Journal of Theoretical and Applied Information Technology
30th November 2024. Vol.102. No. 22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EMOTION ANALYSIS OF TEXTUAL CONTENTS USING NATURAL LANGUAGE PROCESSING AND TEXT MINING

SUBHASH CHANDRA GUPTA[1], NOOPUR GOEL[2]
[1]Research Scholar, Department of Computer Applications, VBS Purvanchal University, Jaunpur, India
[2]Professor, Department of Computer Applications, VBS Purvanchal University, Jaunpur, India
E-mail: [1]csubhashgupta@gmail.com, [2]noopurt11@gmail.com

ABSTRACT

Background: As the internet based e-commerce applications are expanded, a tsunami of comments are arised in the digital world. Most of the comments are textual in nature, and a reflection of emotions made by the people from their day-to-day activities performed on these applications. Analysis of these comments is not an easy task, a highly technological model based on machine learning classifiers is required to perform this task.
Methods: Model's Performance is based on the classifiers, the higher the count of correct predictions made by the classifier, the better performance the model shows. In the proposed NLP based model, seven classifiers used in the model are KNN, Logistic regression, SVM, multinomial Naive Bayes, decision tree, random forest and Gradient Boosting. Input dataset is an example of multiclass, during the preprocessing, noise was removed from the dataset and converted into a binary class dataset. Model uses the term frequency-inverse document frequency (TF-IDF) to vectorize it for feature extraction before applying classifiers to produce its results.
Result and Analysis: Finally a comparative study has been made to analyze the obtained results of all classifiers used in the model. From the analysis, it is found that the SVM classifier is the best scorer for the model and it scores accuracy, F1 score, precision, recall and AUC 96.42%, 96.38%, 96.42%, 96.15% and 96.35 % respectively.
Keywords: *Sentiment Analysis, Imbalanced dataset, NLP, Text Mining, Text summarization.*

1. INTRODUCTION

As the cost of using the internet is decreasing day-by-day in India, it brings the whole world into the palm of a mango people through a device, "mobile". Not only the information related to knowledge, education, research, politics, healthcare and many more, are just a single click away from his hands, but also the markets also reach their door-steps through it. A consumer, whatsoever he needs; those he wants to buy; all may be bought online through e-commerce websites. After shopping online, he shares his joy, anger, disappointments, and satisfaction with the other people about the product, use of product, merits or demerits of products along with the service provided by the producers. Similarly he also shares his words, through online messaging applications, about the current market situation, the political and social prosperity of a society. The words which are shared online, are called reviews, opinions or sentiments[1], [2].

Millions of reviews/ sentiments are added in the digital world every day and available in text form and carry huge amounts of information. Analysis/summarization of reviews are needed everywhere in online marketing[3]. The Internet makes online shopping easy for a customer, but selecting the right product is very difficult. The brand name of a product/ service may help customers a little bit. Most of the customers consider getting feedback from the reviews of products made by other customers. There are thousands of reviews available on the digital world about a specific product. Reading all these reviews about the product and making a perception from it about the product is a very difficult and time taking process[4]. Sometimes, some users also write long reviews and a user could get their required information only after covering the whole review. Thus it is needed to minimize the reviews in a shorter form which has the same meaning as the whole contents [1].

On the other side, an organization/company also wants to read and analyzes the reviews to get feedback from it to take necessary steps to increase

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the customer support. The information hidden in these reviews are valuable and organizations may require it not only to increase their customer support but also to focus on new strategies / concepts to compete their opponents[5]. Today, a company receives a large number of text data in the form of emails, social media and chat box messages, and customer responses. These textual contents are so important for the company and require the needs to analyze and extract the information hidden in it. The text contents may contain information related to reviews of customers about its products/ services or the complaints/feedback or something else[6], [7]. They may take necessary steps such as rewarding his liable customers, or improve their services/ products after getting knowledge obtained from the analysis of text reviews[8].

A sentiment/ review is a textual material which was shared/ given about a product or service by a user digitally. Sentiment analysis is the process of analyzing these digital texts to extract the emotions of messages and categorize them as positive, negative or neutral labels[9]. Due to the nature of textual content, only data mining/ machine learning techniques are unable to extract valuable information from it. This is the situation where techniques related to the text mining, text summarization and text analytics come into the picture and play the role of "knowledge miners" to extract information from it[10].

In the research work, a model is proposed using machine learning classifiers to analyze the sentiments as positive or negative. The model used the emotion dataset[11] downloaded from Kaggle, a public platform to provide dataset, for research work.

Although the samples has six different types of samples, the focus is on the classification of these emotions in two different class labels which are either positive emotions or negative emotions. The objective of the study to emphasis on the method to enhance the analysis ability of the emotion analysis model. The study has included the suitable preprocessing methods, classification algorithms and NLP techniques to achieve its goal and they are also discussed in detail in the different sections of the study. The tuning of classifier's hyperparameter would be better option to enhance the performance, but it is not included in the study and it will be considered in future on the same or different dataset.

Most of the researcher consider the dataset as a multi-class dataset and analyzed their performance for each class label individually. But if generalization of multiclass dataset into a binary class dataset, model may produce better results and take less computation time also. For example a customers have made comments including "joy", "happy" or "enjoy" type of words in their text, then it all are showing positive sentiment for the organization. Similarly a sentences having "pain", "sadness" or anger words are showing negative sentiments. So instead of considering these similar emotion type words in different multi-class labels, they might be mapped into two class labels that is either positive or negative sentiments, i.e. a binary class dataset. In study, this approach has been applied to enhance the performance of model.

2. LITERATURE REVIEW

The mining of information from textual contents is performed by the techniques called text mining and with the advancement in NLP techniques, made it as a hot topics of research amongst the researchers. A number of researches has been done in it and still have the scope of new researches or enhancing the technological performance of the existing researches. Sentiment analysis also known as emotion analysis is a techniques belongs from text mining. In the study, the literature survey has included some of the research articles which are related to sentiment analysis.

Sachin Misal et al [12] has proposed a model to investigate online cell phone surveys using a combination of partner altered K methods calculations, KNN, Naive Bayes other machine learning methods. In their study, LSTM, RNN are shown playing a major role in sentiment analysis and producing a good accuracy for the customer review. Yanying Mao et al [5] have made a systematic literature review of the paper related to the sentiment analysis and emphasized the methods/ techniques used for, the classifiers used, by the content the analysis had made and many more. Their findings also emphasized the importance of AI in automatic text analysis and focused on the importance of analysis of sentiments in the life of people.

Xiaohui Shen et al[13] has developed a model using deep learning techniques to analyze the sentiments of modern Chinese literature. In their paper, they had proposed a novel deep learning framework recurrent neural networks (RNNs)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

termed BERT-LLSTM- DL. In the model framework, they have integrated Bidirectional Encoder Representations from Transformers (BERT), Long Short-Term Memory (LSTM) networks and deep learning techniques. BERT was applied for language representation, LSTM networks for sequential learning and deep learning for feature extraction. The proposed model was evaluated on a dataset containing Chinese literature textual contents and produced promising results in terms of accuracy, precision, recall, and F1-score values across multiple iterations.

Manual summarization of a textual document is always a challenging task for humans. A large number of researchers have been trying to improve the summarization processes digitally in the past some years.

Taneja A. et al [3] and Suryawanshi N. S. [7] has made an effort to review of researched made in this direction in last some years and summarized them according the techniques used, types of summarization, evaluation metrics, standard datasets and future scopes for research.

Semary N et al[14] has emphasized the role of the feature extraction process for a sentiment analysis model. They show by their research work that the performance of an analysis model is affected by the feature extraction process. There are a number of feature extraction methods such as Word2Vector, N-gram, Bag-of-words (BOW), Hashing Vectorizer (HV), Term Frequency- Inverse Document Frequency (TF-IDF), and Global vector for word representation (GloVe). They applied these methods to the dataset to evaluate the ability of each feature extraction method and found that Based on our results, they had found that the TF-IDF technique demonstrated better performance, with a higher accuracy in the Amazon reviews dataset and the Twitter US airlines dataset.

Mohandas Archana et al[1] have proposed an analysis model to analyze the sentiments of customers for electronic products. The input dataset has been collected from the world famous e-commerce organization such as Flipkart and other media. The aim of their research work is to focus on the polarity of consumer comments using machine learning sentiment analysis models on the textual comments/ reviews on electronic items. They transformed, after preprocessing the dataset, the text-based data into the numerical format by applying vectorization methods. In the model, they have divided the dataset into training and test dataset and further used different machine learning classifiers such as Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine (SVM).

The objective of their research is to identify the most significant sentiments within the textual comments/ reviews and check the efficiency and effectiveness of the classifiers used for their model. Most of the time extraction of emotions/ sentiments are not in the form of two predefined binary class labels. Instead, it may be in the form of multiple class labels. Kavitha Subramani et al[15] has developed a model using NLP and machine learning classifiers to analyze a multi-class sentiment analysis from a text dataset. The model is tested by three classifiers SVM, Random Forest, Decision Tree and achieved a 90% accuracy level.

3. THEORETICAL BACKGROUND

3.1. Text Mining

Most of the data mining and machine learning techniques are applicable on structured data or produce better results when input data are in structured form and in numeric. Structured data are those data/dataset which follows a well-defined structured format and can be represented in rows and columns[16], [17]. But more than two-third portion of digital world data is unstructured/ semi-structured and contains text only. The examples of unstructured data files are web pages, text files, pdfs, and chat messages and reviews. Extracting information from these textual contents is a tedious and time taking task.

Text mining[18], [19], also referred as text analysis or text analytics, is the process that extracts information from text and represents it in structured form so as the machine learning techniques may be applicable to get valuable information from it. It is the process to get information using tagging or annotation, lexical resources, NLP and the techniques like association, visualization, and prediction. Machine learning and natural language processing techniques extend (like dimension reduction, latent factor identification) its extraction capability at a much higher level[15]. Text mining is a little bit similar to data mining; both are applied to extract hidden information from the ocean of data by using machine learning techniques. Dissimilarity between them is the data source used for extraction. In data mining[18], [20], used data source is in structured form and mostly contains quantitative data easy to implement machine learning algorithms; while in text mining data source is unstructured or semi-structured and contains text.However, Natural Language Processing along with machine learning and other associative techniques are used to extract information from it.

3.2. Text Summarization

Automatic text summarization is a technique of text mining that produces an accurate, short and fluent summary of text which covers the important points of text documents and have size less than half of the size of original text[21]. It enables a reader to grasp the key points of a document without reading the whole contents of text documents[12].The process of text summarization of a document follows either of two approaches- extractive and abstractive approach[22]. Extractive methods rely on the statistical methods to compute the most important sentences / words of a document and place it in output summarized text[16]. While the abstractive approach uses NLP techniques to produce the summarized text. The obtained text may not present in input text documents but it covers the meanings of whole contents[23].

3.3. Natural Language Processing

It is a machine learning technology to enable computers to understand, generate and manipulate natural language. NLP is the branch of Artificial Intelligence and used in applications such as search engines, chatbots, spam detection, text-based scanning programs and many more[4]. NLP applied during data preprocessing generally in case of textual or voice contents. It breaks the textual contents in smaller size elements or removes irrelevant words. The major operations performed in NLP preprocessing are tokenization, stop word removal, lemmatization and stemming and POS tagging[6], [12], [21]. A model is developed to process the preprocessed dataset.

3.4. Classifiers:

The classification algorithm[24] of machine learning helps to categorize the samples in distinct class labels. It allows prediction classification models to recognize the features of input samples and assign them into predefined labels. Through it a sentiment or opinion can be predicted as a positive or negative in sense for a business or organizations. For the research work, the classification algorithms used are K-Neighbors, multinomial naive bayes, logistic regression, decision tree, random forest, SVM and gradient boosting. Following table shows a brief concept about the classifiers used in the model.

3.4.1 **K- Nearest neighbors**[25][26]is one of the easiest machine learning techniques applicable for classification and regression type of problems. It is a supervised, distance based, non-parametric classifier and takes no prior information about the distribution of data points in the dataset. It uses the similarity measures, exist amongst data, for classification. The Similarity measure, also known as the closeness or distances between data points, is calculated by a specific number of neighbors /data points which assign weights to the data points. KNN uses the weight as important tools to assign an outcome class label to data points. The output class label is a group of the neighbors which are the most nearest to data points.

3.4.2 **Logistic regression**[27] is a statistical, supervised machine learning algorithm used for classification problems. It analyzes the relationship between two data points using a sigmoid function taking inputs as independent variables and produces a value between 0 and 1. By default, logistic regression is used for binary class problems, however using some extensions it can be used for multiclass problems also.

3.4.3 **Multinomial Naive Bayes (MNB)**[7], [28] is a popular and efficient machine learning technique based on Bayes' theorem. It is used for classification problems with discrete features, such as text classification. Multinomial Naive Bayes is a probabilistic classifier for calculating the probability distribution of text data, making it suitable for data with features representing discrete frequencies or numbers of events in various natural language processing (NLP) tasks.

3.4.4 **Decision tree** is a supervised learning algorithm used for classification and regression modeling. Decision tree develops the model in the form of a tree structure, which consists of root node, internal node and leaf node which shows the attribute, decision role and the result of decision tree respectively [10].

3.4.5 **Random forest** is a supervised learning technique that builds a number of decision trees, where each decision tree is issued to represent a classifier. Every classifier predicts about one target class out of all the target classes. Prediction of a

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

class is based on the majority voting of the target class predicted by multiple classifiers. Due to the formation of a high number of decision trees, it generally produces high accuracy. Random forest is used for classification problems and produces a better result in comparison to other techniques due to the reduction in overfitting problems[1], [15], [26], [27].

3.4.6 **SVM** is a non-probabilistic supervised learning approach used to analyze data and put them in different class labels. SVM is used for a number of research works such as data analysis, pattern matching, and classification and regression analysis. SVM can be used for binary class and multiclass classification problems[15]. In a binary class classification problem, the prediction of outcome comes under the category of two class labels that is either 'Yes' or 'No'. SVM primarily determines a hyperplane or set of hyperplanes which categorize the data into distinct classes. There may be an infinite number of hyperplanes to divide the data space, but the hyperplane, which maximizes the margin that is distance between hyperplane and SVM data points, is the best hyperplane. Due to this, SVM searches out the hyperplane with the highest feasible margin. SVM can be used for sentiment analysis[1], [27].

4. WORKING METHODOLOGY OF SENTIMENT ANALYSIS MODEL

On the basis of experience, knowledge and collective or individual beliefs, feelings are expressed in different ways. Therefore, to understand the emotions transmitted through the text, a reliable mechanism is needed that is able to capture and model various linguistic patterns and events.

Sentiment analysis is an application of text mining which applies natural language processing and machine learning techniques to analyze a text (sentiment/review/opinion) into positive, negative or neutral class labels automatically[29]. Sentiment analysis is used to answer the question of what people do feel about a certain topic or a product or a service. The objective of it is to identify the high-quality information that lead a business to make correct decisions. In the research work, a sentiment analysis model is developed on a dataset downloaded from Kaggle[11], a public platform to provide dataset, for research work. In the model, several machine learning classifiers such as KNN, logistic regression, decision tree, random forest, SVM and gradient boosting algorithms has been used to classify the emotions of messages in positive, negative or neutral class labels. Our experimental results show that the proposed method is better than the most modern methods in the tasks of identifying emotions. The workflow of the model is given below briefly.

4.1 **Load the Emotion Dataset from Kaggle**:

The "emotion dataset for NLP", which is downloaded from a well-known public data source repository "Kaggle" website, is designed to support the research and experimental support for the researchers in the field of natural language processing [11]. The dataset contains 18000 samples with two features- text messages, and class features. The text samples of the dataset are diverse in nature and each labeled with the corresponding emotions that it has. The dataset of English messages are with six basic emotions: anger, fear, joy, love, sadness, and surprise.

Journal of Theoretical and Applied Information Technology
30th November 2024. Vol.102. No. 22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

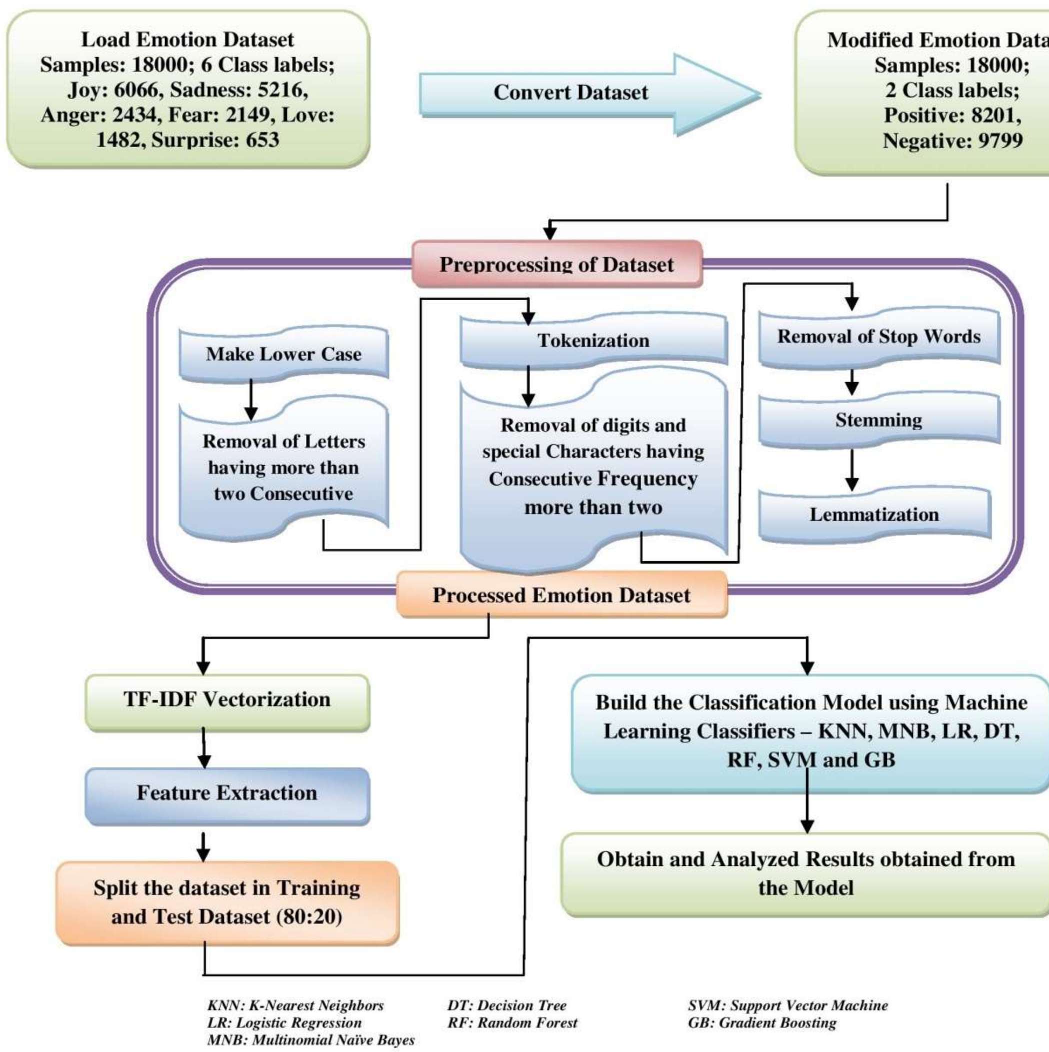

Figure 1: Working Methodology

The dataset is an example of a multiclass dataset and contains a different number of samples. Since the emotion "anger", "fear" and "sadness" are generally considered in negative sense and while the emotions "joy", "love" and "surprise" are in positive sense, so the input multi-class dataset is converted into a binary class dataset by labeled these emotions as either "Positive (1)" or "Negative (0)" [11]. The number of samples of each class label in the actual dataset and the samples in positive and negative, after conversions, are shown in the following figure "fig-1".

4.2 Preprocessing of Dataset

The model built on a correct dataset does produce a better result and when the dataset has only text, it becomes necessary to eliminate the noise and convert it in a proper format before the model implementation. Text Preprocessing is an important step for Natural Language Processing (NLP) tasks [15], [30]. It works on raw text and transforms it after removing superfluous information into a more digestible form so that machine learning algorithms can perform better. Some of the text preprocessing steps included in model for data preparations for the model are Tokenization, standardization of text (such as conversion of text into lower form, removal of digits and special characters, removal of letters having consecutive occurrence more than two times (eg: "joyyyy" to "joy") etc., removal of short words), removal of stop words and lemmatization [2][31].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

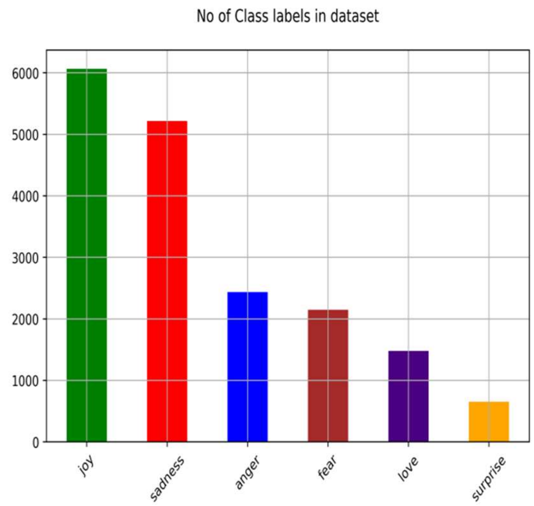

Figure 2(a):

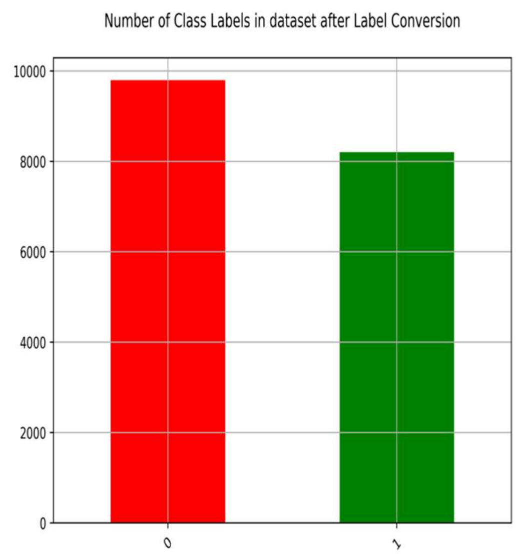

Figure 2(b):

4.2.1 Tokenization is the process of dividing the text into smaller size meaning components[1], [30]. Sometimes, a word has the multiple repetition of a single letter such as joyyyy, my lovvve etc.; which are actually not a part of the real word in an English dictionary, although used by the orator to emphasize those words in the sentences. These letters must be removed from the words and it makes sure that the frequency of such letters is not more than two[15].

4.2.2 Removal of stop words -Stop words are those English words whose appearances are so many in text but they do not add any meaning to sentences (eg- "the", "have")[14], [19] . They are safely removed without losing the meaning of the sentence. In the following figure, the most critical steps for data preparation for text analytics procedure are given.

4.2.3 Stemming: In text, a single word may appear in multiple locations in different forms as a verb, noun, adjectives etc.; for example go, goes, went, gone and going. These are the different forms of a single word "Go". To extract the emotions, it is necessary to identify the root meaning of a word used in a sentence. Stemming is the preprocessing technique to identify the roots of a word in a text [30].

4.2.4 POS Tagging: The process of tagging a word in a text according to their context and definitions is known as Part of Speech tagging (POS tagging). In this process the token words are categorized as a noun, verb, adjective, adverb or prepositions[2], [39]. For example, an adjective is used more to represent the emotions of an opinion. NLTK and Spacy are the most popular POS taggers used for this purpose in python implemented models.

Figure3: Word-cloud

The visualization of the occurrences of words of a textual content is called Word Cloud. It is a visualization technique to show the importance of a word in the text document. The word cloud shows the frequency of a word by its size in the image [26], [44]. A word, having a larger frequency, got a larger shape in the Word Cloud image while the infrequent words have smaller size or not mentioned in the image. A Word cloud also shows the relevance of words by their different formatting. Thus the more important words are bigger, bolder and centrally placed in comparison to the infrequent words [2]. The following figure shows the word cloud formed by the important words used in the sentiments of text documents.

4.3 TF-IDF Vectoriser

Feature selection from textual content is a challenging task for the classification. During the analysis work, a feature is represented as a unit to classify the polarity of sentences in sentiment analysis. A large number of available features does not ensure the classification task becomes easy, although it may cause the model to take more time for classification. For optimal performance, a model must have a small set of accurate and relevant features. A feature extraction technique maps textual contents into vectors which is the set of a word and its frequency. To extract information, hidden in text messages, understanding of text dataset/documents is necessary. TF-IDF is a weighted scheme that measures the importance of a token (word) in a text. The intuition behind TF-IDF is based on the frequency counts of a word. A dataset may contain low and high frequency count words, although TF-IDF considers lower frequency count words are more important to other higher frequency words[14].

In various NLP applications such as Sentiment Analysis, Term frequency is the way to represent more frequent features for information extraction. It considers uni-gram, bi-gram and trigram (a group of one or two or three words), generally known as n-gram words, with their terms counts as representing features. The presence or absence of a term gives a word value either "1" or "0". Term frequency, an integer value, is its count in the given document[18]. TF-IDF Vectoriser converts a collection of text into a matrix of TF-IDF features. The Vectoriser method is usually trained on only the train dataset.

4.4 Building Classification Model

In the proposed model, most of the supervised machine learning algorithms such as K-Nearest neighbors , multinomial Naive Bayes, Logistic Regression, Decision Tree, Random forest, Support vector Machine (SVM) and gradient boosting classifiers have been used to extract the emotions hidden in the textual sentiments of user. The featured dataset is divided into two parts 80% training dataset and 20% test dataset.

5 RESULT AND DISCUSSION

The performance of the proposed model is measured on the standard performance metrics such accuracy, precision, recall and F1 Score. The values of these metrics are calculated from the elements TP, FN, FP and TN parameters obtained from the confusion matrix. Accuracy is the ratio of correct predictions against the total predictions it made. Precision defines the ratio of number of correct positive predictions against the total positive predictions made by the model. Recall defines the ratio of the number of positive predictions made by the model against the total number of positive samples in the dataset. F1 Score is used to make a trade-off between precision and recall. Further an ROC curve is also drawn for all classifiers considering their TPR and FPR[14]. The confusion matrix of KNN, Multinomial Naïve Bayes and decision tree are shown in the following images generated by the python code of model.

In next page “Figure: 4” shows the confusion matrix of about the 3600 samples of the test dataset. The observation of confusion matrix shows that the number of correct prediction made by the KNN, Multinomial NB and decision tree are 3168, 3390 and 3330 respectively samples while 432 (195 FP, 237 FN) , 210 (50 FP, 160 FN) and 270 (154 FP, 116 FN) prediction are goes wrong. To improve the level of their services, if organization focus on only those customer who are unsatisfied from their service and made negative comments or shows negative thoughts, then they will consider about those comments which are actually negative or wrongly classify as positive instead of negative; i.e. False Positives (FP) classification by the model [38], [39]. Considering the False Positives (FP) and True Negative (TN) as important metrics, Multinomial Naïve Bayes is the best amongst these three classifiers, since it predicted only 50 comments as positive while KNN is the worst.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

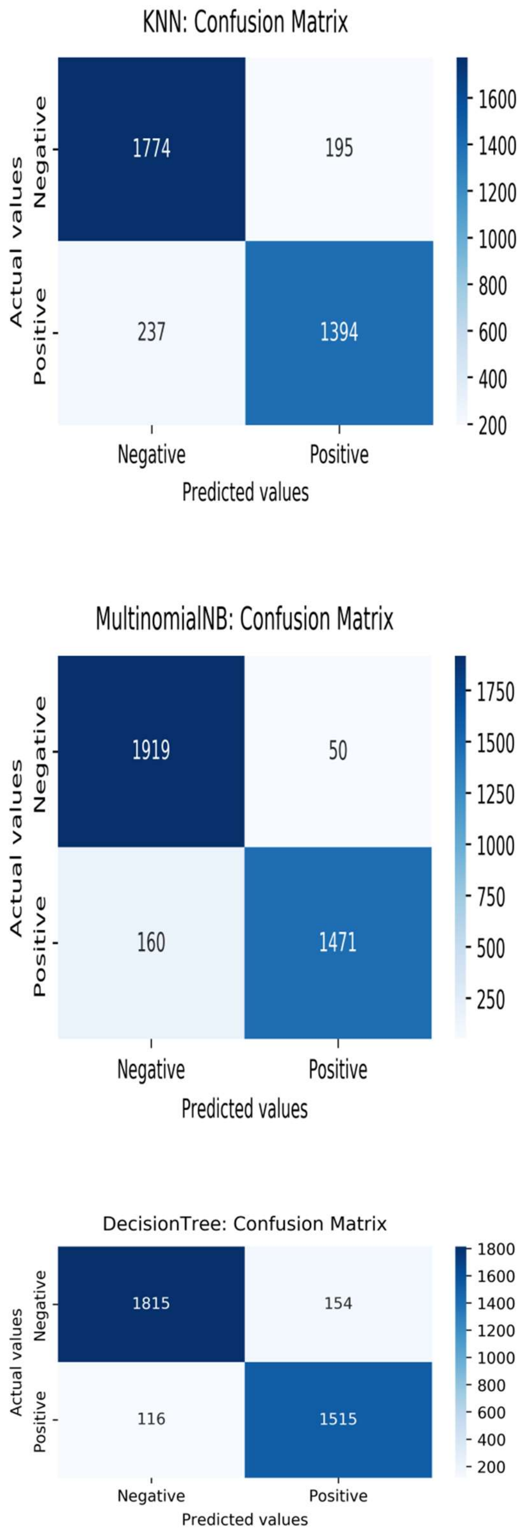

Figure 4: Confusion Matrix of Classifiers

\

The following table also shows the same for all classifiers used in the model.

Table 1: Confusion Matrix Score of Classifier

Classi-fiers	TP	TN	FP	FN
KNN	1394	1774	195	237
MNB	1471	1919	***50***	160
LR	1551	1897	72	80
DT	1515	1815	154	116
RF	1541	1879	90	90
SVM	1560	1911	***58***	71
GB	1014	1914	***55***	617

The above table “Table: 1” shows that the best three performing classifiers, considering their False Positive prediction as a selection metric, are Multinomial Naïve Bayes, Gradient Boosting and SVM since they have made only mistakes 50, 55 and 58 respectively in comparison to others. The following Table shows a comparative chart between the performances of classifiers on these different metrics.

According to the table: 2 shown in the next page, the model shows its best result through the Support Vector Machine classifier when it achieves the maximum value 96.42, 96.38, 96.42, and 96.15 for accuracy, f1 score, precision and recall respectively. The performance of SVM is followed by random forest and logistic regression tree classifiers. The same is also shown in detail by line chart for each performance metrics of all classifiers.

Journal of Theoretical and Applied Information Technology
30th November 2024. Vol.102. No. 22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Performances of Classifiers of Emotion Preprocessing Model

Classifier	*Accuracy*	*F1 Score*	*Precision*	*Recall*
KNN	*88.00*	*87.86*	*87.97*	*87.78*
MultinomialNB	*94.17*	*94.07*	*94.51*	*93.83*
Logistic Regression	*95.78*	*95.75*	*95.76*	*95.72*
Decision Tree	*92.5*	*92.45*	*92.38*	*92.53*
Random Forest	*95.00*	*94.96*	*94.96*	*94.96*
SVM	*96.42*	*96.38*	*96.42*	*96.15*
Gradient Boosting	*81.33*	*80.09*	*85.24*	*79.69*

The following Figure 5 shows the classification report of Logistic regression, SVM and Multinomial Naïve Bayes classifiers. The classification report shows the performance of a classifier with respect to each class label. The major metrics, it considers, are precision, recall, f1 score, accuracy which are calculated at macro-averaging and weighted averaging of each class-wise respectively [40].

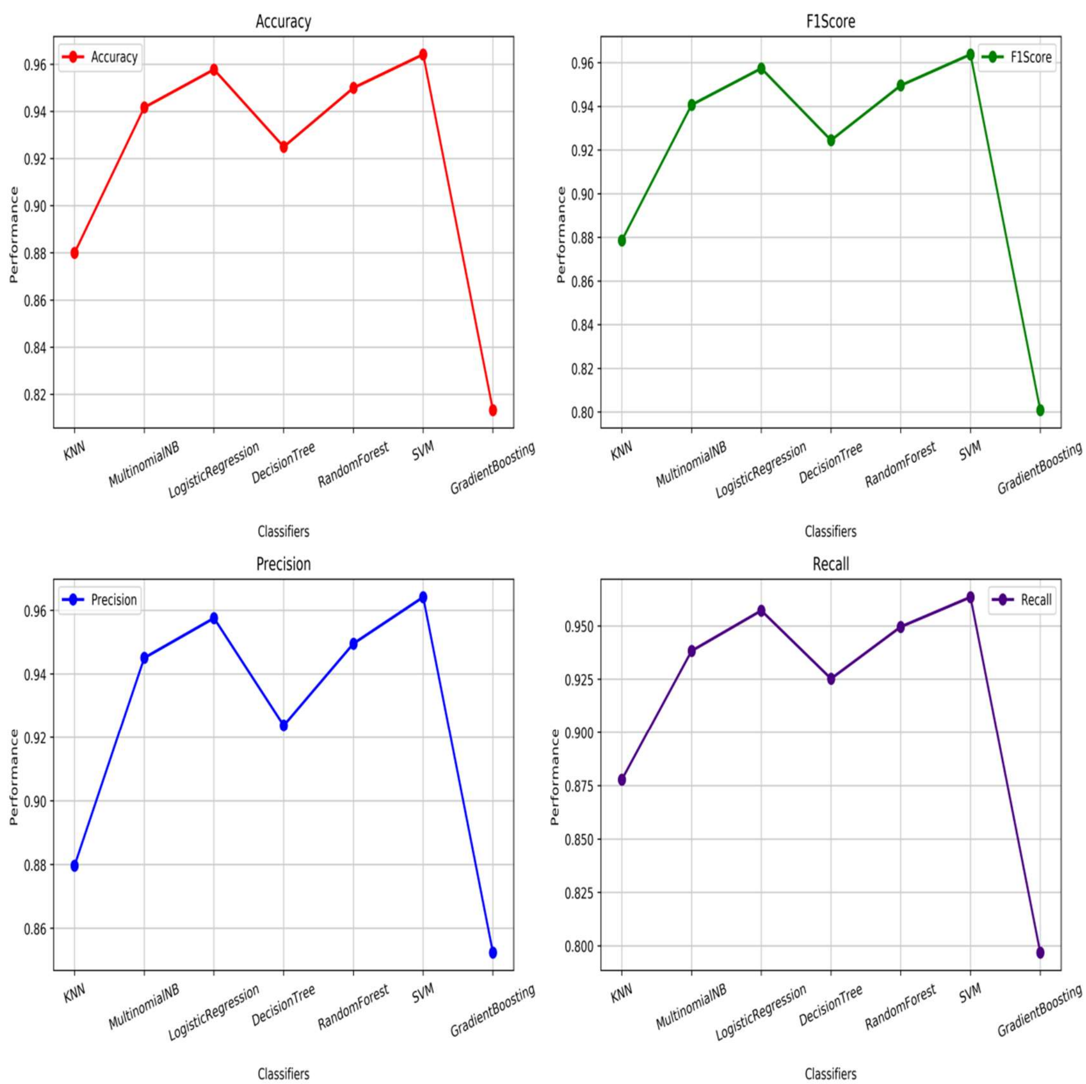

Figure 5: Metric score of the classifiers of Emotion Processing Model

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The macro-average value gives equal weight to each class ignoring the number of samples of each class. Weighted average is a way to calculate the score of every class by providing a weighted representation in the dataset. It is applicable when the dataset is imbalanced. As known, precision is the measuring the accuracy of positive samples against the total positive prediction made by the model while recall is the ratio of accuracy of positive prediction made by the classifier and total positive samples available in the dataset. A high recall indicates that the model is capable of catching the most positive samples of a dataset and a high precision means that most of the positive predictions made by the classifier are actually positive.

```
Classification Report of LogisticRegression
              precision    recall  f1-score   support

           0       0.96      0.96      0.96      1969
           1       0.96      0.95      0.95      1631

    accuracy                           0.96      3600
   macro avg       0.96      0.96      0.96      3600
weighted avg       0.96      0.96      0.96      3600
```

```
Classification Report of SVM
              precision    recall  f1-score   support

           0       0.96      0.97      0.97      1969
           1       0.96      0.96      0.96      1631

    accuracy                           0.96      3600
   macro avg       0.96      0.96      0.96      3600
weighted avg       0.96      0.96      0.96      3600
```

```
Classification Report of MultinomialNB
              precision    recall  f1-score   support

           0       0.92      0.97      0.95      1969
           1       0.97      0.90      0.93      1631

    accuracy                           0.94      3600
   macro avg       0.95      0.94      0.94      3600
weighted avg       0.94      0.94      0.94      3600
```

Figure 6: Classification Report of Classifiers

The above figure-6, classification report of KNN, Multinomial Bayes and Decision tree, shows that the test dataset is the instance of imbalance dataset having support 1961 for negative samples and 1631 for positive. The classification report of Logistic regression and SVM show that their macro and weighted precision , recall and f1- score are same and all are 0.96 while the precision, recall and f1 score all are 0.96 , 0.96 and 0.97 respectively for negative samples and 0.96, 0.95 and 0.95 respectively for positive samples. A comparative data of the classification report of all classifiers are given in the following table. ROC (Receiver operating characteristics) curve is a graph which measures the performance of a classifier by making a plot between true positive rate (TPR) and false positive rate (FPR). The more area occupied by the ROC curve, the better performance is shown by the classifier. In the early 1980s, ROC was mostly used in medical diagnostics, however it also got popularity to measure the performance of machine learning classifiers.

In the ROC curve, the TPR and FPR are shown on the x-axis and y-axis of the graph respectively. Each point of the ROC curve depicts the rate of TPR and FPR at each decision classification threshold. The scale goes from zero to one, where one is ideal for positive class labels and zero for negatives. A point on a ROC curve, closest to (0, 1), represents a range of the best-performing thresholds for the given model[20].

Table 3: A Comparative analysis of the score of classifiers According to each class label

		KNN	Logistic Regression	Multinomial Naïve Bayes	Decision Tree	Random Forest	SVM	Gradient Boosting
Accuracy		0.88	0.96	0.94	0.93	0.95	0.96	0.81
Negative Class Label	Precision	0.88	0.96	0.92	0.94	0.95	0.96	0.76
	Recall	0.90	0.96	0.97	0.92	0.95	0.97	0.97
	F1 Score	0.89	0.96	0.95	0.93	0.95	0.97	0.85
Positive Class Label	Precision	0.88	0.96	0.97	0.91	0.94	0.96	0.95
	Recall	0.85	0.95	0.90	0.92	0.94	0.96	0.62
	F1 Score	0.87	0.95	0.93	0.93	0.94	0.96	0.75
Macro Average	Precision	0.88	0.96	0.95	0.92	0.95	0.96	0.85
	Recall	0.88	0.96	0.94	0.93	0.95	0.96	0.80
	F1 Score	0.88	0.96	0.94	0.93	0.95	0.96	0.80
Weighted Average	Precision	0.88	0.96	0.94	0.93	0.95	0.96	0.84
	Recall	0.88	0.96	0.94	0.93	0.95	0.96	0.81
	F1 Score	0.88	0.96	0.94	0.93	0.95	0.96	0.81

The following image shows the ROC curve drawn for the classification model to analyze the classifier's performances. In the comparison of area under ROC curve, plotted by the different classifiers used in the model, SVM classifier outperforms the others although it is followed by logistic regression by a small margin.

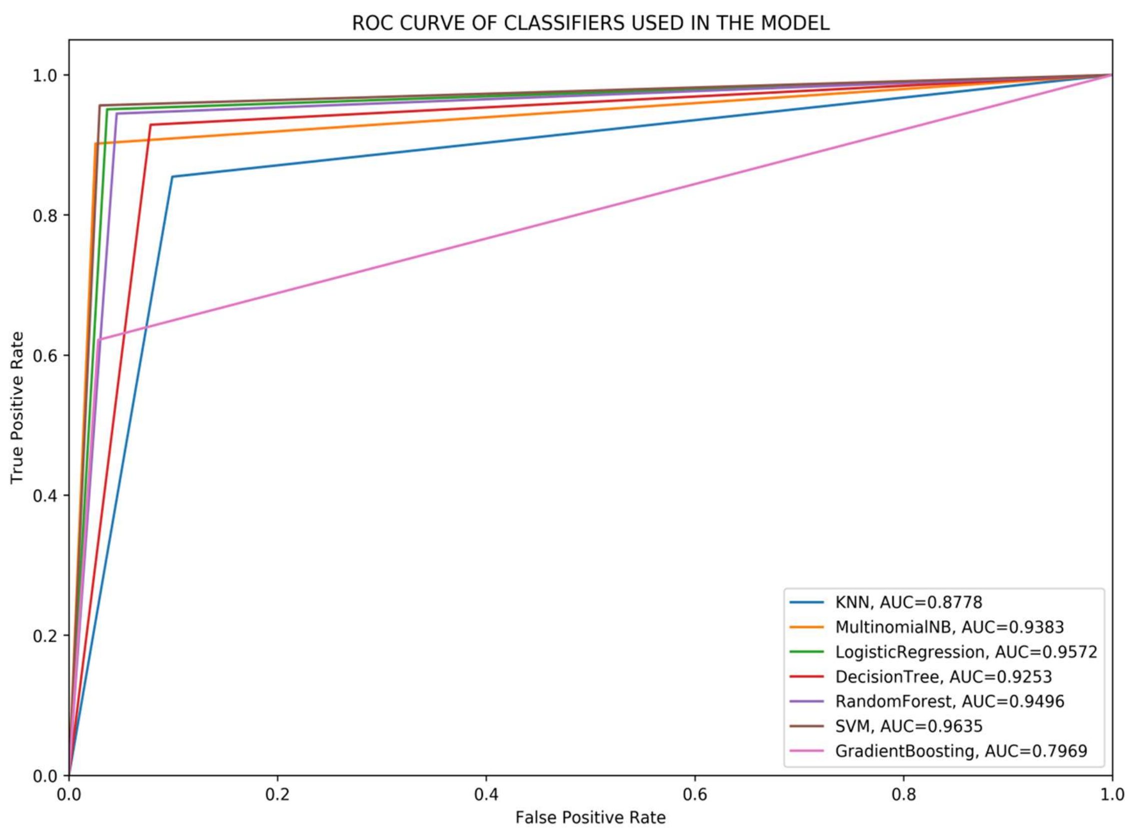

Figure 7: ROC Curve of Model

In the comparison of area under ROC curve, plotted by the different classifiers used in the model, SVM classifier outperforms the others although it is followed by logistic regression by a small margin. The AUC score of SVM and logistic regression classifiers are 0.963 and 0.957 respectively. Such a high value of AUC provides an indication that both classifiers produce optimal

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

results for the model. However KNN and gradient boosting are not doing well for the model and achieve only 0.877 and 0.796 respectively which is the lowest amongst all classifier's AUC.

The model has used NLP methods to preprocess the dataset and further used a number of classification techniques after applying TF-IDF vectorization method for feature selection. The classifier's performances has been judged not only on basic metrics such as accuracy, precision, recall, sensitivity and some advance metrics like f1 score and ROC but these has been also calculated individually for each and every class label and along with at macro and weighted average levels too. Although a lot has been tried to do in the study but these all are done on some selected techniques or methods. If some more classifiers, NLP based feature selection methods and Hypertuning of classifiers be included, the study would get comparatively a larger scope.

6 CONCLUSION

Extracting the correct sentiments from textual contents is not such an easy task due to its importance in today's competitive environments. Text mining provides such types of several techniques to make the task easy. In the research paper, a brief discussion about the text mining, summarization and sentiment analysis is made. As we have seen, sentiment analysis is a technique to deal with the different emotions or human feelings hidden in the textual comments and reviews. In this paper, a model is proposed to analyze the polarity of sentiments as positive or negative for an emotion dataset downloaded from the public domain.

The obtained results are analyzed and evaluated on different metrics such as accuracy, precision, recall, f1 score and AUC. SVM produces the best result for the model by scoring 96.42 % accuracy, 96.38% f1 score, 96.42% precision and 96.15 % recall. It also gets 96.35 % AUC score which indicates that it is doing very well for the model by classifying most of the positive and negative samples correctly. The next best classifier is logistic regression and it is followed by random forest, Multinomial Naïve Bayes and decision tree. The KNN and Gradient Boosting classifiers come in the last two positions respectively in the performance table.

Since the input dataset is a little bit imbalanced, so performances of classifiers has also been analyzed the macro and weighted level averaging. If it is considered the performance at weighted averages, which is better than macro- averages for an imbalanced dataset, LR and SVM shows the best results for the model and interestingly both have equal score (Precision, Recall and F1 Score 0.96, 0.96 and 0.96 respectively) of all metrics. These are followed by random forest, MNB and decision tree classifiers while KNN and gradient boosting are the last two performers of the model on all metrics.The model has made better effort for achieving the goal of enhancing the performance.

Limitation and future scope: The model uses TF-IDF technique to victories the samples for feature selection and it is a statistical approach. In future, deep learning techniques, hyper-parameter tuning of classifiers may apply for the same to enhance the extraction of hidden sentiments from emotion dataset.

REFERENCES

[1] M. Archana and T. Velmurugan, "Enhancing Sentiment Analysis in Electronic Product Reviews Using Machine Learning Algorithms," vol. 16, pp. 547–565, 2024.

[2] A. Hossain, M. Karimuzzaman, M. M. Hossain, and A. Rahman, "Text mining and sentiment analysis of newspaper headlines," *Information (Switzerland)*, vol. 12, no. 10. 2021, doi: 10.3390/info12100414.

[3] A. Taneja, "Sentiment Analysis Using Machine Learning : A Comprehensive Sentiment Analysis Using Machine Learning : A Comprehensive Review," no. July, 2024.

[4] R. Boorugu and G. Ramesh, "A Survey on NLP based Text Summarization for Summarizing Product Reviews," *Proc. 2nd Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2020*, pp. 352–356, 2020, doi: 10.1109/ICIRCA48905.2020.9183355.

[5] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 4, p. 102048, 2024, doi: 10.1016/j.jksuci.2024.102048.

[6] H. P. N, S. G. S, R. S. N, and A. H. S M, "Summarization of Customer Reviews in Web Services using Natural Language Processing," 2022, doi: 10.4108/eai.7-12-2021.2314556.

[7] Nikhil Sanjay Suryawanshi, "Sentiment analysis with machine learning and deep learning: A survey of techniques and applications," *Int. J. Sci. Res. Arch.*, vol. 12,

no. 2, pp. 005–015, 2024, doi: 10.30574/ijsra.2024.12.2.1205.
[8] K. Kothari, A. Shah, S. Khara, and H. Prajapati, "a Novel Approach in User Reviews Analysis Using Text Summarization and Sentiment Analysis: Survey," vol. 8, no. 3, p. 3067, 2020.
[9] N. C. Dang, M. N. Moreno-García, and F. D. la Prieta, "A Survey of Sentiment Analysis Based on Deep Learning," *Int. J. Comput. Inf. Eng.*, vol. 14, no. 12, pp. 473–485, 2020.
[10] T. State, "Hotel reviews on sentiment analysis for using nlp algorithms 1," vol. 03, no. 04, pp. 115–120, 2016.
[11] Kaggle, "Emotions dataset for NLP," 2020. [Online]. Available: https://www.kaggle.com/datasets/praveengovi/emotions-dataset-for-nlp.
[12] S. R. Misal, "Customer Reviews Sentiments Analysis Using Natural Language Processing (Nlp) And Deep Learning," no. December, 2023.
[13] X. Shen, "Sentiment Analysis of Modern Chinese Literature Based on Deep Learning," *J. Electr. Syst.*, vol. 20, no. 6s, pp. 1565–1574, 2024, doi: 10.52783/jes.3075.
[14] N. A. Semary, W. Ahmed, K. Amin, P. Pławiak, and M. Hammad, "Enhancing machine learning-based sentiment analysis through feature extraction techniques," *PLoS One*, vol. 19, no. 2 February, 2024, doi: 10.1371/journal.pone.0294968.
[15] Kavithasubramani, Nivedha U, L. Jabasheela, Divya S, and Dhuneesha E, "Secured Text-Based Emotion Classification Using Machine Learning With NLP Educational Administration: Theory And Practice," *Educ. Adm. Theory Pract.*, vol. 2024, no. 5, pp. 901–910, 2024, doi: 10.53555/kuey.v30i5.2986.
[16] S. C. Gupta and N. Goyal, "Text Mining: Techniques, Applications and Issues," no. 1, pp. 283–289, 2024, doi: 10.55524/csistw.2024.12.1.49.
[17] N. Goel, "A study of text mining techniques: Applications and Issues," *Pramana Res. J.*, vol. 8, no. 12, pp. 307–316, 2020.
[18] D. Antons, E. Grünwald, P. Cichy, and T. O. Salge, "The application of text mining methods in innovation research: current state, evolution patterns, and development priorities," *R D Manag.*, vol. 50, no. 3, pp. 329–351, 2020, doi: 10.1111/radm.12408.
[19] N. Bhanot, H. Singh, D. Sharma, H. Jain, and S. Jain, "Python vs. R: A Text Mining Approach for analyzing the Research Trends in Scopus Database," 2019.
[20] A. Kołcz, "Text Mining for Spam Filtering," *Encycl. Mach. Learn. Data Min.*, pp. 972–977, 2017, doi: https://doi.org/10.1007/978-0-387-30164-8_828.
[21] A. Sharma, "Natural Language Processing and Sentiment Analysis," *Int. Res. J. Comput. Sci.*, vol. 8, no. 10, pp. 237–242, 2021, doi: 10.26562/irjcs.2021.v0810.001.
[22] G. Sharma and D. Sharma, "Automatic Text Summarization Methods: A Comprehensive Review," *SN Comput. Sci.*, vol. 4, no. 1, 2023, doi: 10.1007/s42979-022-01446-w.
[23] P. Janjanam and C. H. P. Reddy, "Text summarization: An essential study," *ICCIDS 2019 - 2nd Int. Conf. Comput. Intell. Data Sci. Proc.*, pp. 1–6, 2019, doi: 10.1109/ICCIDS.2019.8862030.
[24] X. Luo, "Efficient English text classification using selected Machine Learning Techniques," *Alexandria Eng. J.*, vol. 60, no. 3, pp. 3401–3409, 2021, doi: 10.1016/j.aej.2021.02.009.
[25] S. Dang and P. H. Ahmad, "Text Mining : Techniques and its Application Text Mining View project Text Mining: Techniques and its Application," *IJETI Int. J. Eng. Technol. Innov.*, vol. 1, no. March, 2014.
[26] S. C. Gupta, D. K. Singh, and N. Goel, "Enhancing the Performance of Diabetes Prediction Using Tuning of Hyperparameters of Classifiers on Imbalanced Dataset," *Indian J. Comput. Sci. Eng.*, vol. 12, no. 6, pp. 1646–1662, 2021, doi: 10.21817/indjcse/2021/v12i6/211206049.
[27] S. Anis, S. Saad, and M. Aref, "Sentiment Analysis of Hotel Reviews Using Machine Learning Techniques," *Adv. Intell. Syst. Comput.*, vol. 1261 AISC, no. 1, pp. 227–234, 2021, doi: 10.1007/978-3-030-58669-0_21.
[28] P. S. P. Raut, S. D. Chaudhari, and V. B. Waghole, "Automatic Evaluation of Descriptive Answers Using NLP and Machine Learning .," vol. 2, no. 1, pp. 735–745, 2022, doi: 10.48175/IJARSCT-3030.
[29] A. M. Daniel, N. Firoozy, and Z. Qu, "A Review and Analysis of the Literature on Text Summarization and Sentiment Analysis for English and Russian," no. November, 2021.
[30] S. Mhatre, "Text Mining and Sentiment Analysis: Analysis with R," *Red Gate Hub*, pp. 282–290, 2020.
[31] M. Wankhade, A. C. S. Rao, and C. Kulkarni, *A survey on sentiment analysis methods,*

applications, and challenges, vol. 55, no. 7. Springer Netherlands, 2022.

[32] U. The and G. Of, "Sentiment Analysis of Product-Based Reviews Using Machine Learning Approaches," 2018.

[33] A. P. Widyassari *et al.*, "Review of automatic text summarization techniques & methods," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1029–1046, 2022, doi: 10.1016/j.jksuci.2020.05.006.

[34] V. Ramanathan and T. Meyyappan, "Twitter text mining for sentiment analysis on people's feedback about Oman tourism," *2019 4th MEC Int. Conf. Big Data Smart City, ICBDSC 2019*, pp. 1–5, 2019, doi: 10.1109/ICBDSC.2019.8645596.

[35] A. Hossain, M. Karimuzzaman, M. M. Hossain, and A. Rahman, "Text mining and sentiment analysis of newspaper headlines," *Inf.*, vol. 12, no. 10, pp. 1–15, 2021, doi: 10.3390/info12100414.

[36] P. Kunekar, "Restaurant Review Analysis using Natural Language Processing," pp. 4–6.

[37] J. Zhan, H. T. Loh, and Y. Liu, "Automatic summarization of online customer reviews," *Webist 2007 - 3rd Int. Conf. Web Inf. Syst. Technol. Proc.*, vol. SEBEG, no. EL/-, pp. 5–12, 2007, doi: 10.5220/0001266400050012.

[38] R. K. Bakshi, N. Kaur, R. Kaur, and G. Kaur, "Opinion mining and sentiment analysis," *Proc. 10th INDIACom; 2016 3rd Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2016*, pp. 452–455, 2016, doi: 10.1007/978-3-642-19460-3_11.

[39] M. T. Verma* and D. N. Singh Gill, "Email Spams via Text Mining using Machine Learning Techniques," *Int. J. Innov. Technol. Explor. Eng.*, vol. 4, no. 9, pp. 2535–2539, 2020, doi: 10.35940/ijitee.d1915.029420.

[40] E. Ho, M. Schneider, S. Somanath, Y. Yu, and L. Thuvander, "Sentiment and semantic analysis: Urban quality inference using machine learning algorithms," *iScience*, vol. 27, no. 7, p. 110192, 2024, doi: 10.1016/j.isci.2024.110192.