IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Sentiment Analysis with Ensemble Hybrid Deep Learning Model

KIAN LONG TAN[1], CHIN POO LEE[1],(Senior Member, IEEE), and KIAN MING LIM[1],(Senior Member, IEEE), KALAIARASI SONAI MUTHU ANBANANTHEN

[1]Faculty of Information Science and Technology, Multimedia University, Melaka 75450 Malaysia.

Corresponding author: Chin Poo Lee (e-mail: cplee@mmu.edu.my).

This work was supported by the Telekom Malaysia Research and Development Grant RDTC/221045, Fundamental Research Grant Scheme of the Ministry of Higher Education FRGS/1/2021/ICT02/MMU/02/4, and Multimedia University Internal Research Fund MMUI/220021.

ABSTRACT The rapid development of mobile technologies has made social media a vital platform for people to express their feelings and opinions. Understanding the public opinions can be beneficial for business and political entities in making strategic decisions. In light of this, sentiment analysis plays an important role to understand the polarity of the public opinions. This paper presents an ensemble hybrid deep learning model for sentiment analysis. The proposed ensemble model comprises three hybrid deep learning models which are the combination of Robustly optimized Bidirectional Encoder Representations from Transformers approach (RoBERTa), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM) and Gated Recurrent Unit (GRU). In the hybrid deep learning model, RoBERTa is responsible for projecting the textual input sequence into a representative embedding space. Thereafter, the LSTM, BiLSTM and GRU capture the long-range dependencies in the embedding given the class. The predictions by the hybrid deep learning model are then amalgamated by averaging ensemble and majority voting, further improving the overall performance in sentiment analysis. In addition to that, the data augmentation with GloVe pre-trained word embedding has also been applied to alleviate the imbalanced dataset problems. The experimental results show that the proposed ensemble hybrid deep learning model outshines the state-of-the-art methods with the accuracy of 94.9%, 91.77%, and 89.81% on IMDb, Twitter US Airline Sentiment dataset and Sentiment140 dataset, respectively.

INDEX TERMS Sentiment Analysis, Transformers, RoBERTa, LSTM, BiLSTM, GRU, Ensemble learning

I. INTRODUCTION

SENTIMENT analysis is a sensational topic in recent years due to its wide spectrum of applications. Sentiment analysis is also known as opinion mining where it involves natural language processing and deep learning approaches to systematically extract and identify the affective states and subjective information. In simple words, sentiment analysis analyzes the polarity and sentiment of written texts to understand whether it is positive, negative or neutral.

In recent years, social media has penetrated into people's daily life. Social media such as Twitter, Meta (previously known as Facebook), Instagram or any open forum have become the platforms for people to express their feelings and opinions. Hence, analyzing the written text from social media can help to understand public opinions. For example, the business owners get to know what should be improved for their products by analyzing the sentiment of the customer reviews. Other than that, political entities can also apply sentiment analysis to help them in making action plans.

In view of the importance of sentiment analysis, this paper presents an ensemble hybrid deep learning model for sentiment analysis. The ensemble model integrates three hybrid deep learning models that are built upon Robustly optimized Bidirectional Encoder Representations from Transformers approach (RoBERTa), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM) and Gated Recurrent Unit (GRU). The RoBERTa model encodes the input sequence into representative word embeddings. Subsequently, the LSTM, BiLSTM and GRU model captures the long-range dependencies in the word embeddings given the class. The ensemble model fuses the predictions of the hybrid deep learning models by averaging ensemble and majority voting. As the dataset used is imbalanced, word embedding-based data augmentation is also performed to synthesize more samples for the minority classes. The primary contributions of this paper are as follows:

1) An ensemble hybrid deep learning model is proposed for sentiment analysis. The ensemble model combines three hybrid deep learning models, namely RoBERTa-LSTM, RoBERTa-BiLSTM and RoBERTa-GRU. The predictions of the hybrid deep learning models are fused using averaging ensemble and majority voting to harness the strengths of the models and to improve the overall performance in sentiment analysis.
2) The hybrid deep learning models leverage RoBERTa in the encoding of the textual data into a meaningful embedding space. The RoBERTa uses dynamic masking attention mechanism enabling it to produce contextual word embedding. The long-range dependencies of the word embedding are then captured by the LSTM, BiLSTM and GRU, mitigating the vanishing gradient problems.
3) The data augmentation technique with GloVe pre-trained word embedding is performed to solve the imbalanced dataset problem. Data augmentation produces more training samples for better model learning and improved generalization capability.

The rest of the paper is structured as follows: Section II describes the existing ensemble models in sentiment analysis. Section III details the process flow of the proposed ensemble model, including data pre-processing, data augmentation, hybrid deep learning models and ensemble methods. Section IV presents the details of the datasets for performance evaluation. Section V reports the hyperparameter tuning and the optimal hyperparameter values. Section VI compares the experimental results of the proposed ensemble hybrid deep learning model and the existing sentiment analysis methods. Lastly, Section VII concludes the whole paper.

II. RELATED WORK

This section describes some existing ensemble learning methods for sentiment analysis. The ensemble learning methods combine machine learning, deep learning and hybrid learning.

In an early work, Alrehili et al. (2019) [1] performed sentiment analysis by using the ensemble of a variety of machine learning and ensemble learning methods, namely Naive Bayes, Support Vector Machine (SVM), Random Forest, Bagging and Boosting. The dataset consists of 34661 customer reviews collected from the Amazon website. The pre-processing techniques such as case folding, stop word removal, stemming and N-grams were applied to the text. The ensemble method with majority voting using unigram achieved 89.40% accuracy in the experiments.

Bian et al. (2019) [2] performed sentiment analysis with some simple machine learning methods. The authors investigated the performance of Logistic Regression, SVM and K-Nearest Neighbors (KNN) and their ensemble model with majority voting. The dataset used is small where it only consists of 6328 positive and negative samples. The TF-IDF vectorizer was used in the feature extraction. In the 10-fold cross-validation experiments, the ensemble model yielded the highest accuracy of 98.99%.

Likewise, Gifari et al. (2021) [3] also explored some simple machine learning methods, such as Multinomial Naïve Bayes, KNN, Logistic Regression and their ensemble model for sentiment analysis. The dataset was collected from the IMDb website and only contains positive and negative classes. The pre-processing steps include tokenization, stop word removal and word stemming. The TF-IDF vectorizer was selected as the feature extractor to compute the weights of the word occurrence in the document. The experimental results showed that the ensemble model achieved the highest accuracy of 89.40% on the dataset.

The performance of similar machine learning methods was compared in Parveen et al. (2020) [4]. There were five single machine learning models and the ensemble of the models, including Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, Linear Support Vector Classification and Nu-Support Vector Classification. The movie reviews from the corpora community with 1000 positive and 1000 negative samples were leveraged. The most frequent 3000 words were used as the features. The ensemble model with majority voting recorded the highest accuracy of 91% on the dataset.

Varshney et al. (2020) [5] explored the similar machine learning methods, including Logistic Regression, Naïve Bayes, Stochastic Gradient Descent (SGD) and their ensemble learning with majority voting. The Sentiment140 dataset was pre-processed by removing the unnecessary columns, usernames, hyperlinks and null values. The texts were then converted into lowercase and vectorized using TF-IDF feature extraction to capture the important features. The experiments demonstrated that the ensemble learning method achieved the best result of 80% positive class recall.

A more extensive sentiment analysis was done by Aziz & Dimililer (2020) [6], where different feature extraction and classification methods were tested on three datasets. The authors proposed an ensemble learning algorithm with Naïve Bayes, Logistic Regression, SGD, Random Forest, Decision Tree and SVM. Three datasets were adopted, namely SemEval-2017 4A, SemEval-2017 4B and SemEval-2017 4C. The datasets were pre-processed by tokenization, stop word removal, punctuation and number removal, repeated word removal, word substitution and stemming. Subsequently, the authors extracted five different features from the text, namely part-of-speech (POS) tagging, N-grams, Bag of Words (BoW), TF-IDF and lexicon-based features. Two ensemble learning, simple majority voting ensemble and weighted majority voting ensemble were used in the experiment. The empirical results suggested that the ensemble model with weighted majority voting achieved the highest accuracy of 72.95%, 90.8% and 68.89% on SemEval-2017 4A, SemEval-2017 4B and SemEval-2017 4C, respectively.

Athar et al. (2021) [7] proposed an ensemble model of Logistic Regression, Naïve Bayes, XGBoost, Random Forest and Multilayer Perceptron (MLP) to analyze the sentiment of movie reviews. The experiments were conducted on the IMDb that consists of 25000 positive reviews and 25000 negative reviews. Some pre-processing steps, including URLs, punctuation and stop words removal, tokenization and stemming were applied to the dataset. Subsequently, the TF-IDF vectorizer was leveraged in the feature extraction phase. The ensemble model achieved an accuracy of 89.9% on the IMDb.

The deep learning models were also engaged in the ensemble models. Nguyen & Nguyen (2018) [8] performed Vietnamese sentiment analysis by implementing the ensemble of machine learning and deep learning models. For the machine learning, TF-IDF vectorizer was used as the feature extractor and Logistic Regression and SVM were used in classification. As for deep learning, the Vietnamese texts were represented as word2vec embedding before passing into CNN and LSTM for classification. Three types of ensemble techniques were investigated, namely average/mean rule, max rule and voting rule in the ensemble learning. The experimental results demonstrated that the ensemble model with mean rule obtained 69.71% accuracy on Vietnamese Sentiment Dataset (DS1), while the ensemble model with voting rule achieved 89.19% and 92.80% accuracy on Vietnamese Sentiment Food Reviews (DS2) and Vietnamese Sentiment Dataset (DS3), respectively.

Salur and Aydin (2020) [9] conducted Turkish sentiment analysis with the combination of Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM), Bidirectional Long Short-term Memory (BiLSTM) or Gated Recurrent Unit (GRU). The authors collected a Turkish tweets dataset about the Global System for Mobile Communications operators. For CNN, the dataset was represented in the character-level embedding before being fed into CNN for feature extraction. As for the LSTM, BiLSTM or GRU, the dataset was cleaned by case folding, URL removal, and unnecessary words removal. The cleaned texts were then represented as FastText and word2vec before being transmitted into LSTM, BiLSTM or GRU for feature extraction. Subsequently, the features of CNN are concatenated with the features of LSTM, BiLSTM or GRU. The combined features are then classified by a softmax layer. The combination of CNN and BiLSTM with FastText recorded the highest accuracy of 82.14% on the dataset.

Another extensive sentiment analysis was carried out by Kamruzzaman et al. (2021) [10] where three conventional ensemble models and three deep neural network models were proposed. The conventional ensemble models were the voting ensemble, bagging ensemble and boosting ensemble of Logistic Regression, SVM and Random Forest. The neural network ensemble models were 7-Layer CNN + GRU, 7-Layer CNN + GRU + GloVe embedding and 7-layer CNN + LSTM + Attention Layer. The datasets used were Grammar and Online Product Reviews dataset and Restaurant Reviews in Dhaka, Bangladesh dataset. Some pre-processing steps were applied to clean the datasets, including tokenization, lemmatization, case folding, special character removal and stop word removal. For the conventional ensemble models, the TF-IDF vectorizer was used as the feature extractor. The neural network models, on the other hand, represented the text as the embedding. The experimental results reported that the 7-Layer CNN + GRU + GloVe model achieved the highest accuracy of 94.19% on the Grammar and Online Product Reviews dataset whereas the 7-Layer CNN + LSTM + Attention Layer model obtained the highest accuracy of 96.37% on the Restaurant Reviews in Dhaka, Bangladesh dataset.

Later, the sequence models were employed for sentiment analysis. Wazrah & Alhumoud (2021) [11] investigated the performance of the stacked gated recurrent unit (SGRU), stacked bidirectional gated recurrent unit (SBi-GRU), Arabic Bidirectional Encoder Representations from Transformers (AraBERT) and their ensemble model in the sentiment analysis of Arabic Tweets. The dataset adopted was the Arabic Sentiment Analysis dataset that contains 56674 samples with positive, negative and neutral classes. The automatic sentiment refinement (ASR) technique was applied in the pre-processing phase to remove the repeated tweets, hashtags, stop words and unrelated contents before tokenization. The pre-trained word embedding for Arabic text, i.e., AraVec was used to represent the text for the SGRU and SBi-GRU models. The empirical results showed that the ensemble model (SGRU + SBi-GRU + AraBERT) performed the best with 90.21% accuracy.

Another Arabic sentiment analysis was performed by Saleh et al. (2022) [12] using the ensemble of Recurrent Neural Network (RNN), LSTM and GRU. The experiments engaged three datasets: Arabic Sentiment Twitter Corpus, ArTwitter and Arabic Jordanian General Tweets. The samples were pre-processed by tokenization, Arabic stop words removal, stemming, emoticons and Tweets cleaning. The textual samples were first encoded into Continuous Bag of Words (CBoW) embedding. The ensemble model stacked the predictions of the deep learning models before passing into the Random Forest, SVM and Logistic Regression for final prediction. The ensemble model with Logistic Regression yielded the highest accuracy of 92.22% on Arabic Sentiment Twitter Corpus and 86.11% on Arabic Jordanian General Tweets. On ArTwitter, the ensemble model with SVM achieved the highest accuracy of 83.12%.

Alsayat (2022) [13] conducted sentiment analysis with the ensemble of LSTM, Google Sentiment API, Microsoft Azure Sentiment API, and IBM Watson Sentiment API. The datasets were self-collected Twitter Covid-19 dataset, Yelp review dataset, Amazon review dataset, and Web2.0 dataset. The data samples were cleaned by case folding, URLs and username removal, stop words removal, digits removal, and emoticons translation. The cleaned samples were then represented as BoW embedding before being fed into the LSTM. The ensemble model obtained the highest accuracies of 92.65% on Twitter Covid-19 dataset, 96.97% on Yelp review dataset, 97.50% on Amazon review dataset,

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3210182

and 86.4% on the Web2.0 dataset.

Table 1 summarizes the existing ensemble methods for sentiment analysis. The majority of the existing works described the textual data as word frequency-based features, such as TF-IDF and Bag of Words. These features capture the word occurrences normalized by the whole dataset, which might lose the contextual interpretation. Unlike the existing works that leveraged frequency-based features, the textual data in this work is encoded into the RoBERTa embedding with the attention mechanism. The RoBERTa embedding is able to capture the contextual significance of the words in the given sentence, which is important to determine the polarity of the sentence. Subsequently, the embedding is passed into LSTM, BiLSTM and GRU separately for model learning and inference. The predictions of each model are finally assembled by majority voting. Not only that, the text augmentation with pre-trained word embedding is also applied to mitigate the imbalanced dataset issues.

III. SENTIMENT ANALYSIS WITH ENSEMBLE HYBRID DEEP LEARNING MODEL

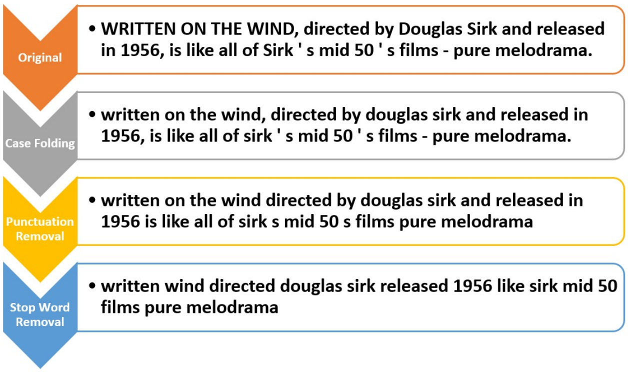
This section describes the process flow of the proposed sentiment analysis with an ensemble hybrid deep learning model. Firstly, some pre-processing steps are performed to clean the unnecessary elements in the textual data. Subsequently, data augmentation is performed on the imbalanced dataset to oversample the minority classes. The textual data are then sent into three hybrid deep learning models, namely RoBERTa-LSTM, RoBERTa-BiLSTM and RoBERTa-GRU for feature extraction and classification. The predictions returned by the hybrid deep learning models are combined via averaging ensemble and majority voting to improve the overall performance.

A. DATA PREPROCESSING

The data pre-processing is essential to clean the noise in the texts to improve the effectiveness of feature extraction and classification. In this work, case folding is first performed where all texts are standardized into lowercase. Thereafter, the punctuation and stop words are removed by excluding the punctuation and stop word set in Natural Language Toolkit from the texts. Figure 1 depicts the flow of data preprocessing with an example from Twitter US Airline Sentiment dataset.

B. DATA AUGMENTATION

Data augmentation synthesizes more sample data from the original data with the aim of improving the generalization capability of deep learning models. Some popular data augmentation techniques for textual data are Thesaurus [14], text generation [15] and word embedding [16]. The Thesaurus technique augments the textual data by substituting the words and phrases with their synonyms. The text generation technique produces the complete sentences instead of just replacing a few words in the sentences. The word embedding technique utilizes KNN and cosine similarity to identify the words with similar word embedding for substitution. The data augmentation with GloVe word embedding technique is leveraged in this work. As the Twitter US Airline dataset is imbalanced, the data augmentation is performed on the dataset to oversample the minority classes so that all classes have the same number of samples.



- WRITTEN ON THE WIND, directed by Douglas Sirk and released in 1956, is like all of Sirk ' s mid 50 ' s films - pure melodrama.
- written on the wind, directed by douglas sirk and released in 1956, is like all of sirk ' s mid 50 ' s films - pure melodrama.
- written on the wind directed by douglas sirk and released in 1956 is like all of sirk s mid 50 s films pure melodrama
- written wind directed douglas sirk released 1956 like sirk mid 50 films pure melodrama

FIGURE 1. The flow of data preprocessing with a sample sequence.

C. ENSEMBLE HYBRID DEEP LEARNING MODEL

The proposed ensemble model consists of three hybrid deep learning models. The hybrid models integrate the Robustly optimized Bidirectional Encoder Representations from Transformers approach and sequence models. The sequence models include LSTM, Bidirectional Long Short-Term Memory (BiLSTM) and GRU. Specifically, the hybrid models are RoBERTa-LSTM, RoBERTa-BiLSTM and RoBERTa-GRU. Figure 2 illustrates the architecture of the proposed ensemble hybrid deep learning model.

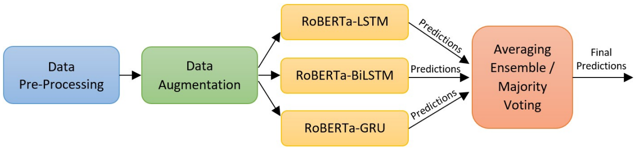


FIGURE 2. The architecture of the proposed ensemble hybrid deep learning model.

1) Robustly optimized BERT approach

Robustly optimized Bidirectional Encoder Representations from Transformers approach (RoBERTa) [17] is the first layer of every hybrid deep learning model. As the name implies, RoBERTa is the optimized version of BERT. BERT and RoBERTa are the members of the Transformer [18] family. The transformer model is a type of deep learning model that utilizes the self-attention mechanism in weighing the significance of the input sequence. Unlike RNNs, Transformer models allow parallel processing as the attention mechanism can provide context for any position in the input sequence. Figure 3 shows the architecture of the Transformer model. The encoder part consists of a self-attention network and a feed-forward network which are responsible for mapping the

TABLE 1. Summary of Related Work

Authors	Features	Ensemble Learning Methods
Alrehili et al. (2019) [1]	Unigram	Naive Bayes + SVM + Random Forest + Bagging + Boosting
Bian et al. (2019) [2]	TF-IDF	Logistic Regression + SVM + KNN
Gifari et al. (2021) [3]	TF-IDF	Multinomial Naïve Bayes + KNN + Logistic Regression
Parveen et al. (2020) [4]	3000 Frequent Words	Multinomial Naive Bayes + Bernoulli Naive Bayes + Logistic Regression + Linear Support Vector Classification + Nu-Support Vector Classification
Varshney et al. (2020) [5]	TF-IDF	Logistic Regression + Naïve Bayes + SGD
Aziz & Dimililer (2020) [6]	POS, N-grams, BoW, TF-IDF, Lexicon	Naïve Bayes + Logistic Regression + SGD + Random Forest + Decision Tree + SVM
Athar et al. (2021) [7]	TF-IDF	Logistic Regression + Naïve Bayes + XGBoost + Random Forest + Multilayer Perceptron
Nguyen & Nguyen (2018) [8]	word2vec	Logistic Regression + SVM + CNN + LSTM
Salur and Aydin (2020) [9]	Character-level Embedding, word2vec, FastText	CNN + LSTM / BiLSTM / GRU
Kamruzzaman et al. (2021) [10]	TF-IDF	CNN + GRU + GloVe, CNN + LSTM + Attention Layer
Wazrah & Alhumoud (2021) [11]	AraVec	SGRU + SBi-GRU + AraBERT
Saleh et al. (2022) [12]	CBoW	RNN + LSTM + GRU
Alsayat (2022) [13]	BoW	LSTM + Google Sentiment API + Microsoft Azure Sentiment API + IBM Watson Sentiment API

input sequence into a contextual encoding whereas the decoder part contains three components including self-attention network, encoder-decoder attention and feed-forward network which are used for the prediction task. In this work, only the encoder part of RoBERTa is trained as it is used to generate a contextual embedding sequence for the input sequence.

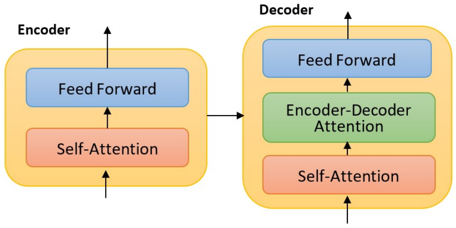


FIGURE 3. The components of the Transformers model.

Although RoBERTa has a similar architecture to BERT, there are some differences between them. Unlike BERT that applies static masking, RoBERTa adopts dynamic masking. With dynamic masking, different masking patterns are applied to the input sequence enabling the model to generate more diversified and contextually specific representation. Besides that, RoBERTa implements byte-level Byte-Pair Encoding (BPE) for tokenization which is more computation resources friendly compared to the character-level BPE of BERT. Apart from that, RoBERTa is trained on 10 times larger datasets for a longer time with larger batches and longer sequences in comparison with BERT. RoBERTa are trained on four datasets, as follows:

1) Book Corpus and English Wikipedia dataset: This dataset contains 16GB of text and is the same dataset that is used for the training of BERT.
2) CommonCrawl (CC)-News: The dataset consists of 63 million English news articles with the size of 76GB. The text was collected between September 2016 and February 2019.
3) OpenWebText: This dataset was collected from the Reddit website with the size of 38GB.
4) Stories: The size of this dataset is 31GB which is a subset of CommonCrawl data filtered to match the story-like style of the Winograd Natural Language Processing task.

The input sequence is tokenized before the input ids and attention mask are implemented. The input ids are responsible to encode the tokenized text into numerical representation sequentially. The attention mask is an indicator that determines the significance of every token when batching sequences together. Finally, the input ids and attention mask will be fed into the RoBERTa layer. The RoBERTa layer consists of 12 encoder layers with 768 hidden states.

2) Long Short-Term Memory

Long Short-Term Memory (LSTM) [19] is a sequence model that was introduced to solve the long-short memory and vanishing gradient problems of standard RNN. There are three gates that play an important role in the gating mechanism of LSTM, enabling it to encode the long-range dependencies of the input. The three gates are the forget gate, the input gate and the output gate. Firstly, the forget gate is used to determine which relevant information should be stored or

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3210182

dropped. Secondly, the input gate is applied to decide which value is important to update at the current step. Lastly, the output gate will decide what information should be passed to the next hidden state. Figure 4 shows the architecture of the RoBERTa-LSTM hybrid model. The calculations in the LSTM unit are as below:

$$f_t = \sigma\left(W_f X_t + U_f h_{t-1} + b_f\right) \tag{1}$$

$$i_t = \sigma\left(W_i X_t + U_i h_{t-1} + b_i\right) \tag{2}$$

$$o_t = \sigma\left(W_o X_t + U_o h_{t-1} + b_o\right) \tag{3}$$

$$\tilde{c_t} = \tanh\left(W_c X_t + U_c h_{t-1} + b_c\right) \tag{4}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c_t} \tag{5}$$

$$h_t = o_t * \tanh\left(c_t\right) \tag{6}$$

where f_t, i_t, o_t, c_t and h_t denote the forget gate, input gate, output gate, cell state and hidden state at the current time step t given the input X_t, respectively. The sigmoid function σ applied on three of the gates helps to control whether to pass or block the information at the current time step. The output of the sigmoid function that closes to 1 allows the information to be passed to the next state; conversely, an output value closes to 0 blocks the information to be passed to the next state. Besides, (W_f, W_i, W_o, W_c, U_f, U_i, U_o, U_c) and (b_f, b_i, b_o, b_c) denote the weight matrices and biases in forget gate, input gate, output gate and cell state, respectively. There are a total of 256 LSTM units in the hybrid RoBERTa-LSTM model, hence the calculations will repeat for 256 times in every training epoch.

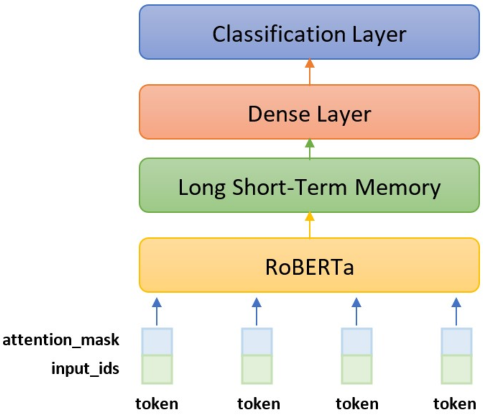


FIGURE 4. The architecture of the RoBERTa-LSTM hybrid model.

3) Bidirectional Long Short-Term Memory

Bidirectional LSTM (BiLSTM) implements the LSTM in both forward pass and backward pass at every time step. In doing so, the BiLSTM not only captures the past information, but also future information. Figure 5 illustrates the architecture of the RoBERTa-BiLSTM hybrid model. In the hybrid RoBERTa-BiLSTM model, the LSTM unit is set to 128 for both directions.

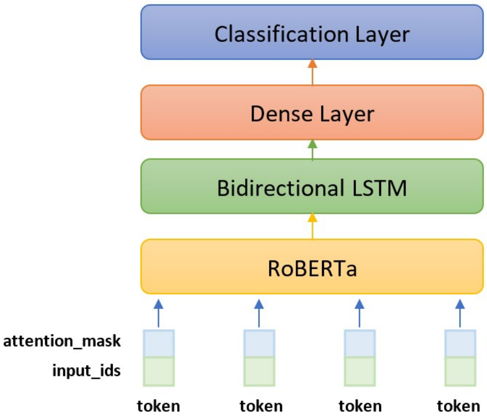


FIGURE 5. The architecture of the RoBERTa-LSTM hybrid model.

4) Gated Recurrent Unit

Gated Recurrent Unit (GRU) [20] is another sequence model that was introduced to solve the vanishing gradient problems of RNN. GRU also employs gating mechanisms to encode long-range dependencies but with a simpler architecture. There are two gates in GRU, namely reset gate and update gate. The reset gate is responsible for forgetting irrelevant information in the input, whereas the update gate will decide which information should be passed along to the future. Figure 6 illustrates the architecture of the RoBERTa-GRU hybrid model. The computations in the GRU unit are defined as:

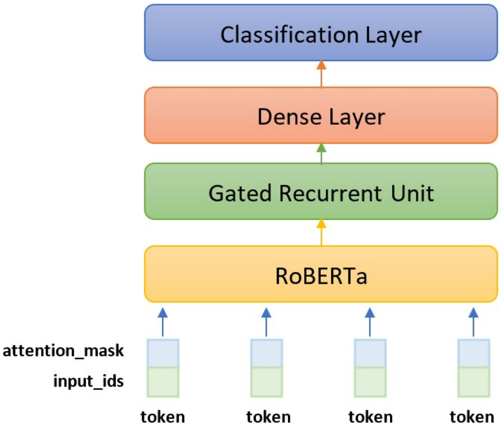


FIGURE 6. The architecture of the RoBERTa-GRU hybrid model.

$$\begin{aligned}
z_t &= \sigma\left(W_z X_t + U_z h_{t-1} + b_z\right) \\
r_t &= \sigma\left(W_r X_t + U_r h_{t-1} + b_r\right) \\
c_t &= \tanh\left(W_c X_t + U_c\left(r_t * h_{t-1}\right) + b_c\right) \\
h_t &= z_t * h_{t-1} + \left(1 - z_t\right) * c_t
\end{aligned} \tag{7}$$

where z_t and r_t denote the update gate and reset gate. The hybrid RoBERTa-GRU model consists of 256 GRU units, thus the computations will repeat 256 times at every time step.

5) Dense layer

The dense layer is also known as the fully connected layer as all neurons have deep connections with the neurons in the preceding layer. There are two dense layers in every hybrid deep learning model. The first layer consists of 256 hidden neurons which are responsible for capturing the class associated relationships in the output of LSTM / BiLSTM / GRU. The second dense layer serves as the classification layer where the softmax activation function is applied to compute the probability distributions of the sentiment classes.

6) Ensemble Methods

Ensemble methods combine the predictions of multiple machine learning models to improve the overall performance than a single machine learning model. In this paper, the averaging ensemble and majority voting ensemble are leveraged to fuse the predictions. The averaging ensemble calculates the mean of the probability distributions of every class, the final class label is determined by the class with the highest average probability. The majority voting ensemble (soft voting) takes the largest probability by the machine learning models as the final class label.

IV. DATASET

The datasets adopted in this work are Internet Movie Database (IMDb), Twitter US Airline Sentiment dataset and Sentiment140 dataset.

IMDb [21] is a balanced dataset that contains 50K movie reviews. The reviews are labeled into 25K positive and 25K negative classes. There are two columns in the dataset, which are review and sentiment.

Twitter US Airline Sentiment dataset was collected by CrowdFlower from Twitter in 2017. The dataset contains customer reviews about six American airline service providers, including American, United, US Airways, South-West, Delta and Virgin America Airlines. The sentiment classes in this dataset are positive, negative and neutral with 2363, 9178 and 3099 samples, respectively. As the data is imbalanced, the data augmentation with GloVe pre-trained word embedding is applied to oversample the minority classes so that all classes have the same number of samples.

Sentiment140 dataset is a large dataset that was collected by Stanford University [22] from Twitter. The dataset has balanced class sample distributions where there are 0.8 million positive samples and 0.8 million negative samples.

Figure 7 presents the class sample distribution of the datasets before and after data augmentation. All datasets are partitioned into 80% training set and 20% testing set. A sample input sequence for every dataset is provided below:

- IMDb: *Basically there's a family where a little boy (Jake) thinks there's a zombie in his closet & his parents are fighting all the time.<br /><br />This movie is slower than a soap opera... and suddenly, Jake decides to become Rambo and kill the zombie.<br /><br />OK, first of all when you're going to make a film you must Decide if its a thriller or a drama! As a drama the movie is watchable. Parents are divorcing & arguing like in real life. And then we have Jake with his closet which totally ruins all the film! I expected to see a BOOGEYMAN similar movie, and instead i watched a drama with some meaningless thriller spots.<br /><br />3 out of 10 just for the well playing parents & descent dialogs. As for the shots with Jake: just ignore them.*
- Twitter US Airline Sentiment: *@VirginAmerica I love the hipster innovation. You are a feel good brand.*
- Sentiment140: *@caregiving I couldn't bear to watch it. And I thought the UA loss was embarrassing . . . . .*

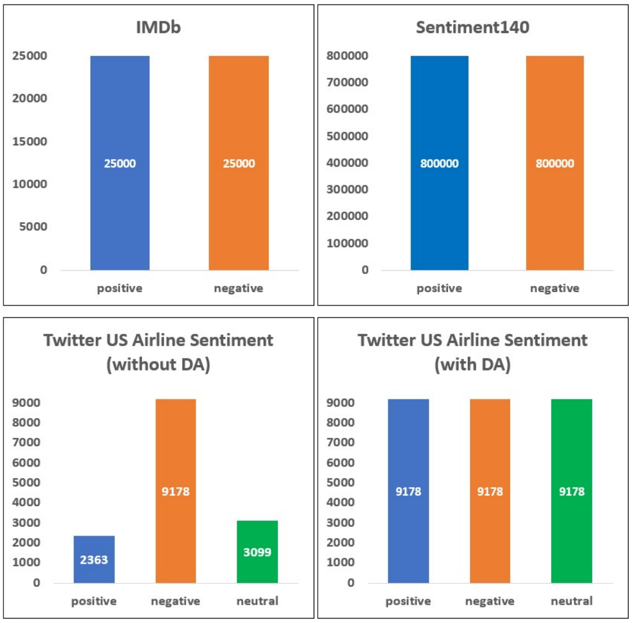


FIGURE 7. The sample distribution of the datasets.

V. HYPERPARAMETER TUNING

Hyperparameter tuning is essential to determine the optimal value of the hyperparameters so that the model performance can be optimized. The hyperparameter tuning is performed on the number of hidden units (LSTM or GRU) and optimizer. The augmented Twitter US Airline Sentiment dataset is leveraged for the hyperparameter tuning as it is the largest among the three datasets. Table 2 and Table 3 present the tested hyperparameter values and the optimal value of each hyperparameter for the RoBERTa-LSTM, RoBERTa-BiLSTM and RoBERTa-GRU, respectively.

TABLE 2. The hyperparameter tuning of the hybrid RoBERTa-LSTM model.

Hyperparameter	Tested Values	Optimal Value
LSTM unit	64, 128, 256	256
Optimizer	Adam, Nadam, SGD	Nadam

TABLE 3. The hyperparameter tuning of the hybrid RoBERTa-BiLSTM model.

Hyperparameter	Tested Values	Optimal Value
LSTM unit	64, 128, 256	128
Optimizer	Adam, Nadam, SGD	Nadam

TABLE 4. The hyperparameter tuning of the hybrid RoBERTa-GRU model.

Hyperparameter	Tested Values	Optimal Value
GRU unit	64, 128, 256	128
Optimizer	Adam, Nadam, SGD	Nadam

Table 5, Table 6 and Table 7 show the experimental results of different hidden units of the RoBERTa-LSTM, RoBERTa-BiLSTM and RoBERTa-GRU models. The RoBERTa-LSTM with 256 LSTM hidden units achieved the highest accuracy of 91.37%. The RoBERTa-BiLSTM model with 128 LSTM hidden units in each direction achieves the highest accuracy of 91.14%. Similarly for the RoBERTa-GRU model, 256 GRU hidden units obtain the highest accuracy of 91.48%.

A deep learning model that consists of too many hidden units may lead to overfitting where the model tends to learn too well on the training data and unable to generalize well to the testing data. The overfitting will cause the high variance problems where the training error is low, but the testing error is high. On the other hand, having too few hidden units may lead to underfitting where the model does not learn well on the training set and unable to generalize well to the testing set. This leads to high bias problems where both training error and testing error are high.

TABLE 5. Experimental results of different LSTM units for the RoBERTa-LSTM model.

LSTM Unit	Accuracy (%)
64	90.63
128	90.58
256	**91.37**
512	90.92

TABLE 6. Experimental results of different LSTM units for the RoBERTa-BiLSTM model.

LSTM Unit	Accuracy (%)
64	90.96
128	**91.21**
256	90.50

TABLE 7. Experimental results of different GRU units for the RoBERTa-GRU model.

GRU Unit	Accuracy (%)
64	91.07
128	91.41
256	**91.52**
512	90.83

Table 8, Table 9 and Table 10 present the experimental results of different optimizers. Among Adaptive Moment Estimation (Adam), Nesterov-accelerated Adaptive Moment Estimation (Nadam) and stochastic gradient descent (SGD) optimizers, the Nadam optimizer yields the best performance on the Twitter US Airline Sentiment dataset with 91.52%, 91.21% and 90.7% accuracy on RoBERTa-LSTM, RoBERTa-BiLSTM and RoBERTa-GRU, respectively. Nadam optimizer [23] is an enhancement of Adam optimizer where it integrates Nesterov momentum into the gradient descent process. Nesterov momentum involves calculating the decaying moving average of the gradients of projected positions rather than the actual positions themselves. In doing so, Nadam is good at accelerating the gradient descent process while optimizing the search when approaching the optima.

TABLE 8. Experimental results of different optimizers for the RoBERTa-LSTM model.

Optimizer	Accuracy (%)
Adam	90.63
Nadam	**91.37**
SGD	83.89

TABLE 9. Experimental results of different optimizers for the RoBERTa-BiLSTM model.

Optimizer	Accuracy (%)
Adam	90.63
Nadam	**91.21**
SGD	83.89

TABLE 10. Experimental results of different optimizers for the RoBERTa-GRU model.

Optimizer	Accuracy (%)
Adam	90.72
Nadam	**91.52**
SGD	84.66

VI. EXPERIMENTAL RESULTS AND ANALYSIS

In the performance evaluation, the performance of different attention-based embeddings is studied. The performance of the proposed ensemble model is also compared with the existing sentiment analysis methods.

A. COMPARATIVE RESULTS OF EMBEDDINGS

As embeddings are prominent in the representation of texts, this study evaluates the performance of different embeddings, namely BERT, A Lite BERT (ALBERT) [24] and RoBERTa. ALBERT is another extension of BERT where ALBERT improves the computational efficiency by introducing three enhancements: factorized embedding parameterization, cross-layer parameter sharing and inter-sentence coherence loss. The experimental results of different embeddings with LSTM, BiLSTM and GRU are presented in Table 11. It is observed that the hybrid models using RoBERTa as the embedding achieve the highest accuracy compared with BERT and ALBERT.

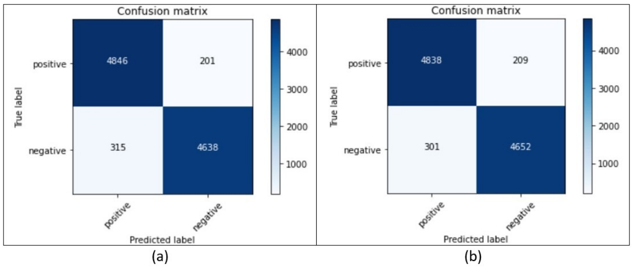
TABLE 11. The comparative results of different embeddings.

Model	Accuracy (%)		
	IMDb	Twitter US Airline Sentiment dataset	Sentiment140
BERT-LSTM	92.18	89.43	86.95
BERT-BiLSTM	92.47	90.52	87.04
BERT-GRU	92.54	89.69	87.09
ALBERT-LSTM	89.94	86.25	86.42
ALBERT-BiLSTM	90.18	84.24	85.33
ALBERT-GRU	90.72	88.51	86.18
RoBERTa-LSTM	**94.62**	**91.37**	**89.62**
RoBERTa-BiLSTM	**94.49**	**91.21**	**89.62**
RoBERTa-GRU	**94.63**	**91.52**	**89.59**

B. COMPARATIVE RESULTS WITH THE EXISTING WORKS

The performance evaluation also includes the existing sentiment analysis methods for a fair comparison. The existing methods include machine learning models, ensemble models and deep learning models. Besides, the experimental results of the individual hybrid deep learning model of the ensemble model are also presented. The evaluation metrics are accuracy, precision, recall and F1-score.

As shown in Table 12, the best performance was achieved by an ensemble model that combined Logistic Regression, Support Vector Machine and Random Forest [25] with an accuracy of 88.55% on the IMDb. The RoBERTa-LSTM, RoBERTa-BiLSTM and RoBERTa-GRU hybrid models record 94.62%, 94.49% and 94.63% accuracy, respectively. The proposed ensemble models with averaging ensemble and majority voting further increase the accuracy to 94.85% and 94.9%, respectively. The experimental results demonstrate a boost in the discriminating power when aggregating the predictions of multiple models. After representing the texts in RoBERTa embedding, RoBERTa-LSTM, RoBERTa-BiLSTM and RoBERTa-GRU demonstrate increments of 9.51%, 8.21% and 6.75% compared to LSTM, BiLSTM and GRU. The improvement substantiates the effects of the RoBERTa embedding in highlighting the significant tokens, thus allowing the sequence models to effectively learn the long-range dependencies. Figure 8 illustrates the confusion matrices of the proposed ensemble model on the IMDb.



(a) (b)

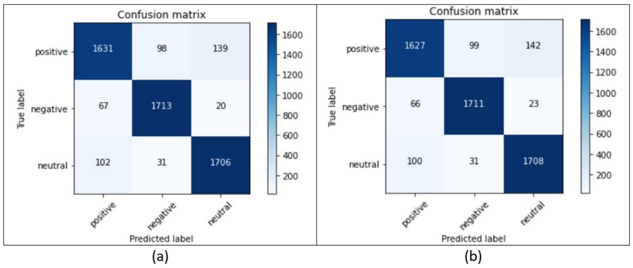
FIGURE 8. The confusion matrix of the ensemble model with majority voting (left) and averaging ensemble (right) on IMDb.

Table 13 presents the experimental results on the Twitter US Airline Sentiment dataset. It is observed that the Logistic Regression model [27] performed the best among the existing methods with 80.5% accuracy. The RoBERTa-LSTM, RoBERTa-BiLSTM and RoBERTa-GRU hybrid models achieve the accuracy of 85.79%, 86.24% and 85.69% on the original dataset without data augmentation. The proposed ensemble model obtains 85.99% and 85.86% with averaging ensemble and majority voting on the unaugmented dataset, respectively.

TABLE 12. The comparative experimental results on the IMDb.

Methods	Accuracy	Precision	Recall	F1-score
Naïve Bayes [26]	87.01	87	87	87
Logistic Regression [27]	87.12	90	90	90
Decision Tree [28]	73.46	74	73	73
KNN [27]	77.37	78	77	77
XGBoost [25]	80.8	81	81	81
AdaBoost [29]	83.37	83	83	83
Ensemble (LR+SVM+RF) [25]	88.55	89	89	89
GRU [30]	87.88	88	88	88
LSTM [30]	85.11	85	85	85
BiLSTM [31]	86.28	87	86	86
CNN-LSTM [32]	86.07	86	86	86
CNN-BiLSTM [33]	86.16	86	86	86
RoBERTa-LSTM [34]	**94.62**	**95**	**95**	**95**
RoBERTa-BiLSTM	**94.49**	**95**	**95**	**95**
RoBERTa-GRU	**94.63**	**95**	**95**	**95**
Ensemble model (average)	**94.85**	**95**	**95**	**95**
Ensemble model (majority)	**94.9**	**95**	**95**	**95**

After data augmentation, the accuracies of RoBERTa-LSTM, RoBERTa-BiLSTM, RoBERTa-GRU, ensemble model with averaging ensemble and ensemble model with majority voting have escalated to 91.37%, 91.21%, 91.52%, 91.47% and 91.77%, respectively. The improvement in performance corroborates the data augmentation with pre-trained word embedding is effective in dealing with the imbalanced dataset problems. Not only that, the data augmentation generates more samples for the model learning thus improving the generalization capability of the model. The confusion matrices of the ensemble model on the Twitter US Airline Sentiment dataset are depicted in Figure 9.



(a) (b)

FIGURE 9. The confusion matrix of the ensemble model with majority voting (left) and averaging ensemble (right) on the Twitter US Airline Sentiment dataset.

TABLE 13. The comparative experimental results on the Twitter US Airline Sentiment dataset.

Methods	Accuracy	Precision	Recall	F1-score
Naïve Bayes [26]	69.5	79	44	45
Logistic Regression [27]	80.5	78	69	72
Decision Tree [28]	71.14	62	56	58
KNN [27]	68.41	60	60	60
XGBoost [25]	73.84	74	55	57
AdaBoost [29]	74.59	67	63	65
Ensemble (LR+SVM+RF) [25]	79.78	80	80	80
GRU [30]	78.55	73	71	72
LSTM [30]	77.56	71	69	69
BiLSTM [31]	77.46	71	69	70
CNN-LSTM [32]	76.02	68	69	69
CNN-BiLSTM [33]	77.32	70	65	67
RoBERTa-LSTM (without data augmentation)	85.79	81	81	81
RoBERTa-BiLSTM (without data augmentation)	86.24	82	81	82
RoBERTa-GRU (without data augmentation)	85.69	81	81	81
Ensemble model (average) (without data augmentation)	85.99	81	82	81
Ensemble model (majority) (without data augmentation)	85.86	81	81	81
RoBERTa-LSTM [34]	**91.37**	**91**	**91**	**91**
RoBERTa-BiLSTM	**91.21**	**91**	**91**	**91**
RoBERTa-GRU	**91.52**	**91**	**91**	**91**
Ensemble model (average)	**91.47**	**91**	**92**	**91**
Ensemble model (majority)	**91.77**	**92**	**92**	**92**

The experimental results in Table 14 demonstrate that the LSTM, BiLSTM and GRU sequence models performed the best among the existing methods on the Sentiment140 dataset. The promising results exhibit the generalization and scalability of LSTM, BiLSTM and GRU in capturing long-range dependencies in the text and learning on extremely large dataset. The RoBERTa-LSTM, RoBERTa-BiLSTM and RoBERTa-GRU hybrid models achieve the accuracies of 89.62%, 89.62% and 89.59% on the dataset. The improvement in the performance when comparing the sequence model alone and hybrid of RoBERTa and sequence model reflects the contributions of RoBERTa embedding in paying more attention to significant tokens. The performance is further enhanced when the predictions are fused in the ensemble models, where both averaging and majoring voting recorded a higher accuracy of 89.81%. Figure 10 shows the confusion matrices of the proposed ensemble model on the Sentiment140 dataset.

TABLE 14. The comparative experimental results on the Sentiment140 dataset.

Methods	Accuracy	Precision	Recall	F1-score
Naïve Bayes [26]	76.57	77	77	77
Logistic Regression [27]	78.01	78	78	78
Decision Tree [28]	62.34	69	62	59
KNN [27]	60.39	66	60	57
XGBoost [25]	68.11	71	68	67
AdaBoost [29]	69.94	71	70	69
Ensemble (LR+SVM+RF) [25]	74.93	75	75	75
GRU [30]	78.96	78	78	78
LSTM [30]	79.10	79	79	79
BiLSTM [31]	78.53	78	78	78
CNN-LSTM [32]	77.53	77	77	77
CNN-BiLSTM [33]	77.58	77	77	77
RoBERTa-LSTM [34]	**89.62**	**90**	**90**	**90**
RoBERTa-BiLSTM	**89.62**	**90**	**90**	**90**
RoBERTa-GRU	**89.59**	**90**	**90**	**90**
Ensemble model (average)	**89.81**	**90**	**90**	**90**
Ensemble model (majority)	**89.81**	**90**	**90**	**90**

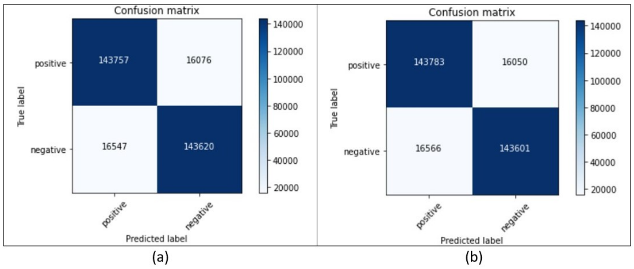


FIGURE 10. The confusion matrix of the ensemble model with majority voting (left) and averaging ensemble (right) on the Sentiment140 dataset.

VII. CONCLUSION

Sentiment analysis plays a vital role in many domains, such as business and politics, to understand the public opinions so that a strategic decision can be made. Therefore, an effective algorithm is required to automatically determine the polarity (positive, negative or neutral) of the opinions. This paper presents an ensemble hybrid deep learning model for sentiment analysis. Every single hybrid deep learning model applies RoBERTa to transform the input sequence into contextual word embedding. Thereafter, the sequence models are utilized to capture the long-range dependencies in the word embedding. Specifically, the sequence models are LSTM in the RoBERTa-LSTM model, BiLSTM in the RoBERTa-BiLSTM model and GRU in the RoBERTa-GRU model. The gating mechanisms of the LSTM, BiLSTM and GRU are effective in retaining the significant information even in the long input sequence, thus mitigating the vanishing gradient problem. The predictions by RoBERTa-LSTM, RoBERTa-BiLSTM and RoBERTa-GRU are then fused by the averaging ensemble and majority voting to improve the overall performance in sentiment analysis. In addition to that, the data augmentation with GloVe pre-trained word embedding is also leveraged to oversample the minority classes in Twitter US Airline Sentiment dataset. The hyperparameter tuning is also performed to optimize the performance of each hybrid model. The experimental results demonstrate that the ensemble hybrid deep learning model with majority voting outshines all methods in comparison with 94.9%, 91.77% and 89.81% on the IMDb, Twitter US Airline Sentiment dataset and Sentiment140 dataset, respectively.

REFERENCES

[1] Ahlam Alrehili and Kholood Albalawi. Sentiment analysis of customer reviews using ensemble method. In 2019 International Conference on Computer and Information Sciences (ICCIS), pages 1–6. IEEE, 2019.

[2] WenShuo Bian, ChunZhi Wang, ZhiWei Ye, and Lingyu Yan. Emotional text analysis based on ensemble learning of three different classification algorithms. In 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), volume 2, pages 938–941. IEEE, 2019.

[3] Muhammad Khaifa Gifari, Kemas M Lhaksmana, and P Mahendra Dwifebri. Sentiment analysis on movie review using ensemble stacking model. In 2021 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS), pages 1–5. IEEE, 2021.

[4] Rubeena Parveen, Neelesh Shrivastava, and Pradeep Tripathi. Sentiment classification of movie reviews by supervised machine learning approaches using ensemble learning & voted algorithm. In 2nd International Conference on Data, Engineering and Applications (IDEA), pages 1–6. IEEE, 2020.

[5] Chaudhary Jagrit Varshney, Ashish Sharma, and Dhirendra Prasad Yadav. Sentiment analysis using ensemble classification technique. In 2020 IEEE Students Conference on Engineering & Systems (SCES), pages 1–6. IEEE, 2020.

[6] Roza H Hama Aziz and Nazife Dimililer. Twitter sentiment analysis using an ensemble weighted majority vote classifier. In 2020 International Conference on Advanced Science and Engineering (ICOASE), pages 103–109. IEEE, 2020.

[7] Ali Athar, Sikandar Ali, Muhammad Mohsan Sheeraz, Subrata Bhattachariee, and Hee-Cheol Kim. Sentimental analysis of movie reviews using soft voting ensemble-based machine learning. In 2021 Eighth International Conference on Social Network Analysis, Management and Security (SNAMS), pages 01–05. IEEE, 2021.

[8] Hoang-Quan Nguyen and Quang-Uy Nguyen. An ensemble of shallow and deep learning algorithms for vietnamese sentiment analysis. In 2018 5th NAFOSTED Conference on Information and Computer Science (NICS), pages 165–170. IEEE, 2018.

[9] Mehmet Umut Salur and Ilhan Aydin. A novel hybrid deep learning model for sentiment classification. IEEE Access, 8:58080–58093, 2020.

[10] Mahammed Kamruzzaman, Mohammed Hossain, Md Rashidul Islam Imran, and Sagor Chandro Bakchy. A comparative analysis of sentiment classification based on deep and traditional ensemble machine learning models. In 2021 International Conference on Science & Contemporary Technologies (ICSCT), pages 1–5. IEEE, 2021.

[11] Asma Al Wazrah and Sarah Alhumoud. Sentiment analysis using stacked gated recurrent unit for arabic tweets. IEEE Access, 9:137176–137187, 2021.

[12] Hager Saleh, Sherif Mostafa, Abdullah Alharbi, Shaker El-Sappagh, and Tamim Alkhalifah. Heterogeneous ensemble deep learning model for enhanced arabic sentiment analysis. Sensors, 22(10):3707, 2022.

[13] Ahmed Alsayat. Improving sentiment analysis for social media applications using an ensemble deep learning language model. Arabian Journal for Science and Engineering, 47(2):2499–2511, 2022.

[14] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. Advances in Neural Information Processing Systems, 28:649–657, 2015.

[15] Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. Data augmentation for visual question answering. In Proceedings of the 10th International Conference on Natural Language Generation, pages 198–202, 2017.

[16] William Yang Wang and Diyi Yang. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2557–2563, 2015.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998–6008, 2017.

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[20] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259, 2014.

[21] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150, 2011.

[22] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12):2009, 2009.

[23] Timothy Dozat. Incorporating nesterov momentum into adam. 2016.

[24] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942, 2019.

[25] Priyanka Meel, Puneet Chawla, Sahil Jain, and Utkarsh Rai. Web text content credibility analysis using max voting and stacking ensemble classifiers. In 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), pages 157–161. IEEE, 2020.

[26] Young Gyo Jung, Kyung Tae Kim, Byungjun Lee, and Hee Yong Youn. Enhanced naive bayes classifier for real-time sentiment analysis with sparkr. In 2016 International Conference on Information and Communication Technology Convergence (ICTC), pages 141–146. IEEE, 2016.

[27] Tanushree Dholpuria, YK Rana, and Chetan Agrawal. A sentiment analysis approach through deep learning for a movie review. In 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), pages 173–181. IEEE, 2018.

[28] Arman S Zharmagambetov and Alexandr A Pak. Sentiment analysis of a document using deep learning approach and decision trees. In 2015 Twelve International Conference on Electronics Computer and Computation (ICECCO), pages 1–4. IEEE, 2015.

[29] M Vadivukarassi, N Puviarasan, and P Aruna. An exploration of airline sentimental tweets with different classification model. International Journal for Research in Engineering Application & Management, 4(02), 2018.

[30] Md Sagar Hossen, Anik Hassan Jony, Tasfia Tabassum, Md Tanvir Islam, Md Mahfujur Rahman, and Tania Khatun. Hotel review analysis for the prediction of business using deep learning approach. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), pages 1489–1494. IEEE, 2021.

[31] Apar Garg and Rohit Kumar Kaliyar. Psent20: An effective political sentiment analysis with deep learning using real-time social media tweets. In 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), pages 1–5. IEEE, 2020.

[32] Praphula Kumar Jain, Vijayalakshmi Saravanan, and Rajendra Pamula. A hybrid cnn-lstm: A deep learning approach for consumer sentiment analysis using qualitative user-generated contents. Transactions on Asian and Low-Resource Language Information Processing, 20(5):1–15, 2021.

[33] Maryem Rhanoui, Mounia Mikram, Siham Yousfi, and Soukaina Barzali. A cnn-bilstm model for document-level sentiment analysis. Machine Learning and Knowledge Extraction, 1(3):832–847, 2019.

[34] Kian Long Tan, Chin Poo Lee, Kalaiarasi Sonai Muthu Anbananthen, and Kian Ming Lim. Roberta-lstm: A hybrid model for sentiment analysis with transformer and recurrent neural network. IEEE Access, 10:21517–21525, 2022.

TAN KIAN LONG received the Bachelor Information Technology(Artificial Intelligence) degree from Multimedia University in 2021. He is currently doing his Master Degree in the same university. His current research interest include Natural Language Processing (NLP), Deep learning, Machine Learning and Sentiment Analysis.

CHIN POO LEE is a Senior Lecturer in the Faculty of Information Science and Technology at Multimedia University, Malaysia. She completed her Masters of Science and Ph.D. in Information Technology in the area of abnormal behaviour detection and gait recognition. Her research interests include action recognition, computer vision, gait recognition, and deep learning.

KIAN MING LIM received B.IT (Hons) in Information Systems Engineering, Master of Engineering Science (MEngSc) and Ph.D. (I.T.) degrees from Multimedia University. He is currently a Lecturer with the Faculty of Information Science and Technology, Multimedia University. His research and teaching interests includes machine learning, deep learning, and computer vision and pattern recognition.

KALAIARASI SONAI MUTHU is a Associate Professor in the Faculty of Information Science and Technology at Multimedia University (MMU), Malaysia. She is currently the coordinator for Master of Information Technology(information System). She received her doctorate in Artificial Intellingence from the University Malaysia Sabah, Malaysia, researching rule extraction from Artificial Neural Network. She has over 30 publications in the areas of artificial neural network, rule extraction, data mining and knowledge management. Her current research interests focus on data mining, opinion mining, neural, network and knowledge management.

• • •