Journal of Big Data

RESEARCH Open Access

Sentiment analysis classification system using hybrid BERT models

Amira Samy Talaat[1*]

*Correspondence:
amtalat@yahoo.com

[1] Computers and Systems Department, Electronics Research Institute, Cairo 12622, Egypt

Abstract

Because of the rapid growth of mobile technology, social media has become an essential platform for people to express their views and opinions. Understanding public opinion can help businesses and political institutions make strategic decisions. Considering this, sentiment analysis is critical for understanding the polarity of public opinion. Most social media analysis studies divide sentiment into three categories: positive, negative, and neutral. The proposed model is a machine-learning application of a classification problem trained on three datasets. Recently, the BERT model has demonstrated effectiveness in sentiment analysis. However, the accuracy of sentiment analysis still needs to be improved. We propose four deep learning models based on a combination of BERT with Bidirectional Long ShortTerm Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU) algorithms. The study is based on pre-trained word embedding vectors that aid in the model fine-tuning process. The proposed methods are trying to enhance accuracy and check the effect of hybridizing layers of BIGRU and BILSTM on both Bert models (DistilBERT, RoBERTa) for no emoji (text sentiment classifier) and also with emoji cases. The proposed methods were compared to two pre-trained BERT models and seven other models built for the same task using classical machine learning. The proposed architectures with BiGRU layers have the best results.

Keywords: Intelligent systems, Sentiment classification, Machine learning system, Emotion classification, Artificial intelligence, Deep learning, BERT model, Data science, Machine learning application

Introduction

Sentiment Analysis (SA) is a branch of Natural Language Processing (NLP) that focuses on analysing people's views, feelings, and emotions. SA is a multi-step procedure that includes data retrieval, extraction, preprocessing, and feature extraction. With the fast expansion of social media comments in numerous industries, there is a strong need to stay aware of this huge volume of internet data and extract useful information automatically. In this task, sentiment analysis models play an important role. Many sectors, including political challenges, marketing, public policy, disaster management, and public health, rely on emotion detection [1]. The use of emotion recognition software based on images and facial expressions is widespread [2–5]. Human action recognition has been used to recognise hand gestures [6]. Human–computer interaction can also be used to investigate emotion recognition models [7–9]. Textual data from social networks that is emotionally rich can be handled for a wide range of real-world uses [10–12].

Pang et al. [13] were the first to research sentiment analysis using a machine-learning technique. They conducted tests on a movie review dataset using supervised classifiers such as Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM). The classifiers perform poorly in classifying sentiment when compared to standard text classification. The major explanation might be that they typically apply to traditional text categorization methodologies, in which words in a document are established as a bag of words (BOW) notion. BOW does not store grammatical structures, word order, or semantic relations between words, which are critical characteristics for SA [14]. Many studies used machine learning techniques for SA, like Bayesian Networks [15], Naive Bayes [16], SVM [17, 18], and decision trees [19, 20], as well as Artificial Neural Networks [21]. The low scalability of new data was a drawback since it needed the availability of labeled data, which may be costly or even prohibitively expensive [22].

Due to the advantages of DL, different researchers have used different types of SA for Artificial Neural Networks. Among these benefits are [23]: Automated Feature Generation by creating new features from a small set of features in the training data so they can generalize better; and scalability, where DL analyses large volumes of data and does numerous computations, which is cost and time effective. [24] proposed a dynamic neural network with the author performing experiments with different CNN variations and a max-pooling dynamic K operator. Socher [25] utilised this model to produce comparable results to DCNN in another study. [26] provided a comprehensive overview of the most common CNN, LSTM, and other deep learning approaches for aspect-level sentiment analysis. Liu and Shen [27] introduced the Gated Alternate Neural Network (GANN) as a separate neural network model to solve several drawbacks in previously presented models involving noise in collecting meaningful sentiment expressions and produced improved results. For financial sentiment research, Akhtar [28] proposed using MLP to blend deep learning and feature-based approaches. They created several deep learning models based on CNN, LSTM, and GRU. In another study, Pan [29] employed a hybrid strategy to localise the text using MLP and got good results. Karakuş [30] analyzed the performance of numerous deep learning models in the training and testing stages by building model variations with varying sizes of the layers as well as the approach of word embedding. They improved their outcomes by using LSTM, CNN, BILSTM, and CNNLSTM. Furthermore, they classified using MLP and achieved a 78% test accuracy on a movie review dataset. In another research, the review dataset from the IDMB is used for sentiment classification, and MLP is used in the training process, and the results are promising [31]. The majority of deep learning sentiment categorization work has been done on reviewed datasets to achieve better outcomes. A hybrid method of sentiment classification using Twitter datasets should be used to acquire the best possible outcomes during the testing phase.

We can use the advantages of both ML and DL to overcome the disadvantages of both approaches, resulting in more accurate and less computationally expensive solutions [32]. Hybrid models show an increase in accuracy and give higher efficiency and performance for systems.

Nimmi [33] presented the AVEDL (Average Voting Ensemble Deep Learning mode) Model, which analyses the contents of calls landed in the emergency response assistance system using pre-trained transformer-based models BERT, DistilBERT, and RoBERTa (ERSS). By achieving a Macro-average F1-score of 85.20% and an accuracy of 86.46%, the AVEDL Model beats typical deep learning and machine learning models. The disadvantage of this model was that it did not account for slang as well as speech when computing COVID-19-relevant emotions.

Adoma [34] investigated the performance of pre-trained transformer models BERT, RoBERTa, DistilBERT, and XLNet in identifying emotions from the text. On the ISEAR data, the implemented models are fine-tuned to distinguish between joy, humiliation, guilt, fear, anger, disgust, and sadness. The models were effective in detecting emotions in text. RoBERTa had the highest accuracy of 0.74. The model should be generalized.

For sentiment analysis, Uddagiri [35] integrates RoBERTa with ABSA (Aspect-Based Sentiment Analysis), RoBERTa, and LSTM (Long Short-Term Memory). Using the Twitter Dataset for the Ukraine Conflict. They examined the emotions of Optimism, Sadness, Anger, and Joy. The accuracy was 94.7 percent.

Bansal [36] presented an attribute-based hybrid technique for analysing consumers' intelligence by identifying features using POS tags. The approach must detect more attribute exhaustive subjects with less computing cost. Also, the study should be expanded to include short text categorization and hybrid approaches, as well as eliminating human labeling of attribute specific words in existing lexicons.

Using Long and Short Term Memory (LSTM) models, Ma et al. [37] developed a classification based on certain aspects utilising common sense knowledge. They improve Sentic LSTM by including a mix of LSTM and a recurrent addictive network. Wang [38] implicitly extracted characteristics using hybrid ARM and employed five strategies to use them. Because implicit words are ignored by the context, only explicit aspects were recovered. F-measure achieved 75.51 percent. shortcomings in this method: There are a few aspects of the hybrid association rule mining that are difficult to control in practice.

Zainuddin [39] employed SVM + PCA with the addition of POS tags as feature extractors and attained high accuracies for the STS and STC datasets. The sentiment classifier approach outperformed the existing baseline sentiment classification methods by 76.55, 71.62, and 74.24%, respectively. The approach does not apply to other social media data sources, such as YouTube and Facebook.

In [40] Combining Bangla-BERT with LSTM yields 94.15% accuracy. They should employ a sophisticated deep learning algorithm to work on a richer and more balanced dataset.

The BERT-DCNN model was proposed by [41], which stacks BERT with a dilated convolutional neural network (DCNN) to produce a stronger sentiment analysis model. For Twitter airline sentiment data, the model had an accuracy of 87.1 percent. This strategy is primarily limited to data received from a single source, rather than data obtained from multiple sources.

Tweets can contain a variety of data types, including news, media, retweets, and replies to postings, and they can be structured as audio, video, or images. By permitting and encouraging discussion between numerous parties on a public platform, Twitter allows you to gain fast feedback from users and future clients. Because everyone can see what you're saying on Twitter, it fosters transparency and responsibility in conversation [42].

Kian [43] Where authors applied a RoBERTa-LSTM based model, on the sentiment Airlines dataset and obtained an accuracy of 89.85% without a data augmentation process, and 91.37% when adding a data augmentation process is done to oversample the minority classes. The datasets are split into 6:2:2 for training, validation, and testing with Adam optimizer and a learning rate of 0.00001 with batch size set to 64 and 30 epochs.

The authors proposed a CNN-LSTM architecture and obtained an accuracy of 91.3% on the sentiment Airlines dataset [44]. They have only considered data obtained from an online platform that is in the form of English sentences. Thus, consumer reviews written in other languages does not include in sentiment analysis. They classified consumer sentiment into 2 classes only (positive and negative). They discard neutral sentiment data from their dataset so their classification result is high.

Barakat [45] proposed a novel ULMFit-SVM model to improve sentiment analysis performance. The model demonstrates an accuracy rate of 99.78% on Twitter US Airlines, 99.71% on IMDB, and 95.78% on GOP debate. The sentiment analysis was restricted to the document level. They did not take into account the sentiment at the aspect level. For the Twitter dataset, each one of the three classes (Positive, Negative, Natural) is split separately into 66% training and 33% testing with Adam optimizer and learning rate of 0.004 and 0.01 for fine-tuning with 64 batches.

Tweets often only include a few words with practical significance, and these words are critical in the classification phase. The BERT model has demonstrated effectiveness in the sentiment analysis of tweets. However, the accuracy still needs to be improved. We propose a hybrid BERT based multi-feature fusion short text classification model. The technique is made up of three parts: BERT, BiLSTM, and BiGRU. BERT is used to train dynamic word vectors to improve short text word representation. The BiLSTM and BiGRU help in extracting and learning sentence sequence characteristics. We also examined the impact of changing the number and location of (BiLSTM and BiGRU) layers to enhance the performance.

The contributions of this study are summarised as follows:

- We propose four hybrid innovative deep learning models for emotion classification applied to three datasets. Four models for RoBERTa and four models for DistilBERT are compared to select the best hybrid model, which has the ability to extract contextual information from text.
- BiGRU and BiLSTM networks are used to extract context information from text for the fine-tuning process.
- We employed training models on emoji datasets and then tested the hypothesis of emojis' advantage as a cue in classification by training the same model but eliminating emojis in the preprocessing stage and observing the impact.

The following is how the paper is arranged. The Methodology Section explains the suggested technique for measuring the emotional elements of tweets. The Experiments and Results Section highlights and discusses the most important findings.

Finally, the Conclusion and Future Work Section summarises the findings and discusses future research directions.

Methodology

This section discusses the methodology that we use to create a framework for predicting the emotions of users based on their tweets. The structure of the framework construction is in Fig. 1.

The techniques we used to construct a framework for predicting users' emotions from their tweets are described in this section.

Dataset

The tweets are represented as feature vectors by two BERT models (BERT and BERT-mini) from the HuggingFace website:

1- "Twitter-RoBERTa-Base-Sentiment", which is "BERTBase": This is a RoBERTa-based model that was finetuned on the emotion dataset for sentiment analysis using the TweetEval benchmark after being trained on 58 million tweets. This model is appropriate for use in English. RoBERTa is BERT with more hyperparameter options therefore they called it Robustly optimised BERT during pretraining.

2- "DistilBERT-Base-Uncased-Emotion", which is "BERTMini": DistilBERT is constructed during the pre-training phase via knowledge distillation, which decreases the size of a BERT model by 40% while keeping 97% of its language understanding. It is faster and smaller than any other BERT-based model, and it was fine-tuned on the emotion dataset.

We used three sets of data freely accessible on the Kaggle website to perform a multilabel emotion classification, as shown in Table 1. We train and test each dataset separately. We transform all the datasets into two columns, i.e., text and label. Text for tweet text and a label for representing sentiment, with 3 classes: 2 indicating a positive sentiment,

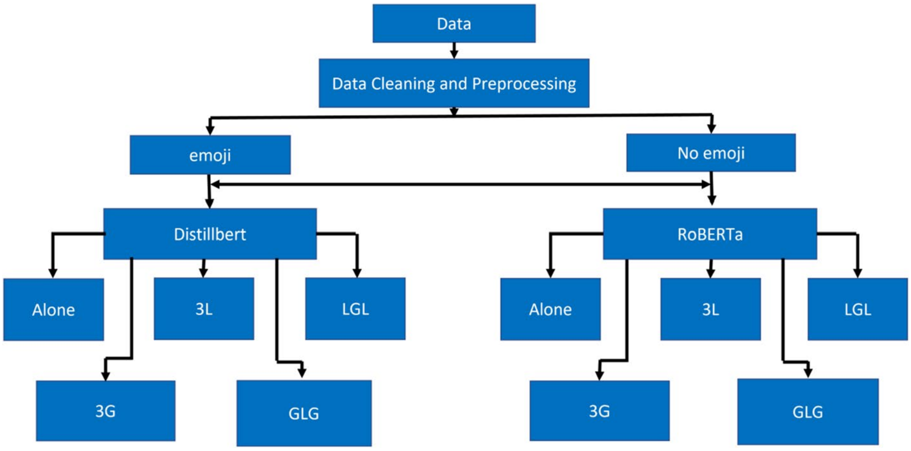


Fig. 1 The structure of the proposed framework

Table 1 Shows the datasets that were used

Dataset	Num of tweets	Num of post tweets	Num of Negative tweets	Num of neutral tweets
Airlines	14,640	2363	9178	3099
CrowdFlower	3804	423	1219	1219
Apple	1630	686	143	801
All	20,074	3472	10,540	6062

1 indicating a neutral tweet or comment, and 0 indicating it is a negative tweet or comment.

1. Airline Sentiment (Airlines) [46]: Analyse how passengers use Twitter to reflect their emotions. A sentiment analysis project concerning airline difficulties. Contributors were requested to categorize tweets into positive, neutral, and negative categories.
2. Apple Sentiment (CrowdFlower) [47]: Tweets with an emotion label that mention Apple computers. Based on tweets including #AAPL, @apple, etc. A look at how people feel about Apple. Contributors were given a tweet and asked if the user thought Apple was positive, neutral, or negative.
3. Apple texts (Apple) [48]: this dataset also includes tweets about Apple computers.

The classical ML approach

Seven classical machine learning approaches are compared with the eight BERT/BiGRU/BiLSTM methods. These methods are the decision tree, k-nearest neighbors, random forest, naive Bayes, support vector machine (SVM), logistic regression, and XGBoost algorithms. In the sentiment classification models, we have used multiple combinations of pretrained BERT Models by stacking them with BiLSTMs and BiGRUs. The main objective of the project is to classify the sentiments as positive, negative, and neutral across multiple datasets.

First, we transform all the datasets into two columns, i.e., text and sentiment. The sentiment has 2 (positive), 1 (neutral), and 0 (negative).

After this, we import the pre-trained BERT classifier models RoBERTa (BERTBase) and DistilBERT (BERTMini) from Hugging Face.

Data cleaning and preprocessing

Now we do some preprocessing on the text, like normalizing Unicode encoding, removing names, trailing whitespaces, hashtags, numbers, punctuation, and URL addresses. Also, emojis are removed in the case of no emoji methods.

After preprocessing the text column, we tokenize each sentence and come up with input ids and attention masks for each line of text.

Proposed models

Now we prepare the actual model by stacking some combinations of BiGRUs and BiLSTMs on different BERT models, as shown in Fig. 2.

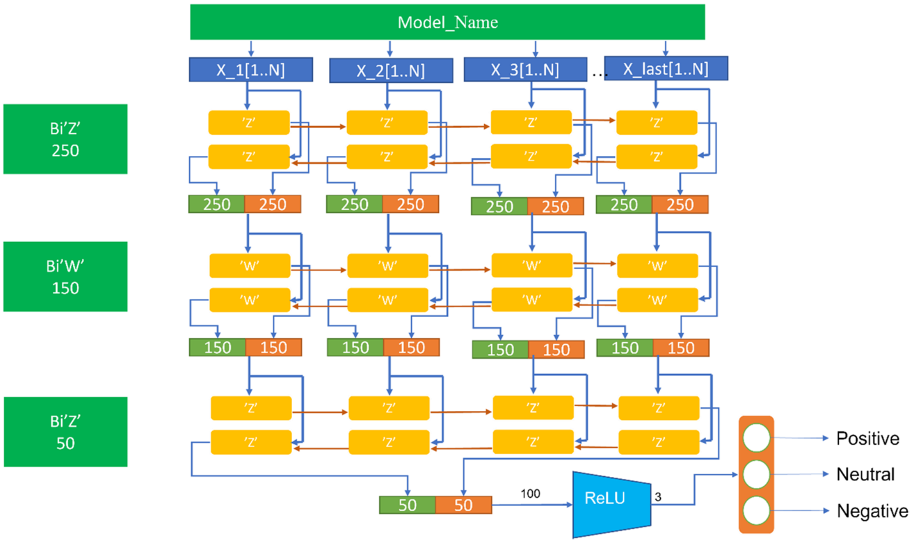


Fig. 2 Hybrid Models Architecture

The following is the eight models used in detail:

- DistilBERT-3G: DistilBERT-3xBiGRU
- DistilBERT-3L: DistilBERT-3xBiLSTM
- DistilBERT-GLG: DistilBERT-BiGRUxBiLSTMxBiGRU
- DistilBERT-LGL: DistilBERT-BiLSTMxBiGRUxBiLSTM
- RoBERTa-3G: RoBERTa-3xBiGRU
- RoBERTa-3L: RoBERTa-3xBiLSTM
- RoBERTa-GLG: RoBERTa-BiGRUxBiLSTMxBiGRU
- RoBERTa-LGL: RoBERTa-BiLSTMxBiGRUxBiLSTM

The parameters of the eight hybrid methods in Fig. 2 are modified as follows:
1. DistilBERT_3xBiGRU (DistilBERT_3G) where:
(Model_name) = DistilBERT,
(Hugging_Name) = ‘Distilbert-Base-Uncased-Emotion’,
(N) = 256,
(’Z’) = (’W’) = ‘GRU’
2. DistilBERT_3xBiLSTM (DistilBERT_3L) where:
(Model_name) = DistilBERT,
(Hugging_Name) = ‘DistilBERT-Base-Uncased-Emotion’,
(N) = 256,
(’Z’)=(’W’) = ‘LSTM’
3. DistilBERT_BiGRUxBiLSTMxBiGRU (DistilBERT_GLG) where:
(Model_name) = DistilBERT,
(Hugging_Name) = ‘DistilBERT-Base-Uncased-Emotion’,
(N) = 256,
(’Z’) = ‘GRU’,
(’W’) =‘LSTM’

4. DistilBERT_BiLSTMxBiGRUxBiLSTM (DistilBERT_LGL) where:
(Model_name) = DistilBERT,
(Hugging_Name) = ‘DistilBERT-Base-Uncased-Emotion’,
(N) = 256,
(’Z’) = ‘LSTM’,
(’W’) =‘GRU’
5. RoBERTa_3xBiGRU (RoBERTa_3G) where:
(Model_name) = RoBERTa,
(Hugging_Name) = ‘Twitter-RoBERTa-Base-Sentiment’,
(N) = 768,
(’Z’) = (’W’) = ‘GRU’
6. RoBERTa_3xBiLSTM (RoBERTa_3L) where:
(Model_name) = RoBERTa,
(Hugging_Name) = ‘Twitter-RoBERTa-Base-Sentiment’,
(N) = 768,
(’Z’) = (’W’)= ‘LSTM’
RoBERTa_BiGRUxBiLSTMxBiGRU (RoBERTa_GLG) where:
(Model_name) = RoBERTa,
(Hugging_Name) = ‘Twitter-RoBERTa-Base-Sentiment’,
(N) = 768,
(’Z’) = ‘GRU’,
(’W’) = ‘LSTM’
7. RoBERTa_BiLSTMxBiGRUxBiLSTM (RoBERTa_LGL) where:
(Model_name) = RoBERTa,
(Hugging_Name) = ‘Twitter-RoBERTa-Base-Sentiment’,
(N) = 768,
(’Z’) = ‘LSTM’,
(’W’) = ‘GRU’
These are the main steps of the eight hybrid methods:

- We first instantiate a (Model_Name) pretrained model from (Hugging_Name) Hugging Face.
- The output of this (Model_Name) model has (N) features, so we create one (Bi’Z’) layer and pass these (N) features inside along with the 250 hidden features.
- From this (Bi’Z’) layer, we got 500 features as an output.
- We create a second (Bi’W’) layer and pass these 500 features inside along with the 150 hidden features.
- From this (Bi’W’) layer, we got 300 features as an output.
- We create a third (Bi’Z’) layer and pass these 300 features inside along with the 50 hidden features.
- From this (Bi’Z’) layer, we got 100 features as an output.
- Then we apply the ReLU activation function and convert that (Bi’Z’) layer into a linear layer with an output of three classes: 2, 0, and 1 (positive, negative, and neutral).

The models are initialized with these predefined parameters:

Optimizer is AdamW, the learning rate is 5e-5, the epsilon value is 1e-8, the number of epochs is 10, and the batch size is 16. Now we set the loss as 'Cross Entropy Loss' and put the model on training. The train function will evaluate the training loss after each batch and the validation loss, and the accuracy after each epoch.

All eight models are run again after removing emojis for the three datasets.

Experiments and results

In this study, we applied three datasets to train the classifier, validate the system, and test it. The data was split into three groups: 80% for training, 10% for validation, and 10% for testing. The work was completed on Kaggle using a 2.3 GHz Intel(R) Xeon(R) CPU, an Nvidia P100 GPU, and 16 GB of RAM.

Evaluation criteria

The overall classification efficiency was done using a variety of evaluation factors.

To assess the sentiments of the text based on neutral, negative, and positive classes, four evaluation criteria were established: accuracy criteria (Eq. (1)), recall (Eq. (2)), precision (Eq. (3)), and F-measure (Eq. (4)).

Four functional accuracy metrics were considered: false positive (FP), true positive (TP), false negative (FN), and true negative (TN).

The following are the testing parameters that were used to analyze the performance of our suggested system:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Percision = \frac{TP}{TP + FP} \tag{3}$$

$$F1 = \frac{2 * Percision * Recall}{(Percision + Recall)} \tag{4}$$

Results and charts

The three datasets are trained on classical ML methods to get the best testing classical classifiers, which are Logistic Regression and SVM, with accuracies of 80.62, 73.73, and 84.05 for the Airlines, CrowfFlower, and Apple datasets respectively, as shown in Table 2.

We run the eight models with and without emojis. All the results are shown in Table 3. A comparison graph between eight models is shown in Figs. 3,4,5. The ROC curves of the best four methods (DistilBERT with emojis, DistilBERT without emojis, RoBERTa with emojis, and RoBERTa without emojis) for each dataset are shown in Figs. 6,7,8.

Table 2 Testing accuracies of classical methods techniques trained on three datasets

Dataset	Decision Tree	KNN	Logistic Regression	SVM	Naive Bayes	Random Forest	XGBoost
Airlines	66.67	73.16	80.62	80.33	68.56	76.91	76.16
CrowdFlower	63.57	69.18	73.73	73.12	72.15	71.8	71.45
Apple	71.57	78.32	81.6	84.05	79.75	82.41	81.39

Table 3 Accuracies of airlines, crowdflower, and apple datasets

Model		with emojis			without emojis		
		Accuracy Airlines	Accuracy CrowdFlower	Accuracy Apple	Accuracy Airlines	Accuracy CrowdFlower	Accuracy Apple
DistilBERT	-	83.5	78.71	84.97	83.4	77.92	86.2
	3G	81.8	76.48	83.74	82.17	76.61	87.12
	3L	81.9	74.9	83.13	81.83	76.74	82.82
	GLG	83.74	80.42	85.89	83.47	79.24	88.04
	LGL	82.55	78.71	86.81	83.27	78.98	87.42
RoBERTa	–	85.72	82.39	91.72	85.69	79.63	90.18
	3G	86	79.63	91.1	85.72	79.63	91.72
	3L	85.66	82.26	90.18	85.66	81.34	89.57
	GLG	85.28	81.21	91.41	85.93	80.55	90.18
	LGL	85.52	81.34	89.88	84.97	80.16	90.49

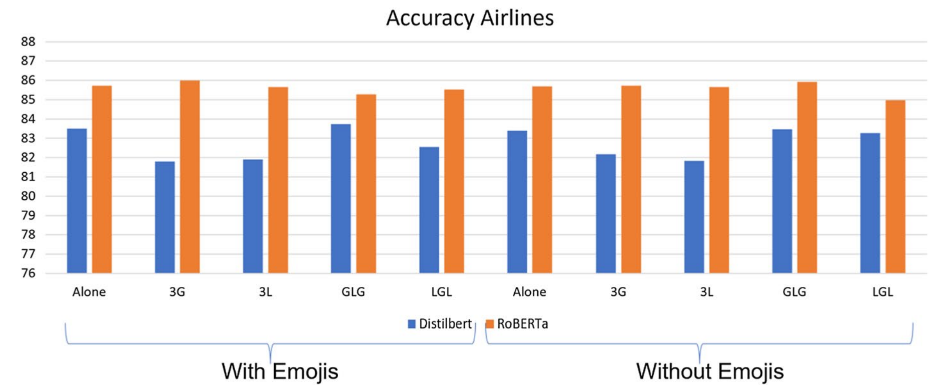


Fig. 3 Airlines Accuracy comparison between models of DistilBERT and RoBERTa for Emojis and no Emojis

We found better accuracies by hybridizing BIGRU layers with (DistilBERT and RoBERTa) because BIGRU is more efficient due to their simpler structure than BILSTM.

For the Airlines dataset Table 3 and Fig. 3, the best accuracy method is GLG with 83.74% and 83.47% for DistilBERT in both emoji and no emoji cases, respectively. In DistilBERT: BiGRU layers work well with large datasets. Here we have two BiGRU layers and one BiLSTM (GLG).

For the Airlines dataset, RoBERTa-3G is the best accuracy method with 86% for the emoji case but when we remove the emoji, GLG is the best accuracy model with 85.93%. As in DistilBERT, RoBERTa works also great with BiGRU layers for large datasets. Here

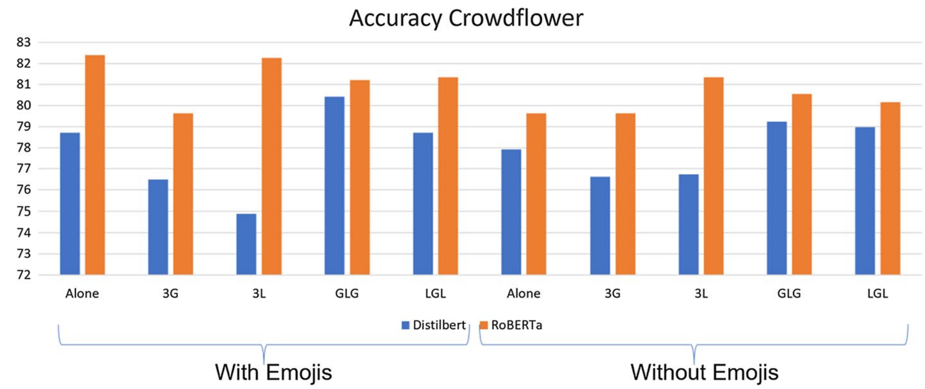


Fig. 4 Crowdflower Accuracy comparison between models of DistilBERT and RoBERTa for Emojis and no Emojis

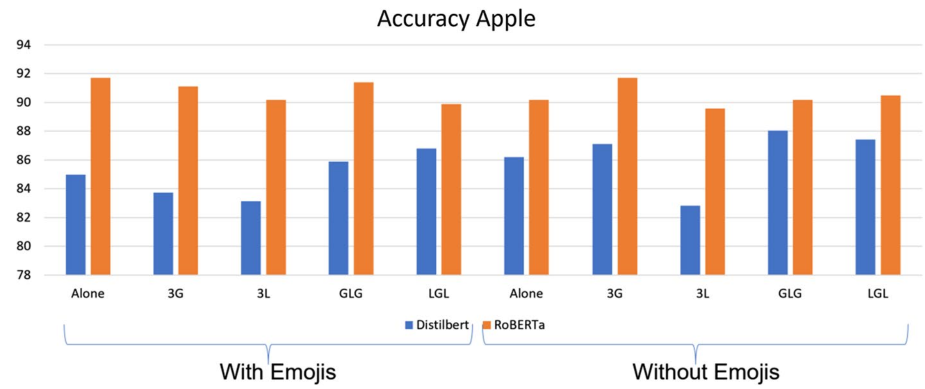


Fig. 5 Apple Accuracy comparison between models of DistilBERT and RoBERTa for Emojis and no Emojis

we have three BiGru layers (3G) for emojis and two BiGru layers with one BiLSTM (GLG) for no emojis.

Airlines without emojis' accuracy are affected and dropped compared to DistilBERT and RoBERTa with emojis.

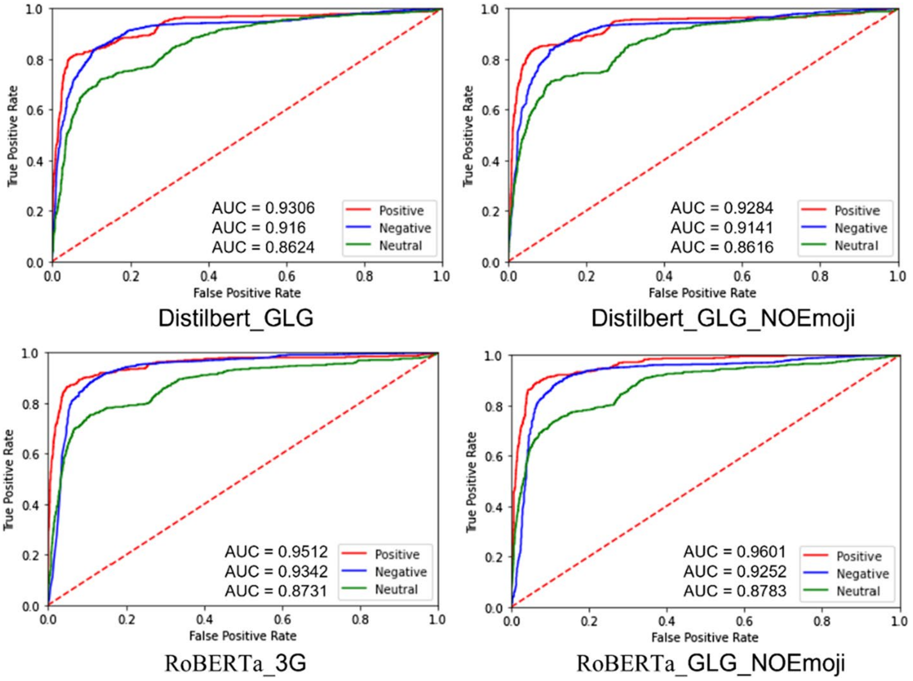
This is expected since the model was learning from emojis, which is undesired behavior (we want a text sentiment classifier).

For the Crowdflower dataset Table 3 and Fig. 4, the best accuracy method is GLG with 80.42% and 79.24% for DistilBERT in both emoji and no emoji cases, respectively. In DistilBERT: BiGRU layers work well with medium size datasets. Here we have two BiGRU layers and one BiLSTM (GLG).

For the Crowdflower dataset, RoBERTa alone has the best accuracy with 82.39% for the emoji case, but when we remove the emoji 3L is the best accuracy model with 81.34%, and GLG is the second model with an accuracy of 80.55%.

Crowdflower without emojis' accuracy is affected and dropped compared to DistilBERT and RoBERTa with emojis. RoBERTa works well with three BiLSTM layers for a medium dataset with no emoji (3G).

For the Apple dataset Table 3 and Fig. 5, which is the smallest dataset, the best accuracy method is LGL with 86.81%, and GLG is the second-accuracy model with 85.89%



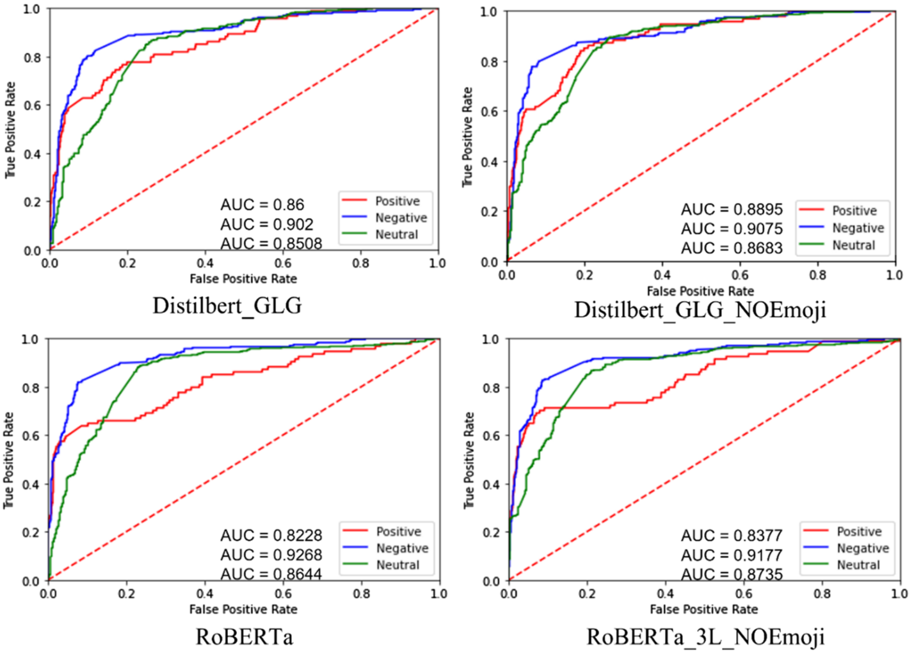
Distilbert_GLG

Distilbert_GLG_NOEmoji

RoBERTa_3G

RoBERTa_GLG_NOEmoji

Fig. 6 Is the ROC curves of best Accuracies of Airlines dataset



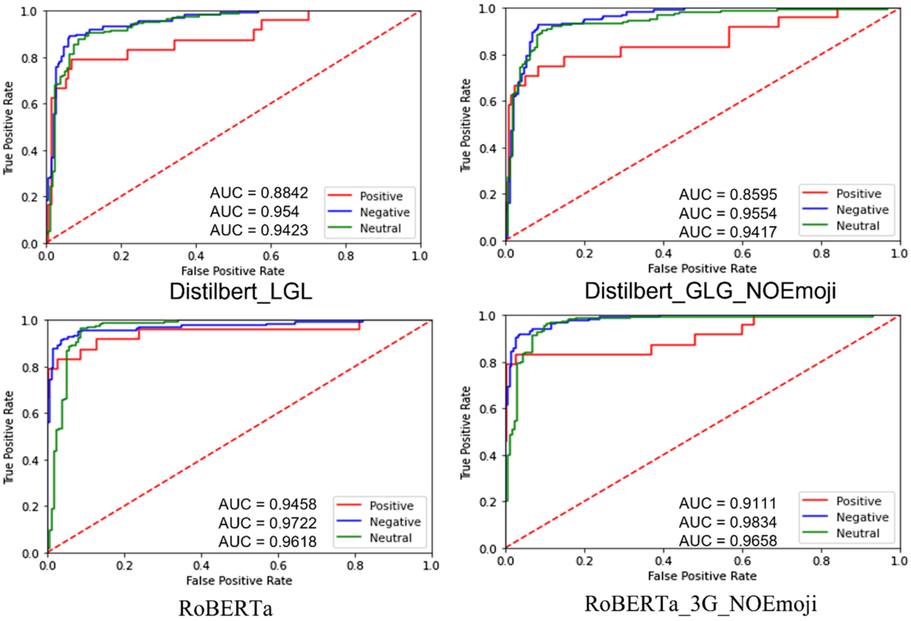
Distilbert_GLG

Distilbert_GLG_NOEmoji

RoBERTa

RoBERTa_3L_NOEmoji

Fig. 7 Is the ROC curves of best Accuracies of Crowdflower dataset



Distilbert_LGL

Distilbert_GLG_NOEmoji

RoBERTa

RoBERTa_3G_NOEmoji

Fig. 8 The ROC curves of best Accuracies of Apple dataset

for DistilBERT with emoji. In DistilBERT's no emoji case, GLG has the best accuracy at 88.04%. In DistilBERT: Double BiLSTM layers with one BiGRU (LGL) work well with a small dataset with emoji but again two BiGRU layers and one BiLSTM (GLG) layers are better for no emoji.

For the Apple dataset, RoBERTa alone is the best for emoji cases with an accuracy of 91.72%, but when we remove emoji, 3G is the best accuracy model with 91.72%, and GLG is the third accuracy model with 90.18%. For RoBERTa with a small dataset, we have three BiGru layers for emoji (3G).

GRU performance is on par with LSTM, but computationally more efficient because they have a less complex structure.

We proved that hybridizing layers of BiGRU (GLG and 3G) are working better with most DistilBERT datasets except for the small dataset (Apple) with emoji.

While hybridizing layers of BiGRU (GLG and 3G) with RoBERTa is working better with no emoji, especially for large (Airlines) and small (Apple) datasets.

In general, RoBERTa has higher accuracies than DistilBERT, as shown in Figs. 3,4,5 because it is a bigger model that improves the performance.

Overall, for the three datasets, DistilBERT_GLG is the best model with and without emoji except for Apple dataset with emojis. Furthermore, without emoji, RoBERTa_GLG has the highest accuracy for Airlines, the second highest accuracy for CrowdFlower, and the third highest accuracy for the Apple dataset.

So each dataset has a different response due to the different subjects of this dataset, but in general RoBERTa GLG is considered good for the three datasets used.

RoBERTa_GLG without emoji is working better for large datasets than for smaller datasets, where RoBERTa_3G and RoBERTa_3L are better. Hence the RoBERTa model without emoji is working well with our proposed method.

DistilBERT_GLG accuracy is higher than DistilBERT alone, with 1.84% for DistilBERT no emoji in the apple dataset. DistilBERT_GLG accuracy is much higher than in Logistic Regression and SVM.

Of the four DistilBERT models, the best one is GLG, with 88.04% accuracy. Of the four RoBERTa models, the best one is 3G, with 91.72% accuracy.

Comparison of performance with other models

A comparison of the accuracies of previous papers is presented in Table 4. The bold models are the proposed models in this research.

When comparing DistilBERT models only, the DistilBERT_GLG is the best model for all the three datasets in emoji and no emoji cases except for only the smallest dataset Apple with emoji case the GLG is the second-best method after LGL.

GRUs is more efficient than LSTM because they have less complex structure. This efficiency appears in our final results.

For DistilBERT, GLG is working better on our datasets except for the small dataset Apple with emoji.

For RoBERTa without emojis multi layers BIGRU (GLG, 3G) are working better for large (Airline) and small (Apple) datasets.

For RoBERTa with emojis three layers (3G) are working better for large datasets only (Airline).

The combination of BIGRU layers with DistilBERT and RoBERTa enhances the accuracy.

The proposed methods were compared to two pre-trained BERT models and seven other models built for the same task using classical machine learning.

The proposed architecture of hybridizing GLG with DistilBERT has more accuracy than DistilBERT alone by 0.24% to 1.84% for the three datasets.

For Dang et al. [50] they removed the neutral class and calculated the accuracies for datasets with two classes (positive and negative), so their classification result is high.

For Kian [43] The datasets are split into 6:2:2 for training, validation, and testing with Adam optimizer and a learning rate of 0.00001 with the batch size set to 64 and 30 epochs. While in our work we split datasets to 80% for training, 10% for validation, and 10% for testing with AdamW optimizer and learning rate of 5e-5 and batch size 16 and 10 epochs.

Barakat [45] For the Twitter dataset, each one of the three classes (Positive, Negative, Natural) is split separately to 66% training and 33% testing with Adam optimizer and learning rate of 0.004 and 0.01 for fine-tuning with 64 batches. While in our work we split datasets to 80% for training, 10% for validation, and 10% for testing with AdamW optimizer and learning rate of 5e-5 and batch size 16.

Jain [44] They classified sentiment into 2 classes only (positive and negative). They discard neutral sentiment data from their dataset so their classification result is high. We used AdamW and the learning rate is 5e-5 but in this paper, neither the optimizer type nor learning rate value is mentioned to compare with it.

Table 4 Comparison between other approaches and ours

Model	Dataset	Accuracy%	Notes
DistilBERT with emojis	Airlines	83.74	GLG
	Crowdflower	80.42	GLG
	Apple	86.81	LGL
DistilBERT without emojis	Airlines	83.47	GLG
	Crowdflower	79.24	GLG
	Apple	88.04	GLG
RoBERTa with emojis	Airlines	86	3G
	Crowdflower	82.39	–
	Apple	91.72	–
RoBERTa without emojis	Airlines	85.93	GLG
	Crowdflower	81.34	3L
	Apple	91.72	3G
Indrayuni et al. [49]	Apple products	85.76	SVM + GA
Dang et al. [50] (two classes)	Sentiment140	80	Word embeeding-RNN
	Tweets SemEval	85	Word embeeding-RNN
	IMDB Movie Reviews (1)	87	Word embeeding-RNN
	IMDB Movie Reviews (2)	86	Word embeeding-RNN
	Cornell Movie Reviews	76	Word embeeding-RNN
	Book Reviews	76	Word embeeding-CNN
	Music Reviews	76	TF-IDF-DNN
	Tweets Airline	90	Word embeeding-RNN
Kumawat et al. [51]	Twitter US Airline Sentiment	81.2	BERT
		80.8	RoBERTa
Xiang [52]	Twitter collection	76.6	BiLSTM(EPA)
	airline dataset	82	BiLSTM(P)
	IMDB review	82.6	BiLSTM-AT(P)
Shuang [53]	Twitter airlines	83.3	M_ARC
	Yelp	79.1	RC
Janjua et al. [54]	Sanders Twitter Corpus (STC)	78.99	MuLeHyABSC + MLP
	Twitter Airline Sentiment (TAS)	84.09	
	First GOP Debate (FGD)	80.38	
	Apple Twitter Corpus (ATC)	82.37	
	Stanford Twitter Sentiment (STS)	84.72	
Kian [43]	Twitter US Airline Sentiment	85.89	RoBERTa-LSTM
	Twitter US Airline Sentiment Augmented	91.37	
	IMDB	92.96	
	Sentiment140	89.70	
Barakat [45]	Airline	99.78	ULMFit-SVM
Jain [44] (two classes)	Airlinequality Airline Sentiment Data	90.2	CNN-LSTM
	Twitter Airline Sentiment Data	91.3	
Thapa et al. [55]	Twitter	60	VADER
	Reddit	70	
Demotte et al. [56]	CrowdFlower US Airline	82.04	GloVe + shallow capsule network with static routing
	Twitter Sentiment Gold	86.87	

Conclusions and future work

Sentiment analysis is critical in many fields, including business and politics, to understand public sentiment and make strategic decisions.

This paper provided hybridizing methods for developing a deep learning model for tweet emotion classification using multiple labels for three datasets. Many preprocessing phases are adapted, such as removing names, trailing, whitespace, hashtags, and numbers. Tokenize each sentence and come up with input ids and attention masks for each line of text. Pre-trained BERT classifier models are used in RoBERTa (BERTBase), and DistilBERT (BERTMini) hybridized with BiGRU and BiLSTM to give better accuracy. Eight hybrid models are proposed, and DistilBERT-GLG without emoji achieved a 1.84% increase over DistilBERT alone for the Apple dataset. DistilBERT-GLG achieved a 0.24% increase over DistilBERT alone for the Airline Dataset. It seems like the presence or absence of emojis can affect model performance in terms of accuracy. The accuracy went from 80.42% to 79.24% after only removing emojis in the preprocessing step for Distilbert_GLG in the CrowdFlower dataset. Also, the RoBERTa model without emoji is working well with our proposed method.

In a conclusion, GLG is working well for DistilBERT for all datasets with no emoji, and the big and medium datasets with emoji. For Roberta, the models with BiGRU layers have better performances than others, especially for large and small Datasets. The combination of BIGRU layers with DistilBERT and RoBERTa enhances the accuracy.

We would like to extend this work in the future by combining it with classical text classification algorithms. To increase the performance of the present system, the most up-to-date approaches to feature extraction and feature selection will be integrated with traditional methods.

Acknowledgements
Not applicable.

Author contributions
AS contributed to the design and implementation of the research, the analysis of the results, and the writing of the manuscript. The author read and approved the final manuscript.

Funding
Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 18 July 2022 Accepted: 17 May 2023
Published online: 23 June 2023

References

1. Mohammad S, et al. Semeval-2018 task 1: affect in tweets. In: Proceedings of the 12th international workshop on semantic evaluation. 2018.
2. Trad C, et al. Facial action unit and emotion recognition with head pose variations. In: International Conference on Advanced Data Mining and Applications. 2012. Springer.
3. Ruiz-Garcia A, et al. A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots. Neural Comput Appl. 2018;29(7):359–73.
4. Wegrzyn M, et al. Mapping the emotional face. How individual face parts contribute to successful emotion recognition. PLoS ONE. 2017;12(5):e0177239.
5. Filippini C, et al. Facilitating the child–robot interaction by endowing the robot with the capability of understanding the child engagement: the case of mio amico robot. Int J Soc Robot. 2021;13(4):677–89.
6. Ozcan T, Basturk A. Transfer learning-based convolutional neural networks with heuristic optimization for hand gesture recognition. Neural Comput Appl. 2019;31(12):8955–70.
7. Constantine L, et al. A framework for emotion recognition from human computer interaction in natural setting. In: 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2016), Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM 2016). 2016.
8. Hibbeln MT, et al. How is your user feeling? Inferring emotion through human-computer interaction devices. MIS Q. 2017;41(1):1–21.
9. Patwardhan AS, Knapp GM. Multimodal affect analysis for product feedback assessment. arXiv preprint arXiv:1705.02694, 2017.
10. Karyotis C, et al. A fuzzy computational model of emotion for cloud based sentiment analysis. Inf Sci. 2018;433:448–63.
11. Giatsoglou M, et al. Sentiment analysis leveraging emotions and word embeddings. Expert Syst Appl. 2017;69:214–24.
12. Abdul-Mageed M, Ungar L. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers). 2017.
13. Pang B, Lee L, Vaithyanathan S, Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070, 2002.
14. Xia R, Zong C, Li S. Ensemble of feature sets and classification algorithms for sentiment classification. Inf Sci. 2011;181(6):1138–52.
15. He Y. A Bayesian modeling approach to multi-dimensional sentiment distributions prediction. In: Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining. 2012.
16. Almatrafi O, Parack S, Chavan B. Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014. In: Proceedings of the 9th international conference on ubiquitous information management and communication. 2015.
17. Maas A, et al. Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. 2011.
18. Das A, Gambäck B. Sentimantics: conceptual spaces for lexical sentiment polarity representation with contextuality. In: Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. 2012.
19. Njølstad PCS, et al. Evaluating feature sets and classifiers for sentiment analysis of financial news. In: 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). IEEE; 2014.
20. Saad F. Baseline evaluation: an empirical study of the performance of machine learning algorithms in short snippet sentiment analysis. In: Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business. 2014.
21. Sharma A, Dey S. A document-level sentiment analysis approach using artificial neural network and sentiment lexicons. ACM SIGAPP Appl Comput Rev. 2012;12(4):67–75.
22. Alessia D, et al. Approaches, tools and applications for sentiment analysis implementation. IJCA. 2015;125(3):26–33.
23. Biswas S. Advantages of deep learning, plus use cases and examples. https://www.width.ai/post/advantages-of-deep-learning. Accessed 10 Nov 2021.
24. Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188, 2014.
25. Socher R, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing; 2013.
26. Do HH, et al. Deep learning for aspect-based sentiment analysis: a comparative review. Expert Syst Appl. 2019;118:272–99.
27. Liu N, Shen B. Aspect-based sentiment analysis with gated alternate neural network. Knowl-Based Syst. 2020;188:105010.
28. Akhtar MS, et al. A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In: Proceedings of the 2017 conference on empirical methods in natural language processing. 2017.
29. Pan Y-F, Hou X, Liu C-L. Text localization in natural scene images based on conditional random field. In: 2009 10th international conference on document analysis and recognition. IEEE; 2009.
30. Ay Karakuş B, et al. Evaluating deep learning models for sentiment classification. Concurrency Computat Pract Exper. 2018;30(21): e4783.
31. Hong J, Fang M. Sentiment analysis with deeply learned distributed representations of variable length texts. Stanford: Stanford University Report; 2015. p. 1–9.
32. Bhattacharya A. Deep hybrid learning—a fusion of conventional ML with state of the art DL. https://towardsdatascience.com/deep-hybrid-learning-a-fusion-of-conventional-ml-with-state-of-the-art-dl-cb43887fe14. Accessed 26 Jul 2020.
33. Nimmi K, et al. Pre-trained ensemble model for identification of emotion during COVID-19 based on emergency response support system dataset. Appl Soft Comput. 2022;122: 108842.

34. Adoma AF, Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2020.
35. Sirisha U, Bolem SC. Aspect based sentiment & emotion analysis with ROBERTa, LSTM. IJACSA. 2022. https://doi.org/10.14569/IJACSA.2022.0131189.
36. Bansal B, Srivastava S. Hybrid attribute based sentiment classification of online reviews for consumer intelligence. Appl Intell. 2019;49(1):137–49.
37. Ma Y, et al. Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis. Cogn Comput. 2018;10(4):639–50.
38. Wang W, Xu H, Wan W. Implicit feature identification via hybrid association rule mining. Expert Syst Appl. 2013;40(9):3518–31.
39. Zainuddin N, Selamat A, Ibrahim R. Hybrid sentiment classification on twitter aspect-based sentiment analysis. Appl Intell. 2018;48(5):1218–32.
40. Prottasha NJ, Sami AA, Kowsher M, Murad SA, Bairagi AK, Masud M, Baz M. Transfer learning for sentiment analysis using BERT based supervised fine-tuning. Sensors. 2022;22:4157.
41. Jain PK, et al. Employing BERT-DCNN with sentic knowledge base for social media sentiment analysis. J Ambient Intell Human Comput. 2022. https://doi.org/10.1007/s12652-022-03698-z.
42. Fredrick H. Why is twitter important?. https://yourbusiness.azcentral.com/twitter-important-5023.html Accessed Jan 2022.
43. Tan KL, et al. RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network. IEEE Access. 2022;10:21517–25.
44. Jain PK, Saravanan V, Pamula R. A hybrid CNN-LSTM: a deep learning approach for consumer sentiment analysis using qualitative user-generated contents. ACM Trans Asian Low-Resour Lang Inf Process. 2021;20(5):84.
45. AlBadani B, Shi R, Dong J. A novel machine learning approach for sentiment analysis on twitter incorporating the universal language model fine-tuning and SVM. Applied System Innovation. 2022;5(1):13.
46. Pranika Jindala Varun Jaiswala and M. Umac, "Opinion Mining ofTwitter Data for Recommending Airlines Services", InternationalJournal of Control Theory and Applications, 2016,Twitter US Airline Sentiment. https://www.kaggle.com/crowdflower/twitter-airline-sentiment. Accessed Jan 2022.
47. Preslav Nakov, Alan Ritter, Sara Rosenthal, FabrizioSebastiani, and Veselin Stoyanov. 2016a. SemEval2016 task 4: Sentiment analysis in Twitter. In Proceedings of the 10th International Workshop on Semantic Evaluation. San Diego, California, USA, SemEval '16, pages 1–18., Apple Twitter Sentiment (CrowdFlower). https://www.kaggle.com/slythe/apple-twitter-sentiment-crowdflower. Accessed Jan 2022.
48. apple_twitter_sentiment_texts. https://www.kaggle.com/seriousran/appletwittersentimenttexts. Accessed Jan 2022.
49. Indrayuni E, Nurhadi A. Optimizing genetic algorithms for sentiment analysis of apple product reviews using SVM. SinkrOn. 2020;4(2):172–8.
50. Dang NC, Moreno-García MN, De la Prieta F. Sentiment analysis based on deep learning: a comparative study. Electronics. 2020;9(3):483.
51. Kumawat, S., et al. Sentiment analysis using language models: a study. In: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE; 2021.
52. Xiang, R., et al. Affection driven neural networks for sentiment analysis. in Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association; 2020.
53. Wen, S. and J. Li. Recurrent convolutional neural network with attention for twitter and yelp sentiment classification: ARC model for sentiment classification. In: Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence. 2018.
54. Janjua SH, et al. Multi-level aspect based sentiment classification of Twitter data: using hybrid approach in deep learning. PeerJ Comp Sci. 2021;7: e433.
55. Thapa B. Sentiment analysis of cybersecurity content on twitter and reddit. arXiv preprint arXiv:2204.12267, 2022.
56. Demotte P, et al. Enhanced sentiment extraction architecture for social media content analysis using capsule networks. Multimed Tools Appl. 2021. https://doi.org/10.1007/s11042-021-11471-1.

Publisher's Note