

A simplified classification computational model of opinion mining using deep learning

Rajeshwari Dembala¹, Ananthapadmanabha Thammaiah²

¹Department of Information Science and Engineering, The National Institute of Engineering, Mysuru, India

²Mysore University School of Engineering, University of Mysore, Mysuru, India

Article Info

Article history:

Received Jul 25, 2023

Revised Oct 17, 2023

Accepted Nov 4, 2023

Keywords:

Bidirectional long short-term memory

Natural language processing

Natural text data

Opinion classification

Preprocessing

ABSTRACT

Opinion and attempts to develop an automated system to determine people's viewpoints towards various units such as events, topics, products, services, organizations, individuals, and issues. Opinion analysis from the natural text can be regarded as a text and sequence classification problem which poses high feature space due to the involvement of dynamic information that needs to be addressed precisely. This paper introduces effective modelling of human opinion analysis from social media data subjected to complex and dynamic content. Firstly, a customized preprocessing operation based on natural language processing mechanisms as an effective data treatment process towards building quality-aware input data. On the other hand, a suitable deep learning technique, bidirectional long short term-memory (Bi-LSTM), is implemented for the opinion classification, followed by a data modelling process where truncating and padding is performed manually to achieve better data generalization in the training phase. The design and development of the model are carried on the MATLAB tool. The performance analysis has shown that the proposed system offers a significant advantage in terms of classification accuracy and less training time due to a reduction in the feature space by the data treatment operation.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Rajeshwari Dembala

Department of Information Science and Engineering, The National Institute of Engineering

Mysuru, India

Email: rajeshwaridresearch@gmail.com

1. INTRODUCTION

Opinion analysis is used for many purposes, such as determining the mood of social media users about a topic, their views on social events, market price, and products [1], [2]. On the other hand, Twitter is widely preferred as a data source in opinion and sentiment analysis studies because it is a popular social network and convenient for collecting data in different languages and content [3]. However, considering opinion analysis as a text and sequence classification problem. However, the social media data comprises short texts and dynamic representation, so the corpus becomes sparse and semi-unstructured [4]. This poses a significant problem regarding system response time and classification performance, especially on large text corpus [5]. For this reason, various preprocessing, text representation, and data modelling techniques are used to address the issues associated with classification performance arising from sparse data quality and high feature space. Text representation can be realized with meaningful information extracted from text content in traditional methods such as bag of words (BoW) skip-gram and N-grams [6]–[8]. In the BoW model, attributes are words extracted from text content, and the order of these is not much important. The n-gram model, which can be applied at the word and character level, is generally more successful at the character level and is robust against situations such as spelling mistakes and using abbreviations because it is language-

independent. On the other hand, the skip gram is an unsupervised mechanism that determines the most relevant words for a given text. Structured or semi-structured texts' structural and statistical properties are also used in text representation. In addition to these traditional methods, various methods based on graphs, linear algebra, and latent Dirichlet allocation (LDA) [9], [10] are used to address text classification problems. In addition, there are studies where the prevalent word-to-vectors method is also used in text representation in lower space [11]. In the word-to-vector method, an n-dimensional vector of each word seen in the document is obtained and clustered [12]. Documents are represented with the number of members (words) in each set of the respective document, and thus texts are represented in lower dimensions. With the advancement in machine learning and deep learning techniques, the existing literature presents various models for mining opinions from the text and, more specifically, from social media data. Among these, recurrent neural networks (RNN) and long-short-term memory (LSTM) are widely adopted in the context of opinion classification from the rich natural language [13], [14].

The RNN and LSTM are the most suitable model for the sequence classification problems. Opinion mining can also be regarded as a sequence classification task, as the text sentences consist of multiple chunks of a word to represent meaningful information. The work carried out by Pergola *et al.* [15] suggested a deep learning model for the topic-oriented attention system towards sentiment analysis. The outcome shows that adopting the RNN offers better accuracy in the prediction phase. Ma *et al.* [16] implemented a variant of RNN, namely LSTM, with a layered attention mechanism for opinion mining. The study exhibited that their model outperforms the existing models for aspect-based opinion analysis. Xu *et al.* [17] designed an advanced word representation technique based on the weighted word vectors and implemented a Bi-LSTM model with a feedforward neural network to classify the sentiment from the comment data. The work done by Alattar and Shaalan [18] presented a filtered-LDA model to reveal sentiment variations in the Twitter dataset. The model adopts various hyperparameters to obtain reasons that cause sentiment variations. Fu *et al.* [19] have suggested an enhanced model that uses an LSTM model combination of sentiment and word embedding to represent better the words followed by an attention vector. Jiang *et al.* [20] proposed a bag-of-words text representation method based on sentiment topic words composed of the deep neural network, sentiment topic words, and context information and performed well in sentiment analysis. Pham and Le [21] suggested a combined approach of multiple convolutional neural networks (CNN), emphasizing word embeddings using different natural language processing (NLP) mechanisms such as Word2Vec, GloVe, and the one-hot encoding. Han *et al.* [22] developed an advanced learning model based on joint operation CNN and LSTM for text representation. The work of Majumder *et al.* [23] exhibited the correlation between sarcasm recognition and sentiment analysis. The authors have introduced a multitasking classification model that improves sarcasm and sentiment analysis tasks. The study of Rezaeinia *et al.* [24] presented an improved word embedding technique based on part-of-speech tagging and sentiment lexicons. This method provides a better form of performance regarding sentiment analysis. Pressurized water reactor (PWR) alerting model built using an LSTM-based neural network is presented by Liu *et al.* Using simulated data from a PWR and analysis methods for both predictable and uncertain data, the model was trained and tested. A combination of deterministic assessment and uncertainty evaluation, the precision of the LSTM model is 96.67% and 98.7%, respectively [25].

Based on the literature review, it has been analyzed that various research works have been done in the context of opinion classification. The prime outline of problem from existing scheme is found that they are more focused on simple preprocessing operations such as tokenization, removal of punctuation, and stop words, which do not provide effective data modelling and are not enough to deal with the high feature space complexity in the training phase of the learning model. Also, an effective data treatment process is one of the primary steps contributing to higher classification accuracy. Most of the existing contributors are found not to emphasize effective data modelling prior to training. Majority of the contributors mainly focuses on adopting deep learning approach and subject them to dataset to arrive at conclusive remarks without much emphasizing on the granularity of mining issues related to dynamic form of opinion. Instead, they just did the regular preprocessing operation. The current research effectively handles the significant problems associated with data quality and classification accuracy. Therefore, the statement of unsolved problem is “developing a simplified and cost-effective analytical model for efficient opinion mining is quite a challenging task knowing the complexities associated with existing learning-based approaches”.

In order to differentiate the proposed study model for addressing the identified research problem, it is now designed with certain contributing characteristics that make them separate from existing models viz: i) Exploratory analysis is carried out to understand the characteristics of data and the need for preprocessing operations, ii) suitable data treatment operation is carried out by performing customized cleaning and data filtering operations based on the requirement of preprocessing, iii) data modelling and preparation are done by performing manual data truncation and padding processes to maintain the length of the text sentences in the training dataset, and iv) implementation of Bi-LSTM learning model followed by suitable training parameters.

The outline of the manuscript is as follows: section 2 discusses an adopted research methodology and implementation procedure adopted in system design and development which showcases the novel cost-effective contribution towards opinion mining. This section highlights the dataset, its cleaning, treatment, and suitable modelling and development of the deep learning model for opinion classification. Section 3 presents the discussion of outcome and performance assessment of the proposed system to showcase its effectiveness and its distinction from existing system, and finally, section 4 concludes the entire work discussed in this paper.

2. PROPOSED METHOD

The prime agenda of the proposed scheme is to evolve up with a novel computational model which is capable to extract essential information using opinion mining approach. The proposed scheme harnesses the potential of deep learning approach in order to carry out the classification of significant opinion. Along with the above value added in proposed scheme, the proposed study aims to present an enhancement in opinion classification from the natural language data (text) using customized preprocessing operations and a suitable deep learning approach. In order to perform opinion analysis, the study considers text data generated on the social media platform, which consists of dynamic information (text, symbol, number, and punctuation). However, the accuracy of any text data-based classification system highly depends on the data quality. Suppose the dataset is ambiguous and consists of complex and dynamic information. In that case, the machine learning or deep learning model may deliver misleading, biased outcomes and seriously harm decision-making processes. In this regard, the proposed study presents customized data preprocessing operation in order to provide a suitable treatment and cleaning process to the text dataset captured from social media. To date, various deep learning algorithms or models are present with their advantages and limitations. However, selecting a suitable deep learning model becomes another significant problem in the context of human behavior (opinion) analysis. Therefore, the current study considers opinion classification as a sequence classification problem since the current study deals with text sentences and attempts to implement a class of recurrent neural networks. The schematic architecture of the proposed system is shown in Figure 1. The proposed model's design and development are carried out to better and more accurately identify people's viewpoints towards entities such as events, topics, products, services, organizations, individuals, and issues. This section presents an exploratory analysis of the dataset and strategy adopted in preprocessing the text data. Also, an implementation procedure is discussed for the deep learning-based opinion mining process followed by the word embedding process and model training.

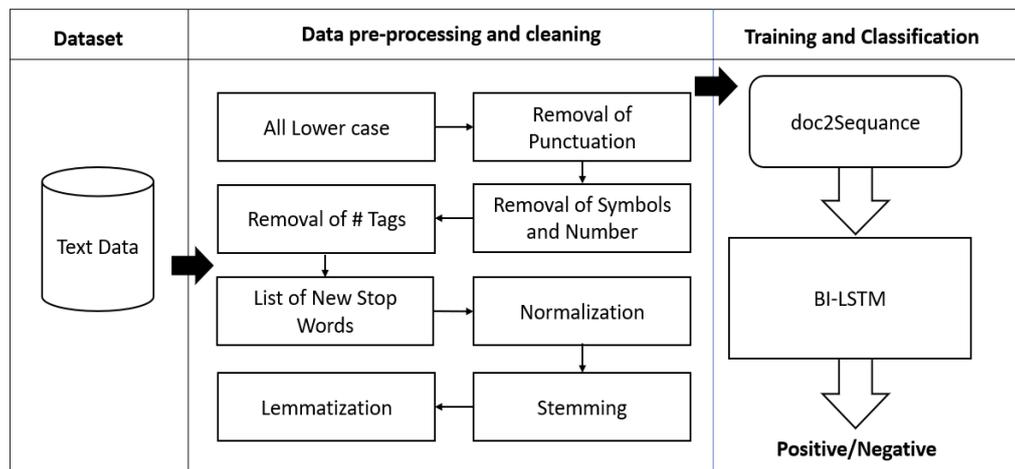


Figure 1. Schematic architecture of the proposed system

It is essential to understand the significant value-added features and their impact towards betterment of proposed opinion mining model. It was already seen that existing studies have complex form of applying deep learning approach where not much preference is offered towards improving data quality. However, proposed model emphasized strongly towards ensuring effective data quality. By adopting this methodology, the scheme not only increases its accuracy but also reduces the computational burden on deep learning approach. More detailing of prime operations under this adopted methodology are as follows:

2.1. Data visualization and analysis

The proposed study makes use of a social media dataset for opinion mining. Table 1 highlights a few samples of data to understand the characteristics and complexity of the dataset considered in the current work from the natural language processing viewpoint. Specifically, the Twitter dataset is considered, which is downloaded from the Kaggle. Twitter data are a unique form of text data that reflects information, opinion, or attitude that the users share publicly. The rationale behind considering tweets for the analysis is that the tweet texts are associated with dynamic representation and are semi-unstructured because they contain different forms of text representation, as shown in Table 1.

The user shares their thoughts in native format, especially with the different styles, containing numbers, digits, punctuation, and subjective context. It is to be noted that some texts are small, and some are capitalized between the sentences. The dataset considered in this study is labelled where each text belongs to two different opinion contexts, i.e., 0 and 1, where zero means negative opinion, and one is subjected to positive opinion. In order to make the dataset more friendly, the labels are updated from numerical to categorical, as shown in Table 2.

The distribution of the opinion class is shown in Figure 2, where 2,464 texts are subjected to the negative opinion class, and 3,204 texts are subjected to the favorable opinion class. Based on the analysis, it is identified that the dataset is imbalanced with a difference of 740, which may lead the learning model to be biased towards positive opinions in the classification process. In order to deal with this imbalance factor, the proposed study focuses more on treating the unstructured and rich natural language (text). Therefore, the proposed study performs no sampling or scaling (up and downscaling) operation over the text data. Instead, it executes a preprocessing operation from the viewpoint of feature engineering, which will precisely represent the input text data and enable better generalization in the training phase of the learning model.

Table 1. Visualization of text data samples

SI. No	Text	Label
1	Friday hung out with Kelsie, and we went and saw The Da Vinci Code SUCKED!!!!	0
2	Harry Potter is AWESOME I don't care if anyone says differently!	1
3	<---Sad level is 3. I was writing a massive blog tweet on Myspace, and my comp shut down. Now it is all lost *lays in fetal position*	0
4	BoRinG): What is wrong with him?? Please tell me.....:-/	0
5	I want to be here because I love Harry Potter and want a place where people take it seriously, but it is still so much fun.	1

Table 2. Visualization of updated label

SI. No	Existing label	Updated label
1	0	Negative
2	1	Positive
3	0	Negative
4	0	Negative
5	1	Positive

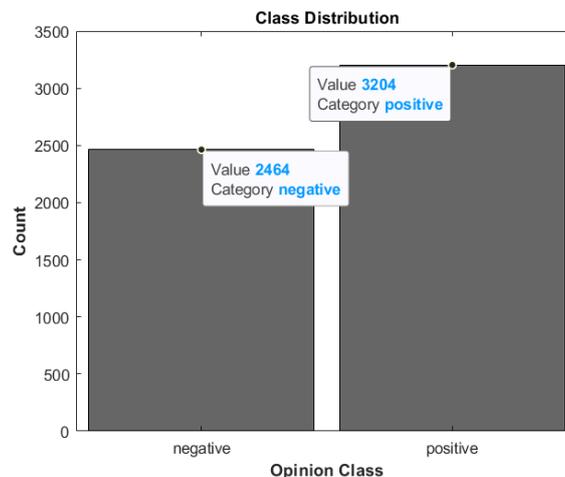


Figure 2. Distribution of the opinion class

In the current analysis, approximately 5,000 texts were considered, split into training, validation, and testing set. Initially, the dataset is split into an 80:20 ratio, where 80% is considered for the training set, and the remaining 20% of the dataset is considered for the testing set. Again, the training set is split into an 80:20 ratio where 80% is kept for model training, and 20% is considered for validation. However, this is a conventional split to ensure that predictive modelling could be simpler to carry out benchmarking with other learning approaches which approximately adopts similar splitting.

2.2. Preprocessing

The discussion in the previous section was all about data visualization and preliminary analysis, which shows that the text dataset consists of dynamic contents, semi-structured sentences, punctuation, numbers, short words/text, and stop words. Therefore, an effective preprocessing operation is required to provide accurate treatment and cleaning operations to correct the text data in a suitable form. In this regard, the preprocessing operation is executed based on the NLP mechanism, and some customized data modelling is carried out to achieve better precision in the data treatment process. The entire data modelling and treatment process is executed in the following manner as shown in Algorithm 1 discussed as follows: i) Removal of unconnected term: In this step of execution, the algorithm considers the elimination of punctuations, URLs, tickers, hashtags (#), numbers, digits, memorable characters, and extra-wide spaces, and the removal of short sentences whose length is limited to only two words. Also, all the text phrases are transformed to lowercase, and emoticons are changed to relevant words; ii) Tokenization: Further tokenization of the text data is carried out, an essential step of the NLP. In this process, the text sentences are split into multiple and smaller elements called tokens. These tokens are strings with known meanings that help in understanding the context; iii) Modelling of stop words: In this process, the proposed study performs modelling of stop words that could carry significant meaning. However, the stop words have less entropy value that does not describe the contextual meaning; therefore, they can be discarded from the text sentences. However, in the proposed preprocessing operation, a customized operation is carried out, where a list of the stop words is edited to choose the stop words (such as are, are not, aren't, did, didn't, and many more) that could carry significant meaning for the opinion mining or classification; and iv) Stemming/lemmatization: stemming and lemmatization are essential steps of text normalization after executing the above data treatment procedures. The stemming operation over text provides a stem representation of the text. Similarly, lemmatization is also performed to get a base form of the text according to the part-of-speech (POS) and grammar protocol. The computing procedure for Twitter dataset preprocessing is discussed in the Pseudo constructs as follows:

Algorithm 1. For data preprocessing

```

Input:Text dataset (T)
Output:Preprocessed Texts (PT)
Start:
Init DF
1. Load:DF → f1(filename)
2. DF.Lable→f2( Label rename:{'0','1'}{'negative','positive'})
3. Split DF: [Training, Test] ← f3(DF,0.2)
4. Split Training: [Train, Val] ← f3(Training,0.2)
5. deffunction: dataclean(text)
6. Text_lower = re.findall('DF.text', '(.[a-z][A-Z])') do
7. lower(DF.text)
8. Text= f4('DF.text', '@\w+', '#', RT[\s] +, 'https?:\S+')
9. Tok= f5(DF. Text)
10. new_stopwords= stopwords
11. Init N // list of not stop words
12. new_stopwords (ismember(new_stopwords, N))=[];
13. Tok= f6(DF. Text, new_stopwords)
14. DF. Text= f7(Tok, Stem)
15. DF. Text= f7(Tok, Lem)
16. return = PT
17. Foreach text from Train do
18. DFp = dataclean(train.text)
End

```

A rich natural language (text) requires an effective data treatment operation to perform precise classification or analysis of the opinion. The above-mentioned algorithmic steps exhibit a vital operation of correcting the raw text data for further analysis. The algorithm takes an input of raw text data (T), and after undergoing several stages of preprocessing operation, it provides precise and cleaned data (PT). Initially, an empty vector is initialized as data frame (DF) that stores text data in a structured format using function $f1(x)$ (line-1). Further, in the next step, the label field of the dataset gets updated in a more friendly manner using

2.3. Opinion classification using a deep learning model

The classification of a public opinion involves several procedures viz. i) word encoding, ii) data padding and truncating, iii) constructing and training deep learning model, and iv) validation of trained model via introducing new text data from the testing set. This section discusses modelling the learning model to perform opinion classification followed by an algorithmic approach. The text data is inherently a combination of sequences of words, which have a dependency, which means they are interconnected via directed links that reveal the association owned by the connected words. Therefore, the proposed study implements a specific class of deep learning techniques: long-short-term memory (LSTM), which is most efficient for learning long-term dependencies between sequences of text sentences. Also, classifying opinions from the sequential data corpus (text dataset) can be treated as a sequence classification problem. The LSTM suits the other deep learning model and shallow machine learning techniques better. LSTM is an improved class of recurrent neural networks (RNN), designed to deal with sequential data by distributing their weights across the sequence. LSTM addresses the problem of gradient vanishing by employing its gate mechanisms and captures long-term relationships. The LSTM can be numerically defined as (1):

$$h_t = f(W_h \cdot x_t + U_t \cdot h_{t-1} + b_h) \tag{1}$$

where x_t denotes current input sequence data, h_t denotes a hidden state of the neural network, W_h and U_t refers to the weights, and b_h denotes bias. The $f(x)$ is a non-linear function (i.e., tangent function) to learn and classify operations. Though LSTM has many advantages, it is also subjected to a limitation that it suffers in considering post-word information as the sequences of the text data are read in a single feedforward direction. Therefore, the proposed study implements a bidirectional LSTM (Bi-LSTM) learning model whose outputs are stacked together, one for forward and one for backward. The hidden states of the feedforward LSTM unit (h_t^F) and hidden states of backward LSTM units (h_t^B) are concatenated to form a single hidden layer of Bi-LSTM (h_t^{Bi}) numerically expressed as (2).

$$h_t^F \oplus h_t^B \rightarrow h_t^{Bi} \tag{2}$$

Figure 4 shows the architecture of the proposed deep learning Bi-LSTM model for the opinion classification from the natural language (text). In order to train the learning model, the sequence of input preprocessed text data needs to be converted into numeric sequences. The study uses a word encoding mechanism that maps the training dataset into integer sequences. The encoding technique adopted in the proposed study is one-hot encoding vectorization. In addition, padding and truncation are carried out to make text data of the same length. However, there is an option in the training process to pad and shorten input sequences automatically. However, this option does not effectively applicable to word vector sequences.

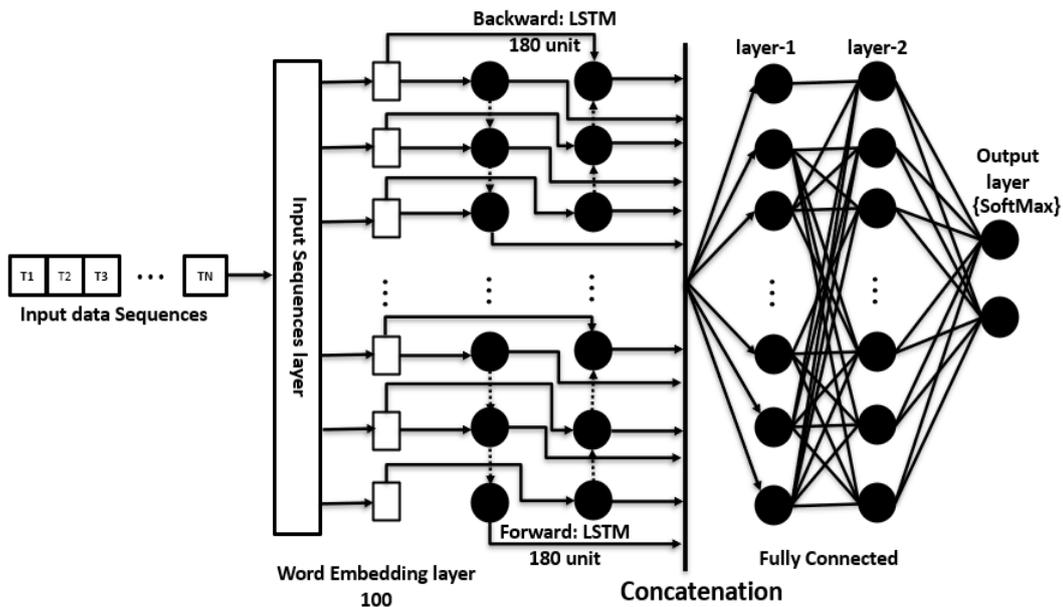


Figure 4. Proposed deep learning model Bi-LSTM for opinion classification

Therefore, the study performs this process manually by determining the length of text sentences, and then the text sequences that are longer than identified target value are truncated, and the sequences shorter than the target value are left-padded. A histogram plot in Figure 5 shows the length of the text sentences that belongs to the training dataset. Based on the analysis from Figure 5, it is analyzed that most of the text sentences have fewer than 23 chunks. Therefore, this length will be considered as a target length to truncate and pad the training dataset.

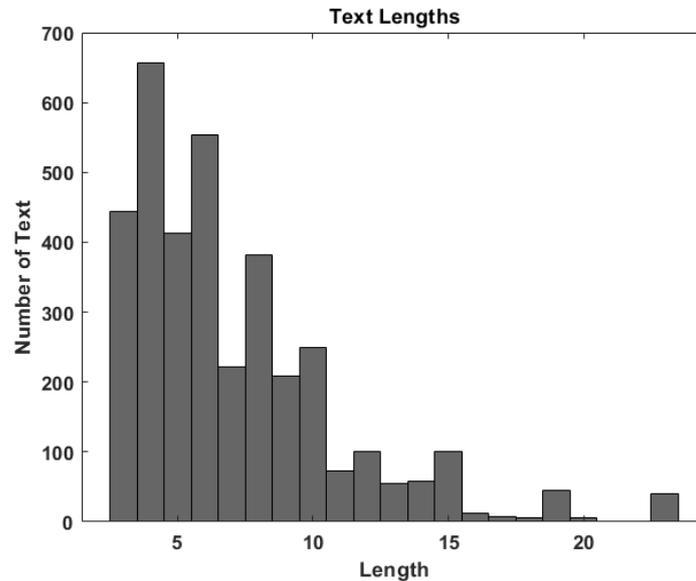


Figure 5. Visualization of the text length

The study also considers embedding layers in modelling the Bi-LSTM learning model. Including a word embedding layer will map a word in the lexicon to numerical vectors instead of a scalar. Also, it obtains semantic information about text phrases, which eventually maps the text phrase or word with similar meanings to have similar vectors. After adding the word embedding layer, a Bi-LSTM layer is considered and added with 180 hidden units as shown in Algorithm 2. Finally, two fully connected layers and one output layer with a softmax function are added for a sequence-to-label classification. The configuration details of the model development and Training options are highlighted in Tables 4 and 5, respectively.

Algorithm 2. For opinion classification

Input: Preprocessed Texts (PT)

Output: Opinion: Positive (P) or Negative (N)

Start:

1. Load: $DF \rightarrow f1(\text{filename})$
 2. Apply Algorithm 1: Preprocessing Training, and validation dataset
 3. $\text{train}_x \leftarrow \text{dataclean}(\text{train.text})$
 4. $\text{val}_x \leftarrow \text{dataclean}(\text{train.text})$
 5. Prepare data for model training
 6. Execute word encoding
 7. $\text{enc} \leftarrow f7(\text{train}_x)$
 8. do truncating and padding
 9. Compute: $\text{target_length} \leftarrow f_{\max}(\text{length}(\text{train}_x))$
 10. $\text{train}_x \leftarrow f8(\text{enc}, \text{train}_x, \text{target_length})$
 11. $\text{val}_x \leftarrow f8(\text{enc}, \text{val}_x, \text{target_length})$
 12. Execute Model Development
 13. Init, I (input layer size), D(), O(), N(), C()
 14. Config layers
 15. Input layer (I)
 16. Embedding layer (D,N)
 17. Bi-LSTM (O, outputmode, 'last')
 18. Fully connected (C)
 19. Output layer (activation function, softmax)
- End

Table 4. Configuration details of the model development

6×1 layers of the proposed learning models		
1	Sequence input	Sequence input with 1 dimension
2	Word embedding layer	Word embedding layer with 100 dimensions
3	Bi-LSTM	180 hidden cells
4	Fully connected	2
5	Classification output	Softmax
6	Loss function	crossentropyex

Table 5. Details of the model training option

Training details		
1	Optimizer	Adam
2	MiniBatchSize	32
3	InitialLearnRate	0.01
4	GradientThreshold	1
5	MaxEpochs	100
6	L2Regularization	0.0006

3. RESULTS AND DISCUSSION

This section discusses the outcome obtained and performance analysis of the proposed learning model for opinion classification. The selection of the performance metric was accomplished on the basis of existing approaches using similar evaluation metric reported in literatures. The core idea of selection of this performance metric is to justify the appropriateness of the adopted predictive approach of deep learning. The analysis of the proposed system performance is evaluated concerning the accuracy, precision, recall rate, and F1-score, briefly described as follows:

3.1. Performance indicator

The assessment of the proposed study model is carried out using a standard performance metric of accuracy, precision, recall, and F1-Score. These are the standard metric which are universally adopted in research work towards assessing the predictive accuracies. As the proposed scheme uses deep learning scheme, hence it becomes more inevitable to adopt the standard metric towards proper study evaluation. The empirical formulation of these metrics is:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}, Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}$$

$$and F1_Score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall} \right)$$

TP is true positive (correctly classified positive value, i.e., actual and classified class=yes)

TN is true negative (correctly classified negative value i.e., actual and classified class=no)

FP is false positive (Actual class=no and classified class=yes)

FN is false negative (Actual class=yes and classified class=no)

The Figure 6 shows a Heatmap of the confusion matrix, which assesses the output of the trained Bi-LSTM model. A closer analysis of the Heatmap shows that out of 1,133 text data, 640 text data belong to true positive, meaning that 640 were classified as positive opinion where there is a total of 645 text that reflects positive opinion. Also, 481 texts were classified as true unfavorable, with 489 texts reflecting negative opinions. In order to better understand the performance of the proposed system, the study performs comparative analysis with other machine learning models concerning multiple performance parameters such as accuracy, precision, recall rate, and F1 score. Table 6 highlights the quantified outcome of the proposed Bi-LSTM-based model and other machine-learning models.

The Figure 7 shows a comparative analysis to validate the proposed work's effectiveness and scope. The comparative analysis concerns the proposed Bi-LSTM, LSTM, support vector machine (SVM), and probabilistic model, i.e., naïve Bayes. The classification accuracy, precision, Recall and F1_Score are analyzed and illustrated in Figures 7(a) to 7(d) respectively. It can be seen that the proposed Bi-LSTM model outperforms LSTM and shallow machine learning models. The SVM requires ample training time, which makes it inefficient and computationally expensive. Also, the SVM is not much scalable to have high feature space in the training process. For the naïve Bayes, it depends on an often-faulty consideration of equally essential and independent features, leading to biases in the prediction.

In addition, the proposed system based on Bi-LSTM is enriched with proposed preprocessing operation that overcome the lexical sparsity and ambiguity issue in the training dataset. This is one of the reasons that attributes towards better data generalization in the training phase. The significance of the

proposed work is the effective data treatment which improves the quality of data and adequate modelling of deep learning technique with suitable training parameters for the precise classification, which considers the contextual information by dealing with both forward and backward dependencies.

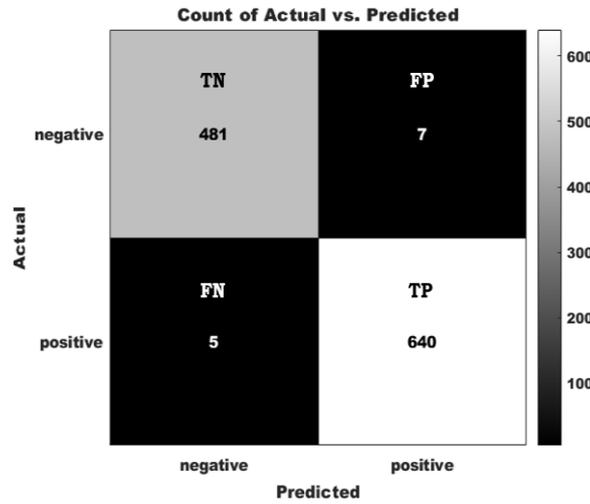


Figure 6. Heatmap of the confusion matrix

Table 6. Quantified outcome for comparative analysis

Model	Accuracy	Precision	Recall	F1_Score
SVM	85.018%	81.815%	78.961%	83.931%
Naïve Bayes	72.235%	79.81%	77.481%	80.063%
LSTM	94.975%	90.11%	94.36%	91.321%
Bi-LSTM	98.940%	98.918%	99.224%	99.071%

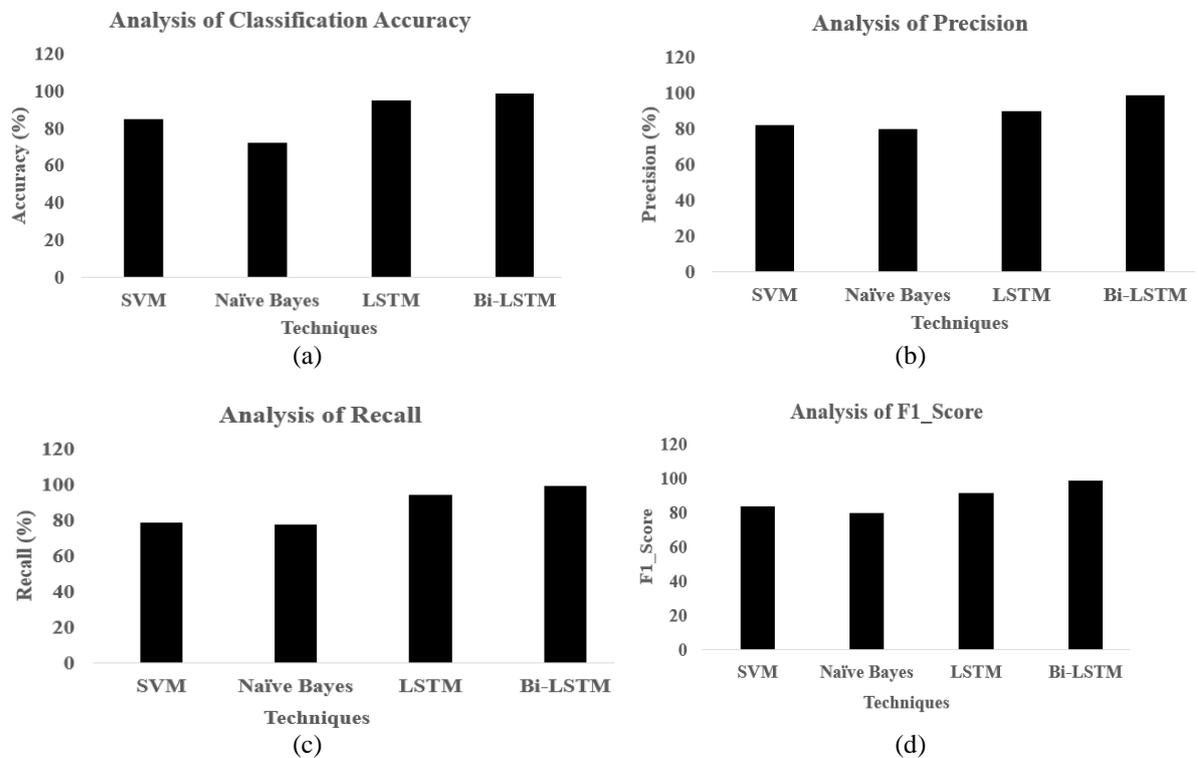


Figure 7. Comparative analysis of (a) classification accuracy, (b) precision, (c) Recall and (d) F1_Score

4. CONCLUSION

The proposed work has addressed the opinion classification problem for rich natural language. The proposed study exploits the advantage of utilizing a deep learning technique and effective data treatment on the performance of the opinion analysis from the rich and dynamic text data. A Bi-LSTM learning model is used to capture contextual information to generalize textual features better. The feature space complexity is reduced significantly before the training process, which is carried out in the preprocessing and after preprocessing, where the padding and the truncating process are performed manually. The results show the scope and effectiveness of Bi-LSTM in the context of sequence classification problems. The proposed system achieved good accuracy, precision, recall rate, and F1-measure results over the existing learning models. The proposed study may extend towards processing different languages emphasizing different lexical approaches in future work.

REFERENCES

- [1] M. Khan, A. Malviya, and S. K. Yadav, "Big data and social media analytics-a challenging approach in processing of big data," in *Lecture Notes in Electrical Engineering*, vol. 698, Springer Nature Singapore, 2021, pp. 611–622.
- [2] A. D. Cheok, B. I. Edwards, and I. O. Muniru, "Human behavior and social networks," in *Encyclopedia of Social Network Analysis and Mining*, Springer New York, 2018, pp. 1025–1034.
- [3] B. Calabrese, M. Cannataro, and N. Ielpo, "Using social networks data for behavior and sentiment analysis," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9258, Springer International Publishing, 2015, pp. 285–293.
- [4] B. Le and H. Nguyen, "Twitter sentiment analysis using machine learning techniques," in *Advances in Intelligent Systems and Computing*, vol. 358, Springer International Publishing, 2015, pp. 279–289.
- [5] A. Sharma and U. Ghose, "Sentimental analysis of Twitter data with respect to general elections in India," *Procedia Computer Science*, vol. 173, pp. 325–334, 2020, doi: 10.1016/j.procs.2020.06.038.
- [6] M. Kanakaraj and R. M. R. Guddeti, "NLP based sentiment analysis on Twitter data using ensemble classifiers," in *2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN)*, Mar. 2015, pp. 1–5, doi: 10.1109/ICSCN.2015.7219856.
- [7] J. Mutinda, W. Mwangi, and G. Okeyo, "Lexicon-pointed hybrid N-gram features extraction model (LeNFEM) for sentence level sentiment analysis," *Engineering Reports*, vol. 3, no. 8, Aug. 2021, doi: 10.1002/eng2.12374.
- [8] Ó. Apolinario-Arzuabe, J. A. García-Díaz, J. Medina-Moreira, H. Luna-Aveiga, and R. Valencia-García, "Comparing deep-learning architectures and traditional machine-learning approaches for satire identification in Spanish tweets," *Mathematics*, vol. 8, no. 11, Nov. 2020, doi: 10.3390/math8112075.
- [9] F. A. Lovera, Y. C. Cardinale, and M. N. Homsí, "Sentiment analysis in Twitter based on knowledge graph and deep learning classification," *Electronics*, vol. 10, no. 22, Nov. 2021, doi: 10.3390/electronics10222739.
- [10] M. F. A. Bashri and R. Kusumaningrum, "Sentiment analysis using latent Dirichlet allocation and topic polarity word cloud visualization," in *2017 5th International Conference on Information and Communication Technology (ICOICT)*, May 2017, pp. 1–5, doi: 10.1109/ICOICT.2017.8074651.
- [11] Q. Chen and M. Sokolova, "Specialists, scientists, and sentiments: Word2Vec and Doc2Vec in analysis of scientific and medical texts," *SN Computer Science*, vol. 2, no. 5, Sep. 2021, doi: 10.1007/s42979-021-00807-1.
- [12] Y. Parikh, A. Palusa, S. Kasthuri, R. Mehta, and D. Rana, "Efficient Word2Vec vectors for sentiment analysis to improve commercial movie success," in *Lecture Notes in Electrical Engineering*, vol. 475, Springer Singapore, 2018, pp. 269–279.
- [13] H. Kaur, S. U. Ahsaan, B. Alankar, and V. Chang, "A proposed sentiment analysis deep learning algorithm for analyzing COVID-19 tweets," *Information Systems Frontiers*, vol. 23, no. 6, pp. 1417–1429, Dec. 2021, doi: 10.1007/s10796-021-10135-7.
- [14] A. Londhe and P. V. R. D. P. Rao, "Aspect based sentiment analysis – an incremental model learning approach using LSTM-RNN," in *Communications in Computer and Information Science*, vol. 1440, Springer International Publishing, 2021, pp. 677–689.
- [15] G. Pergola, L. Gui, and Y. He, "TDAM: a topic-dependent attention model for sentiment analysis," *Information Processing and Management*, vol. 56, no. 6, Nov. 2019, doi: 10.1016/j.ipm.2019.102084.
- [16] Y. Ma, H. Peng, and E. Cambria, "Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, pp. 5876–5883, Apr. 2018, doi: 10.1609/aaai.v32i1.12048.
- [17] X. Xu, T. Gao, Y. Wang, and X. Xuan, "Event temporal relation extraction with attention mechanism and graph neural network," *Tsinghua Science and Technology*, vol. 27, no. 1, pp. 79–90, Feb. 2022, doi: 10.26599/TST.2020.9010063.
- [18] F. Alattar and K. Shaalan, "Using artificial intelligence to understand what causes sentiment changes on social media," *IEEE Access*, vol. 9, pp. 61756–61767, 2021, doi: 10.1109/ACCESS.2021.3073657.
- [19] X. Fu, J. Yang, J. Li, M. Fang, and H. Wang, "Lexicon-enhanced LSTM with attention for general sentiment analysis," *IEEE Access*, vol. 6, pp. 71884–71891, 2018, doi: 10.1109/ACCESS.2018.2878425.
- [20] Z. Jiang, S. Gao, and L. Chen, "Study on text representation method based on deep learning and topic information," *Computing*, vol. 102, no. 3, pp. 623–642, Mar. 2020, doi: 10.1007/s00607-019-00755-y.
- [21] D. H. Pham and A. C. Le, "Exploiting multiple word embeddings and one-hot character vectors for aspect-based sentiment analysis," *International Journal of Approximate Reasoning*, vol. 103, pp. 1–10, Dec. 2018, doi: 10.1016/j.ijar.2018.08.003.
- [22] H. Han, X. Bai, and P. Li, "Augmented sentiment representation by learning context information," *Neural Computing and Applications*, vol. 31, no. 12, pp. 8475–8482, Dec. 2019, doi: 10.1007/s00521-018-3698-4.
- [23] N. Majumder, S. Poría, H. Peng, N. Chhaya, E. Cambria, and A. Gelbukh, "Sentiment and sarcasm classification with multitask learning," *IEEE Intelligent Systems*, vol. 34, no. 3, pp. 38–43, May 2019, doi: 10.1109/MIS.2019.2904691.
- [24] S. M. Rezaeinia, R. Rahmani, A. Ghodsi, and H. Veisi, "Sentiment analysis based on improved pre-trained word embeddings," *Expert Systems with Applications*, vol. 117, pp. 139–147, Mar. 2019, doi: 10.1016/j.eswa.2018.08.044.
- [25] M. Liu, Y. Wei, L. Wang, Z. Xiong, and H. Gu, "An accident diagnosis method of pressurized water reactor based on BI-LSTM neural network," *Progress in Nuclear Energy*, vol. 155, Jan. 2023, doi: 10.1016/j.pnucene.2022.104512.

BIOGRAPHIES OF AUTHORS

Rajeshwari Dembala     has awarded by B.E., M.Tech. and Ph.D. degree in computer science and engineering from Visvesvaraya Technological University, Belagavi, India. Currently she is working as assistant professor, in the Department of Information Science and Engineering, The National Institute of Engineering (North Campus), Mysore. She has Published 13 Journal papers in reputed Journals, 5+ Conference Publications and Book Chapters. Her research interests include data mining, artificial intelligence, internet of things, block chain technology and data analytics. She has around 19+ years of teaching experience. She can be contacted at email: rajeshwaridresearch@gmail.com.



Ananthapadmanabha Thammaiah     has received B.E. degree in EEE from The National Institute of Engineering (NIE), Mysuru, Karnataka affiliated to University of Mysore with 9th rank in 1980. ME in power systems with 1st rank and gold medal and Ph.D. with gold medal in power systems from University of Mysore, Mysuru in 1984 and 1997, respectively. He is currently working as Director of Mysore University School of Engineering, University of Mysore. He has 300+ publications and guided 18 candidates for the doctoral degree. His research interests include soft computing application in power system engineering, reactive power compensation, renewable integration, and distributed generation. He can be contacted at email: drapn2015@gmail.com.