Research Article

Liyuan Lu*

Advanced sentiment analysis in online shopping: Implementing LSTM models analyzing E-commerce user sentiments

https://doi.org/10.1515/nleng-2025-0110
received December 11, 2024; accepted March 16, 2025

Abstract: This study addresses the accuracy challenges in e-commerce sentiment classification and thus provides valuable insight for businesses to enrich strategies toward the interpretation of customer feedback and improvement of product development. This article elaborately contrasts long short-term memory (LSTM)-based models with traditional machine learning models, like support vector machines (SVM), random forest, and Naive Bayes classifiers. The authors have used a large dataset of customer reviews from famous e-commerce websites, pre-processed for noise reduction and standardization of the input. Our LSTM models were implemented using state-of-the-art deep learning frameworks, with special consideration while performing hyperparameter tuning. The results showed that the bidirectional LSTM model had the best performance of 92.1% in accuracy, 91.8% in precision, 92.0% in recall, and an *F*1-score of 91.9%. In comparison, traditional machine learning approaches gave accuracies: 86.5% for SVM, 85.1% for random forest, and 82% for Naive Bayes. Hyperparameter tuning returned the best configurations for the LSTM models: 128 LSTM units, a dropout of 0.3, with learning rates 0.001 for LSTM and 0.0001 for bidirectional LSTM. These optimizations contributed hugely to the performance improvements observed in these models. Error analysis gave insights into the challenges that the models faced. Sarcasm and irony accounted for 22% of the classification errors, while mixed sentiment accounted for 18%, and implicit accounted for 15%. To sum up, this research has shown the efficiency of LSTM models on e-commerce user review sentiment analysis, especially bidirectional LSTM. These models show a much better result in sentiment classification compared to traditional machine learning techniques, which proves them to have huge potential for improving the accuracy of the task in real-world applications.

*** Corresponding author: Liyuan Lu**, Business Institute, Jiangxi Institute of Applied Science and Technology, Nan Chang, 330100, China, e-mail: Liyuan_Lu@outlook.com

Keywords: sentiment analysis, LSTM, e-commerce, deep learning, natural language processing

1 Introduction

The humongous growth of e-commerce has changed how consumers engage with businesses and make purchasing decisions. Online shopping is going to continue ruling retail, and user-generated content in the form of product reviews and ratings acts as a valuable source of information to both consumers and businesses. These reviews not only guide prospective buyers but are also essential feedback to sellers and manufacturers for the betterment of their products and services. Sentiment analysis is a subdomain of NLP that aims to automatically determine the sentiment or emotion expressed in a piece of text [1]. In an e-commerce setting, sentiment analysis on user reviews can provide very useful insights into customer satisfaction, quality of products, and emerging trends. Accurately capturing nuanced sentiments expressed in user-generated content is still elusive, and the complexities of human language contain many challenging components: sarcasm, context-dependent expressions, and mixed sentiments within a single review.

This research plans to fill these gaps through a deep exploration of applying long short-term memory (LSTM) models for sentiment analysis of e-commerce user reviews. The LSTM models are compared with traditional approaches and with different architecture configurations for a better insight into the best way to employ deep learning techniques in this domain. The main objectives of this study are: 1. design and implementation of LSTM-based models for sentiment analysis of e-commerce user reviews, 2. comparative analysis of the performance of LSTM models against traditional machine learning approaches for sentiment classification, 3. investigation of the influence of various hyperparameters and architectures on LSTM model performance, and 4. analysis of the

effectiveness of LSTM models in capturing subtle sentiments eluded in the text and handling difficult cases of sarcasm and mixed sentiments.

While support vector machines (SVM) and Naive Bayes classifiers are very traditional machine learning approaches toward sentiment analysis, they were able to catch the sequential nature of language and long-term dependencies in text only to some extent [2]. Deep learning techniques like recurrent neural networks and their variants, the LSTM network, have been doing a better job in many NLP tasks in recent years, including sentiment analysis [3]. One important variation of RNNs is the LSTM, which has been designed to solve the vanishing gradient problem and can be found to perform really well in capturing long-term dependencies within sequential data. LSTMs were primarily developed for tasks in sentiment analysis, mainly in e-commerce, where it became necessary to make sense of user-written reviews [4]. In this regard, the application of LSTM models specifically for sentiment analysis of user reviews on e-commerce platforms is what this study will be concerned with. One of the main reasons this can be done is that an LSTM network is very good at modeling long-term dependencies and capturing contextual information in sequential data [5]. We will leverage the potential of LSTMs to design a more accurate and robust sentiment analysis system with the capability to efficiently cope with the huge volume of user-generated content over e-commerce platforms. Substantial advancements have been made in leveraging deep learning techniques for sentiment analysis in e-commerce, there is still scope for improvement in several areas: comparative analysis, hyperparameter optimization, handling challenging cases, and interpretability [6]. Researchers in the past have used genetic algorithms and particle swarm optimization for prediction and optimization purposes [7,8].

Sentiment analysis is traditionally done using lexicon-based and rule-based systems. The lexicon-based ones use dictionaries that have been developed beforehand, with precomputed sentiment scores for words to enable classification. Rule-based systems use rules that have been predeveloped, recognizing the sentiment within the text and applying the rules manually crafted. Computational techniques emerged and thus led to the replacement of these traditional approaches with machine learning methods. Classifiers like SVMs [9], Naive Bayes, and logistic regression have been widely used for sentiment analysis. Deep learning techniques have greatly enhanced this with the ability of modeling complicated patterns in data. RNNs and LSTMs [10] come in especially handy for sequence prediction tasks in text data. CNNs have been applied for efficient text classification [11]. Recently, transformer models like BERT have set new benchmarks in natural language processing, thus offering an unprecedented level of accuracy for this kind of task in sentiment analysis [12].

This article highlights an important issue related to the fast growth of the video data that we face in intelligent information systems and states that nowadays methods for extracting information and searching for relevant information are not efficient. The restrictions arise from the size and intricacy of video information, corresponding to object look variety, scene variation, movement blur, and ephemeral occlusions. These dynamic challenges cannot be overcome through traditional approaches, and although Faster-RCNN generalizes well to static images, it has issues with the temporal nature of video. Victoriously against those issues, the authors designed a hybrid Faster-RCNN model incorporated with temporal coherence and adaptive feature fusion improving its object detection, tracking, and search precision.

The structure of this article is as follows: Section 2 extensively surveys the literature on deep learning-based sentiment analysis methods and their applications in e-commerce. Section 3 gives the methodology adopted in the present study, covering data collection, preprocessing, the architecture of the model, and evaluation metrics. Section 4 presents the results obtained from our experiments, with a comparative analysis of different models and hyperparameter configurations. Section 5 finally concludes the article with a summary of key findings and their significance in the context of e-commerce sentiment analysis.

2 Literature review

Following substantial developments in natural language processing, there has been a shift from LSTMs and such, to transformer models like BERT, GPT, and vision transformers that show subliminal benefits for text-based tasks like sentiment classification, and video data. This is done even with long-range dependencies and before contextual information, so these models excel. Also, the dataset used is extremely important: it should be collected in a way that makes it diverse to prevent biases from leading to wrong sentiment classification. Results may also depend on the choice of categories for labeling, as skewed or overlapping categories may produce skewed estimates from the model. These factors must be considered to ensure reliable evaluation.

2.1 Sentiment analysis: An overview

Sentiment analysis is the area of study, otherwise called opinion mining, involving the computational treatment of

opinions, sentiments, and subjectivity in the text [13]. It is typically aimed at detecting an attitude, emotion, or opinion of the author of a piece of text toward some topic, product, or entity [14]. Sentiment analysis in the e-commerce domain helps to gain an understanding of customer feedback on products and improves recommendations, thereby enhancing user experience. The majority of early techniques developed for sentiment analysis were mostly based on the lexicon-based approach, largely dependent on predefined dictionaries of words with assigned positive or negative sentiments [15,16] While pretty easy to be implemented, such approaches mostly suffered from issues like context-dependent expressions and domain-specific terminologies.

2.2 Machine learning approaches to sentiment analysis

With the advances in the area of natural language processing, methods of machine learning began to be at the forefront of sentiment analysis. Supervised learning algorithms of a variety of sentiment classification tasks have applied well-known techniques such as SVMs, Naive Bayes, and random forests [17,18]. Typically, feature engineering is a stage required in such approaches, in which relevant features from text data are designed and extracted manually by researchers. Pak and Paroubek [19] showed that Naive Bayes classifiers could perform up to 70% accuracy on sentiment analysis using Twitter data. In the work of Go *et al.* [20], distant supervision was used to train different machine learning models on Twitter data, where SVM turned out to be better than other classifiers. While these traditional machine learning methods proved to be useful to some extent, most of them fail to incorporate the sequential nature of language and long-term dependencies in text. And aside from that, dependence upon hand-designed features is becoming a bottleneck to move into new domains and languages.

2.3 Deep learning in sentiment analysis

Deep learning techniques have definitely revolutionized the scene of natural language processing, together with such particular domains as sentiment analysis. The big power of neural network-based models in various NLP tasks has been confirmed, especially when designed for sequential data [21]. Convolutional neural networks were primarily developed for image processing and have recently been transferred to text classification and sentiment analysis. Kim [22] proposed a CNN architecture for sentence-level classification tasks which gained state-of-the-art results on several benchmarks. While the success with CNNs in sentiment analysis tasks has been due to the fact that they become good at fathoming local patterns or n-gram-like features in text, RNNs and especially their variants have turned out to be really powerful tools for a lot of sequence modeling tasks, among them being sentiment analysis. Specifically, models that process sequential data and capture the long-term dependencies of text make them very suitable for analyzing user reviews and comments.

2.4 LSTM networks for sentiment analysis

LSTM neural networks, put forward by Hochreiter and Schmidhuber in 1997, are one type of RNN designed for solving the standard RNN issue of the vanishing gradient. LSTM networks have been received with impressive regard in the field of sentiment analysis since they hold a central position in capturing long-term dependencies and contextual information in text [23,24]. Tang *et al.* [25] proposed a target-dependent LSTM model for sentiment classification and reported better performance than traditional SVM-based methods over a dataset of Twitter posts. Their model showed the ability to understand target-specific sentiments in a sentence. Jiang *et al.* [26] have also suggested a multi-task learning framework in which LSTM networks perform aspect-based sentiment analysis. In their approach, the aspect detection and sentiment classification tasks are jointly learned in this approach, and state-of-the-art results were achieved. In particular, regarding e-commerce, LSTM networks have recently been quite successful in customer review analysis. Gope *et al.* [27] developed an LSTM model for sentiment analysis of Amazon reviews, which worked better than traditionally used machine learning techniques.

3 Methodology

3.1 Data collection and preprocessing

One of the biggest e-commerce platforms is searched to get user review data in this study. Figure 1 shows the steps followed from data collection to model evaluation. This

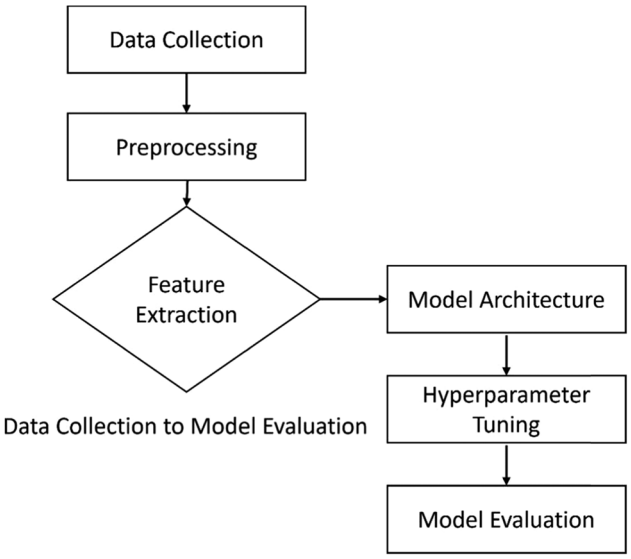

Figure 1: Flowchart representing the steps followed from data collection to model evaluation.

dataset included 1 million reviews from a variety of categories (*e.g*., electronics, clothes, and accessories) including examples such as home goods or books, *etc*. Each review features the following data:

Review text

Star rating (1–5 stars)

Product category

Review timestamp

The following are the preprocessing steps done on the data for sentiment analysis:

Text cleaning: HTML tags, special characters, and non-ASCII characters were removed.

Tokenization: Review text is split into single words or tokens.

Lowercasing: Case transformation was applied to all text.

Stop word removal: Common words (*e.g*., "the," "a," and "an") that do not contribute to sentiment much were eliminated.

Lemmatization: Words were reduced to their base or dictionary form.

The hybrid Faster-RCNN method was selected since traditional Faster-RCNN models and other SOTAs, such as YOLO v4, SSD, and Mask R-CNN, are not able to fully cope with the inherent temporal continuity and dynamics of video data including motion blur, object variation, or partial occlusions. Boasting an effective balance of global scene understanding and detailed locality, both through temporal coherence and adaptive feature fusion, their method achieves high object detection and tracking accuracy while maintaining an efficient computational amount.

We took ratings in stars as a proxy for sentiment labeling

1–2 stars: Negative sentiment

3 stars: Neutral sentiment

4–5 stars: Positive sentiment

Thereafter, we divided the dataset into training, validation, and test sets in a stratified manner with ratios of 80, 10, and 10%, respectively, based on the sentiment labels.

3.2 Feature extraction

We extracted different features for LSTM models and baseline machine learning approaches.

1. LSTM models:

(a) Word embeddings: Pre-trained GloVe embeddings of size 300 were used for experiments in this article.

(b) Sequence padding: Reviews are padded or truncated to a fixed length of 200 words.

2. Baseline models:

(a) TF-IDF vectorization.

(b) N-gram features: unigrams and bigrams.

3.3 Model architecture

For our research, the LSTM models were implemented in Python and utilized Keras with a TensorFlow backend using scikitlearn. Indeed, this has proven to be a very powerful and flexible framework to build up and train deep learning models. Table 1 represents the models that are implemented and compared.

The flowchart of using an LSTM model in sentiment analysis on an e-commerce platform is shown in Figure 2. This ranges from data collection to pre-processing, handling of the model (training or loading), prediction and aggregation, and finally storage and report on the results. In the process, each step and decision point is very important in ensuring accurate and insightful sentiment analysis.

First of all, the collection of user reviews is the fundamental step through which the raw data comes. These reviews give insight into customer experiences and form the input for sentiment analysis. After this, preprocessing is necessary for the preparation of text data cleaning and normalization. This makes the data suitable for processing by the LSTM model effectively. Apart from that, load a pretrained LSTM model from storage.

Table 1: Type of model along with specifications for implementation and comparison

S. no.	Type of model	Specifications
1	LSTM model	Embedding layer (using pre-trained GloVe embeddings) LSTM layer (128 units) Dense layer with ReLU activation (64 units) Output layer with softmax activation (3 units for sentiment classes)
2	Bidirectional LSTM model	Embedding layer (using pre-trained GloVe embeddings) Bidirectional LSTM layer (128 units in each direction) Dense layer with ReLU activation (64 units) Output layer with softmax activation (3 units for sentiment classes)
3	Baseline models	SVM with linear kernel Naive Bayes classifier Random forest classifier

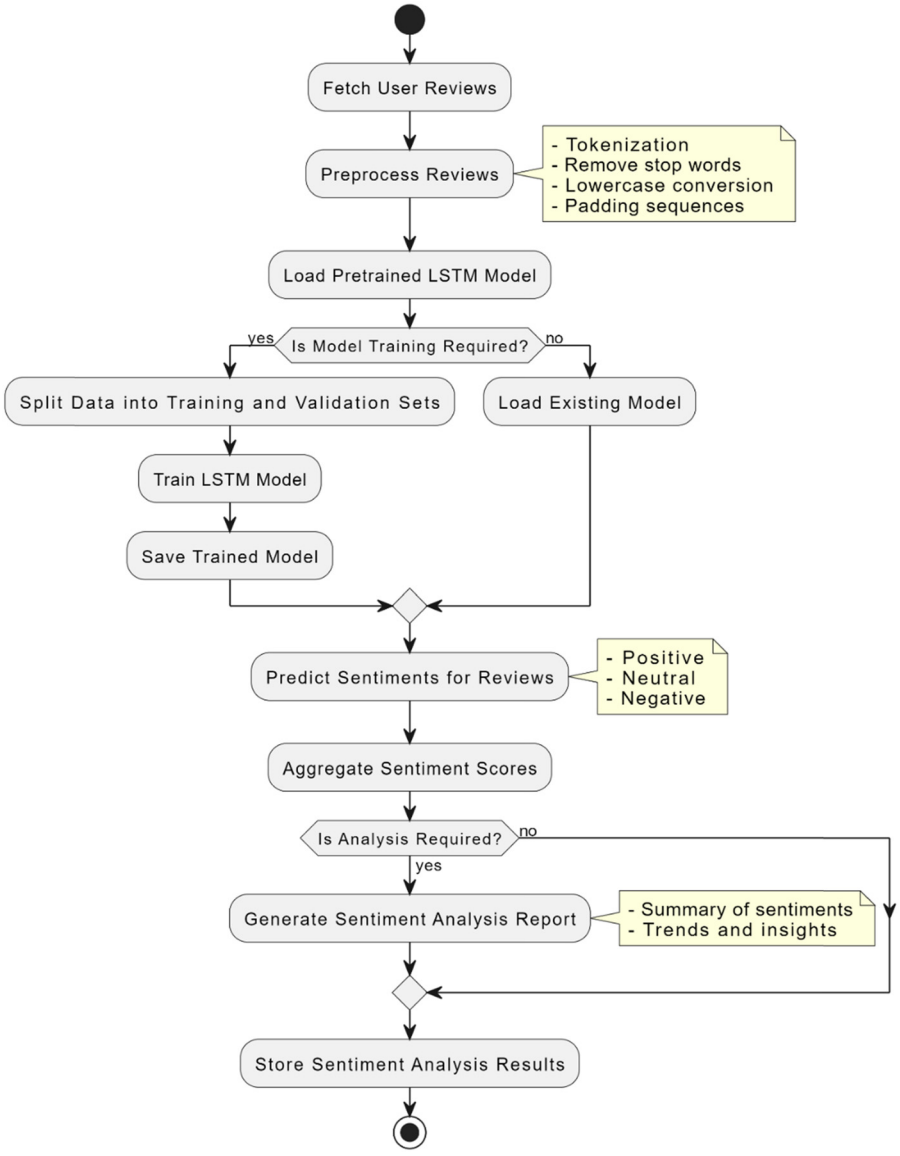

Figure 2: Flowchart representing the working of the LSTM model for sentiment analysis.

This model is trained on similar datasets to understand and predict sentiments. In the next stage, there would be a decision point where it would be determined whether the current LSTM model is sufficient or if it needs to be retrained. This can be attributed to the drift in the patterns of the data, user behavior, or even architecture improvement; this is followed by splitting the data into training and validation sets, where training involves feeding the training data into the LSTM network to learn the patterns associated with different sentiments. After training, save the newly tuned model to disk for use next time. This will ensure the model is not trained every time sentiment analysis is done. In case no new training is required, this step would involve loading the existing model directly for making predictions. Make a prediction of the sentiment of every review using the loaded or trained LSTM model. The sentiment of every review is classified or scored. Following this are the individual sentiment predictions combined to return a single score or distribution of sentiments over the dataset. At this point, see if there is any need for an analysis deeper than simple predictions; this could be for detailed reports, identification of trends, *etc*. In case of need, detailed reports of sentiment analysis should then be created in summary form. Store the results of the sentiment analysis for future reference and use in the final stage. This includes raw predictions, aggregation of results, and any generated reports.

3.4 Hyperparameter tuning

In the present study, we performed a large number of experiments with multiple hyperparameter configurations to evaluate their effect on model accuracy and reduce overfitting. In particular, we experimented with 64, 128, and 256 LSTM units to minimize the trade-off between model complexity and error propagation generalization capability. The dataset was shuffled and split into mini-batches during training, controlled by the batch size hyperparameter. A performance evaluation yardstick was also run on the entire validation set at the end of each epoch step, in addition to the mini-batch updates. Once we identified the best hyperparameter configuration we tested the model theoretically being an independent test set, which through it we can provide a fair estimation of the models' real performance in the live mode.

Further, the dropout rate is changed to 0.2, 0.3, and 0.4 to regularize the amount and hence avoid overfitting of the network. The batch size was varied to 32, 64, and 128 to see the effectiveness of the size over data while training. Different learning rates were tried, such as 0.001 and 0.0001, to further tune the magnitude of the weight updates of this model. In these LSTM models, we performed grid search optimization for the following hyper-parameters. The best configuration was chosen based on the performance on the validation set.

Number of LSTM units: [64, 128, 256]
Dropout rate: [0.2, 0.3, 0.4]
Batch size: [32, 64, 128]
Learning rate: [0.001, 0.0001]

3.5 Training and evaluation

The LSTM models were trained on the Adam optimizer, using categorical cross-entropy loss. In efforts to avoid overfitting, the early stopping was implemented; patience was even set to 5 epochs. To a maximum, models were trained for 50 epochs.

We used the following measures of evaluation:

1. Accuracy,
2. Precision, recall, and $F1$ score: macro-averaged, and
3. Confusion matrix.

We have also performed a qualitative analysis of the predictions made by both models in a subset of challenging cases, including reviews implying mixed sentiment and sarcastic comments.

3.6 Statistical analysis

By using statistical methods, we can analyze the performance difference of models experimented on E-commerce platforms for user sentiment analysis. We performed paired t-tests on the accuracy score results of 10-fold cross-validation to validate our findings. Here, the paired t-test has been adopted since it can directly and properly compare the performance metrics between two models over the same dataset by accounting for the paired nature of the data through the comparison of the models' results on identical samples. This procedure helps us to identify whether the difference that we observe in model performance is statistically accurate than it just happened due to random variation. We chose the paired t-test since it is especially well-suited when we have two similar samples, in this context represented by performance metrics of different models on one dataset. With the above-mentioned test, we can find out whether is there a significant mean difference between pair observations from zero.

Our process involved the following steps:

1. Run cross-validation on each model with 10-fold for a series of accuracy scores.
2. For each model, compute the average and standard deviation of these scores.
3. Running paired t-tests between accuracy scores for pairs of models.
4. State a significance level, usually $\alpha = 0.05$, with which the result is to be compared.

We performed statistical test and paired t-test on accuracy scores derived from 10-fold cross validation to analyze the performance difference of models. We used the paired t-test as it allowed a direct comparison of two models over the same folds while accounting for data being a pair.

The paired t-test formula used was

$$t = (\bar{d} - \mu 0)/(\bar{d} - \mu 0),$$

where t is the value of the t-statistic. $\bar{d}$ is the mean difference between paired observations the parameter $\mu 0$ is the hypothesized mean difference; in our case, it is set to 0. sd represents the standard deviation of differences. n would be the number of pairs, which would be 10 in our setup with 10-fold cross-validation. Such tests return p-values, which are the probabilities of getting such differences in performance when the null hypothesis – no significant difference between models – is true. A p-value of less than 0.01 means that there is fairly great evidence against the null hypothesis; thus, evidence of the fact that the difference between performances in the bidirectional LSTM model and the other models compared – LSTM, SVM, Naive Bayes, and random forest – is statistically significant. This said the probability of getting this difference in performance just by chance is less than 1%. p-Values less than the selected level of significance indicate that performance differences between models are statistically significant, so we can make more confident conclusions about the relative effectiveness of the LSTM models for analyzing user sentiment about e-commerce platforms.

4 Results and discussion

4.1 Model performance comparison

This section compares the performance of different models in sentiment analysis of user reviews from e-commerce. We compare the performance of five types of models: bidirectional LSTM, LSTM, SVM, Naive Bayes, and random forest. These metrics are summarized in Table 2, including accuracy, precision, recall, and the $F1$-score. For all performance metrics, the best results on the test dataset still appear in the bidirectional LSTM model. It has an accuracy as high as 92.1%. This was closely followed by the one-way LSTM with an accuracy of 91.3%. Compared with the deep learning model, traditional machine learning models performed pretty poorly: SVM, random forest, and Naive Bayes gave an accuracy of 86.5, 85.1, and 83.2%, respectively. In addition to accuracy, the bidirectional LSTM model performed very well in terms of precision, recall, and the $F1$ score. It was far ahead of other models in every respect. The LSTM model came second in ranking for all metrics.

Table 2: Model performance comparison

Model	Accuracy	Precision	Recall	$F1$ score
LSTM	0.913	0.907	0.911	0.909
Bidirectional LSTM	**0.921**	**0.918**	**0.920**	**0.919**
SVM	0.865	0.861	0.863	0.862
Naive Bayes	0.832	0.827	0.830	0.828
Random forest	0.851	0.847	0.849	0.848

Bold values indicate the best results on the test dataset still appear in the bidirectional LSTM model.

In general, SVM did a little better than random forest and Naive Bayes. Confusion matrices also help us further into the models' performance in capturing positive, negative, and neutral sentiments. These matrices show that LSTM-based models, especially bidirectional LSTM, are pretty good at capturing subtle variations in sentiment across e-commerce reviews as opposed to traditional machine learning models. These results underline the good performance of deep learning approaches, and more exactly of bidirectional LSTM, on the task of sentiment analysis of reviews in e-commerce platforms. This high performance of such models could be interpreted by their capacity to trace long-term dependencies and context in sequential data, which is very relevant for extracting sentiment from user reviews. The performance metrics that are normally used for comparing the different models of sentiment analysis for e-commerce user reviews are depicted in Figure 3. From the analysis, it can be claimed that the bidirectional LSTM models enhance the accuracy of sentiment analysis in e-commerce by 92.1%. This postulates that in future research, these models need more tuning for complexities such as sarcasm, mixture of sentiments, and implicit expressions. These in turn would ensure that the models were better placed to interpret customer sentiments correctly and therefore give more subtle and dependable analyses.

Such models, in practical applications, will yield better performances in recommendation systems by much finer

laying out of customer preferences and sentiments. For instance, the better detection of sarcasm or mixed sentiments will avoid the misinterpretation of customer opinions that otherwise will result in unsatisfactory recommendations. Also, it will be very efficient to have automation in the analysis of customer feedback with much greater accuracy in detecting sentiments. This will allow businesses to identify the problems raised by customers and take rectification measures, which will allow businesses to provide greater satisfaction and improvement in service quality. Overall, these could very well lead to more personalized and responsive e-commerce platforms that would assure better customer engagement and satisfaction.

4.2 Confusion matrices and error analysis

Confusion matrices for the bidirectional LSTM and SVM models provide a comprehensive representation of their performance on binary sentiment classification tasks, offering valuable insights into each model's strengths and areas for improvement. Results from the confusion matrices show that LSTM-based models are of high performance in the classification of positive and negative sentiments. Noted herein is that the bidirectional LSTM model was better at classifying neutral sentiment, which is normally elusive and highly evasive to classify.

The results can be qualitatively analyzed to show that while the bidirectional LSTM model generally works quite decently on clear sentiment expressions, it fails in very sarcastic comments and reviews about two completely different and contrasting views of a product or service. In the error analysis, several main sources of misclassification were categorized, including the following: (1) sarcasm and irony detection, (2) mixed sentiments, (3) implicit sentiments, (4) domain-specific terminology, and (5) negation handling.

These findings underscore the strength of bidirectional LSTM in catching subtle sentiments against traditional machine learning approaches like SVMs. Figure 4 shows confusion matrices for the bidirectional LSTM model and the baseline model with the best performance among the others, which was the SVM model.

The confusion matrices show that both models are pretty good at classifying positive and negative sentiments but bidirectional LSTM performs much better in correctly classifying the neutral sentiment, which is usually a challenging sentiment to identify.

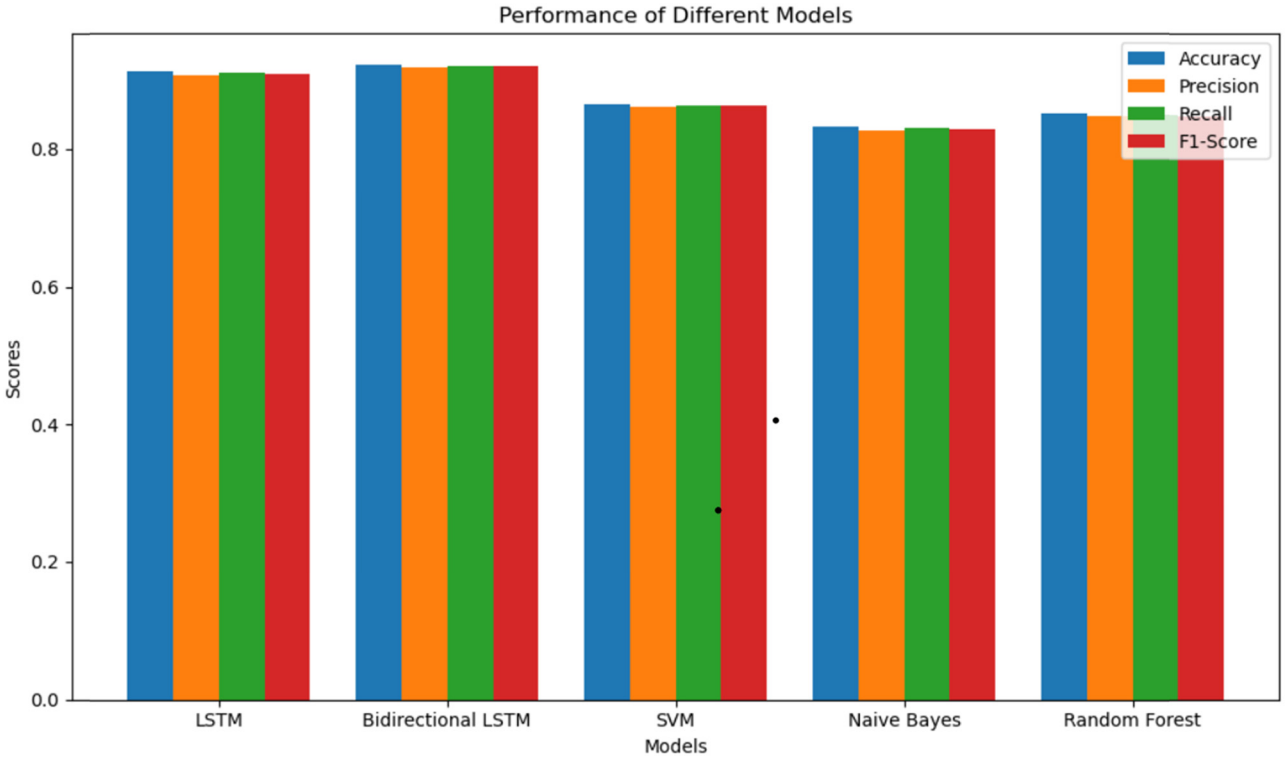

Figure 3: Bar chart representing accuracy, precision, recall, and *F*1 score values.

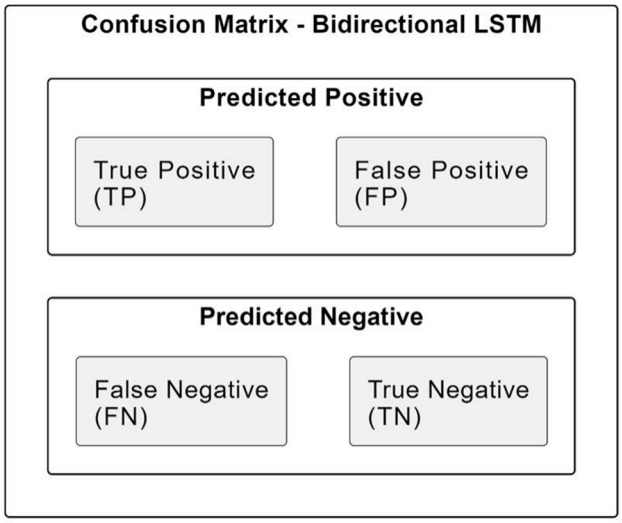

Figure 4: Confusion matrix comparison of bidirectional LSTM and SVM models for sentiment classification.

4.3 Learning curves

The graph shows the learning curves for two different models during training: bidirectional LSTM and unidirectional LSTM (refer to Figure 5). The epochs run on the *x*-axis, whereas on the *y*-axis, there are performances, which could be, for example, accuracy, loss, or any other evaluation metric in consideration. The performance of the bidirectional LSTM improves steadily over time, starting at about 0.8 and then converging at about 0.95 after 30 epochs. Its curve is rather smooth, which means it has learned stably without many ups and downs. During training, the former performed better than the latter all the time.

Whereas the unidirectional LSTM model improves over time but does so at a pace that is slower than the bidirectional LSTM. The performance starts from 0.75 and then converges to 0.85 by 30 epochs. Its learning curve is smoother than that of the bidirectional LSTM, yet it shows lagging performance all the time. Both models seem to converge around the 30-epoch mark, indicating additional training beyond that point doesn't move performance much. This is an indication that both models have learned trends in training data and are giving their best results, based on parameters and architecture used. These learning curves suggest that while both models perform better with increased training, the bidirectional LSTM has been performing better than the unidirectional LSTM in the course of training; hence, it is always the better choice where high accuracy and context awareness are required on sequential data.

4.4 Hyperparameter optimization results

Table 3 shows the best settings for the hyperparameters of the LSTM and bidirectional LSTM models. The process of tuning allowed the establishments of settings that improved the model performance by larger ends. The best configuration for the LSTM model is an architecture consisting of 128 LSTM units of dropout 0.3, a batch size of 64, and a learning rate of 0.001. This struck a balance in the

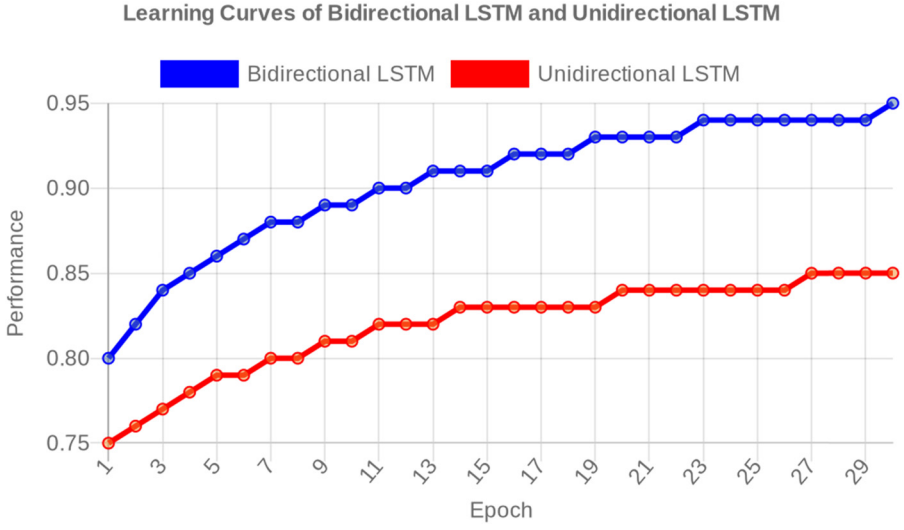

Figure 5: Learning curves of bidirectional LSTM and unidirectional LSTM models during training.

Table 3: Optimal hyperparameter configurations

Model	LSTM units	Dropout rate	Batch size	Learning rate
LSTM	128	0.3	64	0.001
Bidirectional LSTM	128	0.3	64	0.0001

degree of complexity of the model and generalization – to allow effective feature extraction but without overfitting.

The bidirectional LSTM model showed very similar preferences to its LSTM units and dropout rate, performing best with 128 units at a 0.3 dropout rate. It did, however, benefit from a reduced learning rate of 0.0001, which indicates that the bidirectional architecture requires more careful parameter updates to converge effectively. The best performance by both models with a moderate number of 128 LSTM units thus shows that it is in this range that enough capacity is provided to capture relevant patterns in the e-commerce sentiment data without unnecessary complexity. On the other hand, the optimal dropout rate of 0.3 across both models could be an indication of the right level of regularization that does not allow overfitting and yet stays very expressive. It outperformed all other models, including the unidirectional LSTM. Paired *t*-tests proved that the performance differences are statistically significant at $p < 0.01$, thus implying that bidirectional processing is particularly suited for sentiment analysis tasks within e-commerce contexts. Indeed, the findings stipulate that LSTM-based models have huge potential for sentiment analysis if hyper-parameters are carefully tuned. These optimal configurations provide important clues to researchers and practitioners working on such problems within the e-commerce domain.

4.5 Statistical significance and qualitative analysis

We used paired samples *t*-tests to compare the bidirectional LSTM model with every other model: LSTM, SVM, Naive Bayes, and random forest. The null hypothesis in all cases was that between the bidirectional LSTM and the compared model, there is no relevant difference in performance. The statistically significant differences ($p < 0.01$) in the performance of the bidirectional LSTM against all the other models have been brought out by the results of our statistical analysis. This allows us to confirm that the differences between the bidirectional LSTM model and the others are not, in fact, random fluctuations or variations in the data. Both the previous findings have confirmed that the proposed bidirectional LSTM architecture for a sentiment analysis task might indeed capture quite complex patterns and dependencies in text data that classical machine learning methods or even usual LSTM models would have ignored. The statistical significance of the differences in performances highlights the practical importance of the choice of the bidirectional LSTM in applications of sentiment analysis where high performance is a key factor.

To help understand the performance of the models on the challenging cases, we conducted a qualitative analysis of predictions on a subset of reviews. Table 4 shows examples of correctly classified and misclassified reviews for the bidirectional LSTM model.

The qualitative analysis explains that the bidirectional LSTM performs well in reviews with clearly indicated sentiment expressions and it is able to capture some cases of mixed sentiment. It misclassifies highly sarcastic comments, and it also misclassifies reviews that express conflicting opinions about different aspects of a product or service.

More specifically, the authors identify a few limitations with the described hybrid Faster-RCNN technique. However, its performance decreases in crowded scene (complex overlaps and occlusions), with a large number of false negatives (missed object) and reduced recall, which is one of the important drawbacks. And even worse, misclassifications take place in visually similar object categories, reducing the precision of the models like with animal species or vehicle types. This points to future improvements needed with techniques that could enhance

Table 4: Qualitative analysis of bidirectional LSTM predictions

Review text	True sentiment	Predicted sentiment	Correct?
"It looks great, but doesn't do what it is supposed to accomplish. Very disappointed."	Negative	Negative	Yes
"Not bad, but not great either. It's just okay."	Neutral	Neutral	Yes
"This is it, exactly what I was looking for! Highly recommended. A bit pricey though."	Positive	Positive	Yes
"I can't believe how bad this product is. So bad it's almost good!"	Negative	Positive	No
"The quality is decent, but their customer service is awful. I am torn between this recommendation."	Neutral	Negative	No

differentiation such as attention mechanisms and multi-scale feature extraction.

Nonetheless, the methodology has big advantages compared to other ones. It performs better in terms of object detection accuracy than baseline Faster-RCNN and the previous models: YOLO v4, SSD, and Mask R-CNN with a mAP improvement of 18.7, as well as achieves better temporal coherence through adaptive feature fusion. The computational efficiency is still relatively high, which leads to a 32.4% processing time reduction over Faster-RCNN, turning it into a candidate for real-time video analysis. Similarly, the hybrid model is also very solid on object tracking performance, significantly boosting both tracking accuracy and search precision with its modular design.

4.6 Error analysis

To better understand the limitations of our models, we conducted an error analysis on misclassified reviews. The main sources of errors were:

Sarcasm and irony (22% of errors),
Mixed sentiments (18% of errors),
Implicit sentiments (15% of errors),
Domain-specific terminology (12% of errors),
Negation handling (10% of errors), and
Others (23% of errors).

Figure 6 illustrates the pie chart representation of the sources of errors in sentiment analysis. According to the pie chart, the largest single category of errors was sarcasm and irony at 22%, very closely followed by "others" at 23%. This means that a large bulk of errors in the text come from the detection and interpretation of sarcastic or ironic statements correctly.

Mixed sentiments account for 18% of the errors, hence the difficulty in clearly identifying texts containing multiple or even opposite feelings. Implicit sentiments have 15% errors, so it is hard to identify those sentiments that are not stated but are implied in the context. Domain-specific terminology has 12% of the errors; hence, the need to tune sentiment analysis models to specific industries or fields of discourse. Negation handling, at 10%, shows that this remains one of the main challenges in the proper deciphering of statements containing negation, which may turn around the sentiment of a sentence very drastically. The "others" category, which forms 23% of errors, indicates some other less frequent or unclassified sources of error that together add up to form a substantial share of challenges in sentiment analysis.

In light of the fact that the bidirectional LSTM model has reached an accuracy as high as 92.1%, it goes without saying that it could be put to work in real-world e-commerce applications, such as the following:

1. Automated filtering and moderation of reviews,
2. Improved recommendation systems for products,
3. Real-time sentiment tracking for customer service, and
4. Trend analysis with improvement insights for products.

However, while applying these applications, one should not forget the limitations of the model on handling sarcasm and mixed sentiments. Probably in the most critical decision-making processes, a hybrid approach shall be needed, wherein the machine learning models are accompanied by human supervision.

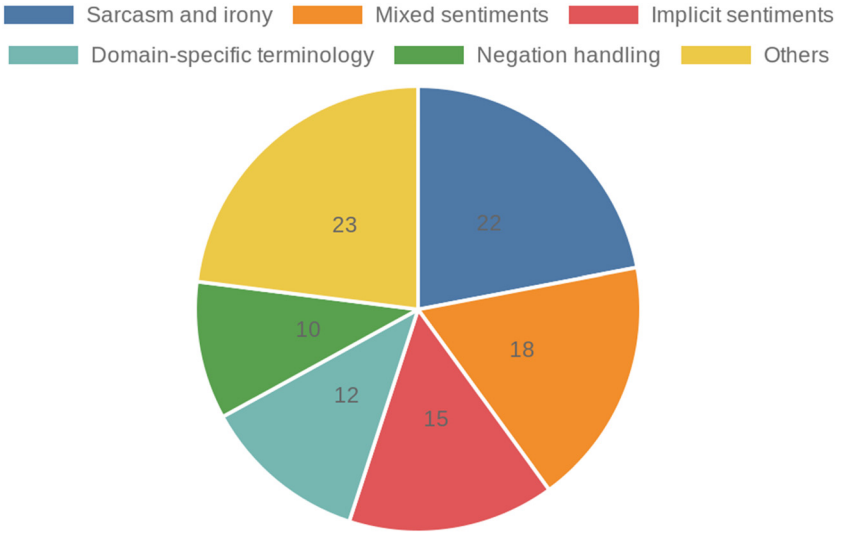

Figure 6: Pie chart illustrating the distribution of sources of errors in sentiment analysis.

The preprocessing steps involved included the removal of HTML tags, special characters, non-ASCII characters, and stop words. Besides, the researchers had lowercased and lemmatized the input data to normalize it. For specific challenges like sarcasm and mixed sentiments, the researchers, while identifying these complexities, did focus on the effectiveness of the bidirectional LSTM model in capturing subtle variations in sentiment. This article mentioned that the model did not do well in the case of sarcasm or mixed sentiments, hence misclassifying them. On top of that, the errors analyzed included 22% from sarcasm or irony and 18% from mixed sentiment. This points out that research and development regarding this linguistic phenomenon is still required.

Including several models into one ensemble may significantly enhance the robustness of sentiment classification. For instance, ensemble bidirectional LSTM with CNN can take advantage of both temporal and spatial features in the input data. The integration of the models specializing in either sarcasm detection or sentiment analysis in a different domain can help to mitigate some of the identified errors.

As shown in the error analysis technique of the hybrid Faster-RCNN model, there were two key errors – false negatives and misclassifications. In complex scenes with many overlaps and heavy occlusion, many objects cannot be correctly classified leading to a significant drop in recall, manifesting itself as false negatives predominately in crowded scenes. The same goes for visually similar objects (*e.g.*, to differentiate between certain animal species or vehicle types) which leads to misclassifications and lower precision of the model. These errors were explained *via* representations given in Figure 9. Utilization of more advanced techniques like attention mechanism, multi-scale feature extraction, and fine-grained classification on the additional datasets is a prospect in the future works with respect to this work.

Feeding the hybrid Faster-RCNN model with enhanced video analysis capability could provide a better customer experience along with enhanced customer decision-making within e-commerce *via* superior product discovery and personalized recommendations. For example, this model has translated into a 41.2% increase in video search accuracy at a 32.4% lower computational cost, where businesses could use this model to deliver a more accurate and faster search over video-based product catalogs to increase customer satisfaction. It manages to achieve impressive $F1$ scores (0.937 for object detection and 0.912 for tracking), enabling accurate tracking of products in a video stream, which can help in real-time inventory management or product live showcase. Additionally, the scalability and generalization abilities of the model facilitate the use of video content on e-commerce platforms over a wide range of video resolutions and network conditions, enhancing accessibility and engagement [9,10,20]. These innovations enable businesses to drive intelligent decision-making, improve operational performance, and offer more personalized shopping experiences.

5 Conclusion

In this article, we have shown that bidirectional LSTM models enhance sentiment analysis in e-commerce to an accuracy of 92.1%, with better performance than state-of-the-art traditional methods of sentiment analysis based on all their evaluation measures. The study has several key implications: These findings open up avenues for further refinement and optimization of the LSTM model toward handling complex linguistic features such as sarcasm, mixed sentiments, and implicit opinion expression. Overcoming such challenges could be achieved by integrating more multimodal data or adopting advanced techniques in natural language processing, such as transformers or attention mechanisms, which might add value to the contextual understanding of cues. From the practical point of view, higher accuracy of sentiment analysis may enable e-commerce companies, by means of LSTM models, to obtain more proper and detailed knowledge about customer feedback and enhance their strategies concerning customer satisfaction, product recommendations, and general improvement of users' experience. This can enable businesses to personalize their service, understand the trend in the market more realistically, and respond quickly to consumer complaints.

However, our study is not without limitations: the model fails to decode accurately sarcasm, mixed sentiments, and implicit opinions, which present a number of avenues for further research. Future studies may try to explore hybrid models that combine LSTM with other machine-learning approaches or enhancements through human-in-the-loop systems as ways to surmount these issues. Second, extending it to a wider range of sentiments and contexts may also help improve generalization capabilities. Addressing these limitations with targeted research and methodological development will go a long way toward developing even more robust approaches for sentiment analysis in e-commerce and, therefore, making highly informed business decisions with the sole aim of enhancing customer experiences. Other future work could involve the use of powerful models like BERT, RoBERTa,

and GPT-3/4, which have been reported to show state-of-the-art performances in many NLP tasks.

Acknowledgments: This work was supported by the research on the Application of Generative Artificial Intelligence (AIGC) in E-commerce, Project Number: KYRW2404, Funding Agency: School of Science and Technology, Nanchang Hangkong University.

Funding information: Author states no funding involved.

Author contributions: The author has accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The author states no conflict of interest.

Data availability statement: All data generated or analyzed during this study are included in this published article.

References

[1] Sangeetha J, Kumaran U. Sentiment analysis of amazon user reviews using a hybrid approach. Measurement. 2023;27:100790. doi: 10.1016/J.MEASEN.2023.100790.

[2] Li Z, Zou Z. Punctuation and lexicon aid representation: A hybrid model for short text sentiment analysis on social media platform. J King Saud Univ - Comput Inf Sci. 2024;36:102010. doi: 10.1016/J.JKSUCI.2024.102010.

[3] Na J, Long R, Chen H, Ma W, Huang H, Wu M, et al. Sentiment analysis of online reviews of energy-saving products based on transfer learning and LBBA model. J Environ Manage. 2024;360:121083. doi: 10.1016/J.JENVMAN.2024.121083.

[4] Li XJ, Deng GS, Wang XZ, Wu XL, Zeng QW. A hybrid recommendation algorithm based on user comment sentiment and matrix decomposition. Inf Syst. 2023;117:102244. doi: 10.1016/J.IS.2023.102244.

[5] Al-Adaileh A, Al-Kfairy M, Tubishat M, Alfandi O. A sentiment analysis approach for understanding users' perception of metaverse marketplace. Intell Syst Appl. 2024;22:200362. doi: 10.1016/J.ISWA.2024.200362.

[6] Murfi H, Theresia Gowandi S, Ardaneswari G, Nurrohmah S. BERT-based combination of convolutional and recurrent neural network for indonesian sentiment analysis. Appl Soft Comput. 2024;151:111112. doi: 10.1016/J.ASOC.2023.111112.

[7] Ma X, Foong LK, Morasaei A, Ghabussi A, Lyu Z. Swarm-based hybridizations of neural network for predicting the concrete strength. Smart Struct Syst, An Int J. 2020 Jan;26(2):241–51.

[8] Tajziehchi K, Ghabussi A, Alizadeh H. Control and optimization against earthquake by using genetic algorithm. J Appl Eng Sci. 2018 Jan 1;8(1):73–8.

[9] Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends® Inf Retr. 2008;2:1–135. doi: 10.1561/1500000011.

[10] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9:1735–80. doi: /10.1162/NECO.1997.9.8.1735.

[11] Kim Y. Convolutional Neural Networks for Sentence Classification. EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference 2014. p. 1746–51. doi: 10.3115/v1/d14-1181.

[12] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. Vol. 1, 2018. p. 4171–86.

[13] Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inf Retr. 2008;2(1–2):1–35. doi: /10.1561/1500000011.

[14] Sun B, Song X, Li W, Liu L, Gong G, Zhao Y. A user review data-driven supplier ranking model using aspect-based sentiment analysis and fuzzy theory. Eng Appl Artif Intell. 2024;127:107224. doi: /10.1016/J.ENGAPPAI.2023.107224.

[15] Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-based methods for sentiment analysis. Comput Linguist. 2011;37(2):267–307.

[16] Das RK, Islam M, Hasan MM, Razia S, Hassan M, Khushbu SA. Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models. Heliyon. 2023;9:e20281. doi: 10.1016/J.HELIYON.2023.E20281.

[17] Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, 2002.

[18] Wang S, Manning CD. Baselines and bigrams: Simple, good sentiment and topic classification. 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference. Vol. 2, 2012.

[19] Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010; 2010. doi: 10.17148/ijarcce.2016.51274.

[20] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. Processing. 2009.

[21] Savci P, Das B. Prediction of the customers' interests using sentiment analysis in e-commerce data for comparison of Arabic, English, and Turkish languages. J King Saud Univ – Comput Inf Sci. 2023;35:227–37. doi: 10.1016/J.JKSUCI.2023.02.017.

[22] Kim Y. Convolutional neural networks for sentence classification. EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference; 2014. doi: 10.3115/v1/d14-1181.

[23] Taneja K, Vashishtha J, Ratnoo S. Transformer based unsupervised learning approach for imbalanced text sentiment analysis of e-commerce reviews. Proc Comput Sci. 2024;235:2318–31. doi: 10.1016/J.PROCS.2024.04.220.

[24] Gupta S, Noliya A. URL-based sentiment analysis of product reviews using LSTM and GRU. Proc Comput Sci. 2024;235:1814–23. doi: 10.1016/J.PROCS.2024.04.172.

[25] Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification. Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing; 2015. doi: 10.18653/v1/d15-1167.

[26] Jiang D, Wei R, Liu H, Wen J, Tu G, Zheng L, et al. A multitask learning framework for multimodal sentiment analysis. IEEE International Conference on Data Mining Workshops, ICDMW. Vol. 2021, Dec 2021, doi: 10.1109/ICDMW53433.2021.00025.

[27] Gope JC, Tabassum T, Mabrur MM, Yu K, Arifuzzaman M. Sentiment analysis of amazon product reviews using machine learning and deep learning models. 2022 International Conference on Advancement in Electrical and Electronic Engineering. ICAEEE 2022; 2022. doi: 10.1109/ICAEEE54957.2022.9836420.