SENTIMENT ANALYSIS ON FACEBOOK COMMENT USING MACHINE LEARNING

Nivedita Mishra (1812010069)
Institute of Technology and Management Gida, Gorakhpur

Priya Srivastava (1812010053)
Institute of Technology and Management Gida, Gorakhpur

Saumya Srivastava (1812010094)
Institute of Technology and Management Gida, Gorakhpur

Shikha Gupta (1812010102)
Institute of Technology and Management Gida, Gorakhpur

Mr. Mahendra Sonker (1812010069)
Institute of Technology and Management Gida, Gorakhpur

Abstract

In this work, I present a methodology that uses social media to analyze a person's sentiment and emotion. Face-book is utilizing machine learning. This initiative will assist them in understanding their situation and improving their emotional stability. The goal of this study is to retrieve and pre-process social media data in order to do sentiment analysis, which is a type of natural language processing. For improved sentiment analysis, the most important aspect of this article is to demonstrate how people feel about certain social media statuses, which will be used to classify them. The amount of data created has continually increased, and an ever-increasing variety of data types are being stored in unstructured or semi-structured pages: arranged configurations. Slant Analysis is a technique for extracting emotional information from online data. Assumption testing allows computers to automate human-like tasks by making decisions based on assumptions made in comments or posts on social media sites. In this research, we used five different machine learning approaches to assess the sentiments of Face-book comments: Nave Byes, SVM, Random Forest, KNN, and Decision tree. Different performance measures such as Precision, Recall, and F1-score were used to evaluate these five classifiers.

Keywords: -

Sentiment Analysis, Machine learning, Classifiers Precision, Recall, F1 Score.

Introduction

Sentiment analysis has a big role in computer science. Facebook allows the user to post real time short messages called as comments. These comments are restricted to 140 characters in length [2, 14, 16]. Peoples on a social network share the various points of view related to any subject or product with their friends, relatives, and followers etc. Users can post their views on a special social issues belonging form national or international in the from of text, audio, video, photos (comments). Sentiment can be expressed in terms of feeling, attitude, opinion, and emotions. As we know, online social networking website are growing as new communication platforms that give huge freedom for people to communication and share their feeling. Sentiment analysis is a natural processing language. Processing and information extraction task that analysis a huge number of documents in order to extract the thoughts reflected in positive or negative comments, queries and requests. It determine the polarity and intensity of the emotion. Sentiments are categorized as objective(Facts), positive(denotes a state of happiness, bliss or contentment part), or negative (denotes a state of sorrow, disappointment part). The statement can also be given a score based on how positive negative objective they are Because of the rapid increase in the number of social media users, a massive amount of online data is generated every day. People now feel free to share their thoughts, feelings, and opinions on social media in the form of comments or posts. And these comments or posts can be used as product reviews to understand 1808 about a product, to learn public opinion on a social problem, and so on. For these objectives, the text's sentiments must be examined. The extraction of sentiments expressed in it is known as sentiment analysis. It's a strategy for determining whether a sentence is good, negative, or neutral.

There are different level of sentiment analysis are:

Document level: -sentiment of an entire document.

Sentence level: - sentiment of a single sentence.

Sub sentence level: -sentiment of sub expressions within a sentence.

We attempted to determine the sentiment polarity of Face-book comments, i.e. positive, negative, or neutral, in this paper. Sentence level sentiment analysis was the scope of this technology. On the same dataset, we used multiple machine learning approaches for classification and compared their performance using an accuracy metric. Nave Bayes, Support Vector Machine, Decision Trees, Random Forest, and K Nearest Neighbors were among the classifiers utilized (KNN). The remaining sections of this work are organized as follows: The works that are connected are discussed in Section II. Section III contains the proposed system's methodology. Section IV contains a brief discussion of the classifiers utilized in the implementation. The result analysis is presented in Section V, and the conclusions are presented in Section VI.

Literature Review: -

Related Work: -

According to Monali et al.[6], the writers' attitudes in the text can be used to make decisions. It is a method that provides the user's opinion, whether good, negative, or neutral. Proposed the technology extracts feelings using a Lexicon-based technique. All of the issues were discussed. Sentiment analysis tools, including lexicon-based and machine-learning techniques and, according to the authors, Lexicon-based results are quick and accurate without requiring a lot of effort training.

Basheer, Shakila, and colleagues [7] conducted an assessment of different sentiment analysis methodologies, including vocabulary-oriented and machine-learning approaches. SVM is used in the proposed method to analyse sentiment in smart phone product reviews. They determined that SVM is a superior and more robust technique, with an accuracy of 90.99 percent, based on the findings.

According to K.Srividya et al. [8,] feelings from Face-book data reveal public opinion on a topic. Based on user ratings, the Naive Bayes classifier is used to classify the most popular Android version.

Using machine learning algorithms such as decision trees, KNN, and the Naive Bayes method, Achmadet al.[9] performed sentiment analysis on tweets concerning online markets. According to the findings, Naive Bayes classification performed effectively and with high accuracy.

Sultana, H Parveen, and colleagues[10] created a technique for sentiment analysis of Amazon product reviews. The proposed system uses SVM, Random Forest, and a hybrid of SVM and RF called SVMRF to complete the process. The results reveal that the combined categorization approach has a higher accuracy rating.

R.Soundarya et al. [11] and Ram lingam et al. [12] have used sentiment analysis to construct several machine learning algorithms for depression diagnosis.

Methodolgy: -

The main goal of this method is to determine if Facebook comments are favourable, negative, or neutral in terms of mood. The suggested system's architecture is depicted in Figure 1.

STEP 1: Input data set - The Kaggle data set of 1000 comments from Facebook users is used in this method.

STEP 2: Pre-Processing — Before processing, several changes and content removal were performed to improve the performance of the classifiers. Pre-processing is done on the input data set with the help of the NLTK library.

Tokenizing: To begin, break down the statement into words.

Lowercase: Change all of the letters to lowercase letters. Punctuation: To eliminate sounds, all punctuation symbols that are irrelevant to the sentiment were removed from the text.

Stop Words: Stop words, such as am, a, the, and others, are words that have no unique meaning in the phrase. These stop words were deleted in the next step of Pre-processing.

POS Tagging: The Part Of Speech tag analyses each word and assigns syntactic labels such as Verb, Noun, or Adjective to each one.

Lemmatization: Finding the natural word root or its base form, 'lemma,' is the first step in lemmatization. The word 'classification,' for example, has the lemma 'classify.'

STEP 3: Extraction of Features: The suggested approach makes use of the Count Vectorizer from the sklearn-library. Creates a sparse matrix that contains all of the words in a document. For example:

I am happy I am sad

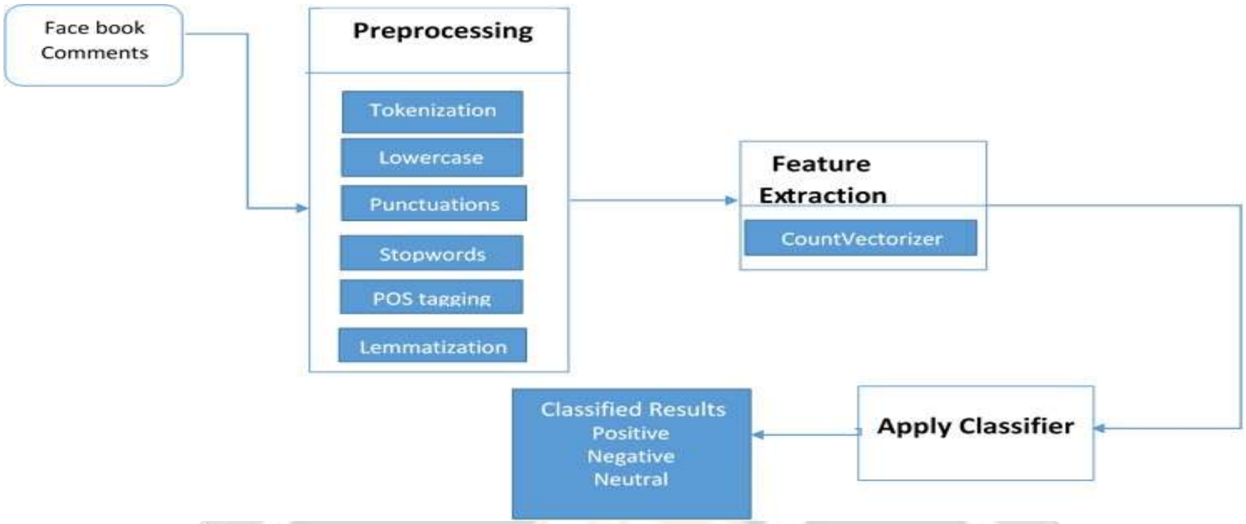
Word index:{I:0,am:1,happy:2,sad:3}

Matrix:	I	am	happy	sad
0	1	1	1	0
1	1	1	0	1

Here, Each vector contains all of the words in the document, which automatically increases the size of the matrix, but the max-count parameter of the Count vectorizer can be used to limit the number of features.

STEP 4: Classification Process: The data is divided into two groups, with 80% of the data being used to train the algorithm and 20% being used to test the taught model. First, each machine learning method was given its own model, which was then used to train the data. Following the learning process, we compared the results to test data and evaluated the model's performance. This procedure was repeated five times for each machine leaning method. The models employed here are explained in depth in the following section.

STEP 5: The Final Product: This is the final step in the architectural process. The Facebook comments are categorised as Positive, Negative, or Neutral.



Classification Models:-

a) **Innocent Bayes (NB)-** This is one of the simplest and most powerful regulated grouping calculations available. The Bayes hypothesis is the foundation of the NB classifier. The Innocent Bayes classifier assumes that predefined properties are unrelated to the presence of another component. The Credulous Bayes Model is useful for large informational collections and is based on the Bayes hypothesis. It determines the relationship between the probabilities p of two events c and Z, referred to as P(c) and P(Z), and the contingent likelihood of event c moulded by event Z and vice versa, referred to as P(c | Z) and P(Z | c). As a result, Baye's Formula is:

$$P(c/Z)=(P(c) \times P(Z/c)) / P(Z)[1]$$

It works well with data with a lot of dimensions and assumes that characteristics are independent. In the NB categorization procedure, less training is required. The Multinomial NB Classification approach was employed in this system, which employs the frequency of words encountered in the document as a feature.

b) Support Vector Machines (SVMs) - give strong responses to high-dimensional input spaces, such as text analysis. SVM also works well with the fact that archive vectors are scarce. The Support Vector Machine's (SVM) goal is to find a hyper plane that best isolates our two independent unmitigated components (positive and negative class), the order problem. Instinctively, the model tries to find a choice limit that can 'best' separate the information esteems based on the probable objective assessment of our 'positive' and 'negative' classes[2] by widening the mathematical edge. This approach avoids overfitting even more precisely, especially when dealing with high dimensional data. In any event, the model preparation will take more time.

(c)Random Forest (RF)- This is a learning calculation based on an ensemble tree. It is concerned with the guideline 'Number of powerless assessors when consolidating structures'. The Random Forest is made up of a few different trees, each of which will be fully developed, so it won't have to cut corners. The more trees it has, the more precise the result will be, and it won't overfit. The irregular woodland calculation will calculate the overall gauge, and it has the advantage of pre-programmed inclusion options, among other things. [4]. There are two types of random forest models:

I) RF forecast for an order issue

f(x)= Majority vote of all anticipated classes over B trees.

ii)RF expectation for a relapse issue

f(x)= Sum of all sub tree expectations partitioned over B-Trees.

Random forests computation has a number of advantages, including being one of the most precise characterization calculations and performing brilliantly in a large informative index.

The classifier's standard is set to 'entropy,' and 50 is given as the estimated number of n-assessors, which represents the number of trees in the woods.

d) K-Nearest Neighbor (KNN)- KNN is used to calculate text arrangement and works well when the preparation set is huge. Consider the set of M marked examples ai, bi1M and the vector An. On the predefined N classes, the classifier predicts An's class name. Using the lion's share vote, the KNN arrangement computation finds An's k nearest neighbours and determines the class name of An. Text order and example acknowledgment are two areas where it is widely used. The KNN works in the following way: It is determined how the test information and all of the preparation tests are separated. Any common approach might be used to determine the separation. Example Separation based on Euclidean geometry. If the separation of the preparatory tests from the question is not exactly or equivalent to the Kth least separation, the K closest neighbour may be used. Then we put together a specific element estimation of all the closest neighbours who are preparing tests. We take the lion's share of this incentive as a projection and set up our new test data[5].

e)Decision Trees(DT)- A decision tree is a type of chart that represents decisions and their results. The diagram's hubs represent an event, while the edges represent a condition of choice. By sectioning branches, we can understand their puzzling challenges. A choice tree model is created using preparation data, and a few approval arrangements are used to check and improve the choice tree model's presentation [13].

Mostly two sorts of Decision Trees are there:

- Characterization Trees-Decision variable is clear cut or discrete.
- Relapse Tress-Decision factors take constant qualities.

The Choice Tree Classifier starts at the top and splits the data into the components that result in the most data gain (IG). This step repeats until all of the leaf hubs are clean. We can avoid overfitting in this model by putting a breaking point on the tree's depth.

Result Analysis: -

We used five different machine learning algorithms to determine the sentiment polarity of Facebook comments in this paper. Performance measures such as Precision, Recall, and F1-score of algorithm were used to evaluate the performance of different classifiers.

Precision=TP/(TP+FP)

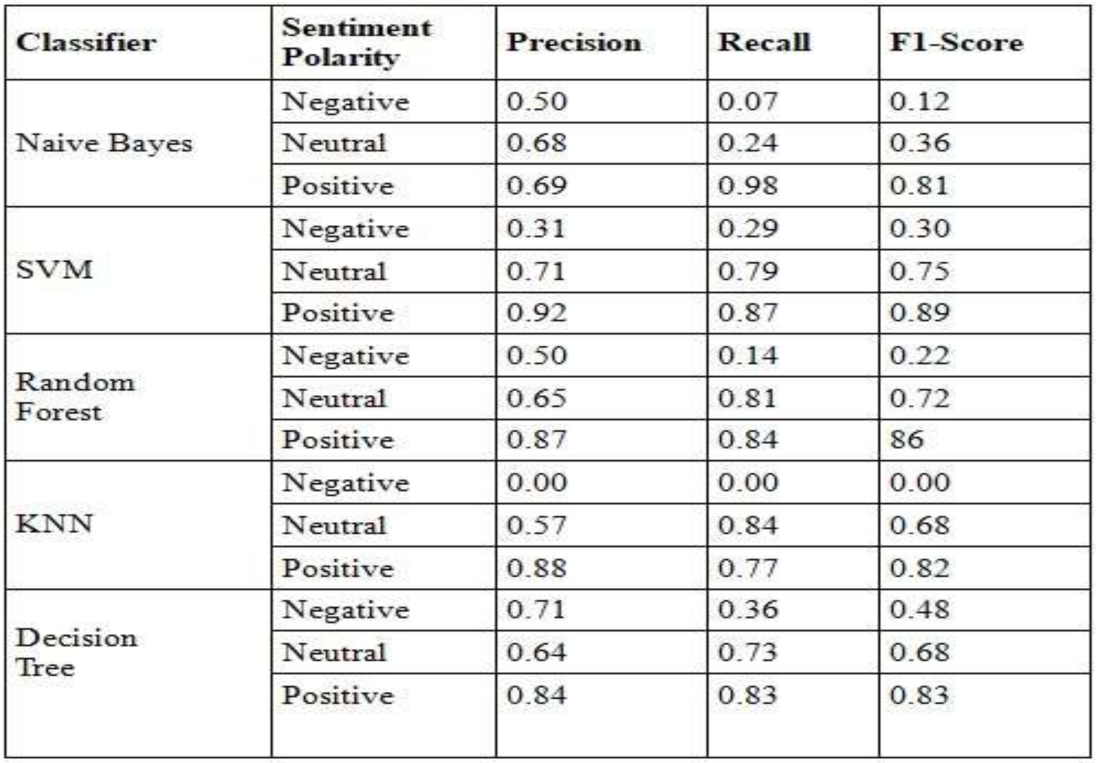
Recall=TP/(TP+FN)

F1-score= 2x ((Precision x Recall)/(Precision+Recall))

Where TP – No. Of True Positives

FP -No. of False Positives

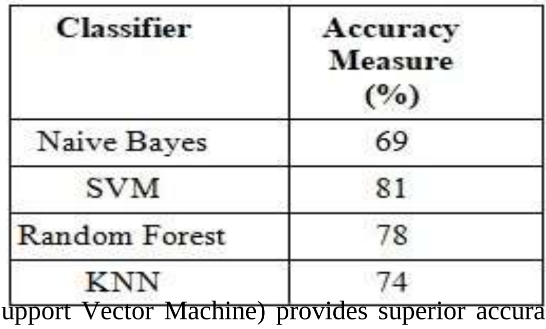
FN – No. Of False Negatives.



Classifier	Sentiment Polarity	Precision	Recall	F1-Score
Naive Bayes	Negative	0.50	0.07	0.12
	Neutral	0.68	0.24	0.36
	Positive	0.69	0.98	0.81
SVM	Negative	0.31	0.29	0.30
	Neutral	0.71	0.79	0.75
	Positive	0.92	0.87	0.89
Random Forest	Negative	0.50	0.14	0.22
	Neutral	0.65	0.81	0.72
	Positive	0.87	0.84	86
KNN	Negative	0.00	0.00	0.00
	Neutral	0.57	0.84	0.68
	Positive	0.88	0.77	0.82
Decision Tree	Negative	0.71	0.36	0.48
	Neutral	0.64	0.73	0.68
	Positive	0.84	0.83	0.83

Table1. Performance Measures of different Classifiers

In the proposed system, Accuracy measure is taken to evaluate the overall performance of different classifiers. Accuracy of all classifiers are In Table 2.



Classifier	Accuracy Measure (%)
Naive Bayes	69
SVM	81
Random Forest	78
KNN	74

We can conclude that SVM (Support Vector Machine) provides superior accuracy than the other four classifiers based on these data acquired after the implementation of classifiers to determine the sentiment polarity of Facebook comments. We simply developed the basic form of classifiers in the proposed system, but it is feasible to increase prediction accuracy by tuning parameters of different classifier models.

Conclusion: -

The current study has provided a review of previous contributions with various discrete information machine learning techniques. The current review looked at 20 research papers that covered a variety of sentiment analysis implementations. Initially, the evaluation focused on determining the contribution of each task and determining the type of machine learning method used. The evaluation also looked at the types of data that were used. The context in which each contribution was created, as well as the performance indicators covered in each contribution, were then examined. Finally, the study gaps and obstacles were discussed, which were helpful in identifying the non-saturated implementation for which sentiment analysis was needed in future studies.

References: -

[1] "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval Vol. 2, No. 1-2 (2008) 1–135, 2008.

[1] R. V. Ogutu, R. Rimiru, and C. Otieno, "Target Sentiment Analysis Model Using Nave Bayes and Support Vector Machines for Product Review Classification," International Journal of Computer Science and Information Security, vol. 17, no. 7, pp. 1–17, 2019.

[1] A. Nisha Jebaseeli and E. Kirubakaran, "M-learning sentiment analysis using data mining approaches," International Journal of Computer Science and Telecommunications, vol. 3, no. 8, pp. 45–48, 2012.

[2] "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," vol. 5, 2017. W. Lin, Z. Wu, L. Lin, A. Wen, and J. I. N. Li, "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," vol. 5, 2017.

[3] B. G. Priya, "Emoji Based Sentiment Analysis Using Knn," vol. 07, no. 04, 2019, pp. 859–865.

[4] "Comparative Analysis of Sentiment," R. S. Jagdale and V. Shirsat, vol. 5, no. February, pp. 1190–1195, 2018.

[5] Shakila Basheer; S Mariyam Aysha Bivi; S Jayakumar; Arpit Rathore; Balajee Jeyakumar. Journal of Computational and Theoretical Nanoscience, Volume 16, Numbers 1-2, Machine Learning-Based Classification of Cervical Cancer Using K-Nearest Neighbour, Random Forest, and Multilayer Perceptron Algorithms 5-6, May 2019, pp. 2523-2527(5).

[6] K. Srividya and A. M. Sowjanya, "Sentiment analysis of Facebook data using nave Bayes classifier Assistant Professor, Department of Computer Science and Engineering, AU College of Engineering (A), Andhra," vol. 15, no. 1, pp. 179–186, 2017.

[7] A. Bayhaqy, S. Sfenrianto, K. Nainggolan, and E. R. Kaburuan, "Sentiment Analysis of E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Nave Bayes," 2018 Int. Conf. Orange Technol. ICOT 2018, pp. 1–6, doi: 10.1109/ICOT.2018.8705796.

[8] Sultana, H Parveen; Shrivastava, Nirvishi; Dominic, Dhanapal Durai; Nalini, N; Balajee, J.M. Sultana, H Parveen; Shrivastava, Nirvishi; Dominic, Dhanapal Durai; Nalini, N; Balajee, J.M. Journal of Computational and Theoretical Nanoscience, Volume 16, Numbers 56, May 2019, pp. 2541-2549, Comparison of Machine Learning Algorithms to Build Optimized Network Intrusion Detection System (9).

[9] R. S. Soundariya, M. Nivaashini, R. M. Tharsanee, and P. Thangaraj, "Application of various machine learning techniques in sentiment analysis for depression detection," International Journal of Innovative Technology and Engineering, vol. 8, no. 10 Special Issue, pp. 292–297, 2019, DOI: 10.35940/invitee.J1052.08810S19.

[10] D. Ramalingam, V. Sharma, and P. Zar, "Depression analysis using machine learning techniques," Int. J. Innov. Technol. Explore. Eng., vol. 8, no. 7C2, pp. 187–191, doi: 10.35940/ijitee.h7163.0881019, 2019.

Vinoth Kumar V, Karthikeyan T, Praveen Sundar P V, Magesh G, Balajee J.M.

[11] Vinoth Kumar V, Karthikeyan T, Praveen Sundar P V, Magesh G, Balajee J.M. (2020). Quantum Key Distribution is used in a quantum approach to LiFi security. 2345-2354 in International Journal of Advanced Science and Technology, 29(6s).

[12] Using Machine Learning Methods in the Financial Market for Technical Analysis Based on Hybrid Models. Sabbagh Lalimi A H, Damavandi H. 2020; 2(4):1-11; sjambok. 2020; 2(4):1-11; sjambok. 2020; 2(4):1-11