Sentiment Analysis towards COVID-19 on Twitter

Chenglong Wei (z5375926), Yewei Huang (z5459400), Gorjan Muratov (z5677486), Sidhanth Kafley (z5504979), Ziyi Ding (z5610550)

I. INTRODUCTION

The COVID-19 pandemic has sparked widespread discussion on social media platforms, particularly on Twitter. This project aims to analyze COVID-19 related tweets across different sentiments (positive, negative, or neutral) using natural language processing (NLP) techniques. Ultimately, trying to answer the question, “How can Deep Learning and Neural Networks be utilised to improve sentiment analysis of COVID-19 Tweets?”.

Two approaches will be explored to attempt to answer this question from multiple aspects. Approach one will explore the effect Deep Learning has on sentiment analysis at the model building phase. Comparing the results between rule-based, machine learning, and deep learning models. The second approach will analyse the effect Neural Networks have on the embedding portion of the task. Evaluating the difference between frequency based and neural network based embedding approaches and the effect the different strategies have on sentiment analysis.

Tweets are short, informal, and have highly variable context yet hold much information within them that can be used to understand the stance society has towards a certain topic. During the COVID-19 pandemic, news and public awareness about the virus spread rapidly, much like the pandemic itself. Online opinions played a significant role in shaping public understanding. Analysing the sentiment of COVID-19 related tweets is important for stakeholders such as policymakers, crisis managers, and public health authorities as it gives them insights into how individuals perceive the pandemic.

II. RELATED WORK & LITERATURE REVIEW

“COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis” [1] crawls COVID-19 related tweets from February to March 2020 and manually annotates English tweets, introducing a sentiment dataset COVIDSenti. The paper compares a broad range of ML, DL, hybrid and transformer-based models, finding that BERT achieves the best performance. Additionally, the authors conduct exploratory analyses, including keyword trends, topic modelling, and sentiment timelines, revealing shifts in public opinion toward government policies as the pandemic evolved.

"Machine Learning Techniques for Sentiment Analysis of COVID-19 related Twitter Data"[2] conducts a literature review of 40 peer-reviewed papers published between October 2019 and January 2022, focusing on machine learning techniques for sentiment analysis of COVID-19-related Twitter data. The studies reviewed employed three main approaches: lexicon-based methods, traditional ML, and DL models. While lexicon-based methods are simple but lack contextual nuance. ML models offer better performance but require vast amounts of labelled datasets to achieve the desired results. DL models, especially transformer-based architectures like BERT and RoBERTa, achieved the highest accuracies when fine-tuned on Twitter-specific data. The paper also observed a strong limitation of transferability of models due to linguistic and semantic differences between languages, as most studies relied on English datasets.

III. METHODS

Approach one will compare and analyse different models and their ability to provide accurate sentiment analysis. Standardizing everything and varying the model is a logical start to accurately assess the effect deep learning will have on sentiment analysis of COVID-19 tweets. As a benchmark, Vader and Textblob were chosen. These rule-based models provide a quick and simple implementation which requires no training time and is applicable to many use cases. Such models could be preferred by stakeholders who are less technical as these models are extremely explainable. Vader has the added benefit of being specifically designed for sentiment analysis on social media-oriented information. This will be followed by traditional machine learning models such as Bernoulli Naive Bayes, Multinomial Naive Bayes, and Logistic Regression. These models also have the benefit of being highly explainable to stakeholders as they are extremely popular and well known. They provide a more hands on approach to solving the problem as training and testing is done by the developer. Such an approach is preferred by most professions. These models are also predicted to have better performance in sentiment analysis as they learn from the data they are provided instead of following generic rules. Finally, a neural network and deep learning approach will be attempted through the implementation of LSTM and GRU models. These models were chosen because of their ability to learn, understand, and remember information long term. Both models utilise gates to remember crucial information and discard information that doesn't hold much value. Because of this, these models are expected to perform the best for this task.

Rule-based models

TextBlob: This method assigns a polarity score to each word based on a predefined sentiment lexicon, with values ranging from –1 (most negative) to +1 (most positive). We used it as a simple baseline, hypothesizing that a dictionary-based approach could provide a quick and interpretable measure of public opinion, even without training on the dataset.

VADER: This method incorporates heuristic rules to account for punctuation, capitalization, degree modifiers, and even emojis. We hypothesized that VADER would outperform generic dictionary-based approaches on Twitter data due to its specific tuning for online conversational language.

Traditional Machine Learning Models
To convert the raw text into numerical features, we used the CountVectorizer from the scikit-learn library with the parameter binary=True. Each tweet was converted into a binary vector, where each feature represents whether a sentiment word appeared in the tweet (1 for presence, 0 for absence). This binary representation is suitable for the Bernoulli Naive Bayes model.

Bernoulli Naive Bayes (BNB): This method focuses on whether a given sentiment appears, rather than how often it appears. We hypothesized it would work well for sentiment analysis since detecting whether a word like 'happy' or 'sad' appears might be enough.

Multinomial Naive Bayes (MNB): This is similar to BNB, both being Bayesian algorithms. However, unlike BNB, MNB focuses on frequency rather than presence. This model uses word frequency and assumes a multinomial distribution. By considering the frequency of sentiment words, it helps capture the intensity of sentiment.

Logistic Regression (LR): This is a linear classifier that learns a weight for each word and estimates the probability of it belonging to different sentiment categories. Unlike Naive Bayes, it does not assume independence between features, thus capturing interactions between words. Given the high dimensionality of the input features, we set its max_iter parameter to 1000 to ensure model convergence.

Deep Learning Models
LSTM is an enhanced version of the RNN, and it can handle long-term dependencies in sequential data with an extra memory cell and three gates. Because of this characteristic, it can remember information over extended periods and is suitable for NLP tasks.

The LSTM model was configured with 40 embedding dimensions, 64 hidden units, and 3 output units. It was trained using the Adam optimizer with a learning rate of 0.001 and a weight decay of 0.00005. With a batch size of 16 and 20 training epochs, the model achieved its best performance, reaching approximately 94% across all evaluation metrics. For comparison, increasing the batch size to 64 and removing weight decay led to a performance drop, with metrics falling to around 88%. We also compared two tokenizer methods - one that preserved punctuation with tokens and one that removed it. The tokenizer that retained punctuation consistently produced better results, even when paired with a smaller network.

The GRU is the final deep learning model that was explored. Like the LSTM it is also an RNN network that utilises gates to learn and was created to solve the vanishing gradient problem. GRU's, however, are a simpler alternative to the LSTM and have fewer parameters while having similar performance. This means that the model offers efficient training and a faster inference time, which is the reason why the team wanted to investigate its performance.

The GRU managed to attain 92% across the board for all four metrics that are being evaluated which implies that the model was able to learn the sentiment task well fully understanding the sarcasm, temporal differences, and overall context of the tweets. A grid search was performed to determine the optimised hyper-parameters for the model. The grid search yielded the following, embedding dimension 100, hidden dimension 128, learning rate 0.001, number of layers 3. The optimal model was trained on 20 epochs.

Approach two will investigate how different text embedding strategies affect the performance of sentiment analysis. For this approach we used Logistic regression model with Count Vectorizer as the baseline model and compared the effects of TF-IDF, Word2Vec and GloVe embeddings to it. The reason for why we are used logistic regression model was because there are many companies and stakeholders who would like to perform sentiment analysis of COVID-19 related tweets but don't want to or have the computational resources to run large deep learning models. For such stakeholders, we would like to investigate how much of an uplift can be achieved by using dense word embeddings along with a light machine learning model like logistic regression.

TF-IDF: Is a frequency based embedding method which weighs each word by its frequency in the document and scales it inversely by its frequency across the corpus.

Word2Vec: Is a neural network-based word embedding method that learns dense vector representations for words based on their context in the corpus.
GloVe: Is an extension of Word2Vec model and produces dense word embeddings by not only looking at the context of word but also the cooccurrence of word in the corpus.

IV. EXPERIMENTS

Data Overview

The COVID sentiment dataset can be found at the following link: **https://github.com/usmaann/COVIDSenti.git** The dataset consists of approximately 90,000 entries, where 74.9% are neutral, 18.2% are negative, and 7% are positive. It clearly exhibits a notable imbalance in the sentiment label distribution. Neutral sentiment accounts for the largest proportion, followed by negative and positive categories. Such a pronounced class imbalance, with substantial dominance of the neutral category, has the potential to introduce bias during model training and reduce the classifier's ability to accurately predict minority classes. As a result of this metrics must not only be viewed at an average level but rather class by class to understand how the model truly performs.

Analysis of text feature distributions indicates that tweet lengths approximately follow a normal distribution. Median lengths across negative, neutral, and positive categories are relatively similar, and average lengths by label show a consistent pattern. It is fair to assume that the tweets are similar in this regard. Examination of social interaction features, particularly the density of user mentions, reveals that mention density is generally higher in negative and neutral tweets than in positive ones, suggesting differences in communication patterns across sentiment categories. High-frequency terms in the dataset are predominantly associated with the COVID-19 pandemic, including words directly referencing the virus. This concentration of pandemic-related vocabulary reflects the dataset's thematic focus on discussions concerning COVID-19.

Evaluation

The metrics chosen to evaluate the efficiency of the approaches are accuracy, precision, recall, and F1 score. Accuracy indicates the proportion of correctly predicted tweets and shows the overall correctness of the approach. However, accuracy can be misleading and may hide class imbalances. Due to this, precision, recall, and F1 are also considered. Precision measures the reliability of predictions. For example, given the positively predicted tweets, how many are actually positive. Recall identifies the model's ability to determine the relevant tweets. For example, for all the tweets that are truly one sentiment, how many of them did the model identify correctly. Finally, F1 was chosen as it is a score that combines both precision and recall and finds a harmonic meaning between them. F1 is specifically useful when class imbalances are present. The combination of these metrics was used to evaluate the models, how often they were correct, and how they performed for each sentiment class. To ensure standarisation the dataset was split into train, validation, and test sets in the 70:20:10 ratio respectively.

Error Analysis

Representative tweets were examined to compare the predictions of different models against the true sentiment labels. The results show that the LSTM model demonstrates a superior ability to capture long-range contextual dependencies and implicit connotations, leading to more contextually appropriate predictions. In contrast, Rule-Based Vader and Logistic Regression model primarily rely on local word neighborhoods and fail to effectively link the beginning and end of sequences, resulting in misclassifications in cases involving nuanced sentiment or figurative language.

Tweet	Rule-Based Vader	Logistic Regression	LSTM	True Label
when i said i wanted to die,, i didnt mean to die from coronavirus lmao no lawd ive suffered enuf plz	Pos	Neg	Neu	Neu
oh great, two vancouver school are closed because there are kids with presumptive positive cases of coronavirus f me	Pos	Neg	Neu	Neu
even a mild case of covid-19 is extremely serious (might even include pneumonia). the word mild is very misleading.	Neg	Neg	Neu	Neu

V. RESULTS

Approach 1

The results show clear performance differences across model types and indicate that deep learning approaches outperform traditional and rule-based methods in capturing complex semantic cues. This can be attributed to the deep learning model's ability to learn long term context and remember it. The gates within the GRU and LSTM models allow them to remember the important information while disregarding the less substantial terms. This in turn results in both the models learning the social cues, temporal differences, sarcasm, and context within the tweets to achieve really good results at determining the sentiment. The LSTM outperforms all the other models and is clearly the best performing model.

	Precision	Recall	F1	Accuracy
TextBlob	71%	48%	52%	48%
Vader	71%	48%	52%	48%
Bernoulli Naïve Bayes	76%	79%	74%	79%
Multinomial Naïve Bayes	80%	81%	77%	81%
Logistical Regression	91%	91%	91%	91%
LSTM	94%	94%	94%	94%
GRU	92%	92%	92%	92%

Approach 2

The results show that embedding strategy has a significant impact on machine learning models' performance with the simple frequency-based embeddings such as Count Vectorizer and TF-IDF Vectorizer outperforming the dense neural network-based word embeddings for our specific dataset. Among the tested models, Count Vectorizer delivers the highest precision, recall, F1 and accuracy, while the Word2Vec embedding performed the worst across all the metrics. This could be because our dense word embedding models might be overfitting to the training dataset and not generalizing well with the test dataset. Another reason could also be because of the limited and short tweet length which would make it difficult for our word embedding models to better understand the context given the short and limited tweet length.

	Precision	Recall	F1	Accuracy
Count Vectorizer	91%	91%	91%	91%
TF-IDF	87%	87%	86%	87%
Word2Vec	71%	75%	68%	75%
GloVe	73%	77%	72%	77%

VI. CONCLUSION

Approach 1

This project focused on sentiment analysis towards COVID-19-related tweets, aiming to evaluate the performance of statistical models, machine learning models, and deep learning models. The study demonstrated that while statistical models such as TextBlob and VADER provide fast and interpretable sentiment analysis without training, their performance on informal, imbalanced tweets is limited. Machine learning models significantly improve their classification accuracy by learning feature weights from the dataset. And deep learning models outperform all other methods, leveraging sequential modeling to capture complex linguistic patterns and sentiment nuances. The deep learning models are also capable of leveraging their gate structure to understand the tweets and the underlying context, sarcasm, and dependencies.

Due to the above, the recommended model to use for sentiment analysis of COVID-19 tweets is an LSTM. Being able to choose what information within the sequence to remember and what to forget serves as a major advantage in learning the overall context of tweets. The model is also quick to train and easy to evaluate. A limitation of the LSTM approach is that implementing the proposed model might be challenging for individuals who are not very technical. If used in a professional setting the models' complicated nature makes explainability more difficult for non-technical stakeholders. If looking to implement, these are limitations that should be considered for the use case. To achieve greater uplift, future research could explore transformer technologies to assess their impact on sentiment analysis.

Approach 2

This project also explored how different word embedding techniques influence sentiment classification performance. The study demonstrated that stakeholders could achieve a high level of performance with sentiment analysis of COVID-19 related tweets using a simple logistic regression model with a simple frequency-based embedding like Count Vectorizer. While such approaches may not match the performance of large deep learning models, they offer a practical balance between computational resources and performance which can be useful particularly for stakeholders who want to conduct sentiment analysis on COVID-19 related tweets but don't have the computational resources to run large deep learning models. However, using large deep learning models such as LSTM still provides a significant uplift in performance with regards to sentiment analysis and deep learning models are essential if we want the best performance and highest accuracy. Future research could investigate how the different embedding techniques affect the performance of deep learning models such as LSTM.

REFERENCES

[1] Naseem, Usman, et al. "COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis." IEEE transactions on computational social systems 8.4 (2021): 1003-1015

[2] Braig, Niklas, et al. "Machine learning techniques for sentiment analysis of COVID-19-related twitter data." IEEE Access 11 (2023): 14778-14803.