H., Guevara, N., Herndon, N., Caragea, D., Neppalli, K., Caragea, C., Squicciarini, A., amp; Tapia, A. H. (2015, May 27). Twitter Mining for Disaster Response: A Domain Adaptation Approach. Kansas State University. Retrieved December 8, 2021, from https://www.researchgate.net/publication/348616136_Deploying_Spatial_Data_for_Coastal_Community_Resilience_A_Review_from_the_Managerial_Perspective.

Pennington, J., Socher, R., amp; Manning, C. D. (2015). GloVe: Global Vectors for Word Representation. Glove: Global vectors for word representation. Retrieved December 8, 2021, from https://nlp.stanford.edu/projects/glove/.

Robinson, T. (2016). Disaster Tweet classification using parts-of-speech tags: A domain adaptation approach. K. Retrieved December 8, 2021, from https://krex.k-state.edu/dspace/handle/2097/34531.

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Statista. (2021, November 19). Twitter: Most users by country. Statista. Retrieved December 8, 2021, from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countr