Predicting Natural Disaster Tweets Using Binary Classification

Tammy Dahl
Tommy Kieu
Kyle Krick
Roger Reinhardt
Jordan Spencer

Problem statement and motivation

We will implement and evaluate binary classification machine learning (ML) models to predict whether a given tweet describes a natural disaster. We chose this project because it allows us to apply skills that we have already learned from class assignments to a real-world problem while also requiring us to learn techniques concerning natural language processing (NLP). ML models like the one that we are proposing to design could potentially be incorporated into existing efforts to detect and respond to natural disasters by collecting and analyzing public social media posts. This could be especially useful in California wildfire emergency responses. Social media apps have access to rich data such as user-tagged locations, which could be helpful in establishing the locality of a natural disaster. Furthermore, given the recent breakthroughs in NLP, namely ChatGPT and other large language models (LLM), developing our skills in the NLP space will prepare us for future work using and improving these types of ML models. This is very likely to be a significant component of our careers, as suggested by San Diego State University's 2024 AI Student Survey, where over 70% of students agree that generative artificial intelligence will become an essential part of most professions [1].

Dataset description

Our dataset will come from Kaggle using the Natural Language Processing with Disaster Tweets Which provides 10,0000 labeled tweets. Each tweet is classified as either related to a disaster or not. For example, if there is a disaster, predict a "1;" otherwise, predict a "0." The dataset includes text data along with metadata, such as the locations and keyword features. We will preprocess the dataset by removing stopwords, performing tokenization, and applying other techniques like word embeddings to convert text into numerical representations.

Planned methodology

Our project applies binary classification models to determine whether a tweet describes a real natural disaster. After collecting and formatting the dataset, we will clean the text by removing stop words, hashtags, mentions and special characters, and convert all words to lowercase to standardize the input.

We will transform the processed tweets into numerical features using TF-IDF. For modeling, we will focus on Logistic Regression and Support Vector Machines (SVM), both of which are well-suited for high-dimensional text data. These models will be trained and evaluated using a split between training, validation, and test sets.

Time permitting, we will consider adding sentiment analysis scores, tweet length, or keyword presence as features. Also, if feasible, we will experiment with pre-trained word embeddings and contextual embeddings to determine if deep learning improves performance over traditional models.

Evaluation metrics

To assess model performance, we will rely on accuracy, precision, and recall. Should the dataset contain considerably fewer disaster tweets than non-disaster tweets, we will consider implementing oversampling or undersampling to balance the dataset. Cross-validation will be used during development to tune model parameters and help reduce overfitting. The goal is to determine which of the implemented models performs best for classifying disaster-related tweets.

We will evaluate the performance of our binary classification models by checking the F1 score: $F1 = 2 * ((Pr * Re) / (Pr + Re))$, where Pr = precision = $TP / (TP + FP)$, Re = recall = $TP / (TP + FN)$, TP = true positive, FP = false positive, and FN = false negative. We will also use the area under the receiver operating characteristic curve (AUC-ROC) to visualize the models' performance.

Beyond standard evaluation metrics, it is worthwhile to analyze misclassified tweets to improve the model. For example, we should be able to identify whether the tweet uses sarcasm or metaphorical language.

Expected outcomes

We expect the SVM and Logistic Regression models to perform well, acknowledging their effectiveness in text classification. Aiming for an F1 score above 0.75 will ensure a good balance between precision and recall. If time permits, we may explore additional NLP techniques, such as deep learning models, to improve classification performance.

The project will result in a trained model capable of predicting whether a given tweet is about a real disaster, which could potentially be used in real-world disaster response applications.

Work distribution among group members

Data Collection and Processing: Team Lead: Tommy Kieu
Week 1 Responsibilities:

- Find and clean a dataset of disaster-related tweets
- Manage text cleaning, remove stop words, and preprocess for model training
- Format data into a usable structure

Model Development: Team Lead: Kyle Krick
Week 2 Responsibilities:

- Choose and implement a binary classification model
- Train and validate the model, fine-tuning for accuracy

Natural Language Processing: Team Lead: Jordan Spencer
Week 3 Responsibilities:

- Convert raw text into numerical representations
- Experiment with various NLP techniques to improve classification accuracy

Evaluation and Performance Analysis: Team Lead: Roger Reinhardt
Week 4 Responsibilities:

- Compare different ML models using evaluation metrics
- Identify limitations and suggest improvements

Software Developer: Team Lead: Tammy Dahl
Week 5 Responsibilities:

- Create an app to showcase the model's predictions
- Automate certain parts of the workflow

Expected challenges

We realize that we may run into computational limitations and plan to utilize the team members who have high-end computers at home and Kaggle's free GPUs and ensure that we optimize our model parameters to be able to train our model effectively.

Due to limited experience using NLP tools, we will likely incorporate pre-trained sentiment analysis tools to handle the ambiguity in language within the tweets of our dataset.

References

[1] J. Frazee, D. Goldberg, S. Hauze, and E. Sobo, "Academic Applications of AI: Using Student Survey Data to Develop a Data-Informed Strategy," presented at the EDUCAUSE Annual Conference 2024, San Antonio, TX, USA, Oct. 2024. [Online]. Available: https://events.educause.edu/annual-conference/2024/agenda/academic-applications-of-ai-using-student-survey-data-to-develop-a-datainformed-strategy