ROBERTA-ENHANCED SENTIMENT ANALYSIS OF PANDEMIC TWEETS: A COMPARATIVE STUDY MACHINE LEARNING MODELS ON LABELLED AND SEMI-SUPERVISED DATASETS

HAFSA R MOHAMMED
University of Greenwich
Computing and Mathematical Sciences

Contact Information: School of Computing and Mathematical Sciences
University of Greenwich
Old Royal Naval College, Park Row, London SE10 9LS

Phone: **020 8331 8000**
Email: hm6942o@gre.ac.uk

Abstract

Sentiment analysis of tweets has become an important tool for understanding public opinion about various topics, including the pandemic. In this project, we developed a semi-supervised sentiment analysis model for pandemic-related tweets using RoBERTa-embeddings machine learning algorithms. We analyzed sentiment on tweets linked to the pandemic using a semi-supervised learning technique that blends labelled and unlabelled data to increase the precision of the sentiment analysis model. RoBERTa model was used to generate contextual embeddings for the unlabelled tweets, and these embeddings were used to compute sentiment scores for each tweet. These scores were then used as labels for training and evaluating different machine learning models. The tf-idf vectorizer was used to convert the text data into numerical features that can be used as inputs to the machine learning models. The comparison of models using both labelled and semi-supervised datasets, ml models used were svm, mlp, lr, cnn and lr was the best model, which didn't overfit. The results of our study suggest that the combination of RoBERTa embeddings and semi-supervised learning can significantly improve the accuracy of sentiment analysis on pandemic-related tweets. The established model could be used to help government and healthcare organizations better gauge public opinion about the epidemic and make informed decisions.

Introduction

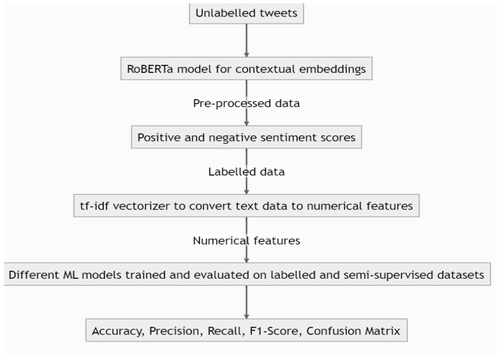
The COVID-19 pandemic has had a significant impact on society, affecting all aspects of life. During the pandemic, social media platforms such as Twitter were often used as a means of communication and information sharing, providing valuable insights into the attitudes and behaviors of the general public. Sentiment analysis of pandemic-related tweets has become increasingly important in this context. Sentiment analysis involves identifying the emotional tone or attitude of a text and categorizing opinions into "positive," "negative," or "neutral" (Kharde & Sonawane, 2016). Our proposed methodology for analyzing pandemic-related tweets using RoBERTa model-generated contextual embeddings and machine learning models such as SVM, MLP, LR, and CNN with tf-idf vectorization can provide useful information for public health authorities, government bodies, and healthcare professionals to prepare for future outbreaks. This poster will present the objectives, methodology, and results of our study, as well as its potential limitations and avenues for future work.

Project Idea Objectives

The objective of this project is to utilize machine learning techniques to analyse the sentiments of tweets related to pandemics, specifically the COVID-19 pandemic. The project aims to achieve the following objectives:

- Develop a methodology to generate contextual embeddings using RoBERTa model for unlabelled tweets.
- Utilize the generated embeddings to compute sentiment scores for each tweet.
- Train and evaluate various machine learning models using the sentiment scores as labels, including SVM, MLP, LR, and CNN.
- Compare the performance of these models using both labelled and semi-supervised datasets.
- Identify the most effective machine learning model for sentiment analysis of pandemic-related tweets.
- Visualize the results through a dashboard built on Tableau.
- By accomplishing these objectives, the project can provide a useful tool for public health authorities, governmental bodies, healthcare professionals, and researchers to better understand the public's sentiment during pandemics and to inform decision-making.

Project Lifecycle



Methodology

The methodology used in this project involved several steps to build a sentiment analysis model for pandemic-related tweets. The following steps were followed:

- Data Collection: Tweets were extracted for the unlabelled data from 2021 to 2022 December. Labelled tweets were from year 2020.
- Data Preprocessing: Preprocessing of the tweets was done to remove unwanted information such as URLs, mentions, and stop words. The remaining text was then cleaned using techniques like stemming and lemmatization.
- Embedding Generation: The RoBERTa model was used to generate contextual embeddings for the unlabelled tweets. The embeddings captured the meaning of the text in a higher-dimensional space, making it easier for the machine learning models to understand the context of the tweets.
- Sentiment Score Calculation: The generated embeddings were used to compute sentiment scores for each tweet. These scores ranged from -1 to +1, with negative scores indicating negative sentiment and positive scores indicating positive sentiment.
- Feature Engineering: The TF-IDF vectorizer was used to convert the text data into numerical features that can be used as inputs to the machine learning models.
- Machine Learning Model Training: The sentiment scores were used as labels for training and evaluating different machine learning models such as SVM, MLP, LR, and CNN. Both labelled and semi-supervised datasets were used for model training and comparison.
- Tableau Dashboard: All the visual graphs from Jupyter Notebook were put into a Tableau dashboard to make it more visually appealing and interactive for users.

Figure 1: Word Cloud

These terms capture the present worries and issues surrounding the epidemic, such as the dangers posed by the virus, efforts to contain its spread, and legislative responses and length of tweets.

Results

The results indicate that all three models achieved high accuracy rates on the training data, with the MLP model achieving 100% accuracy. However, the MLP model showed overfitting on the training dataset, resulting in lower accuracy rates on the validation and test datasets. The SVM and LR models, on the other hand, achieved consistent performance across all datasets. After hyperparameter tuning, the SVM model showed the most improvement in accuracy, with an increase from 82.19% to 90.24% on the validation dataset and from 82.62% to 90.04% on the test dataset. The LR model also showed slight improvement in accuracy after hyperparameter tuning. Overall, the semi-supervised learning approach that combined labelled and unlabelled data resulted in accurate sentiment analysis on pandemic-related tweets.

Models	Accuracy (1)	Accuracy (2)
MLP	~~100~~%	~~100~~%
SVM	49.76%	83.81%
LR	75.06%	92.58%

Table 1: Test Accuracy (before & After Hyperparameter Tuning) of MLP, SVM, and LR

The models were evaluated on their accuracy, precision, recall, and F1-score

The best model, LR, had an accuracy of 75% on the labelled dataset and 92%on the semi-supervised dataset

The precision, recall, and F1-score were also calculated for each model

Model	LR train	LR test	MLP train	MLP test	SVM train	SVM test	DL train	DL test
Accuracy	93.52	75.06	100.0	73.94	100.0	49.76	51.22	51.00
Error Rate	6.48	24.94	0.0	26.06	0.0	50.24	48.78	48.94
Sensitivity/Recall	96.98	86.27	100.0	80.24	100.0	99.88	0.00	0.00
Specificity	89.32	64.37	100.0	67.93	100.0	1.95	100.00	100.00
Precision	91.70	69.79	100.0	70.48	100.0	49.29	0.00	0.00
F1	94.26	77.16	100.0	75.04	100.0	66.00	0.00	0.00

Table 2: Labelled Metrics

Model	Metric	Train	Test	Hyperparameter Tuned Train	Hyperparameter Tuned Test
LR	Accuracy	96.76	92.58	100.0	93.41
	Error Rate	3.24	7.42	0.0	6.59
	Sensitivity	94.37	88.11	100.0	91.18
	Specificity	99.11	97.25	100.0	95.74
	Precision	99.05	97.10	100.0	95.73
	F1	96.65	92.39	100.0	93.40
MLP	Accuracy	100.0	92.32	100.0	92.19
	Error Rate	0.0	7.68	0.0	7.81
	Sensitivity	100.0	90.86	100.0	90.68
	Specificity	100.0	93.84	100.0	93.78
	Precision	100.0	93.92	100.0	93.85
	F1	100.0	92.37	100.0	92.23
SVM	Accuracy	100.0	83.81	100.0	90.08
	Error Rate	0.0	16.19	0.0	9.92
	Sensitivity	100.0	68.46	100.0	81.79
	Specificity	100.0	99.87	100.0	98.76
	Precision	100.0	99.82	100.0	98.57
	F1	100.0	81.22	100.0	89.40
DL	Accuracy	51.22	51.00	N/A	N/A
	Error Rate	48.78	48.94	N/A	N/A
	Sensitivity	0.0	0.00	N/A	N/A
	Specificity	100.0	100.00	N/A	N/A
	Precision	0.0	0.00	N/A	N/A
	F1	0.0	0.00	N/A	N/A

Table 3: Semi-supervised Metrics

Unigram, Bigram, Trigram

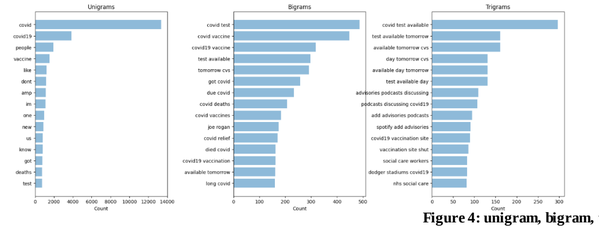


Figure 4: unigram, bigram, trigram.

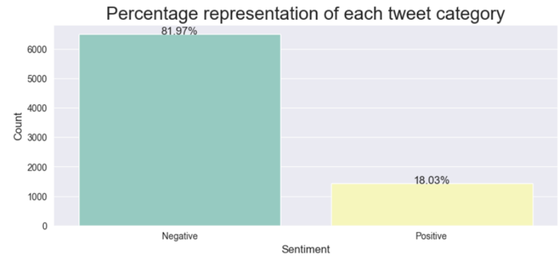


Figure 5: Percentage representation.

Negative – 81.91%, Positive – 18.03%

Evaluation

The Logistic Regression model seems to be the best appropriate for sentiment analysis, according on the results of tests done on the labelled and semi-supervised datasets. On both datasets, the model had excellent accuracy, a low error rate, high specificity, and high precision. On the other hand, the MLP model, which performed well on the labelled dataset, overfit on the semi-supervised dataset and did not generalise effectively to new data.

Conclusions

In conclusion, our experiment proved that contextual embeddings produced by RoBERTa may be utilised to precisely categorise sentiment in tweets. Our research suggests that the LR model is a good and efficient machine learning method for social media sentiment analysis jobs. The tf-idf vectorizer also proved to be a useful tool for transforming text data into numerical characteristics that can be utilised as inputs to machine learning models.

Limitations Further Work

- High computational power required for using RoBERTa model to generate contextual embeddings may limit practicality for some users or applications
- Performance of machine learning models varied depending on type of model used, with MLP and SVM overfitting and logistic regression achieving lower performance after hyperparameter tuning
- CNN model achieved lowest results, suggesting it may not be the best fit for sentiment analysis in social media
- Small size of labelled dataset may limit generalizability of results
- Use of web scraping for unlabelled tweets may have introduced bias or missing data and resulted in small sample size or inadequate representation of target audience, impairing generalizability of findings.

References

Kharde, V. and Sonawane, P., 2016. Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971.

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L. and Neves, L.T., 2020. Unified benchmark and comparative evaluation for tweet classification. Findings of the Association for Computational Linguistics.