Twitter Sentiment Analysis

Om Koli, Shubhada Londhe, Prasad Kharche, Sharad Sawant

Department of Artificial Intelligence and Data Science
All India Shri Shivaji Memorial Society, Institute of Information Technology

Abstract- The sentiments behind opinions in texts that might have a public interest must be analyzed. Sentiment analysis is opinion mining as a task of preparing sentiments on varying subjects. An emotional tone behind a textual voice makes it easy for the readers to clarify the specific topics. Sentiment analysis uses the Natural Language Processing (NLP) approach to identify the emotional tone from the textual context. Sentiment analysis is used to have positive and negative or neutral opinions. The different applications that give us a briefer idea or express sentiments on public topics are TWITTER, INSTAGRAM, and HIKE, etc. Being specific about Twitter it's one of the most popular applications used between the age group of 25 to 34 years, with a grouping accounting for 38.5 percent of social platforms worldwide user base in the year 2021. Additionally, the male population has made it up 56.4 percent of Twitter's audience. Let's come down to the statistical ratio of political reviews on Twitter that needs to have a sentimental analysis to have a good approach towards political ideologies. Twitter generates a huge amount of text containing political fights, and insights, that need to be mined to analyze. The political reviews, insights, and distributions demonstrate divided politically divided groups on political views. Political popularity may reflect the sentiments of people towards that person, organization, or place. Such political reviews need to be analyzed to predict an exact positive approach to a particular political group. Twitter data has much more attention over the last decade and involves dissecting the tweets into positive and negative sentiments. This enormous popular microblog depicts the client's voice needs sentimentally analyzed. This paper involves the sentiment analysis applied to Twitter comments and their outcomes.

Index Terms- Logistic Regression, Tweet Analysis, Text Pre-processing, Naïve Bayes

I. INTRODUCTION

There are a tonne of these unanalysed opinions on Twitter. These Twitter-related thoughts can be characterized as textual emotions that call for sentimental investigation. Imagine a political figure using a social media platform to make his case in public. However, while the opinions of each leader are favourable to their supporters, they are viewed negatively by their rivals. There may be areas where the subject is seen as neutral. For a clearer understanding of the current politics of the general population, these viewpoints need to be examined. Sentiment analysis, commonly referred to as opinion mining, is a branch of natural language processing (NLP) that includes robotically locating and extracting the sentiment or opinion included in a text. It is used to identify whether a text is neutral, positive, or negative towards a certain thing, like a good, service, or subject. The objective of sentiment analysis is to create machine learning algorithms that can reliably and quickly sift through massive amounts of text data, including social media postings, product reviews, news stories, and consumer feedback, to determine the overall sentiment that is being represented in the text. Numerous uses for sentiment analysis exist, including brand monitoring, customer support, market research, and political analysis.

I. BASIC IDEA OF TWITTER SENTIMENT ANALYSIS

Let's take into consideration a basic architectural diagram used in sentiment analysis. Sentiments are words or sentiments that represent a view or opinion that is held by the user expressing positive, negative, or neutral emotions. Sentiment analysis includes features like emoticons, neutralization, negation handling, and capitalization as they are a huge part of feature extraction. Sentiment analysis includes data collection, data pre-processing, data extraction, and data and data classification.

Taking data collection into consideration, Twitter consists of various sentiments, and emotions being expressed every second by thousands of Twitter handlers. The emotions vary from every field on every single topic. To uncover these sentiments, one needs to collect the data, process, classify these sentiments for analysis.

The process of data collection is explained in detail on page this of this research paper. However, taking a broader view of the same.
Data collection is the second step comes in sentiment analysis. Framing the problem, we want to solve is the first step for the same. This forms the backbone for the rest of the phases. Social media is a great place to collect data, especially for sentiment analysis. According to statistics for 2017, Twitter has 313000000 active users which gives a large amount of data for analysis. Almost 500 tweets are sent per day on Twitter. The fairly generous APIs of Twitter allow us to load these data datasets using specific guidelines. A few such tools are TAGS, APIFY, and TWARC.

II. DIFFERENT ALGORITHMS USED FOR SENTIMENT ANALYSIS

Now let's see how different algorithm work:

A. Logistic Regression

Logistic regression is a supervised machine-learning algorithm that is mainly used for classification. It transforms linear regression function continuous value output into categorical value output using a sigmoid function, it maps a set of independent variables input into values between 0 and 1.

For our project of tweet sentiment analysis, it classifies tweets in positive or negative. To use this algorithm, we first pre-process the dataset.

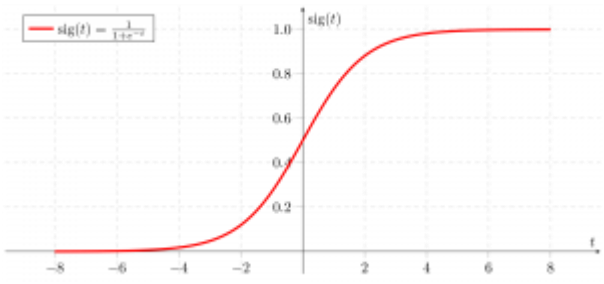
steps involved in pre-processing: -

1. Remove all the stop words.
2. remove all unwanted characters and symbols
3. convert all text into lowercase
4. Apply stemming

we use the sigmoid function as the activation function in logistic regression.

sigmoid activation function: -

sigmoid activation curve looks like S - shape. We use sigmoid function is because it exists between (0 t o1). Therefore, it is especially used for models where we have to predict the output. Sigmoid is the best choice to choose.



Logistic Regression

Logistic regression is used for the classification of discrete variables. Here we are going to classify them into Positive or Negative tweets, so after pre-processing we are going to apply the logistic regression model.

Steps performed in logistic regression:

1) Data Pre-processing
2) Fitting Logistic Regression to the Training set.
3) Predicting the test result.
4) Test accuracy of the result.
5) Visualizing the test set result.

B. Random Forest

Random forest is a commonly used machine learning algorithm. Random forest is trademarked by Leo Bierman and Adele Cutler. It combines the output of multiple decision trees to reach a single result.

Random forest is extremely popular and is used for Classification and Regression problems in Machine Learning. It is based on the concept of ensemble learning. It is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.
A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

The following steps explain the working Random Forest Algorithm:
Step 1: Select random samples from a given data or training set.
Step 2: This algorithm will construct a decision tree for every training data.
Step 3: Voting will take place by averaging the decision tree.
Step 4: Finally, select the most voted prediction result as the final prediction result.
Features of a Random Forest Algorithm

- It can produce a reasonable prediction without hyper-parameter tuning.
- It solves the issue of overfitting in decision trees.
- In every random forest tree, a subset of features is selected randomly at the node's splitting point.
- It is unexcelled in accuracy among current algorithms.
- It runs efficiently on large data bases.
- It can handle thousands of input variables without variable deletion.
- It's more accurate than the decision tree algorithm.
- It provides an effective way of handling missing data
- It gives estimates of what variables are important in the classification.
- It generates an internal unbiased estimate of the generalization error as the forest building progresses.

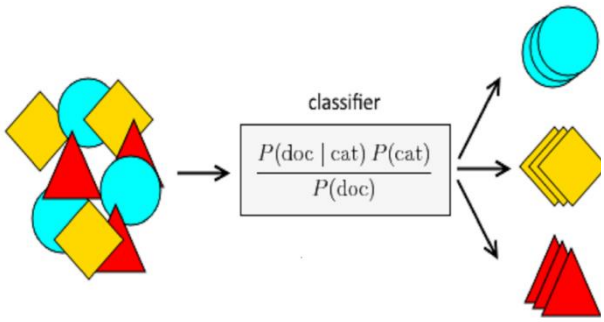
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the It data are missing.
- It has methods for balancing error in class population unbalanced data

C. Naïve Bayes

Naive Bayes is a simple but surprisingly powerful algorithm for predictive modelling. Naive Bayes is a simple supervised machine learning algorithm that uses the Bayes' theorem with strong independence assumptions between the features to procure results. For the algorithm, phrases like 'I like Harry Potter', 'Harry Potter like I', and 'Potter I like Harry' are the same. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.

$$P(h|d) = (P(d|h) * P(h)) / P(d)$$

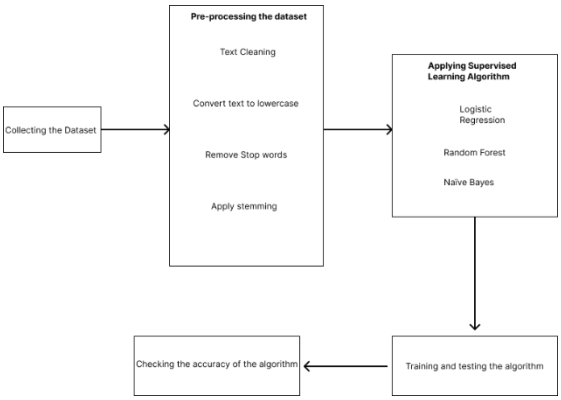
- P(h): the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h.
- P(D): the probability of the data (regardless of the hypothesis). This is known as the prior probability.
- P(h|D): the probability of hypothesis h given the data D. This is known as posterior probability.
- P(D|h): the probability of data d given that the hypothesis h was true. This is known as posterior probability.



Types of Naïve Bayes Classifiers –

1. Multinominal Naïve Bayes Classifier
2. Bernoulli Naïve Bayes Classifier
3. Gaussian Naïve Bayes Classifier

III. SENTIMENT ANALYSIS ARCHITECTURE



Let's take into consideration a basic architectural diagram used in sentiment analysis. Sentiments are words or sentiments that represent a view or opinion that is held by the user expressing positive, negative, or neutral emotions. Sentiment analysis includes features like emoticons, neutralization, negation handling, and capitalization as they are a huge part of feature extraction. Sentiment analysis includes data collection, data pre-processing, data extraction, and data and data classification.

Taking data collection into consideration, Twitter consists of various sentiments, and emotions being expressed every second by thousands of Twitter handlers. The emotions vary from every field on every single topic. To uncover these sentiments, one needs to collect the data, process, classify these sentiments for analysis.

The process of data collection is explained in detail on page this of this research paper. However, taking a broader view of the same.

Data collection is the second step comes in sentiment analysis. Framing the problem, we want to solve is the first step for the same. This forms the backbone for the rest of the phases. Social media is a great place to collect data, especially for sentiment analysis. According to statistics for 2017, Twitter has 313000000 active users which gives a large amount of data for analysis. Almost 500 tweets are sent per day on Twitter. The fairly generous APIs of Twitter allow us to load these data datasets using specific guidelines. A few such tools are TAGS, APIFY, and TWARC.

IV. COMPARISON OF ALGORITHMS

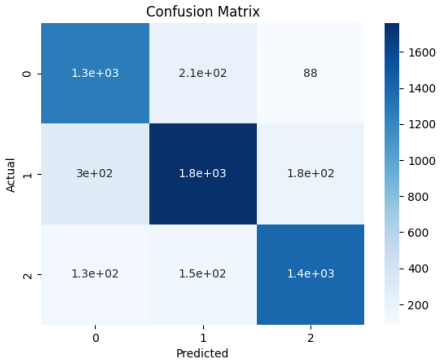
In the project we used above mentioned algorithms to see difference in perform for each of them.

We have the labelled dataset of different kind of tweets. In total we have 27481 rows and 4 columns which are textID, text, selected_text, sentiment.
We tested each mentioned algorithm on the labelled dataset.

1) Logistic regression:

Accuracy – We tested 20% percent of the datasets and trained the algorithm with 80% of the datasets, after training the dataset **we get the accuracy of 80.71%** using logistic regression.

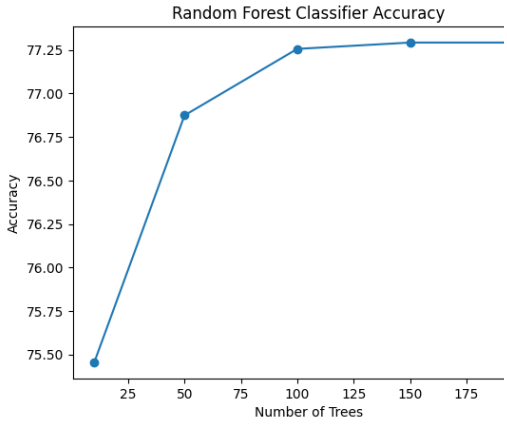
Visual representation using confusion matrix:



2) Random Forest:

Accuracy – Here also we have tested 20% percent of the datasets and trained the algorithm with 80% of the datasets, after training the dataset **we get the accuracy of 77.26%** using random forest algorithm.

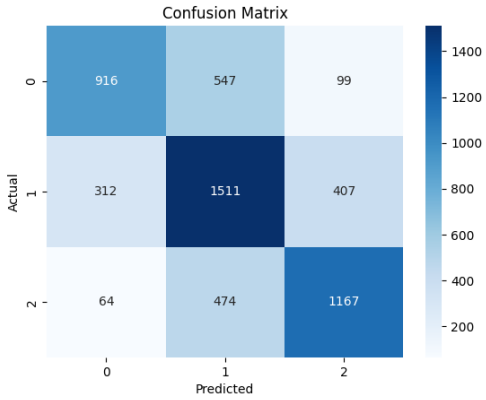
Visual representation using decision trees:



3) Naïve Bayes:

Accuracy – Same as above two here we have tested 20% percent of the datasets and trained the algorithm with 80% of the datasets, after training the dataset **we get the accuracy of 65.40%** using Naïve Bayes.

Visual Representation using confusion matrix:



V. CONCLUSION

As you can see, we have test three different algorithms on the tweet dataset of 27481 rows and 4 columns for sentiment analysis those are Logistic Regression, Random Forest and Naïve Bayes with the accuracy score of 80.71%, 77.26% and 65.40% respectively. Logistic Regression algorithm outperforms Random Forest and Naïve Bayes in Sentiment analysis, because Logistic Regression is the linear model and it tries to find a linear boundary to separate the classes whereas, Naïve Bayes and Random Forest are non-linear model, which may not be as effective in finding the optimal boundary.
Logistic Regression works well with small datasets and handles noise and irrelevant features as well. Logistic Regression's simplicity and ability to handle noise and irrelevant features make it an effective algorithm for sentiment analysis.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to all those who supported and contributed to the completion of this research paper on tweet analysis. First and foremost, we would like to thank our guide, **Mrs. Mayura Shelke**, for her invaluable guidance, support, and encouragement throughout the entire research process. Her expertise and insightful feedback have been instrumental in shaping the direction of this study.

We are grateful to our AISSMS - IOIT for providing us with the necessary resources and facilities to carry out this research. The access to computational resources and the availability of relevant literature greatly contributed to the success of this project.
We would also like to extend our appreciation to the participants who generously shared their tweets for analysis. Without their cooperation and willingness to contribute their data, this research would not have been possible.

LITERATURE REVIEW

- *Systematic reviews in sentiment analysis: A tertiary study*

Alexander Ligthart, Cagatory Catal, and Bedir Tekinerdogan explain the computational study of analyzing people's feelings and opinions for an entity which is called sentiment analysis. The research paper was published in the journal named "ARTIFICIAL INTELLIGENCE REVIEW" in the year 2021. The paper explains how this field of sentiment analysis is a matter of extensive research in the past decade. In this paper, the author presents the results of a tertiary study, which aims to investigate the current state of research in this field by synthesizing the results of published secondary studies on sentiment analysis.
The tertiary study follows the guidelines of systematic literature reviews(SLR) and covers only secondary studies. Different features, algorithms, and datasets used in sentiment analysis are mapped in this research paper. According to the analysis made by the authors, LSTM and CNN algorithms are the most used deep learning algorithms for sentiment analysis.
ORIGINAL LANGUAGE: ENGLISH
JOURNAL: Artificial intelligence review
VOLUME: 54
ISSUE NO: 7
PUBLISHED STATUS: PUBLISHED 2021

- *SENTIMENT ANALYSIS OF STUDENT FEEDBACK USING MULTI-HEAD ATTENTION FUSION MODEL OF WORD AND CONTEXT EMBEDDING FOR LSTM*

This article was published in the year 2022 by an Indian author named K.Sangeetha and
D. Prabha in the Journal of Ambient Intelligence and Humanized Computing. The article explains how nowadays sentiment analysis using deep learning models has gained good performance. This especially ensemble LONG SHORT–TERM MEMORY (LSTM) with attention layers giving more attention to the emotions. In the proposed method, input sequences of the sentence are processed parallel across a multi-head attention layer with fine-grained embeddings and tested with different dropout rates to increase accuracy. Later in this paper, the author concludes the fusion of multiple layers accompanied by LSTM improves the result over a common Natural Language Processing method.

- *TWEET ANALYSIS BASED ON DISTINCT OPINOPINIONSSOCIAL MEDIA USERS*

This paper was published by authors S. Geetha and Kaliappan Vishnu Kumar in the year 2018. The paper is a part of the ADVANCES IN INTELLIGENT SYSTEM AND COMPUTING book series(AISC. VOLUME 750. The paper explains how the state of mind is expressed via Emojis and test Messages for the huge population. It says that the supervised text classifier is used for sentiment analysis in both general and specific emotion detection with more accuracy. The main objective remains to include intensity for prediction of different text formats from Twitter, by considering a text context associated with the emoticons and punctuations. The paper maps the novel "FUTURE PREDICTION ARCH ARCHITECTURE IN EFFICIENT CLASSIFICATION (FPAEC) which was designed with various classification algorithms such as Fisher's linear discriminant classifier, Support Vector Machine (SVM), Nïve Bayes classifier and Artificial Neural Network along with the clustering algorithms, The preliminary stage is to analyze the distinct classification algorithms efficiency, during the prediction process and then the classified data will be clustered to extract the required information from the trained dataset BIRCH method, for predicting the future.

Finally, the performance of the text analysis can get improved by using an affine client classification algorithm

ORIGINAL LANGUAGE: ENGLISH

JOURNAL: Advances in intelligent system and computing

VOLUME: 750

CITATIONS: 3

PUBLISHED STATUS: PUBLISHED 2018

- *LEARNING SENTIMENT -SPECIFIC WORD EMBEDDING FOR TWITTER SENTIMENT CLASSIFICATION*

The paper was published by six authors named Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin at the RESEARCH CENTER FOR SOCIAL COMPUTING AND INFORMATION RETRIEVAL, HARBIN INSTITUTE OF TECHNOLOGY, CHINA & UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CINA, HEFEI CHINA. The paper presents the methods that learn word embedding for Twitter sentiment

classification in this paper. Most existing algorithms for learning continuous word representation typically only model the syntactic context of words but ignore the sentiment of the text. This is problematic for sentiment analysis. The authors have undertaken the regiments on applying SSWE to a benchmark. Twitter sentiment classification dataset in SemEval 2013 shows that (1) the SSWE feature performs comparably with hand-crafted features in the top-performed system;(2) the performance is further improved by concatenation SSWE with an existing feature set.

ORIGINAL LANGUAGE: ENGLISH

JOURNAL: Advances in intelligent system and computing

VOLUME: 1

CITATIONS: 800

PUBLISHED STATUS: PUBLISHED 2014

REFERENCES

AUTHORS

First Author – Om Koli, B.E, AISSMS Institute Of Information Technology and omkoli3114@gmail.com

Second Author – Shubhada Londhe, B.E, AISSMS Institute Of Information Technology and londheshubhada43@gmail.com

Third Author – Prasad Kharche, B.E, AISSMS Institute Of Information Technology and prasadkharche02@gmail.com

Fourth Author – Sharad Sawant, B.E, AISSMS Institute Of Information Technology and sharadsawant433@gmail.com.