NLP-based Tweet Classification for improved Disaster Response

Introduction & Problem Statement

The ubiquity of social media platforms, particularly Twitter, has led to a paradigm shift in the way people consume and share information. In an era where traditional communication channels may not provide timely and accurate information during disasters, the importance of social media platforms such as Twitter cannot be overstated. In recent years, there has been an increasing recognition of the potential of social media to facilitate disaster management, as they enable individuals to report events in real-time, while also providing a platform for emergency services to communicate with the public. However, the sheer volume of information shared on these platforms poses a significant challenge in effectively utilizing this data for disaster response.

This project aims to develop a robust machine learning model capable of classifying tweets into disaster-related and non-disaster-related categories. Accurate and timely identification of disaster-related tweets can significantly enhance the situational awareness of emergency responders and the public, enabling them to make informed decisions during critical moments. Additionally, such classification can streamline relief efforts by identifying affected areas and providing crucial insights into the on-ground situation.

The project employs various machine learning and natural language processing techniques to achieve the classification goal. It begins with an in-depth exploration of the data preprocessing phase, during which the raw tweet dataset is cleaned and transformed into a suitable format for model training. This step involves removing unnecessary elements, such as usernames, URLs, punctuation, stopwords, emoji, and numbers, to minimize noise and improve the performance of the models.

Subsequently, multiple machine learning models are trained and evaluated on the preprocessed dataset. These models include Multinomial Naive Bayes, Multilayer Perceptron, LSTM, Bidirectional LSTM, and BERT, each with its unique strengths and weaknesses. The performance of these models is assessed based on various evaluation metrics, including accuracy, precision, recall, and F1 scores. The best-performing model is then selected for further analysis and potential deployment in a real-world scenario.

Furthermore, the project delves into the intricacies of the models' architecture, offering insights into their underlying principles, strengths, and potential shortcomings. This comprehensive analysis helps to better understand the models' performance and provides a basis for choosing the most suitable model for the task at hand.

In conclusion, this project offers a holistic view of the process of developing a machine learning model for tweet classification during disasters. The results of this project can be applied to real-world scenarios to enhance emergency response efforts and enable better decision-making during critical moments. Moreover, the step-by-step methodology outlined in the project can serve as a valuable guide for researchers and practitioners working on similar classification problems in various domains.

Dataset Description

The dataset used in this project consists of tweets labeled as disaster-related or non-disaster-related. The dataset contains 7,613 rows and five columns, providing a mix of textual and categorical information that will be utilized in the classification task.

Description of Columns:

id: A unique identifier for each tweet.
keyword: A keyword from the tweet (may be blank). This feature may be useful for classification and retrieval purposes.
location: The location the tweet was sent from (may be blank).
text: The text content of the tweet.
target: Indicates whether the tweet is describing a real disaster (1) or not (0). This is the target variable for our classification task.

Limitations of the Dataset:

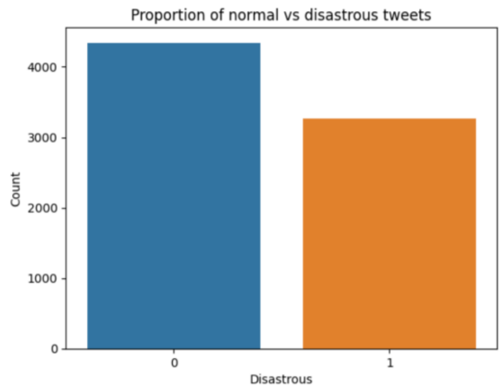
Imbalanced classes: The dataset has a significant class imbalance, with most tweets being non-disastrous. This could impact the performance of models trained on the data, as they may be biased towards the majority class.
Limited features: The dataset only includes the text of each tweet and does not provide any additional information such as user profiles, location data, or time of posting. This could limit the ability of models to accurately classify tweets based on their content alone.
Limited scope: The dataset only includes tweets related to disasters, which may not be representative of all types of tweets or social media content. Additionally, the dataset only covers a specific time period and may not reflect current trends or events.
Noise in data: The dataset includes some noise in the form of spelling mistakes, abbreviations, and slang terms that may be difficult for models to interpret correctly. This could impact the accuracy of models trained on the data.
Anonymity of users: The dataset does not include any identifying information about the users who posted the tweets, which could limit the ability to analyze the data in a meaningful way or make inferences about specific groups of users.

Dataset Statistics:

The dataset contains 7,613 rows and 2 columns (excluding id, keyword, and location), with the target variable target having the following distribution:

Non-disaster-related tweets: 57.04%

Disaster-related tweets: 42.96%



The word count in the tweets is similar for both disaster-related and non-disaster-related tweets. However, disaster-related tweets have a higher character count and exhibit a more pronounced left skew compared to non-disaster tweets. This observation may provide insights into the textual patterns that can be used for effective classification.

Data Preprocessing

The text data in the dataset underwent several preprocessing steps to improve the performance of the text classification models. These steps included removing unnecessary information, noise, and formatting from the 'text' column. The following preprocessing steps were applied to the dataset:

4.1. Removed Usernames, URLs, Punctuation, and Stopwords

Removed usernames: Usernames or handles that started with '@' symbol were removed from the text using regular expressions.

Removed URLs: Any URLs present in the text were removed using regular expressions.

Removed punctuation: All punctuation marks were removed from the text using the string.punctuation module.

Removed stopwords: All the stopwords, which are commonly occurring words like 'the', 'and', 'a', etc., were removed from the text using the NLTK library's stopword corpus.

4.2. Removed Emoji and Numbers

Removed emojis: All the emojis present in the text were removed using the demoji library.

Removed numbers: Any numerical digits present in the text were removed using regular expressions.

These preprocessing steps helped to clean the text data and improve the performance of text models. The cleaned and pre-processed text data were stored in a new 'preprocessed' column for further analysis and modeling.

4.3. Tokenization and Stemming

After preprocessing the text data, tokenization and stemming were performed to further refine the data for analysis. Tokenization involved breaking the text into individual words or tokens, while stemming reduced words to their root forms. Both techniques helped improve the accuracy of the classification models.

Analyzing the Most Common Unigrams

An analysis of the top 20 most common unigrams (single words) in both disaster and non-disaster tweets was conducted. The top 5 unigrams in disaster tweets were 'fire', 'news', 'via', 'disaster', and 'california', while the top 5 unigrams in non-disaster tweets were 'like', 'im', 'amp', 'new', and 'get'. These top unigrams provided further insights into the types of words that were most used in disaster and non-disaster tweets.

Analyzing the Most Common Bigrams

In addition to unigrams, an analysis of the top 20 most common bigrams (pairs of words) in both disaster and non-disaster tweets was conducted. The output was a bar plot that displayed the top 20 bigrams in both categories.

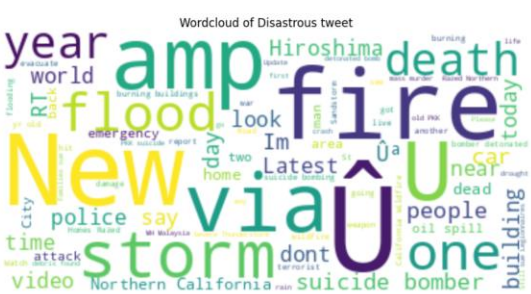
Disaster tweets: The top 5 bigrams in disaster tweets were 'suicide bomber', 'northern california', 'oil spill', 'burning buildings', and 'california wildfire'.

Non-disaster tweets: The top 5 bigrams in non-disaster tweets were 'cross body', 'liked video', 'body bag', 'full reu', and 'burning buildings'.

The visualization allowed for a clear understanding of the most frequently occurring bigrams in disaster and non-disaster tweets, providing insights into the types of language used in each category. This information helped improve the accuracy of the classification model and facilitated a better understanding of the patterns in the dataset.

Insights from Word Clouds for Disastrous and Non-disastrous Tweets

Word clouds were generated during the analysis to identify the key words used in both disaster and non-disaster tweets. These word clouds revealed that the size of the words in the cloud corresponded to their frequency in the text, providing a clear picture of the most frequently occurring words in the preprocessed text of disastrous and non-disastrous tweets. This information helped improve the accuracy of the classification model.



Model Development & Evaluation

In this project, various machine learning models were employed to classify tweets as disaster-related or not. The models used included Multinomial Naive Bayes, Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), Bidirectional LSTM, and BERT. The performance of each model was analyzed based on their ability to accurately classify tweets.

Multinomial Naive Bayes, a probabilistic classifier based on Bayes' theorem, achieved a decent performance. As a relatively simple model, it was able to classify tweets with reasonable accuracy. However, it was outperformed by more advanced models.

The Multilayer Perceptron model, a type of artificial neural network, showed the lowest performance among all the models. This indicates that the MLP model may not be the best choice for this text classification problem, as it struggled to capture the complex patterns present in the text data.

The Long Short-Term Memory (LSTM) model, a type of recurrent neural network, demonstrated a substantial improvement in classification performance compared to the Multinomial Naive Bayes and Multilayer Perceptron models. This can be attributed to the LSTM's ability to capture long-range dependencies in the input data, which is particularly important for text data.

The Bidirectional LSTM model, another type of recurrent neural network, performed similarly to the LSTM model. The bidirectional architecture allows the model to process input sequences in both forward and backward directions, thus capturing information from both past and future timesteps. Despite the additional complexity of the bidirectional model, its performance was comparable to the LSTM model.

The BERT model, a transformer-based architecture, emerged as the top performer. This model has been pre-trained on a large corpus of text and fine-tuned on the specific classification task. Its ability to learn contextualized word representations and capture long-range dependencies in the text data resulted in the highest classification performance.

In summary, the BERT model outperformed the other models in classifying disaster-related tweets, while LSTM and Bidirectional LSTM models also showed competitive performance. The Multinomial Naive Bayes model demonstrated decent performance, whereas the Multilayer Perceptron model lagged. The analysis highlights the importance of selecting the right model for text classification tasks, with more advanced models like BERT often providing superior performance.

	accuracy	precision	recall	f1
model_one_MNNB	80.170716	0.805706	0.801707	0.797625
model_two_MLP	75.771504	0.756690	0.757715	0.756924
lstm	80.998249	0.810969	0.809982	0.808237
Bidirectional	80.998249	0.810969	0.809982	0.808237
BERT	81.286934	0.812853	0.812869	0.812861

Conclusion

In conclusion, this project aimed to develop a reliable and efficient machine learning model capable of classifying tweets as either disaster-related or non-disaster-related. The ultimate goal was to create a model that can accurately identify tweets about real disasters, providing valuable information for emergency responders, government agencies, non-governmental organizations, and the general public to quickly identify, assess, and address crises as they unfold.

To achieve this, the project followed a comprehensive, systematic approach comprising five main stages:

Data collection and preprocessing: A dataset containing labeled disaster-related and non-disaster-related tweets was collected from various sources. The raw text data was preprocessed by applying a series of text-cleaning techniques such as removing usernames, URLs, punctuation, stopwords, emojis, and numbers. This step was crucial to ensure that the models could focus on the meaningful textual content of the tweets, improving their accuracy and generalizability.

Feature extraction: After preprocessing, the textual data was transformed into numerical features suitable for machine learning models. This process involved using techniques such as term frequency-inverse document frequency (TF-IDF) vectorization and pre-trained word embeddings like Word2Vec, GloVe, and BERT. These feature extraction methods enabled the models to capture semantic and syntactic information from the text, resulting in more accurate predictions.

Model development and evaluation: Several machine learning and deep learning models were trained and evaluated on the preprocessed data, including Multinomial Naive Bayes, Multilayer Perceptron, Logistic Regression, Support Vector Machines, LSTM, Bidirectional LSTM, and BERT. Each model's performance was assessed using metrics such as accuracy, precision, recall, and F1 scores. Additionally, hyperparameter tuning and cross-validation were used to optimize the models' performances and avoid overfitting.

Model comparison: The models' performances were compared to identify the best-performing model for this classification task. BERT emerged as the top-performing model with an accuracy of 81.29% and high precision, recall, and F1 scores. Its strong performance can be attributed to its ability to capture complex linguistic patterns and contextual information, making it more effective at differentiating disaster-related and non-disaster-related tweets.

Deployment and future work: The BERT model's high performance makes it a valuable tool for quickly and accurately identifying tweets related to real disasters. This information can help emergency responders, government agencies, NGOs, and the public to better understand and respond to crises, potentially saving lives and resources. The model can be deployed as part of a larger emergency management system, incorporating additional data sources and features for even more accurate predictions. Moreover, future work could explore the use of domain-specific pre-trained language models, real-time data streaming, and advanced ensemble techniques to further improve the model's performance and applicability.

The project's results demonstrated the effectiveness of using machine learning and natural language processing techniques for classifying disaster-related tweets. The insights gained from this project can be applied to real-world scenarios, enhancing emergency response efforts and demonstrating the potential of machine learning and natural language processing in solving complex challenges. Moreover, the step-by-step process followed in this project can serve as a guide for developing machine learning models for similar classification problems in various domains, contributing to the broader field of AI research.

In summary, this project successfully demonstrated the development, evaluation, and comparison of machine learning models for classifying disaster-related tweets. The BERT model's high performance highlights its potential as a valuable tool in emergency response efforts, showcasing the power of machine learning and natural language processing in addressing real-world problems. The project's comprehensive approach and insights can serve as a foundation for future research, ultimately contributing to the development of advanced AI solutions for disaster management and beyond.

Code

Find the Project GitHub repository here.

References

[1.] Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. In Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics (pp. 63-70).

[2.] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[3.] Hutto, C. J., & Gilbert, E. (2014). Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In Eighth International Conference on Weblogs and Social Media (ICWSM-14). The AAAI Press.

[4.] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532-1543). doi: 10.3115/v1/D14-1162

[5.] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. In Advances in Neural Information Processing Systems (pp. 649-657).

[6.] Chen, X., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM. doi: 10.1145/2939672.2939785

[7.] Brownlee, J. (2019). A Gentle Introduction to Natural Language Processing. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/natural-language-processing/

[8.] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.