International Journal for Modern Trends in Science and Technology
Volume 10, Issue 04, pages 01-11.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol10issue04.html
DOI: https://doi.org/10.46501/IJMTST1004001

A Features Based Machine Learning Prediction Model for Sentiment Analysis on Social Media

Jessica Sarah Deen[1] | Juan Mark Deen[2] | Amisha Michelle Danny[3] | Arien Maxwell Danny[4] | Marc Ruben Danny[5]

[1]Department of computer science and Engineering, Vellore Institute of Technology, Kotri Kalan, Ashta, Near, Indore Road, Bhopal,Madhya Pradesh 466114, India,[1]Email ID: jessica.sarah2020@vitbhopal.ac.in

[2,4] Department of computer science and Engineering and Bioinformatics, Vellore Institute of Technology, Vellore Campus, Tiruvalam Rd, Katpadi, Vellore,Tamil Nadu 632014, India,[2]Email ID:juanmark0521@gmail.com,[4]Email ID: arien.danny@gmail.com

[3]Department of Computer Science and EngineeringKalinga Institute of Industrial Technology, KIIT Road, Patia, Bhubaneswar, Odisha751024,India, [3]Email ID: amisha.danny@gmail.com,

[5]Department of BBA in Logistics, Retailing & E-Commerce, INDIAN MARITIME UNIVERSITY - [IMU-K] A central university.South End Reclamation Area (Located on NH47- A), Near Alexander Parambithara Bridge, Matsyapuri P.O., Willingdon Island, Kochi- 682029, Kerala, India. [5]Email ID: marcruben.danny@gmail.com

To Cite this Article
Jessica Sarah Deen, Juan Mark Deen, Amisha Michelle Danny, Arien Maxwell Danny and Marc Ruben Danny, A Features Based Machine Learning Prediction Model for Sentiment Analysis on Social Media, International Journal for Modern Trends in Science and Technology, 2024, 10(04), pages. 01-11.https://doi.org/10.46501/IJMTST1004001

Article Info
Received: 16 March 2024; Accepted: 02 April 2024; Published: 03 April 2024.

ABSTRACT

Sentiment analysis is discovering the current ideology opinion of a group of people and their thoughts. The Sentiment analysis based onthe natural reaction of people on social media platform to reflect their mantel status and state. The main poupose of sentiment analysis is to dealing with society's environment and its impact effects towards the media world and surrounding environment. However, this is the key task of understanding every part of the world. The evolution of feeling simulates the sentiment behaviours to shows different direction of reactions and feeling across time. It can help users obtain a more advanced and detailed understanding of the views and attitudes represented in the content provided by users. The development of social media platforms, such as journals, forums, blogs, micro-blogs, Twitter, and social networks, has fostered sentiment analysis. Competitive advantages for organizations are collecting corporate social media and implementing machine learning algorithms to get valuable insights. In this study, our tasks are to show Bag of Words (BoW) and Term-Frequency-Inverse-Document-Frequency (tf_idf) feature-based machine learning prediction models that can help with sentiment analysis and figure out what their customers need and want from company items. Market research is perhaps the most important field for sentiment analysis applications, aside from brand perception and customer opinion surveys and feedbacks. This study results analysis shows the crucial way of classifying social media tweets feedback into positive or negative categories via using the classifier as a baseline to demonstrate in what manner comments are important based on features for any business model and their result.

KEYWORDS: Sentiment analysis, Random forest, Naive Bayes, Logistic regression, Decision Tree

1. INTRODUCTION

Present world more focuses on digital information which will be shared on different media platform. Social media is the one of powerful sharing platform for many of the people or by the group of people from around the world. Social media holding huge information of different opinions comments as likes and dislikes, these databases updated in every millisecond. In order to observed these information resources and data based on public opinions is known as sentiment analysis. Emotion of humans are analyses with a set of practices to identifying public opinions and extracting them for use of particular media environments. These emotion/sentiments are like a political, personal, region, religion, technological, health, entertainment, social network, science, education, news, and business purposes. Machine learning algorithms explore in-depth, thoughtful information due to data mining. Sentiment analysis aims to determine the global attitude of a social media user and their concerning on the overall tonality of a document, notes, blogs, video clips, comments, and image reactions and any topics. Analysis of sentiment opinion based on user comments mining, it reproduces nearly the different opinion but same meaning categories. Sentiment analysis recognizes emotion expressed in a natural language text, and opinion mining actively extracts the opinion from the text/action clicks shorts/emojis'/images. To perform sentiment analysis, we first need to identify the subjective and objective tests as shown in Figure.1. The only subjective text holds the sentiments [7],[9],[10].

1.1 LEVEL OF SENTIMENT ANALYSIS

1.1.1 Document level

The subjective document plays a powerful role in sentiment analysis to classify the level of the entire document based on opinion into different sentiment. In this order various informatics services based on products and their quality. The opinion of the documents is classified into a positive, negative, or neutral.

1.1.2 Sentence level

Sentiment level are classified based on public opinion on social media platform. A level of sentiment identifies based on users' comments. Each comment sentence expresses a positive, negative or neutral opinion for various product and their quality and this is viewed by reviews of user's comments.

1.1.3 Entity and Aspect level

The entity and aspect level of sentiments are concerns to identify and extracting any product features from the source data level or by identifying and judging to their level of feature data. For example, any company launch their new product in the market to fulfils the customer needs according to their brand values, thus this means the information are given by any business entity is satisfied to the desired aspect/feature of the users/customers/public. In these types of sentiment mining find out the reason for the mindset of the audience/ users/customers/public statements views/satisfaction levels for accurately predicted the status of any new launch product.

Expanded product analytics

When users/customers/public start giving biased insight of any company/business/other information related to brand value of product can mislead the looking information, or if they are giving truly insides will helps to gain a competitive edge with other company. These insights of sentiment analysis information led to actual success rate and helps to both the companies/customers to understand what customers are looking for themselves from a particular company's products. After implementing it properly, the owner of the company gets the direction to make changes in their product, so every company while launching their new product has to satisfy customers desire based on sentiment analysis and it helps them to finds usefulness and benefit of their new products and upcoming event.

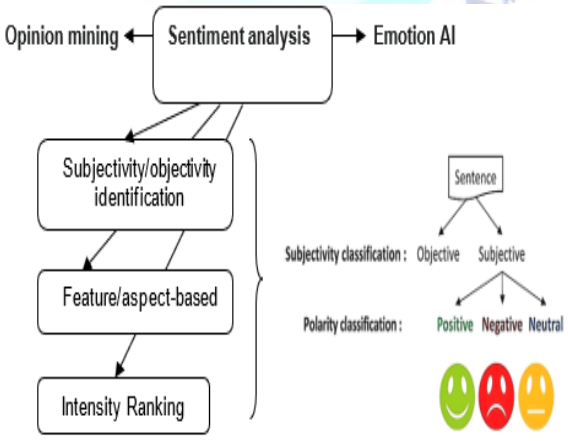

Figure 1. Sentiment analysis workflow.

Examples of Sentiment Analysis

Status Management - Social Media Monitoring - Brand Sentiment Analysis

Today's world business market is monitoring social media analysis based on their reputation and brand values, it is very important for the user as well for compony too, brand sentiment analysis is common in social media to monitoring their reputation in different markets. It is also important to understand what consumers think about the company's brand/product/service. Brand sentiment analysis is equally helpful for any like tech companies, marketing agencies, fashion brands, media organizations and many more. These assessment and presentation of ideas brings additional flexibility and insight into all areas. In addition, it gives some recommendations to companies like: -

1. To track concept/review of customers for their brand
2.User's/views/reviews are indicating the specific details about the brand.
3. It helps to discover the business outlines and trends
4. To keep a continue watch on comments or by very influential persons.

Content analysis on brand mentions media monitoring strategy/testing to reviews on numerous platforms like blogs, social media, clubs, customer reviews, forums, newsgroups and so on.

Monitoring stage operation are:

1. Initially, the company reacts to and adjusts to the incoming data.
2. Sentiment analysis can shift the course of action from responding to managing perception over time.

Market Research, Competitor Analysis

Market research is perhaps the most common application of sentiment analysis, aside from brand reputation and consumer opinion investigation. It's worth noting that sentiment analysis isn't the only approach accessible for market research and their needs. However, it can provide a different viewpoint on the market and a few useful insights shows ground level. Furthermore, the organisation might employ a similar strategy to assess the competition and their marketing activities. Like: -

1. To identifying what types of business products are employed in sentiment analysis
2. Collect data from a variety of sources were comments are created by the users (contents, reviews, etc.)
3. Headlines, articles, content stuff from influencers, and content from competitors

Such feature-based sentiment analysis may help companies better understand their customers' requirements and demands, and adapt their value package to hit all the appropriate notes and fulfils the market desires.

Product Analytics

Most of the sectors are using data and business analytics for their product, the application of analysis on product can be traced back to the reputation of brand values and management analytics. It's conceptually comparable to brand monitoring. However, rather than brand mentions, it focuses on particular comments and observations on the product and its performance in certain areas (user interface, feature performance, etc.). This insight is critical during the early stages of development when the product is being tested by fire (i.e., actual users) and polished as much as possible. The most basic approach to use sentiment analysis at this point is to collect and analyse comments for future improvements. The sentiment analysis algorithm can undertake the necessary business of determining what type of feedback comes from which audience segment and where it points. Typically, the entire thing is separated into the following types:

1. Keywords related to a brand.
2. Customer requirements, Customer sentiment
3. Examine your competitors (based on similar criteria)

As a result, this can play a crucial role in the product's successful market entry.

Example: How Apple is doing it

For example, as the Apple brand presents its products and establishes them in the international market, it is successful in attracting a particular segment. This experiment is a strong example of sentiment analysis applied to the advantage of market research and competitor analysis. It shows how well a product's strong points fit the general satisfaction or dissatisfaction of different segments of the user. For example, the user will not pay attention to these points, such as bad design, poor privacy, low battery life, storage capacity, utility and other negative parameters, if any user uses a branded product because he/she trusts the said brand. Such factors have to be highlighted by examining competitors and their market moves in general in terms of specific aspects. As an example:

1. Value proposition for a brand
2. Dealing with varied issues

3. Adding new features
4. Announcement of milestones
5.Customer care

Combining this information from various market maps allows for the calculation of an extra perspective on how to differentiate and strengthen its value offer. As a result, Apple is a trillion-dollar firm because it pays close attention to each consumer.

OBJECTIVES

SANTIMENT ANALYSIS CHALLENGES

The common sentiment analysis challenges are basically based on credibility/behaviour, sarcasm, grammatically incorrect words, review author segmentation, incremental approach and parallel computing for massive data. An analysis of a large amount of real-time data is very challenging due to the need for an all-time synthesis of data. An incremental approach disagrees with an existing result as if not updated with the current and past event. Authors [17], applies the new sentiment information-based network model (SINM) for Transformer encoder and LSTM as model components via help of Chinese emotional dictionary, they proposed system inevitably find sentiment knowledge in Chinese text. In SINM, they designed a hybrid task learning method to learn valuable emotional expressions and predict sentiment tendencies. Suppose data split the computation into sub-tasks or processes into small segments, which can be performed simultaneously. In that instance, the usage of parallelism may result in a performance boost. It is vital to achieving this in trend and sentiment analysis for enormous data of social media, where massive amounts of instant messages are released every day, in order to fully utilise the overall processing power. The authors [18] proposed creating a multilingual sentiment analysis method that translates words for words using a sentiment dictionary in any language that the user speaks. Text morphological analysis, sentiment extraction from each word using a sentiment dictionary, and text sentiment extraction based on word emotion comprise. This methodology they demonstrate via a test their capacity to identify sentiment in tweets written in English, German, French, and Spanish.

2. RELATED WORK

SOCIAL NETWORK ANALYTICS IMPACT ON SENTIMENT ANALYSIS

People's lives are impacted by social media platforms such as blogs, forums, and social network sites. Individuals are using these virtual places on a regular basis to share thoughts and information, as well as to preserve or develop their relational network. Businesses should not compromise speed, scale, or accuracy in order to comprehend what customers are saying about their products. Sentiment analysis and opinion mining tools combine AI with actual human intelligence to give the most accurate findings, allowing product reviews and other people's opinions to accomplish large-scale sentiment research projects in days. The unprecedented use of online social networks, and the variety of data generated through them, have grown the technological and business communities' immersion in them dramatically. They were coming from dealing with the complexities of natural language without seeing the data collected across social networks as networked data. The majority of sentiment analysis work is focused solely on textual information provided in internet posts and comments. In the most recent research, the first efforts to overcoming this amazing constraint are appearing. [1], [2] authors reported that linking may attempt to access information on social relationships between individuals. However, these characteristics are only inferred from the encrypted rich relationship structure in some online social networks and they proposed the purpose of multi-class classification of online posts from Twitter users is to demonstrate how appropriate the classification is based on views. A hypothesized long short-term memory network (LSTM) that is strengthened by lexicon is presented in article [6]. Before obtaining the sentiment embeddings of the words, the model critically evaluates the collected information using the sentiment lexicon to pre-train a word sentiment classifier. It is possible to improve the accuracy of the word representation by integrating the emotion setting and word establishment. Writing tests on English and Chinese datasets indicate that prediction models could perform as well as or better than current models. According to the paper [7], social networks using machine learning techniques for contextual analysis (CA) contain a system that links terms and sources as they show the tree structure is based on the

Hierarchical Knowledge Tree (HKT). To learn emotion-specific word embeddings from Arabic tweets, the authors of the article [10] describe a unique method for sentiment classification based on Arabic Twitter word embeddings. The publication [12] presented the results of a comprehensive, methodical review of the literature on the approaches and techniques applied in cross-domain sentiment analysis. The writers are making an effort to focus their investigation on articles that were out in 2010 and 2016. In the framework [16], the researchers introduced a unique framework for examining emotions through user comment analysis on Indian Railways tweets. This type of framework is known as a domain-specific framework. The use of various classifiers such as SVM, C4.5 and Random Forest is suggested to leverage business intelligence.

3. PROPOSED METHODOLOGY

PROBLEM STATEMENT

Social networking websites like Facebook and Twitter have changed any information devastated or shared instantly. They became a platform for social network meetings and information exchange. In the process, there is a possibility that people express their valuable opinions on products and services. Thus, social feedback is made rapidly available readily. Any organization cannot afford to have a blind eye to this feedback. However, collecting content from social media and using it to discover business intelligence from the information received is a challenging task. Sentiment analysis in the micro-blogging domain is a relatively new research topic, so there is still a lot of room for further research in this area. The most effective approach to sentiment analysis is to use classification techniques to observe such occurrences in great detail. Machine learning approaches like KNN, Random Forest, Naive Bayes, Decision Tree, Logistic Regression, and Support Vector Machines have been recommended by numerous researchers to accomplish these tasks [8], [9], [11], and [15] are more widespread in used. Due to the quick advancement of Internet technology in recent years, online shopping has gained popularity as a means of making purchases and consuming goods. User happiness can be efficiently increased by sentiment analysis of a huge number of user evaluations on e-commerce sites. A sentiment lexicon analysis using a convolutional neural network (CNN) with a bidirectional gated recurrent unit is shown in article [20].

PROPOSED METHODS

In this study, the proposed methodology has been

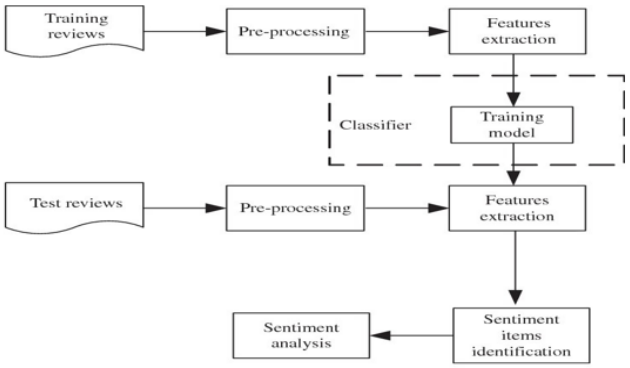

Figure 2. Diagram of Proposed Methodology

used for sentiment classification based on tweets datasets—data samples collection from Twitter accounts of Apple products. Twitter API is used to have live connectivity collected. Afterward, the tweet datasets are pre-processed, and split into the training and testing data samples. The proposed aim for sentiment analysis is to use different machine learning algorithms like Logistic Regression, Naive Bayes, SVM, and Decision Tree as shown in Figure.2.

DATASETS

In social network is very vast it almost covers around the globe. many of users uses to pay attentions via their positive and negative feedbacks or comments on various product, news articles, political views, business industries, entertainment sectors, technological aspects and other prominent sectors etc. However, these comments information are stored in different websites like as using the Twitter API, one can have real-time access to Twitter and gather tweets from the Kaggle website and the UCI Machine Learning Repository. This study uses a balanced data sets of apple product; it should be equally positive and negative tweets—the two sentiment columns corresponding to each class label {features are bow and tf_idf}. Taking a binary value {0,1}, 0 if the tweet is negative, 1 if the tweet is positive as shown in Table 1.

TABLE 1. Description of Datasets

Name	Apple_Datasets
Total no. of samples	16000
Total no. of Features	2
BoW	PositiveFeature 1 Negative Feature 0
tf_idf	Positive Feature 1 Negative Feature 0

DATA PREPROCESSING

In this study pre-processing includes cleaning (remove redundant data), normalization, and splitting. Class A for positive and class B for the negative tweet; both are consistent, uniform sizes of each sample. The split () method in Python is used for selected training datasets as shown in Table 2. The training model implies a validation set used to help tune the hyperparameters of the trained model and lead to better performance.

TABLE 2. DATA AFTER SPLITTING

DATASETS	NO. OF TWEETS SAMPLE (UCI and Kaggle)	NO. OF FEATURES
Train Data	13000	2
Test Data	3000	2

FEATURE EXTRACTION

Word Embedding is one such approach that allows us to represent text using vectors. The most common types of word embeddings are these two common techniques, BOW and TF-IDF, are used for feature extraction. The steps involved in extracting features; first have to convert the text into a numerical form; second the documents apply the classifier into two sets: positive {1}and negative {0} for sentiment text analysis.

BAG OF WORDS

A document is viewed as an unstructured set of words that ignores word order and grammar. In the proposed model, during the training phase, a dictionary is constructed based on the training data and utilized to characterize the positive and negative posts in the procedure. Each document's feature vector has a dimension of either 0 or 1. The most basic ways of representing text in numbers is the Bag of Words (BoW) model. A phrase can be represented as a vector of words, or a bag of words, just like the term itself. Here we have considered the three kinds of Apple user reviews:

Review 1: This Apple product is not user friendly

Review 2: This Apple product is very costly and unrepairable

Review 3: This Apple product user friendly and good.

Using each of the three distinct terms from the aforementioned reviews, this study first creates a lexicon. Twelve words make up the vocabulary: as table 3 illustrates, "this," "Apple," "product," "is," "very," "not," "user," "friendly," "unrepairable," "good," "and," and "costly".Here at Table 3 at coloum 1 comments are R1 isReview 1,R2 -Review2 and R3 is Review3.

TABLE 3. BoW(Bag of word) Model

comments	1.This	2.Apple	3.Product	4.is	5.not	6.user	7.friendly	8.very	9.costly	10.and	11.unrepairable	12.good	Length of the
R 1	1	1	1	1	1	1	1	0	0	0	0	0	0 7
R 2	1	1	1	1	0	0	0	1	1	1	1	0	0 8
R 3	1	1	1	0	0	1	1	0	0	1	0	1	0 7

These terms are highlighted with 1s and 0s in the three product consumers' reviews shown above. Three vectors for three reviews will result from this:

Review #1's vector is [1 1 1 1 1 1 0 0 0 0 0]

Review #2's vector is [1 1 1 1 0 0 0 1 1 1 1 0]

Review #3's vector is [1 1 1 0 0 1 1 0 0 1 0 1]

likewise, a Bag of Words (BoW) model is based on that key principle.

TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)

"Term-frequency-Inverse-document-frequency" is the complete form of tf-idf, where term frequency is the

number of times a word appears in a particular document. If a particular word occurs very frequently in a particular document, then that document is considered relevant for that word(query). The frequency value is known as term-frequency. The basis of word frequency shown positive and negative trends in social media. If a word occurs only in a few documents, then it gets a higher IDF value, while if the given the word occurs in most of the documents (means not relevant) gets a lower IDF value. In this way, infrequent words get highlighted, and frequent non-useful words are penalized by inverse document frequency value. Thus, it solves the issue with frequent irrelevant words. In other words, it does not care about the frequency of a word within a document.

TFIDF score tells the importance of a given word in a given document. For example, a given word query you can actually rank the documents w.r.t. (with respect to) to the relevance tf-idf score. tf-idf score of a term (t), in a given document (d) with respect to a set of documents (D), is defined as eq. (1).

$$tft, d=\frac{n_{t,d}}{Numberoftermsint \quad hedocument} \text{eq. (1)}$$

The term frequency-inverse document frequency is a metric that quantifies a word's significance within a set of textual documents.

Term Frequency (TF)

Term Frequent = TF. t = measure of frequent term, document = d, n = No. of times the term "t" appears in the document "d". For example consider review 2 to calculate the TF, here review two comment is" This Apple product is very costly and unrepairable " 'This', 'apple', 'product', 'is','very', 'costly', 'and', 'unrepairable'. Total length of words in Review 2 = 8 as shown in table 3. Count tf in comment of reviewer 2 is as given eq. 2. TF for the word "this" is equal to (number of words in review 2) / (number of times "this" appears in review 2), which equals 1/8. likewise, count for all words.

- Term Frequent ('apple') = 1/8
- Term Frequent ('product') = 1/8
- Term Frequent ('is') = 1/8
- Term Frequent ('not') = 0/8=0
- Term Frequent ('user') = 0/8=0
- Term Frequent ('friendly') = 0/8 = 0
- Term Frequent ('very') = 1/8
- Term Frequent ('costly') = 1/8
- Term Frequent ('and') = 1/8
- Term Frequent ('unrepairable') = 1/8
- Term Frequent ('good') = 0/8 = 0

This feature evaluates, as indicated in the Table no. 4, the term frequencies for each term and each review.

TABLE 4. Term Frequencies Model

TERM	Review#1	Review#2	Review#3	TF Review#1	TF Review#2	TF Review#3
This	1	1	1	1/7	1/8	1/7
Apple	1	1	1	1/7	1/8	1/7
product	1	1	1	1/7	1/8	1/7
is	1	1	0	1/7	1/8	0
not	1	0	0	1/7	0	0
user	1	0	1	1/7	0	1/7
friendly	1	0	1	1/7	0	1/7
very	0	1	0	0	1/8	0
costly	0	1	0	0	1/8	0
and	0	1	1	0	1/8	1/7
unrepairable	0	1	0	0	1/8	0
good	0	0	1	0	0	1/7

Inverse Document Frequency (IDF)

The IDF is a metric for determining how essential a phrase is. The IDF value is required for this study since computing the TF alone is insufficient to evaluate the significance of words:

$$idft, =\log \frac{numberofdocument}{Numberofdocumentwit \quad hterm\ 't'} \text{eq. (2)}$$

The IDF values for each word in Review#2 are determined using equation (2) as follows: - IFDF('this') = log of number of documents/number of documents that include the word 'this') = log (3/3) = log (1) = 0; likewise: -

- For 'apple' IDF is equals to log (3/3) = 0
- For 'product' IDF is equals to log (3/3) = 0
- For 'is' IDF is equals to log (3/2) = 0.18
- For 'not' IDF is equals to log (3/1) or log (3) = 0.48
- For 'user' IDF is equals to log (3/2) = 0.18
- For 'friendly' IDF is equals to log (3/2) = 0.48
- For 'very' IDF is equals to log (3/1) or log (3) = 0.48
- For 'costly' IDF is equals to log (3/1) or log (3) = 0.48
- For 'and' IDF is equals to log (3/2) = 0.18

- For 'unrepairable' IDF is equals to log (3/1) or log (3) = 0.480
- For 'good' IDF is equals to log (3/1) or log (3) = 0.48

The IDF values for each word in this study are determined by feature extraction algorithms in inverse document frequency manner. As a result, the IDF values displayed in the Table no.5 would apply to the entire lexicon.

TABLE 5. IDF Value of comment

Term	Review#1	Review#2	Review#3	IDF#
This	1	1	1	0.00
Apple	1	1	1	0.00
product	1	1	1	0.00
is	1	1	0	0.18
not	1	0	0	0.48
user	1	0	1	0.18
friendly	1	0	1	0.18
very	0	1	0	0.48
costly	0	1	0	0.48
and	0	1	1	0.18
unrepairable	0	1	0	0.48
good	0	0	1	0.48

This study extracts the feature from tweeter comments of apple products to determine each word's TF-IDF score inside the quantity. The above table indicates that terms that have a higher score are considered more significant than those that have a lower value. As a result, we observe that terms like "Apple," "this," and "Product" are negligible and have little significance, whereas terms like "is", "and", "user", "friendly" are words with more importance and those IDF value is 0.48 have a higher value.

$$(tf\_idf)_{t,d} = tf_{t,d} * idf_t \qquad \text{eq. (3)}$$

Using eq. (3), determine the TF-IDF score for each word in Review #2. TF-IDF ('this', Review#2) = TF ('this', Review#2) * IDF('this') = 1/8 * 0 = 0

In a similar vein,

- For IDF 'apple' by Review #2 equals to 1/8*0= 0
- For IDF 'product' by Review#2 equals to 1/8*0= 0
- For IDF 'is', Review#2equalsto 1/8*0.18=0.023
- For IDF 'not', Review#2 equals to 1/8* 0.48=0.06
- For IDF 'user', Review#2 equals to 1/8*0.18=0.023
- For IDF 'friendly', Review#2 equals to 1/8*0.18=0.023
- For IDF 'very', Review#2 equals to 1/8*0.48=0.06
- For IDF 'costly', Review#2 equals to 1/8*0.48=0.06
- For IDF 'and', Review#2 equals to 1/8* 0.18=0.023
- For IDF 'unrepairable', Review#2 equals to 1/8*0.480=0.06
- For IDF 'good', Review#2 equals to 1/8* 0.48=0.06

The TF-IDF scores for each word in relation to each review were calculated in the same manner, as indicated in the Table.6.

TABLE.6 Tf-idf Scores of All Reviews

Term	Review #1	Review #2	Review #3	IDF	TI_IDF Review# 1	TI_IDF Review# 2	TI_IDF Review #3
This	1	1	1	0.00	0.00	0.00	0.00
Apple	1	1	1	0.00	0.00	0.00	0.00
product	1	1	1	0.00	0.00	0.00	0.00
is	1	1	0	0.18	0.025	0.023	0.00
not	1	0	0	0.48	0.068	0.060	0.00
user	1	0	1	0.18	0.025	0.023	0.025
friendly	1	0	1	0.18	0.025	0.023	0.025
very	0	1	0	0.48	0.00	0.060	0.00
costly	0	1	0	0.48	0.00	0.060	0.00
and	0	1	1	0.18	0.00	0.023	0.025
unrepairable	0	1	0	0.48	0.00	0.060	0.00
good	0	0	1	0.48	0.00	0.060	0.068

The TF-IDF scores for the above vocabulary and classifiers train to classify these feature vectors in sentiment comments as views on social media datasets. It is noted that in this case, the TF-IDF likewise provides higher values for less common terms and is high when both the IDF and TF values are high, meaning the word is rare throughout all of the documents taken together but frequent in one particular document. In article [19], they design a weight distributing method combining the two methods for text sentiment analysis, by which the sentence vectors obtained can both highlight words with sentiment meanings while retaining their text information and rule-based sentiment dictionary method shows their results are 7.7% better than those of the TF-IDF weighting method. But also, they acknowledged

that the suggested approach still has certain drawbacks and restrictions and that it has to be enhanced, sentiment analysis at the sentence level is frequently insufficient for applications because it does not identify opinion targets or assign sentiments to such targets. The suggested method in [19] is unable to break the boundary of a single sentence, which breaks the connection between text and contexts.

4. CLASSIFICATION TECHNIQUES

NAIVE BAYES

Classifying Naive Bayes is a collection of Bayes' classification algorithms. Bayes Theorem finds the likelihood of an event. The mathematical description of Bayes' theorem is as follows as eq. (4):

$$P(A/B) = \frac{P(B/A)\,P(A)}{P(B)} \qquad \text{eq. (4)}$$

Determine how likely it is that event A is true, given that event B is true. Event B is referred to as proof. The priori of A is P(A). An attribute value of an unidentified instance (in this case, event B) serves as the proof. When evidence is observed, the event probability is known as P(A|B), or the posterior probability of B [3].

LOGISTIC REGRESSION

Logistic regression is used to group observations into distinct classes. Here, the outcomes of linear regression are obtained using two sentiment classes. As illustrated in Figure. 3, logistic regression maps its output to both discrete classes [5] by means of the logistic sigmoid function.

Figure 3. Block Diagram of Logistic Regression.

SUPPORT VECTOR MACHINE

A separating hyperplane serves as the formal definition of a discriminative classifier, or support vector machine (SVM). As seen in Figure.4, it provides labeled training data. The method produces a desirable hyperplane that classifies new data points. This hyperplane is a line that splits a plane into two sections in two dimensions, with each class lying on one side [6].

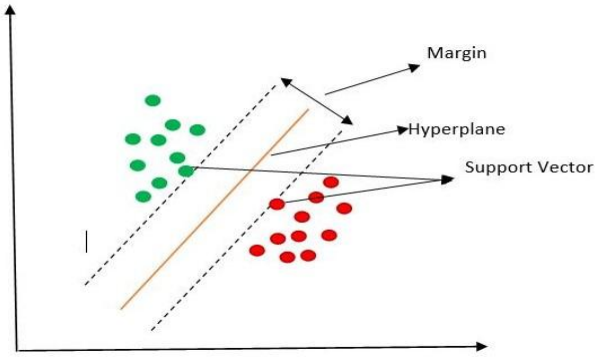

Figure 4. Block Diagram of SVM.

DECISION TREE

A decision tree is a decision guidance tool that makes use of an analogy or graph of decisions that forms a tree. It's one method of presenting an algorithm with just conditional control statements. The decision tree of each branch indicates the test result, while the leaf nodes indicate the class labels (decisions made after calculating all characteristics). The routes connecting the root and the leaf signify the rules for classification.

5. EXPERIMENTAL SETUPS AND RESULTS

SYSTEM SPECIFICATION

The following are the basic system requirement for the better performance of the software and the tools used in the experiment. As the system specification as shown in Table. 7, plays an important role in determining how the software and tool works.

TABLE. 7 System Specification

Operating System	Windows 10
Processor	Intel(R) Core (TM) i3 5th Gen
RAM	8.00 GB
Hard Disk	1 TB
System Type	64-bit Operating System, x64 based processor.

SOFTWARE AND TOOLS USED

For this suggested experiment, Python, the most widely used general-purpose high-level programming language, is utilized. Figure.5 illustrates the workflow used in web development, data visualization, and analytics today. A vast array of libraries for data analysis and machine learning are available in Python.

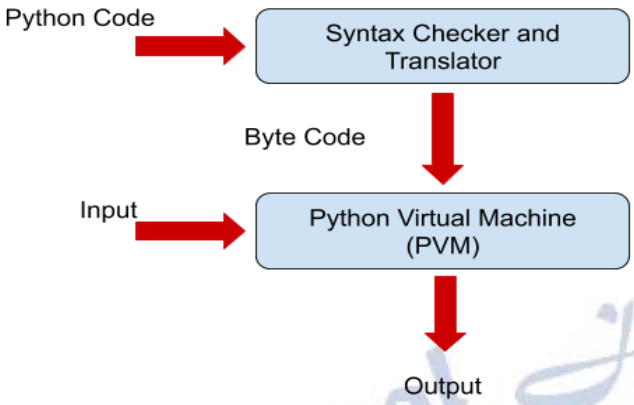

Figure 5. Working of Python.

Python library used

sklearn is a machine learning library for the Python programming language. It may be used with the Python scientific and numerical libraries NumPy and SciPy, and it has a number of classifications, regression, and clustering algorithms built in.

Pandasone of the most popular Python libraries for data analysis and visualization. It offers back-end source code performance that is highly optimized. The DataFrame is its primary data structure, and it is based on the NumPy library.

NLTK is a top platform for developing Python applications that interact with data in human languages, along with a collection of text processing tools for parsing, tokenization, classification, stemming, tagging, and semantic reasoning.

NumPy is a broadly useful exhibit preparing bundle. It is the essential package for Python logical processing. sophisticated capabilities for broadcasting.

JUPYTER NOTEBOOK

A Jupyter notebook is open-source software for interactive computing across many different languages. As seen in Figure. 6, the Jupyter Notebook App is a client-server program that enables us to edit and execute the notebook document through a web browser.

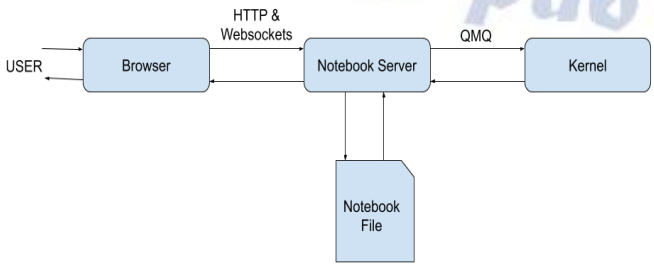

Figure 6. Working of Jupyter Notebook.

PARAMETER OF EVALUATION

The evaluation parameter, eq. (5) to eq. (8) as terms used below [13],[14].

TP - Samples that were categorized as positive but are actually positive are known as true positive samples (TP).

FP - Samples that were categorized as positive but ought to have been categorized as negative are known as false positive samples (FP).

FN - False-negative samples (FN), which ought to have been positive, were labeled as negative.

TN - Samples that were first categorized as negative but are actually negative are known as true negative samples (TN).

Process Description

The following diagram makes it easier to understand how we proceed.

Precision $P = \frac{TP}{TP+FP}$ eq. (5)

Recall $R = \frac{TP}{TP+FN}$ eq. (6)

Accuracy $A = \frac{TotalPositive}{TotalSample}$ eq. (7)

F-score $F - score = 2\frac{P*R}{P+R}$ eq. (8)

Receiver Operating Characteristics (ROC) curve

A graphical plot known as a receiver operating characteristic curve, or ROC curve, shows how well a binary classifier can diagnose problems. Plotting the true positive rate (TPR) against the false positive rate (FPR) yields the ROC curve. And accurate positive findings among all the positive samples that are available for the test are determined by the TPR. FPR, on the other hand, indicates the proportion of false positive results that arise from all of the test's available negative samples. FPR and TPR define a ROC space as the x and y axes, respectively.

BoW COMPARISION OF VARIOUS CLASSIFIER

Machine learning algorithms are trained, and their performance is analyzed using an Apple dataset that was gathered from the UCI Machine Learning Repository and Kaggle. The study employed Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), and Naive Bayes (NB) algorithms for comparison; pre-processing and implementation procedures are discussed for comparison. The accuracy score, precision, recall, and f-score of SVM, LR, NB, and DT are compared in Table 8. Achieved 97.91% height accuracy in the suggested SVM. Figure 8 displays the ROC of the bow comparison of different classifier

methods, while Figure 7 displays the different classifier bar chart of BoW.

TABLE 8. Comparisons of various Classification Algorithms for BoW.

Algorithms	Accuracy Score	Precision	Recall	F-score
LR	92.92	0.93	0.90	0.84
SVM	97.91	0.97	0.92	0.98
NB	95.27	0.89	0.91	0.93
DT	90.09	0.93	0.90	0.90

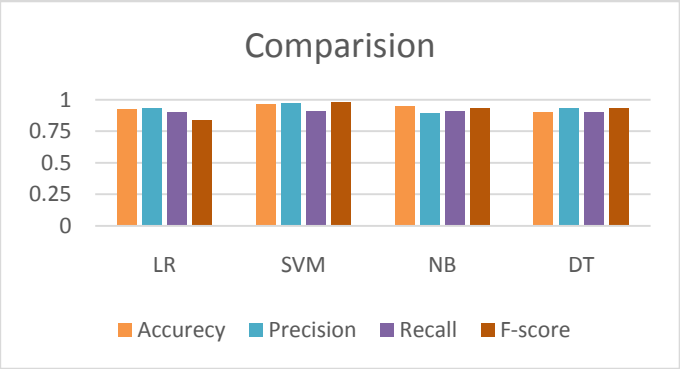

Figure 7. Comparison of Various Algorithms for BOW.

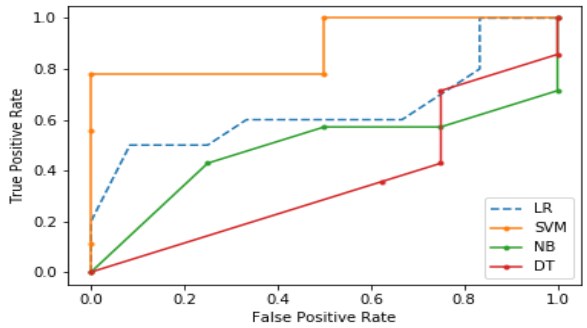

Figure 8. The graphical representation of the accuracy score, precision, recall and f-score of algorithms represented BoW.

The x-axis represents the various classification algorithms and the y-axis represents the accuracy score, f-score, precision and recall of algorithms.

COMPARISION OF VARIOUS CLASSIFIER ALGORITHMS USED ON TF-IDF

The details about the dataset and implementation steps are discussed. Performance analysis of machine learning algorithms for apple sentiment is performed. For comparison, we have employed the techniques of Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), K-nearest Neighbor (KNN), Decision Tree (DT), and Naive Bayes (NB). An analysis of the accuracy score, precision, recall, and f-score of SVM, LR, NB, and DT is presented in Table 9. When compared to other algorithms, the SVM has the highest accuracy (92.17%). Figures 9 and 10 display the ROC of classifier comparisons and the bar chart of different classifiers for tf_idf.

TABLE 9. Comparision of Various Classification Algorithms for TF-IDF.

Algorithms	Accuracy Score	Precision	Recall	F-score
LR	91.36	0.89	0.93	0.87
SVM	89.68	0.90	0.89	0.82
NB	92.17	0.91	0.80	0.85
DT	87.91	0.82	0.81	0.82

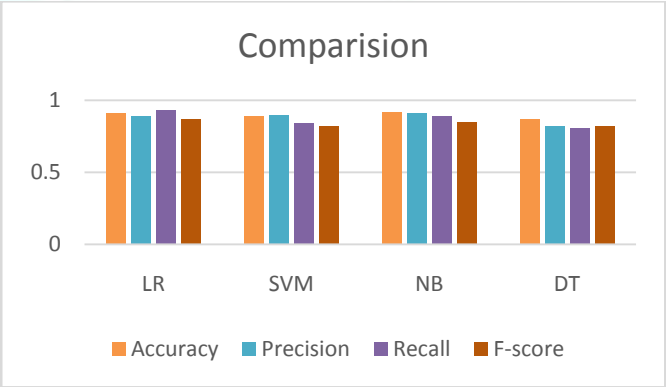

Figure 9. Comparison of Various Algorithms for TF-IDF.

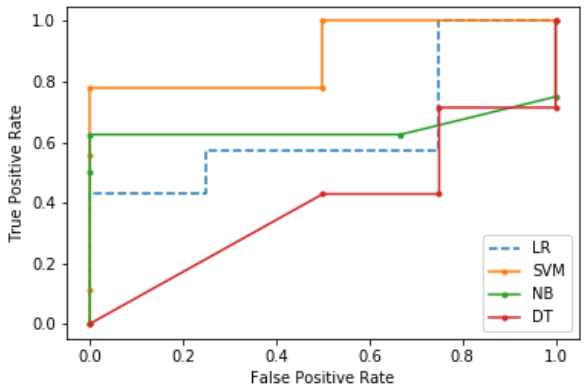

Figure 10. The graphical representation of the algorithm's accuracy, precision, recall, and f-score are represented Tf_IDF

The accuracy score, precision, recall, and f-score of SVM, LR, NB, and DT are displayed graphically in Figure 10. The different classification methods are represented by the x-axis, and the algorithm's accuracy, precision, recall, and f-score are represented by the y-axis of Tf_IDF.

6. FUTURE SCOPE AND CONCLUSION

This study presents a new technique for analyzing Twitter data for sentiment in social media. The enhancement of model Bag-of-words (BoW) and term frequency-Inverse document frequency(tf-idf) comments on English language sentiments. The proposed technique can improve accuracy and visualize the understanding of the implicit and explicit meaning of sentiments. This study measures the proposed learning techniques very effectively. Here, studies attempted to demonstrate the fundamental process of categorizing tweets into positive or negative groups by utilizing working language models based on characteristics and the classifier as a baseline. By attempting to extract two distinct features from the tweets and adjusting the classifier's parameters to suit the objectives. A further study area might be utilizing active learning techniques to detect Twitter sentiments and increase decision-makers confidence as the TF-IDF has significant limitations as well because it skips sentiment tendency. In the future, using more hybrid methods can provide a more accessible way of feature selection and direction for sentiment analysis. If the hybrid technique is appropriate in regions where sentiment refinement is highly required, it may prove to be a more effective means of classifying sentiment into positive and negative categories in the future.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] MondherBouazizi,TomoakiOhtsuki, "Multi-class sentiment analysis Twitterter: Classification performance and challenges", Big Data Mining and Analytics,2019

[2] HuizhiLiang;Umarani, Ganesh babu, " A Dynamic Bayesian Network Approach for Analysing Topic-Sentiment Evolution" Thomas Thorne IEEE Access,2020.

[3] JianfeiYu; Jing Jiang; Rui Xia," Entity-Sensitive Attention and Fusion Network for Entity-Level Multimodal Sentiment Classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 28) Page(s): 429 - 43, 06 December 2019

[4] M. Bouazizi and T. Ohtsuki, "A Pattern-Based Approach for Multi-class Sentiment Analysis in Twitter", in Proc. IEEE ACCESS, pp. 20617-20639, 2017.

[5] Shihab Elbagir Saad; Jing Yang, "Twitter Sentiment Analysis Based on Ordinal Regression", DOI 10.1109/ACCESS.2019.2952127, IEEE Access, 2019

[6] XianghuaFu;JingyingYang;JianqiangLi; Min Fang;Huihui Wang, "Lexicon-Enhanced LSTM With Attention for General Sentiment Analysis", 10.1109/ACCESS.2018.2878425,2018.

[7] Azwa Abdul Aziz; Andrew Starkey, "Predicting Supervise Machine Learning Performances for Sentiment Analysis Using Contextual-Based Approaches", IEEE Access,2020.

[8] JieshengWu;Kui Lu; ShuzhiSu;Shibing Wang "Chinese Micro-Blog Sentiment Analysis Based on Multiple Sentiment Dictionaries and Semantic Rule Sets", IEEE Access,2019

[9] LakshmishKaushik; Abhijeet Sangwan; John H. L. Hansen "Automatic Sentiment Detection in Naturalistic Audio", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017

[10] Nora Al-Twairesh; Hadeel Al-Negheimish, "Surface and Deep Features Ensemble for Sentiment Analysis of Arabic Tweets", IEEE Access, 2019.

[11] ZhengjieGao;AoFeng;XinyuSong; Xi Wu," Target-Dependent Sentiment Classification With BERT, IEEE Access,2019

[12] Tareq Al-Moslmi;NazliaOmar;Salwani Abdullah; Mohammed Albared, "Approaches to Cross-Domain Sentiment Analysis: A Systematic Literature Review", IEEE Access Year: 2017.

[13] Guixian Xu; YuetingMeng;XiaoyuQiu; Ziheng Yu; Xu Wu, " " Sentiment Analysis of Comment Texts Based on BiLSTM" IEEE Access ,2019

[14] Bo Zhang; Duo Xu; Huan Zhang, Meizi, "LiSTCS Lexicon: Spectral-Clustering-Based Topic-Specific Chinese Sentiment Lexicon Construction for Social Networks", IEEE Transactions on Computational Social Systems, 2019

[15] D. Krishna Madhuri (2019) A Machine Learning based Framework for Sentiment Classification: Indian Railways Case Study ISSN: 2278-3075, Volume-8 Issue-4, February 2019.

[16] Ankur Goel, Jyoti Gautam, Sitesh Kumar, "Real Time Sentiment Analysis of Tweets Using Naive Bayes", 2nd International Conference on Next Generation Computing Technologies, Dehradun, 2016.

[17] Gen Li; Qiu Sheng Zheng; Long Zhang; Su Zhou Guo; Li Yue Niu," Sentiment Information based Model for Chinese text Sentiment Analysis" 2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE) 20-22 November 2020, DOI: 10.1109/AUTEEE50969.2020.9315668, Conference Location: Shenyang, China.

[18] Keita Fujihira, Noriko Horib, et.al.," Multilingual Sentiment Analysis for Web Text Based on Word-to-Word Translation", 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI),01-15 September 2020, DOI: 10.1109/IIAI-AAI50415.2020.00025.

[19] Hao Liu, Xi Chen, Xiaoxiao Liu," A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis", IEEE Access, Vol.10, Pp: 32280 – 32289, 16 March 2022, ISSN: 2169-3536, DOI: 10.1109/ACCESS.2022.3160172.

[20] Li Yang; Ying Li; Jin Wang; R. Simon Sherratt," Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning", IEEE Access, Vol.8, Pp: 23522 – 23530, 28 January 2020, ISSN: 2169-3536, DOI: 10.1109/ACCESS.2020.2969854.