Journal of Industrial Information Technology and Application *ISSN(Online): 2586-0852*
Vol.7 No.2

Performance Analysis of Classification Methods for Sentiment Analysis using Customer Reviews based Text Data

P.C. Sridevi[1)], M.Archana[2)] and T.Velmurugan[3,*)]
[1)2)3)]Department of Computer Science, Dwaraka Doss Govardhan Doss Vaishnav College, Chania, Tamil Nadu 600106, India.

Abstract: Social media archives have enormous amounts of a wide range of unstructured data kinds. Among them are text data, audio, video, and visual media. It also includes sentiments, medical information, debate topics, and client testimonials. The data also includes client evaluations of goods and services. There are countless amounts of online reviews. Because of this, it may be challenging for a potential merchant to analyze them. It also makes it difficult for the product's creator to keep track of and manage user reviews. Sentiment analysis is a technique that helps with the challenging process by looking at the emotions expressed in so many online evaluations. Above all, sentiment analysis yields beneficial outcomes based on facts, enabling you to decide for your organization from the most important feelings present in social media. Sentiment analysis (SA) is a method of natural language processing that seeks to identify emotions related to a given topic and extract views about that topic from a vast corpus of data. The objective of this work is to examine the sentiment analysis technique like bag-of-word, Word distribution – inverse document frequency, Vader sentiment Analysis and evaluate the performance of classification algorithms for the analysis of twitter poco customer review sentiments. The classification techniques Tree Logistic Model Tree (LMT), Lazy Bayesian Rules (LBR), Hoefflin tree classifier, and Naive Bayes classifier are employed in this study work to examine consumer sentiments. This work determines which algorithms are most appropriate for the analysis of text data.

Keywords: Twitter Sentiment Analysis, Bag-of-Words method, Vader sentiment Analysis, Word distribution – inverse document frequency method.

1. Introduction

The two most essential logical substances in human existence are emotions as well as opinion. These opinions can be utilized to gather useful information that impacts

*Corresponding author: velmurugan_dgvc@yahoo.co.in
Received: Feb 27, 2023; Accepted: Mar 30, 2023; Published: Jun 30, 2023

decision making. The volume of data on the internet has expanded intensely since the origin of the Internet. Though this large amount of information is beneficial and the majority of which is in the form of online recourse, people have a difficulty in identifying the most relevant information or expertise. Various databases hold a huge amount of information that can be analyzed for product and customer reviews. This can be accomplished by reviewing client comments. A customer's attitude could be a decision or evaluation, an emotional experience, or the destined sentimental interaction. Customers are more likely to share their real feelings on social media or through the company's regular feedback system. As a result, social media and organizational data repositories are evaluated in order to extract a clear picture of the customer.

Twitter is a fantastic place to start for social web analytics because to its intrinsic openness for public consumption, easy and extremely well-written API, wealth of development tools, and attraction to people from all walks of life. Twitter offers a surplus of social media statistics. Tweets are incredibly fascinating while moving at the "speed of thought" because they represent the broadest spectrum of mankind on such a huge scale, are by their very nature complex, and are ready for consumption almost instantly. Through tweets and Twitter's "following" feature, users are connected in a variety of ways, from brief but frequently crucial conversation dialogues to interest graphs that relate back to people's interests in particular topics.

To mine opinions and derive features from an input source, techniques like NLP (Natural Language Processing), text analysis, and sentiment analysis are frequently utilized [2]. Therefore, the consumer review information that is available on social media platforms like Twitter is unstructured information that must be processed before it can be utilized by businesses to help them. Text categorization therefore helps to solve this issue. The process of extracting a collection of input documents is known as text analytics. Automatic sentiment categorization is the process of automatically classifying unclassified text materials. Because it retrieves high-quality information, text mining is a crucial word in data mining [3]. When unexpected content is entered into the system, it automatically assigns it to the category that best fits the content. Large-scale data classification is necessary for effective archiving. The categorization model and text representation are the main areas of research in text mining. In accordance with this, this study will first introduce the hefting, tree LMT, Naive Bayes, and LBR classic models.

A method for collecting corpora completely automatically that can be used to train a classification model has been extensively studied. This research work thus clarifies the preprocessing, classification, summarization, and evaluation steps involved in the preparation of raw text with opinions acquired from social media sites for presentation to top management. In order to emphasize the value of preprocessing for any sentiment evaluation of text content records, the research work addresses several trials with

applying this to pertinent facts tools. A literature review is found in Section II, study materials and procedures are covered in Section III, experiment results are covered in Section IV, and the conclusion is found in Section V.

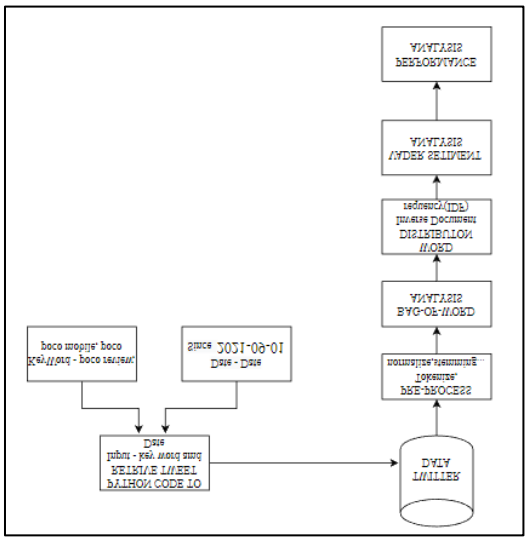


Figure 1. The process flow of sentiment analysis

2. Literature Survey

Various approaches for categorization of documents in Twitter for opinion mining have been proposed. It's been such a commonly utilized field for decades at least, and there is always more to be examined into the prognostic factors of Twitter material. Having provided several methods for data organization, this work provides a fairly wide review of prior work and is covered throughout this section. Here as result, this section examines some of the works concerning the text categorization and categorization of Social media and Twitter text data. In the investigation done by Mean M.G. Younis [4] expresses that Text analysis and sentiment classification for Twitter feeds would be used to evaluate consumer evaluations for goods or services. The author Salton G et al., [5] suggested that automatic extraction methodologies outperformed traditional extraction in terms of sentiment analysis. Maulana et al., [6] work aims to anticipate the length of a student's study session. The research was done on student data set and determined that the Decision Tree J48 and LMT algorithms generate a tree structure that would be used to, categories and forecast the duration of study of college students. The Decision Tree J48 method delivers 62.8 percent validity of the data mining modeling, whereas LMT produces 71% accuracy.

Table 1. COMPARISON OF SURVEY RESULTS

S. No	Names of Authors	Data Set Used	Methods Applied	Results
1	Maulana, Mohamad et al.,	Student academic data	Decision Tree J48, Decision Tree J48	68% Accuracy for Decision Tree and 71 % LMT
2	Valliant Ramanathan et al.,	Twitter	Naïve Bayes	With existing lexicon the accuracy is 66.62 ,for domain specific ontology it is 72.77 and with proposed combined lexicon the accuracy is 85.54
3	Sayyed Mohammad et al.,	BBC News Data set	SVM-NN(Hybrid of SVM and NN)	Precision for BBC new feeds is 97.84 and 20 Newsgroup is 94.93
4	Alec Go et al.,	Twitter dataset	SVM, Maxent, Naïve Bays Twitter	Naïve Bayes 82.7, Ent 82.7 %, SM 81.6 in terms of bigram and 81.3%, 80.5%, and 82.2% in terms of unigram
5	Arnaiz Zubeida et al.,	Twitter dataset	SVM	Precision for news 0.634, Ongoing event 0.829, Memes 0.731, Commemoratives 0.132
6	Prathima Singh et al.	Demonetization tweet Data set	Analysis based on scoring method	Analyzed the effect of the Demonetization and concluded that excluding the hurdle the people were happy with Demonetization
7	Ahuja, Ravinder, et al.,	SS- Tweet data set	SVM,LR, Naïve Bayes, Random Forest, Decision tree, K-NN	When compared to N-Gram features, TF-IDF features produce superior results (3-4%).

Researchers Valliant et al., [7] proposed a new technique for Twitter customer feedback using the Nave Bayes classification approach. The information retrieved from lexicons, domain specific ontologies, and a new ontology which is a combo of domain - independent and lexicon ontologies is compared. In the end it is concluded, the novel technique outperformed the conventional method. Dadar [8] proposed a novel method that integrated the SVM and nearest neighbor methods in classifying the BBC and 20 Newsgroup news feeds, and the evaluation result shows that the novel method produced a precision rate of 97.84 for BBC new feeds and 94.93 for 20 Newsgroup. Alec Go et al.,[8] proposed a method for classifying tweet data that was using machine learning algorithms such as SVM, Maximum Entropy, and Nave Bayes and found that the accuracy was 80 percent or higher when trained with emotion data. They also suggested a method for processing Twitter data. Arkaitz Zubeida et al., [10] proposed a method to classify the tweets based on their topology of topic and they evaluated their topology using the SVM classifier.

Prathima Singh et al., [11] used a sentiment analysis approach and Twitter data to analyses the government policy of demonetization from the perspective of regular people. Tweets are collected using a specific #demonetization hashtag based on spatial location,

which is a collection of tweets by state. The emotional analysis Tool from Meaning Cloud was used to sort the states. In the work by Ahuja, Ravinder, et al., [12], the frequency-inverse document frequency (also known as TF-IDF) is a well-known approach for determining the relevance of a word in a text. The term frequency of a specific term (t) is determined as the ratio of the number of times a term appears in a document to the total number of words in the document. IDF stands for Inverse Document Frequency.

3. Material And Methods

Natural language processing systems are used to educate machines well how to analyze, comprehend, and produce content. As a branch of NLP text analytics uses categorization as being one of those approaches. For categorization supporting methods are needed to carry out the categorization in a healthy was there in subsequent sections, we will go over each of the techniques utilized in the Twitter-based customer review of twitter poco c31 categorization.

A. Data Set from Twitter Tweets

Jack Dorsey founded Twitter [13] in 2006 as an online social networking and micro blogging service that allows users to send and receive brief messages (called tweets) of up to 280 characters. Tweeter's 280-character message limit caters to today's busy people's desire for information in a concise and timely manner. Reading brief tweets contains far less meaningless minutia. People can spend over 5 - 10 minutes on Twitter to learn about what is going on in the world. As a result, over 2 decades, Twitter has evolved to be one of the most popular social networking sites, with 500 million tweets shared every day on average. This translates to 6,000 tweets each second, 350,000 tweets per minute, and almost 200 billion tweets per year. Extract from the Twitter website with the Twitter API the "raw" data that was collected. Among numerous data this research work uses the dataset known as twitter poco c31 customer review Dataset which is downloaded from developer account using twitter API. Below is the sample input given to the code so that the data containing poco mobile review from 2021-01-01.

Enter Twitter Hash Tag to search for

Poco

Enter Date since The Tweets are required in yyyy-mm--dd

2021-09-01

B. Preprocessing

Twitter account information the following twitter post about something like a product review of poco mobile is taken from twitter using twitter API. This work collects 3000 pieces of data. Information has noise, and because a data preprocessing step is required for classification task. This research employs annotation to handles basic cleanup processes, including such expelling rare circumstances or jarring components for the subsequent phases of analysis and legitimizing a few misspelt words. In attempt supply hardly necessary details, a smooth tweet doesn't include URLs, hash tags (e.g., #Flight), or mentions. Tabs and line breaks also should be supplanted to blanks, and quotation marks with apexes. Ever since this phase, all punctuation is excluded except apexes, which are part of grammar constructs like the genitive.

C. Bag of Word Analysis

Text modeling is difficult because it is unstructured, and approaches such as machine learning algorithms demand well-defined stationary inputs and outputs. Algorithms for machine learning typically deal closely on raw text; the content needs to be turned to figures and number vectors, to be specific. A bag-of-words model, abbreviated as Bow, is a method of extracting textual characteristics to be used in analysis, like machine learning approaches. The method is quite basic and adaptable, and it may be used this to glean data from texts in a variety of ways. A bag-of-words is a text representation that specifies the appearance of words in a document. It entails two steps, which are 1) A list of well-known terms and 2) A metric for the presence of well-known terms.

D. Word Distribution – Inverse Document Frequency (Idf)

In Information Extraction, inverse document frequency (IDF) is often utilized [14]. IDF typically given by -log2 dfw/ D, whereby D is just the collection size while df w is indeed the document rate, or the number of documents that contain w. There has been definitely a substantial link between document frequency, dfw, and word frequency, fw. Figure 1 depicts the association between 10glO fw and IDF for 193 words chosen from a 50-million-word corpus of 1989 Associated Press (AP) Newswire articles (D = 85, 432 tales). IDF, which quantifies the in formativeness of a term t (word), is the inverse of document frequency. Because stop words are used so frequently in texts and because N (count of corpus)/ df (document frequency) assigns such a low value to them, their IDF will be quite low when we calculate IDF for them. This ultimately provides a relative weight, which is what we desire.

$$idf(t) = \frac{N}{df} \tag{1}$$

There are a few other issues with the IDF, such as the IDF value exploding for huge corpus sizes, like N=10000. Therefore, we use the IDF log to lessen the effect. When a

word is not in the vocabulary at inquiry time, it will simply be disregarded. But occasionally, when we employ a fixed vocabulary and only a few of the vocabulary terms are present in the text, the df will be 0. We smooth the value by adding 1 to the denominator because we are unable to divide by 0.

$$idf(t) = \log(N/(df+1)) \tag{2}$$

TF-IDF score is obtained by multiplying the values of TF and IDF. TF-IDF comes in a wide variety of versions, but for the time being, let's focus on this fundamental one.

$$tf - idf(t,d) = tf(t,d) * \log(N/(df+1)) \tag{3}$$

E. Vader Sentiment Analysis

In the work carried out it uses Vader sentiment analysis for analysis sentiment for each tweet. This method was created by Gilbert [15], VADER (Valence Aware Dictionary for Sentiment Reasoning), a basic rule-based approach for sentiment analysis in general. In the work for social media sphere, the VADER lexicon functions exceedingly well. The correlation coefficient demonstrates that VADER (r = 0.881) matches ground truth as well as individual human raters (r = 0.888). (Aggregated group mean from 20 human raters for sentiment intensity of each tweet).

Following that, the work uses a scoring rule to assess whether the overall sentiment polarity in each tweet fell into one of five categories: high positive, positive, neutral, negative, or high negative. The scoring rule is utilized in the existing method to categories tweets into five sentiment classes, as follows: Examine the general tone of the tweet. If the score value is 1, calculate the overall tweet polarity as: If the positive value is greater than 0.5, assign tweet polarity = +2. Otherwise: (positive value 0.5) polarity of tweet = +1. If (score value) equals -1: The overall tweet polarity is calculated as follows: If (negative value > 0.5), assign tweet polarity = -2. Otherwise, (negative value 0.5) set tweet polarity to -1. Assign tweet polarity = 0 if (score value = 0). To sum up the value are assigned as Positive sentiment: (compound score >= 0.05), neutral sentiment: (compound score > -0.05) and (compound score < 0.05), negative sentiment: (compound score <= -0.05).

F. Classification Technique

The selection of appropriate pre-processing methods proves to be an important step in enhancing classification's accuracy. The goal of this work is to classify tweet sentiment using the for models the Tree Logistic Model Tree (LMT) model, the Lazy Bayesian Rules (LBR) model, the Hoefflin tree classifier, and the Naive Bayes classifier.

Logistic Model Tree (LMT): A Logistic Model Tree (LMT) [16] is like a normal regression model; however, the leaves have multinomial logistic features. Each internal node, such as in classical decision trees, has been connected with just a test on one of

the features. The node really does have t child nodes for a nominal attribute with k values, as well as occurrences are managed to sort beneath several of the t branches based just on feature values. [17][18]Its entity has children, and the experiment consists of comparing the variable to a criterion; an instance is organized down the left group only if the value for such an attribute is less than the criterion, and organized down the right branch otherwise. In much more technical terms, a logistic regression tree is a binary tree composed of a series of internal or non-terminal nodes N and a set of leaves or terminal nodes T. The LMT [19] [20] is the combination of decision tree with regression function and it is represented as below.

S denotes the Total Instance.
T terminal nodes and t belongs to T
St Sub division of S
I(x ∈ St) is 1if x ∈ St else 0.

$$a^{j}_{vk}=0 \text{ for } vk \notin Vt.$$

$$s = \bigcup_{t\in T} s_t \quad St \cap S_{t'=\emptyset \text{ for } t\neq t'} \tag{4}$$

For a given model class probability and fj (x) the model is represented as

$$Fj(x) = a^{j}_{0} + \sum_{k=1}^{m} a^{i}_{k}.vk \tag{5}$$

$$f(x) = \prod_{t\in T} ft(x) \cdot I(x \in St) \tag{6}$$

Naïve Bayes: The Naive Bayes classifier performs magnificently in some kind of a variety of complex application areas, including character recognition [21, 22], key retrieval [23] and clinical diagnosis [24]. Bayesian methods [25] provide a natural and principled way of combining previous information to records inside a sturdy decision conceptual framework model, through comparison to conventional techniques. A previous proportion for future work can indeed be formed by incorporating historical knowledge about such a factor. It's [26] possible to think of the Bayesian Naïve Bayes Classification algorithm as a production sequence. The Bayes Rules [27] specify that the probability of just an instance is given as G = (a1, a2, a3....an), where a1, a2.... an, are attribute values in a tuple.

Let V denote the classification parameter, and v represent this same variable's value. There are three classes in this paper: positive, negative, and neutral.

$$p(v|G) = \frac{p(G|v)p(v)}{p(G)} \tag{7}$$

If and only if,

$$f_b(G) = \frac{p(V=+|G)}{p(V=-|G)} \geq 1 \tag{8}$$

If it is classified as class V= +.
$f_b(G)$ is the Bayesian classifier. And if all the classifier are independent

$$p(G|v) = p(a1, a2, \cdots, an|V) = \prod_{i=1}^{n} p(a_i|v) \tag{9}$$

Then the NB $fnb(G)$ classifier is given as

$$\text{fnb}(G) = \frac{p(V=+)}{p(V=-)} \prod_{i=1}^{n} \frac{p(a_i|V=+)}{p(a_i|V=-)} \quad (10)$$

HOEFFDING: The Hoefflin bonded technique for estimating a confidence threshold for an unknown class iteratively [28]. Incremental induction algorithms are based on Decision tree. All of this takes a while to become able to use the data. [29] It uses the Hoefflin confined to determine how many examples of an instance are required to reach a certain level of confidence. This bonded claims that perhaps the true mean of the variable is at least $\bar{r}$- $\in$ to probability 1-δ and therefore is determined using the formula below.

$$\in = \sqrt{\frac{R^2 \ln 1/\delta}{2n}} \quad (11)$$

R is the range with r
The real-valued random variable is r.
The variable n is the independent observation. And its mean value is $\bar{r}$

LAZY BAYESIAN RULES (LBR): LBR is a Nave Bayes Classifier (NB) variant [30]. The collection in NB Tree represents the set of test attributes along the NB Tree's path to the leaf, whereas the collection in LBR represents the feature extractor in the rule's antecedent. For each training sample, a Bayesian rule is created lazily by LBR [31]. The cause of a Bayesian rule is in fact a set of attribute-value pairs, and the rule's unavoidable effects are in fact a local NB that also makes use of the features that weren't included in the predictors. LBR calculates the value of P(y, x).

$$\hat{P}(y, x) = \hat{P}(y, q) \prod_{i \in s} \hat{P}(xi \,|y, q) \quad (12)$$

4. Experimental Results

In this study, the bag-of-word, word distribution - inverse document frequency (idf), Vader sentiment analysis and classification technique are used on the poco mobile review data set. The same is stated in the next section.

A. Tweeter Data Set

The description of the twitter data set chosen for this analysis is explained. It is a twitter based flip kart poco mobile Customer review containing a weight of 3000 dataset. With the attribute like username, description, location, following, followers, total tweets, re-tweet count, text, hash tags. Among which username, text, has tags alone considered for this research work. The table 2 shows the sample feed taken for analysis. Table 3 shows the opinions again for posts on twitter together with their weight.

Table 2. EXAMPLE TWEETS FOR EACH SENTIMENT

S. No	TEXT	HASHTAGS	SENTIMENT
0	@IndiaPOCO, Can we use 33 Watt fast Charger on poco C31? Your regular charger is taking 3hr from 0 to 100+ charges....”	#POCO C31 (64 GB) (4 GB RAM)	Neutral
1	@flipkartsupport @Flipkart I have ordered a poco c31 but have received some powder and some stickers. I have trying to solve the issue for last 15 days, but replacement is getting cancelled frequently. please help in this regard https://t.co/GPsoVL1qBh	#POCO	Negative
2	ðŸ”¥ [Plus Member only]	#POCO C31 (64 GB) (4 GB RAM)	Neutral
5	Too much hangingflies, request filpkart for return. Battery backup is very bad. Quality. @POCOSupport @IndiaPOCO @Flipkart	#POCO C31 (64 GB) (4 GB RAM)	Negative
6	value for money recommend poco c31	#POCO C31 (64 GB) (4 GB RAM)	Positive
7	#poco can't believe poco brand is a bad quality brand. I buy poco c31, after buy i notice mobile is too hanged. Camera is very low And I request filpkart for return. Battery backup is very bad. Quality. @POCOSupport @IndiaPOCO @Flipkart	#POCO	Negative
8	Fashioner Printed Soft Silicone Designer Pouch Mobile Back Cover for Poco C31 / Mi Redmi 9 / Redmi 9 Active Case and Covers | for Boys & Girls -P028 - Multi-Colored https://t.co/dPnpRNN2T6 via @amazon https://t.co/YcvfmfOClw	#POCO C31 (64 GB) (4 GB RAM)	Neutral

B. Preprocessing

Tweets are cleaned during the preprocessing step. First, the emoticons are recognized and eliminated from the text using regular expressions. We also remove any links, URLS, or user names that can be identified since their initial character is the symbol @. We do not make changes to words that begin with hash tags (that is, the symbol #) because they can be directly related to the topic of the text. Finally, we lowercase the words and eliminate all non-letter characters and stop words found in tweets. Below are the pre-processing work carried out on a sample tweet data set.

Table 3. PREPROCESSING APPROACHES

Preprocessing method	Description	Text Before preprocessing	Text After Preprocessing
URL Removal	In tweets, users include a URL to provide context for the text, such as “https://t.co/dPnpRNN2T6 via @amazon https://t.co/YcvfmfOClw". During sentiment analysis, these URL links become noise data. These URL need	#Poco can't believe poco brand is a bad quality brand. I buy poco c31, after buy I notice mobile is to hang. Camera is very low And I request filpkart for return. Battery backup is very bad. Quality. @POCOSupport	#Poco can't believe poco brand is a bad quality brand. I buy poco c31, after buy I notice mobile is to hang. Camera is very low And I request filpkart for return. Battery backup is very bad. Quality. @POCOSupport

Preprocessing method	Description	Text Before preprocessing	Text After Preprocessing
	to be removed before it is to be used for preprocessing.	@IndiaPOCO @Flipkart https://t.co/dPnpRNN2T6	@IndiaPOCO @Flipkart
Username Removal	I.There're many usernames in tweets which begin well with symbol "@," such as "@Hepburn," where the symbols appear to suggest the user id or identify user, to be removed.	#Poco can't believe poco brand is a bad quality brand. I buy poco c31, after buy I notice mobile is to hang. Camera is very low And I request filpkart for return. Battery backup is very bad. Quality. @POCOSupport @IndiaPOCO @Flipkart	#Poco can't believe poco brand is a bad quality brand. I buy poco c31, after buy I notice mobile is to hang. Camera is very low And I request filpkart for return. Battery backup is very bad. Quality.
Hash Tag	II.Hash tags with sign "#" indicate that posts were also connected to a particular topic and also include statements made inside the twitter posts. Just the symbol "#" was deleted, leaving the contents intact.	III.#Poco can't believe poco brand is a bad quality brand. I buy poco c31, after buy I notice mobile is to hang. Camera is very low And I request filpkart for return. Battery backup is very bad. Quality.	V.can't believe poco brand is a bad quality brand. I buy poco c31, after buy i notice mobile is too hang. Camera is very low And I request filpkart for return. Battery backup is very bad. Quality.
Character normalization	In tweets, phrases with successive characters, such as "takeeeee," are much more usual. The term "consecutive characters" refers to characters featured and over three times in some kind of a word. To give a formal representation, this must be normalized.	I'm flying your #fabulous #Seductive skies again! U takeeeee all the #stress away from travel http://t.co/ahlXHhKiyn	I'm flying your skies again! U take all the from travel
Punctuation	V.Those punctuation signs, such as ",","#","$"," % ","*", "?","/," etc., must be omitted since they have no impact on the sentiment of posts on twitter.	VI.I'm flying your #fabulous #Seductive skies again! U takeeeee all the #stress away from travel http://t.co/ahlXHhKiyn	I'm flying your skies again U take all the from travel
Stop-words	Stop-words are perhaps the most commonly utilized words in twitter posts, including all, an, the, a, as, be, for, and so on. We got rid of such English stop - words because they don't add much to the sentiment of the twitter post.	can't believe poco brand is a bad quality brand. I buy poco c31, after buy i notice mobile is to hang. Camera is very low And I request filpkart for return. Battery backup is very bad. Quality.	II.can't believe poco brand bad quality brand. buy poco c31, notice mobile hang. Camera is low request filpkart return. Battery backup bad. Quality.
Stemming	A next procedure is to extract base from a word normalizes the words: for instance, flags need to be written as flag are becoming identical.	I'm flying your #fabulous #Seductive skies again! U takeeeee all the #stress away from travel http://t.co/ahlXHhKiyn	fly sky travel

Finally, the tweet which we obtained from twitter "#poco can't believe poco brand is a bad quality brand. I buy poco c31, after buy i notice mobile is too hang. Camera is very low And I request filpkart for return. Battery backup is very bad quality.

@POCOSupport @IndiaPOCO @Flipkart ". Will be preprocessed and the result is given as poco brand bad quality hang camera low battery backup bad return bad quality. Thus the preprocessed text will contain data that is much required for analysis.

C. Bag of Word Analysis

Bag of words characteristics can represent data or tweet classes [32]. Preprocessing was performed prior to the extraction of Bag of Words features, which included data cleaning, stemming, filtering, and tokenization. Since any detail about the sequence or structure of words in the document is deleted, it is referred to as a "bag" of words. The system is mainly worried as whether recognized terms appear in the document, not about where they appear. Word Cloud for the poco filpkart review is obtained before and after preprocessing and is presented in the below table. From the image that is the word cloud it is evident that the pre-processed bag-of-word contains word that is much useful for further work.

Table 4. SAMPLE WORD LIST BEFORE PREPROCESSING

@, ., poco, #, indiapoco, himanshu_poco, ,, to, is, my, pro, 5g, pocodiwalimadness, t, x4, win, wish, answer, /, co, https, ://, i, the, ', and, :, stereo, speakers, a, you, for, -, in, this, of, on, dual, it, camera, phone, join, s, not, congratulations, ", me, pocosupport, with, (, update, after,), are, but, tag, x3, that, like, flip kart, m4, no, un, joined, don, akhil_slimshady, dead, lee, !, x2, ', storage, they, if, pocoglobal, ?, all, internal, ;, from, at, please, have, sir, charging, now, 1, will, ..., your, what, xiaomi, glass, issue, &, service, working, one, am, or, we, sonic, so, 0, problem, buy, ",, by, phones, mi, can, .., was, get, Redmi, wide, be, make, got, free, new, tagging, mobile, ram, do, amp, Maui, help, just, 3, 13, same, ois, dance, 2, as, he, need, snapdragon, f4, samim_busy, any, Billy, front, time, get, good, best, de, also, more, 12, when, m, gorilla, there, back, out, many, using, friends, thank, well, why, mamuddct, go, merch, 5, motherboard, version, m5, quakily, poco_lee21, worst, system, bad, guys, has, mother, want, winner, amole, price, people, gab, day, only, know, ultra, u, review, updated, Pocono, than, rt, najim_msd, steady, then, love, liquid, cooling, us, very, facing, center, which, 🎉, software, 4, design, still, bro, how, device, year, android, money, green, ₹, angle, too, 🔄, follow, dear, tags, issues, vahgar, thanks, xiaomiindia, day, 9, give, centre, today, pay, realme, !!, speaker, 10, its, say, even, check, been, about, sale

The list of words in the 3000 tweets that were located is in table 4, and it includes every word that could be found. There are 9549 words in total among these 3000 tweets.

Table 5. WORD LIST AFTER PREPROCESSING

5g 5g, 5g indiapoco, 5g poco, 5g pocodiwalimadness, 5g pro, 5g wish, answer 5g, answer answer, answer indiapoco, answer poco, answer pocodiwalimadness, answer speaker, answer stereo, answer wish, indiapoco answer, indiapoco indiapoco, indiapoco poco, indiapoco pocodiwalimadness, indiapoco pro, indiapoco speaker,

indiapoco stereo, indiapoco win, indiapoco wish, poco 5g, poco answer, poco indiapoco, poco poco, poco pocodiwalimadness, poco pro, poco speaker, poco win, poco wish, poco x4, pocodiwalimadness answer, pocodiwalimadness indiapoco, pocodiwalimadness poco, pocodiwalimadness pocodiwalimadness, pocodiwalimadness win, pro 5g, pro poco, pro pocodiwalimadness, pro pro, pro speaker, pro stereo, pro win, speaker indiapoco, speaker poco, speaker pocodiwalimadness, speaker speaker, speaker win, speaker wish, stereo speaker, stereo wish, win poco, win pocodiwalimadness, win pro, wish indiapoco, wish poco, wish win, wish x4, x4 5g, x4 poco, x4 pro, x4 x4, Charging,, hate, bad, happiness , good , fantasy, dance, dead

Table 4 contains a list of words that were found in the 3000 tweets that were found, including words that had been preprocessed. These 3000 tweets include 138 words in total.

Figure 2. Bag of Word Analysis before preprocessing

The word cloud from the bag of words analysis is shown in figures 2 and 3. The word cloud is shown in figure 2 without any processing. Similar to that, after preprocessing, the figure 3 includes a word cloud.

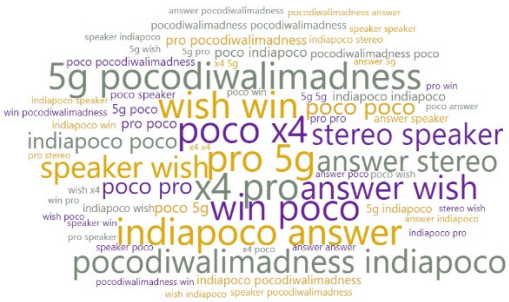


Figure 3. Bag of Word Analysis after preprocessing

In order to identify the word instances in the corpus used for research, the bag of words analysis approach is used. In order to identify word instances in the corpus that

was taken before and after preprocessing, and to record the word instances that were available, this work used a bag-of-word technique.

D. Word Distribution – Inverse Document Frequency (IDF)

According to Beel et al., in [33], 83% of academic digital libraries use TF-IDF as a content screening approach. And for automatic keyword detection in information retrieval and text mining, a term frequency-inverse document frequency (tied) approach is used Rajaraman and Ullman [34]. The concepts underlying this method were first proposed in [35]. This algorithm makes advantage of the three numerical features listed as 1) term frequency is the number of times ft (t, d) that the word t appears in a document d. 2) the total number N of documents in a given corpus D; 3) Thus the document frequency is defined as the number of documents df(t) that contain the specified phrase t. The approach computes the quantity inverse document frequency idf (t, D) def = ln N df(t) based on the above last two properties. We then choose words t with the highest value of the product tf-idf (t, d, D) def = tf(t, d) idf as keywords characterizing a particular document d. (t, D).

Table 6. WORD DISTRIBUTION COUNT

Word	Word count	tf-idf count
pocodiwalimadness	319	0.35
Good	270	0.31
wish poco	252	0.31
Stereo wish	247	0.29
Bad	240	0.29
Charging	233	0.29
Happiness	227	0.29
Dead	204	0.27
Dance	201	0.25

TF = (Frequency of a word in the document)/ (Total words in the document)

IDF = Log ((Total number of docs)/ (Number of docs containing the word))

This research work uses the tf-idf weighting method to count the word distribution in the tweet dataset. The tf-idf word distribution count is employed in this work to count the number of occurrences of each word in the complete corpus chosen for research, and it is discovered that in the 3000 tweets taken, the term pocodiwalimadness has a higher word count with 319 word counts and a tf-idf count rate of 0.35 and the word good appears 270 times, with a tf-idf count of 0.3. Similarly, this count is done for the occurrence of each word in the tweet data set considered for this research.

E. Vader Sentiment Analysis

The emotions expressed in the tweets were labeled. The dataset was analyzed using the VADER Sentiment Analyzer. VADER employs a mix of a lexicon is indeed a

collection of lexical features (e.g., words) that are categorized as positive or negative based on their semantic orientation. VADER not just to reports the Positive and Negative scores, as well as how positive or negative a sentiment is. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based sentiment analysis tool and language for expressing emotions through social networks [36].

Table 7. SAMPLE VADER SENTIMENT ANALYSIS

Content	Author	Pos	neg	neu	Compound
1	I am not sure what to do. Is there any fix for this issue? @IndiaPOCO @POCOSupport @Himanshu_POCO @iamprasadtech @prasadyoutuber @GyanTherapy @geekyranjit @TechWiser. Please help. If I apply for a replacement through @Flipkart & I am not sure whether I will receive a good piece.	0.176	0.088	0.736	0.6093
2	@JoJoVSBot Poco dyed his hair 👐	0	0	1	0
3	Want a smartphone with a touch of magic? It's time to pull #POCOM5s out of our early bird hat! Go here: https://t.co/TWxipUCOO9 https://t.co/NIwnWy3ddr	0.074	0	0.926	0.1511
4	Park Jimin, just so you know... my therapist and Pinterest board will be hearing this. U MAKE ME GO UN POCO LOCO https://t.co/4kDaaefHS3	0	0	1	0
5	@DONJAZZY @ayrastarr @boy_director "make you dance like Poco lee" 😳😳💔💔💔	0.217	0	0.783	0.3612
6	Me no Getty time for the hate and the bad energy Got mi mind on my money Make you dance like Poco Lee Steady green like broccoli Steady on my grind, no wan hear what they want tally me Kodi my c (yeah, yeah) Dem wan day check if my tap e no rush	0.125	0.205	0.67	-0.7506
7	she's special. loved the poco cameo.	0.623	0	0.377	0.765
8	@POCOSupport my poco X2's camera is dead and sometimes it facing touch issue and many others and service centre said they can't do anything, never expected this kind of experience from a brand like poco please help.!! @POCOSupport @POCOGlobal @cmanmohan @Xiaomi @XiaomiIndia	0.113	0.09	0.797	-0.2695
9	@ujjwalkhanna20 @IndiaPOCO @Himanshu_POCO @GyanTherapy @stufflistings @yabhishekhd @KaroulSahil @TheGeekyMonk @geekabhishek_ @Harshit_vermaaa @technicaldost Yes Lalo	0.184	0	0.816	0.4019
10	@Poco_lee21 @carterefe__ @BRODAshaggiNG @AbhiDewanCEC @AfricanGodling @BoySpyce_ I wish so many good things for you today on your special day. May each moment meet you with tenderness, filling your heart with endless happiness. God bless ur new age with sufficient grace	0.38	0	0.62	0.9669

The VADER lexicon provides a predetermined sentiment score to each characteristic, which can be a word, an acronym, or an emoticon, ranging from -4 (most extreme negative) to 4 (most extreme positive). VADER created a language based on valence that can detect both the intensity and polarity of sentiments. To begin, in this research work a sentiment intensity analyzer to classify dataset is used. The emotion was therefore determined to use the polar ratings approach. The preprocessed texts were classified as positive, negative, neutral, or compound using the VADER Sentiment Analyzer. The compound value is a useful indicator for assessing sentiment in a group provided tweet. The method uses threshold limit to classify the tweet dataset as good, bad, or balanced. A tweet with just a rating larger than just the threshold was deemed a positive tweet in the current study, while a tweet with a compound value less than the threshold was considered a bad tweet. The tweet was deemed impartial inside the rest circumstances. From the 3000 instances of dataset used for the research work there are 9549 distinct words are identified using word count and after pre-processing 138 words scrutinized by VADER. In the given dataset, the tweet content is analysed, and a score is given in accordance with the Vader rule using VADER, which scores each text sample in a column of text in a CSV file.

Total Word instances: 9549
Features: Word Count
Meta attributes: Word

According to the VADER sentiment analysis, the overall positive sentiment score for the dataset used for this research is 0.011; the total negative sentiment score is 0.027; the total neutral sentiment score is 0.960; and the compound score for all 3000 tweets is -0.05 and since the compound score is -0.05 the overall sentiment of these 3000 tweets is neutral.

F. Performance Evaluation Algorithms

Countless text classification methods have indeed been recommended that use machine learning techniques, probabilistic models, and other methods. They frequently take different approaches. Despite the fact that many strategies have already been suggested, automatically generated text categorization remains a significant research topic, owing to the fact that the effectiveness of current automated text classifiers is not faultless and still needs improvements. In this work data the prediction of Vader sentiment analysis has been evaluated using LMT, Naïve Bayes, Hoefiding, and LBR classifiers in such a training dataset as well as cross-validation.

The classifier LMT, Naïve Bayes, Hoefiding, and LBR classifiers are used in the work. For the purpose of evaluation, the metrics like precision, recall and accuracy is

used. The cross-validation test and the training data set are the two test alternatives used in this work's evaluation method. Cross-validation is a statistical technique that separates a data set into two parts: one for learning or training a model and the other for validating it. It is used to evaluate and compare classification algorithms. Training data is the term for the data that is used to "construct the model." Cross-validation is employed in LMT trees and Hoeffding trees, whereas training data sets are used by the models LBR and Naive Bayes. In essence, the measures used to evaluate the performance of the classifier include precision, F-measure, and recall. The accuracy P has been calculated in comparison to positive, neutral, and negative comments and relates to the ratio of expected positive samples that have been correct. In the work, the classifiers LMT, Naive Bayes, Hoefiding, and LBR are employed, and metrics such as accuracy, precision, and other metrics are utilized to assess the performance of the classification algorithm.

Table 8. PERFORMANCE MEASURE FOR NAÏVE BAYES

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Neutral	0.945	0.014	0.966	0.945	0.956	0.998
Positive	0.968	0.016	0.956	0.968	0.962	0.998
Negative	0.973	0.026	0.965	0.973	0.969	0.997
Weighted Average	0.965	0.024	0.965	0.965	0.965	0.995

The table 8 contains the results of Naïve Bayes Algorithm for the given data set poco mobile review. The Precision for the class Neutral is .966, for positive it is .956, for negative it is .973. Similarly, the Recall for each classis is .94, .968, .973 respectively.

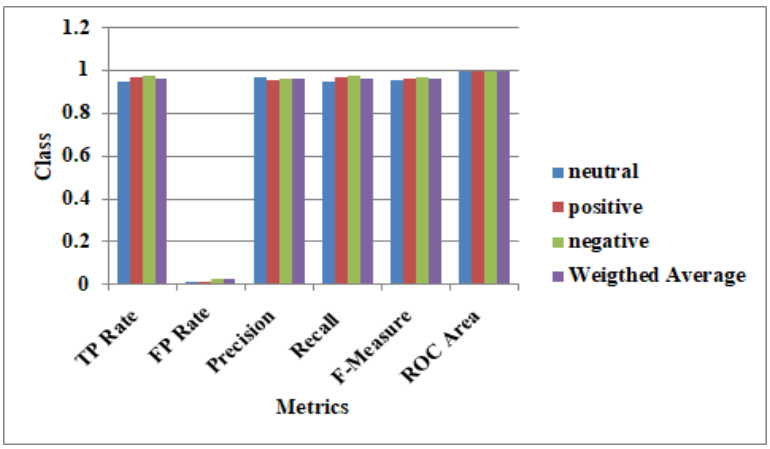


Figure 4. Measures of Naïve Bayes Classification method

Figure 4 shows a visual display of the Naive Bayes Classification findings for the poco mobile reviews data set. The TP rate, FP rate, Precision, Recall, F-Measure, and ROC area for the classes of neutral, positive, and negative are shown in the figure.

Table 9: PERFORMANCE MEASURE FOR LMT

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Neutral	0.809	0.123	0.744	0.809	0.809	0.671
Positive	0.797	0.098	0.746	0.797	0.771	0.929
Negative	0.864	0.023	0.965	0.864	0.912	0.971
Weighted Average	0.856	0.059	0.866	0.856	0.859	0.953

The findings of the LMT algorithm for the provided data set are shown in table 9 poco mobile review. Precision is .744 for the class Neutral, .746 for positive, and.965 for negative. Similar to this, each class's recall is.809, .771, and.859 accordingly.

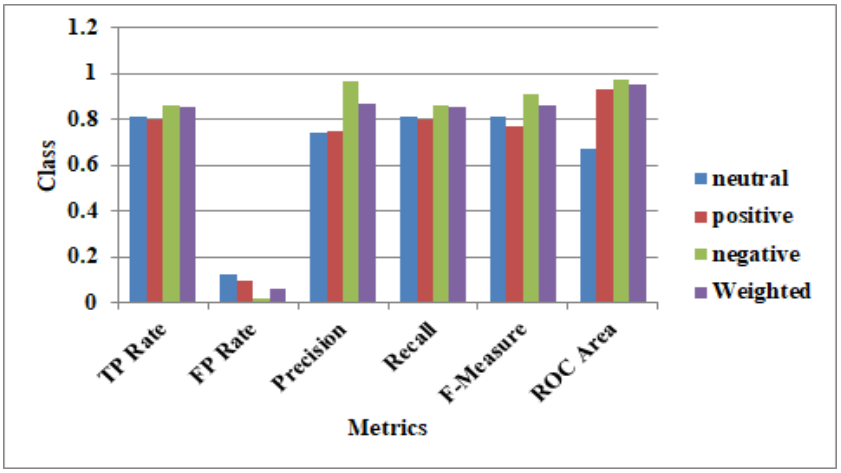


Figure 5. Measures of LMT Classification method

Figure 5 displays the graphs of the LMT Classification algorithm for the collection of reviews for the Poco Mobile device. For the classifications neutral, positive, and negative, the figure shows the TP rate, FP rate, and Precision, Recall, F-Measure, and ROC area.

Table 10. PERFORMANCE MEASURE FOR LBR

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Neutral	0.951	0.01	0.978	0.951	0.964	0.999
Positive	0.968	0.005	0.987	0.968	0.978	0.999
Negative	1	0.023	0.97	1	0.985	1
Weighted Average	0.856	0.059	0.866	0.856	0.859	0.953

In table 10 of the poco mobile review, the results of the LBR method for the given data set are displayed. Precision for the class Neutral is 0.978, for positive it is 0.98, and for negative it is 0.97. Similar to this, the recall for each class is.964,.978, and.985 respectively.

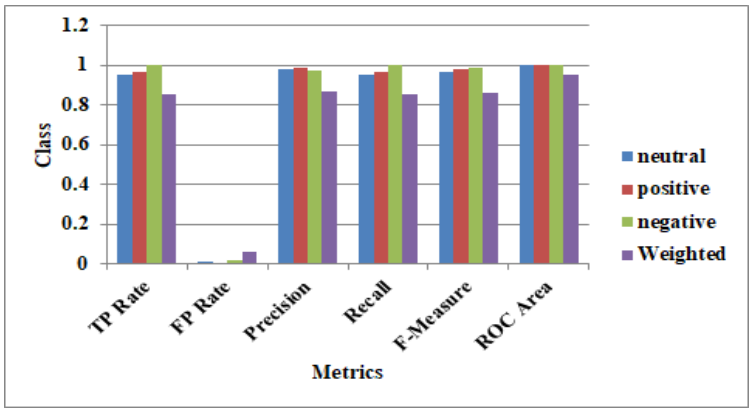


Figure 6. Measures of LBR Classification method

The graphic representation of the LBR Classification algorithm's findings for the data set of Poco Mobile reviews is shown in Figure 6. The figure displays the TP rate, FP rate, Precision, Recall, F-Measure, and ROC area for the classes neutral, positive, and negative.

Table 11. PERFORMANCE MEASURE FOR HOEFFIDING

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Neutral	0.945	0.014	0.966	0.945	0.956	0.998
Positive	0.968	0.016	0.956	0.968	0.962	0.998
Negative	0.973	0.026	0.965	0.973	0.969	0.997
Weighted Average	0.96	0.024	0.96	0.96	0.96	0.995

Table 11 of the poco mobile review shows the results of the Hoeffiding method for the supplied data set. Precision is 0.966 for the Neutral class, 0.956 for the Positive class, and 0.965 for the Negative class. Similar to this, the recall for each class is.945,.968, and.973. And the data in the above figure and table reveal that the accuracy of the LMT algorithm is significantly higher than that of others. Similarly, LBR has a better precision rate than the other three models. To summarize, LBR outperforms other algorithms with values of 0.973, 0.018, 0.974, 0.973, 0.973, 0.998 for measures such as TP rate, FP rate, and Precision, Recall, F-Measure, and ROC area respectively for the supplied poco mobile review dataset from twitter.

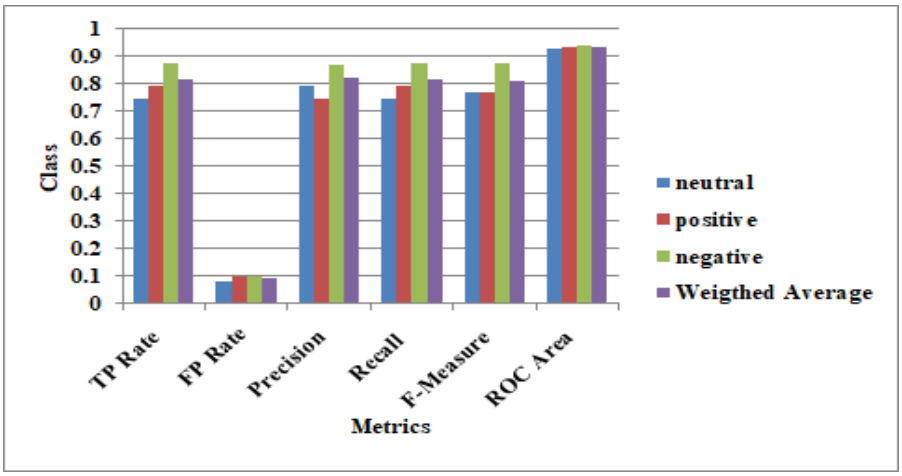


Figure 7. Measures of Hoeffiding Classification method

The Hoeffiding Classification algorithm's findings for the data set of Poco Mobile reviews are shown in Figure 7. The figure displays the TP rate, FP rate, Precision, Recall, F-Measure, and ROC area for the classes neutral, positive, and negative. The performance of each algorithm in comparison to the Poco Mobile review is shown in Figures 4 to 7. The values of each metric for the classification methods Naive Bayes, LMT, LBR, and Hoeffding are similarly shown in tables 8 through 11.

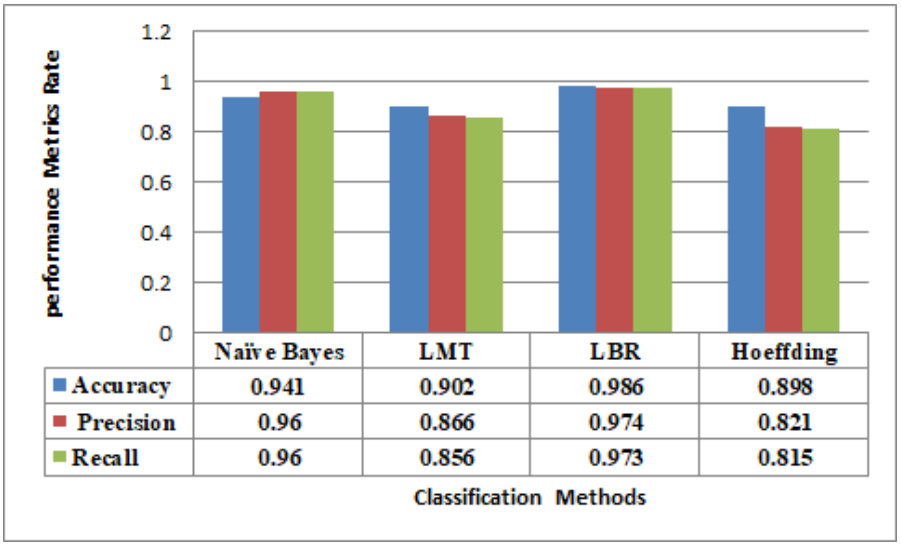
Table 12. PERFORMANCE MEASURE FOR HOEFFIDING

Classification Method	Accuracy	Precision	Recall
Naïve Bayes	0.94	0.96	0.96
LMT	0.90	0.866	0.856
LBR	0.98	0.974	0.973
Hoeffding	0.89	0.821	0.815

The table 12 implies that we add up the total number of forecasts that were correctly classified as Positive (TP) or Negative (TN), and divide it by all categories of predictions—both accurate (TP, TN) and inaccurate (TP, TN)—to arrive at the final result (FP, FN). The accuracy is between 0 and 1. These uncommon situations involve either completely missing or always drawing conclusions accurately. The accuracy, precision and recall of the classification methods are shown in the figure 8. The accuracy obtained from the analysis is 94.11 % for Naïve Bayes, 90.23% for LMT, 98.62% for LBR and For Hoeffding it is 89.86%. So, for the given poco dataset Customer Survey the LMT Algorithm produces better result.

Precision, Recall, and Specificity are commonly used by Data Scientists to overcome the limitations of Accuracy. Precision indicates what percentages of positive predictions were correct. It accomplishes this by dividing the total positive predictions, correct or

incorrect, by the sample size accurately predicted as positive (TP) (TP, FP). The figure 8, shows that the performance measure obtained by each of the classification method. For each method the measures like Accuracy, Precision, Recall is given in the below figure.



	Naïve Bayes	LMT	LBR	Hoeffding
Accuracy	0.941	0.902	0.986	0.898
Precision	0.96	0.866	0.974	0.821
Recall	0.96	0.856	0.973	0.815

Figure 8. Results of Classification method

As per figure 8, The LBR model has highest precision rate compared to the other model. Thus, the LBR has the precision rate of 0.96, Naïve Bayes with 0.974, LMT with 0.866 percent and Hoeffding with 0.821.

5. Conclusions

A vital part of predicting a business's future and ensuring that an organization succeeds through its client reviews based on their opinions is researching and evaluating social media data. More attention is being paid to the customer, and all companies are working hard to delight customers with their goods and services. The business must first comprehend how clients will respond to the product and service in order to offer them such a service and products. It is able to learn about the opinions of the customers by using social media sites like Twitter and others. The analysis of these customers' sentiments, which must have only recently become a large and important factor to enhance their business, is greatly aided by the categorization algorithms. Any firm must evaluate consumer sentiment in order to understand its benefits and shortcomings. The firm would be able to predict market trends with the help of sentiment analysis of the

unstructured text data. The poco review Twitter Sentiments dataset was utilized in this study to examine the effectiveness of various preprocessing techniques. Additionally, classification of the performance is done utilizing algorithms like Naive Bayes, LMT, LBR, x and Hoofing. Finally, it has been determined from this research that, for the pico mobile review data set, the LBR algorithm performs better than the other techniques for Vader sentiment analysis. The outcome of word distribution counts the no of word occurrence in the data set is 9549 and the stance of the tweet is neutral as the compound sentiment value is -0.05.

Acknowledgements

The authors thank the management for permitting us to do the research work.

References

[1] Pak, Alexander, and Patrick Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining", Proceedings of the Seventh International Conference on Language Resources and Evaluation, 2010, pp. 1320-1326.

[2] Sheshasaayee, Ananthi, and R. Jayanthi, "A text mining approach to extract opinions from unstructured text", Indian Journal of Science and Technology, 2015, Vol. 8, Issue 36, pp. 1-4.

[3] Yogapreethi.N, Maheswari.S,"A Review On Text Mining in Data Mining, International Journal on Soft Computing", 2016, Vol.7, No.2, pp.1–8.

[4] Eman Younis, "Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study", International Journal of Computer Applications, 2015, Volume 112, No. 5, pp. 44-48.

[5] Salton G, Singhal A, Mitra M, Buckley C,"Automatic text structuring and summarization", Information processing & management", 1997, Vol.33, No.2, pp.-193-207.

[6] Maulana, Mohamad & Defriani, Meriska, "Logistic Model Tree and Decision Tree J48 Algorithms for Predicting the Length of Study Period", Penelitian Ilmu Komputer Sistem Embedded and Logic , Volume, no.1, pp. 39-48.

[7] Ramanathan, Vallikannu, and T. Meyyappan. "Twitter text mining for sentiment analysis on people's feedback about oman tourism", MEC International Conference on Big Data and Smart City, pp. 1-5.

[8] Dadgar et al.,"A novel text mining approach based on TF-IDF and Support Vector Machine for news classification", 2016, IEEE International Conference on Engineering and Technology, pp. 112 – 116.

[9] Go, Alec, Richa Bhayani, and Lei Huang, "Twitter sentiment classification using distant supervision", CS224N project report, Stanford Vol.1, Issue 12,pp.1-6.

[10] Zubiaga, Arkaitz and Spina, Damiano and Mart"Real-time classification of twitter trends", Journal of the Association for Information Science and Technology, Vol. 66, No.3, 2015, pp. 462-473.

[11] Singh, Prabhsimran and Sawhney, Ravinder Singh and Kahlon, Karanjeet Singh,"Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government", ICT Express, 2018,Vol 4, No. 3, pp. 24-9.

[12] http://www.twitter.com. Retrieved: September, 2022

[13] Sparck Jones, K., "A Statistical Interpretation of Term Specificity and its Application in Retrieval", Journal of Documentation, 1972, vol. 28, Issue 1, pp. 11-21.

[14] Hutto, C., & Gilbert, E, "Vader: A parsimonious rule-based model for sentiment analysis of social media text", Proceedings of the international AAAI conference on web and social media, Vol. 8, No. 1, pp. 216-225.

[15] Pham, Hoang, "Springer handbook of engineering statistics", Vol. 49. London: Springer, 2006,pp. 537–549

[16] Siqueira, Henrique and Barros, Flavia, "A feature extraction process for sentiment analysis of opinions on services" International Workshop on Web and Text Intelligence, pp. 404-413. 2010.

[17] Yuvaraj N and Sabari A, "Twitter sentiment classification using binary shuffled frog algorithm", Intelligent Automation & Soft Computing, Vol. 23, Issue 2, pp.373-381.

[18] Landwehr, Niels , "Logistic Model Trees", extended version of a paper Proceedings 14th European Conference on Machine Learning, pp.16-18, 2004.

[19] Landwehr, Niels and Hall, Mark and Frank, Eibe, "Logistic Model Trees", Machine Learning, Springer Science + Business Media, Inc. M, Volume 59, pp. 161–205, 2005.

[20] Rennie, Jason D and Shih, Lawrence and Teevan, Jaime and Karger, David R, "Tackling the poor assumptions of naive bayes text classifiers", Proceedings of the 20th international conference on machine learning, pp. 616-623, 2003.

[21] Bird, Steven and Klein, Ewan and Loper, Edward,(2009), "Natural language processing with Python: analyzing text with the natural language toolkit ", O'Reilly Media, Inc, pp. 245-250

[22] AytuğOnan, SerdarKorukoğlu, HasanBulut, "Ensemble of keyword extraction methods and classifiers in text classification", Expert Systems with Applications, Volume 57,pp . 232-247, 2016.

[23] Fauziyyah, N. A., S. Abdullah, and S. Nurrohmah, "Reviewing the consistency of the Naïve Bayes Classifier's performance in medical diagnosis and prognosis problems" AIP Conference Proceedings, vol. 2242, no. 1, pp. 030019(1-5), 2020.

[24] Congdon P,"Applied bayesian modeling", John Wiley & Sons; 2014, pp. 1-30

[25] Xu S,"Bayesian Naïve Bayes classifiers to text classification", Journal of Information Science, 2018, Volume 44, No.1, pp.48-59.

[26] Zhang H," The optimality of naive Bayes" Proceedings17th international Florida artificial intelligence research society conference, 2004, pp. 562–567.

[27] Doan T, Kalita J, "Overcoming the challenge for text classification in the open world", 2017, IEEE 7th Annual Computing and Communication Workshop and Conference, pp. 1-7.

[28] Akhter, Muhammad Pervez and Jiangbin, Zheng and Naqvi, Irfan Raza and Abdelmajeed, Mohammed and Mehmood, Atif and Sadiq, Muhammad Tariq, "Document-level text classification using single-layer multisize filters convolutional neural network", IEEE Access, Volume 8, 2020, 42689-42707.

[29] Al-Aidaroos KM, Bakar AA, Othman Z. "Naive Bayes variants in classification learning", International Conference on Information Retrieval & Knowledge Management ,2010 ,pp. 276-281, IEEE.

[30] Zijian Zheng and Geoffrey I. Webb, "Lazy learning of Bayesian rules", Machine Learning, Vol. 41, Issue 1, pp.53–84, 2000.

[31] Permatasari, R. I., Fauzi, M. A., Adikara, P. P., & Sari, E. D. L. (2018, November), "Twitter sentiment analysis of movie reviews using ensemble features based Naïve Bayes", International Conference on Sustainable Information Engineering and Technology (SIET) ,pp. 92-95, IEEE.

[32] Beel, Joeran and Gipp, Bela and Langer, Stefan and Breitinger, Corinna, "Paper recommender systems: a literature survey", 2016, International Journal on Digital Libraries, Vol. 17, No. 4,pp. 305-338.

[33] Rajaraman, A., and J. D. Ullman, 2011, "Mining of Massive Datasets", Cambridge: Cambridge University Press.

[34] Havrlant, Luk and Kreinovich, Vladik,"A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)." International Journal of General Systems, 2017, Vol. 46, No.1, pp. 27-36.

[35] Hutto, C. and Gilbert, E., 2014, May, "Vader: A parsimonious rule-based model for sentiment analysis of social media text", International AAAI conference on web and social media Vol. 8, No. 1, pp. 216-225.

[36] Ahuja, Ravinder and Chug, Aakarsha and Kohli, Shruti and Gupta, Shaurya and Ahuja, Pratyush, "The impact of features extraction on the sentiment analysis", Procedia Computer Science, 2019, Vol. 152, pp. 341-348.