The Journal of Supercomputing
https://doi.org/10.1007/s11227-021-04169-6

A deep learning approach in predicting products' sentiment ratings: a comparative analysis

Vimala Balakrishnan[1] · Zhongliang Shi[1] · Chuan Liang Law[2] · Regine Lim[1] · Lee Leng Teh[3] · Yue Fan[1]

Accepted: 21 October 2021

Abstract

We present a benchmark comparison of several deep learning models including Convolutional Neural Networks, Recurrent Neural Network and Bi-directional Long Short Term Memory, assessed based on various word embedding approaches, including the Bi-directional Encoder Representations from Transformers (BERT) and its variants, FastText and Word2Vec. Data augmentation was administered using the Easy Data Augmentation approach resulting in two datasets (original versus augmented). All the models were assessed in two setups, namely 5-class versus 3-class (i.e., compressed version). Findings show the best prediction models were Neural Network-based using Word2Vec, with CNN-RNN-Bi-LSTM producing the highest accuracy (96%) and *F*-score (91.1%). Individually, RNN was the best model with an accuracy of 87.5% and *F*-score of 83.5%, while RoBERTa had the best *F*-score of 73.1%. The study shows that deep learning is better for analyzing the sentiments within the text compared to supervised machine learning and provides a direction for future work and research.

Keywords Sentiment rating · Deep learning · Word embeddings · Customer reviews · Ensemble models

✉ Vimala Balakrishnan
vimala.balakrishnan@um.edu.my

[1] Faculty of Computer Science and Information Technology, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

[2] Malayan Banking Berhad, 50050 Kuala Lumpur, Malaysia

[3] Datium Insights, 59200 Kuala Lumpur, Malaysia

Published online: 05 November 2021

1 Introduction

Online shopping has grown tremendously, significantly more during the on-going COVID-19 pandemic, which resulted in many countries enforcing stay-at-home orders among their citizens. With the closure of most retail shops and fear of COVID-19 infections, online shopping has become the main means for customers to satisfy their consumption needs. It is common for online retailers to solicit customer reviews on products and services through textual reviews and/or ratings [1, 2]. These online reviews play a great role in influencing the purchasing decisions made by customers while providing more insights to the sellers. As online platforms including social media contain voluminous data, sentiment analysis provides an easy and fast mechanism to categorize the reviews, hence providing useful insights to both customers and sellers on the feedback of the products and services [3, 4].

Sentiment analysis generally elicits a sentiment orientation (i.e., positive, neutral, negative) of textual information, which can improve decision-making processes for multitude domains including businesses such as finance and stock market [5–7], digital payment services [4], retails [2, 8], and products [1, 3, 9], among others. Scholars investigating sentiment analysis based on textual communications have also examined or attempted to determine the sentiment ratings, often using scales ranging from 1 to 5 or 10 (i.e., higher scores indicate more positive reviews) [10]. Though often performed using machine learning approaches, deep learning has gained momentum in sentiment analysis in recent years showing promising results [6, 10]. Further, scholars have also explored various word embedding techniques including the popular Word2Vec and its variants to the more advanced and state-of-art transformer-based pre-trained models such as Bi-directional Encoder Representations from Transformers (BERT) [10–13] that have displayed much better results in text classifications. Nevertheless, as shown later in Sect. 2.2, studies exploring deep learning algorithms, particularly those exploring and comparing various embedding techniques are lacking, both for English and non-English datasets [10, 12, 13]. Moreover, recent reviews show studies exploring data augmentation techniques in supervised deep learning algorithms to improve prediction improvements [14]. The technique, which is generally a regularization technique that synthesizes new data from existing data has been widely used in computing vision [14, 15]; however, works relating to textual data is limited due to the difficulty of establishing standard rules for automatic transformations of textual data while conserving the quality of the annotations [14, 16, 17], except for a few. For example, authors in [17] explored various data pre-processing and regularization techniques to analyze the sentiments of Vietnamese users on Twitter with results indicating data augmentation to be a promising solution to boost the accuracies of classifiers.

To address the gaps identified above, this study aims to predict the customer review ratings using deep learning models based on an e-commerce dataset containing reviews for women's clothing. Specifically, this is achieved through data pre-processing and data augmentation to increase the variability of the dataset.

Several word embedding techniques were examined including Word2Vec, FastText, BERT model and its variants (i.e., RoBERTa and ALBERT) in order to identify the best embedding technique along with the deep learning algorithms. Several Neural Network (NN) classifiers were then used such as Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Bi-directional Long Short Term Memory (Bi-LSTM), on two different setups, that is, 5-class versus 3-class. The models were evaluated through performance metrics. Further, we also validated our models against several machine-learning algorithms including Naïve Bayes, Logistic Regression and Support Vector Machine (SVM), etc. The paper contributes to extensively analyzing various well-known deep learning models along with the more recent and advanced BERT variants in order to identify the best sentiment review prediction model using both the original and augmented datasets. The remainder of the paper consists of background, methodology, results, discussion and conclusion.

2 Background

2.1 Sentiment analysis and customer reviews

Sentiment refers to 'a feeling or an opinion, especially one based on emotions' [18] while sentiment analysis is the process of analyzing people's sentiment expressed toward services, products, mandates, organizations, etc. [19]. A sentiment rating on the other hand, refers to the use of numerical values (or stars) to indicate the intensity of one's sentiment [10]. As mentioned previously, sentiment analysis has been studied and applied in various fields primarily to gauge how people feel about something, and its popularity among research scholars can be attributed to the proliferation of social media.

Sentiment analysis is often performed using machine learning, lexicons or hybrid approach [20]. Machine learning remains to be the widely used approach in sentiment analysis as the algorithms demonstrate high accuracy of classifications, however, the classifiers are very domain-dependent. On the other hand, lexicon-based approach uses opinion lexicons to determine the semantic orientation of the words as negative or positive with the help of scores [20]. Although this approach does not require labeled data and learning procedures, powerful linguistic resources are usually required, which are not always available especially for non-English datasets. The hybrid approach is a combination of the machine learning and lexicon-based approaches.

Studies investigating sentiment analysis based on user reviews using machine learning approaches are many, the majority of which have used supervised algorithms such as Naïve Bayes, Decision Tree, SVM, etc. For instance, Haque and colleagues [21] applied a semi-supervised approach on Amazon review dataset of three different categories of products, using pool-based active learning for data labeling. Their experiments show Linear Support Vector Machine to produce the highest accuracy. Scholars have also attempted to improve sentiment analysis based on specific features, such as Pang et al. [22] who analyzed the performance of Naive

Bayes, Maximum Entropy and SVM on movie reviews with ratings (i.e., a number of stars) whereas authors in [23] examined the effect of word lengths for airline reviews. Despite the popularity of the machine learning approaches, researchers have noted the need for more advanced and robust sentiment analysis approaches to better understand customers and their needs [24].

2.2 Deep learning approaches to reviews

Deep learning refers to 'neural networks with multiple layers of perceptrons inspired by the human brain' [20] and has been shown to bring benefits toward text generation, word representation estimation, sentence classification and feature presentation [25]. The approach has been successfully used to analyze sentiments for reviews [10, 11, 26], stock price prediction [5, 6] and also non-English datasets [27–29], with popular algorithms including RNN, CNN, Bi-LSTM and integrated versions of the algorithms.

To further elaborate, [29] performed sentiment analysis using RNN based on Word2Vec embedding on reviews extracted from the Indonesian Traveloka website. Their proposed model reported an accuracy of 91.9%. Hameed and Garcia-Zapirain [30] used Bi-LSTM on three datasets, namely IMDB, Movie Review and Stanford Sentiment Treebank (SST2) with accuracy results of 85.8%, 80.5% and 90.6%, respectively. The authors found Bi-LSTM to be computationally efficient and well-suited for sentiment analysis tasks as well. A similar approach was adopted by Xu and colleagues [31] who used Word2Vec along with Bi-LSTM, LSTM, RNN and CNN to extract sentiments of Chinese hotel reviews, with Bi-LSTM emerging as the best model with an F-score of 92%. On the other hand, [32] compared CNN, RNN and deep NN (DNN) using Word2Vec and Term Frequency-Inverse Document Frequency (TF-IDF) on 13 different datasets, with results showing the models to have the best performance when Word2Vec was used across all the metrics. Also, RNN using Word2Vec emerged as the best model although computationally expensive compared to the others.

Others have merged several deep learning models in improving sentiment analysis, for example, [33] proposed an LSTM-CNN grid-search (GS) model to predict sentiment analysis on two datasets, namely Amazon and IMDB movie reviews. The authors specifically implemented a grid-search approach in their proposed work and compared their model against several baseline algorithms such as CNN, LSTM, CNN– LSTM, etc., with results indicating their model to have outperformed the baselines with an overall accuracy of 96%. A similar work was accomplished by [26] using Amazon reviews in which topic modeling was first administered with Fuzzy C-means prior to classifying sentiments using CNN. The authors reported their proposed model to have an enhanced accuracy between 6 and 20% compared to the traditional systems.

Literature also revealed studies exploring the more advanced embedding technique, BERT and its variants in improving sentiment analysis for reviews. For instance, [34] improved sentiment analysis for commodity reviews using BERT-CNN with F-score results indicating the combination of BERT-CNN (84.3%) to

be the best compared to BERT (82%) and CNN (70.9%). Similarly, [12] developed SenBERT-CNN to analyze JD.com (mobile phone merchant) reviews by combining BERT and CNN, the latter of which was used to extract deep features of the text. The authors found BERT-CNN to have the highest accuracy (95.7%) compared to LSTM, BERT and CNN.

On the other hand, [10] used Neural Network (NN) models to predict drug reviews using a dataset from Drugs.com. The reviews had a score ranging from 0 to 9 indicating satisfaction level of patients. The authors proposed several NN models including BERT-LSTM on two setups (i.e., 10-class and 3-class, which is the compact version of the dataset), with results showing BERT-LSTM to be the best for the 3-class setup with an average F-score of 82.37%, albeit with a very high training time. Others include the work of [11] who examined several NN models along with BERT for a movie review dataset with results indicating BERT to produce the best accuracy while [13] used BERT for Twitter sentiment analysis, which transformed jargons into plaintext for BERT training. A summary of the studies using deep learning algorithms to predict sentiment analysis based on user reviews is given in Table 1.

3 Methodology

This section provides the methodology adopted in this study, outlining the datasets used, data pre-processing steps, feature extractions, sentiment review and rating classifications, experimental setups and evaluations. Figure 1 illustrates the overall methodology.

3.1 Dataset

The dataset for this study comprised customer reviews on women's clothing, consisting of 23, 486 observations, including clothing ID, age of the reviewers, title of the reviews, review text, rating, recommended indicators, positive feedback counts, division name, department name and class name. The review text is used to predict the rating given to the products (i.e., 1: extremely negative – 5: extremely positive). The dataset is available at Kaggle [35]. A preliminary check revealed approximately 845 missing reviews, hence these were removed resulting in a final sample size of 22, 641. Figure 2 illustrates the word cloud for the two extreme ratings in the dataset.

3.2 Data augmentation

Data augmentation is commonly used to enrich the training dataset such that the trained models are robust and produce improved performance for deep learning models, and the technique has been widely used in computer and speech processing [14, 15], with interests in textual data augmentation increasing over the last few years [14, 36]. As textual communications are inherently more complex (i.e., syntax and semantic constraints), several data augmentation techniques have been proposed

Table.1 Summary of studies using deep learning approaches in review sentiment analysis

References	Datasets	Technique/Algorithms	Result
[29]	Traveloka—Indonesian	RNN – Word2Vec	Accuracy: 91.9%
[31]	Hotel reviews—Chinese	Bi-LSTM; LSTM; RNN; CNN – Word2Vec	*F*-score: 92% for Bi-LSTM
[10]	Drug reviews	CNN; LSTM BERT-LSTM	*F*1-score: 82.37% for BERT-LSTM
[11]	Movie Reviews – Rotten Tomatoes English	RNN; RNTN; CNN Bi-LSTM; BERT	Accuracy: Bert Base 94.0% – BERT large 94.7%
[33]	Amazon and IMDB—English	LSTM-CNN-GS	Accuracy – 97.8% *F*-score – 97.2%
[26]	Eight different Amazon products: amazon instant video, books, electronics, home and kitchen, movie review, media, kindle, and camera—English	Selective Memory based CNN	Average accuracy – 92.85
[34]	JD.com—Chinese	BERT-CNN	*F*-score – 84.3%
[12]	JD.com—Chinese	BERT-CNN	Accuracy—95.7%
[13]	Twitter – Italian	BERT; ALBERTo	Average *F*-score: 75% for BERT

BERT, Bi-directional Encoder Representations from Transformers; RNN, Recurrent Neural Network; CNN, Convolutional Neural Network; LSTM, Long Term Short Memory

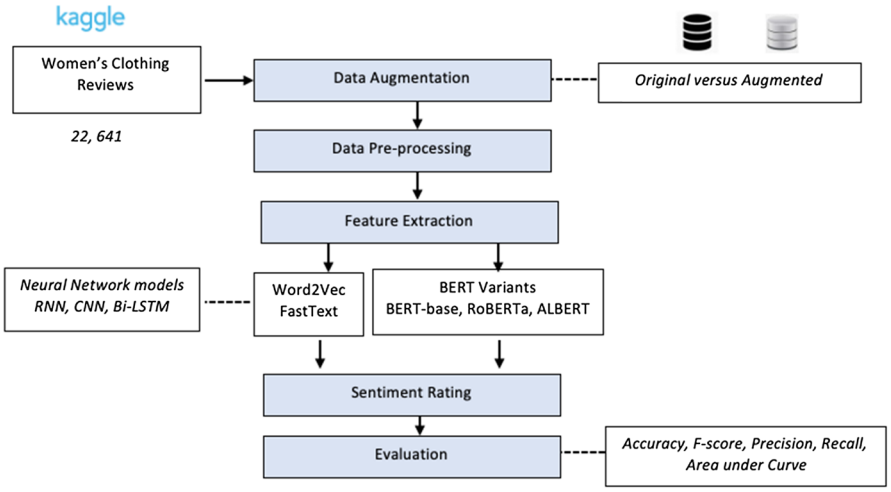

Fig. 1 Overall proposed methodology

Fig. 2 Word cloud overview for reviews with scores 1 and 5

by scholars including translations [17, 37, 38], question answering [39], synonym replacement [16], etc. The present study adopted one of the recent methods introduced in [39], that is, Easy Data Augmentation (EDA) comprising four NLP operations, namely random deletion, random insertion, random swap, and synonym replacement (see Table 2 for explanation and examples). EDA is known for its simplicity and ease of use as it does not require any predefined datasets, and often yield promising results [17, 36, 40]. For instance, Xiang and peers [40] compared various data augmentation techniques on several datasets and found EDA to perform better than DICT (i.e., a synonym replacement thesaurus [41]) but was outperformed by their proposed POS-based augmentation technique. The present study adopted the default setting recommended for EDA by [36], that is, up to four augmented sentences were generated for each original sentence using a learning rate of 0.1. We administered all four operations listed in Table 2 on each sentence, hence generating

Table 2 Types of data augmentation used in the present study [36]

Operations	Description	Example
Original text	–	The quick brown fox jumps over the lazy dog
Random swap	Two words are randomly selected and swapped	The ***lazy*** quick fox jumps over the dog ***brown***
Random deletion	Randomly remove a word from the sentence	The quick brown jumps over the lazy dog
Random insertion	Randomly introduce and insert a new word	The quick ***sluggish*** brown fox jumps over the lazy dog
Synonym replacement	Selecting *n* number of non-stop word(s) and replacing them with its synonym randomly	The quick sluggish ***umber*** fox jumps over the lazy dog

Bold words refer to the changes made as per the operation listed

four different variations for each. As the augmentations were done according to the pre-evaluated and recommended parameters, the resulting augmented dataset closely represents the original sentences, hence maintaining the meaning of the original data and conserving the true labels [36]. The EDA technique resulted in a single augmented dataset, and was used to train and evaluate the sentiment rating prediction models, along with the original dataset.

3.3 Data pre-processing

Common natural language processing tasks were then incorporated, that is, canonicalization, which involves conversion of text into lowercases, removal of leading and trailing spaces, numbers, punctuations and stop words (i.e., common words in English that carries little information about the context of the texts such as 'a,' 'an,' 'the,' etc.). These were then followed with tokenization (i.e., splitting sentences into singular words) and lemmatization, which reduces the words into its root forms (e.g., 'silky' to 'silk,' 'happened' to 'happen'). Additionally, index encoding and zero padding were performed to ensure all the matrixes were of the same size, accomplished using the Keras library. Table 3 illustrates a hypothetical case for the pre-processing steps.

3.4 Feature extraction

Features are individual measurable properties or dimensions for algorithms to process whereas feature extraction is the process of translating the processed texts into informative format. In general, the feature extractions techniques are dependent on the prediction models used in a sentiment analysis. In this study, word embeddings (i.e., vector representations of a particular word) were extracted as features, through several techniques, namely:

Table.3 Pre-processing steps

Pre-processing steps	Examples
Convert the text to lowercase	I love these dresses SO MUCH!!!!
	I love these dresses so much!!!!
Remove leading and trailing spaces	I love these dresses so much!!!!
	I love these dresses so much!!!!
Remove of punctuations, numbers, special characters	I love these dresses so much!!!!
	I love these dresses so much
Remove of stop words	I love these dresses so much
	I love dresses so much
Lemmatization	I love dresses so much
	I love dress so much

1. Word2Vec: a pre-trained model that learns the relationship between the words in a corpus, and returns an embedded vector for each word in the text [42],
2. FastText: an extension of Word2Vec that breaks words into n-grams (smaller parts), e.g., 'apple' to 'app' with the intention of learning the morphology of the words. The model also returns a bag of embedded vectors for each word in the text [43].

Word2Vec and FastText might not handle polysemous words (i.e., words with multiple meaning) as they are deemed to be context-free (i.e., map the same word to the same embedding vector). For example, 'fire' would have the same representation in 'building on fire' and 'fire someone.' To mitigate this problem, scholars have begun to explore transformer-based embeddings, including BERT and its variants. BERT-variant models were pre-trained by incorporating the context of the word within the text in Wikipedia and BooksCorpus [44], and the embedding are then used through a classifier for predictions. As they produce contextualized word embeddings, they produce state-of-the-art results on Natural Language Processing tasks [12, 34]. The BERT-base model is a bi-directional (both left-to-right and right-to-left direction) transformer for pre-training over a lot of unlabeled textual data to learn a language representation that can be used to fine-tune for specific classification tasks (see [44] for further details). One of its popular variant is RoBERTa (Robustly Optimized BERT approach), which was introduced by Facebook. It is basically an improved version of BERT, capable of handling more data with higher computing power. Compared to BERT, RoBERTa has been shown to have a higher prediction power. Finally, Google and Toyota developed a smaller/smarter BERT variant known as A Lite BERT (ALBERT), which is dramatically smaller in size compared to BERT. The present study examined BERT-base model and two of its variants, that is, RoBERTa and ALBERT.

3.5 Sentiment review predictions using deep learning models

Three well-known NN algorithms were identified from the literature, namely CNN, RNN and Bi-LSTM. NN models are basically made up of artificial neurons organized in layers, known as input (i.e., predictors), output (i.e., predictions) and hidden layers. In a feed-forward multilayer NN model (see Fig. 3), each layer receives inputs from the previous layers, and the inputs are combined using adaptive weights that are calibrated through a training process [45]. There is an activation function for each neuron, with popular ones including tangent sigmoid, logarithmic sigmoid and Softmax [45].

RNN belongs to a class of NN that are good at modeling sequence data and processing for predictions. It is a word-based vector and deals with long-term dependencies among words in a text corpus. RNN processes sequential data using its internal memory and allows the network to retain the information that has been processed before the current stage [46]. In the current study, we used an LSTM layer with 256 units, a dropout rate of 0.3 and learning rate of 0.001. Softmax, which converts a vector of values to a probability distribution, was used as the activation function.

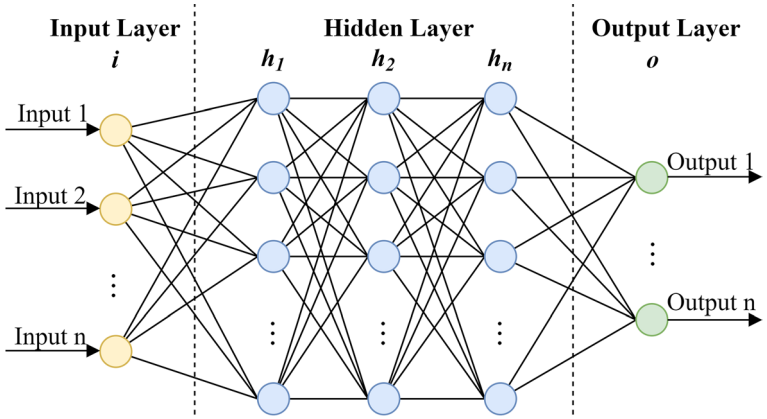

Fig. 3 General Neural Network Architecture [45]

CNN, on the other hand, is designed to adaptively learn spatial hierarchies of features, typically composed of three layers, that is, convolution, pooling, and fully connected layers. The first two layers perform feature extraction, whereas the fully connected layer maps the extracted features into a final output [46, 47]. The extracted features can hierarchically and progressively become more complex, hence parameters are often optimized through algorithms [47]. We used a convolution layer with 256 filters with a window size of 3, 4 and 5-word vectors, along with a linear rectification unit (ReLU) as the activation function. Further, a kernel regularizer that applies an L1 regularization penalty with a value of 0.01 was also applied, along with a dropout rate of 0.3.

Finally, the Bi-LSTM is an improvised version of RNN that processes the input text storing the semantics of the previous and future context information. It is composed of LSTM units that operate in both directions, consisting of recurrently connected memory blocks with each memory cell containing three gates, namely the input gate (controls if the information is allowed in), forget gate (controls the length of time information remains in the memory) and output gate (controls the output of the memory cells [48]. We used two types of dense layers, namely a layer with 64 units using ReLU as the activation function, and another with 3 and 5 classes using Softmax as the activation function. A similar dropout rate of 0.3 was used for the Bi-LSTM model as well.

3.6 Machine learning models

It is to note that we carried out additional analyses using conventional machine learning algorithms to compare their performance with the deep learning models. Specifically, five well-known machine learning algorithms used in sentiment analysis studies [7, 21–24] were selected, namely SVM, Naive Bayes, Random Forest, Logistic Regression and Decision Tree. Naïve Bayes is one of the simplest and widely used probabilistic algorithms for classification problems, requiring only a

small amount of training data. In other words, it returns a probability based on the class that has the 'maximum posterior' [3, 49]. SVM on the other hand, attempts to find the best hyperplane for classification purpose, and is known to work well with high-dimensional datasets. However, it requires a substantial amount of time to determine the optimal kernel functions [50].

The Decision Tree is a powerful classification algorithm that describes the relationship of attributes and targets in the form of a tree using a 'if–then' rule-based structure [51]. It has the ability to deal with large datasets compared to other machine learning algorithms; however, it also suffers from an instability issue where a small change in the training samples tend to cause a large difference in the classification results [52]. An improvement to Decision Tree is Random Forest, which is one of the best known algorithms for classifications, often yielding good accuracy results without any overfitting issues. Random Forest produces a number of individual trees and makes a final prediction by aggregating the decisions of the individual trees [53]. Finally, the boosting approach merges weak classifiers to improve classification performance, and studies have shown that the approach is superior to other machine learning algorithms such as SVM and Decision Tree [54].

3.7 Experiment

The experiments were conducted in several setups and scenarios, as follows:

There were two experiment setups based on the labels:

1. 5-class: refers to the original rating scale from 1 to 5 (i.e., 1 – extremely negative; 2 – negative; 3 – neutral; 4 – positive; 5 – extremely positive) [35],
2. 3-class: Ratings 1 and 2 were combined to reflect negative sentiments, 3 as neutral, 4 and 5 combined as positive sentiment [10].

The scenarios of the experiments are as follows:

1. RNN, CNN, Bi-LSTM using Word2Vec and FastText using both the original and augmented datasets, tested in the 5- and 3-class setups;
2. BERT variants (i.e., BERT, RoBERTa and ALBERT) using both the original and augmented datasets, tested in the 5- and 3-class setups. This excludes the Word2Vec and FastText embedding.

Upon the identification of the best setups from the experiments above (i.e., original versus augmented, word embedding techniques (Word2Vec, FastText and BERT variants) and class setups (i.e., 5-class versus 3-class) (Sects. 4.1 and 4.2), we carried out further modeling using ensemble models, namely CNN-RNN, CNN – Bi-LSTM, RNN – Bi-LSTM and CNN-RNN-Bi-LSTM, using the majority voting technique to choose the best prediction (Sect. 4.3). Further, to validate the findings, the best class and word embedding setup was also used against several machine learning algorithms, that is, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, and SVM (Sect. 4.4).

All the NLP tasks and model developments were accomplished using Python and Keras. We used AdamW as the compiler and CrossEntropyLoss as the loss functions. For validation purpose, we adopted the k-fold cross-validation in which the data will be partitioned into k disjoint folds, with one of the folds used for testing while the remaining k-1 folds used for training. We used $k = 10$, hence 10 different models were trained and tested over 10 iterations before the final value is averaged. In classification approaches, it is common to use $k = 5$ or 10 [1, 4, 23].

3.8 Evaluation

The standard performance metrics for classification problems were used to assess all the models, namely:

1. Accuracy—the proportion of the total number of correct predictions over the total number of cases examined, as given in Eq. (1):

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

   where TP – true positive; TN – true negative; FP – false positive; FN – false negative [55].
2. Precision—the ratio of true positive results over the total number of positive predictions (including true positive and false positive) by the model (Eq. 2).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

   where TP – true positive; FP – false positive [55].
3. Recall—the proportion of actual positive cases which are correctly identified (Eq. 3).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

   where TP – true positive; FN – false negative [55].
4. F-measure—the harmonic mean between precision and recall, and the range of F-measure is between 0 and 1. Greater value of F-measure indicates better performance of the model. The formula for determining F-measure is:

$$F - \text{measure} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{4}$$

5. Area under the ROC (Receiver Operating Characteristic) curve (AUC-ROC)—Similar to $F1$-score, AUC has also the range of 0 and 1. The higher the score for AUC, the better the performance. ROC curve is a graph that shows the plot between sensitivity (true positive rate) and (1-specificity) (false positive rate).

4 Results and discussion

4.1 NN-based models using Word2Vec and FastText

Table 4 presents the results of the experiments involving all the NN models using the original dataset in both 5- and 3-class setups, along with Word2Vec and FastText techniques. This was accomplished to assess the performance of the various NN models based on the two word embedding techniques and class setups. Conversely, the results for the augmented dataset for the same embedding and class setups are provided in Table 5.

It can be generally observed that Bi-LSTM based on Word2Vec consistently outperformed other NN models, regardless of the setups, followed very closely by RNN. The results for the augmented dataset produced a more consistent pattern where RNN emerged to be the best model using Word2Vec, for both the setups (see Table 5). This is in accordance with other studies that found RNN using Word2Vec to be the best model in sentiment classification [29, 32], however in contrast to [31] who found CNN-Word2Vec to be the best model. Studies using RNN have generally found the use of word embedding techniques to produce better prediction models compared to other techniques such as TF-IDF [29].

Further, it can also be observed that the performance of the models were better for the augmented dataset compared to the original, across all the metrics. This is probably because data augmentation, which is one of the most useful interfaces to train NN models, is able to prevent overfitting by shuffling particular forms of language. Therefore, it mitigates NN models from learning spurious correlations and memorizing high-frequency patterns that do not generalize [36]. Similar observations have

Table.4 Performance of NN models in percentage (%) for the original dataset: Word2Vec versus FastText

Feature extraction	Model	Precision	Recall	*F*-score	AUC	Accuracy
5-class						
Word2Vec	CNN	37.96	32.96	31.56	59.07	61.65
	RNN	43.80	41.47	41.92	64.76	62.22
	Bi-LSTM	45.95	42.51	**42.59**	**65.44**	**63.10**
FastText	CNN	37.45	32.94	31.11	58.78	61.14
	RNN	43.36	40.34	41.05	64.16	62.08
	Bi-LSTM	44.93	40.90	41.33	64.48	62.83
3-class						
Word2Vec	CNN	58.56	50.99	52.30	65.16	80.79
	RNN	60.89	57.49	**58.65**	70.65	81.42
	Bi-LSTM	60.51	57.99	58.31	**70.99**	**81.54**
FastText	CNN	57.14	49.84	50.36	64.27	80.52
	RNN	60.75	57.10	58.27	70.29	81.47
	Bi-LSTM	60.18	57.68	58.02	70.92	81.22

Best scores in bold

Table.5 Performance of NN models in percentage (%) for the augmented dataset: Word2Vec versus FastText

Feature extraction	Model	Precision	Recall	*F*-score	AUC	Accuracy
5-class						
Word2Vec	CNN	72.52	68.05	69.77	80.89	80.00
	RNN	83.55	83.60	**83.52**	**89.89**	**87.45**
	Bi-LSTM	75.32	76.24	75.62	85.70	83.20
FastText	CNN	66.52	59.89	62.03	76.00	75.25
	RNN	74.38	73.99	74.11	84.14	81.25
	Bi-LSTM	66.76	67.38	66.76	80.52	77.95
3-class						
Word2Vec	CNN	83.63	78.51	80.72	85.15	90.98
	RNN	88.85	90.76	**89.77**	**93.73**	**94.84**
	Bi-LSTM	84.06	87.42	85.57	91.63	92.87
FastText	CNN	76.05	68.34	71.41	78.18	87.60
	RNN	83.72	85.27	84.45	90.09	92.36
	Bi-LSTM	81.89	84.11	82.94	89.51	91.74

Best scores in bold

been reported in other studies that have compared the use of EDA in textual communications both in English [40] and non-English languages [17].

4.2 BERT variants for sentiment review prediction

Tables 6 and 7 show the performance results for the BERT variants, where a consistent pattern was noted for RoBERTa for both the datasets and setups. It can also be observed that prediction performance is better in the 3-class setup as opposed to 5-class setup. In fact, the same pattern was found in the NN models (Tables 3 and 4), probably due to a more refined classification when the number of classes/categories are smaller. A similar result was reflected in [10] where the authors reported an improved *F*-score in their 3-class setup as opposed to the 10-class.

Table.6 Performance of BERT models in percentage (%) for the original dataset: 5- versus 3-class setups

Class	Model	Precision	Recall	*F*-score	AUC	Accuracy
5-class	BERT	57.03	51.18	52.28	77.30	69.34
	ALBERT	52.45	51.36	51.44	75.96	67.88
	RoBERTa	55.37	54.79	54.69	77.74	69.99
3-class	BERT	68.93	68.39	68.54	84.92	85.48
	ALBERT	66.73	67.84	67.19	84.95	84.55
	RoBERTa	70.08	71.49	**70.64**	**86.79**	**86.29**

Best scores in bold

Table.7 Performance of BERT models in percentage (%) for the augmented dataset: 5- versus 3-class setups

Class	Model	Precision	Recall	*F*-score	AUC	Accuracy
5-class	BERT	57.48	53.32	53.03	80.45	73.55
	ALBERT	54.76	51.54	52.49	76.34	69.03
	RoBERTa	59.26	57.02	57.79	79.13	72.44
3-class	BERT	71.34	70.89	71.01	86.31	86.65
	ALBERT	69.27	68.51	68.83	84.53	85.57
	RoBERTa	73.25	73.04	**73.09**	**87.08**	**87.68**

Best scores in bold

The BERT variants were found to perform better in the augmented dataset as well, akin to the NN models however, with lower metric scores. Of all three variants, RoBERTa produced the best results, though marginally close to BERT. This is in line with [56] who found RoBERTa to outperform the BERT model, achieving a 2 to 20% increase in model performance on the majority of NLP tasks. However, this result is also in contrast with [57] who found BERT to perform better than RoBERTa for sentiment analysis task, with the author attributing this to the quality of data and features extracted for their sentiment analysis task.

In conclusion, the results in Sects. 4.1 and 4.2 revealed RNN-Word2Vec to be the best model using the 3-class setup and augmented dataset. Therefore, the rest of the experiments was executed using Word2Vec and 3-class setup on the augmented dataset.

4.3 Ensemble NN models for sentiment rating prediction

Table 8 depicts the results for the ensemble models based on the best setup (i.e., 3-class) using the augmented dataset and Word2Vec. This was done to assess the performance of the merged NN models in predicting the sentiment reviews as opposed to individual models in Sects. 4.1 and 4.2.

Our results indicate all the ensemble models to perform better than the NN models individually (see Table 5), with the CNN-RNN-Bi-LSTM to have the best accuracy (i.e., 96%) and *F*-score of 91.1%. This pattern of observation have been reported in other studies as well, whereby multi-models were generally found to

Table.8 Performance for the ensemble models in percentage (%) for the augmented dataset

Model	Precision	Recall	*F*-score	AUC	Accuracy
CNN-RNN	90.9	87.5	89.2	98.8	94.8
CNN – Bi-LSTM	88.8	86.2	87.3	98.5	93.8
RNN – Bi-LSTM	90.8	91.3	91.1	99.1	95.6
CNN – RNN—Bi-LSTM	**91.9**	**90.4**	**91.1**	**99.1**	**96.0**

Best scores in bold

perform better than individual models [26, 33], regardless of the datasets used. Though the metric differences between the ensemble models are not significantly large, our results provide evidence that the use of ensemble models (which aims to improve predictions) helps to improve the overall review prediction results compared to the traditional approach of using deep learning models.

4.4 Machine learning models

Finally, to validate our findings against the machine learning approach, the same setup as in Sect. 4.3 was used with several machine learning models, as shown in Table 9. All the models were found to have performed poorly as opposed to the deep learning models, with at least 20% of differences in terms of the accuracy results. Based on these results, the study concludes that the more robust deep learning models are better suited to perform sentiment rating predictions compared to the conventional machine learning approach.

5 Conclusion, limitations and future directions

This study contributed to the research domain of online customer reviews using several deep learning algorithms based on various embedding techniques. Our findings show that all the prediction models work better in a setup with fewer and more refined classes (3-class versus 5-class), and using augmented dataset improves the prediction compared to the original dataset. As for the context-free embeddings, Word2Vec was found to produce better results than FastText, though the differences were minimal. Similarly, RoBERTa produced the best results compared to BERT and ALBERT. Finally, our results also show the ensemble models to produce the best results compared to the individual models, and also against the machine learning models.

We identify several limitations. The dataset used in this study was not checked for spams or fake reviews, hence this may have affected the predictions to a certain extent. Thus, an additional step in automatically detecting fake reviews and spams could be included in the pre-processing stage [26]. The scope of the study is also limited to English reviews, thus the proposed models and findings may not be applicable in a multi-lingual setting. This is considered important as online customers are

Table.9 Performance for the machine learning models in percentage (%)

Model	Precision	Recall	*F*-score	AUC	Accuracy
Logistic regression	43.93	35.97	37.68	64.14	62.26
Naïve Bayes	43.76	38.40	39.90	66.15	62.12
Decision tree	43.88	30.27	30.84	66.30	60.20
Random Forest	46.15	26.27	24.80	55.02	59.17
Support vector machine	37.71	36.21	36.82	64.23	56.11

known to originate from all around the world, and there is a tendency to communicate in languages other than English, such as Chinese, Spanish, etc. In future studies, other languages could be further explored by enhancing the current proposed framework in order to handle languages other than the English language.

We experimented with well-known NN models, using various embedding techniques including the more advanced BERT and its variants. However, other approaches could be explored such as the use of lexicons, which can be merged with NN and BERT-variant models, such as lexicon enhanced BERT and lexicon-RNN. Moreover, the present study did not consider the proportion of polysemous words for BERT, in line with numerous other studies that have shown BERT-derived representations could reflect words' polysemy level and their partitionability into senses [58–60]. Nevertheless, it would be interesting to further investigate this notion by considering the proportion of polysemous words for BERT variants.

Further, our results indicate machine learning algorithms performed considerably poorly compared to the NN models in the same setup. Although deep learning models are generally known to perform better than machine learning models, they are however, computationally expensive. Therefore, future studies could explore optimization techniques or use other ensemble boosting approaches to improve the prediction performance of the machine learning models. In addition, predicting review ratings based on real-time data and applications would be an interesting and important direction as well considering the popularity of online shopping that is gaining momentum during the COVID-19 pandemic which has dramatically changed the shopping landscape globally.

References

1. Zhang J, Zhang A, Liu D, Bian Y (2021) Customer preferences extraction for air purifiers based on fine-grained sentiment analysis of online reviews. Knowl-Based Syst. https://doi.org/10.1016/j.knosys.2021.107259
2. Wu JJ, Chang ST (2020) Exploring customer sentiment regarding online retail services: a topic-based approach. J Retail Consum Serv 55:102145
3. Xu F, Pan Z, Xia R (2020) E-commerce product review sentiment classification based on a Naïve Bayes continuous learning framework. Inf Process Manage. https://doi.org/10.1016/j.ipm.2020.102221
4. Balakrishnan V, Lok PY, Rahim HA (2021) A semi-supervised approach in detecting sentiment and emotion based on digital payment reviews. J Supercomput 77:3795–3810
5. Carosia AE, Coelho GP, Silva AE (2021) Investment strategies applied to the Brazilian stock market: a methodology based on sentiment analysis with deep learning. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2021.115470
6. Jing N, Wu Z, Wang H (2021) A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2021.115019
7. Yadav A, Jha CK, Sharan A, Vaish V (2020) Sentiment analysis of financial news using unsupervised approach. Proced Comput Sci 167:589–598. https://doi.org/10.1016/j.procs.2020.03.325
8. Zhan Y, Han R, Tse M, Ali MH, Hu J (2021) A social media analytic framework for improving operations and service management: a study of the retail pharmacy industry. Technol Forecast Soc Change 163:120504
9. Taparia A, Bagla T (2020) Sentiment analysis: predicting product reviews' ratings using online customer reviews. Available at Soc Sci Res Netw https://doi.org/10.2139/ssrn.3655308

10. Colón-Ruiz C, Segura-Bedmar I (2020) Comparing deep learning architectures for sentiment analysis on drug reviews. J Biomed Inform. https://doi.org/10.1016/j.jbi.2020.103539
11. Munikar M, Shakya S, Shrestha A (2019) Fine-grained sentiment classification using BERT, 2019 Artificial intelligence for transforming business and society (AITB). Kathmandu, Nepal: IEEE.
12. Wu F, Shi Z, Dong Z, Pang C, Zhang B (2020) Sentiment analysis of online product reviews based on SenBERT-CNN, 2020 International Conference on Machine Learning and Cybernetics (ICMLC), pp 229–234, https://doi.org/10.1109/ICMLC51923.2020.9469551.
13. Pota M, Ventura M, Catelli R, Esposito M (2021) An effective BERT-based pipeline for twitter sentiment analysis: a case study in ITALIAN. Sensors 21:133. https://doi.org/10.3390/s21010133
14. Shorten C, Khoshgoftaar TM, Furht B (2021) Text data augmentation for deep learning. J Big Data 8:101. https://doi.org/10.1186/s40537-021-00492-0
15. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
16. Kobayashi S (2018) (2018) Contextual augmentation: data augmentation bywords with paradigmatic relations. In NAACL HLT 2:452–457. https://doi.org/10.18653/v1/n18-2072
17. Duong HT, Nguyen-Thi TA (2021) A review: preprocessing techniques and data augmentation for sentiment analysis. Comput Soc Netw 8:1. https://doi.org/10.1186/s40649-020-00080-x
18. Alaoui ME, Bouri E, Azoury N (2020) The determinants of the U.S. consumer sentiment: linear and nonlinear models. Int J Financ Stud 8:38. https://doi.org/10.3390/ijfs8030038
19. Pathak AR, Pandey M, Rautaray S (2021) Topic-level sentiment analysis of social media data using deep learning. Appl Soft Comput. https://doi.org/10.1016/j.asoc.2021.107440
20. Birjali M, Kasri M, Beni-Hssane A (2021) A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl-Based Syst 226:107134. https://doi.org/10.1016/j.knosys.2021.107134
21. Haque TU, Saber NN, Shah FM (2018) Sentiment analysis on large scale Amazon product reviews, 2018 IEEE International Conference on Innovative Research and Development (ICIRD). IEEE. https://doi.org/10.1109/ICIRD.2018.837629
22. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: Sentiment classification using machine learning techniques, In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10. Association for Computational Linguistics, pp 79–86
23. Kaur W, Balakrishnan V (2018) Improving sentiment scoring mechanism: a case study on airline services. Ind Manag Data Syst 118(8):1578–1596
24. Lee H, Lee N, Seo H, Song M (2019) Developing a supervised learning-based social media business sentiment index. J Supercomput. https://doi.org/10.1007/s11227-018-02737-x
25. Ain QT, Ali M, Riaz A, Noureen A, Kamran M, Hayat B, Rehman A (2017) Sentiment analysis using deep learning techniques: a review. Int J Adv Comput Sci Appl 8(6):424–433. https://doi.org/10.14569/IJACSA.2017.080657
26. Mandhula T, Pabboju S, Gugalotu N (2019) Predicting the customer's opinion on amazon products using selective memory architecture-based convolutional neural network. J Supercomput. https://doi.org/10.1007/s11227-019-03081-4
27. Al-Dabet S, Tedmori S, Al-Smadi M (2021) Enhancing Arabic aspect-based sentiment analysis using deep learning models. Comput Speech Lang. https://doi.org/10.1016/j.csl.2021.101224
28. Pasupa K, Ayutthaya TS (2019) Thai sentiment analysis with deep learning techniques: a comparative study based on word embedding POS-tag, sentic features. Sustain Cities Soc. https://doi.org/10.1016/j.scs.2019.101615
29. Kurniasari L, Setyanto A (2020) Sentiment analysis using recurrent neural network. J Phys: Conf Ser. https://doi.org/10.1088/1742-6596/1471/1/012018
30. Hameed Z, Garcia-Zapirain B (2021) Sentiment classification using a single-layered BiLSTM model. IEEE Access 8:73992–74001
31. Xu G, Meng Y, Qiu X, Yu Z, Wu X (2019) Sentiment analysis of comment texts based on BiLSTM. IEEE Access 7:51522–51532
32. Dang NC, Moreno-Garcia MN, de la Prieta F (2021) Sentiment analysis based on deep learning: a comparative study. Electronics 9:483. https://doi.org/10.3390/electronics9030483
33. Priyadarshini I, Cotton C (2021) A novel LSTM–CNN–grid search-based deep neural network for sentiment analysis. J Supercomput. https://doi.org/10.1007/s11227-021-03838-w
34. Dong J, He F, Guo Y, Zhang H (2020) A commodity review sentiment analysis based on BERT-CNN model, 5th International Conference On Computer And Communication Systems (ICCCS), pp 143–147, https://doi.org/10.1109/ICCCS49078.2020.9118434.

35. Kaggle (2021) Women's clothing reviews, https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews.
36. Wei, J, Zou, K (2019) EDA: easy data augmentation techniques for boosting performance on text classification tasks, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). https://doi.org/10.18653/v1/d19-1670
37. Fadaee M, Bisazza A, Monz C. (2017) Data augmentation for low-resource neural machine translation, In Proceedings of the 55th annual meeting of the association for computational linguistics, vol 2: Short Papers. Vancouver. 2017; pp. 567–573, https://doi.org/10.18653/v1/P17-2090.
38. Sennrich R, Haddow B, Birch A (2016) Improving neural machine translation models with monolingual data, In Proceedings of the 54th annual meeting of the association for computational linguistics, vol 1: Long Papers, Berlin. 2016; pp. 86–96, https://doi.org/10.18653/v1/P16-1009.
39. Fader A, Zettlemoyer L, Etzioni O (2013) Paraphrase-driven learning for open question answering. In Proceedings of ACL (pp 1608–1618).
40. Xiang R, Chersoni E, Lu Q, Huang CR, Li W, Long Y (2021) Lexical data augmentation for sentiment analysis. J Assoc Inf Sci Technol 72:1432–1447
41. Zhang X, Zhao J, LeCun Y (2015) Character-level convolutional networks for text classification. In Advances in neural information processing systems (pp 649–657). Montreal, Quebec, Canada: Curran Associates, Inc. (Jun 2016).
42. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space, Available at [cs.CL]
43. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching Word Vectors with Subword Information. Trans Assoc Comput Linguist 5:135–146
44. Devlin, J, Chang, MW, Lee, K, Toutanova, K (2019) BERT: pre-training of deep bidirectional transformers for language understanding, Proceedings of NAACL-HLT 2019, pp 4171–4186 Minneapolis, Minnesota, Association for Computational Linguistics
45. Bre F, Gimenez JM, Fachinotti VD (2018) Prediction of wind pressure coefficients on building surfaces using artificial neural networks. Energy Build 158:1429–1441
46. Zhang L, Wang S, Liu B (2018) Deep learning for sentiment analysis: a survey. WIREs Data Min Knowl Discov. https://doi.org/10.1002/widm.1253
47. Yamashita R, Nishio M, Do RKG, Togashi K (2018) Convolutional neural networks: an overview and application in radiology. Insights Imaging 9:611–629. https://doi.org/10.1007/s13244-018-0639-9
48. Jang B, Kim M, Harerimana G, Gaspard K, Kang SU, Kim JW (2020) Bi-LSTM model to increase accuracy in text classification: combining Word2Vec CNN and attention mechanism. Appl Sci 10(5841):1–14
49. Feldman R, Sanger J (2007) The text mining handbook: advanced approaches in analyzing unstructured data. United States of America
50. Cheng MY, Peng HS, Wu YW, Chen TL (2010) Estimate at completion for construction projects using evolutionary support vector machine inference model. Autom Constr 19(5):619–629
51. Roe BP, Yang HJ, Zhu J, Liu Y, Stancu I, McGregor G (2005) Boosted decision trees as an alternative to artificial neural networks for particle identification. Nucl Instrum Methods Phys Res, Sect A 543(2–3):577–584
52. Nie CY, Wang J, He F, Sato R (2015) Application of J48 decision tree classifier in emotion recognition based on chaos characteristics. In: Proceedings of the 2015 International Conference on Automation, Mechanical Control and Computational Engineering. https://doi.org/10.2991/amcce-15.2015.330
53. Zhou J, Li X (2015) Mitri, HS (2015) Comparative performance of six supervised learning methods for the development of models of hard rock pillar stability prediction. Nat Hazards 79(1):291–316
54. Shin Y, Kim T, Cho H, Kang KI (2012) A formwork method selection model based on boosted decision trees in tall building construction. Autom Constr 23:47–54
55. Goldberg Y (2016) A primer on neural network models for natural language processing. J Artif Intell Res 57:345–420
56. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized BERT pretraining approach, arXiv preprint.
57. Narayanaswamy GR (2021) Exploiting BERT and RoBERTa to improve performance for aspect based sentiment analysis, dissertation. Technological University Dublin, Dublin. https://doi.org/10.21427/3w9n-we77

58. Soler AG, Apidianaki M (2021) Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. arXiv preprint
59. Sun C, Huang L, Qiu X (2019) Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. arXiv preprint .
60. Gao Z, Feng A, Song X, Wu X (2019) Target-dependent sentiment classification with BERT. IEEE Access 7:154290–154299