Contributions on Individual Report

The “Natural Language Processing with Disaster Tweets” dataset sourced from Kaggle consists of 10873 tweets with three features (not including “id”) and the target variable [0 for unrelated disaster tweet, 1 for disaster tweet]. We chose to do this text classification problem due to the increase in misinformation on today’s social media platforms. The tweets from the dataset could either be related to the disaster, consisting of natural disasters (like earthquakes, hurricanes, or wildfires), extreme weather events (like heat waves and storms), or man-made disasters (like oil spills and transportation accidents). Or, the tweets could be completely unrelated to disasters, which would be classified as fake disasters. The three features of our dataset include ‘text’ (the body of the tweet), ‘location’ (where the tweet was sent from), and ‘keyword’ (the most notable word of the tweet relating to a disaster).

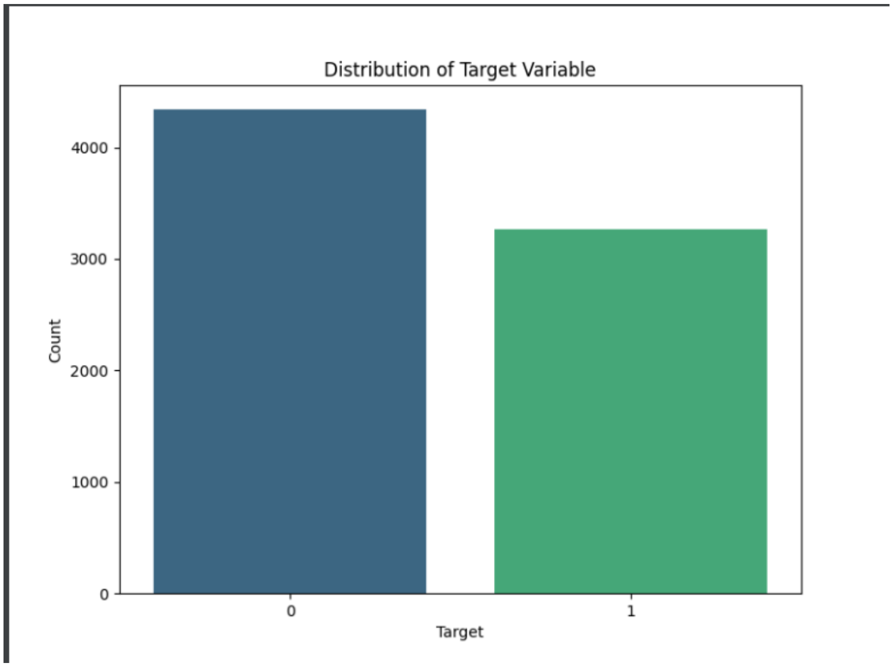


Figure 1: Distribution of Target Classes in the Dataset

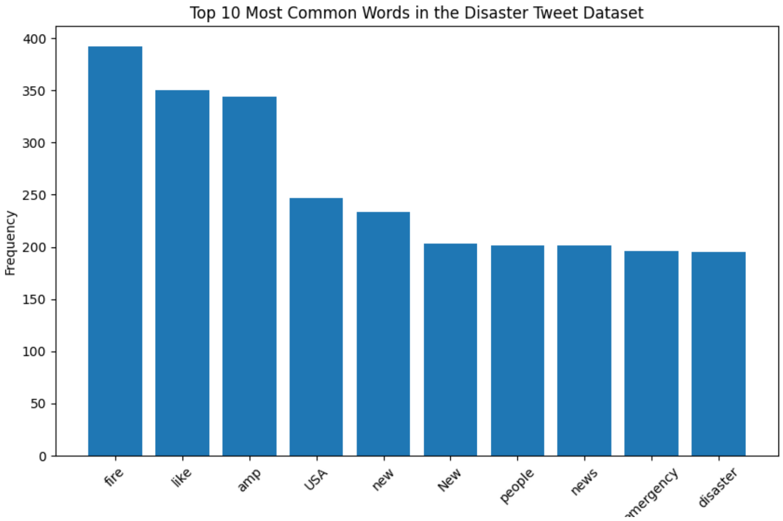
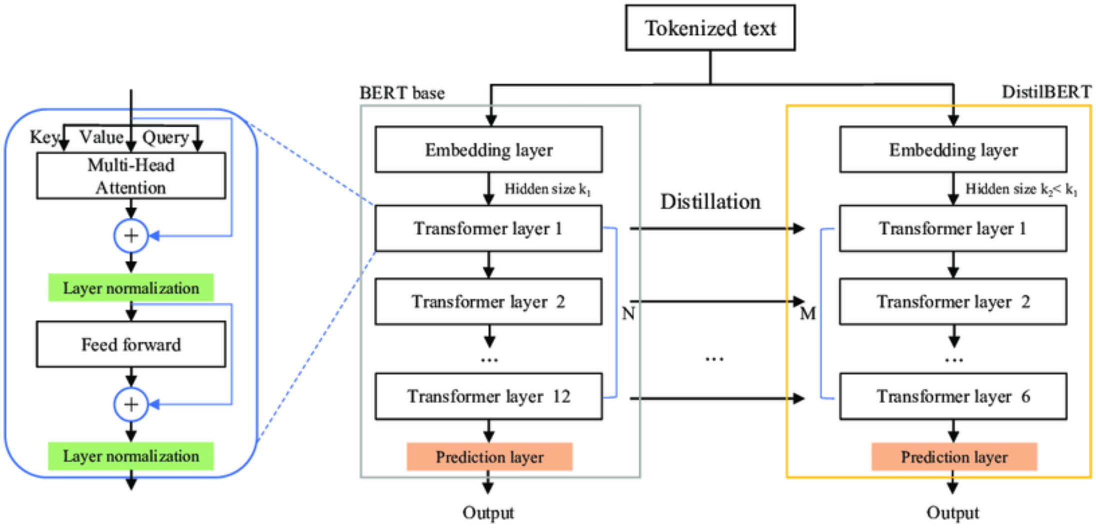


Figure 2: Most common words in the Dataset

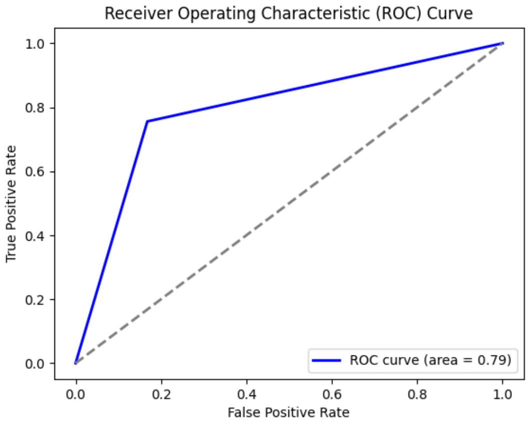
As depicted in Figure 1, our dataset exhibits a balanced distribution, with 4342 instances for target 0 and 3271 instances for target 1. From Figure 2, we can see that the most common words in the dataset often relate to emergencies or disasters that would be of interest to the target classes in the tweets. “Fire”, “disaster”, “emergency”, and “news” were often in the dataset, and the dataset was also centered around regions in the USA.

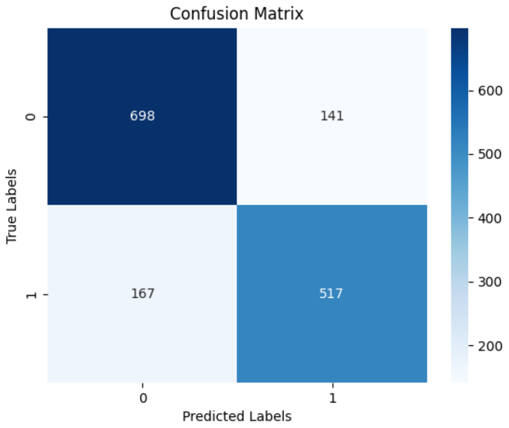
DistilBERT:

DistilBERT is a transformer-based model created from knowledge distillation, trained to learn the logits of the original BERT model. For example, while BERT’s distribution of a multiclass classification problem could have a softmax distribution of .6, .2, and .2 for three different classes, the training goal of DistilBERT is to score a .6, .2, and .2 for the same data observation. In doing so, the hope is that the knowledge distillation will allow DistilBERT to track the same context and patterns that the original and larger BERT model used to create the scores in the first place. DistilBERT has 40% less parameters than the original BERT model, trains 60% quicker, and scores 95% or above of what the original BERT model’s scores on language capability benchmarks. The transformer layers of DistilBERT are half of the transformer layers of the original BERT base (as you can see from the figure below), and in the paper “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter” by Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, it was noted that it’s language capabilities would often outperform an ELMo encoder model with two bidirectional LSTMs.



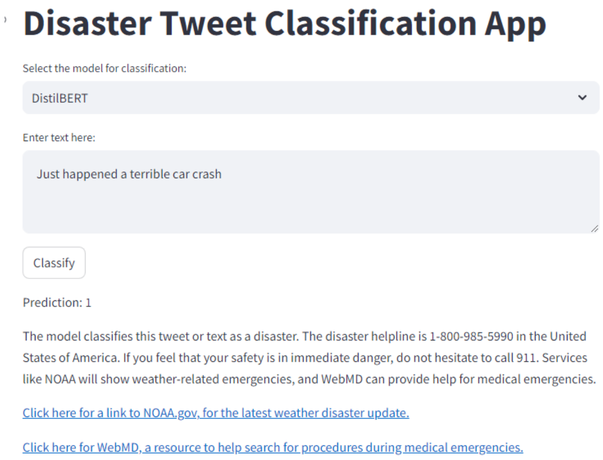
From the ROC curve and confusion matrix below, we can see that, after training for 5 epochs, the AUC was .79, and the confusion matrix is fairly well distributed across true positives and true negatives, along with false positives and false negatives. The f1 score reached 0.797, and the accuracy was .798. Due to the distribution of the confusion matrix, we can see that the model is not heavily leaning towards classifying true negatives nor true positives, and although the false positives were more than the false negatives, this was also due to the original distribution of the dataset having more class 0s than class 1s.





Implementation and Experimental Setup

Our procedure to create, compare, and test our models was to train through a few epochs before comparing our model results. We kept the preprocessing (other than specific tokenizers or embedding layer setup) the same for each of our models.



Disaster Tweet Classification App

Select the model for classification: DistilBERT

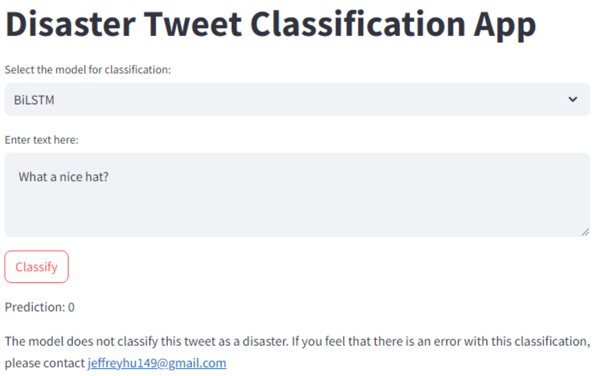
Enter text here: Just happened a terrible car crash

Prediction: 1

The model classifies this tweet or text as a disaster. The disaster helpline is 1-800-985-5990 in the United States of America. If you feel that your safety is in immediate danger, do not hesitate to call 911. Services like NOAA will show weather-related emergencies, and WebMD can provide help for medical emergencies.

Click here for a link to NOAA.gov, for the latest weather disaster update.

Click here for WebMD, a resource to help search for procedures during medical emergencies.



Disaster Tweet Classification App

Select the model for classification: BiLSTM

Enter text here: What a nice hat?

Prediction: 0

The model does not classify this tweet as a disaster. If you feel that there is an error with this classification, please contact jeffreyhu149@gmail.com

Summary and Conclusions

From our project, we can show that an affordable and light gpu model can be constructed to tackle this fake and real disaster tweet problem. We did run into a few problems with preprocessing, for example, due to filtering out stopwords using Spacy, past and present tense were ambiguous to the model in some situations, so past disasters and ongoing disasters were treated the same (I was in a car accident vs. I am in a car accident).

In the future, we plan to pay closer attention to the specific words/show embeddings and associations of the disaster and do additional data analysis on the focus/robustness of our models.

For future exploration, we could automate disaster classification on hashtagged tweets to filter out false or mistakenly tagged tweets from interfering with the information provided by the true disaster tweets and updates. We could also further classify positively predicted tweets between which type of disaster it is (fire, hurricane, man-made disasters, etc.), and give referral links/phone numbers as to which resource is best suited to the disaster.