IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 29, 2022, accepted February 16, 2022, date of publication February 18, 2022, date of current version March 2, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3152828

RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network

KIAN LONG TAN, CHIN POO LEE, KALAIARASI SONAI MUTHU ANBANANTHEN, AND KIAN MING LIM, (Senior Member, IEEE)
Faculty of Information Science and Technology, Multimedia University, Melaka 75450, Malaysia

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

This work was supported in part by the Fundamental Research Grant Scheme of the Ministry of Higher Education under Award FRGS/1/2019/ICT02/MMU/03/7, and in part by the Multimedia University Internal Research Fund under Grant MMUI/210112.

ABSTRACT Due to the rapid development of technology, social media has become more and more common in human daily life. Social media is a platform for people to express their feelings, feedback, and opinions. To understand the sentiment context of the text, sentiment analysis plays the role to determine whether the sentiment of the text is positive, negative, neutral or any other personal feeling. Sentiment analysis is prominent from the perspective of business or politics where it highly impacts the strategic decision making. The challenges of sentiment analysis are attributable to the lexical diversity, imbalanced dataset and long-distance dependencies of the texts. In view of this, a data augmentation technique with GloVe word embedding is leveraged to synthesize more lexically diverse samples by similar word vector replacements. The data augmentation also focuses on the oversampling of the minority classes to mitigate the imbalanced dataset problems. Apart from that, the existing sentiment analysis mostly leverages sequence models to encode the long-distance dependencies. Nevertheless, the sequence models require a longer execution time as the processing is done sequentially. On the other hand, the Transformer models require less computation time with parallelized processing. To that end, this paper proposes a hybrid deep learning method that combines the strengths of sequence model and Transformer model while suppressing the limitations of sequence model. Specifically, the proposed model integrates Robustly optimized BERT approach and Long Short-Term Memory for sentiment analysis. The Robustly optimized BERT approach maps the words into a compact meaningful word embedding space while the Long Short-Term Memory model captures the long-distance contextual semantics effectively. The experimental results demonstrate that the proposed hybrid model outshines the state-of-the-art methods by achieving F1-scores of 93%, 91%, and 90% on IMDb dataset, Twitter US Airline Sentiment dataset, and Sentiment140 dataset, respectively.

INDEX TERMS Sentiment, transformer, RoBERTa, long short-term memory, LSTM, recurrent neural network, RNN.

I. INTRODUCTION

Sentiment analysis is a process that studies the emotion, sentiment, and attitude of people from their written language. Sentiment analysis is one of the famous topics in the Natural Processing Language (NLP) field due to its importance [1]. The influence of sentiment analysis has penetrated business and even social media nowadays. Due to the rapid development of social media, everyone can express their feelings and opinions on the internet. Hence, sentiment analysis plays a vital role in understanding what consumers or reviewers think. Apart from that, sentiment analysis also acts as an important tool for investigating the public reaction in the political aspect. The sentiment of people's voice influences the decisions made by the political parties.

The sentiment analysis is challenging mainly due to the long-distance dependencies and lexical diversity of texts. Many machine learning methods were proposed for sentiment

The associate editor coordinating the review of this manuscript and approving it for publication was Zhouyang Ren.

analysis, especially sequence models that are able to encode the long-distance dependencies of the text. The sequence models are however less computationally efficient where the processing is serialized. In contrast, the Transformer models improve the computation by implementing parallelized processing. In view of this, a hybrid deep learning model that leverages the strengths of Transformer models [2] and sequence models is proposed. To be more specific, the proposed method integrates Robustly optimized BERT approach (RoBERTa) [3] from the Transformer family and Long Short-Term Memory (LSTM) [4] from the Recurrent Neural Networks family. The RoBERTa models stand out in sequence-to-sequence modeling to effectively generate representative word embedding for the texts. On the other hand, the LSTM excels in temporal modeling to encode the long-distance dependency of the given input. Apart from that, the data augmentation with synonym replacements are incorporated to enrich the lexical diversity of the corpus. The data augmentation also alleviates the skewed dataset problem. The contributions of this paper are three folds:

1) A hybrid deep learning model is proposed for sentiment analysis that incorporates the RoBERTa and LSTM model. The RoBERTa model serves the purpose of word or subword tokenization and word embedding generation, whereas the LSTM model encodes the long-distance temporal dependencies in the word embedding.
2) A data augmentation technique with pre-trained word embedding is leveraged to synthesize lexically diverse samples and to oversample the minority classes. In doing so, the generalization ability of the model is improved with more lexically rich training samples as well as the imbalanced dataset problem is solved.
3) The hyperparameter tuning is performed to determine the optimal hyperparameter settings that yield the superior results for sentiment analysis. The empirical results show that the proposed RoBERTa-LSTM method records a huge leap in the performance compared to the state-of-the-art methods.

This paper is structured into seven sections. Section I gives some background of the research and outlines the contributions of this work. The existing methods for sentiment analysis are described in Section II. Section III presents the processes, architecture, and technical details of the proposed RoBERTa-LSTM. Section IV specifies the sentiment analysis datasets in the performance evaluation. The hyperparameter tuning details are provided in Section V. Section VI presents the experimental results and analysis of the proposed method in comparison with the state-of-the-art methods. Finally, Section VII concludes the paper.

II. RELATED WORKS

This section describes the state-of-the-art methods for sentiment analysis. The methods can be broadly categorized into machine learning and deep learning methods.

A. MACHINE LEARNING

The authors in [5] compared three machine learning methods, including Naïve Bayes, Support Vector Machine (SVM) and K-Nearest Neighbour (KNN). The dataset was crawled from Twitter related to the 2019 presidential candidates of the Republic of Indonesia. The data samples were labeled as positive and negative classes. Some preprocessing techniques, namely text parsing, tokenization, and text mining were done before the training process. The dataset was split into 80% for training and 20% for testing. The paper shows that the Naïve Bayes method obtained the highest accuracy of 75.58%, followed by KNN with 73.34% and SVM with 63.99%.

Likewise, the authors in [6] compared the performance of Multinomial Naïve Bayes and Support Vector Classifier (SVC) with linear kernels in sentiment analysis. The dataset used was taken from Twitter which is related to Airline reviews. The dataset contains approximately 10K tweets from positive, negative, and neutral classes. First, the preprocessing techniques, such as stemming, URLs removal, and stop words removal, were applied to clean the data. In the work, 67% of data is used as the training data and 33% of data is used as the testing data. In the experiments, SVC achieved 82.48% accuracy while Multinomial Naïve Bayes obtained 76.56% accuracy.

In [7], the authors performed feature extraction to extract the useful features from the tweets before passing into Multinomial Naïve Bayes for sentiment analysis. They performed sentiment analysis on the sentiment140 dataset. The dataset contains 1,600,000 tweets that were labeled with three classes, i.e. positive, negative, and neutral. The sentiment140 dataset was split into several ratios in the experiments, which are 6:4 for experiment 1, 7:3 for experiment 2, 8:2 for experiment 3 and 9:1 for experiment 4. The Multinomial Naïve Bayes method recorded the highest accuracy of 85% in experiment 4.

The work [8] compared four machine learning methods in the sentiment analysis. The dataset used is the Indian Railways Case Study that was collected from Twitter. The data samples were labeled with positive, negative or neutral. The experimental results showed that the accuracy achieved by C4.5, Naïve Bayes, Support Vector Machine, and Random Forest are 89.5%, 89%, 91.5%, and 90.5%, respectively.

The authors of [9] proposed an AdaBoost model for the US Airline Twitter data sentiment analysis. Data preprocessing was performed to eliminate the unnecessary data from the text. Besides, some data mining skills were applied to understand the relationship among the elements in the dataset. The dataset was collected from Skytrax and Twitter using keywords which are related to the top 10 US Airlines. In the experiments, 75% of the data was allocated for training and 25% was allocated for testing. The authors combined the boosting and bagging methods in AdaBoost to obtain the highest F-score of 68%.

The research work [10] used six machine learning models for the sentiment analysis of the US Airlines twitter data. Preprocessing steps including stop word removal, punctuation removal, case folding, and stemming were performed. The Bag of Words was used in feature extraction. The dataset used was collected from CrowdFlower and Kaggle, which are related to six US Airlines. The dataset contains 14640 samples with three classes, i.e., positive, negative, and neutral. The dataset was split into 70% training and 30% testing. The accuracy achieved by Support Vector Machine, Logistic Regression, Random Forest, XgBoost, Naïve Bayes, and Decision Tree are 83.31%, 81.81%, 78.55%, 75.93%, and 73%, and 70.51%, respectively.

B. DEEP LEARNING

The authors in [11] proposed two deep learning methods for sentiment analysis of the multilingual social media text. The dataset was taken from Twitter during the general election of Pakistan in 2018. The dataset contains 20375 tweets in two languages, i.e., English and Roman Urdu. The data samples were classified into positive, negative, and neutral classes. The dataset was split into 80% for training and 20% for testing. The authors evaluated the performance of two Bidirectional Encoder Representations from Transformers (BERT), specifically Multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R). In the hyperparameter tuning, the learning rate of mBERT was set to 2e-5 whereas the learning rate of XLM-R was 2e-6. The experimental results showed that mBERT obtained 69% accuracy and XLM-R recorded 71% accuracy.

A character-based Deep Bidirectional Long Short-Term Memory (DBLSTM) method was leveraged for sentiment analysis of self-collected Tamil tweets in [12]. The dataset consists of 1500 tweets from the positive, negative, and neutral classes. Firstly, data preprocessing was done to remove unnecessary symbols, special characters, and numbers in the text. The cleaned data was then represented by the word embedding of DBLSTM which is based on the Word2Vec pre-trained model. The dataset was divided into 80% for training and 20% for testing. The DBLSTM method recorded an accuracy of 86.2% in the experiments.

The paper [13] performed sentiment analysis on movie reviews taken from IMDb. The dataset contains 3000 reviews with positive or negative labels. Some preprocessing steps were carried out to remove less meaningful characters, symbols, repeating words, and stop words. After that, CountVectorizer was used in the feature extraction. After that, the features were fed into several models for training and testing. The dataset was split into 75% for training and 25% for testing. Among Naïve Bayes, Support Vector Machine, Logistic Regression, K-Nearest Neighbour, Ensemble model, and Convolutional Neural Network (CNN), the CNN model records the highest accuracy of 99.33%.

In [14], the sentiment of the Saudi dialect was analyzed with a deep learning approach. The dataset comprises 60000 Saudi tweets of positive and negative class. Data preprocessing steps were done to remove the numbers, punctuation, special symbols and non-Arabic letters. Besides that, text normalization was also applied where the word forms were replaced by their lemmas. Subsequently, Word2Vec pre-trained models were adopted as the word embedding to encode the text. The data was randomly split into 70% for training and 30% for testing. The data was then trained and tested on Long-Short Term Memory (LSTM) and Bidirectional Long-Short Term Memory (Bi-LSTM). The empirical results suggest that the Bi-LSTM model performed better with 94% accuracy.

The authors in [15] proposed the residual learning by using 1-dimensional CNN (1D-CNN) and Recurrent Neural Networks for sentiment analysis. The dataset was taken from the IMDb dataset which contains 50000 movie reviews from positive and negative class. The data was split into 50% for training and 50% for testing. The convolutional layers of the 1D-CNN have 128 filters and 256 filters. The Recurrent Neural Networks layers contain 128 units for LSTM, Bi-LSTM and Gated Recurrent Unit (GRU). The experimental results demonstrated that the 1D-CNN with GRU model yielded the highest accuracy of 90.02%.

Likewise, [16] proposed a Sentiment Analysis Bidirectional Long-Short Term Memory (SAB-LSTM). The model consists of 196 Bi-LSTM units, 128 Embedding layers, 4 dense layers and classification layer with SoftMax activation function. The paper suggested that the additional layers can avoid the overfitting problem and optimize the model parameters dynamically. The dataset used consists of 80689 samples from five sentiment classes that were collected from social media reviews such as Twitter, YouTube, Facebook, news articles, etc. The dataset was split into 90% for training and 10% for testing. The experiments revealed that the SAB-LSTM model performed better than the common LSTM models.

The work [17] compared the performance of Support Vector Machine, Multinomial Naïve Bayes, LSTM, and Bidirectional Encoder Representations from Transformers (BERT) in sentiment analysis. Some preprocessing steps were applied, including tokenization, stemming, lemmatization, and stop words and punctuation removal. The dataset used consists of 1.6 million tweets with positive or negative class. The dataset was split into 80% for training and 20% for testing. The study concluded that BERT performed the best with 85.4% accuracy.

The authors of [18] adopted LSTM to deal with the sentiment analysis on 5000 Tweets in Bangla. The dataset was cleaned by space and punctuation removal. In the experiments, the dataset was split into 80% for training, 10% for training, and 10% for testing. After hyperparameter tuning, the architecture of 5 LSTM layers of size 128, batch size of 25 and learning rate of 0.0001 yielded the highest accuracy of 86.3%.

Table 1 provides a summary of the related works, with the methods and datasets used. Most existing works leveraged machine learning methods and recurrent neural networks.

TABLE 1. Summary of Related Works.

Article	Methods	Dataset
[5]	NB, SVM and KNN	Indonesia tweets
[6]	SVC, MNB	Airline Reviews
[7]	MNB	Sentiment140
[8]	C4.5, NB, SVM and RF	Indian Railways Tweets
[9]	SVM, DT, RF, Bagging, Boosting and AdaBoost	Twitter US Airline
[10]	SVM, LR, RF, XGB, NB, DT	Twitter US Airline
[11]	mBERT, XLM-R	Pakistan Election Tweets
[12]	DBLSTM	self-collected Tamil Tweets
[13]	NB, SVM, LR, KNN, Ensemble model and CNN	IMDb (3000 Reviews)
[14]	LSTM, Bi-LSTM	Saudi Tweets
[15]	1D-CNN	IMDb
[16]	SAB-LSTM	Social Media Review
[17]	SVM, MNB, LSTM, BERT	Sentiment140
[18]	LSTM	Bangla Tweets

The Transformers models are yet to be widely explored. Moreover, recurrent neural networks and Transformers models have their own strengths. Recurrent neural networks perform well in encoding long-range dependencies while Transformers improves the computation by parallelized processing. Hence, this paper explores the integration of Transformers and recurrent neural networks.

III. SENTIMENT ANALYSIS WITH RoBERTa-LSTM

This section describes the phases of the proposed RoBERTa-LSTM for sentiment analysis. Firstly, preprocessing steps are performed on the corpus to remove the unnecessary tokens or symbols in the text. Subsequently, data augmentation is carried out to address the imbalanced dataset problem. The data augmentation oversamples the minority class to make the sample size equally large as the majority class. The balanced dataset is then passed into the RoBERTa-LSTM model for training and classification.

The proposed model is the hybrid of Robustly optimized BERT approach (RoBERTa) [3] and Long Short-Term Memory (LSTM), referred to as the RoBERTa-LSTM model. The proposed model utilizes the pre-trained RoBERTa weights to efficiently map the tokens into meaningful embedding space. The output word embeddings are then fed into the LSTM to capture the salient semantic features.

A. DATA PREPROCESSING

In text analytics applications, data preprocessing is essential to remove the noise in the corpus. In this work, several preprocessing steps are performed. Firstly, the texts are standardized into lowercase. Subsequently, texts that are less necessary in the sentiment analysis, including stop words, numbers, punctuation, and special symbols are eliminated. Stop words are the common words, such as a, an, the, etc., that are syntactically important but semantically less important. Apart from that, stemming is applied to change the words into their lemmas. For example, the word 'collected' is converted into 'collect' after the stemming operation.

B. DATA AUGMENTATION

Data augmentation increases the sample size by synthesizing additional lexically similar samples from the original corpus. In doing so, the deep learning models perform training on more data samples, hence improving the generalization capability. The data augmentation method can also be used to oversample the minority classes in an imbalanced dataset.

There are a few data augmentation techniques in the text analytics field, namely Thesaurus [19], text generation [20], word embedding [21], etc. Thesaurus is an efficient text augmentation technique that uses synonyms to substitute the words or phrases. The text generation technique synthesizes a new sentence to replace the original sentence. The word embedding technique adopts K-nearest-neighbor (KNN) and cosine similarity in the word embedding to substitute the word vector with the closest word vector. Several popular pre-trained word embedding, such as Global Vectors for Word Representation (GloVe), Word2Vec and fastText, are widely leveraged in the text analytics field. In this research, the data augmentation with pre-trained word embedding, GloVe is applied due to its computation efficiency. As Twitter US Airline Sentiment dataset is imbalanced, the data augmentation is mainly applied on the dataset to oversample the minority classes.

C. RoBERTa-LSTM

This subsection gives a brief background of RoBERTa and explains the phases in the proposed RoBERTa-LSTM model. Figure 1 illustrates the architecture of the proposed RoBERTa-LSTM model.

1) RoBERTa

The RoBERTa model is an extension of Bidirectional Encoder Representation from Transformers (BERT). The BERT and RoBERTa fall under the Transformers [2] family that was developed for sequence-to-sequence modeling to address the long-range dependencies problem.

Transformer models comprise three components, namely tokenizer, transformers, and heads. The tokenizer converts the raw text into the sparse index encodings. Then, the transformers reform the sparse content into contextual embedding for deeper training. The heads are implemented to wrap the transformers model so that the contextual embedding can be used for the downstream tasks. The components of the Transformers are depicted in Figure 2.

BERT is slightly different from the existing language models where it can learn the contextual representation from both ends of the sentences. For the tokenization part, BERT used 30K vocabulary of character level Byte-Pair Encoding. In contrast, RoBERTa used a byte-level Byte-Pair Encoding with a larger vocabulary set that consists of 50K subword units. Apart from that, the RoBERTa model fine tunes the BERT model by training on more data, longer sequences, and longer time. RoBERTa was trained on 4 different corpora, as follows:

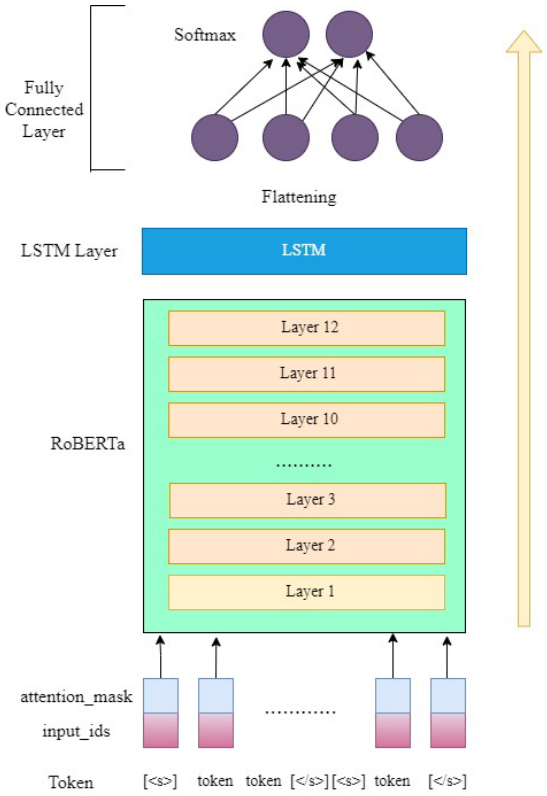


FIGURE 1. The architecture of the proposed RoBERTa-LSTM model.

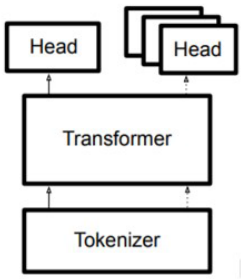


FIGURE 2. The components of the Transformers model.

1) BookCorpus + English Wikipedia: This dataset was also used to train BERT. The BookCorpus was released by Zhu *et al.* in 2015 [22].
2) CC-News: This dataset contains 63 million English news articles which were taken from CommonCrawl News.
3) OpenWebText: This is an open-source recreation which was released by Gokaslan and Cohen in 2019. It contains 38 GB of data.
4) Stories: Trinh and Le published this dataset in 2018 which includes 31GB of CommonCrawl data that was filtered to match the story-like style of Winograd schemas.

In the proposed RoBERTa-LSTM model, the cleaned text is first tokenized into words or subwords to be easily encoded into word embeddings. In this work, the RoBERTa tokenizer is used. The RoBERTa tokenizer has some special tokens, such as the <s> and </s> tokens to indicate the beginning and end of the sentence, <pad> token to pad the text to reach the maximum length of the word vector.

In the RoBERTa model, the byte-level Byte-Pair Encoding tokenizer is leveraged to partition the text into subwords. With this tokenizer, the frequently used words will not be split. However, the words that are rarely used will be split into subwords. For instance, the word 'Transformers' will be split into 'Transform' and 'ers'.

To make the model understand the text, the words need to be translated into a meaningful numerical representation. The RoBERTa tokenizer encodes the raw text with input ids and attention mask. The input ids represent the token indices and numerical representation of the token. On the other hand, the attention mask is used as an optional argument to batch the sequence together. The attention mask indicates which tokens should be attended and which should not.

The input ids and attention mask are passed into the RoBERTa base model. There are 12 RoBERTa base layers, 768 hidden state vectors, and 125 million parameters in the RoBERTa base model. The RoBERTa base layers aim to create a meaningful word embedding as the feature representation so that the following layers can easily capture the useful information from the word embedding.

2) LONG SHORT-TERM MEMORY

Subsequently, the output from the dropout layer is fed into the Long Short-Term Memory (LSTM) [4] model. The LSTM model is able to store the previous information thus capturing the prominent long-range dependencies in the given input. LSTM has performed significantly in sequence modeling tasks, such as text classification, sentiment analysis, time series prediction, etc.

There are three important elements in the LSTM model, including forget gate, input gate, and output gate. The forget gate will decide to forget or discard irrelevant information from the previous cell state and new input data. A sigmoid function is used in the training to return the values between [0, 1]. A value close to zero means that the information is less important to remember. The input gate plays the role of a filter to decide which information is worth remembering, thus to be updated into the next state. The value close to zero means that it is less important to be updated. The output gate determines the information that should be the output in the next cell state.

The calculations of the single LSTM unit at a single time step t in the forget gate f_t, input gate i_t, output gate o_t and cell state c_t, are defined as follows:

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

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

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

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

$$c_t = f_t * c_{t-1} + i_t * \tilde{c_t} \quad (5)$$
$$h_t = o_t * \tanh(c_t) \quad (6)$$

where σ is the sigmoid function, X_t denotes the input, (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 the forget gate, input gate, output gate, and cell state, correspondingly. h_{t-1} and c_{t-1} are the output of the LSTM at time step $t-1$. The operation $*$ is the element wise multiplication.

3) FLATTEN LAYER

The flatten layer serves the purpose of reshaping the input from 3D tensor to 2D tensor so that it can fit into the following dense layer.

4) DENSE LAYER

The dense layer is also known as the fully connected layer. The dense layer connects all the input from the preceding layer to all activation units in the subsequent layer. There are two dense layers in the proposed RoBERTa-LSTM model. The first dense layer consists of 256 hidden neurons to capture the relationship between the input and the classes. The last dense layer serves as the classification layer where the number of hidden neurons corresponds to the number of classes in the dataset. In the classification layer, the SoftMax activation function is applied to produce the probabilistic distribution of the classes in the sentiment analysis dataset.

5) MODEL TRAINING PARAMETERS

In the training of the proposed RoBERTa-LSTM model, the Adaptive Moment Estimation (Adam) optimizer is adopted to optimize the gradient descent process. The Adam optimizer utilizes the moving average of the gradients to avoid being stuck in the local minima, thus improving the gradient descent process. Not only that, the Adam optimizer is able to handle the sparse gradients on the noisy problems.

The loss function is part of the optimization algorithm where it is used to estimate the model loss in every training epoch. Since the sentiment analysis is a multi-class problem, the categorical cross entropy is selected as the loss function. The categorical cross entropy is defined as:

$$\text{categorical cross entropy}\ (p, t) = -\sum_{n=1}^{M} y_{o,c} \log\left(p_{o,c}\right) \quad (7)$$

where M denotes the number of classes, $y_{o,c}$ and $p_{o,c}$ denote the true and predicted class label for the observation o in class c.

IV. DATASET

In this work, three sentiment analysis datasets are used, namely IMDb, Sentiment140 and Twitter US Airline Sentiment dataset.

The IMDb dataset was collected from IMDb and published by Maas *et al.*, (2011) [23]. The dataset consists of two columns, i.e., review and sentiment. There are 50000 movie reviews with 25000 labeled as positive and 25000 labeled as negative. IMDb is a balanced dataset and is available at https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.

The Twitter US Airline Sentiment dataset was collected and released by CrowdFlower in February 2015. The dataset contains 14640 tweets with three class labels, including positive, negative, and neutral. The dataset is imbalanced where there are 9178 negative reviews, 2363 positive reviews and 3099 neutral reviews. The tweets in the dataset are related to six American airlines, namely United, US Airways, South-West, Delta, and Virgin America. The purpose of collecting this dataset was to analyze the sentiment of the customers from each airline. There are 10 columns in the dataset, including tweets id, sentiment, text, airline name, etc. However, only the text and sentiment column will be used in the sentiment analysis. The dataset can be accessed at https://www.kaggle.com/crowdflower/twitter-airline-sentiment.

The Sentiment140 dataset was collected from Twitter by Stanford University [24]. The dataset is a balanced dataset with 0.8 million positive reviews and 0.8 million negative reviews. There are six columns in the dataset, including target, id, text, flag, user, and id. In the experiment, the target and text column are used for the training and testing. The dataset can be downloaded from https://www.kaggle.com/kazanova/sentiment140.

The IMDb and Sentiment140 datasets have almost the equal number of samples for both positive and negative classes. However, the Twitter US Airline dataset is imbalanced where the sample size of the negative class is much higher than the positive and neutral class. In view of this, data augmentation with pre-trained word embedding, GloVe is applied on the Twitter US Airline Sentiment dataset. After the data augmentation, the numbers of samples in all three classes are equal and balanced. Figure 3 illustrates the sample distribution of the IMDb dataset and IMDb dataset without data augmentation, and the sample distribution of the Twitter US Airline Sentiment dataset before and after the data augmentation.

V. HYPERPARAMETER TUNING

The hyperparameter tuning is performed to determine the optimal hyperparameter values that yield the best performance. As the performance of all three datasets is monotonic, the Twitter US Airline Sentiment dataset after data augmentation is used in the hyperparameter tuning. Table 2 lists the hyperparameters, the tested values, and the optimal value of the hyperparameters. The experimental results of the proposed RoBERTa-LSTM with different LSTM units, optimizers and learning rates are presented in Table 3, Table 4, and Table 5, respectively.

The experimental results demonstrate that the optimal hyperparameter settings of the RoBERTa-LSTM model are 256 LSTM units, Adam optimizer and learning rate of 1e-5. The LSTM layer with 256 units exhibits the best

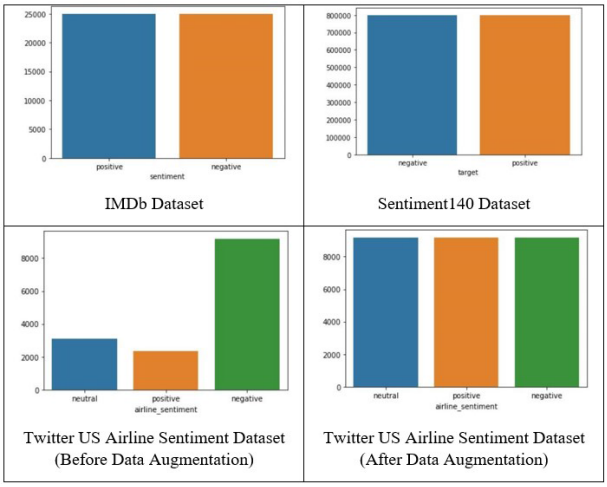


FIGURE 3. The sample distribution of the datasets.

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

Hyperparameter	Tested Values	Optimal Value
LSTM unit	64, 128, 256, 512	256
Optimizer	RMSProp, Adam, SGD	Adam
Learning rate	0.001, 0.0001, 0.00001, 0.000001	0.00001

TABLE 3. Different LSTM units (optimizer = Adam, learning rate = 0.00001).

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

TABLE 4. Different optimizers (LSTM unit = 256, learning rate = 0.00001).

Optimizer	Accuracy (%)
RMSprop	90.34
Adam	**91.37**
SGD	83.71

performance which demonstrates the capability to capture long-range dependencies. Besides that, the Adam optimizer outperforms the Root Mean Squared Propagation (RMSProp) and Stochastic gradient descent (SGD) in the gradient optimization process where Adam takes into account the past gradients and momentum. Learning rate is an important hyperparameter in model learning. Setting a learning rate too high might cause the model converges too fast leading to suboptimal solutionThe learning rate of 1e-5 allows the model to converge at an appropriate speed leading to an optimal performance.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents some experimental settings and the performance comparison of the proposed RoBERTa-LSTM model with the state-of-the-art methods.

TABLE 5. Different learning rates (LSTM unit = 256, optimizer = Adam).

Learning Rate	Accuracy (%)
0.001	32.69
0.0001	59.65
0.00001	**91.37**
0.000001	90.21

TABLE 6. The experimental results on the IMDB dataset.

Methods	Accuracy	Precision	Recall	F1-score
Naïve Bayes [7]	87.01	87	87	87
Logistic Regression [13]	87.12	90	90	90
Decision Tree [25]	73.46	74	73	73
KNN [13]	77.37	78	77	77
AdaBoost [26]	83.37	83	83	83
GRU [27]	87.88	88	88	88
LSTM [27]	85.11	85	85	85
BiLSTM [28]	86.28	87	86	86
CNN-LSTM [29]	86.07	86	86	86
CNN-BiLSTM [30]	86.16	86	86	86
RoBERTa-LSTM	**92.96**	**93**	**93**	**93**

TABLE 7. The experimental results on the Twitter US Airline Sentiment.

Methods	Accuracy	Precision	Recall	F1-score
Naïve Bayes [7]	69.5	79	44	45
Logistic Regression [13]	80.5	78	69	72
Decision Tree [25]	71.14	62	56	58
KNN [13]	68.41	60	60	60
AdaBoost [26]	74.59	67	63	65
GRU [27]	78.55	73	71	72
LSTM [27]	77.56	71	69	69
BiLSTM [28]	77.46	71	69	70
CNN-LSTM [29]	76.02	68	69	69
CNN-BiLSTM [30]	77.32	70	65	67
RoBERTa-LSTM (without data augmentation)	85.89	82	81	81
RoBERTa-LSTM	**91.37**	**91**	**91**	**91**

The training epoch is set to a maximum of 100, however, the early stopping mechanism is applied to prevent the overfitting problem. The observation metric of the early stopping is set to validation accuracy with the patience set to 30 epochs. In all experiments, the datasets are split into 6:2:2 for training, validation, and testing. The batch size is set to 32.

For a fair comparison, several machine learning and deep learning methods for sentiment analysis are included in the experiments. The machine learning methods include Naïve Bayes [7], Logistic Regression [13], Decision Tree [25], K-nearest neighbor (KNN) [13], and AdaBoost [26]. The deep learning methods include Gated Recurrent Unit (GRU) [27], Long Short-Term Memory (LSTM) [27], Bidirectional Long Short-Term Memory (BiLSTM) [28], Convolutional Neural Network-LSTM (CNN-LSTM) [29], and Convolutional Neural Network-BiLSTM (CNN-BiLSTM) [30]. Table 6, Table 7, and Table 8 present the experimental results on the IMDb dataset, Twitter US Airline Sentiment dataset, and Sentiment140 dataset.

The experimental results demonstrate that the proposed RoBERTa-LSTM model outperforms the state-of-the-art

TABLE 8. The experimental results on the Sentiment140 dataset.

Methods	Accuracy	Precision	Recall	F1-score
Naïve Bayes [7]	76.57	77	77	77
Logistic Regression [13]	78.01	78	78	78
Decision Tree [25]	62.34	69	62	59
KNN [13]	60.39	66	60	57
AdaBoost [26]	69.94	71	70	69
GRU [27]	78.96	78	78	78
LSTM [27]	79.10	79	79	79
BiLSTM [28]	78.53	78	78	78
CNN-LSTM [29]	77.53	77	77	77
CNN-BiLSTM [30]	77.58	77	77	77
RoBERTa-LSTM	**89.70**	**90**	**90**	**90**

methods on all three datasets. Referring to Table 6, on the IMDb dataset with 50K movie reviews, the methods in comparison record F1-scores within the range of 73% – 90%. Nevertheless, the proposed RoBERTa-LSTM model yields a higher F1-score of 93%. In terms of accuracy, the performance has increased from 87.88% recorded by the GRU model to 92.96%.

As observed in Table 7, on the Twitter US Airline Sentiment dataset with approximately 14.6K tweets, the proposed RoBERTa-LSTM has exhibited a huge leap in performance where the F1-score has increased from 45% – 72% to 91%. There is an improvement of 10.87% in accuracy compared to the second most competitive method, i.e., logistic regression. Additionally, the data augmentation has greatly improved the accuracy by 5.48%, hence demonstrating the effects of data augmentation in addressing the imbalanced dataset problems.

Apart from that, the proposed RoBERTa-LSTM model has also outshined the state-of-the-art methods on the Sentiment140 dataset with 1.6 million reviews. As shown in Table 8, the F1-score has improved from the range of 57% - 79% to 90%. The proposed RoBERTa-LSTM model yields an increment of 10.6% in accuracy compared to the best method in comparison, i.e., LSTM.

The experimental results corroborate the effectiveness of the proposed RoBERTa-LSTM model in sentiment analysis. The RoBERTa model performs exceptional in tokenizing and encoding the text sequence in word embeddings representation. The LSTM model, on the other hand, is capable of learning long-distance dependencies in the given input. The proposed RoBERTa-LSTM model integrates the strengths of RoBERTa and LSTM. The RoBERTa model plays the role of creating useful and representative word embedding as the features to facilitate the LSTM in capturing the temporal information.

Also, it is noteworthy that the data augmentation with GloVe pre-trained word embedding has greatly enhanced the performance of the proposed RoBERTa-LSTM on the imbalanced Twitter US Airline Sentiment dataset. The accuracy has improved from 85.89% to 91.37% after data augmentation is done to oversample the minority classes, thus improving the generalization ability of the proposed RoBERTa-LSTM model.

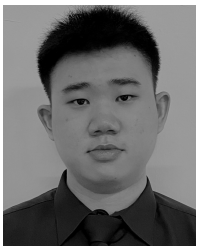
VII. CONCLUSION

In the big data era, having an effective sentiment analysis tool is essential in many aspects, especially economics and politics. The feedback of the sentiment analysis drives the decision making of the interested parties. The existing works on sentiment analysis mostly focus on machine learning methods and Recurrent Neural Networks. There are limited works that adopt Transformer for sentiment analysis. Henceforth, this paper presents a hybrid model of Transformer and Recurrent Neural Network, referred to as the RoBERTa-LSTM model. Firstly, data augmentation with GloVe pre-trained word embedding is used to generate more lexically similar samples and to oversample the minority classes. Subsequently, text preprocessing is performed to normalize the text and remove less important words. The cleaned corpus is then passed into the proposed RoBERTa-LSTM model for training and sentiment analysis. The proposed RoBERTa-LSTM model benefits from the strengths of both RoBERTa and LSTM, where RoBERTa efficiently encodes the words into word embedding while LSTM excels in capturing the long-distance dependencies. The experimental results demonstrate that the proposed RoBERTa-LSTM model outshines the state-of-the-art methods in sentiment analysis on IMDb dataset, Twitter US Airline Sentiment dataset, and Sentiment140 dataset.

REFERENCES

[1] S. Tam, R. B. Said, and Ö. Ö. Tanriöver, "A ConvBiLSTM deep learning model-based approach for Twitter sentiment classification," *IEEE Access*, vol. 9, pp. 41283–41293, 2021.
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Red Hook, NY, USA. Long Beach, CA, USA: Curran Associates, 2017, pp. 6000–6010.
[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[5] M. Wongkar and A. Angdresey, "Sentiment analysis using naive Bayes algorithm of the data crawler: Twitter," in *Proc. 4th Int. Conf. Informat. Comput. (ICIC)*, Oct. 2019, pp. 1–5.
[6] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of naive Bayes and SVM algorithm based on sentiment analysis using review dataset," in *Proc. 8th Int. Conf. Syst. Modeling Adv. Res. Trends (SMART)*, Nov. 2019, pp. 266–270.
[7] Y. G. Jung, K. T. Kim, B. Lee, and H. Y. Youn, "Enhanced naive Bayes classifier for real-time sentiment analysis with SparkR," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2016, pp. 141–146.
[8] D. K. Madhuri, "A machine learning based framework for sentiment classification: Indian railways case study," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 4, pp. 1–5, 2019.
[9] E. Prabhakar, M. Santhosh, A. H. Krishnan, T. Kumar, and R. Sudhakar, "Sentiment analysis of US Airline Twitter data using new AdaBoost approach," *Int. J. Eng. Res. Technol.*, vol. 7, no. 1, pp. 1–6, 2019.
[10] A. I. Saad, "Opinion mining on US Airline Twitter data using machine learning techniques," in *Proc. 16th Int. Comput. Eng. Conf. (ICENCO)*, Dec. 2020, pp. 59–63.
[11] A. Younas, R. Nasim, S. Ali, G. Wang, and F. Qi, "Sentiment analysis of code-mixed Roman Urdu-English social media text using deep learning approaches," in *Proc. IEEE 23rd Int. Conf. Comput. Sci. Eng. (CSE)*, Dec. 2020, pp. 66–71.
[12] S. Anbukkarasi and S. Varadhaganapathy, "Analyzing sentiment in Tamil tweets using deep neural network," in *Proc. 4th Int. Conf. Comput. Methodol. Commun. (ICCMC)*, Mar. 2020, pp. 449–453.

[13] T. Dholpuria, Y. K. Rana, and C. Agrawal, "A sentiment analysis approach through deep learning for a movie review," in *Proc. 8th Int. Conf. Commun. Syst. Netw. Technol. (CSNT)*, Nov. 2018, pp. 173–181.
[14] R. M. Alahmary, H. Z. Al-Dossari, and A. Z. Emam, "Sentiment analysis of Saudi dialect using deep learning techniques," in *Proc. Int. Conf. Electron., Inf., Commun. (ICEIC)*, Jan. 2019, pp. 1–6.
[15] N. K. Thinh, C. H. Nga, Y.-S. Lee, M.-L. Wu, P.-C. Chang, and J.-C. Wang, "Sentiment analysis using residual learning with simplified CNN extractor," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, pp. 335–3353.
[16] D. A. Kumar and A. Chinnalagu, "Sentiment and emotion in social media COVID-19 conversations: SAB-LSTM approach," in *Proc. 9th Int. Conf. Syst. Modeling Adv. Res. Trends (SMART)*, Dec. 2020, pp. 463–467.
[17] K. Dhola and M. Saradva, "A comparative evaluation of traditional machine learning and deep learning classification techniques for sentiment analysis," in *Proc. 11th Int. Conf. Cloud Comput., Data Sci. Eng.*, Jan. 2021, pp. 932–936.
[18] A. H. Uddin, D. Bapery, and A. S. M. Arif, "Depression analysis from social media data in Bangla language using long short term memory (LSTM) recurrent neural network technique," in *Proc. Int. Conf. Comput., Commun., Chem., Mater. Electron. Eng.*, Jul. 2019, pp. 1–4.
[19] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 649–657.
[20] K. Kafle, M. Yousefhussien, and C. Kanan, "Data augmentation for visual question answering," in *Proc. 10th Int. Conf. Natural Lang. Gener.*, 2017, pp. 198–202.
[21] W. Y. Wang and D. 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 *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2557–2563.
[22] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 19–27.
[23] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and A. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 142–150.
[24] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *Stanford*, vol. 1, no. 12, p. 2009, 2009.
[25] A. S. Zharmagambetov and A. A. Pak, "Sentiment analysis of a document using deep learning approach and decision trees," in *Proc. 12th Int. Conf. Electron. Comput. Comput. (ICECCO)*, Sep. 2015, pp. 1–4.
[26] M. Vadivukarassi, N. Puviarasan, and P. Aruna, "An exploration of airline sentimental tweets with different classification model," *Int. J. Res. Eng. Appl. Manage.*, vol. 4, no. 2, pp. 1–6, 2018.
[27] M. S. Hossen, A. H. Jony, T. Tabassum, M. T. Islam, M. M. Rahman, and T. Khatun, "Hotel review analysis for the prediction of business using deep learning approach," in *Proc. Int. Conf. Artif. Intell. Smart Syst. (ICAIS)*, Mar. 2021, pp. 1489–1494.
[28] A. Garg and R. K. Kaliyar, "PSent20: An effective political sentiment analysis with deep learning using real-time social media tweets," in *Proc. 5th IEEE Int. Conf. Recent Adv. Innov. Eng. (ICRAIE)*, Dec. 2020, pp. 1–5.
[29] P. K. Jain, V. Saravanan, and R. Pamula, "A hybrid CNN-LSTM: A deep learning approach for consumer sentiment analysis using qualitative user-generated contents," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 5, pp. 1–15, Sep. 2021.
[30] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A CNN-BiLSTM model for document-level sentiment analysis," *Mach. Learn. Knowl. Extraction*, vol. 1, no. 3, pp. 832–847, 2019.



KIAN LONG TAN received the Bachelor of Information Technology degree in artificial intelligence from Multimedia University, in 2021, where he is currently pursuing the master's degree. His current research interests include natural language processing (NLP), deep learning, machine learning, and sentiment analysis.

CHIN POO LEE received the Master of Science and Ph.D. degrees in information technology in the area of abnormal behavior detection and gait recognition. She is currently a Senior Lecturer with the Faculty of Information Science and Technology, Multimedia University, Malaysia. Her research interests include action recognition, computer vision, gait recognition, and deep learning.

KALAIARASI SONAI MUTHU ANBANANTHEN received the Ph.D. degree in artificial intelligence from the University Malaysia Sabah, Malaysia, researching rule extraction from artificial neural network. She is currently the Co-ordinator for Master of Information Technology (information system). She is also an Associate Professor with the Faculty of Information Science and Technology, Multimedia University (MMU), Malaysia. She has over 30 publications in the areas of artificial neural networks, rule extraction, data mining, and knowledge management. Her current research interests include data mining, opinion mining, neural networks, and knowledge management.

KIAN MING LIM (Senior Member, IEEE) received the B.I.T. (Hons.), Master of Engineering Science (M.Eng.Sc.), and Ph.D. (I.T.) degrees in information systems engineering from Multimedia University. He is currently a Lecturer with the Faculty of Information Science and Technology, Multimedia University. His research interests include machine learning, deep learning, computer vision, and pattern recognition.