IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Sentiment Analysis in Twitter Data

[1]K.L.Sudha, [2]Prerana B Patil

[1]Professor, DSCE, Bengaluru, [2]PG student, DSCE, Bengaluru

Abstract: **Nowadays, people from different parts of the world use social media sites to share messages. For example, Twitter is a platform in which users send, read posts known as 'tweets' and network with different groups. Users share their daily lives, post their feelings on everything such as brands and places. Companies can benefit from this massive platform by collecting data related to opinions on them. The aim of this paper is to propose a model that can perform sentiment analysis of real data collected from Twitter. Data in Twitter is highly unstructured which makes it difficult to analyze. Our proposed model combines the use of supervised and unsupervised machine learning algorithms to do the task. Tweet is extracted directly from Twitter API, then cleaning and discovery of data is performed. After that, the data will be fed to several Machine learning models for the purpose of training. Each tweet extracted is classified based on its sentiment whether it is a positive, negative or neutral.**

Index Terms **- Machine learning, twitter, classification algorithms**

I. INTRODUCTION

Sentiment is defined as thought, view or attitude especially one based mainly on emotion instead of reason as per dictionary. Emotions play an important role in stating the feeling among human beings. Sentiment analysis or opinion mining is the technique that is used to study people's emotions, attitudes, sentiments, evaluations, moods, and opinions. Sentiment analysis is the stem of natural language processing and machine learning methods, which is the current trending research area in the text mining. It is the important source of decision making and it can be extracted, Identified, evaluated from the on line sentimental reviews. Social media is a platform that gives a chance to express the opinion of the people. The sentiment analysis can be applied on the social media platforms such as twitter, instagram, facebook, etc. The impression or opinion of the people with respect to some particular topic varies from person to person, thus it is necessary to collect all kinds of the opinions for finding the exact sentiment about a topic.

Twitter is an American micro blogging and social networking service, on which users post and interact with messages known as "tweets". Twitter is a reservoir for a large amount of data. Twitter is one of the open source social media that provides the chance to access its data. Sentiment analysis focuses on the polarity of a text such as positive, negative, neutral but it also goes beyond polarity to detect specific feelings and emotions like angry, happy, sad, etc, whether something is urgent, or not urgent and even intentions such as interested, not interested. So, this data is extremely useful for predicting results of political activities, new initiatives led by government, or research and deciding on what content to share with the audience. These opinions can be tapped and used as business intelligence for various uses such as marketing, prediction, etc. Thus there is a need for efficient algorithms on sentimental analysis. Machine learning algorithms like SVM, Naive Bayes etc are commonly used in predicting the polarity of the sentence.

II. MACHINE LEARNING ALGORITHMS FOR SENTIMENT ANALYSIS

Machine learning (ML) is a subfield of artificial intelligence (AI) that allows computers to learn to perform tasks and improve performance over time without being explicitly programmed. There are a number of important algorithms that help machines such as compare data, find patterns, or learn by trial and error to eventually calculate accurate predictions with no human intervention.

Commonly used machine learning algorithms for sentiment analysis are

A. Logistic Regression is a technique for predicting the outcome of a situation. Logistic regression is a model that may be used to solve both regression and classification issues. It can be used to classify tweets into Positive and Negative categories. There are two types of regression models: linear and logistic. For randomly selected observations, the Logistic Regression measures the chance of a word occurring with the probability that the term does not occur.

B. Naive Bayes multinomial algorithm is a probability algorithm based on the Nave Bayes theorem which includes a multinomial Nave Bayes component. It is a Bayes Theorem-based categorization algorithm. The multinomial distribution is used by Multinomial Nave Bayes to determine the frequency of a specific word in a given text.

C. Random Decision Forest or Random Forest algorithm is an ensemble learning method that can be used for classification or regression, depending on the situation. The root, splitting, decision nodes, and leaf of a decision tree combine to generate a structure that can be utilized to identify the problem area. The random forest method works by combining many decision trees to produce superior results.

D. Support Vector Machine algorithm, commonly known as SVM algorithm is a supervised machine learning technique that can be used to solve classification and regression problems. The SVM method is a mathematical classification strategy that aims to maximize the margin between instances and the separation hyper-plane. SVM classifies data by selecting a hyper-plane that divides the groups in the n-dimensional space.

III. DEEP LEARNING APPROACH FOR SENTIMENT CLASSIFICATION

Deep learning (DL) is considered as evolution of machine learning. It combine together algorithms that aim to simulate how the human brain works. With the help of deep learning, sentiment analysis models can be trained to understand text beyond simple definitions, read for context, sarcasm, etc., and understand the actual mood and feeling of the writer. Following are few DL algorithms which are tried in sentiment analysis.

A. **Bidirectional Encoder Representations from Transformers** approach also known as Bidirectional Encoder Representations from Transformers (BERT) can be used to pre-train a general-purpose classification model and fine-tune it to a specific goal. BERT takes undescribed text as an input, masking 15% of the words, and then attempting to predict the remaining words. By training a simple job of predicting the input sentence, the BERT Model additionally keeps track of the context between sentences.

B. Long Short-Term Memory (LSTM) networks are a type of recurrent neural network that can learn long-term dependencies. RNNs are made up of a sequence of replicating neural network modules, each of which has a simple structure, such as a single tanh layer. Although LSTM has a series-like layout, the repetition module has a distinct appearance. For encoding written texts, LSTM modules are widely employed. The word embedding layer, LSTM layer, fully connected layer, and sigmoid activation layer are the four layers that make up the LSTM model architecture. Rather than establishing a single neural network layer, it acts independently.

IV. LITERATURE REVIEW:

As said before, Sentiment analysis or opinion mining or emotion extraction is the classification of emotions within a textual data. This technique has been widely used over the years in order to determine the sentiments, emotions within a particular textual data. Many authors in recent years are working on this to improve accuracy of decision.

Authors in paper by Sheresh Zahoor et.al,[1], have collected tweets from twitter for a number of events, analyzed them using a number of Machine Learning algorithms like Naive Bayes, SVM, Random Forest classifier and LSTM and compared the results. Finally, the authors concluded that the sentiments can be predicted with more accuracy using Machine Learning and Deep Learning algorithms especially Naive Bayes, SVM, Random Forest Classifier and LSTM.

Paper by Yash Indulkar et.al,[2] considered 3 algorithms for sentiment analysis and these are Logistic Regression, Multinomial Naive Bayes & Random Forest on the Uber & Ola datasets. The reason they choose these three algorithms was that the potential that these machine learning algorithms could have. The number of tweets extracted from Twitter is 3000. These tweets are cleaned & tokenized using python. It is observed that from the three algorithms used, the best accuracy was generated from Random Forest for the respective datasets.

Anam Yousaf et.al,[3] considers Twitter data to collect views about products, trends, and politics. Seven Machine Learning models are implemented for emotion recognition by classifying tweets as happy or unhappy. This paper proposed a novel combination of LR and SGD as a voting classifier for emotion recognition by classifying tweets as happy or unhappy.

The rating from the online shopping website flipkart.com is analyzed in paper by P. Karthika et.al,[4]. Based on the aspects of the product, the rating is classified as positive, neutral and negative. The proposed work is analyzed by using Machine Learning algorithm called Random Forest and simulated by using SPYDER. Random Forest gives the best accuracy of 97% among other algorithms and the Support Vector Machine (SVM) gives the accuracy of 92%.

Paper by Teddy Mantoro et.al,[5] analyses Crime information by choosing some appropriate keywords. Eight keywords have been chosen which represented the most viral topic. The keywords in this study were analyzed regarding their sentiment from the hashtags in twitter posts. The Machine Learning algorithms were utilized such as Multinomal Naive Bayes, Random Forest Classifier, Linear SVM, and Nearest neighborhood (kNN) finding a better classifier.

The process of data cleaning and data preparation for sentiment analysis is discussed by Mr. Kaushik Dhola et.al,[6]. The authors present experimental findings that demonstrate the comparative performance analysis of various classification algorithms. The authors concluded a comparative study on various machine learning as well as deep learning techniques to perform the sentiment classification.

Paper by Roza H. Hama Aziz et.al,[7] provides an ensemble of classifiers framework for sentiment analysis. They proposed weighted majority voting ensemble method, which combines six models including Naive Bayes, Logistic Regression, Stochastic Gradient Descent, Random Forest, Decision Tree and Support Vector Machine to form a single classifier. Weights of the individual classifiers of the ensemble are chosen as accuracy or F1-score by optimizing their performance. It can be observed that the proposed method achieved highest performance compared with all individual models and simple majority voting.

Sentiment on the YouTube video comments is analyzed by Abbi Nizar Muhammad et.al,[8]. The process of understanding, extracting, and processing textual data automatically to obtain sentiment information contained in one sentence of YouTube video comment is explained in this paper. Naive Bayes and Support Vector Machine is extensively used as a basic line in tasks related to texts but the performance varies significantly in all variants, features, and numbers of data collection.

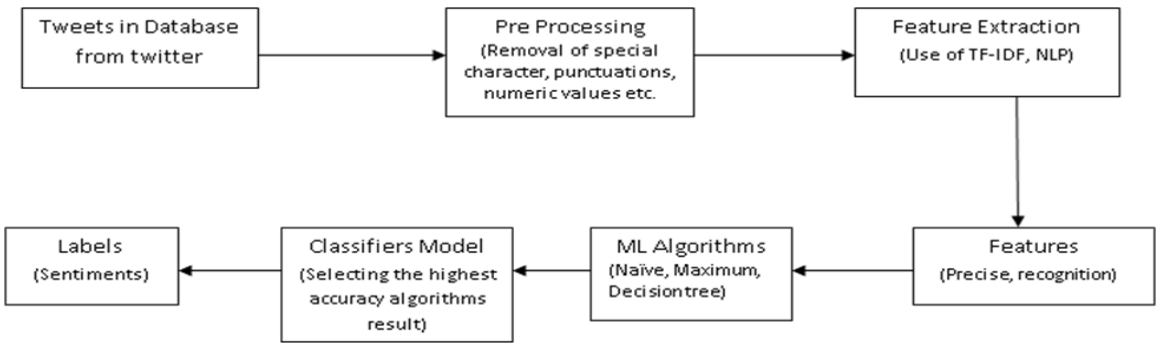
Table 4.1: Comparison of research on sentiment analysis

Sl. No	Reference paper	Data Set considered	Algorithms used	Accuracy achieved	Further improvements
1.	[1]	Twitter data related to different political parties	Naive Bayes, SVM, Random Forest classifier and LSTM	88-98%	Suitability of algorithms for different applications not compared.
2.	[2]	Twitter Uber & Ola datasets.	Logistic Regression, Multinomial Naive Bayes & Random Forest	84-92%	vocabulary problems to be solved
3.	[3]	IEMOCAP and CMU-MOSEI dataset	Multimodal emotion recognition methods	Around 75%	More robust and prediction models can be used
4.	[4]	Twitter data of flipcart products	voting classifier(LR-SGD) with TF-IDF	79% with good stability	Accuracy to be improved
5.	[5]	Crime information	Multinomal Naive Bayes, Random Forest Classifier, Linear SVM, and Nearest neighborhood (kNN)	82-92%	non-text posts are not considered
6.	[6]	Publicly available dataset which contains the Twitter posts	ML & deep learning techniques	>96% with deep learning method	Can use improved DL algorithms
7.	[7]	SemEval-2017, Task 4 dataset extracted from twitter	weighted majority voting ensemble method-based on 6 ML algorithms	Outperform all earlier algorithms	imbalanced dataset not considered
8.	[8]	YouTube video comments	Naive Bayes and Support Vector Machine	Precision of 91% achieved	DL can be tried to improve performance

In these studies, researchers have considered twitter data related to different aspects or products and have tried to analyse the sentiment of people who have posted tweets.

V. PROPOSED METHODOLOGY:

Figure (5.1) shows the process followed to analyze sentiment in twitter data. Different feature sets and machine learning classifiers will be considered to determine the best combination for sentiment analysis of twitter. Pre-processing steps like – removal of punctuations, emotions, twitter specific terms and stemming will be considered for experimenting. From the literature, the following features - unigrams, bigrams, trigrams and negation detection are identified as important. Finally classifier will be trained using various machine-learning algorithms - Naive Bayes, Decision Trees and Maximum Entropy. A new feature vector will be used for classifying the tweets as positive, negative and extract peoples' opinion about products.



Figure(5.1) Process Flow Diagram

DATA COLLECTION:

Twitter equips appliance to a mass data through its Application Programming Interface (API). The streaming mechanism grabs the input information Tweets and operates any anatomizing, percolating, or aggregation mandatorily anterior to accumulating the outcome to a data store. Tweets collected using Twitter API are manually annotated as positive or negative. A dataset is created by taking 600 positive tweets and 600 negative tweets.

Pre-processing of Tweets

A pre-processing step is performed before feature extraction. Pre-processing steps include removing URL, avoiding misspellings and slang words. Misspellings are avoided by replacing repeated characters with 2 occurrences. Slang words contribute much to the emotion of a tweet. Hence, a slang word dictionary is maintained to replace slang words occurring in tweets with their associated meanings.

Creation of Feature Vector:

Feature extraction is done in two steps. In the first step, twitter specific features are extracted. Hash tags and emoticons are the relevant twitter specific features. Emoticons can be positive or negative. So, they are given different weights. Positive emotions are given a weight of +1 and negative emotions are given a weight of -1. There may be positive and negative hash tags. Therefore, the count of positive hash tags and negative hash tags are added as two separate features in the feature vector. Twitter specific features may not be present in all tweets. So, a further feature extraction is to be done to obtain other features. After extracting twitter specific features, they are removed from the tweets. Tweets can be then considered as simple text. Then using unigram approach, tweets are represented as a collection of words. In unigrams, a tweet is represented by its keywords. A negative keyword list, positive keyword list and a list of different words are maintained that represent negation. Counts of positive and negative keywords in tweets are used as two different features in the feature vector. Presence of negation contributes much to the sentiment.

Adverb or verb shows more emotions. If a relevant part of speech can be determined for a keyword, then that is taken as special keyword. Otherwise a keyword is selected randomly from the available keywords as special keyword. If both positive and negative keywords are present in a tweet, selection of any keyword having relevant part of speech will be done. If relevant part of speech is present for both positive and negative keywords, none of them is chosen. Special keyword feature is given a weight of '1' if it is positive and '-1' if it is negative and '0' in its absence. Part of speech feature is given a value of '1' if it is relevant and '0' otherwise. Thus, feature vector is composed of 8 relevant features. The 8 features used are part of speech (POS) tag, special keyword, and presence of negation, emoticon, number of positive keywords, number of negative keywords, number of positive hash tags and number of negative hash tags. After creating a feature vector, classification is done using Naive Bayes, Support Vector Machine, Maximum Entropy and Ensemble classifiers and their performances are compared.

To apply ML algorithms, proper data set need to be considered. The DATASET may have a lot of contradictory tweets in it. Using the symbols 1 and 0, each record is categorized as joyful or unhappy depending on its sentimental polarity. English-language tweets are remembered in the final dataset. There are a variety of features in the dataset.

DATA VISUALIZATION is a term that refers to the visualization of data. Data visualization aids in the discovery of hidden patterns within a dataset. It aids in qualitatively obtaining additional information about the dataset by showing the attributes' properties.

Random Forest is a supervised learning technique that is used to solve problems like regression and classification. A random forest is simply a collection of trees, each of which is distinct from the others. It creates numerous decision trees and then merges them to provide an absolute and stable result, which is mostly used for training and class output. The algorithm flow is as given below

Step 1: Load flip kart dataset and apply random forest algorithm.

Step 2: The required records were selected and the decision tree is created depending on the record.

Step 3: The decision making process is done based on the class value.

Step 4: If the class value is less than the threshold value then it is considered as false or else it is considered as true.

Step 5: The performance of random forest algorithm is compared with SVM algorithm based on the Performance Metrics such as accuracy, precision, F-measure and recall.

In this study, two types of data sets will be used: training (1600 data points) and testing (1600 data points). The data provides user comments or reviews on a given phrase from Natural Language Toolkits (NLTK) in the Python library in English, indicating whether they are satisfied or not. Before it was used to test the sentiment of a review on a specific topic, the Nave Bayes approach was used to train both positive and negative remarks.

Naïve Bayes method uses probability to classify the data into some categories. Compared with SVM the Naïve Bayes not only states the positive or negative sentiments but also present their probability.

The Jupyter collaboration tool in Python language will be used in this work, and it can be accessed through its official website at https://colab.research.google.com. Google's GPU server, as well as certain associated libraries, is used in this Jupyter-based online Python. Before the training phase, the Nave Bayes tool and the corpus were loaded from NLTK. The NLTK corpus is also available in the Jupyter collaboration tool via the download function in NLTK after importing the libraries.

Accuracy Assessment will be done at the end to compare the results of true prediction.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Where, TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative respectively.

VI. **CONCLUSION:** Sentiment analysis or opinion mining is the technique that is used to study people's emotions, attitudes, sentiments, evaluations, moods, and opinions on social media. This paper reviews the work done in this area recently and proposes a model for analyzing twitter data to identify opinions of people on different products. Collecting the data and classifying it for the required accuracy is done with the help of machine learning algorithms.

REFERENCES

[1]. Sheresh Zahoor, Rajesh Rohilla "Twitter Sentiment Analysis using Machine Learning Algorithms: A Case Study", IEEE International Conference on Advances in Computing, Communications and Materials (ICACCM), 21 August 2020.

[2]. Yash Indulkar, Abhijit Patil "Comparative Study of Machine Learning Algorithms for Twitter Sentiment Analysis", IEEE International Conference on Emerging Smart Computing and Informatics (ESCI), 5 March 2021.

[3]. Anam Yousaf, Muhammad Umer, Saima Sadiq, Saleem Ullah, Seyedali Mirjali, Vaibhav Rupapara and Michele Nappi "Emotion Recognition by Textual Tweets Classification Using Voting Classifier (LR-SGD)", IEEE Access, volume 9, pp no. 6286 – 6295, 28 December 2020.

[4]. P. Karthika, Dr. R. Murugeswari, Mrs. R. Manoranjithem "Sentiment Analysis of Social Media Network Using Random Forest Algorithm", IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 13 April 2019.

[5]. Teddy Mantoro, Rahmadya Trias Handayanto "Machine Learning Approach for Sentiment Analysis in Crime Information Retrieval", IEEE 3rd International Conference on Computer and Informatics Engineering (IC2IE), 15 September 2020.

[6]. Mr. Kaushik Dhola, Mr. Mann Saradva "A Comparative Evaluation of Traditional Machine Learning and Deep Learning Classification Techniques for Sentiment Analysis", IEEE 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 28 January 2021.

[7]. Roza H. Hama Aziz, Nazife Dimililer "Twitter Sentiment Analysis using an Ensemble Weighted Majority Vote Classifier", IEEE International Conference on Advanced Science and Engineering (ICOASE), 23 December 2020.

[8]. Abbi Nizar Muhammad, Saiful Bukhori, Priza Pandunata "Sentiment Analysis of Positive and Negative of YouTube Comments Using Naive Bayes - Support Vector Machine (NBSVM) Classifier", IEEE International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), 16 October 2019.