COMPUTOLOGY: Journal of Applied Computer Science and Intelligent Technologies
Volume 3, Issue 2, Jul-Dec 2023, pp. 43-68
DOI: 10.17492/computology.v3i2.2303
https://www.journalpressindia.com/cjacsit

A New Novel Approach for Sentiments Analysis using Contextual Mining and Supervised Learning

Sachin Vyawhare[1]*, *Shashi Bhushan*[2], *Sapna Tayde*[1] *and Mrunali Jaiswal*[1]

ABSTRACT

Textual data mining is used to anticipate the sentiment of a user based on a similar book. Using conditional probability distributions, the rating similarity between books can be quantified by textual mining. In supervised learning, the input and output data are sent to the machine learning model in tandem to maximize accuracy. In this paper, a cloud-based suggested system is incorporated which detects and recommends similar types of content based on the ranking of books. A cloud recommendation system is a means of determining which service is best suited to the user's tastes and requirements. The work uses specific formats such as graphics or text, and the networks built throughout the process may help uncover the links between the words. Textual mining is used to calculate book-ranging similarities and recommendations. From the comparison result, the proposed logistic regression technique has maximum accuracy which is 82% as compared to the existing technique The model shows all recommended books that are like the input book are displayed, while all other types of books are hidden from view. Also, positive, or negative masking can be determined by comparing the intensity of the extracted images.

Keywords: *Sentiments Analysis; Cloud-based Recommendation System; Contextual Mining; Text Mining; Supervised Learning.*

1.0 Introduction

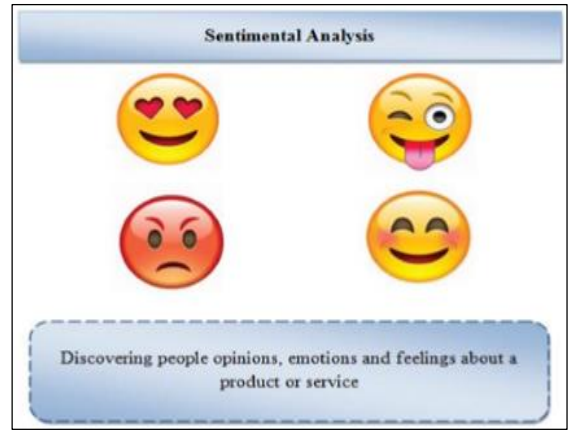
Sentiments Analysis (SA) is a Natural Language Processing (NLP) technique that determines a text's emotional tone. SA or opinion mining is an active subject of study in text mining field.

[1]*Computer Science & Engineering, Sanmati Engineering College, Washim, Maharashtra, India*
[2]*Computer and Information Sciences, UTP, Seri Iskandar, Perak, Malaysia*
**Corresponding author e-mail: vyawhare01@gmail.com*

SA is the computer examination of individuals' thoughts, attitudes, and feelings concerning an object [1]. As the name suggests, this kind of research focuses on how people's written language expresses how they feel about events and how they behave in response to them. Machine learning and deep learning, as well as the integration of these sciences with statistical methods to NLP, have made it one of the most active areas of research [2]. SA might also be used for financial markets, news events, and political debates as well as product evaluations [3]. These types of websites are regarded as particularly reliable sources of information because users can openly express and debate their thoughts and ideas on a subject on social networking sites and microblogging sites [4] [5]. Twitter, Short Message Service (SMS), as well as other community networks have grabbed the interest of the investigation industry and society as significant real-time opinion sources. It is one of the key aims of NLP to construct and build computer systems based on Artificial Intelligence (AI) [6-9]. Text sources, both structured and unstructured, might be now mined for previously unknown patterns of interest and information by use of these automated methods [10-12]. This technique, known as Sentiment Analysis, seeks to ascertain if a text's context is polarizing (positive, negative, or neutral) by analyzing its tone [13-16].

In the most recent few years, several applications and improvements on SA algorithms have been presented. Individuals can disseminate their insights, experiences, and views to the rest of the world by using social media platforms such as blogs, forums, wikis, review sites, social media, tweets, and so on [17]. Figure 1 depicts the different types of emotions [18].

Figure 1: Different Types of Sentiments [18]



Supervised learning is a procedure that involves giving the machine learning model both the appropriate input data and the proper output data. Supervised learning has a variety of real-world applications, including fraud detection, risk assessment, image categorization, and spam filtering, amongst others. An opinion is essentially a favorable or negative feeling, perspective, attitude, emotion, or evaluation of an entity or an element of the entity, and it is held by someone at a certain moment by someone who has the opinion [19] [20] [9]. Monitoring and analyzing social phenomena, such as recognizing potentially hazardous situations and evaluating the overall tone of the blogosphere, are two more applications of sentiment analysis. To predict the polarization of mindsets based on learning as well as test data sets, the machine learning technique is used as an approach. It makes use of a predetermined set of phrases, each of which is linked to a certain emotion.

The use of sentiment analysis is most common in a variety of disciplines, including sociology, politics, and marketing. The marketing industry uses it to build its strategy, find customer views on items or products, how people react to marketing campaigns and new products, and why customers don't purchase products [21].

1.1 Analysis of sentiments from different perspective

Feelings may be analyzed from some perspectives, the specifics of which are dictated by the needs of the area that is under consideration. In the course of our investigation, we will be considering the following factors:

1.1.1 Customer review analysis

The analysis of customer reviews is carried out by contrasting positive and negative comments made about individual aspects of a product [22]. This procedure is undertaken after the results of sentiment label propagation have been compiled. The meanings of the phrases are distributed across the various parts of a product in the same manner that is distributed among the many topics. As a consequence of this, three indices—controversy, complaint, and dissatisfaction—are developed to further emphasize negative feedback. The term controversy refers to the question 'how often are the subjects raised?'. If a specific subject is said to have a high level of controversy, this indicates that the subject is debated by most reviewers or individuals [23]. How seriously the subject being complained about is indicated by the level of complaint. If a certain subject is significantly adversely complained about in comparison to other subjects in review papers, then that subject must be given considerable consideration since it is likely to be the cause of unhappiness on the part of consumers. For example, the amount of discontent with an item or service may be measured using this composite index

known as dissatisfaction, which is a measure that includes both controversy and complaints [24].

1.2 Natural Language Processing

NLP focuses a lot of its attention on developing procedures that can recognize ideas and the polarization of sentiments in the content of everyday social acts [25]. Contextual mining is the concept of mining the context from the corpus for this purpose NLP is used. The sentiments are extracted from the NLP to the numeric value as the computer understands the binary language so the conversion of text to numeric value is understood as the processing of natural language [26]. Since 2018, NLP has entered a new phase. As a result of Bidirectional Encoder Coming from the Transformer (BERT) by Google's impressive performance in 11 NLP tasks, the strategy of "pre-training + fine-tuning" has become one of the most often used in this domain [27]. Pre-training word vectors have been significantly transformed by BERT. More context information would be gleaned from sentences using BERT than the prior approach did from words using BERT. An NLP job is easier to carry out if the vector is broken down into sentences and the approaches used in NLP are as follows:

1.2.1 Bert

Representations of the BERT by using many of the unannotated corpora in general NLP tasks, it has been shown that pre-trained word vectors are very valuable and powerful [28]. However, the word vectors used in the studies only let a separate perspective-free interpretation be used for every word.

1.2.2 Word Embedding

One of the most common ways that NLP represents the vocabulary of a text is via the use of word embedding. It can determine the context of a word inside a document, including the semantic and syntactic similarities as well as the link to other words [30-32]. Projections in the continuous space of words are meant to ensure semantic and grammatical similarity between words embedded in the same space.

2.0 Literature of Review

This portion describes the specific research studies associated with SA in the Deep Learning region. SA tasks are effectively carried out by running different models and approaches, for example, DL models, which have recently been increased.

A sentiment analyzer dashboard implementation is suggested in [33]. Twitter data was subjected to a sentiment analysis using keywords, hashtags, and usernames submitted by users. To do sentiment classification, many different techniques for sentiment analysis were merged. The results are displayed in the form of a dashboard. The data on the dashboard is shown in appropriate plots and charts thanks to the incorporation of raw data from the sentiment analysis findings. Since not all implementation goals are fulfilled within the specified time frame, there is always room for improvement in future initiatives. Sentiment analysis algorithms are now dependent on lexicons, and this would change soon. Ontology or idea vectors might be used to compare and train machine learning algorithms for domain-specific comparison and training. Multilingual compatibility is also a consideration for future projects.

A technique for analyzing consumer sentiment is investigated and it concluded sentiment propagation as well as an examination of customer reviews [24]. To put into this practice, a term graph is created by using a process known as word embedding, inside another network, and semi-supervised learning is used. The strategy that was developed was used on 3,11,550 reviews that were spread among five vehicles and 10 online forums. As a consequence of the practical example, it has been determined which components of automobiles are responsible for consumer dissatisfaction, and hence which aspects of automobiles require further investment and inquiry.

The principles of automatically identifying the feelings stated in the English text for Amazon and Flipkart items utilizing Naive Bayes, SentiWordNet, Logistic Regression, Random Forest, and K-Nearest Neighbor (KNN) algorithms are discussed in [34], as well as a review of those concepts. Using five primary parameters Additionally, the Product Comment Summarizer and Analyzer (PCSA) system was suggested in the study. It provides a synopsis of the comments and divides them into the favorable, negative, and neutral categories that have been predefined. Consequently, the parameters, classifiers, and accuracy ratings used to evaluate their performance would be used in the evaluation.

Four different machine learning models, namely multilayer perceptron neural networks, random forests, logistic regressions, and a naïve Bayes classifier have been discussed to provide empirical proof of the possibility of emotion classification from contextual information in [35]. As a point of reference, a system based on either digital-world contextual information, or a mix of digital-world and real-world data may build individual models, generic models, and gender-specific models with equivalent performance to those acquired by an emotion recognition system. Contextual factors that may help with binary emotion categorization have also been included in the research.

[36] stated that Tweet data is collected and then analyzed by classifiers trained in machine learning. A voted classification technique was used to identify the 'tweet's class and confidence level after categorization by distinct classifiers. The fraction of good and negative tweets may also be determined using the polarity approach for categorization. As a last option, Deep Learning Models have been suggested to categorize the Tweets. The tweets have been classified using Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN-RNN) models. Deep Learning models like CNN-RNN, LSTM, etc., and their various combinations have shown improved functioning compared to the machine learning algorithms. Additional to this, Deep Learning Models were developed using datasets from a wide range of industries to improve their accuracy when applied to real-world data in those same industries. Doing so guarantees that the final model takes into account all of the social media's available variations. One of the models would make the final decision on what to do.

The Sentiment Analysis and Prediction (SAP) approach using KNN for Text Trend is presented in [37]. Tokens and stop words are first converted to tokens and deleted. Weighted words and intensity clauses, as well as emotion shifters, control the polarity in the phrase, paragraph, and the whole text. This stage's gleaned qualities were critical in improving the results. Finally, a KNN classifier was used to predict the trend of the input text based on the collected attributes. Use public datasets like movie reviews on Twitter to train and assess your model. The results of the experiments showed that the new solution was superior to the old one. The text analytics may also be carried out using a Graphics User Interface (GUI) Hello World-based text analysis framework.

[38] revealed a sentiment analysis-based electronic product recommendation system. To create predictions about things, the recommendation algorithms heavily depend on user ratings. These kinds of assessments are often deficient and very restricting. A sentiment-based model for contextual information is proposed for recommended systems that use user statements and preferences to generate a suggestion. The domain-sensitivity problem in suggestion, a term ambiguity prevention technique, is the goal of this approach. An evaluation of RMSE and MAE results demonstrates that the suggested sentiment-based model outperforms the usual collaborative filtering technique when it comes to product suggestion. The results showed that the best performance was achieved when = 0.70, demonstrating that the act rating adds to the optimal performance of both sentiment CF and context CF. That's why the last set of experiments all had a $\lambda = 0.70$.

An abstract model that might be used for sentiment analysis without the need for a specific lingua franca has been evaluated and presented in [39]. To get the attributes, researchers used natural language processing to extract them and feed them into different machine learning classification models. Bangla phrases were taken from the web using SVM, and the authors offered several different combinations. This model was limited to only two categories—positive and negative—for categorization purposes. The major purpose was to design a wide model. Among other things, stochastic gradient descent for optimization was included in the design. As a result, fresh user data may be introduced to the model without having to retrain the whole model. An Internet Movie Database (IMDB) review dataset was used to train the estimators, and different evaluation metrics were calculated for our estimators to analyze their performance. To translate the data into Bangla, researchers utilized Google Translator.

Study in [40] revealed that the quantity of unstructured text data produced by mankind increases in general, and in Cyberspace so does the demand to analyze it logically and extract various kinds of information from it. CNN and RNN have been used to improve NLP systems, and the results have been impressive. The CNN is an excellent method for extracting higher-level characteristics that are not affected by local translation. Multiple convolutional layers must be stacked to capture long-term relationships in neural networks, as these layers are local. To solve this issue, the author suggests a framework that combines CNN and RNN.

[41] stated that the Twitter sentiment analysis method allows polling public opinion on events or items that are relevant to them. Utmost of the recent research focuses on extracting sentiment characteristics from lexical and syntactic data. Emoticons, exclamation marks, Sentiment words, and other symbols are used to convey these characteristics. Word embeddings created from unsupervised learning on big Twitter datasets use latent contextual semantic connections as well as statistical co-occurrence properties between words in tweets, according to the work of the authors.

Study in [42] recommended CNN with Data Augmentation Technology (DAT), a hybridized Neural Network (NN) model architecture, outperformed many solo NN models. The suggested DAT improved the recommended model's generalization ability. Results of trials demonstrate that utilizing the recommended DAT with the NNs model may provide excellent SA or short text classification results without any handcrafted features. It was put to the test on a corpus of Chinese news headlines and a dataset of Chinese online comments. It outperformed some contemporary models.

Text articles containing misleading assertions, particularly news, have lately been a source of irritation for Internet users. These pieces are widely circulated, and

readers are having a hard time distinguishing truth from the narrative. Earlier research on integrity evaluation has concentrated on objective analysis and language characteristics. The differentiation between the characteristics of genuine and fraudulent articles is the task's primary difficulty. The author presented a new method called Credibility Outcome (CREDO) in the work, which attempts to score an article's credibility in an open domain environment [43].

A classifier for determining the polarity of text data in both Bangla and English that is now available online. Polarity strength may be estimated using the Support Vector Machine (SVM). The translation method is used to create a dataset for Bangla. Classifying data is a procedure that removes unnecessary noise from the data. SVM was utilized as a supervised learning strategy by academics who achieved positive results. SVM was shown to be superior to Naive Bayes in the classification of Bangla words. To ensure the accuracy of the classifier, it is also used individual assessments, which have been proven accurate. Accuracy for SVM and Nave Bayes increases to 82.0% and 78.8%, respectively when negative results are ignored [44].

An approach that generates tagged data that may be used to forecast election outcomes has been presented [45]. However, even if the model isn't satisfactory to predict the findings on its own, it becomes crucial when combined with additional statistical models and offline methods (like exit polls). The model is suggested based on data gathered from three days of tweeting. An automated framework for mining months of data might be added to the model in the future since election result prediction is an ongoing process that demands studies conducted over long periods. The development of an active learning model, in which the model itself suggests the labeling of data, is highly recommended. It would reduce the time and effort required for labeling while ensuring that contextual relevance is not compromised.

[46] stated the stock market's mood may be studied by obtaining Sensex and Nifty live server data values at different time intervals and then comparing those values to historical averages. For the aim, Python, a programming language with a fast execution environment, is used. the helps investors estimate which stocks to invest in. It would also help maintain the stock market's economic balance. Python scripts with more advanced functionality may be used to do this in the future.

SentiRobo, a supervised machine learning algorithm for automated sentiment analysis of Twitter material for use in education and airport management has been suggested in [47]. Sentiment Clustering (SentiRobo) was a new technique for enhancing pure Nave Bayes' performance on large datasets. An overall accuracy rate of 71% and

79% was achieved by SentiRobo when predicting the sentiment value of mixed English and Malay tweets on education and airport administration respectively.

A framework for Twitter message polarity analysis that incorporates both methodologies plus an automated contextual module has been discussed in [48]. To evaluate the performance of the suggested system, four text datasets from the scientific literature are used. Decision Trees (J48), Naive Bayes (NB), SVM, and KNN were the five types of classifiers that were analyzed and compared. According to the results of the study, this framework can automate the whole polarity analysis procedure, with high accuracy and low false-positive rates.

Study in [49] revealed contextual valence analysis may be used to measure the sentiment of Bangla texts. The valence of a verb in linguistics refers to the amount of satellite noun phrases that a verb uses. The WorldNet and SentiWordNet were utilized by researchers to determine the prior valence (i.e. polarity) of each word and the senses associated with each one. Researchers are responsible for determining a statement's positive, negative, or neutral tone. Researchers used valency analysis to design a new method for determining the emotion of the Bangla text. There are enough examples and tests to demonstrate the approach.

An evaluative study of a rule-based paradigm called Valence Aware Dictionary of Sentiment Reasoning (VADER) for broad sentiment analysis is presented in [50]. This model is compared to eleven typical state-of-the-art benchmarks, such as Linguistic Inquiry and Word Count (LIWC), Affective Norms for English Words (ANEW), the SentiWordNet, General Inquirer, and machine learning-oriented techniques based on Maximum Entropy, Naive Bayes, and SVM algorithms. Qualitative as well as quantitative methods were used to build and test lexical features (as well as their related sentiment intensity measures) that are sensitive to sentiment in microblog-like settings. Combine this vocabulary with five fundamental guidelines that comprise grammatical and syntactic criteria for expressing and strengthening feelings of intensity. Results demonstrate that VADER surpasses human raters in terms of F1 Classification Accuracy (i.e., 0.96 and 0.84, respectively) with the use of the basic rule-based model for evaluating the sentiment of tweets

[51] suggested word embedding for sentiment classification is described in the research paper. To learn continuous word representations, most algorithms just consider syntactic context and ignore the text's emotional tone. A problem for sentiment analysis, which maps words with the same syntactic context but differing emotion polarity, such as positive and negative, to nearby word vectors. As a workaround for this problem, the technique known as Sentiment Specific Word Embedding (SSWE) has been developed.

There are three neural networks constructed that could include supervision from the polarity of text (such as sentences or tweets) into their loss functions. As a means of obtaining vast training corpora, researchers use unsupervised learning to extract SSWE from massive collections of tweets. SSEW is similar to handcrafted features in the best-performing system, and concatenating SSWE with an existing feature set improves performance even more.

A personalized recommendation system using sentiment analysis and learning automata is suggested in [52]. Here, the Learning Automata-based Sentiment Analysis (LASA) framework is applied to recommend places around the user's current location. The average penalty is calculated using the probability vector. It has several vulnerabilities, including insecure recommender systems. The LASA framework could be used to look at the security of recommender systems in the future. – LASA's framework could be used in additional issue domains, including cellular networks, ad-hoc network sensors, and IEEE 802.11-based networks.

[53] revealed that sentiment classification performance might be improved by using polarity shifting detection. Heuristic rules are built using multiple triggers from the corpus on polarity changing, which extract from the corpus. Polarity shifting is employed in the term-counting approach to categorize sentiment. Term-counting with polarity shifting has been shown in empirical studies to yield much better outcomes than regular term-counting. Combined with a machine-learning-based classifier, the combined classifier performs better than each one working alone does.

Collaborative filtering techniques are primarily used in recommendation systems to suggest items based on user-item interactions, while contextual mining and supervised learning are employed in sentiment analysis to classify sentiment in textual data, taking into account linguistic context and features extracted from the text.
The results of the literature review are outlined in Table 1, that might be seen below.

Table 1: Comparison of Literature Review

References	Technique Used	Outcome
[33]	A sentiment analyzer dashboard	The data on the dashboard is shown in appropriate plots and charts thanks to the incorporation of raw data from the sentiment analysis findings.
[24]	Word Embedding method with semi-supervised Machine Learning Technique	Identified unsatisfied and unhappy users from complete data.
[34]	Naive Bayes, Logistic Regression, SentiWordNet,	Supported suggested technique, the Product Comment Summarizer, and Analyzer (PCSA)

[35]	Four different machine learning models, namely multilayer perceptron neural networks, random forests, logistic regressions, and a naïve Bayes classifier	Results provide empirical proof of the possibility of emotion classification from contextual information
[36]	A voted classification technique	CNN-RNN and LSTM are two examples of deep learning models that have outperformed machine learning algorithms in distinct ways.
[37]	Sentiment Analysis and Prediction	The results of the experiments showed that the new solution was superior to the old one. The text analytics may also be carried out using a Graphics User Interface (GUI) Hello World-based text analysis framework.
[38]	Electronic product recommendation system	The results showed that the best performance was achieved when = 0.70, demonstrating that the act rating adds to the optimal performance of both sentiment CF and context CF. That's why the last set of experiments all had a $\lambda = 0.70$
[39]	Support Vector Machine	As a result, fresh user data may be introduced to the model without having to retrain the whole model.
[40]	CNN and RNN	A combination of used methods is suggested
[41]	Word embeddings approach derived from unsupervised learning	Employs dormant contextual semantic connections, and statistical co-occurrence features between words.
[42]	Recommended CNN with data augmentation technology	Results of trials demonstrate that utilizing the recommended DAT with the NNs model may provide excellent SA or short text classification results without any handcrafted features.
[43]	Credibility Outcome (CREDO)	The semantic similarity module was removed, which resulted in a decrease of 32.7%. Sentiment Analysis's accuracy rate was 7.1%.
[44]	A classifier based on SVM	Accuracy for SVM and Nave Bayes increases to 82.0% and 78.8%, respectively when negative results are ignored.
[45]	An approach generates tagged data	Every candidate's PvT Ratio, or proportion of positive tweets, would offer a good indication of their popularity.
[46]	Python, a scripting language	To keep the stock market's economy in balance, this project would be beneficial.
[47]	SentiRobo	SentiRobo was able to accurately forecast the sentiment value of mixed English-Malay tweets on education and airport management, with an overall accuracy rate of 71% and 79%
[48]	A polarity analysis framework for Twitter messages	It is possible to automate the whole polarity analysis process using this framework, with high accuracy and low false-positive rates, according to the results of the study.

[49]	WorldNet and SentiWordNet	It is possible to utilize both WordNet and SentiWordNet with Bangla text if the combinatory potential of words and sentences in both languages is equal.
[50]	VADER	The model was able to demonstrate that VADER surpasses human raters in terms of F1 Classification Accuracy (i.e., 0.96 vs. 0.84, respectively).
[51]	SSWE	SSEW is similar to handcrafted features in the best-performing system, and concatenating SSWE with an open characteristic set improves performance even more
[52]	LASSA framework, and an average penalty in probability vector	Reduced vulnerabilities and loopholes.
[53]	Polarity shifting detection	The findings of empirical research show that phrase counting with polarity shifting outperforms the regular word-counting technique substantially.

3.0 Background Study

SA has risen to prominence in NLP research in recent years as a result of the popularity of social networking sites like Facebook and Twitter. The attention mechanism in the deep neural network model has shown to be a powerful tool in the field of target-based SA. Extracting the message's sentiment or opinion is a common method for acquiring relevant information. As a result of their ability to learn from the training dataset, machine learning technologies have become more popular in sentiment classification. When the dataset is large, certain techniques may not work well. A naive Bayes classifier (NBC) is being evaluated for its capacity to handle large datasets in this study. Researchers utilized NBC instead of a traditional library (such as Mahout) to fine-tune the analytic process. In addition, this research develops a mechanism for analyzing Big Data. NBC's accuracy rises to 82 percent when the dataset size is doubled, which is an encouraging finding. A growing flow of movie reviews has shown that NBC can analyze the emotion from millions of such reviews [54].

4.0 Research Objective

- To combine both sentiment analysis and cloud recommendation systems to generate the most accurate recommendations for users.
- To increase the Cloud Service's accuracy.
- Assist cloud providers to promote their services and cloud users to identify services that satisfy their QoS criteria.

- To remove possible bottlenecks that limit the ability of people from taking benefit from cloud computing.

5.0 Problem Formulation

The easy availability and adaptability of the digital media have made the fake news transfer over social media a trend that results in rapid transfer/spreading of news which might be done by mistake or even intentionally which misleads the group/community. As per the present part is being considered about the format of the content over the social media then mainly it is in embedded formats like text/objects embedded over the image, the reason being to make the things understand easily by the target group. The particular forms, such as graphics or text, are taken into consideration in the work. Individuals can be processed in any format, and the majority of text/object embedded images used as classifiers to pull out emotions as negative/positive or real/fake is used in the suggested method.

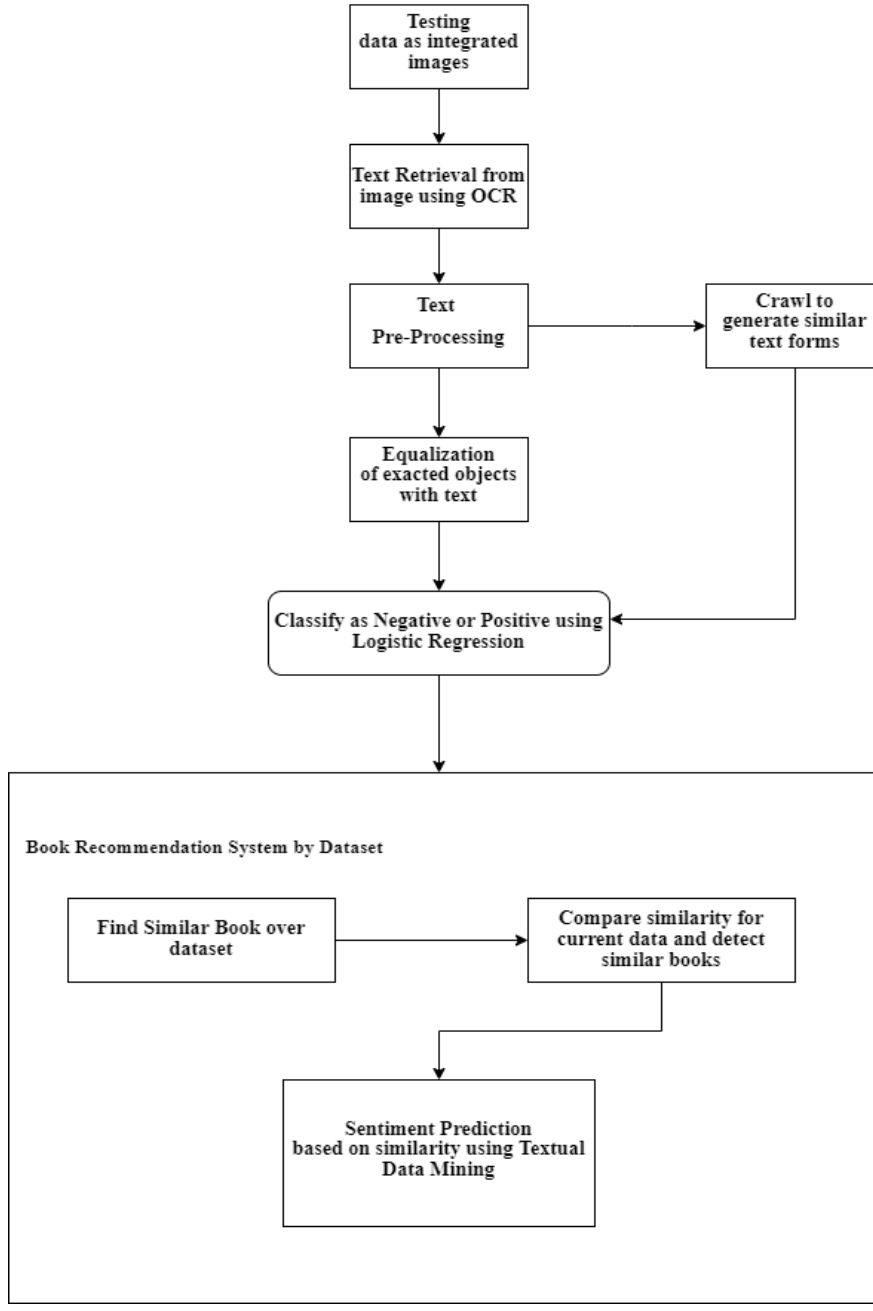
6.0 Research Methodology

The present effort incorporates a cloud-based suggested system that uses book rating to discover and recommend similar sorts of content. SA is carried through using textual data mining. The generated networks allow for the identification of word relationships. For book ranging similarity is computed and recommendations are computed using textual mining.

In the suggested methodology n-gram model is being used for the prediction process, the model is very important in testing and finding the crus which are applied in language modeling and NLP fields. Words, bytes, syllables, and characters might all be used to organize the data. The n-gram model is most often used to categorize a piece of text.

In the case when the dataset is being counted then it is the standby or the downloaded data from the social media containing the images embedded with text over it. The reason for considering the same is just because most of the news floated all around over social media is the textual image. The textual image can be further divided into parts an image embedded with only text, an image embedded with objects, and textual content both.

Figure 2: Proposed Methodology



The complete methodology goes with like first the digital image is being considered as input then the same is considered for textual content detection and object detection. For the textual detection from the image, the Optical Character Reader (OCR) is being used and for the detection of the object, the masking technique is being counted.

Textual OCR is defined as the electronic or mechanical conversion of the text data over the images in a character-by-character manner, the text can be in digital print, handwritten, or scanned documents and the can be the text embedded over the image as shown in Figure 2.

The general steps for the suggested methodology are as follows:

Step 1: At the initial phase of the suggested system, the image provided as input must go through certain processing steps. Pre-processing techniques such as stop word removal, stemming, etc. are used to clean up the text when it is extracted from the image. OCR technology is used to extract text from an image in the suggested project.

Step 2: Further, text preprocessing is done. Cleaning and preparing text data is known as text preprocessing. Pre-processing may be used to remove extraneous data from the received information.

Step 3: Next, crawl the images to generate similar text forms. The work of the crawler is to find undiscovered Web pages that may be entered into the Cloud Service Identifier. Most Web pages viewed are unlikely to be cloud services; hence a significant number of books would need to be crawled before a service can be found.

Step 4: For the validation of the suggested methodology different images collected from Google and other social media platforms and a dataset from the website Kaggle is being considered. In the fourth step, the extracted objects from the images are checked for positive or negative for which masking is checked with the help of intensity. For the objects embedded over the image intensity value is being checked like if the intensity levels of either the objects match or not which is being done using the Logistic Regression classifier [29].

Step 5: Then in the next phase, similarity values between the active books and other books are computed. A similar book can be identified by computing similarity values between the current active book and another training book.

Step 6: Finally, sentiment prediction is computed by using textual data mining based on a similar book. Textual mining is used to determine the most likely recommendations using conditional probability distributions to quantify the rating similarity amongst books.

7.0 Implementation Results

This section deals with the implementation results of the study. The results were founded by using the Python tool.

7.1 Tool

- Python: Python is a high-level, interpreted programming language that may be used for a wide variety of tasks. Code readability is a top priority, and it achieves this goal via substantial indentation. Python is a garbage-collected and dynamically typed system. The Python programming language is utilized in the implementation of the results. It is an object-oriented high-level, actively semantic, construed language. Dynamic linking and dynamic typing combine with their built-in high-level data structures to provide a perfect scripting language for quickly combining existing pieces during implementation. Readability is one of Python's primary focuses, which lowers the cost of maintaining programmers [55].

Result 1: Firstly, the data from the book is imported and imported data must be in text form. OCR was performed on the picture data to extract text data from it, and the data had book names and reviews accessible in image format (Figure 3).

Figure 3: Text Retrieval by OCR Technique

```
2]: config = ('-l eng --oem 1 --psm 3')
    pytesseract.pytesseract.tesseract_cmd = 'C:/Program Files/Tesseract-OCR/tesseract.exe'

3]: path=glob.glob("Book_Data/*.jpg")
    cv1=[]
    for i in path:
        # Dataset/*.jpg
        n=cv2.imread(i)
        cv1.append(n)

4]: TEXT=[pytesseract.image_to_string(iu, config=config) for iu in cv1]
```

Result 2: Figure 4 illustrates the preliminary processing that was done to the data. The data that was extracted might then have unnecessary data removed from it with the aid of pre-processing.

Figure 4: Text Pre-processing of Image

```
# Prepossessing of Data
import re
nstr = re.sub(r'[?|$|.|!]',r'',str(TEXT))
nestr = re.sub(r'[^a-zA-Z0-9 ]',r'',nstr)
# nestr
```

```
import nltk
nltk_tokens = nltk.word_tokenize(str(nestr))
```

Result 3: Crawl the images to generate similar text forms. In Figure 5, similar text forms of the image are represented in the form of a matrix. Similarity matrix generated crawl to generate similar text forms and equalization of exacted objects with text. For validation of the suggested methodology, Google, social media photos, and a Kaggle dataset are used.

Figure 5: Similar Text Forms of the Image

```
vect = TfidfVectorizer(min_df=1, stop_words="english")
tfidf = vect.fit_transform(nltk_tokens)
pairwise_similarity = tfidf * tfidf.T
pairwise_similarity
```

```
<17247x17247 sparse matrix of type '<class 'numpy.float64'>'
        with 79183 stored elements in Compressed Sparse Row format>
```

```
pairwise_similarity.toarray()
```

```
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 1., 0., ..., 0., 0., 0.],
       [0., 0., 1., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 1., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 1.]])
```

Result 4: The data that was crawled is saved in files that use the Comma-separated values (CSV) format. The values of the data that was stored are displayed in Figure 6 using Python code.

Figure 6: Data Store on CSV

```
import csv
f = open('Orc data.csv', 'w')
```

```
writer = csv.writer(f)

# write a row to the csv file
writer.writerow(nltk_tokens)

# close the file
f.close()
```

```
data = pd.read_csv("ocr_data.csv")
df = data
```

Result 5: Figure 7 represents the suggested model. In this stage, the suggested model is built. For objects included in an image, the logistic regression classifier checks if the intensity levels match. The accuracy of the model is 0.822 and the precision value F1 is 0.723 which gets from the implementation of the model.

Figure 7: Suggested Model

Building our Model

```
lr = LogisticRegression(random_state = 42, max_iter=1000)
lr.fit(train_feature_set,y_train)
y_pred = lr.predict(test_feature_set)
print("Accuracy: ",round(metrics.accuracy_score(y_test,y_pred),3))
print("F1: ",round(metrics.f1_score(y_test, y_pred),3))
```

```
Accuracy:  0.822
F1:  0.723
```

```
cm1 = confusion_matrix(y_test, y_pred)
cm1
```

```
array([[2120,  280],
       [ 362,  838]], dtype=int64)
```

```
cm2 = confusion_matrix(y_test, y_pred,normalize='true')
cm2
```

```
array([[0.88333333, 0.11666667],
       [0.30166667, 0.69833333]])
```

```
feature_importance = lr.coef_[0][:10]
for i,v in enumerate(feature_importance):
    print('Feature: ', list(cv.vocabulary_.keys())[list(cv.vocabulary_.values()).index(i)], 'Score: ', v)
```

Result 6: Then, in the following stage, provide a recommendation system that is based on rankings. Predicting Whether or not the book has received favorable, unfavorable, or favorable reviews. After that, suggest a book in the same area (Figure 8).

Figure 8: Suggested Book

```
# cloud storag
myurl = 'http://books.toscrape.com/index.html'
uClient = uReq(myurl)
page_html = uClient.read()
uClient.close()
```

```
page_soup = soup(page_html, "html.parser")
bookshelf = page_soup.findAll("li", {"class": "col-xs-6 col-sm-4 col-md-3 col-lg-3"})
filename = ("Books.csv")
f = open(filename, "w")

headers = "Book title, Price\n"
f.write(headers)
```

```
18
```

```
for books in bookshelf:

    book_title = books.h3.a["title"]
    book_price = books.findAll("p", {"class": "price_color"})
    print("Title of the Book :" + book_title)
```

```
Title of the Book :A Light in the Attic
Title of the Book :Tipping the Velvet
Title of the Book :Soumission
Title of the Book :Sharp Objects
```

Result 7: This whole section does not need any visuals, since its only purpose is to enable users to input the name of a book and determine if the book's reviews are favorable or bad before recommending another book in the same area (Figure 9).

Result 8: Figure 10 shows the overall process of the model. In the developed model, the name of the book is input first. After that, the system displays all suggested books that are comparable to the book that was entered, while hiding any other sorts of books. Then, following the instructions, displays the reviews, which may be positive or negative.

Figure 9: Working Process of the Model

```
# recommend_me_books('The Fellowship of the Ring (The Lord of the Rings, Part 1)')
senti=str(input('Enter The Book Name '))

# recommend_me_books('Message in a Bottle')
recommend_me_books(senti)
test_review = cv.transform([senti])
s = lr.predict(test_review)
if  s==1:
    print("Sentiment: The book has Positive Reviews")
elif s==0:
    print("Sentiment: The book has Negative Reviews")
```

```
Enter The Book Name  The Mulberry Tree
You Like : The Mulberry Tree.
You may also like these :
Forever... : A Novel of Good and Evil, Love and Hope
Midnight Bayou
Three Fates
Killjoy
Sanctuary
Sentiment: The book has Positive Reviews
```

Figure 10: Resultant Process of the Model

```
Enter The Book Name  Message in a Bottle
You Like : Message in a Bottle.
You may also like these :
Nights in Rodanthe
The Mulberry Tree
A Walk to Remember
River's End
Nightmares &amp;amp; Dreamscapes
Sentiment: The book has Positive Reviews
```

8.0 Comparative Results

Figure 11 given below shows the comparison in classifiers used in the base paper with the other one. The classifier is compared namely the technique used in the base paper [54] i.e., Naïve Bayes Classifier (NBC). The Naive Bayes Classifier is less accurate than the suggested Logistic Regression method. The accuracy of Logistic Regression is 82%, while NBC accuracy is just 80%. F1 accuracy of Logistic Regression

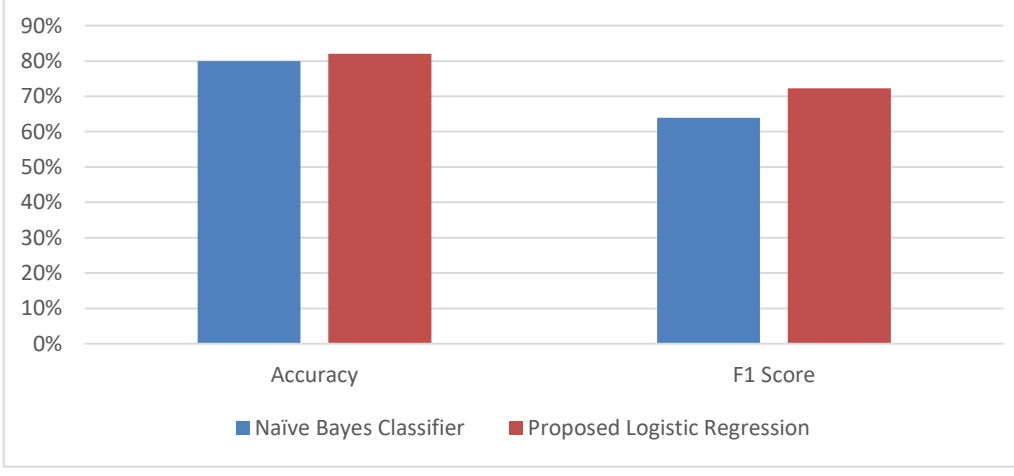
is 72.3% while the accuracy of NBC is 63.9%. The comparison of the two different methods is shown in Table 2.

Table 2: Comparison of Techniques

Techniques	Accuracy	F1 Score
Naïve Bayes Classifier	80%	63.9%
Proposed Logistic Regression	82%	72.3%

Figure 11 shows the comparison graph which has been described and given below. From the graph, the proposed logistic regression technique has maximum accuracy which is 82% as compared to the existing technique.

Figure 11: Comparison Graph



9.0 Conclusion and Future Scope

Text SA is a technique for detecting the user's feelings from a text into different feelings, such as positive, negative, or neutral, or emotions, such as delighted, sad, annoyed, or disgusted, to study the user's actions toward a given subject or entity. In the work, a cloud-based system is based on the ranking of books, discovers, and presents comparable types of content to the user. Mining of text is done to carry out SA. Following the completion of the pre-processing of the text, the OCR technology is

applied to enable the text to be recovered. From the comparison results, the suggested model's accuracy is 82% higher than the comparative to Analytical Nave Bayes. According to the suggested paradigm, if a user searches for a book, then the system will recommend all comparable books. Also, sentiment displays either positive or negative reviews about the book. In future work, it has the capability of expanding the classification categories to accomplish greater results. It can begin working on additional book languages such as Hindi, Russian, and Arabic to provide sentiment analysis to a larger number of users.

References

[1] Yu, L.-C., Jheng-Long, W., Pei-Chann, C. & Hsuan-Shou, C. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems, 41*, 89-97.

[2] Ligthart, A., Cagatay, C. & Bedir, T. (2021). Systematic reviews in sentiment analysis: A tertiary study. *Artificial Intelligence Review, 54*(7), 4997-5053.

[3] Hagenau, M., Michael, L. & Dirk, N. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems, 55*(3), 685-697.

[4] Maks, I. & Piek, V. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems, 53*(4), 680-688.

[5] Ravi, K. & Vadlamani, R. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches, and applications. *Knowledge-based Systems, 89*, 14-46.

[6] Heidari, M. & Setareh, R. (2020). Using transfer learning approach to implement convolutional neural network model to recommend airline tickets by using online reviews. *In 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization* (SMA, pp. 1-6). IEEE.

[7] Zirpe, S. & Bela, J. (2017). Polarity shift detection approaches in sentiment analysis: A survey. *In 2017 International Conference on Inventive Systems and Control (ICISC),* pp. 1-5. IEEE.

[8] Heidari, M. & Setareh, R. (2020). Semantic convolutional neural network model for safe business investment by using bert. *In 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 1-6. IEEE.

[9] Wang, H. & Jorge, A. C. (2015). Sentiment expression via emoticons on social media. *In 2015 IEEE International Conference on Big Data*, pp. 2404-2408. IEEE.

[10] Heidari, M., James, H. J. & Ozlem, U. (2020). Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter. *In 2020 International Conference on Data Mining Workshops (ICDMW)*, pp. 480-487. IEEE.

[11] Lunn, S., Jia, Z. & Monique, R. (2020). Utilizing web scraping and natural language processing to better inform pedagogical practice. *In 2020 IEEE Frontiers in Education Conference (FIE)*, pp. 1-9. IEEE.

[12] Rezapour, R., Lufan, W., Omid, A. & Jana, D. (2017). Identifying the overlap between election result and candidates' ranking based on hashtag-enhanced, lexicon-based sentiment analysis. *In 2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pp. 93-96. IEEE.

[13] Heidari, M. & James, H. J. (2020). Using bert to extract topic-independent sentiment features for social media bot detection. *In 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pp. 0542-0547. IEEE.

[14] Heidari, M., Samira, Z., Brett, B. & Setareh, R. (2021). Ontology creation model based on attention mechanism for a specific business domain. *In 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1-5. IEEE.

[15] Farahani, A., Sahar, V., Khaled, R. & Hamid, R. A. (2021). A brief review of domain adaptation. *Advances in Data Science and Information Engineering*, 877-894. Retrieved from https://www.researchgate.net/publication/344551768_A_Brief_Review_of_Domain_Adaptation

[16] Yang, Z., Diyi, Y., Chris, D., Xiaodong, H., Alex, S., & Eduard, H. (2016). Hierarchical attention networks for document classification. *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480-1489.

[17] Gao, M., Renli, T., Junhao, W., Qingyu, X., Bin, L., & Linda, Y. (2015). Item anomaly detection based on dynamic partition for time series in recommender systems. *PloS one, 10*(8), e0135155.

[18] Pandarachalil, R., Selvaraju, S. & Mahalakshmi, G. S. (2015). Twitter sentiment analysis for large-scale data: An unsupervised approach. *Cognitive Computation, 7*(2), 254-262.

[19] Tsytsarau, M. & Themis, P. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery, 24*(3), 478-514.

[20] Alaei, A. R., Susanne, B. & Bela, S. (2019). Sentiment analysis in tourism: Capitalizing on big data. *Journal of Travel Research, 58*(2), 175-191.

[21] Smith, A. N., Eileen, F. & Chen, Y. (2012). How does brand-related user-generated content differ across YouTube, Facebook, and Twitter. *Journal of Interactive Marketing, 26*(2), 102-113.

[22] Suresh, P. & Gurumoorthy, K. (2022). Mining of customer review feedback using sentiment analysis for smart phone product. *In International Conference on Computing, Communication, Electrical and Biomedical Systems*, (pp. 247-259). Springer, Cham.

[23] Jeon, W., Yebin, L. & Youngjung, G. (2021). Airline service quality evaluation based on customer review using machine learning approach and sentiment analysis. *The Journal of Society for e-Business Studies, 26*(4), 15-36.

[24] Park, S., Cho, J., Park, K. & Shin, H. (2021). Customer sentiment analysis with more sensibility. *Engineering Applications of Artificial Intelligence, 104*, 104356.

[25] Bera, A., Mrinal, K. G. & Dibyendu, K. P. (2021). Sentiment analysis of multilingual tweets based on Natural Language Processing (NLP). *International Journal of System Dynamics Applications (IJSDA), 10*(4), 1-12.

[26] Deepa, D. (2021). Bidirectional encoder representations from transformers (BERT) language model for sentiment analysis task. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(7), 1708-1721.

[27] Devlin, J., Ming-Wei, C., Kenton, L. & Kristina, T. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[28] Alaparthi, S. & Manit, M. (2021). BERT: A sentiment analysis odyssey. *Journal of Marketing Analytics, 9*(2), 118-126.

[29] Fatemeh, H. & Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review, 52*(3), 1495-1545.

[30] Alamoudi, E. S. & Norah, S. A. (2021). Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *Journal of Decision Systems, 30*(2-3), 259-281.

[31] Ghannay, S., Benoit, F., Yannick, E. & Nathalie, C. (2016). Word embedding evaluation and combination. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 300-305.

[32] Dang, N. C., Moreno-García, M. N. & Fernando, D. P. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, *9*(3), 483.

[33] Lien, A. K., Lars, M. R., Hans, P. F. T. & Maryam, E. (2022). OSN dashboard tool for sentiment analysis. arXiv preprint arXiv:2206.06935.

[34] Dadhich, A. & Thankachan, B. (2021). Social & juristic challenges of AI for opinion mining approaches on Amazon & Flipkart product reviews using machine learning algorithms. *SN Computer Science, 2*(3), 1–21.

[35] Salido, O., Martin, G., Luis-Felipe, R. & Gutierrez-Garcia, J. O. (2020). Towards emotion recognition from contextual information using machine learning. *Journal of Ambient Intelligence and Humanized Computing, 11*(8), 3187-3207.

[36] Chandra, Y. & Antoreep, J. (2020). Sentiment analysis using machine learning and deep learning. *In 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1-4. IEEE.

[37] Abdul, R., Asim, M., Ali, Z., Qadri, S., Mumtaz, I., Khan, D. M. & Niaz, Q. (2019). Text sentiment analysis using frequency-based vigorous features. *China Communications, 16*(12), 145-153.

[38] Osman, N. A., Noah, S. A. M. & Darwich, M. (2019). Contextual sentiment-based recommender system to provide recommendation in the electronic products domain. *International Journal of Machine Learning and Computing, 9*(4), 425-431.

[39] Hasan, M., Ishrak, I. & Hasan, K. M. A. (2019). Sentiment analysis using out of core learning. *In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1-6. IEEE.

[40] Hassan, A. & Ausif, M. (2018). Convolutional recurrent deep learning model for sentence classification. *IEEE Access, 6*, 13949-13957.

[41] Jianqiang, Z., Gui, X. & Zhang, X. (2018). Deep convolution neural networks for twitter sentiment analysis. *IEEE Access, 6*, 23253-23260.

[42] Sun, X. & Jiajin, H. (2018). A novel approach to generate a large scale of supervised data for short text sentiment analysis. *Multimedia Tools and Applications*, 1-21.

[43] Choudhary, N., Singh, R., Bindlish, I. & Shrivastava, M. (2018). Neural network architecture for credibility assessment of textual claims. arXiv preprint arXiv:1803.10547.

[44] Sabuj, M. S., Zakia, A. & Hasan, K. M. (2017). Opinion mining using support vector machine with web based diverse data. *In International Conference on Pattern Recognition and Machine Intelligence*, pp. 673-678. Springer, Cham.

[45] Ramteke, J., Shah, S., Godhia, D. & Shaikh, A. (2016). Election result prediction using Twitter sentiment analysis. *In 2016 International Conference on Inventive Computation Technologies (ICICT)*, 1, 1-5. IEEE.

[46] Bhardwaj, A., Narayan, Y. & Dutta, M. (2015). Sentiment analysis for Indian stock market prediction using Sensex and nifty. *Procedia Computer Science*, *70*, 85-91.

[47] Rohani, V. A. & Shayaa, S. (2015). Utilizing machine learning in sentiment analysis: SentiRobo approach. *In 2015 International Symposium on Technology Management and Emerging Technologies (ISTMET)*, pp. 263-267. IEEE.

[48] Lima, A., Carolina, E. S., Leandro, N. C. & Juan, M. C. (2015). A polarity analysis framework for Twitter messages. *Applied Mathematics and Computation, 270*, 756-767.

[49] Hasan, K. M. A. & Mosiur, R. (2014). Sentiment detection from bangla text using contextual valency analysis. *In 2014 17th International Conference on Computer and Information Technology (ICCIT)*, pp. 292-295. IEEE.

[50] Hutto, C. & Eric, G. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *In Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216-225.

[51] Tang, D., Furu, W., Nan, Y., Ming, Z., Ting, L. & Bing, Q. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. *ACL, 1*, 1555-1565.

[52] Venkata, K. P., Misra, S., Joshi, D. & Obaidat, M. S. (2013). Learning automata-based sentiment analysis for recommender system on cloud. *In 2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pp. 1-5. IEEE.

[53] Li, S., Zhongqing, W., Sophia, Y. M. L. & Chu-Ren, H. (2013). Sentiment classification with polarity shifting detection. *In 2013 International conference on Asian language processing*, pp. 129-132. IEEE.

[54] Chen, S., Chao, P., Linsen, C. & Lanying, G. (2018). A deep neural network model for target-based sentiment analysis. *In 2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-7. IEEE.

[55] Hu, Z., Peng, J., & Zhao, H. (2021). Dynamic neural orthogonal mapping for fault detection. *International Journal of Machine Learning and Cybernetics, 12*(5), 1501-1516.