Identify Disaster Events from Twitter Stream

Tarun Singh Bondili
tarunsingh@vt.edu
Virginia Tech

Abstract

In the modern era, social media platforms like Twitter have begun to play a vital role in real-time disaster reporting. Rapidly detecting disaster events from Twitter posts enables faster response by emergency first responders. However, automatically classifying disaster-related tweets from irrelevant chatter poses a challenging natural language processing problem. In this deep learning project, we will develop models to accurately identify disaster messages from Twitter streams. We will extract features from the text data using techniques such as bag-of-words, TF-IDF vectors and count vectorization. These feature representations will be used to train classifiers, including ridge regressions, Naive Bayes, SVM, BERT-based CNN model and LSTM Networks. Additionally, we will experiment with pre-trained word embeddings like GloVe and advanced language models such as BERT. We will compare our approaches to state-of-the-art AI systems including GPT-4.

Keywords: CNN, Naive Bayes, SVM, LSTM, datasets

ACM Reference Format:
Tarun Singh Bondili. 2024. Identify Disaster Events from Twitter Stream. In . , 4 pages.

1 INTRODUCTION

Social media has transformed how people communicate adn share information during emergency events. Platforms like Twitter enable eyewitnesses to broadcast disaster events in real-time, providing vital on-the-ground reporting as events unfold. During natural disasters like hurricanes, floods and wildfires, people increasingly use Twitter to post updates, call for help, and coordinate relief efforts.

However, mining actionable information from Twitter's massive volumes of posts poses a major challenge. The vast majority of tweets have no relation to disasters. Automatically distinguishing disaster-related tweets from irrelevant chatter is a non-trivial but impactful application of natual langauge processing.

Developing models to detect disaster tweets accurately would allow first responder organizations to monitor Twitter to identify emergencies faster. Detecting relevant disaster tweets could also provide vital insights to first responders that could aid in their efforts. News agencies could also use these techniques to find eyewitness accounts and breaking updates to cover. More lives could be saved and suffering alleviated if disasters are detected early using social media data. We aim to build disaster detection models to enable these use cases. And so, effective disaster tweet identification remains an open research problem. Performance gains in this area could directly improve emergency response.

2 CONTRIBUTIONS

The current work makes several contributions to the task at hand for disaster event detection.

We have developed a multi-model classification system that leverages both traditional and modern approaches. Implementation of an efficient pre-processing pipeline specifically for the type of data chosen. The application of few-shot learning using GPT-3.5 for rapid disaster detection with minimal training data.

3 RELATED WORK

The problem of classifying disaster-related tweets is essentially a text classification challenge that may require sentiment analysis. Alam et al. [2] analyzed Twitter data from major hurricanes to provide insights into effective machine learning techniques for disasters. Imran et al. [4] released an early large crisis tweet dataset and corresponding event-specific word2vec embeddings.

Relevant work includes research by [8] on using NLP to aid relief efforts in the 2011 Japan earthquake. They developed systems for word segmentation, NER, and tweet classification, using an SVM classifier with 10-fold cross-validation, achieving 86.13% accuracy on disaster tweet classification. Other relevant text classification tasks include identifying spam tweets. Research by [3] used an SVM for tweet spam detection, achieving 95-97% accuracy. Sarcasm detection in tweets [7] has been approached with Naive Bayes, also giving good performance. These studies suggest traditional machine learning approaches like SVMs and Naive Bayes are simple yet effective for tweet classification problems similar to disaster detection. Deep learning methods will potentially improve upon these baseline models, The latest pre-trained transformer models have shown promise for many text classification tasks as well, warranting exploration for disaster tweet classification. State-of-the-art approaches apply transfer learning to large pretrained language models such as BERT, but little research specifically evaluates such models [5,9]. Spiliopoulou et al. [6] developed an adversarial BERT

, ,
© 2024 .
ACM ISBN

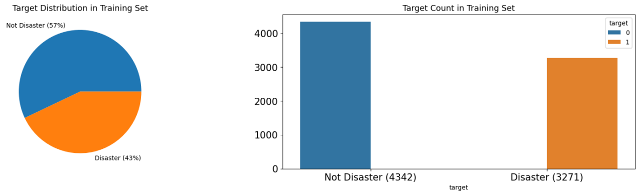

Figure 1. Data distribution

model to reduce event-specific bias for social media text classification related to disasters. Their model improved critical tweet identification, demonstrating the value of transfer learning.

Recent advances in generative pretrained language models like Flan-T5 Suite, GPT-3.5, GPT-4 offer new possibilities for few-shot text classification. By leveraging vast pretrained knowledge, models fine-tuned on just dozens of examples can match or exceed the performance of state-of-the-art discriminative models trained on thousands of labeled instances. However, few studies have evaluated the effectiveness of this approach for disaster-related tweet classification specifically. Our work conducts controlled experiments to analyze the few-shot learning capabilities of models like GPT-3.5 on disaster tweet dataset from Kaggle [1].

4 PROPOSED WORK

4.1 Data Source & Description

We will use the dataset from Kaggle, a well-known platform for machine learning competitions and datasets. Specifically, we obtained the dataset from Kaggle competition titled "NLP Getting Started", available at the following url: https://www.kaggle.com/competitions/nlp-getting-started/data. This data set is designed for a classification task, where the primary objective is to predict whether a given tweet is related to disaster even or not. The dataset contains 5 columns out of which the "text" column contains the tweet data and "target" column a label that the given tweet as a disaster-related or not.

4.2 Preprocessing Steps

Before applying any models to the data, a series of preprocessing steps are essential to ensure the data quality and compatibility with the chosen models. THe preprocessing steps include:

4.2.1 Data Cleaning.

- Removal of special characters, punctuation, and unnecessary white spaces.
- Text normalization, including lowercasing all text to maintain consistency.

4.2.2 Tokenization.

- Tokenization of tweet text into individual workds or subword tokens, depending on model requirements.

4.2.3 Handling Missing Values.

- Identification and handling of missing or null values in the dataset, ensuring that the data is complete and ready for modeling.

4.3 Methodology

We aim to conduct a comprehensive comparison of various machine learning models to tackle the tweet classification task. The intended strategy of work can be summarized as:

4.3.1 Traditional ML Classifiers.

- We will implement and evaluate the performance of traditional ML classifiers including Logistic Regressions, Naive Bayes and Support Vector Machines.
- Hyperparameter tuning will be conducted to optimize the performance of these classifiers.

4.3.2 BERT Fine-Tuning.

- We plan to fine-time a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model, a state-of-the-art deep learning model for natural language understanding.
- Various fine-tuning strategies and hyper-parameters will be explored to enhance the model's performance.

4.3.3 GPT Models.

- We will utilize GPT-3.5 and GPT-4, both powerful language models developed by OpenAI for the classification.
- We sample small training sets of 16,32, and 64 examples per class to fine-tune the generative models and systematically compare against the conventional ML and neural network baselines requiring full dataset training.

The primary goal of our modelling efforts is to optimize the classification accuracy of the models in determining whether a given tweet pertains to a disaster event or not. We will measure our model's success in achieving this through various evaluation metrics, discussed in the subsequent sections. These results also provide insights into whether modern generative AI can achieve competitive disaster tweet classification with less labelled data than existing methods.

5 EVALUATION

Since it is a classification problem, we would be considering the following evaluation metrics:

5.1 Available Metrics

- **Accuracy:** it is a metrics that measures how often a machine learning model correctly predicts the outcome. You can calculate accuracy by dividing the number of correct predictions by total number of predictions.

- **Precision:** It is a mertic that measures how often a model correctly predicts the positive class. You can calculate precision by dividing rue positives by the total number of instances the model predicted as positive (both true and false positives).
- **Recall:** It is a metric that measures how often a machine learning model correctly identifies true positives from all the actual positive samples in the dataset. You can calculate recall by dividing the number of true positives by the number of positive instances.
- **F1 Score:** The F1 score combines precision and recall using their harmonic mean, and maximizing the F1 score implies simultaneously maximizing both precision and recall.

5.2 Choosing the Evaluation Metric

Accuracy is not good for imbalanced datasets like Twitter, which receives millions of tweets daily, but only a few are positive. We believe Recall and F1 Score are excellent choices for this use case because they are good for imbalanced datasets. We want to have very few false Negatives in these cases. In case of a disaster, we do not want to miss classifying it.

- Precision = $\frac{TP}{TP+FP}$

- Recall = $\frac{TP}{TP+FN}$

- F1 = 2 $* \frac{precision*recall}{precision+recall}$

6 EXPERIMENTS

To handle the class imbalance, a stratified split of the original data was adopted in a 8:2 ratio. The major split of 80% serves as the training set, while the minor split is used as the test dataset. For GPT-3.5 turbo, we employed in-context learning without further fine-tuning. This decision was due to financial considerations, as the fine-tuning process incurs substantial costs. To establish meaning benchmarks, traditional machine learning models such as Naive Bayes, Logistic Regression, and SVM were included. These models were chosen as baselines against which we the performance of fine-tuned BERT and GPT models were compared.

6.1 Evaluation Results

The evaluation focused on metrics that are crucial for disaster detection and so the metrics 'precision' and 'recall' were chosen to be evaluated.

The performance of traditional Machine Learning models can be seen in [Table.1].

The parameters chosen and the evaluated metrics for the pre-trained BERT model are as follows:

- Model Used: bert-base-uncased
- Batch Size: 16s
- Total epochs: 3
- Learning Rate: 2e-5

Table 1. Performance of Traditional ML Models

Model	Precision	Accuracy	Recall	F1
LogisticRegression	0.78	0.78	0.79	0.76
Naive Bayes	0.73	0.75	0.81	0.77
SVM	0.77	0.77	0.78	0.78

The evaluation metrics can be seen in [Table.2]

Table 2. Performance of BERT

Model	Precision	Accuracy	Recall	F1
Pre-trained BERT	0.82	0.85	0.86	0.84

Finally, prompt engineering was performed using carefully selected examples and using the few-shot learning technique with the GPT-3.5 model. A few samples were tested against the GPT model using prompts and the results can be seen below:

Table 3. Performance of GPT-3.5

Model	Precision	Accuracy	Recall	F1
GPT-3.5 Turbo	0.75	0.85	1.0	0.8

The BERT model outperforms the other models with the highest scores, indicating superior generalization capabilities on unseen data compared to the other models. However, traditional ML models demonstrate a slightly lower score indicating good overall performance but with scope for improvement.

7 DISCUSSIONS

The implementation of automated disaster detection system carries both significant benefits and potential risks. On the positive side, faster disaster detection can lead to more rapid emergency response, potentially saving lives and reducing suffering. The system's ability to work with minimal training data makes it particularly valuable for detecting emerging or unexpected types of disasters.

However, we must consider potential false positives and their implications for emergency resource allocation. Additionally, over-reliance on automated systems could lead to missed contexts or nuances that human operators might catch. We recommend implementing this system as a support tool for human decision-makers rather than a completely automated solution.

8 CONCLUSION

Our multi-model approach to disaster tweet classification demonstrates the complementary strengths of traditional machine learning, transformer models, and few-shot learning techniques. The BERT-based model achieved the highest overall performance, while the GPT-3.5 few-shot approach showed promising results with minimal training data. This research provides a practical foundation for implementing robust disaster detection systems that can be rapidly deployed and adapted to new scenarios.

9 TIMELINE AND MILESTONES

The estimated timeline while letting the project evolve as we develop our methodology, is as follow:

- **10/21/2024**: Submission of project proposal
- **10/26/2024**: Data preparation (missing values, noise etc)
- **11/04/2024**: Analyse review comments and make improvements
- **11/06/2024**: Data analysis and visualization (plots)
- **11/10/2024**: Model development
- **11/20/2024**: Evaluation and documentation
- **12/09/2024**: Submission of final report and presentation

References

[1] Disaster Identification from Tweets. https://www.kaggle.com/competitions/nlp-getting-started/data.

[2] Firoz Alam, Ferda Ofli, and Muhammad Imran. 2018. A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria. (05 2018)

[3] Sagar Gharge and Manik K. Chavan. 2017. An integrated approach for malicious tweets detection using NLP. 2017 International Conference of Inventive Communication and Computational Technologies (ICICCT) (2017), 435-438. https://api.semanticscholar.org/CorpusID:10975476

[4] Muhammad Imran, Pransenjit Mitra, and Carlos Castillo. 2016. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. CoPP abs/1605.05894 (2016). arXiv:1605.05894 http://arxiv.org/abs/1605.05894

[5] Guoqin Ma. 2019. Tweets Classification with BERT in the Field of Disaster Management. https://api.semanticscholar.org/CorpusID:204771273

[6] Salvador Medina Maza, Evangelia Spiliopoulou, Eduard Hovy, and Alexander Hauptmann. 2020. Event-Related Bias Removal for Real-time Disaster Events. In Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics,Online,3858–3868. https://doi.org/10.18653/v1/2020.findings-emnlp.344

[7] Shubhadeep Mukherjee and PRADIP BALA. 2017. Detecting Sarcasm in Customer Tweets: An NLP based approach. Industrial Management Data Systems 117 (08 2017). https://doi.org/10.1108/IMDS-06-2016-0207

[8] Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. 2011. Safety Information Mining — What can NLP do in a disaster—. In Proceedings of 5th International Joint Conference on Natural Language Processing. Asian Federation of Natural Language Processing, Chiang Mai, Thailand, 965–973. https://aclanthology.org/I11-1108

[9] Congcong Wang, Paul Nulty, and David Lillis. 2021. Transformer-based Multi- task Learning for Disaster Tweet Categorisation. CoRR abs/2110.08010 (2021). arXiv:2110.08010 https://arxiv.org/abs/2110.08010