Enhancing Sentiment Analysis with Hybrid Deep Learning Architectures

[1]Vikash Sawan, [2]Durga Prasad Roy

Research Scholar
Monad University
Hapur, Uttar Pradesh, India.

***Abstract*- Enhancing sentiment analysis on public opinion expressed in social networks, such as Twitter or Facebook, has been developed into a wide range of applications, but there are still many challenges to be addressed. Hybrid techniques have shown to be potential models for reducing sentiment errors on increasingly complex training data. This paper aims to test the reliability of several hybrid techniques on various datasets of different domains. Our research questions are aimed at determining whether it is possible to produce hybrid models that outperform single models with different domains and types of datasets.**

Hybrid deep sentiment analysis learning models that combine long short-term memory (LSTM) networks, convolutional neural networks (CNN), and support vector machines (SVM) are built and tested on eight textual tweets and review datasets of different domains. The hybrid models are compared against three single models, SVM, LSTM, and CNN. Both reliability and computation time were considered in the evaluation of each technique. The hybrid models increased the accuracy for sentiment analysis compared with single models on all types of datasets, especially the combination of deep learning models with SVM. The reliability of the latter was significantly higher.

***Keywords*: Sentiment Analysis; Deep Learning; Transformer; LSTM, SVM & ReLU.**

1. Introduction

The shortcoming of short text in deep learning models. Besides, the study by Qian et al. [16] revealed that LSTM behaves efficiently when used on different text levels of weather-and-mood tweets. After reviewing some recent studies [1, 11, 12], we found that CNN and RNN are outperforming methods with a relatively high overall accuracy. Both shallow neural networks and deep neural networks are capable of approximating any function. However, when contrasted to shallow neural networks, deep neural networks have the advantage of being able to do the feature extraction in the process of learning on large datasets. This is primarily because the deep models are able to extract/build better features than shallow models, using the intermediate hidden layers to achieve this [21]. For the same level of accuracy, deep neural networks can be much more efficient in terms of computation and number of parameters.

Deep neural networks are able to create deep representations; at every layer, the network learns a new, more abstract representation of the input. regardless of the types of social network datasets; providing an experimental study to evaluate the performance of hybrid deep learning models; and detailing a performance comparison of sentiment analysis methods with some state-of-art methods. the paper is organized as follows. Section 2 presents an overview of related work; Section 3 describes the methodology in this research area; Section 4 contains the proposed hybrid models; Section 5 describes and discusses the results of our experiments; and Section 6 offers our conclusions.

2. Related Work

The purpose of this study is to build hybrid models for sentiment analysis that can improve accuracy. We have previously examined the methods proposed and applied in other studies, which are discussed as follows. There are many ways to build hybrid models. In [26–28], the authors combined a CNN model and SVM that can improve the accuracy in image recognition. A convolutional network layer is used for extracting features and SVM functions as a recognizer. Original CNN is used with Softmax functions. Srinidhi et al. [36] proposed a hybrid model that combined LSTM and SVM with a radial basis function kernel for the textual classification of positive and negative sentiments. The hybrid model was evaluated on the IMD movie review datasets. These models are combined from single deep learning models with SVM for classification. Some of them are applied for image recognition. Our research combines two deep learning models and then uses SVM or ReLU for classification. Akhtar et al. [37] built a hybrid deep learning architecture, which is highly efficient for sentiment analysis in resource-poor languages. They used CNN for learning sentiment embedded vectors and SVM for sentiment classification. The model was tested on four Hindi datasets covering varied domains. Vo et al. [31] used a multichannel LSTM-CNN model for sentiment analysis on reviews/ comments from e-commerce sites. In addition, hybrid CNN–LSTM models are applied for sentiment analysis on movie reviews by Rehman et al. [30]. The same techniques are used in several works, for example, [29, 38–40]. Kaur et al. [41] designed an algorithm called a hybrid heterogeneous support vector machine (H-SVM).

They performed sentiment analysis on Twitter data related to COVID-19. Kastrati et al. [42] employed three different deep learning models such as CNN, LSTM, and CNN-LSTM for classifying Facebook comments related to the COVID-19 pandemic. They used pretrained word embedding method called FastText (an extension to Word2vec proposed by Facebook in 2016) and a

contextualized word embedding model, BERT, to learn and generate word vector. Both research scored tweet/comment as positive, negative, or neutral. However, these models were individually tested on different datasets in a particular domain or tested on few sample datasets. Therefore, their validity is not generally proven. A study by Jnoub et al. [19] focused on providing a generalized model for sentiment analysis that combined CNN with their own algorithm to transform reviews to were considered for the selection including the ability to avoid privacy concerns [52], acceptance in the research community, diversity of sources and topics, and size. The selected datasets enable a comprehensive comparison of the sentiment analysis approaches examined in this paper. The aim of the experiment is to understand whether the models give consistently accurate results regardless of the dataset type and size. The experiments were conducted using eight datasets. Three datasets contain tweets (Sentiment140, Tweets Airline, and Tweets SemEval) and five datasets contain reviews (IMDb movie reviews (1) and (2) and Cornell movie review). Among the tweets datasets, Sentiment140 [53], the largest, has 1.6 million tweets, each one labelled as either positive or negative sentiment.

While the others, Tweets Airline [54] and Tweets SemEval [55], contain 14,640 and 17,750 tweets, respectively, labelled as positive, negative, or neutral. The five review datasets include a total of 125,000 comments from user reviews of movies (IMDb movie reviews (1) [56], IMDb movie reviews (2) [57], and Cornell movie reviews [58]), books, and music (book and music reviews [59]), labelled as either positive or negative sentiments.

They are discussed in more detail in [1]. After examining the collected datasets, we saw that six out of eight datasets are initially labelled as positive and negative, and the sample on each label is relatively equal. The two datasets Airline and Tweet SemEval contain not only positive and negative labels but also neutral label. Having a balanced class distribution is important to ensure that prior probabilities are not biased for training models and doing classification [60].

In this research, we focus on polarity sentiment analysis, based on two classes positive and negative. The size of these datasets was reduced by removing the neutral labels. The remaining positive and negative classes are readjusted to be balanced. In addition, we applied k-fold cross-validation to the data in order to evaluate the models. In this way, the tests cover all instances of the datasets avoiding bias towards a particular subset of the data. Table 1 shows the number of samples (positive and negative) taken from each of dataset for performing experiments.

3. Methodology

Considering all of the advantages and potential of hybrid models and aiming at improving the performance of sentiment analysis techniques, our paper evaluates four hybrid models. The methodology is focused on three main components: the data to be used; process to build the feature vectors; building of hybrid methods for an appropriate sentiment analysis solution. These algorithms are applied to predict the sentiment polarity of the text and classify it according to that polarity.

3.1. Datasets

Our study does not focus on solving a problem in a particular domain but on providing an evaluation for general application models. In this study, we used several public datasets instead of generating and labelling new datasets of a specific application domain. Multiple criteria and includes some advances over Word2vec, such as support for out-of-vocabulary (OOV) words. Word2vec was published in 2013 by Tomas Mikolov at Google [62].

Table 1: Number samples of datasets.

#	Datasets	Number of samples
1	Sentiment140 (10%)	160.000
2	Tweets Airline	4.726
3	Tweets SemEval	9.300
4	IMDb movie reviews (1)	50.000
5	IMDb movie reviews (2)	25.000
6	Cornell movie reviews	10.662
7	Book reviews	2.000
8	Music reviews	2.000

This unsupervised learning model has trained datasets from a large corpus.
The dimension of Word2vec is much less than the dimension of one-hot encoding, with a matrix NxD, with N being the number of documents and D being the dimension of word embedding. Word2vec contains two models: skip-gram and continuous bag-of-words (CBOW). Both models are based on the probability of words occurring in proximity to each other. Skip-gram allows us to start with a word and predict words that likely surround it. However, one of the major drawbacks of using Word2vec is a lack of support for out-of-vocabulary words. To work around this issue, we use the special token [UNK] for words not found in the vocabulary. In addition, we also retrain the Word2vec model according to our vocabulary datasets with all words that appear more than five times, reducing the use of the special token. One issue in conducting sentiment analysis modelling is the varying length of the samples of the dataset. While deep learning models require fixed input vectors. Figures 1 and 2 show histograms of the datasets of reviews and tweets after they were cleaned. The x-axis represents the length of the data samples, and the y-axis is the frequency of appearance. Some histograms are rather ragged because we chose different types of datasets from different sources.

Standardizing data by smoothing outlines based on sample size could well fit the models [63]. In this study, we keep nearly raw data for sentiment analysis with the purpose of creating the necessary conditions to compare the efficiency of other models. We can see in Figures 1 and 2 that the data samples are quite widely varied in length. Therefore, it is necessary to set the data samples to the same length. The conversion of data samples adjusted to the same length is done as follows. For each dataset, we select a fixed length called d; for samples shorter than d, we add zeroes to the end of the vector and vice versa, in samples with length greater than d, the back will be cut off.

3.2. Preprocessing and Building the Feature Vector.

Sentiment classification can be carried out on three levels of extraction: the document, sentence, and aspect or feature [61]. In our experiments, we applied document-based sentiment analysis with word embedding techniques on eight datasets of tweets and reviews. Sentiment analysis requires that text-training data be cleaned before using as input for classification models. Irrelevant information in text or sentence data, including white space, punctuation, and stop words, is removed. Two techniques commonly used for this task are TF-IDF and word embedding. Our proposal uses the latter because it provides better results than TF-IDF [1]. We then used word embedding models, BERT and Word2vec, to build the feature vector.

BERT is a language model for nature language processing, and it was published by researchers at Google AI Language in 2018 [46]. BERT was developed after Word2vec the dataset. This is commonly done in other works [30].The fixed length d is selected as follows: datasets related to tweets usually have a small length variation due to the limit of tweets to a maximum of 280 characters; thus, this fixed length d is chosen to be the maximum length of the sample in the dataset. For the remaining datasets, the length d is selected from 300 to 500, based on the histogram of every single dataset. It could be possible to take a fixed length d, instead of different lengths, for tweets and reviews. However, if set length d is larger, it will waste much memory, and if set length d is smaller, it will miss some review data.

3.3. Hybrid Methods

There are numerous methods to build up a hybrid model for sentiment analysis. In this study, we tested the combination of several successful approaches. As shown in Figure 3, we start by using Word2vec or a pretrained BERT model to create the feature vector. We then vary the order of the CNN and LSTM models used in the next stages: Word2vec/BERT - > CNN - > LSTM or Word2vec/BERT - > LSTM - > CNN. We also vary the final stage of the model, using a ReLU function or using an SVM. Combining these two types of variation yields the four hybrid approaches that we have tested:

(1) Word2vec/ BERT - > CNN - > LSTM - > ReLU
(2) Word2vec/BERT - > LSTM - > CNN - > ReLU
(3) Word2vec/BERT - > CNN - > LSTM - > SVM
(4) Word2vec/BERT - > LSTM - > CNN - > SVM

Two approaches were used in our experiments to create feature vectors. The first approach was Word2vec initialized with random weights to learn the embedding for all words in our training datasets. Because Word2vec does not include contextual analysis to handle complex semantical or polymorphic cases in natural languages, our second approach was BERT.

A pretrained BERT model was used in this study. After adjusting the parameters, the BERT model was used as a feature extractor to generate input data for the proposal of hybrid models. The tweets and reviews data were fed into the BERT model to generate the feature vectors, which are the input to the hybrid models that perform the classification.
Enhancing sentiment analysis on information from social networks, such as Twitter or Facebook, is a research topic of growing interest today. Although much work has been done in this area, there are still many challenges to be addressed, including improving model reliability, reducing processing time, and applying techniques developed for specific types of data and specific data domains [1]. In recent years, deep learning models have been extensively applied in the field of sentiment analysis, where their great potential has been demonstrated.

Several studies are focused exclusively on building a single model from a single (or some) dataset(s) in a particular domain, such as marketing strategies [2], financial forecasting [3–5], and medical analysis [6, 7]. For social network applications, sentiment polarity-based deep learning applied to tweets is described thoroughly in [8-14]. Hassan and Mahmood [15] proved that CNN and recurrent neural networks (RNN) models can overcome Although a single machine learning method is relatively reliable when applied within certain domains, each deep learning approach has its own advantages and disadvantages. LSTM normally yields better results but requires more processing time than CNN, and CNN requires fewer hyperparameters and less supervision. Meanwhile, the LSTM performs more accurately for long sentences but requires a longer time to process [1]. The approach of combining two (or more) methods is introduced [23–25] as a means of incorporating the advantages of both and thus fills some shortcomings of individual methods. Alfrjani et al. [25] combined machine learning and semantic knowledge base for improving accuracy of sentiment analysis on reviews (improvement 1% to 6%).

In another case, Gupta and Joshi [23] proposed a hybrid method that combines lexicon and machine learning for sentiment analysis on tweets (improvement 2% to 6%).

A hybrid system with collaborative functions, therefore, is better able to address potential pitfalls, if any exist, associated with one single system. The effectiveness of the integrated models may vary based on different tasks. The CNN enhanced by SVM [26–28], CNN with RNN [29–32], and Lexicon-based analysis with machine learning [33, 34] showed an enhanced result. The combination of CNN, LSTM, and SVM aims to take advantage of the two deep network architecture models and SVM algorithms. When performing sentiment analysis on different domains and types of datasets. Moreover, there are different types of input data obtained from social networks, such as tweets and reviews. Within and across these types, the input data also contains differences, for example, the distribution of the lengths of the tweets and reviews, the diversity of topics in each dataset, the sample size, and the greater or lesser presence of explicit sentiments and irrelevant information. Some approaches may be unable to perform well in different domains, with inadequate accuracy and performance in sentiment analysis [1, 35]. As a result, certain approaches may be ill-suited and difficult to apply to certain types of input data.

A question raised in our study is whether hybrid models perform better than single models regardless of the characteristics of the datasets. therefore, our work examines how selected hybrid models behave with different types of datasets from different domains. In this work, we evaluated and validated the combination of three models CNN, LSTM, and SVM.

We considered the relationship between models and its advanced capacities to extract characteristics, store past information and nodes, and classify text. First, in the initial stages of the model, two possible variations in the sequence of CNN and LSTM are introduced. Then, for each of these alternatives, two new variations are introduced: the use of CNN with ReLU function or SVM. We applied these models with word embedding on eight datasets, including tweets and reviews. The results of our experiments showed that the combined models increased the accuracy of sentiment analysis. This paper offers three important contributions to the literature by highlighting four hybrid deep learning models for sentiment analysis that results in improved accuracy Complexity.

vectors. The model was evaluated on three different datasets: IMD, movie reviews, and their own dataset collected from Amazon reviews. Ombabi et al. [43] proposed a hybrid deep learning model that combines CNN and LTSM. In addition, FastText is used for word embedding and SVM for classification in the Arabic language. In our work, both Word2vec and BERT were applied for word embedding. We proposed four types of hybrid deep learning models based on CNN, LSTM, and SVM for classifying both tweets and reviews. Furthermore, other studies combine Lexicon-based analysis with machine learning [33] or sentiment lexicons and polarity shifting devices [44].

The research by S´anchez-Rada and Iglesias [24] deals with the problem of user and content sentiment classification. they proposed a hybrid model that merges features from different levels of social context. the model is evaluated in different datasets.

A study from Wang et al. [45] presented a hybrid approach, in which sentiment analysis of reviews about movies is used to improve a preliminary recommendation list obtained from the combination of collaborative filtering and content-based methods.

In the same approach, the use of a sentiment classifier induced from movie reviews as a second filter after collaborative filtering was proposed by Pandey et al. [10].

These research projects use traditional techniques to perform sentiment analysis. Our research applies deep learning techniques for improving the accuracy of sentiment classification. Recently, transfer learning has been successfully applied in sentiment analysis, in which lower network layers are trained on high-resource supervised datasets, such as BERT (proposed by researchers at Google AI language in 2018 [46]) and XLNET [47]. Examples can be found in [48–51], where BERT and XLNET were applied for sentiment analysis. the evaluation of different datasets and languages provides significant results.

However, it also requires sufficiently powerful hardware, large datasets, and long processing times when applying these techniques. For example, BERT-Base model has 110M parameters, and BERT-Large model has 340M parameters: pertaining is fairly expensive, requiring four days on 4 to 16 cloud TPUs.

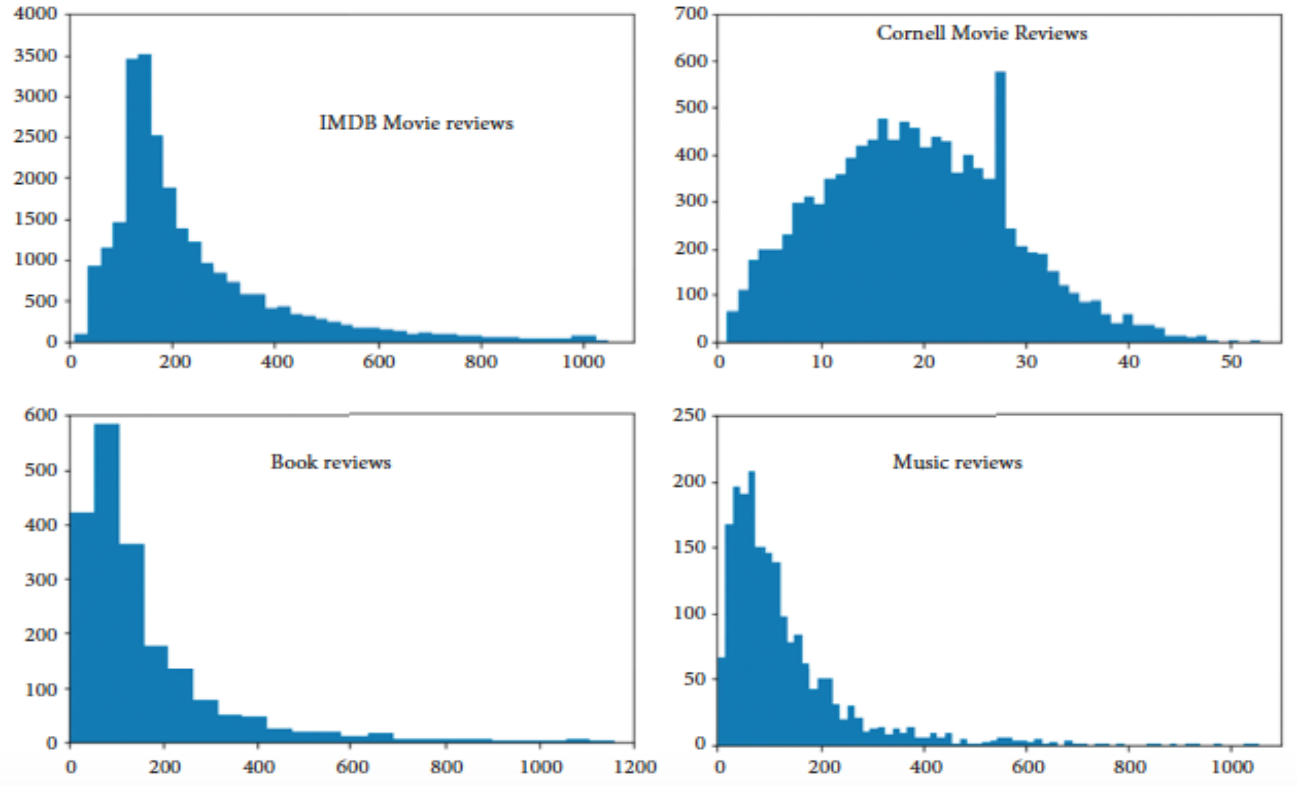

Figure 1: Histograms for different length data samples of reviews datasets.

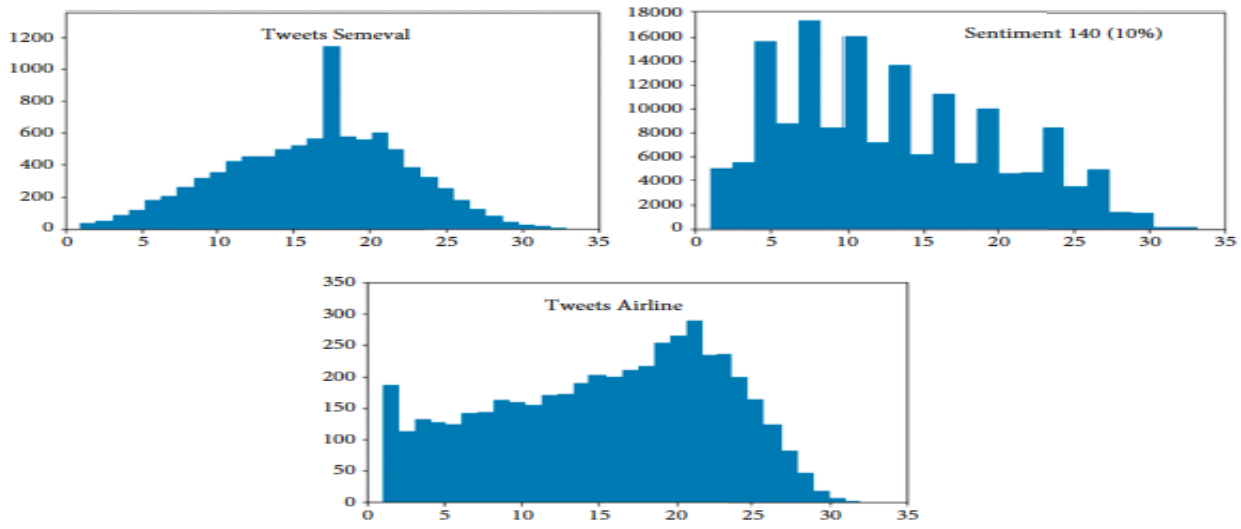

Figure 2: Histograms for different length data samples of tweets datasets.

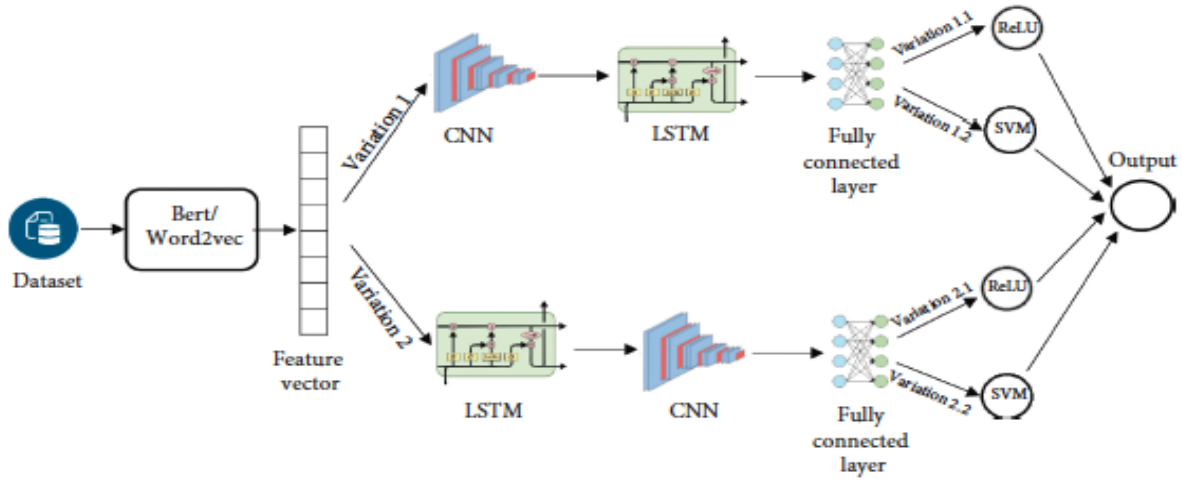

Figure 3: Process of methodology for sentiment analysis

LSTM can remember forward information of the sequence, and multilayer CNN can catch and learn local information sufficiently. So, the combination makes use of the best of both worlds, the spatial and temporal worlds. The final stage is classification. We use the activate function of ReLU instead of Sigmoid because of the high convergence. In addition, SVM was chosen for classification because of its efficiency in word processing, especially in high dimensional contexts, such as natural language processing. Support vector machine [69] is a supervised machine learning algorithm that can be used for both classification and regression tasks. It has been widely exploited with positive results in many areas. In our research, we have applied linear SVMs for classification with the proposed hybrid deep learning models. We extracted feature vectors from the top hidden layer and fed it to SVM that will classify for prediction ("positive" and "negative").

4. Proposed Hybrid Models

In this section, we proposed four hybrid deep learning models on variations in the use of CNN and LSTM in deep learning layers and variations of CNN and SVM in the classifier layers. The architecture of these hybrid models is shown in Tables 2 and 3, and the details are discussed as follows.

4.1. Scenario Combination

The first hybrid model combines CNN and LSTM models. The visualization of the model connection, the connection process, and the data processing flow are indicated in Table 2. The function embedding is the embedding layer that is initialized with random weights, which will learn the embedding for all words in the training datasets. The first layer of the hybrid model is the CNN, which receives the vector produced by word embedding. It has three convolution layers consisting of 512, 256, and 128 filters, respectively, with a kernel size 3, which receive and process data before feeding it into next deep learning layer.

4.2. Scenario Combination

The second hybrid model combines LSTM and CNN models. The visualization of the Complexity model connection, the connection process, and the data processing flow are indicated in Table 3. The input data is preprocessed to reshape data for the embedding matrix. The first layer of the hybrid model is the LSTM layer. That output has a matrix 13 × 500 and is fed into the second model of the hybrid deep learning model. The next layer of the hybrid model is the CNN. It has three convolution layers consisting of 512, 256, and 128 filters, respectively, with a kernel size 3, which are in charge of receiving and processing data before feeding it into the next layer. The CNN output is flattened and transferred to a fully connected layer. Finally, the hybrid model's classifier is a CNN composed of two continuous fully connected layers with 128 nodes and the ReLU activation function as the output layer.

4.3. Scenario Combinations 3 and 4. Our final hybrid model is based on the hybrid models from scenarios 1 and 2. We used the deep learning stages from those models (CNNLSTM and LSTM-CNN) but replaced the classifier. While there are multiple alternatives to the CNN-based ReLU function used, we have chosen to use SVM for the replacement classifier. Scenario 3 is based on CNN-LSTM, and Scenario 4 is based on LSTM-CNN. An architectural overview of the model is shown in Tables 2 and 3.

The second layer of the hybrid model is the LSTM, which produces a 1 × 500 matrix that is fed into the classifier. Next, the hybrid model's classifier is composed of two continuous, fully connected layers with 128 nodes and, finally, the output layer with a ReLU activation function

Table 2: A hybrid CNN-LSTM model.

Layer (type)	Output shape	Param #
embedding_1 (embedding)	(None, 38, 100)	1,808,900
conv1d_1 (Conv1D)	(None, 38, 512)	154,112
conv1d_2 (Conv1D)	(None, 38, 256)	393,472
conv1d_3 (Conv1D)	(None, 38, 128)	98,432
lstm_1 (LSTM)	(None, 500)	1,258,000
dense_1 (dense)	(None, 128)	64,128
dense_2 (dense)	(None, 128)	16,512
dense_3 (dense)	(None, 1)	129
Total params: 3,793,685		
Trainable params: 1,984,785		
Nontrainable params: 1,808,900		

Table 3: A hybrid LSTM-CNN model.

Layer (type)	Output shape	Param #
embedding_2 (embedding)	(None, 38, 100)	1,808,900
lstm_2 (LSTM)	(None, 38, 500)	1,202,000
conv1d_3 (Conv1D)	(None, 38, 512)	768,512
conv1d_5 (Conv1D)	(None, 38, 256)	393,472
conv1d_6 (Conv1D)	(None, 38, 128)	98,432
flatten_1 (flatten)	(None, 4864)	0
dense_4 (dense)	(None, 128)	622,720
dense_5 (dense)	(None, 128)	16,512
dense_6 (dense)	(None, 1)	129
Total params: 4,910,677		
Trainable params: 3,101,777		
Nontrainable params: 1,808,900		

The value of k is chosen to ensure that each train or test sample is large enough to represent the dataset. Furthermore, this procedure ensures that the k models in the cross-validation are induced from training sets of the same size and that the k test sets in all validations are also of the same size. It is recommended to split data into equal samples, so that the performance of the models is equivalent.

5. Experimental Results

In this section, we present the experiments conducted to compare the performance of the proposed hybrid models. Moreover, we also examine other common deep learning models (SVM, CNN, and LSTM). All of them were tested with the eight datasets introduced in subsection 3.1 that have been preprocessed with text processing techniques. Accuracy, AUC, and F-score were the metrics used to evaluate the performance of the models through all experiments. Since F-score is derived from recall and precision, we also show these two measures for reference purposes. The results are shown, discussed, and analysed in Sections 5.2 and 5.3.

5.1. Performance Comparison

Before performing the experiments, the configuration of related parameters, hardware devices, and the necessary library facilities were carried out. We used Google Colab Pro with GPU Tesla P100-PCIE- 16GB or GPU Tesla V100-SXM2-16GB [70] and the Keras [71] and TensorFlow libraries [72].

In all the experiments, we configured the parameter for our code, such as echoes = 4, k-fold = 10, and batch size = 32 with reviews and 128 with tweets. The common values for K-fold validation method are k = 3, k = 5, and k = 10, and by far, the most popular value used in applied machine learning to evaluate models is k = 10. the latter value is used when the dataset is large enough for the subsets to have a significant number of examples.

5.3. Discussion

As seen in Figures 4 to 8, using pretrained BERT produces better results than using Word2vec for sentiment analysis with all models and datasets. Focusing on the results of hybrid models, we see that, for each dataset, the best results are given by a hybrid model.

Hybrid models. produced better results than single models using either Word2vec or BERT. With the use of Word2vec, the results of accuracy from hybrid models are higher than the ones from single models. Using BERT, the results have also improved although by a smaller amount since these models have reached a relatively high accuracy, mostly more than 90%.

The text in a review is normally longer than the text in a tweet, which suggests that LCNN-SVN performs better than other hybrid models on longer textual sample (Table 4).

In selected datasets, when examining the distribution of the textual length of samples, the length of review ranges from 1 to 800 words. However, the Cornell movie reviews range from only 1 to 50 words. Besides, the length of a tweet ranges from 1 to 40 words; however, the distribution of sample length on Sentiment140 dataset is right skewed.

It is observed that the results on two datasets, Sentiment140 and Cornell movie reviews, are lower than those of the remaining datasets.

This is the case of the datasets used in this work. Thus, nine parts are used as training set and one as test set in each of the 10 validations.

5.2. Results

The results of eight sets of experiments are shown: three baseline models (SVM, CNN, and LSTM) and four hybrid models: CNN and LSTM, LSTM and CNN, CNN-LSTM and SVM, LSTM-CNN and SVM referred to as C-LSTM (or C-L), L-CNN (or L-C), CLSTM-SVM (or CL-S), LCNN-SVM (or LC-S), respectively.

A comparison analysis between the results obtained from the proposed hybrid methods against the baseline methods is also included. Our experiments were run twice: once using Word2vec to train word embedding and once using a pretrained BERT model to train word embedding. The results were consistently better when BERT was used, so Tables 4–8 provide details on the experimental results using Word2vec and BERT. Figures 4–8 illustrate the comparative results obtained with Word2vec and BERT using side-by-side bar charts.

The accuracy results shown in Table 4 are very high for all datasets and classification models when using a pretrained BERT model to extract a feature vector, around 90%, especially, 92.9% in Tweets Airline, and 93.4% in IMDb movie reviews (1). Moreover, the results prove that hybrid models show higher (or equal) accuracy than single deep learning models (SVM, CNN, or LSTM) for seven out of eight datasets.

Regarding the use of Word2vec in the music review and book review datasets, CNN's accuracy results given in Table 4 are 76.4% and 76.5%, respectively.

By comparison, when using the LCNN-SVM model, the results significantly improve to 83.7% and 82.7%, which represent an improvement of 7.3% and 6.2%, respectively. For the F-score (Table 7), hybrid models provided higher (or equal) values than single deep learning models for seven out of eight datasets. Regarding the AUC value in Table 8, the hybrid models also perform better than the single deep learning models.

The hybrid models using SVM for classification achieved the best results for six out of eight datasets using Word2vec. Among the datasets, the Tweets Airline dataset and IMDb movie reviews (1) are the datasets that show the highest values for all metrics in all cases. Book reviews and music reviews work well with hybrid LSTMCNN and LCNN-SVM models.

The Sentiment140 dataset has low accuracy in all models. In Figures 1 and 2, we can see the distribution of the total number of samples with the length data sample in the dataset.

The Sentiment140 dataset is also different from the other datasets. Number samples of length data are so much different.

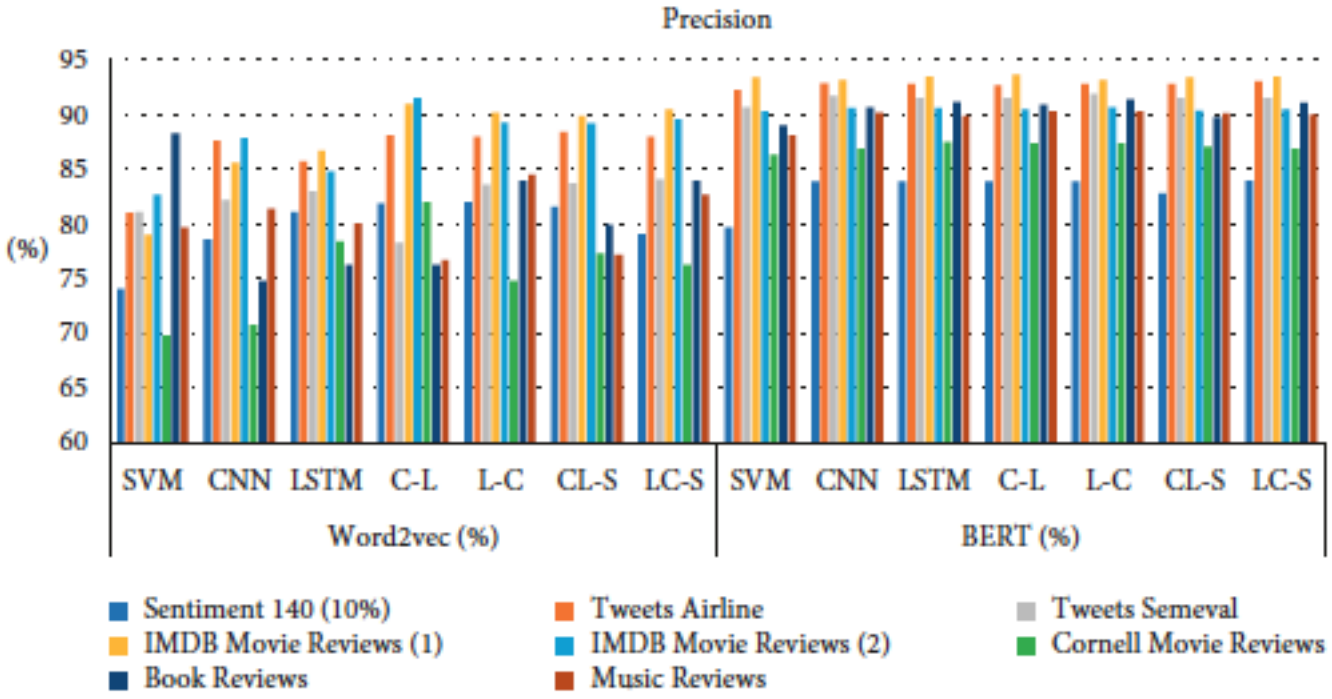

Figure 6: Precision values of deep learning models with Word2vec and BERT for different datasets.

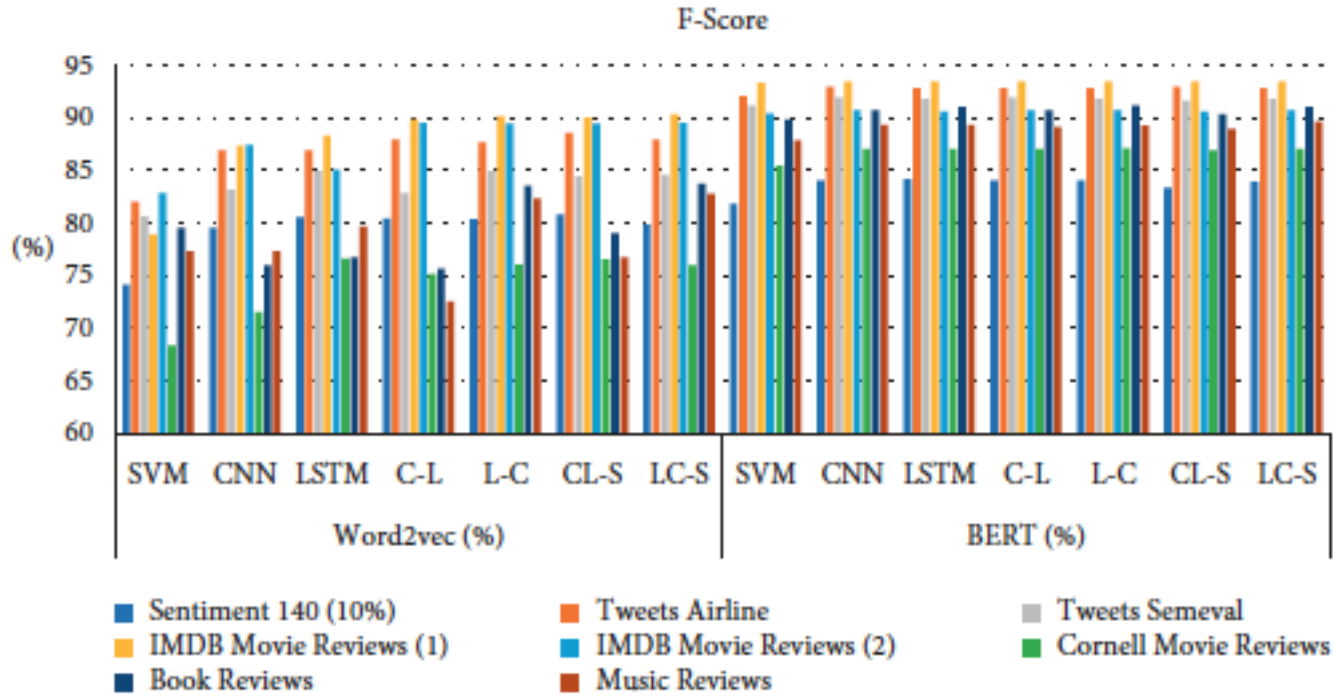

Figure 7: F-score values of deep learning models with Word2vec and BERT for different datasets.

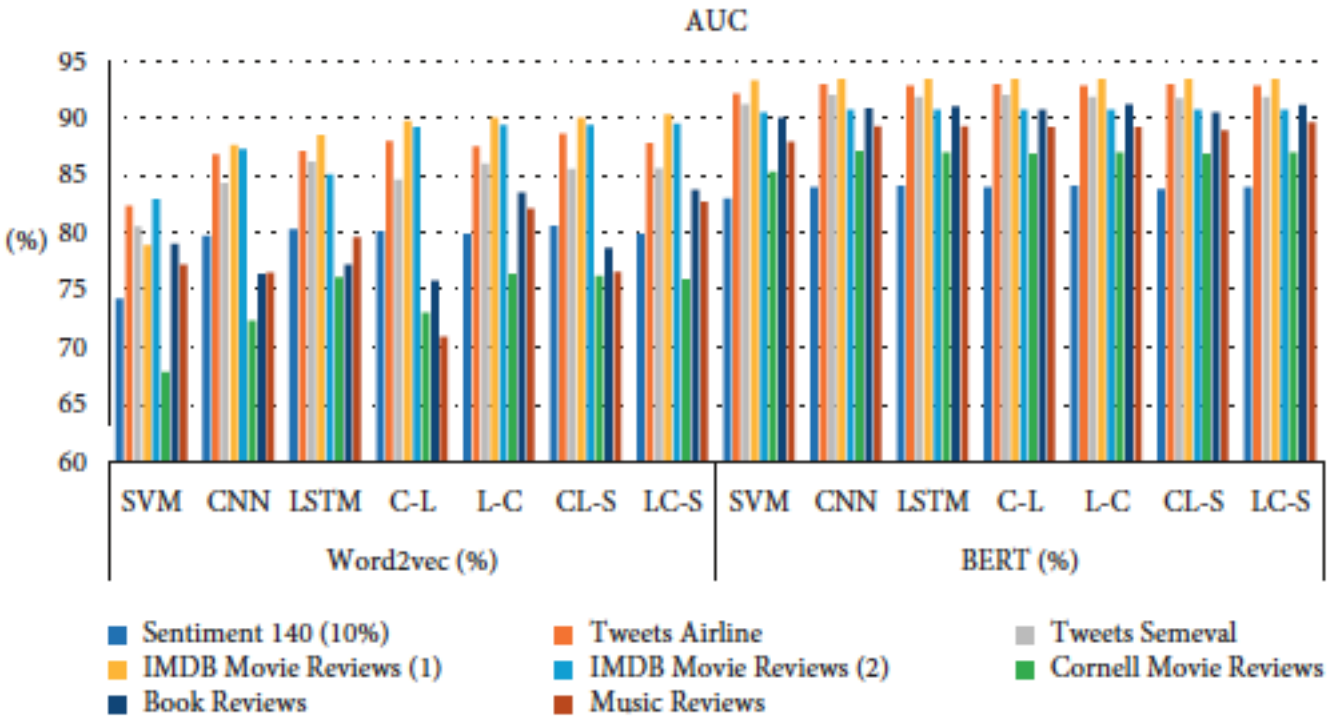

Figure 8: AUC values of deep learning models with Word2vec and BERT for different datasets.

Table 9: A comparison based on the proposed models and state-of-the-art approaches on datasets.

Study	Model	Dataset	Accuracy (%)
Kim and Jeong [18]	CNN	Cornell movie reviews	81
Maulana et al. [75]	SVM-IG	Cornell movie reviews	85.65
Proposed hybrid model	LCNN-SVM	Cornell movie reviews	**87**
Jnoub et al. [19]	SNN/CNN	IMDb	87/81
McCann et al. [20]	Char + CoVe-LSTM	IMDb	92.1
Tang et al. [76]	L-GRNN/Conv-GRNN	IMDb	45.3/42.5
Maltoudoglou et al. [49]	BERT	IMDb	92.28
Yang et al. [47]	XLNET	IMDb	96.21
Proposed hybrid model	CNN-LSTM	IMDb	**93.4**
Baziotis et al. [77]	Bi-LSTM + attention	Tweets SemEval	67.7 (F1)
Cliché [78]	LSTM-CNN	Tweets SemEval	68.5 (F1)
Proposed hybrid model	CNN-LSTM	Tweets SemEval	**91.9** (F1)
Abid et al. [12]	Bi-LSTM/CNN	Sentiment140	87.21/72.42
Han et al. [79]	FK-SVM	Sentiment140	87.2
Proposed hybrid model	LSTM-CNN	Sentiment140	**84.1**
Rane and Kumar [80]	AdaBoost	Tweets Airline	84.5
Duan et al. [81]	SVM and Naive Bayes	Tweets Airline	80
Monika et al. [17]	LSTM	Tweets Airline	80
Proposed hybrid model	CLSTM-SVM	Tweets Airline	**92.9**
Blitzer et al. [82]	SCL-MI	Music reviews/book reviews	79.7
Uribe [83]	Logistic/SVM	Music reviews/book reviews	87/89
Proposed hybrid model	LSTM-CNN/LC-S	Music reviews/book reviews	**91.1/89.5**

Some other studies performing sentiment analysis by using a single dataset of tweets or reviews are presented in [29, 33, 34, 37–39, 73, 74]. Note that the hybrid models provide much improved results in terms of processing time and accuracy.

In addition, the overall accuracy of these hybrid models was given with eight different types of datasets, which give an objective view of overall accuracy. Among the state-of-the-art approaches shown in Table 9, most of our hybrid models proposal got higher accuracy results on six datasets. On Sentiment140, however, Han et al. [79] and Abid et al. [12] achieved a better accuracy of around 87%.

The XLNet method for sentiment analysis with IMDb dataset, performed by Yang et al. [47], resulted in 96.21% accuracy. On the other hand, Akhtar et al. [37] tested a hybrid model of combined CNN and SVM on both tweet and review datasets; however, the

results showed a lower accuracy in comparison to hybrid methods, which were only tested on a single type of dataset (58.62% accuracy on tweet dataset and 77.16% accuracy with review dataset).

The comparison details with the state-of-the-art approaches are shown in Table 9. It includes the authors' names, methods, datasets, and accuracy (or F1 for some studies that only provide the F1 measure).

In addition to the evaluation of the reliability of the models, it is also important to evaluate the performance of the algorithms in terms of resource utilization.

We also discussed the importance of feature extraction in [1], where TF-IDF and word embedding techniques for feature extraction were analysed. These improved results are high and stable at the expense of some increase in processing time, The table shows that the hybrid model required longer computational time than the single models, because hybrid models are complex and feature many more parameters than single models. While the computational times are longer, they do not preclude analysis of the trade-offs between processing time and accuracy of results. Our aim is to build a hybrid deep learning model for sentiment analysis that works well on various datasets of domains. However, when building the classification models, there are many parameters that must be defined before, so they can be suitable for a given dataset but not for others. Therefore, the results obtained are positive and highly reliable because they have been evaluated on many datasets with different topics. Finally, general summaries of the results achieved in the experiments referenced earlier are discussed as follows:
(i) The hybrid models increased the accuracy for sentiment analysis compared with a single model performance on all types of datasets, although the computation time of SVM models is longer.
(ii) The combination helped to take advantage of the strengths of CNN, LSTM, and SVM, where CNN has the capability to extract characteristics, LSTM has capability to store past information at the state nodes (cell state), and SVM has capability to classify.
(iii) Using SVM as the classification method improved the results of both L-CNN and C-LSTM. SVM is effective in multidimensional data stratification and helps minimize local minima of neural networks.
There is very little work evaluating the computational complexity of deep learning models although there is some proposal [84] that considers some factors, such as the number of layers, the size of the input matrix, and other factors depending on the specific algorithm. In CNN, the number and size of convolution kernels and the number of output channels of each layer are considered. In view of this, it is clear that the higher reliability of hybrid models comes at the cost of higher complexity. Since time is one of the most valuable resources and the most taken into account when evaluating the performance of algorithms, we include the analysis of the computational time of the models involved in the comparative study, as this is a reflection of the time complexity.

It contains the time processing required for all datasets involved in the experiments. Processing time is calculated for the entire process of training and testing models using Word2vec and BERT. It includes time for data division and time to create the classification model (initialize the number of layers of the neural network, the number of nodes per layer, etc.) but does not include the time used to display the classification results. When using the hybrid models with the BERT technique for feature extraction, the accuracy generally is higher than with Word2vec, but the processing time is longer. In general, the hybrid methods provide better results than single deep learning models. Most hybrid networks provide higher (or equal) scores in all datasets. Moreover, from the good result of Maltoudoglou et al. [49] (Table 9), we saw that the feature extraction plays an important role in sentiment classification.

6. Conclusions

In this paper, we proposed the use of hybrid deep learning models for sentiment analysis from social network data. We tested the performance of mixing SVM, CNN, and LSTM, using two-word embedding techniques, Word2vec and BERT, on eight textual datasets of tweets and reviews. Afterwards, we compared four generated hybrid models with single models. These experiments are conducted to understand the adaptability of hybrid models, whether hybrid approaches can adapt in a wide range of dataset types and sizes. We studied the influence of different types of datasets, feature extraction techniques, and deep learning models on reliability of sentiment polarity analysis. Our experiments reveal that the reliability of hybrid models outperformed among all tested models for sentiment polarity analysis. Combining deep learning models with the SVM technique yields better results than using an individual model for performing sentiment analysis. In most of the tested datasets, the reliability of hybrid models using SVM is higher than that of the ones not using it; however, the computational time is much longer for the ones with SVM. We also observed that the effectiveness of the algorithms depends largely on the characteristics and quality of the datasets. We are aware that the context of the dataset has a large impact on the choice of sentiment analysis models. We intend to study the performance of hybrid approaches for sentiment analysis on hybrid datasets and multiple or hybrid contexts in order to gain deeper insight in a specific topic, such as business, marketing, or medicine. Its application derives from associating sentiments to relevant context in order to provide detailed personal feedback and recommendation for users

REFERENCES:

[1] L. Yang, Y. Li, J. Wang and R. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning", IEEE Access, vol. 8, pp. 23522-23530, 2020. DOI: 10.1109/access.2020.2969854.
[2] J. Park, "Framework for Sentiment-Driven Evaluation of Customer Satisfaction With Cosmetics Brands", IEEE Access, vol. 8, pp. 98526-98538, 2020. DOI: 10.1109/access.2020.2997522.

[3] T. U. Haque, N. N. Saber, and F. M. Shah, “Sentiment analysis on large scale Amazon product reviews,” 2018 IEEE Int. Conf. Innov. Res. Dev. ICIRD 2018, no. May, pp. 1–6, 2018, DOI: 10.1109/ICIRD.2018.8376299.
[4] N. Nandal, R. Tanwar and J. Pruthi, "Machine learning based aspect level sentiment analysis for Amazon products", Spatial Information Research, vol. 28, no. 5, pp. 601-607, 2020. DOI: 10.1007/s41324-020-00320-2.
[5] P. Jain, R. Pamula and G. Srivastava, "A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews", Computer Science Review, vol. 41, p. 100413, 2021. DOI: 10.1016/j.cosrev.2021.100413.
[6] M. Hu and B. Liu, “Mining and summarizing customer reviews,” KDD-2004 - Proc. Tenth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 168–177, 2004, DOI: 10.1145/1014052.1014073.
[7] X. Fang and J. Zhan, “Sentiment analysis using product review data,” J. Big Data, vol. 2, no. 1, 2015, DOI: 10.1186/s40537-015-0015-2.
[8] K. Jindal and R. Aron, "A systematic study of sentiment analysis for social media data", Materials Today: Proceedings, 2021. DOI: 10.1016/j.matpr.2021.01.048.
[9] J. Keilwagen, I. Grosse, and J. Grau, “Area under precision-recall curves for weighted and unweighted data,” PLoS One, vol. 9, no. 3, pp. 1–13, 2014, DOI: 10.1371/journal.pone.0092209.
[10] Z. Liu, L. Liu, and H. Li, “An Empirical Study of Sentiment Analysis for Chinese Microblogging,” Elev. Wuhan Int. Conf. E-bus., 2012.
[11] J. R. Ragini, P. M. R. Anand, and V. Bhaskar, “Big data analytics for disaster response and recovery through sentiment analysis,” Int. J. Inf. Manage., vol. 42, no. September 2017, pp. 13–24, 2018, DOI: 10.1016/j.ijinfomgt.2018.05.004.
[12] P. Singh, R. S. Sawhney, and K. S. Kahlon, “Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government,” ICT Express, vol. 4, no. 3, pp. 124–129, 2018, DOI: 10.1016/j.icte.2017.03.001
[13] P. Pugsee, P. Sombatsri, and R. Juntiwakul, “Satisfactory analysis for cosmetic product review comments,” ACM Int. Conf. Proceeding Ser., vol. Part F1287, pp. 0–5, 2017, DOI: 10.1145/3089871.3089890.
[14] D. A. Kristiyanti and M. Wahyudi, “Feature selection based on Genetic algorithm, particle swarm optimization and principal component analysis for opinion mining cosmetic product review,” 2017 5th Int. Conf. Cyber IT Serv. Manag. CITSM 2017, 2017, DOI: 10.1109/CITSM.2017.8089278.
[15] P. Pugsee, V. Nussiri, and W. Kittirungruang, Opinion mining for skin care products on twitter, vol. 937. Springer Singapore, 2019.
[16] R. Ren, D. D. Wu, and D. D. Wu, “Forecasting stock market movement direction using sentiment analysis and support vector machine,” IEEE Syst. J., vol. 13, no. 1, pp. 760–770, 2019, DOI: 10.1109/JSYST.2018.2794462.
[17] N. Thessrimuang and O. Chaowalit, “Opinion representative of cosmetic products,” 20th Int. Comput. Sci. Eng. Conf. Smart Ubiquitos Comput. Knowledge, ICSEC 2016, 2017, DOI: 10.1109/ICSEC.2016.7859945.
[18] T. Chatchaithanawat and P. Pugsee, “A framework for laptop review analysis,” ICAICTA 2015 - 2015 Int. Conf. Adv. Informatics Concepts, Theory Appl., 2015, DOI: 10.1109/ICAICTA.2015.7335358.
[19] E. Haddi, X. Liu, and Y. Shi, “The role of text pre-processing in sentiment analysis,” Procedia Comput. Sci., vol. 17, pp. 26–32, 2013, DOI: 10.1016/j.procs.2013.05.005
[20] Y. Zhang, R. Jin, and Z. H. Zhou, “Understanding bag-of-words model: A statistical framework,” Int. J. Mach. Learn. Cybern., vol. 1, no. 1–4, pp. 43–52, 2010, DOI: 10.1007/s13042-010-0001-0.
[21] B. K. Bhavitha, A. P. Rodrigues, and N. N. Chiplunkar, “Comparative study of machine learning techniques in sentimental analysis,” Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2017, no. Icicct, pp. 216–221, 2017, DOI: 10.1109/ICICCT.2017.7975191.
[22] J. Keilwagen, I. Grosse, and J. Grau, “Area under precision-recall curves for weighted and unweighted data,” PLoS One, vol. 9, no. 3, pp. 1–13, 2014, DOI: 10.1371/journal.pone.0092209.
[23] Hossen, M.S.; Jony, A.H.; Tabassum, T.; Islam, M.T.; Rahman, M.M.; Khatun, T. Hotel review analysis for the prediction of business using deep learning approach. In Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 25–27 March 2021; pp. 1489–1494
[24] Vimali, J.; Murugan, S. A Text Based Sentiment Analysis Model using Bi-directional LSTM Networks. In Proceedings of the 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 8–10 July 2021; pp. 1652–1658.
[25] Vadivukarassi, M.; Puviarasan, N.; Aruna, P. An exploration of airline sentimental tweets with different classification model. Int. J. Res. Eng. Appl. Manag. 2018, 4, 72–77.
[26] Dholpuria, T.; Rana, Y.; Agrawal, C. A Sentiment analysis approach through deep learning for a movie review. In Proceedings of the 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 24–26 November 2018; pp. 173–181.
[27] Maas, A.; Daly, R.E.; Pham, P.T.; Huang, D.; Ng, A.Y.; Potts, C. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 142–150.
[28] Harjule, P.; Gurjar, A.; Seth, H.; Thakur, P. Text classification on Twitter data. In Proceedings of the 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), Jaipur, India, 7–8 February 2020; pp. 160–164.