Article

Sentiment Analysis of Consumer Reviews Using Deep Learning

**Amjad Iqbal [1], Rashid Amin [1,2,*] , Javed Iqbal [1], Roobaea Alroobaea [3] , Ahmed Binmahfoudh [4]
and Mudassar Hussain [5]**

1 Department of Computer Science, University of Engineering and Technology, Taxila 47080, Pakistan
2 Department of Computer Science, University of Chakwal, Punjab 48800, Pakistan
3 Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia
4 Department of Computer Engineering, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia
5 Department of Computer Science, University of Wah, Wah Cantt 47040, Pakistan
* Correspondence: rashid4nw@gmail.com

Abstract: Internet and social media platforms such as Twitter, Facebook, and several blogs provide various types of helpful information worldwide. The increased usage of social media and e-commerce websites is constantly generating a massive volume of data about image/video, sound, text, etc. The text among these is the most significant type of unstructured data, requiring special attention from researchers to acquire meaningful information. Recently, many techniques have been proposed to obtain insights from these data. However, there are still challenges in dealing with the text of enormous size; therefore, accurate polarity detection of consumer reviews is an ongoing and exciting problem. Due to this, it is challenging to derive exact meanings from the textual data from consumer reviews, comments, tweets, posts, etc. Previously, a reasonable amount of work has been conducted to simplify the extraction of exact meanings from these data. A unique technique that includes data gathering, preprocessing, feature encoding, and classification utilizing three long short-term memory variations is presented to address sentiment analysis problems. Analysing appropriate data collection, preprocessing, and classification is crucial when interpreting such data. Different textual datasets were used in the studies to gauge the importance of the suggested models. The proposed technique of predicting sentiments shows better, or at least comparable, results with less computational complexity. The outcome of this work shows the significant importance of sentiment analysis of consumer reviews and social media content to obtain meaningful insights.

Keywords: sentiment analysis; consumer reviews; artificial intelligence; deep learning

1. Introduction

Nowadays, the world is considered a global village due to the progress of science and information technology [1]. More than 50% of the world's population uses social media. They not only use it for entertainment but they also use social media for information, marketing, other online activities, etc. Thus, we can say that this is a digital era, and we depend a lot on information technology [2]. This dependence produces an increased volume of data in tweets, posts, and customer reviews related to various products. These redundant and inconsistent data create a separate issue for the users and disturb any system's overall performance, using up massive space in memory. Due to this redundancy, the data have polarity issues. Companies must know their customers' needs, emotions, and behaviours to succeed [3]. Sentiment analysis is responsible for overcoming the related matters. It imparts a vital role in the classification of text and polarity detection.

Sentiment analysis is a technique for identifying ambiguity in language, opinions, etc. It is also known as "opinion mining". Sentiment analysis reveals how a spokesperson and a user feels about a particular topic [4]. The choice or expressive mood while writing

involves one's opinions or emotions. Many algorithms have recently been presented to analyse, anticipate, and assess sentiments from text data, such as product or customer evaluations. Sentiment analysis is a procedure that can be incredibly useful in polarity detection. Along with these difficulties, it has issues with spam and bogus data, domain dependence, negation, the overhead of natural language processing, bi-polar terms, and a vast lexicon. It is crucial to resolve the problems mentioned earlier to increase the effectiveness and efficiency of the data mining process [5].

Various scholars have already investigated sentiment analysis and its difficulties. Due to its relevance and influence on creating several cutting-edge apps, we have selected sentiment analysis of user evaluations using deep learning [6]. Sentiment analysis through classification seeks to address the mentioned issues by extracting subjective information from the given text, such as consumer reviews. Due to their popularity and successes, deep learning methods have become practical to achieve satisfactory accuracy. Throughout this study, the classification of online reviews using a deep learning method describes the overall semanticization of customer reviews by their correct category into positive and negative sentiments. The data employed in this work are a compilation of Amazon cell and accessory product reviews obtained from the Snap dataset [7].

The system helps to enhance the sentiment analysis process in web-based living after understanding the importance of sentiment analysis. The suggested method produces results that are superior, or, at the very least, identical, with the highest degree of assurance and with the least amount of computing complexity [8]. We have examined and investigated the effect of various preprocessing tasks such as data cleaning, normalizing, removal of hashtags, punctuation removal, converting the text to lowercase, and tokenization on consumer reviews. The detail of all the preprocessing steps used in this work is presented in Section 3. The major contributions of this work helped with data selection, preprocessing, and classifying. The details are as follows.

First, we examined and investigated the impact of different preprocessing activities such as data cleaning, normalization, punctuation removal, text tokenization, stop word removal, superfluous space removal, POS tagging, and emotion conversion into meaningful text. Since accurate classification and analysis rely heavily on data collecting and selection, we employed numerous benchmark datasets that other academics have already used.

Secondly, selecting a proper feature encoding method is crucial for the numeric representation of customer reviews in the classification and analysis process. This technique represents each dataset's samples as a numeric feature vector. Since the text in the reviews might be of varying sizes, the feature encoding is used for converting each review into a fixed-length vector. It has been demonstrated that using an appropriate embedding layer is very important in sentiment classification.

Thirdly, we used deep learning-based LSTM models with different layers and parameters to classify data into classes and identify their exact sentiment. When compared to previous approaches, these models produced comparable or better results. These models performed well in terms of accuracy, specificity, precision, and F1 measures.

The rest of the paper is structured as follows: A summarized relevant literature review about sentiment analysis is presented in Section 2. Section 3 outlines the proposed methodology to complete the necessary assignment of classification of consumer reviews. Section 4 shows the experiment's result of three models developed by changing network architecture and parameters. These models are named Model 1, Model 2, and Model 3, respectively. The last section presents the conclusion, contribution of this work, and key findings.

2. Related Work

Several researchers have put forth techniques to address problems with sentiment analysis and the analysis and mining of consumer evaluations. A thorough analysis of previous work is provided in this section. The authors in [9] suggested a quick, flexible, all-encompassing technique for sentiment analysis from the text that displays people's

feelings in various languages. The textbook is processed and evaluated using ConvLstm architecture and word embedding with the aid of a deep learning technique.

Consumer review data that are text-based on digital platforms have dramatically expanded. Marketing researchers have used various techniques to analyze text reviews. The authors in [10] investigated the empirical trade-off between diagnostic and predictive skills. They discovered that machine learning techniques based on neural networks provide the most precise predictions. However, topic models are poorly adapted for producing predictions, whereas neural network models are unfit for diagnostic purposes.

Sentiment analysis examines the detailed reviews left by customers for any product. To offer conclusive suggestions, the area of aspect-based sentiment analysis (ABSA) analyses and classifies the views stated on the many aspects included in these opinions. To broaden an interpretation of the Hindi text submitted, this essay examines the evolution of the ABSA model in Hindi reviews [11].

Authors in [12] used Natural Language Processing (NLP) and machine learning to identify the sentiments of the reviews in our dataset. They also employed business intelligence (BI), namely PowerBI from Microsoft, to assist enterprises who sell these goods in streamlining operations and enhancing customer happiness. The two claims above connect thanks to evaluations made by customers who have previously purchased the product. Analysing and obtaining insights from such active feedback is essential to the potential customers and companies developing the products. This paper discusses how Sentiment Analysis and Business Intelligence can benefit customers and companies. It presents various use cases for the producers and customers, an overview of how their products or services perform in the market, and customer satisfaction.

Sentiment analysis studies how individuals feel, think, and behave in response to a situation or problem. These thoughts were examined using a variety of machine learning and Natural Language Processing (NLP) based techniques. This paper's Long Short-Term Model (LSTM) prediction of the customer review's opinion has a 93.66 percent accuracy rate. Additionally, a comparison of the deep LSTM model with current models has been provided [13].

Sentiment expression and classification technologies have recently gained much popularity. In this paper [14], numerous feature extraction techniques are utilized, including the Gabor filter, the Histogram of Oriented Gradient, the Local Binary Pattern, the Discrete Cosine Transform, and many more. Most methods typically employ the entire text as their input, extracting the features and creating several subspaces that are then used to analyze various independent and dependent components [14].

In [15], the authors compared different methodologies for sentiment analysis. It was concluded from their work that there is a need for successful techniques to perform the task of classification. The work presented in [16] shows that semantic sentence analysis can improve the methods, precision, and consistency. The key finding of this work was that consumer reviews work as a medium in which the users of consumer reviews can share their feeling and thoughts, etc., on various forums and social media platforms [17].

The authors have proposed two methods in their research for integrating vectors and function subsets [18]. The standard integration of various function vectors (OIFVs) was suggested to achieve a novel function vector. A frequency-based ensemble method was proposed in this study. Four well-known Text Classification algorithms have also been used to classify feature subsets in the wrapper method. These findings showed that categorizing speech patterns was beneficial in ranking accuracy relative to unigram-based accommodating. In their research, the authors have used different datasets such as those for movies, books, and music for sentiment analysis. The word2vec technique is used for representation [19]. The process results showed that part of speech (POS)-based features are efficient.

The deep mastering learning methods used for sentiment analysis have emerged as common approaches [20]. Deep learning is a technique that learns through a few different layers with state-of-the-art statistics and prediction results [21]. In their analysis,

the authors concluded that the GoogleNet performed higher than baseline, which was used for analyzing the performance [22]. According to this, the topic nature, Negation, and Domain dependency are the limitations of Deep Learning Sentiment Analysis. The authors addressed recent advancements in recurrent neural networks for broader linguistic development [23]. Additionally, they used a neural network to oversee the issues and difficulties. The work findings showed that vocabulary sizes and long-term language structure were two critical issues in this research.

The authors suggested deep learning sentiment analysis methods to categorize reviews of Twitter data [24]. Significant findings show that deep learning performs better than traditional practices, including Naive Bayes and SVM without Maximum Entropy. The authors have applied LSTM and DCNN models in their study. To train word vectors for the DCNN and LSTM models, they used word2vec [25]. The Twitter dataset was used in this study. This paper showed that DNN is a better technique than LSTM for conducting sentiment analysis using deep learning [26]. Moreover, a large, approach significant data sample is critical for mining.

The authors [27] then presented the ConvLstm architecture after finishing their study. This design displays words above vectors based on Long Short-Term Memory (LSTM) [27]. According to their research, the merging layer of CNN may be replaced with a convolution-neural network to regulate long-term memory better, minimize the reliance on trustworthy local input, and manage long-term corpus dependency. This study thoroughly evaluated several lexical semantics tasks across various parameter settings [28]. They claim that content prediction is a novel and intriguing area of research.

For the sentiment analysis of tweets, the authors used Recursive Neural Networks (RNN) [29]. As a form of communication heavily reliant on symbols and brief brevity, tweets presented unique difficulties for sentiment analysis [30]. They experimented with neural network architectures such as RNN, Recursive Neural Tensor (RNTN), and hidden layer RNN. They examined the users' feelings, attitudes, and views during their analysis [31]. They also tried to create a vocabulary of terms used in customer feedback. This research aims to demonstrate that consumer review data are interpreted effectively [32]. Twitter, a popular consumer reviews website, is valuable for facts mining due to its incidence and reputation among well-known folks.

In this paper, the authors [33] presented a sentiment evaluation device that accommodates the following two capabilities: sentiment analysis amongst Twitter tweets, locating tremendous, pessimistic, and impartial tweets from records sources [33]. This work focuses on reading tweets from those consumer reviews. This survey analyzed and compared lexicon and deep learning-based approaches for opinion mining [34]. This survey uses several deep learning algorithms such as NB, SVM, and ME. The authors provided their experiments on the Twitter dataset. According to their findings, the accuracy rate of the NB is 75%, and the accuracy rate of the SVM is 77.73%. When compared, the SVM produced better outcomes. Finally, the literature review shows that sentiment analysis and assessment of social media content still present many challenges affecting their classification accuracy and performance.

3. Research Methodology

This section describes in depth the methodology of the proposed work. This section is further divided into various subsections. Section 3.1 provides the programming environments utilized to implement the proposed methodology. Section 3.2 presents details about the data used in experiments and preparation mechanisms. The architecture and experimental details of deep learning classifiers are explained in the last section.

3.1. Programming Environments

Python is a programming language primarily used in deep learning and science-related projects. It provides an extensive collection of libraries that can be used to implement various deep learning algorithms. Due to its flexibility, the abundance of open-source

libraries, and ease of use, we have used Python 3.5 in this study. The following libraries were helpful in our work:

TensorFlow is a large-scale machine learning library for numerical computation. Seaborn is a matplotlib-based Python data visualization library. BeautifulSoup is used for pulling data out of HTML XML. Keras is a Python-listed library of open-source neural networks. It can work on Microsoft Cognitive Toolkit, R, Theano, or PlaidML. Figure 1 shows the block diagram of sentiment analysis.
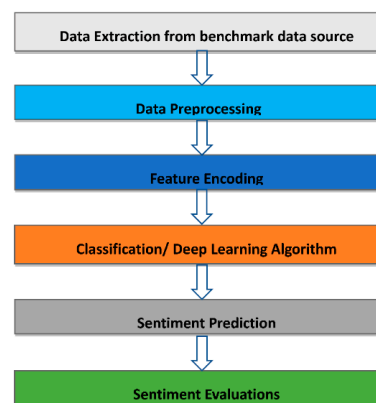
Figure 1. Overview of the sentiment analysis system block diagram.

3.2. Deep Learning-Based Classification/Learning Algorithms

During this phase, the recurrent neural network-based LSTM classifiers are used for the task of classification. Three different types of models were developed by changing network architecture and parameters. These models are considered Model 1, Model 2, and Model 3, respectively. In our studies, a corpus of 1,194,704 reviews was considered for the training dataset, and the remaining 512,016 reviews were used to assess the performance of selected classifiers. Training and testing portions of the data were divided into two separate groups. The classification algorithms were trained and evaluated twice for each analysis, once on the training and once on the testing section reviews. In the subsequent step, the review text was encoded into a numerical feature vector before being taken care of by any classification algorithm. This was conducted by utilizing the word embedding vector model. The third step was to train LSTM-based classifiers. In the last step, we applied the trained classifiers to the test dataset to evaluate their classification performance against the predicted and actual labels that were not seen before the classification algorithms. The general details of the entire sentiment classification/prediction process are shown in Figure 2.
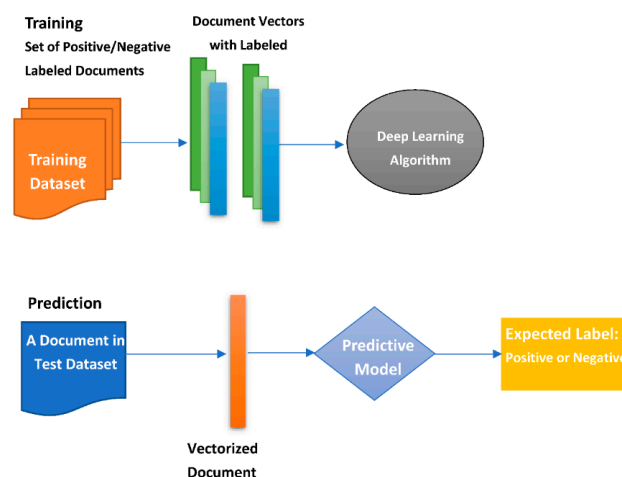
Figure 2. Example of the sentiment classification by supervised deep learning algorithms.

3.3. Extraction of Benchmark Data

The initial step in the process is to obtain review data from the source of benchmark data. The information is taken from postings, comments, reviews, and tweets. Before extracting data from the needed media, search parameters were established for themes and customer evaluations. Twitter tweets, movie reviews, news feeds, product reviews, and Facebook postings are some of the sources of frequently utilized datasets. The extracted data are fed into the system at this step, which is used for data mining and analysis. This stage serves as the central component of the sentiment analysis process. The following datasets of customer reviews have been selected for this purpose and are displayed in Table 1.

Table 1. Experimental datasets applied in the proposed study.

Sr#	Datasets	Positive Instances	Negative Instances	Total Instances
1	Amazon-Fine-Food-Reviews	21,784	3216	25,000
2	Cell Phones and Accessories	88,516	11,484	100,000
3	Amazon-Products	13,251	749	14,000
Yi-Jen Mon	Yi-Jen Mon	25,000	25,000	50,000
5	Yelp	3266	734	4000

3.4. Data Preprocessing

Data preprocessing is a crucial phase in text data analysis [35]. Due to the repetitions and redundancies in tweets, blogs, reviews, and other types of text, text data become more complicated. Data normalization uses data preprocessing as a filtering method. The normalization, word tokenization [36], removing stop words, removing extra spaces, padding, changing the text data to lowercase, and removing hashtagging are examples of data preprocessing, etc. This work implemented various tasks to achieve the data in the desired format.

3.4.1. Erase Punctuation

Punctuation accounts for almost 40 to 50 percent of the text in a written document. Punctuation has no bearing on the outcome of any sentiment analysis model. It is critical to remove these punctuations since they have no bearing on the sentiment analysis.

In this stage, we removed all punctuation from the text and presented the data in their normalized form. The resultant text is simplified and summarised. All punctuation was deleted from the document. This example demonstrates the need to remove punctuation from the text: "Good day, everyone!!!!! I've been with IDS since 2012.", can be transformed into "Hello everyone I've been with IDS since 2012".

3.4.2. Convert the Text Data to Lowercase

In the customer reviews, consumers enter material without following grammar norms, in that the entered text contains both lower and upper case characters. Many of the methods utilized in the study are case-sensitive. As a result, the classifier has difficulty determining the polarity of the provided text. Such an issue may be avoided simply by changing the entire text to a standard format. Conversely, if we wish to perform the same process manually, the lower (txt) statement is used. It transforms all upper case text to lower case while leaving the other characters untouched. The following example illustrates how to convert text data to lower case: "I Am A Senior Big Data Analyst in Islamabad" may be translated to lowercase as "I am a senior big data analyst in Islamabad".

3.4.3. Tokenization of the Text

Tokenization is a technique for dividing text streams into phrases or tiny chunks of textual material. Tokens are fragmented pieces of text This technique aims to make complex textual contents straightforward to solve. As tokens are used, the data mining

process becomes more straightforward. Tokenization is vital in the lexical evaluation and beneficial in semantics and sentiment analysis. Tokenization is an important step in the whole NLP pipeline. We cannot start creating the model without first cleaning up the text. Tokenization is further subdivided into two categories: words tokenize and phrase tokenize. This tokenized form may be used to:

- Count the number of words in the text;
- Count the frequency of words. The text statistics are split into words in this stage. A large and complex record is broken down into little packets of words or symbols. The above text, for example, may be transformed into tokens such as "I am a data analyst in Islamabad", which contains seven tokens.

3.4.4. Removal of Stop Words

Many words in text files appear repeatedly. As a result, it is critical to delete the stop words. Indeed, stop words never provide significance to the written substance. These kinds of words often appear in large numbers in the text. Due to this, the text mining process has become difficult, and the classifiers have produced unexpected results. The stop words are deleted from the selected data in this stage. This strategy minimizes textual content facts while improving overall system efficiency. For example, after deleting the stop words, the preceding statement may read as "I data analyst Islamabad".

3.4.5. Removal of the Hyperlink

In any dataset, hyperlinks no longer have meaning. The linkages are solely functionally helpful. However, with the acquired data, we solely utilise tweets, comments, and reviews to represent thoughts and feelings to fine-tune the text's polarity. As a result, it is critical to delete the linkages from the datasets.

3.4.6. Removal of Hash Tag

Hashtagging is also popular these days. In consumer reviews, hashtagging is often used. Hashtags take up a significant amount of memory. Hashtags are ineffective when it comes to sentiment analysis. These just add to the uncertainty for the classifiers. As a result, it is critical to remove hashtags. The hashtags are deleted from the datasets, making the training data clearer and more succinct.

3.4.7. Removal of Unnecessary Spaces

There is extra space in raw datasets, which causes an issue for the classifier during sentiment analysis. To avoid this problem during the preprocessing step, all unneeded spaces are deleted so that they do not impair the performance of the classifiers and save a lot of time during sentiment analysis.

3.4.8. Padding

There are both extremely short and very long reviews in the consumer review databases, which causes problems for the classifier during sentiment analysis. The quantity of pixels added to the review when evaluated by the network is called CNN-related padding. Padding is merely an addition of layers of zero to our input review at the end to ensure that each consumer review has the same length.

3.4.9. POS Tagging

POS is a strategy for categorizing words in training data with a specified grammatical form. This group uses the context of the words. POS labelling is not an easy task to complete. POS labelling is not a solution to the extreme discovery issue in opinion examination, but it does aid notably in rearranging many concerns. A product review gathered several aspects and views in this process. The modified POS tagger is used to specify a particular functionality. The grammar relations' POS tagging tool is offered in the review. The tags noun (N), proper noun (P), verb (V), article (DET), and adjective (ADJ) should be used to

establish the speech part in the examination. Furthermore, substantive and appropriate substantive were recognized as candidate concerns by POS tags. This POS process may be described as using a Concealed Markov Model (HMM), in which tags are hidden and the Observable Output is produced. When POS is tagging, we always aim to identify a tag sequence (C) that optimizes mathematically as:

$$P\,(C\,|\,W)$$

where C denotes C1, C2, C3, . . . , CT, and W denotes W1, W2, W3, and WT.

3.5. Feature Encoding for Numerical Representation of Textual Data

The obtained datasets might not be in a format suitable for statistical or mathematical calculations. A proper function encoding method is needed to extract numerical characteristics from available text data [37]. We must propose a mathematical model which correctly depicts each review in the sample and captures the accurate or true semanticist word or sentence therein. During the next step of the processing and analysis approach, the proposed numerical features are then used.

Word Embedding

Every word is represented numerically and in vector form through word embedding. Word embedding refers to texts with exact representations of words with the same meaning. In particular, word embedding is unsupervised learning of word representation, which is relatively similar to semantic similarity. This refers to words in a coordinated scheme in which similar terms are put closer together, based on a set of relationships [38].

3.6. Deep Learning-Based Classification/Learning Algorithms

Presently, a massive amount of personal data appear in consumer reviews; classification is becoming popular in sentiment analysis and evaluation [39]. During this phase, the recurrent neural network-based LSTM and Deep LSTM classifiers were used for the task of classification. The LSTM network consists of LSTM units next to the input and output layer. An LSTM framework allows long and short recording values and uses no device in several parts of its action [40]. A three-layer LSTM stack has been developed to build a deep RNN [25]. Moreover, peephole connections in a similar cell between its internal partitions and the entrances can also be used for the cutting-edge LSTM Design to evaluate precise performance [41].

Deep LSTM RNNs (DNN) have been commonly used for the more resounding speech recognition architecture [42]. Using Deep LSTM RNN in the regular LSTM, parameters can be best optimized by spreading them across many layers [43]. This study uses a Deep LSTM model with one input layer, two LSTM layers in a row, two dense layers, and either one output layer or two dense layers. Three models were developed by changing the LSTM network architecture and parameters. These models were considered Model 1, Model 2, and Model 3, respectively. The following subsections provide a summary of each model.

3.6.1. The Architecture of Model 1

One LSTM layer, an embedding layer with vocabulary size, embedding vector length, and maximum review length, and a dense layer with a fully coupled sigmoid activation function make up the architecture of Model 1. A binary cross-entropy loss is used in the model's construction and training according to the nature of our challenge. A better optimization tool is Adam (faster and more reliably reaching a global minimum when minimizing the cost function in training neural nets). The network design of Model 1 is shown in Figure 3.
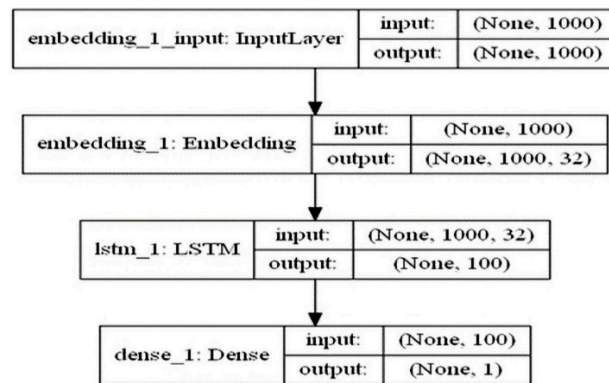
Figure 3. Sample network layout of Model 1.

3.6.2. The Architecture of Model 2

Model 2's architecture consists of two dense hidden layers with an ReLU activation function and one LSTM layer with 0.5 intermediate dropouts. The model contains a thick layer with a sigmoid activation function and an embedding layer with parameters for vocabulary size, embedding vector length, and maximum review length. Total parameters and trainable parameters are 234,449. The Adam Optimizer was used to create and train the model for binary cross-entropy loss. We also evaluated the accuracy, which helps us assess model output more precisely. Figure 4 shows the network layout of Model 2.
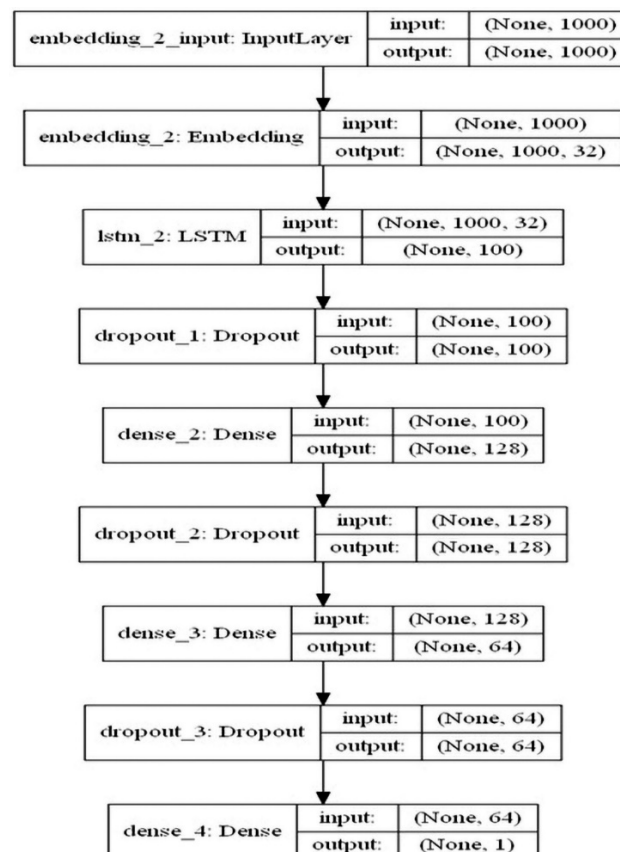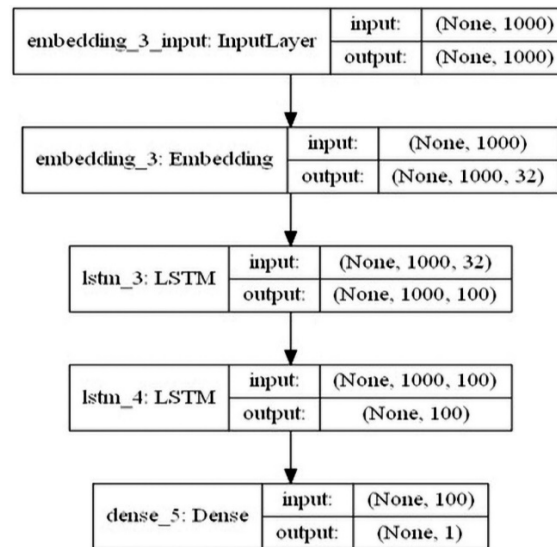

Figure 4. Sample network layout of Model 2.

3.6.3. The Architecture of Model 3

Two LSTM layers make up Model 3's architecture. A deep LSTM network or stacked LSTM network is another name for it. This model combines a dense layer with a sigmoid activation function and an embedding layer added with vocabulary size, embedding vector

length, and maximum review length. The binary entropy loss was used to build the model, and an Adam Optimizer was used to train it. To evaluate model output more accurately, we additionally observed and evaluated accuracy. Figure 5 depicts Model 3's network architecture.

Figure 5. Sample network layout of Model 3.

3.7. Sentiment Prediction

Predicting sentiments from the supplied input data is helpful in this process stage [44]. Several process cycles may be necessary for the algorithms to become more generic. The sentiment prediction's findings and the sentiment's outcomes are connected. It boosts the sentiment analysis's productivity.

3.8. Sentiment Evaluations

We can define the polarity of the texts after all the above-described stages as analysts. The analytical results are listed in this step. The formation of the text's polarity occurs at this point. The words can be positive or negative. This is often called opinion mining and tracks the sender's attitude. The findings of the proposed approach are analyses against the current best literature approaches. The system's overall efficiency is assessed using common factors or parameters. Each performance measure is defined in the following way.

3.8.1. Accuracy

The most well-known performance metric is accuracy. It is convenient and straightforward to compute and identify. Accuracy assesses a predictor's capacity to correctly identify all samples, regardless of their effectiveness or unfavourability [45].

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

where TN = True negative, FN = False negative, FP = False positive, TP = True positive, P = Positive, and N = Negative.

3.8.2. Sensitivity/Recall

The true positive rate or recall can be used to define sensitivity. The true positive percentage may be quickly identified by following a few simple methods [46]. Less false

negatives result from higher sensitivity, whereas more false negatives result from lower sensitivity. Sometimes, as sensitivity increases, accuracy falls off.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{P}} \tag{2}$$

3.8.3. Precision

The precision demonstrates the accuracy of the classifier. Low accuracy and high precision both result in lower accuracy and fewer false positives. The improvement in precision is the cause of reduced sensitivity and is inversely proportional to the sensitivity.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

3.8.4. F1-Measure

F1-Measure is the blend of accuracy and sensitivity. This is the weighted harmonic way to be sensitive and accuracy/precision. The F1 measurement is proven as effective as precision.

$$\text{F1} - \text{Score} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

4. Results and Discussion

The discussion of the experimental findings is presented in this section. The chosen datasets are assessed using several deep learning methods, including various models based on the LSTM classifier for sentiment classification and assessment. In this study, deep learning-inspired long short-term memory and recurrent neural network-based models were created for the precise and trustworthy classification and analysis of sentiment. Three different models were created based on these deep learning approaches. The datasets above assessed these models' performance using several performance indicators, such as accuracy, precision, recall, and F1-score. Additionally, the outcomes of our tests were contrasted with those of earlier methods. The outcomes were superior to or on par with those of earlier methods. When the Amazon-Fine-Food-Review dataset is evaluated using several classification methods, the results are shown in Figure 6 (i.e., Model 1, Model 2, and Model 3). Figure 7 depicts the evaluation of the Cell Phones and Accessories dataset using Models 1, 2, and 3. The assessments of performance metrics when these classifiers were applied to the Amazon-Products, IMDB, and Yelp datasets are shown in Figures 8–10.
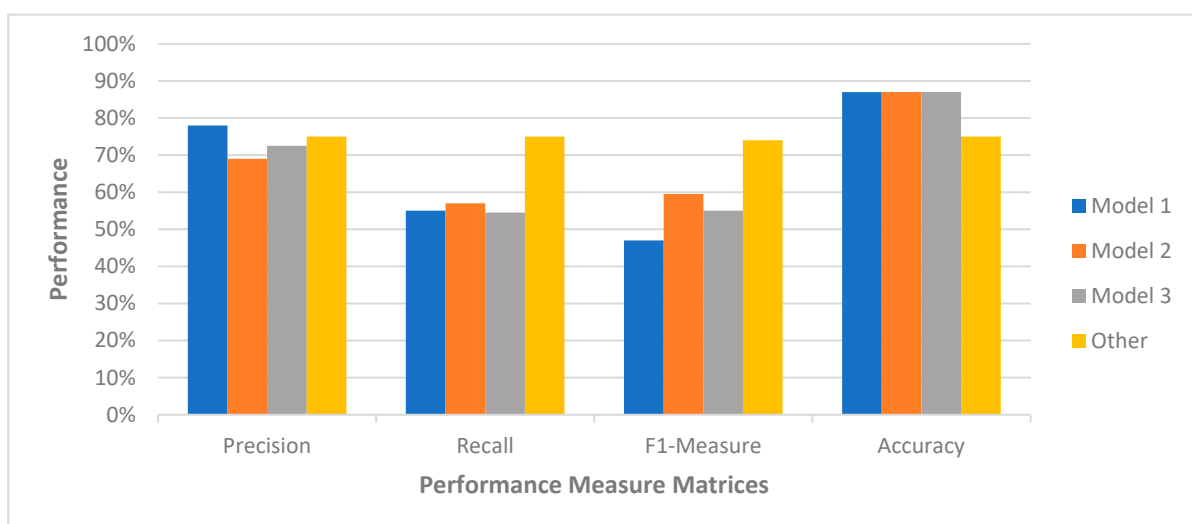
Figure 6. Comparison of performance measure matrices on the Amazon Fine-Food-Reviews dataset.
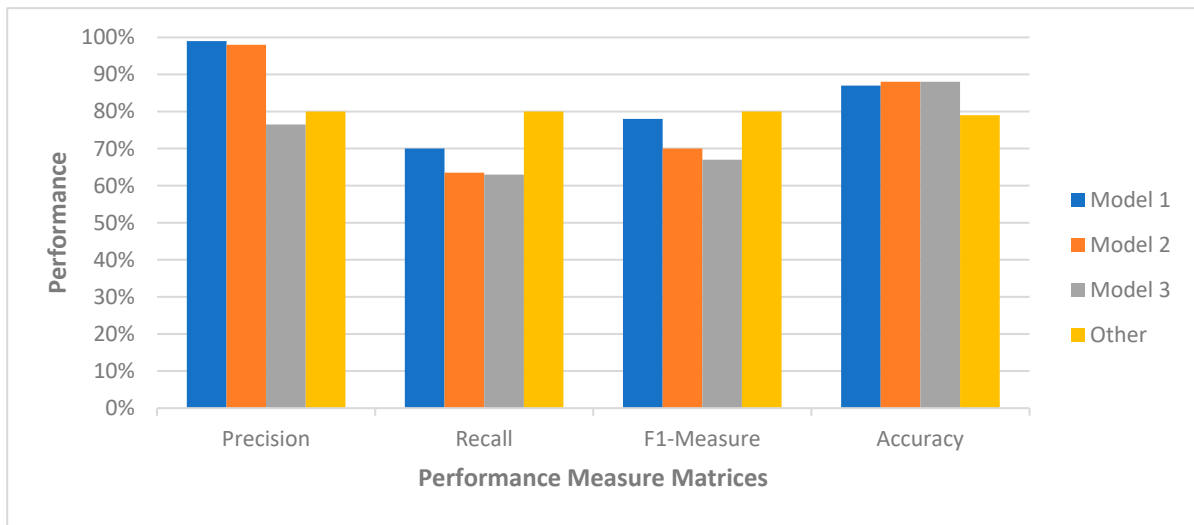
Figure 7. Comparison of performance measure matrices on the Cell Phones and Accessories dataset.
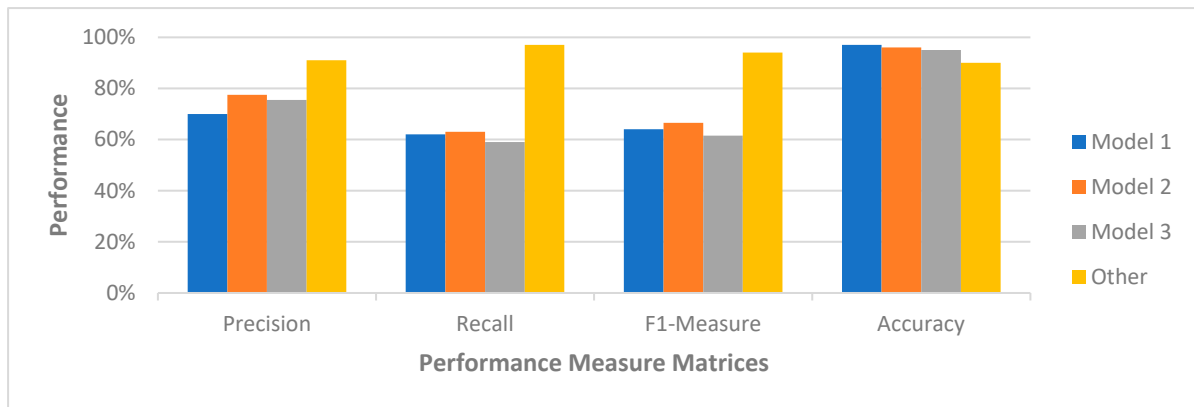

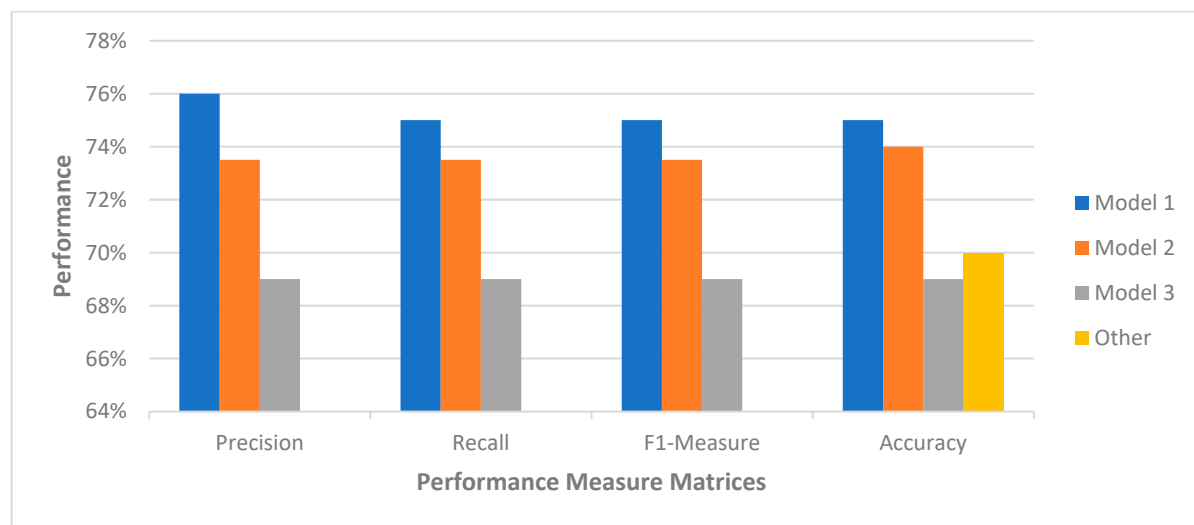**Figure 8.** Comparison of performance measure matrices on the Amazon-Products dataset.

Figure 9. Comparison of performance measure matrices on the Movies dataset.
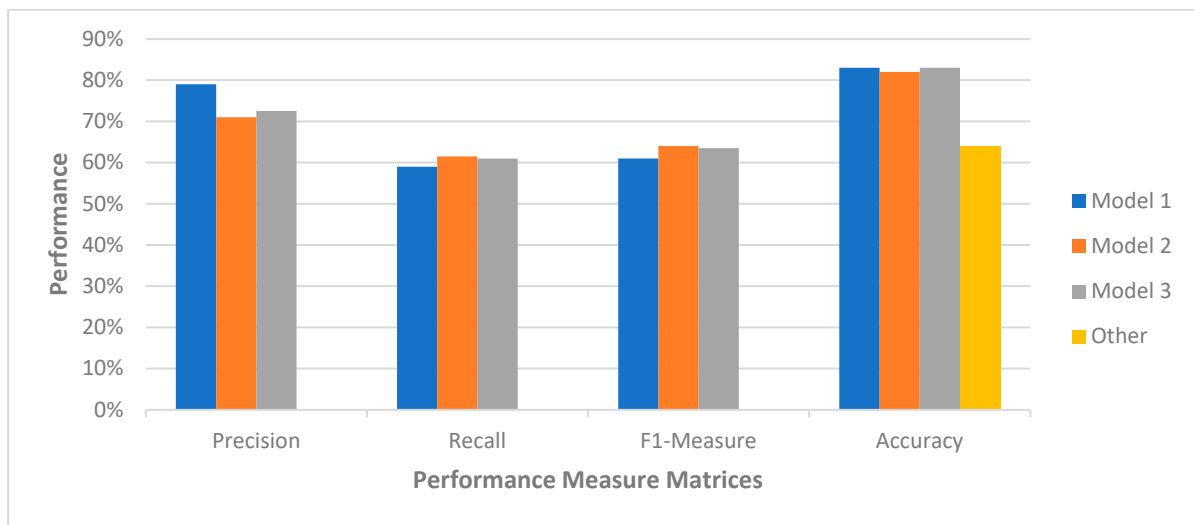
Figure 10. Comparison of performance measure matrices on the Yelp dataset.

4.1. Comparison of Classification Results with LSTM and Deep LSTM

Figure 11 compares the outcomes of Model 1, Model 2, and Model 3 classification methods. The combined experimental results for the three classifiers are shown in Table 2. The experimental findings demonstrate that the performance of the chosen classifiers is superior in every way. Each dataset contains reviews for the binary classifications Positive and Negative. The analysis of all these levels is followed by the display of the total average output. Figure 8 demonstrates that the accuracy for Models 1, 2, and 3 is 97%, 96%, and 95%, respectively. The accuracy of the three classifiers in Figure 7 is 99%, 98%, and 77%, respectively. Figure 9 shows that the Model 1, Model 2, and Model 3 recall rates are 75%, 74%, and 69%, respectively. In Figure 7, the three classifiers' respective F1 Measures are 78%, 70%, and 67%, respectively. As a result, we deduced that Model 1 outperformed other models regarding prediction rate. We can see that while models 2 and 3 are nearly identical to Model 1, they perform a little worse. Model 1 performed 87%, 87%, 97%, 75%, and 83% on the chosen data, whereas Model 2 performed 87%, 88%, 96%, 74%, and 82%, which is a little less well than Model 1. The findings for Model 1 are better if the accuracy levels for each classification are mentioned. Table 3 and Figure 12 evaluate the proposed and current methodologies. After analysing the selected data using the methods above, the results are summarized in Figure 11.
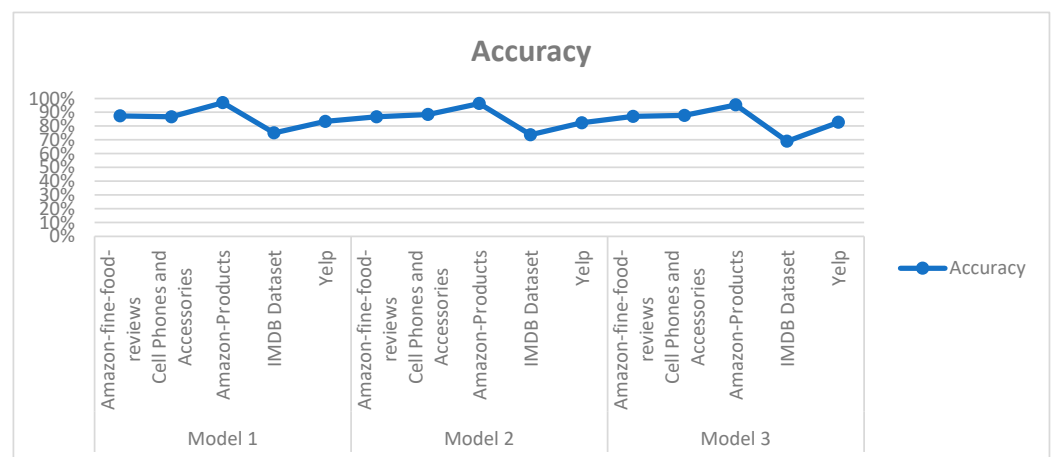

Figure 11. Comparison of accuracy results with different techniques.

Table 2. Comparison of sentiment classification results through different classifiers.

| Classifier	Dataset	Accuracy
Model 1	Amazon-Fine-Food-Reviews	0.87
	Cell Phones and Accessories	0.87
	Amazon-Products	0.97
	IMDB	0.75
	Yelp	0.83
Model 2	Amazon-Fine-Food-Reviews	0.87
	Cell Phones and Accessories	0.88
	Amazon-Products	0.96
	IMDB	0.74
	Yelp	0.82
Model 3	Amazon-Fine-Food-Reviews	0.87
	Cell Phones and Accessories	0.88
	Amazon-Products	0.95
	IMDB	0.69
	Yelp	0.83

Table 3. Comparison of performance measure matrices with previous work.

Classification Models	Datasets	Precision	Recall	F1-Measure	Accuracy
Model 1	Amazon-Fine-Food-Reviews	0.78	0.55	0.47	0.87
	Cell Phones and Accessories	0.99	0.70	0.78	0.87
	Amazon-Products	0.70	0.62	0.64	0.97
	IMDB	0.76	0.75	0.75	0.75
	Yelp	0.79	0.59	0.61	0.83
Model 2	Amazon-Fine-Food-Reviews	0.69	0.57	0.60	0.87
	Cell Phones and Accessories	0.98	0.64	0.70	0.88
	Amazon-Products	0.78	0.63	0.67	0.96
	IMDB	0.74	0.74	0.74	0.74
	Yelp	0.71	0.62	0.64	0.82
Model 3	Amazon-Fine-Food-Reviews	0.73	0.55	0.55	0.87
	Cell Phones and Accessories	0.77	0.63	0.67	0.88
	Amazon-Products	0.76	0.59	0.62	0.95
	IMDB	0.69	0.69	0.69	0.69
	Yelp	0.73	0.61	0.64	0.83
J. Wu and T. Ji (2016) [47]	Amazon-Fine-Food-Reviews	0.75	0.75	0.74	0.75
S. A. Aljuhani and N. S. Alghamdi [48]	Cell Phones and Accessories	0.80	0.80	0.80	0.79
H. Nguyen, A. Veluchamy, M. Diop, and R. Iqbal (2018) [49]	Amazon-Products	0.91	0.97	0.94	0.90
J. Hong and M. Fang (2015) [50]	IMDB	-	-	-	0.89
N. Asghar (2016) [51]	Yelp	-	-	-	0.64

Figure 6 shows the results of the experimental evaluation. The *y*-axis shows the evaluated performance values, and the *x*-axis shows the Performance Measure Matrices for the Amazon-Fine-Food-Reviews dataset.

Model 1 shows that all measures of Accuracy, Precision, Recall, and F-Measure are 87%, 78%, 55%, and 47%, respectively. Model 2 represents that the performance of all Accuracy, Precision, Recall, and F-Measure measures is 87%, 69%, 57%, and 60%, respectively. Model 3 indicates that all measures of Accuracy, Precision, Recall, and F-Measure are 87%, 73%, 55%, and 55%, respectively. We have also compared and evaluated the previous performance with some other classifiers. This is also low compared to the LSTM, with RNNMS (Recursive Neural Network for Multiple Sentences) performances. It gives 75%, 75%, 75%, 74%

performance for Accuracy, Precision, Recall, and F-Measure, respectively [47]. These results provide the best value as compared with other classifiers.
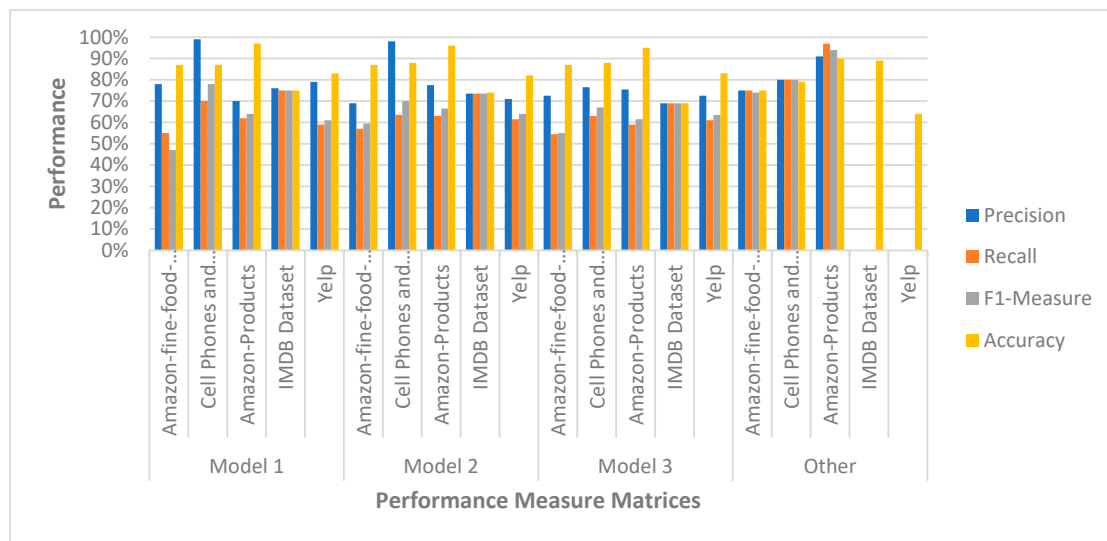
Figure 12. Comparison of performance measure matrices with previous work.

Figure 7 represents the results of the experimental evaluation. The *y*-axis shows the evaluated performance values, and on the *x*-axis, Performance Measure Matrices are placed on the Cell Phones and Accessories dataset. Model 1 shows that all Accuracy, Precision, Recall, and F-Measure measures are 87%, 99%, 70%, and 78%, respectively. Model 2 shows that all Accuracy, Precision, Recall, and F-Measure measures are 88, 98%, 64%, and 70%, respectively. Model 3 indicates that all Accuracy, Precision, Recall, and F-Measure measures are 88%, 77%, 63%, and 67%, respectively.

We have also compared and evaluated the previous performance with some other classifiers. This is also low as compared to the LSTM while compared with CNN. It gives 79%, 80%, 80%, 80% performance for Accuracy, Precision, Recall and F-Measure, respectively [48]. These results are the best value as we compared with other classifiers.

Figure 8 represents the results of the experimental evaluation. The *y*-axis shows the evaluated performance values, and on the *x*-axis, Performance Measure Matrices are placed on Amazon-Products dataset. Model 1 shows that the performance of all measures Accuracy, Precision, Recall, and F-Measure on positive reviews is 97%, 70%, 62%, and 64%, respectively. Model 2 displays that the performance of all measures Accuracy, Precision, Recall, and F-Measure on positive reviews is 96%, 78%, 63%, and 67%, respectively. Model 3 indicates that the performance of all measures, Accuracy, Precision, Recall, and F-Measure, on positive reviews is 95%, 76%, 59%, and 62%, respectively.

We have also compared and evaluated the previous performance with some other classifiers. This is also low compared to the LSTM and Logistic Regression performance. It gives 90%, 91%, 97%, and 94% performance for Accuracy, Precision, Recall, and F-Measure, respectively [49]. These results are the best value as we compared with other classifiers.

Figure 9 represents the results of the experimental evaluation. The *y*-axis shows the evaluated performance values, and Performance Measure and Matrices are planned on the *x*-axis for the Movies dataset.

Model 1 shows that all measures of Accuracy, Precision, Recall, and F-Measure are 75%, 76%, 75%, and 75%, respectively. Model 2 displays that all Accuracy, Precision, Recall, and F-Measure performance measures are 74%, 74%, 74%, and 74%, respectively. Model 3 indicates that the performance of all measures, Accuracy, Precision, Recall, and F-Measure, is 69%, 69%, 69%, and 69%, respectively. We have also compared and evaluated the previous performance with some other classifiers. It gives 70% performance for accuracy. These results are the best value as compared with other classifiers.

Figure 10 represents the results of the experimental evaluation. The *y*-axis shows the evaluated performance values, and on the *x*-axis, Performance Measure Matrices are placed on the Yelp dataset.

Model 1 shows that the performance of all measures, Accuracy, Precision, Recall, and F-Measure, is 83%, 79%, 59%, and 61%, respectively. Model 2 displays that the performance of all measures of Accuracy, Precision, Recall, and F-Measure is 82%, 71%, 62%, and 64%, respectively. Model 3 indicates that the performance of all measures, Accuracy, Precision, Recall, and F-Measure, is 83%, 73%, 61%, and 64%, respectively. We have also compared and evaluated the previous performance with some other classifiers. It gives 64% performance for accuracy. These results are the best value as compared with other classifiers.

Figure 11 compares the outcomes of Model 1, Model 2, and Model 3 classification methods. It shows that the accuracy of Model 1, Model 2, and Model 3 on Amazon-Fine-Food-Reviews is 87%, respectively, on Cell Phones and Accessories it is 87%, 88%, 87%, respectively, and on Amazon-Products, the accuracy of all models is 97%, 96%, and 95%, respectively. On the IMDB dataset, the result of all models in the form of accuracy is 75%, 74%, and 69%, respectively. In the Yelp dataset, the accuracy of all our models is 83%, 82%, and 83%.

The results of the Model 1, Model 2, and Model 3 categorization techniques are contrasted with earlier research in Table 3. The accuracy of models 1, 2, and 3 on Amazon-Fine-Food-Reviews was 87%, when it was just 75% in an earlier study. The accuracy was 87%, 88%, and 87% for accessories and cell phones, respectively. It was 79%, much like in earlier work. Thus, all models on Amazon-Products have accuracy rates of 97%, 96%, and 95%, respectively. The accuracy results for all models on the IMDB dataset are 75%, 74%, and 69%, respectively. Therefore, it was 89% in earlier work. Our models' accuracy in the Yelp dataset is 83%, 82%, and 83%, compared to 64% in the prior study.

Figure 12 compares the results of the Model 1, Model 2, and Model 3 categorization processes to those of earlier investigations. The accuracy of models 1, 2, and 3 on Amazon-Fine-Food-Reviews was 87% compared to past studies and 87%, 88%, and 87% for cell phones, accessories, and other items. It was 79 percent, the same as before. This leads to 97%, 96%, and 95% accuracy for each model on Amazon-Products, respectively. According to the IMDB dataset, the accuracy rates for each model are 75%, 74%, and 69%, respectively. Therefore, 89% was the figure from earlier studies. Our models' accuracy in the Yelp dataset is 83%, 82%, and 83%, up from 64% in the previous research.

4.2. Major Findings from the Experimental Results

Within various datasets, different models produce varying accuracy outcomes. As mentioned above, the complexity and magnitude of the datasets are the primary causes of this difference. Additionally, many factors also impact how well the trial classifier functions. The factors listed below might have an impact on the classifiers' accuracy. Pre-processing is crucial and essential to the examination of certain types of data. It might be challenging for the classifier to produce correct findings on the supplied data if the preprocessing step is not carried out appropriately. Noise frequently reduces the classifier's performance. Due to the noise, classifiers struggle and deliver subpar results. Appropriate encoding of the chosen dataset is also very critical for the performance of the system. Various techniques in literature are present, and this work uses the best among them.

Over-fitting and under-fitting of the models may significantly affect the performance of the classification models. To avoid this, the dataset should be balanced. The size of a dataset should not be so small or huge. The design and characteristics of the chosen model might have an impact on how well it performs. Any model must be adequately trained, which is also essential. It is impossible to assess classification accuracy using a single experiment. Assuring the effectiveness of classifier cross-validation is an excellent activity. In cross-validation, many tests are carried out, and the overall average accuracy is used as the final, legitimate accuracy.

5. Conclusions

Today, every business depends appreciably on consumer feedback gathered from various social media platforms such as Amazon, Facebook, Twitter, etc. The primary emphasis of this work was on the sentiment analysis of products and customers' reviews on social media and products' online webpages. This work employed five different categories of study datasets from benchmark data sources. These well-known datasets included the IMDB, Yelp, Cell Phones and Accessories, Amazon-Products, and Amazon-Fine-Food-Reviews datasets. It has also been found from experiments that selecting suitable feature encoding methods plays a critical role in the numeric representation of consumer reviews in the process of classification and analysis. It has been shown that using this embedding layer is crucial for classifying sentiments. This study implemented deep learning-inspired long short-term memory and recurrent neural network-based models for sentiment classification and analysis. Three models were proposed based on these deep learning approaches with architecture and parameter tuning. On the datasets mentioned above, the effectiveness of these models was assessed using several performance measures, including accuracy, precision, recall, and F1-score. The final results were compared with results from existing techniques. The proposed results were better than or comparable to previous approaches. In the future, this work can involve other deep learning architectures, such as transformers.

References

1. Levy, P.; Bononno, R. *Becoming Virtual: Reality in the Digital Age*; Da Capo Press: Boston, MA, USA, 1998.
2. Khan, S.; Muhammad, K.; Mumtaz, S.; Baik, S.W.; de Albuquerque, V.H.C. Energy-efficient deep CNN for smoke detection in foggy IoT environment. *IEEE Internet Things J.* **2019**, *6*, 9237–9245. [CrossRef]
3. Ajmal, A.; Aldabbas, H.; Amin, R.; Ibrar, S.; Alouffi, B.; Gheisari, M.J.C.I. Stress-Relieving Video Game and Its Effects: A POMS Case Study. *Comput. Intell. Neurosci.* **2022**, *2022*, 4239536. [CrossRef] [PubMed]
4. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
5. Akhtar, M.J.; Ahmad, Z.; Amin, R.; Almotiri, S.H.; Al Ghamdi, M.A.; Aldabbas, H.J.C. An efficient mechanism for product data extraction from e-commerce websites. *Comput. Materi. Contin.* **2020**, *65*, 2639–2663. [CrossRef]
6. Iqbal, M.S.; Ahmad, I.; Bin, L.; Khan, S.; Rodrigues, J.J. Deep learning recognition of diseased and normal cell representation. *Trans. Emerg. Telecommun. Technol.* **2020**, *32*, e4017. [CrossRef]
7. Leskovec, J.; Sosic, R. SNAP: A General Purpose Network Analysis and Graph Mining Library. *ACM Trans. Intell. Syst. Technol.* **2016**, *8*, 1–20. [CrossRef]
8. Aleem, S.; Huda, N.u.; Amin, R.; Khalid, S.; Alshamrani, S.S.; Alshehri, A.J.E. Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions. *Electronics* **2022**, *11*, 1111. [CrossRef]
9. Giatsoglou, M.; Vozalis, M.G.; Diamantaras, K.; Vakali, A.; Sarigiannidis, G.; Chatzisavvas, K.C. Sentiment analysis leveraging emotions and word embeddings. *Expert Syst. Appl.* **2016**, *69*, 214–224. [CrossRef]
10. Alantari, H.J.; Currim, I.S.; Deng, Y.; Singh, S. An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. *Int. J. Res. Mark.* **2021**, *39*, 1–19. [CrossRef]
11. Yadav, V.; Verma, P.; Katiyar, V. E-commerce product reviews using aspect based Hindi sentiment analysis. In Proceedings of the 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 27–29 January 2021; pp. 1–8.

12. Desai, Z.; Anklesaria, K.; Balasubramaniam, H. Business Intelligence Visualization Using Deep Learning Based Sentiment Analysis on Amazon Review Data. In Proceedings of the 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 6–8 July 2021; pp. 1–7.

13. Mohbey, K.K. Sentiment analysis for product rating using a deep learning approach. In Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 25–27 March 2021; pp. 121–126.

14. Darokar, M.S.; Raut, A.D.; Thakre, V.M. Methodological Review of Emotion Recognition for Social Media: A Sentiment Analysis Approach. In Proceedings of the 2021 International Conference on Computing, Communication and Green Engineering (CCGE), Pune, India, 23–25 September 2021; pp. 1–5.

15. Devika, M.; Sunitha, C.; Ganesh, A. Sentiment Analysis: A Comparative Study on Different Approaches. *Procedia Comput. Sci.* **2016**, *87*, 44–49. [CrossRef]

16. Mangold, W.G.; Faulds, D.J. Social media: The new hybrid element of the promotion mix. *Bus. Horiz.* **2009**, *52*, 357–365. [CrossRef]

17. Foster, J.; Çetinoglu, Ö.; Wagner, J.; Le Roux, J.; Hogan, S.; Nivre, J.; Hogan, D.; Van Genabith, J. # hardtoparse: POS Tagging and Parsing the Twitterverse. In Proceedings of the AAAI 2011 Workshop on Analyzing Microtext, San Francisco, CA, USA, 7–8 August 2011; pp. 20–25.

18. Yousefpour, A.; Ibrahim, R.; Hamed, H.N.A. Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis. *Expert Syst. Appl.* **2017**, *75*, 80–93. [CrossRef]

19. Xia, R.; Zong, C. A POS-based ensemble model for cross-domain sentiment classification. In Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, 8–13 November 2011; pp. 614–622.

20. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learning*; MIT Press Cambridge: Cambridge, MA, USA, 2016; Volume 1.

21. Dang, S.; Wen, M.; Mumtaz, S.; Li, J.; Li, C. Enabling Multi-Carrier Relay Selection by Sensing Fusion and Cascaded ANN for Intelligent Vehicular Communications. *IEEE Sens. J.* **2020**, *21*, 15614–15625. [CrossRef]

22. Cambria, E.; Poria, S.; Bajpai, R.; Schuller, B. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–16 December 2016; pp. 2666–2677.

23. Jozefowicz, R.; Vinyals, O.; Schuster, M.; Shazeer, N.; Wu, Y. Exploring the limits of language modeling. *arXiv* **2016**, arXiv:1602.02410.

24. Vateekul, P.; Koomsubha, T. A study of sentiment analysis using deep learning techniques on Thai Twitter data. In Proceedings of the 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen, Thailand, 13–15 July 2016; pp. 1–6.

25. Pal, S.; Ghosh, S.; Nag, A. Sentiment Analysis in the Light of LSTM Recurrent Neural Networks. *Int. J. Synth. Emot.* **2018**, *9*, 33–39. [CrossRef]

26. Miao, Y.; Gowayyed, M.; Na, X.; Ko, T.; Metze, F.; Waibel, A. An empirical exploration of CTC acoustic models. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2623–2627.

27. Hassan, A.; Mahmood, A. Deep Learning approach for sentiment analysis of short texts. In Proceedings of the 2017 3rd International Conference on Control, Automation and Robotics (ICCAR), Nagoya, Japan, 24–26 April 2017; pp. 705–710.

28. Baroni, M.; Dinu, G.; Kruszewski, G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), Baltimore, MD, USA, 23–25 June 2014; Volume 1, pp. 238–247.

29. Yuan, Y.; Zhou, Y. Twitter sentiment analysis with recursive neural networks. In *CS224D Course Project*; Stanford University: Stanford, CA, USA, 2015.

30. Vinodhini, G.; Chandrasekaran, R. Sentiment analysis and opinion mining: A survey. *Int. J.* **2012**, *2*, 282–292.

31. Singh, R.; Kaur, R. Sentiment Analysis on Social Media and Online Review. *Int. J. Comput. Appl.* **2015**, *121*, 44–48. [CrossRef]

32. Hemalatha, I.; Varma, G.; Govardhan, A. Automated Sentiment Analysis System Using Machine Learning Algorithms. *IJRCCT* **2014**, *3*, 300–303.

33. Marks, C.; Allen, L.; Gigliotti, F.; Busana, F.; Gonzalez, T.; Lindeman, M.; Fisher, P. Evaluation of the tranquilliser trap device (TTD) for improving the humaneness of dingo trapping. *Anim. Welf.* **2004**, *13*, 393–400.

34. Kharde, V.; Sonawane, P. Sentiment analysis of twitter data: A survey of techniques. *arXiv* **2016**, arXiv:1601.06971.

35. Shoukry, A.; Rafea, A. Preprocessing Egyptian dialect tweets for sentiment mining. In Proceedings of the Fourth Workshop on Computational Approaches to Arabic Script-Based Languages, San Diego, CA, USA, 1 November 2012; p. 47.

36. Carus, A.B. Method and Apparatus for Improved Tokenization of Natural Language Text. U.S. Patent 5,890,103, 30 March 1999.

37. Kantorov, V.; Laptev, I. Efficient feature extraction, encoding and classification for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 23–28 June 2014; pp. 2593–2600.

38. Mandelbaum, A.; Shalev, A. Word Embeddings and Their Use In Sentence Classification Tasks. *arXiv* **2016**, arXiv:1610.08229.

39. Mohamed, E. Morphological Segmentation and Part of Speech Tagging for Religious Arabic. In Proceedings of the Fourth Workshop on Computational Approaches to Arabic Script-Based Languages, San Diego, CA, USA, 1 November 2012; p. 65.

40. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]

41. Gers, F.A.; Schraudolph, N.N.; Schmidhuber, J. Learning precise timing with lstm recurrent networks. *J. Mach. Learn. Res.* **2003**, *3*, 115–143. [CrossRef]

42. Graves, A.; Jaitly, N.; Mohamed, A. Hybrid speech recognition with Deep Bidirectional LSTM. In Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 8–12 December 2013; pp. 273–278.

43. Hermans, M.; Schrauwen, B. Training and analysing deep recurrent neural networks. *Adv. Neural Inform. Proc. Syst.* **2013**, *26*, 190–198.

44. Kim, J.; Yoo, J.-B.; Lim, H.; Qiu, H.; Kozareva, Z.; Galstyan, A. Sentiment Prediction Using Collaborative Filtering. In Proceedings of the 2013 International AAAI Conference on Weblogs and Social Media (ICWSM), Cambridge, MA, USA, 8–11 July 2013.

45. Caruana, R.; Niculescu-Mizil, A. Data mining in metric space: An empirical analysis of supervised learning performance criteria. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; pp. 69–78.

46. Schell, M.J.; Yankaskas, B.C.; Ballard-Barbash, R.; Qaqish, B.F.; Barlow, W.E.; Rosenberg, R.D.; Smith-Bindman, R. Evidence-based target recall rates for screening mammography. *Radiology* **2007**, *243*, 681–689. [CrossRef]

47. Wu, J.; Ji, T. *Deep Learning for Amazon Food Review Sentiment Analysis*; Stanford University: Stanford, CA, USA, 2016.

48. Aljuhani, S.A.; Alghamdi, N.S. A Comparison of Sentiment Analysis Methods on Amazon Reviews of Mobile Phones. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 608–617. [CrossRef]

49. Nguyen, H.; Veluchamy, A.; Diop, M.; Iqbal, R. Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches. *SMU Data Sci. Rev.* **2018**, *1*, 7.

50. Hong, J.; Fang, M. Sentiment analysis with deeply learned distributed representations of variable length texts. *Stanf. Univ. Rep.* **2015**, *2015*, 1–9.

51. Asghar, N. Yelp dataset challenge: Review rating prediction. *arXiv* **2016**, arXiv:1605.05362.