INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Using XAI Techniques to Persuade Text Classifier Results: A Case Study of Covid-19 Tweets

S Hemdan Hamed[1]*, Hazem Elbakry[1], Haitham Elghareeb[1], Sara Elhishi[1]

1 Department of Information Systems, Faculty of Computers and Information, Mansoura University, Mansoura, Egypt

OPEN ACCESS

Received: 17-02-2022

Accepted: 14-07-2022

Published: 09-08-2022

Citation: Hamed SH, Elbakry H, Elghareeb H, Elhishi S (2022) Using XAI Techniques to Persuade Text Classifier Results: A Case Study of Covid-19 Tweets. Indian Journal of Science and Technology 15(30): 1484-1494. https://doi.org/10.17485/IJST/v15i30.397

* **Corresponding author**.

sehamhemdan@mans.edu.eg

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background: To offer a transparent decision support system able of classifying tweets' sentiment into positive, neutral, and negative sentiment and explains the prediction result by XAI techniques **Methods:** We started by data preprocessing phase. For data representation, we used TF-IDF, and we applied four machine-learning algorithms including Naive Bayes, random forest, logistic regression, and support vector machine, as well as four deep learning RNN, LSTM, GRU, and Bi-directional RNN. To raise model trust, we used LIME and SHAP to improve model explainability. **Findings:** The empirical findings show that the Logistic Regression model and SVM model using the TF-IDF feature extraction approach have the best performance when compared to the other models, with an average accuracy of 84% and 86% respectively. The data balancing step pushed the accuracy of the Random Forest model from 47% to 73%, other models slightly changed. The performance of deep learning models was better than traditional machine learning models, LSTM and GRU achieve approximately 78%, and Bi-directional RNN achieve 79% for dataset 2. **Novelty and applications:** we propose a highly accurate approach for SA which has been tested on two datasets. Also, to increase trust in model prediction, we explain the predicted sentiment.

Keywords: Explainable Artificial Intelligent (XAI); Sentiment Analysis; Covid19; Deep Learning; machine learning

1 Introduction

With the expansion and broad adoption of the Internet among users, it has become simple to access the Internet from anywhere in the world. On social media sites like Twitter and Facebook, people may share their thoughts and opinions on any issue. As a result, such ideas and perspectives create a vast volume of data via blogs, forums, and social media[(1)]. From the content of the opinion, this data may be utilized to deduce the user's sentiments about a product so demand for sentiment analysis has increased. Sentiment analysis (SA) is a type of natural language processing that may be used to forecast underlying sentiments from data[(2)]. To get competitive outcomes in shallow learning, a lot of focus is made on data preparation and feature engineering.

Deep learning classification models have surpassed traditional models in various areas, and they are now being used for sentiment analysis applications. They may be thought of as sophisticated computational models for extracting sentiments from text without using feature engineering [3].

RNNs are a common choice for any type of sequential data; According to Tang et al. RNNs are powerful because they combine two characteristics: (i) distributed hidden states, which allow them to store information from previous computations efficiently; and (ii) non-linear dynamics, which better fits the non-linear nature of data [4]. The most prevalent RNN variant is long short-term memory networks (LSTMs), which solve the gradient vanishing or exploding challenges that normal RNN designs have. GRU (Gated Recurrent Unit) was developed by Cho et al. in 2014 to resolve the vanishing gradient issue that arises with RNN [5]. GRU can also be viewed as a variation of the LSTM due to the similarities in their designs. The Bi-directional RNN is made up of two RNNs, one of which starts at the beginning of the data sequence and goes forward, and the other of which starts at the end and moves backward. That is, it is useful for preserving and analyzing both past and future events.

The area of Natural Language Processing (NLP) has seen a revolutionary change as a result of recent developments in neural architectures like the Transformer pushing the state of the art for a number of NLP tasks such as sentiment analysis, The performance of models like GPT and BERT using this Transformer design has completely surpassed that of the prior state-of-the-art networks.

Despite the excellent performance of deep learning models, they are still unable to explain why a certain prediction was made [6]. As a result, the field of Explainable Artificial Intelligence has sparked scientific attention (XAI), a field that produces more explainable models while maintaining high learning performance [7].

COVID-19 has been the subject of a lot of disinformation. So Bangyal et al utilized machine learning and deep learning methods to detect fake news on Twitter data, CNN and BiLSTM achieved the greatest accuracy of 97 % [8], also Ayoub et al proposed an explainable model using BERT and SHAP with an accuracy of 97%.

Cirqueira et al proposed an application that could be adopted to provide retailers with explanations by LIME and SHAP techniques and insights on their user needs during social media crises [9], this study evaluated many machine learning models and neural network multilayer perceptron (MLP) that obtain the best performance, but when compared with the size of the dataset, it was the lowest performance. Also, this study proposed an MLP model to predict sentiment that achieved an accuracy of 75.6% and applied the LIME technique to interpret the prediction made by the model [10]. Chakraborty et al. Take another path and proposed the use of a fuzzy rule base based on a Gaussian membership function to properly detect sentiment in tweets, and the accuracy of the model yields up to a maximum of 79 % [11].

Gite et al. proved that LSTM networks have shown to be an effective tool for learning and predicting temporal data with long-term dependencies. The prediction was explained by the LIME technique only [12], also SOURAV et al presented a new polarity-popularity paradigm, and sentiments were assigned to popular terms. Then, using both types of rated terms, they trained an LSTM model to predict sentiment on GST tweets, with an overall accuracy of 84.51 percent [13]. Wisesty et al introduced a comparative study for the sentiment analysis method, their study reached that the BERT classifier obtained the best performance with an f1-score of 85% [14], Reshi et al in their research, used hyper deep model(LSTM GRNN) for COVID-19 vaccination data classification with high accuracy 95 but without explanation to prediction result [15], Vaibhav Kumar also used hyper deep model (BiLSTM CNN)for COVID-19 tweets data classification with accuracy 87.38%, also without explanation to prediction result [16].

The following research questions are addressed in this study:

- Q1: How successful is the proposed approach for determining the sentiment of tweets?
- Q2: How to explain prediction results for decision-makers?

This study used two datasets first, the global Twitter dataset to answer the research questions and examine people's global perceptions and viewpoints on COVID-19, and second, the Stanford sentiment140 dataset from a Kaggle repository [17]. Natural language processing (NLP) and machine learning are used in the study to examine sentiments and evaluate the suggested methodology's effectiveness (ML). For sentiment analysis, TextBlob was utilized as an NLP lexicon-based approach, coupled with four machine learning models: Naive Bayes, random forest, logistic regression, and support vector machine as well as four deep learning models RNN, LSTM, GRU, and Bi-directional RNN. Furthermore, the study's sentiment analysis findings may help decision-makers create effective judgments and improve public awareness measures in the event of an epidemic.

2 Material and methods

This research offered a unified approach for doing sentiment analysis on a huge dataset as shown in Figure 1and Figure 2. The COVID-19 data was gathered using various hashtags on Twitter. Only the tweet ID is contained in the dataset, the tweet IDs are hydrated to obtain the tweet text. The dataset was preprocessed to remove unnecessary words, unneeded punctuation, stop words, and special symbols, making it clean and appropriate for training machine learning and deep learning models.

Preprocessed data is better for the models in terms of getting better classification results. TextBlob, which is a lexicon-based technique, was used to annotate the dataset. For the chosen models, the labeled data was divided into training and testing sub-datasets. The training data was used to train machine learning and deep learning models, and their performance was improved by optimizing a few parameters. Tested data was kept hidden from the models and used to evaluate their performance. To persuade decision-makers, the output predictions are explained using LIME and SHAP approaches.

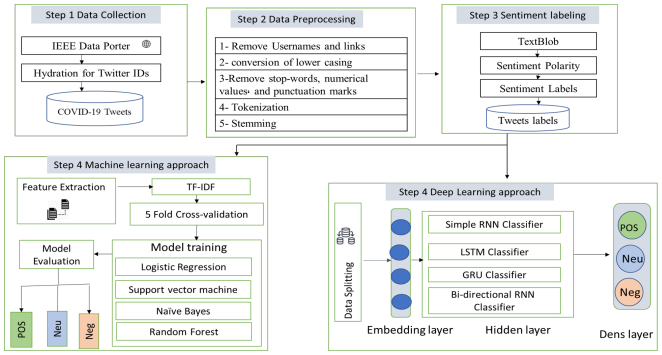

Fig 1. Proposed framework

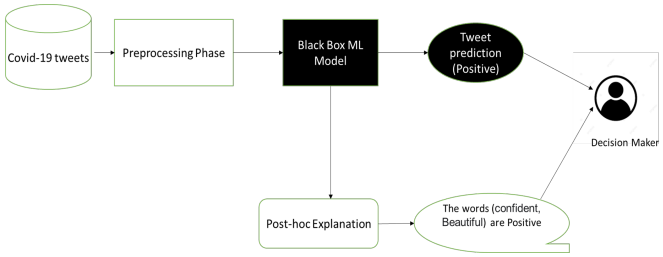

Fig 2. Explainable Sentiment Analysis

2.1 Data Collection

In our research we chose Twitter, a social media network, for extracting tweets on covid-19 for the first dataset, there are two types of Twitter APIs: search API and streaming API. Developers can search through a sample of tweets from the previous seven days with the Standard Version of the search API, as the Premium and Enterprise editions allow them to search through tweets from the past 30 days (30-day endpoint) or as far back as 2006 (Full archive endpoint), Tweets from the real-time Twitter feed are accessed via the streaming API [18]. In our study, we use the dataset from IEEE data porter on May 2020, 20 [19]. It provides 831327 tweet IDs and sentiment scores. Table 1 Shows an example of data.

Table 1. A sample of a dataset accessible on IEEE data port

Tweet IDS	Sentiment Score
1240808782000720000	0.6
1240808782550050000	0
1240808782894100000	-0.4

Because the distribution of Twitter data except IDs is restricted by Twitter policy, this dataset only contains tweet IDs. We hydrate tweets ids to get the tweets data by making a Twitter developer account and then accessing Twitter API. The API generates completely hydrated Tweet instances up to a hundred tweets per request. And the maximum number of requests each window is 900, the endpoint sends out approximately 8,640,000 tweets every two days. Here's how to do it: (1440/15) * 900 * 100, where 15 is the overall number of 15-minute windows for a day. The outcome of the hydration is a JSON file, this JSON file is converted to a CSV file and gets about 488570 tweets.

2.2 Preprocessing Phase

Two datasets are preprocessed after tweets are collected. Preprocessing is a crucial phase that affects the accuracy of learning models also in the explanation process. We remove username, URLs, Punctuations, Numbers, Special Characters, duplicated tweets, and short words and apply tokenizing and stemming in tweets. Table 2 Shows an example of original data corresponding to data after preprocessing for dataset 1.

Table 2. A sample of original tweets and tweets after preprocessing

No	Original Tweets	Tweet after Preprocessing
1	@MarcSkulnick @BoyGeorge Nice! I'll add that @BoyGeorge track to my self-isolation, store clean-up playlist for tomorrow. Thanks! And I wouldn't have expected you to be grooving to any other artis.	nice add that track self isol store clean playlist for tomorrow thank and wouldn have expect you groov ani other artist marc hope you and the fam are well despit the circumst stay safe mark
2	Isolation from virus. https://t.co/YlO9A6NtFR https://t.co/qqtdcylll3	isol from viru
3	@cloudsteph Totally pissed off with it. Off one, on to the next. Isolation, it's not	total piss off with off one the next isol not

2.3 Sentiment Labeling

We use TextBlob to find sentiment scores after the data has been preprocessed. The sentiment values are divided into three categories: positive (greater than 0), neutral (equivalent to 0), and negative (just under 0). Table 3 Displays the score of the sentiment, the number of tweets in each of these classes, and the textblob algorithm. We use Wordcloud to show the most frequent words in each class as shown in Figure 3.

Fig 3. (a) Wordcloudfor Positive tweets. (b) Wordcloudfor Neutral tweets. (c) Wordcloudfor Negative tweets

After finding the sentiment score for dataset 1, we found the number of samples in each class was unbalanced as shown in Table 3. When each class has nearly the same number of samples, most machine learning methods work well; because the majority of algorithms are created to improve accuracy while reducing errors. However, if the data set is unbalanced, we can achieve a high level of accuracy only by forecasting the majority class, but you'll fail out on the minority class, which is normally the model's primary goal [20].

Table 3. The score of the sentiment, number of tweets in each class, and textblob algorithm

Tweet text	Polarity	Sentiment score by textblob	Number of tweets
A lot of you would actually benefit from taking a good look introspectively at your issues with women.	0.350000	Positive	125756
this quarantine has kicked my depression up a couple notches thanks to my work and routine being void now and im effectively avoiding my phone now bc everyone is nuts sending corona stuff	-0.066667	Negative	113693
notes is seen where he crossed out "Corona" and replaced it with "Chinese" Virus	0.000000	Neutral	249121

Continued on next page

Table 3 continued

TextBlob algorithm for sentiment analysis.
Input: tweets about the COVID-19
Result: is Positive if polarity score greater than zero
is neutral if polarity score equal zero
is negative if polarity score less than zero
Calculate the polarity score ,loop foreach tweet
Start condition
If (Polarity Score > 0), then Sentiment = Positive;
else, if (Polarity Score = 0), then Sentiment = Neutral;
else, Sentiment = Negative;
condition end
loop end

For dealing with imbalanced data in machine learning models, some techniques are deployed. we use the -Synthetic Minority Over Sampling- Technique (SMOTE)[21], It is based on simply producing data points on the line segment between a randomly picked data point and one of its K-nearest neighbors to sample data from the minority class. Because this strategy is very simple and extremely effective in practice, it has become very popular.

Once getting the sentiment scores, the dataset is divided into an 80:20 ratio training and testing set. Then feature extraction method is applied to text TF-IDF for machine learning models.

2.4 Word Embedding

In NLP, the input of the deep learning models needs to be word embedding, word embedding is an approach used to provide dense vector representation of words that capture some context words about their own[22]. In the word embedding technique, each word is represented using dimension. GLOV is a word embedding introduced by Stanford University[23]. In our work, we use the glove as word embedding for deep learning models.

2.5 Model Training and Explanations

By 80 % and 20 %, the data has been divided into training and validation, respectively. Then various deep neural network was created to predict sentiments. We use stratified 5-fold cross-validation (CV) to assure that our models get the correct pattern from the data, then various machine learning models are used as the random forest model, linear SVC, the multinomial naïve Bayes, additionally the logistic regression to predict sentiment. The result shows both the linear SVC model and logistic regression obtain higher accuracy than both the random forest model and the naïve Bayes model.

For the Twitter dataset, a deep neural network was created to predict sentiments. The basic components of a deep neural network are the input layer, the hidden layers, and the output layer. Hidden layers are used in the model to extract features and increase the model's complexity. The output layer gives the probability of each class ('positive', 'negative', or 'neutral'), which must add up to 1.0. As a result, the highest-probability target (sentiment), for the matching tweet is the result of the prediction.

As shown in Table 4, the architecture and hyperparameters of the deep learning models for dataset 1 and dataset 2, the first simple RNN model consists of three layers, one layer is the embedding layer with vocabulary size equal to word index plus 1and output size 100, the RNN layer is followed by embedding layer with 100 units, the output layer for dataset1 is dense layer with 3 neurons softmax activation function, while output layer for datase2 is a dense layer with 1 neuron sigmoid activation function. LSTM, GRU, and Bi-directional models have the same embedding input layer that uses the embedding matrix of the glove, the dropout layer was used to avoid the overfitting issue, they were fitted with 20 epochs and compiled with binary CrossEntropy loss function for dataset 1 and compiled with categorical CrossEntropy loss function for dataset 2, and 'Adam' optimizer was used for the optimization.

After the model's training, the model's output is going to be explained utilizing two methods LIME and SHAP. LIME is an acronym for Local Interpretable Model agnostic Explanations. It is a technique for describing Machine Learning model predictions that were created by Marco Ribeiro in 2016[24] LIME belongs to model agonistic approaches that are created to be universally applicable. They are flexible enough to operate simply based on matching a model's input to its outputs. LIME approximates an opaque model locally, in the area of the prediction we're interested in explaining, by generating either a linear model or a decision tree around the opaque model's predictions, and then utilizing the resultant model as a surrogate to explain the more complicated one[25]. LIME technique is based on explanation by simplification, there are other techniques based on feature relevance like SHAP (SHaShapleyditive exPlanations)[26]. The goal is to create a linear model around the instance to be

Table 4. The architecture of the proposed deep neural models

Layers	RNN	LSTM	GRU	Bi-directional RNN
Input layer	Embedding(len(word_index)+1,100)	Embedding(len(word_index) + 1,300, weights=[embedding_matrix])		
Hidden layers	Simple RNN(100)	LSTM (100, dropout=0.3, recurrent_dropout=0.3)	SpatialDropout1D(0.3)) GRU(300)	Bidirectional (LSTM(100, dropout=0.3, recurrent_dropout=0.3))
Output layer for dataset1	Dense (3, activation=softmax)) Loss equal CategoricalCrossEntropy' Optimizer equal 'adam' epochs equal 20			
Output layer for dataset2	Dense (1, activation='sigmoid')) Loss equal 'BinaryCrossEntropy' Optimizer equal 'adam' epochs equal 20			

described, then interpret the coefficients as the importance of the feature. LIME and SHAP are similar; but, SHAP has a set of useful theoretical capabilities. Its mathematical foundation is based on Shapley values and coalitional Game Theory (Shapley, 1952).

3 Results and Discussion

In this section, we show the classification result and how XAI results remove ambiguous predictions by applying a variety of XAI to help different stakeholders.

3.1 Experiment setup

For our experiments, we use Google CoLab as a platform. It's a cloud-based Python notebook platform for Python versions higher than 3. In only a few lines of code, you can import an image dataset, train an image classifier on it, and test the model with CoLab. CoLab notebooks run code on Google's cloud servers, allowing you to take advantage of Google hardware, such as GPUs and TPUs, regardless of your machine's capabilities.

3.2 Classification Result for Machine Learning Models

Table 5 shows the average accuracy for machine learning models before and after data balancing, also the average accuracy for the second dataset, we use another evaluation metric, the f1 score on dataset1 before balancing because accuracy is not an appropriate measure to evaluate model performance, it is observed that balancing data improves the performance of the random forest classifier but the other classifiers don't improve, the performance of the SVM decreases in dataset2 as data amount increases, but the naïve Bayes model works well with dataset2. Not in all cases, opaque models give the highest accuracy, for example, the random forest model is considered an opaque model, and it is less accurate than the models. Figure 4 shows the confusion matrix for Linear SVC after and before data balancing. The trained model's average accuracy is 86 percent for balanced data.

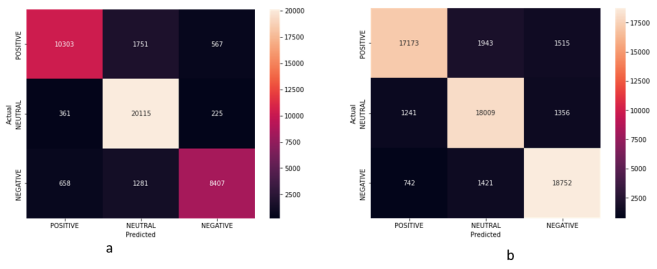

Fig 4. (a) Confusion matrix for Linear SVC with imbalanced data. (b) Confusion matrix for Linear SVC with balanced data

Table 5. Accuracy mean after and before data balancing

Model name	Balancing data	Accuracy AVG for dataset1	F1 score AVG for dataset1 before data balancing	Accuracy AVG for dataset2
Linear SVC	unbalanced	0.88	0.87	0.76
	balanced	0.86		
Logistic regression	unbalanced	0.88	0.87	0.78
	balanced	0.84		
Multinomial NB	unbalanced	0.79	0.77	0.76
	balanced	0.77		
Random Forest	unbalanced	0.47	0.21	0.68
	balanced	0.73		

3.3 Classification Result For deep neural network

Due to the imbalanced nature of the first dataset, models are evaluated using the f1 score, the experimental result is obtained in Table 6, According to the results, the Bi-directional RNN performed better than other models, with an accuracy of 97.9% for the dataset 1 and an f1 score of 79% for dataset 2. Both LSTM and GRU provide very similar outcomes, however, GRU has a lower training cost than LSTM. Bi-directional is similarly highly expensive to train but produces higher performance. Figures 5 and 6 show the model's performance (accuracy and loss, respectively) for different epochs. It is observed that overall deep neural networks achieve better accuracy than machine learning models.

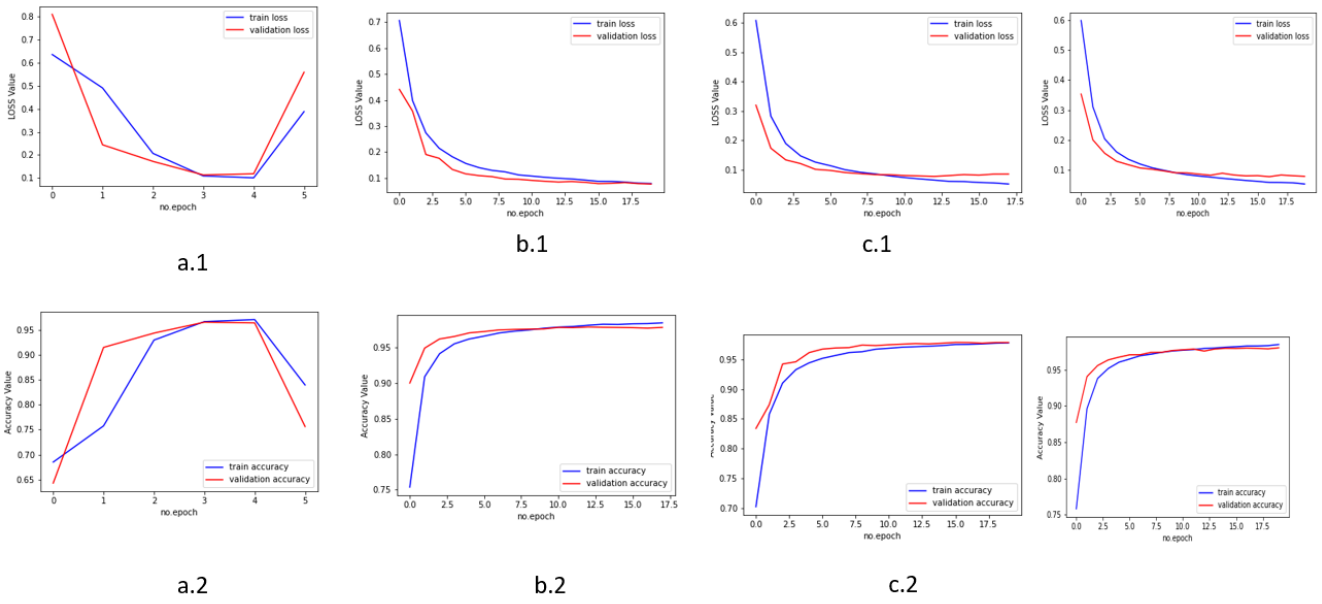

Fig 5. Model performance for each model in dataset1: a 1,2 RNN loss and accuracy, b 1,2 LSTMloss and accuracy, c 1,2 GRU loss and accuracy, d 1,2 bi-directional RNN lossand accuracy

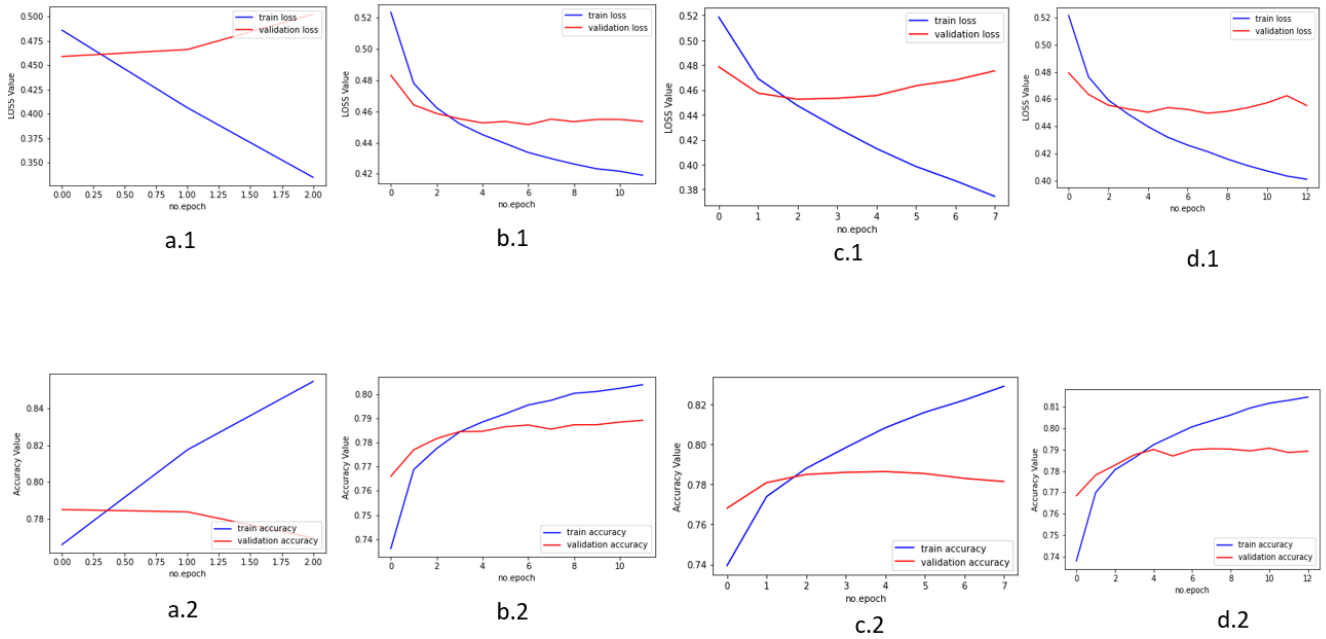

Fig 6. Model performance for each model in dataset2: a 1,2 RNN loss and accuracy, b 1,2 LSTM loss and accuracy, c 1,2 GRU loss and accuracy, d 1,2 bi-directional RNN loss and accuracy

Table 6. Accuracy and f1 score for each deep neural model

Deep neural model	Accuracy for Dataset2	F1 score for Dataset1
Simple RNN Model	78.4%	69.4%
LSTM Model	78.8%	97.4%
GRU Model	78.6%	97.5%
Bi-Directional RNN's Model	79%	97.9%

3.4 Model Interpretation –XAI

In this section, we show the results of LIME and SHAP techniques on opaque models (SVM model and deep neural models) and the transparent model (logistic regression). Figure 7a, shows the output result for the negative tweet, It can be seen that words like 'sick' increase the likelihood of a negative tweet, but their impact on positive tweets is small. In Figure 7b, the original prediction for this tweet to be negative is 60%, after removing the most feature that contributes to classification, like 'stupid', the prediction became 13%.

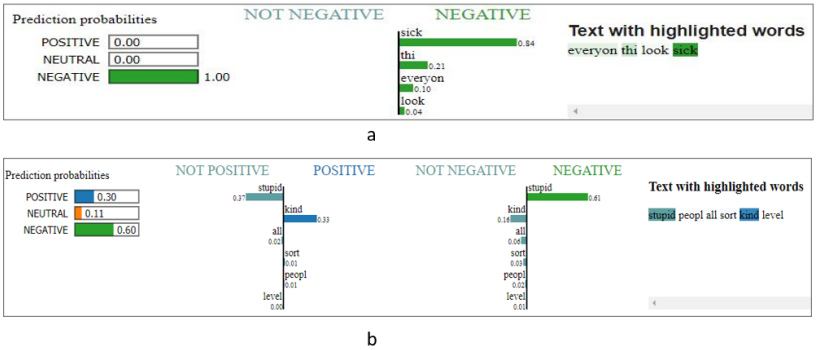

Fig 7. Explanation by LIME for two instances

In Figure 8 we use the SHAP technique to extract the features that impact the classification result. Words like 'love' and 'more' have a great impact on the positive and neutral classes, while their impact on the negative category is very little.

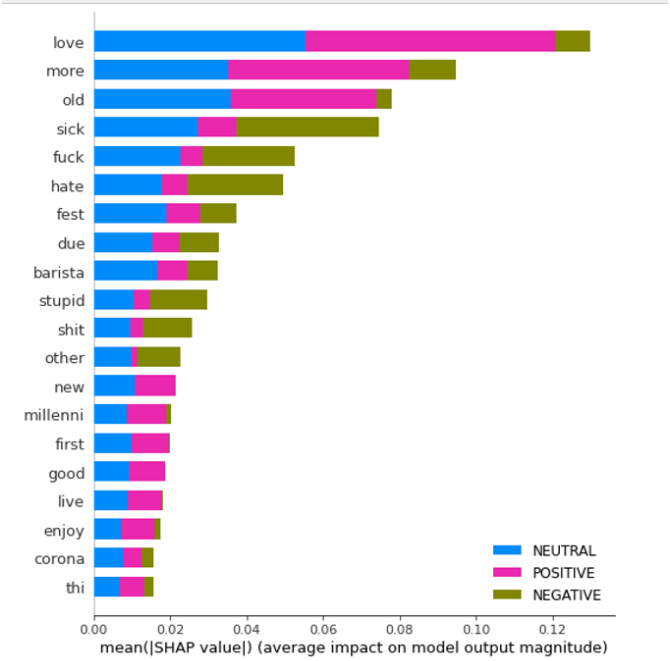

Fig 8. The most features that impact the classification result by SHAP technique

3.5 Comparison with State-of-the-Art Studies

The study in [9] developed an Explainable Sentiment Analysis (XSA) application for Twitter data and MLP achieves performance better than machine learning models, Similarly, using MLP, the study in [10] developed an explainable DL-based sentiment analysis model to predict tweets. This study uses a fuzzy rule-based model that achieves an accuracy of 79% [11]. In this study [12] using an article about financial news, they developed an explainable model for stock prices, also in this study [14] the researchers developed sentiment classification using a BERT classifier that achieves better performance than LSTM. This study [15] developed a classifier for sentiment classification related to covid-19 vaccination tweets without explaining predictions, similarly, the study [16] developed a hybrid deep model for sentiment classification related to covid-19 tweets also without explanation for prediction results. Our approach is better than other studies' in accuracy measurement as shown in comparison Table 7.

Table 7. Performance comparison with previous studies

Performance	XAI technique	Model	Dataset size	Study
Accuracy 79%	Not used	Deep learning(fuzzy-rule-based model)	Two datasets D1 226668 tweets D2 23000 tweets	[11] 2020
F1-Score SVM (75%) RF (74%) XGBoost (79%) MLP (81%)	LIME and SHAP	SVM RF XGBoost MLP	Electronic products dataset(1467)	[9] 2021
Accuracy 75.6 %	LIME	Feedforward neural network (with two hidden layers)	Stanford sentiment140 dataset 1,045,576 tweets) 80%,20%	[10] 2021

Continued on next page

Table 7 continued

Accuracy LSTM (88.73%) LSTM-CNN(74.76%)	LIME	LSTM LSTM-CNN	Financial datasets from PLUS combined with Yahoo finance data(19.736)	[12] 2021
Accuracy CNN and LSTM (97%)	Not used	Naive Bayesian, Adaboost, -nearest neighbor and other machine learning methods CNN, LSTM, RNN, and GRU	COVID Fake News Dataset 10202 news from multiple social media platforms	[8] 2021
F1-score 85%	Not used	BERT classifier	Covid-19 tweets (44955 tweets)	[14] 2022
Accuracy 95%	Not used	Hybrid deep models(LSTM-GRNN)	COVID-19 vaccination	[15] 2022
Accuracy 87.38%	Not used	hybrid Deep Learning (DL) model BiLSTM+CNN	Covid-19 Tweets data(~128000 tweets)	[16] 2022
Accuracy for D2 LSTM and GRU 78% Bi-directional RNN 79%	LIME and SHAP	Machine learning (LR, SVM, Naïve Bayes, RF) Deep learning (RNN, LSTM,GRU,Bi-directional RNN)	Two dataset D1 Covid-19 tweets (488570 tweets) D2 Stanford sentiment140 dataset (1600000 tweets)	Our study

The following are the primary elements that contribute to greater accuracy results: -

1. Preprocessing contributes to improving data quality, since strong data quality leads to good accuracy values and good explanation.
2. Balancing data for machine learning models to get not biased accuracy.
3. Adjusting the hyper parameters' values, such as the number of layers, number of the epoch, and so on, takes a substantial amount of time and effort because it is a trial-and-error experiment.

4 Conclusion

AI systems are considered a black-box approach because it makes critical judgments in an obfuscated manner without explaining the reasoning behind them, especially in areas where we do not wish to delegate decision-making to machines, so, our proposed framework, provides a transparent decision support system that can classify negative words by various machine learning models that the best model obtains average accuracy 86% is SVM model, and deep-neural models obtain accuracy 78.4% for RNN model,78.8% for LSTM model,78.6% for GRU model, and 79% for Bi-directional RNN model, and support its decision with the reasons behind the decision, by applying XAI technology. The lack of transparency in prediction results led to the decision-maker not trusting the model. So our presented work opens the black box of the AI system, this system assists many stakeholders with explanations and insights into user opinion. From our experiment, we reach that accuracy is not everything to persuade the decision-makers with the prediction, accuracy is a crucial measure to evaluate, but it does not always provide the complete picture. Our future work is on how to retrain the model with the consideration of XAI results. And we will implement the transformer model to obtain higher performance.

References

1) Giachanou A, Crestani F. Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *J ACM Comput Surv*. 2016;49(2). Available from: https://doi.org/10.1145/2938640.
2) Pang B, Lee L. Opinion Mining and Sentiment Analysis. . *Foundations and Trends® in Information Retrieval*. 2008;2. Available from: https://doi.org/10.1561/1500000011.
3) Duncan B, Zhang Y. Neural networks for sentiment analysis on Twitter. *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*. 2015;2015:6–8.
4) Tang D, Qin B, Feng X, Liu T. Effective LSTMs for Target-Dependent Sentiment Classification2016 dec. In: The COLING 2016 Organizing Committee. 2016. Available from: https://aclanthology.org/C16-1311.
5) Cho K, Van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. 2014. Available from: https://aclanthology.org/D14-1179.
6) Yang CC. Explainable Artificial Intelligence for Predictive Modeling in Healthcare. *Journal of Healthcare Informatics Research*. 2022;6(2):228–239. Available from: https://doi.org/10.1007/s41666-022-00114-1.

7) Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*. 2018;6:52138–52160. Available from: https://doi.org/10.1109/ACCESS.2018.2870052.
8) Bangyal WH, Qasim R, Rehman NU, Ahmad Z, Dar H, Rukhsar L, et al. Detection of Fake News Text Classification on COVID-19 Using Deep Learning Approaches. *Computational and Mathematical Methods in Medicine*. 2021;2021:1–14. Available from: https://doi.org/10.1109/ACCESS.2018.2870052.
9) Cirqueira D, Almeida F, Cakir G, Jacob A, Lobato F, Bezbradica M, et al. Explainable Sentiment Analysis Application for Social Media Crisis Management in Retail. *Proceedings of the 4th International Conference on Computer-Human Interaction Research and Applications*. 2020. Available from: https://doi.org/10.5220/0010215303190328.
10) Malhotra D, Saini P, Singh AK. Explaining Deep Learning-Based Classification of Textual Tweets. Data Analytics and Management. Singapore; Singapore. Springer. 2021. Available from: https://doi.org/10.1007/978-981-15-8335-3-18.
11) Chakraborty K, Bhatia S, Bhattacharyya S, Platos J, Bag R, Hassanien AE. Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. *Applied Soft Computing*. 2020;97:106754–106754. Available from: https://doi.org/10.1016/j.asoc.2020.106754.
12) Gite S, Khatavkar H, Kotecha K, Srivastava S, Maheshwari P, Pandey N. Explainable stock prices prediction from financial news articles using sentiment analysis. *Peer Journal of Computer Science*. 2021;7:e340–e340. Available from: https://doi.org/10.7717/peerj-cs.340.
13) Das S, Das D, Kolya AK. Sentiment classification with GST tweet data on LSTM based on polarity-popularity model. *Sādhanā*. 2020;45(1). Available from: https://doi.org/10.1007/s12046-020-01372-8.
14) Wisesty UN, Rismala R, Munggana W, Purwarianti A. Comparative Study of Covid-19 Tweets Sentiment Classification Methods. *2021 9th International Conference on Information and Communication Technology*. 2021;2021:3–5.
15) Reshi AA, Rustam F, Aljedaani W, Shafi S, Alhossan A, Alrabiah Z, et al. COVID-19 Vaccination-Related Sentiments Analysis: A Case Study Using Worldwide Twitter Dataset. *Healthcare*. 2022;10(3):411–411. Available from: https://doi.org/10.3390/healthcare10030411.
16) Kumar V. Spatiotemporal sentiment variation analysis of geotagged COVID-19 tweets from India using a hybrid deep learning model. *Scientific Reports*. 2022;12(1). Available from: https://doi.org/10.1038/s41598-022-05974-6.
17) Kaggel. Sentiment140 dataset with 1.6 million tweets. . Available from: https://www.kaggle.com/datasets/kazanova/sentiment140.
18) Lamsal R. Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence*. 2021;51(5):2790–2804. Available from: https://doi.org/10.1007/s10489-020-02029-z.
19) Dataport. . Available from: https://ieee-dataport.org/access-covid-19-datasets.
20) Padurariu C, Breaban ME. Dealing with Data Imbalance in Text Classification. *Procedia Computer Science*. 2019;159:736–745. Available from: https://doi.org/10.1016/j.procs.2019.09.229.
21) Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002;16:321–357. Available from: https://doi.org/10.1613/jair.953.
22) Jiao Q, Zhang S. A Brief Survey of Word Embedding and Its Recent Development. *2021 IEEE 5th Advanced Information Technology Electronic and Automation Control Conference (IAEAC)*. 2021;2021:12–14. Available from: https://doi.org/10.1109/IAEAC50856.2021.9390956.
23) Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. 2014. Available from: https://aclanthology.org/D14-1162.
24) Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. Association for Computational Linguistics. 2016;p. 1135–1179.
25) Arrieta AB, Díaz-Rodríguez N, Ser JD, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*. 2020;58:82–115. Available from: https://doi.org/10.1016/j.inffus.2019.12.012.
26) Lundberg SM, Lee SI, Guyon I, Luxburg UV, Bengio S, Wallach H, et al. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. 2017;30:4765–74.