# Status updates Sorting and Analyzing Data using Machine Learning Models

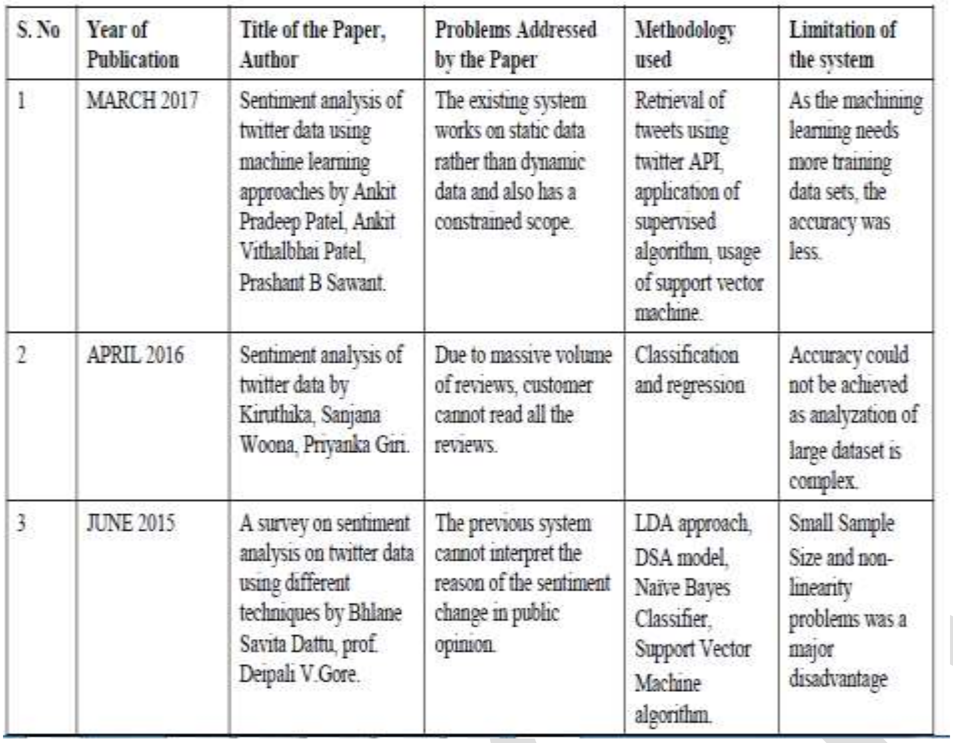
Mr.P Viswanatha Reddy , Mr.A Srinivasan , Mrs.T Sarada
Assistant Professor[1,2,3]
Department of CSE,
Viswam Engineering College (VISM) Madanapalle-517325 Chittoor District, Andhra Pradesh, India

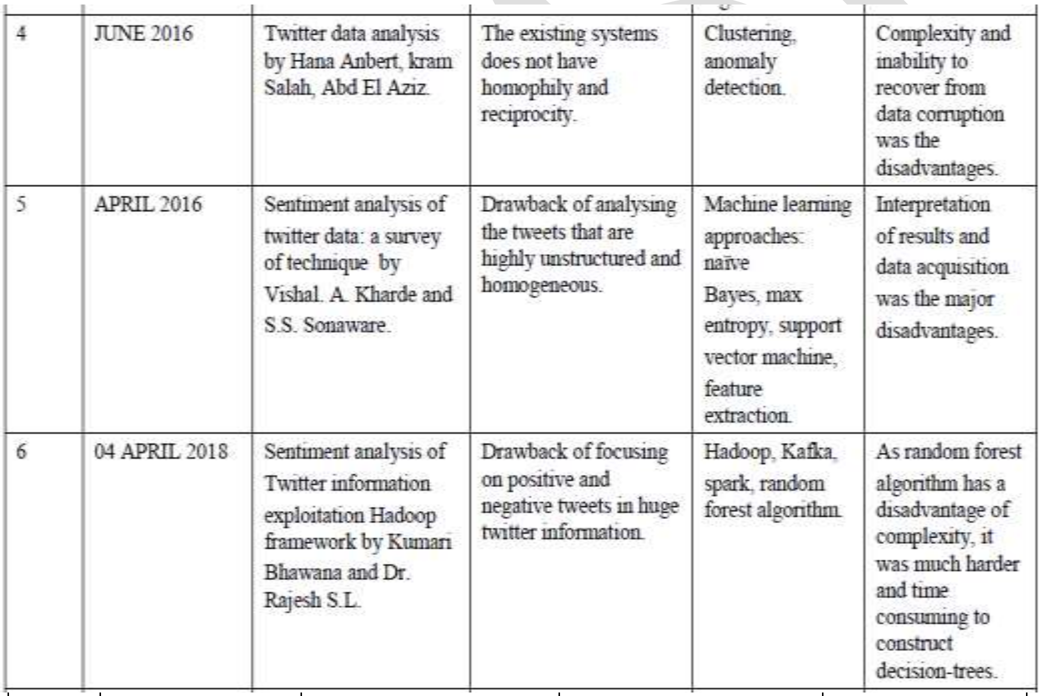
Abstract

Every day, people utilize several social media sites to disseminate information about what's hot in the world. Twitter, in particular, is used extensively by individuals all over the globe to voice and disseminate their thoughts and ideas. This has a profound effect on many issues, and individuals do it to express their agreement or disagreement. Logistic Regression, Decision Tree, and XGBOOST from Machine Learning, and TF-IDF and Bag of Words from Natural Language Processing, are the five basic methodologies utilized when thinking about Twitter sentiment analysis. First, raw data is prepared for analysis, then a word cloud is constructed, features are extracted, and finally, a comparison of multiple Machine Learning models is generated. Preprocessing, visualization, and feature extraction are some of the most important aspects of machine learning. Preprocessing, visualization, and feature extraction are some of the most important aspects of machine learning.

## Introduction

Opinion mining, or sentiment analysis, is a technique for systematically quantifying sentiments via the use of methods from natural language processing, computational linguistics, and biometrics. Information about people's thoughts, both good and negative as well as neutral, is abundant on social media. In order to foretell the reviews of any product, service, or person, sentiment analysis is a beneficial tool. As well as having applications in information retrieval and text mining, the Term- Frequency-Inverse-Documentation-Frequency technique may be used to analyze a term's occurrence frequency by assigning relative weights to the words that include it. When it comes to classic algorithms for processing natural language and evaluating word frequency, the Bag-Of-Words is a staple. When making predictions from unstructured data, the XGBOOST might help. It is an ensemble machine learning technique based on decision trees. It compiles a dictionary of all the distinct terms. To determine the likelihood that a particular data point belongs to category "1," statisticians use a regression model known as logical regression. In order to determine what category or value each input data point corresponds to, a training model is developed using a decision-tree method. When analyzing Twitter handles for sentiment, the aforementioned algorithms are put to use.





| S. No | Year of Publication | Title of the Paper, Author | Problems Addressed by the Paper | Methodology used | Limitation of the system |
|---|---|---|---|---|---|
| 1 | MARCH 2017 | Sentiment analysis of twitter data using machine learning approaches by Ankit Pradeep Patel, Ankit Vithalbhai Patel, Prashant B Sawant. | The existing system works on static data rather than dynamic data and also has a constrained scope. | Retrieval of tweets using twitter API, application of supervised algorithm, usage of support vector machine. | As the machining learning needs more training data sets, the accuracy was less. |
| 2 | APRIL 2016 | Sentiment analysis of twitter data by Kiruthika, Sanjana Woona, Priyanka Giri. | Due to massive volume of reviews, customer cannot read all the reviews. | Classification and regression | Accuracy could not be achieved as analyzation of large dataset is complex. |
| 3 | JUNE 2015 | A survey on sentiment analysis on twitter data using different techniques by Bhlane Savita Dattu, prof. Deipali V.Gore. | The previous system cannot interpret the reason of the sentiment change in public opinion. | LDA approach, DSA model, Naïve Bayes Classifier, Support Vector Machine algorithm. | Small Sample Size and non-linearity problems was a major disadvantage |
| 4 | JUNE 2016 | Twitter data analysis by Hana Anbert, kram Salah, Abd El Aziz. | The existing systems does not have homophily and reciprocity. | Clustering, anomaly detection. | Complexity and inability to recover from data corruption was the disadvantages. |
| 5 | APRIL 2016 | Sentiment analysis of twitter data: a survey of technique by Vishal. A. Kharde and S.S. Sonaware. | Drawback of analysing the tweets that are highly unstructured and homogeneous. | Machine learning approaches: naïve Bayes, max entropy, support vector machine, feature extraction. | Interpretation of results and data acquisition was the major disadvantages. |
| 6 | 04 APRIL 2018 | Sentiment analysis of Twitter information exploitation Hadoop framework by Kumari Bhawana and Dr. Rajesh S.L. | Drawback of focusing on positive and negative tweets in huge twitter information. | Hadoop, Kafka, spark, random forest algorithm. | As random forest algorithm has a disadvantage of complexity, it was much harder and time consuming to construct decision-trees. |
| 7 | 03 MARCH 2017 | Survey on sentiment analysis for twitter by Ankita Gupta and Jyothika Pruthi. | Review on various tools and techniques that has been used in existing literature for sentiment analysis of tweets. | Semantic orientation, Naïve Bayesian, support vector machine | Large training datasets was required to attain accuracy. |
| 8 | JANUARY 2019 | Systematic literature review of sentiment analysis using soft | Increase in the feasibility, scope than Existing systems. | Evaluating the use of soft computing | Lower speed, longer run time and lack of real |

| | | | | | |
|---|---|---|---|---|---|
| | | computing techniques by Akshi Kumar and Arunima Jaiswal. | | techniques such as Fuzzy logic and Bayesian statistics. | time response was the major disadvantages. |
| 9 | MARCH 2020 | A review on sentiment analysis techniques and applications by Mold Ridzwan Yaakub, Muhammad Iqbal Abu Latiffi and Liyana Safra Zaabar. | Drawbacks in analyzation of large amount of reviews. | Natural language Processing techniques: support vector machine, max entropy, Bayesian networks. | The accuracy was not achieved as max entropy and Bayesian networks does not give accurate result. |
| 10 | JUNE 2019 | Systematic literature review on context based on sentiment analysis in social ultimedia by Akshi Kumar and Geetanjali Garg. | Intend to explore and analyse the existing work on content- based sentiment analysis and to report gaps. | Fuzzy logic and Bayesian statistics. | Restricted number of usage of inputs variables was a disadvantage. |

3.Analysis of the System's Architecture

A collection of concepts, elements, and components is represented graphically in an architectural diagram. The system architecture diagram depicts the system as a whole and details how data and processing are distributed across the many components.

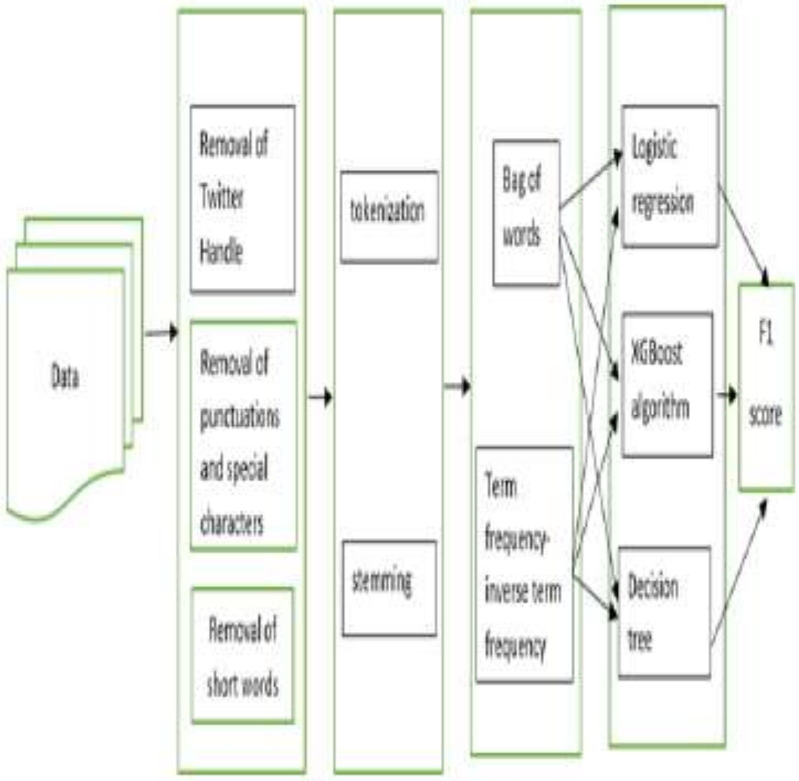


Figure 1: Architecture Diagram of Sentiment Analysis using Twitter.

Figure 1 is a schematic depicting the infrastructure used to do Twitter sentiment analysis, including the steps of gathering data, extracting features, and classifying them using a machine learning model.

Four. Acquiring the Required Information

Data collection refers to the process of amassing information for use in training and testing ML models. Information gathered here consists of both favorable and negative tweets from Twitter. Analytics Vidhya is the source for this data collection.

ID: Unique identifier for each tweet in the specified dataset.

Tweets: aggregated tweets from different sources, with good or bad connotations.

A tweet with a label of "0" is considered to have a positive sentiment, whereas a tweet with a label of "1" is considered to have a negative emotion.

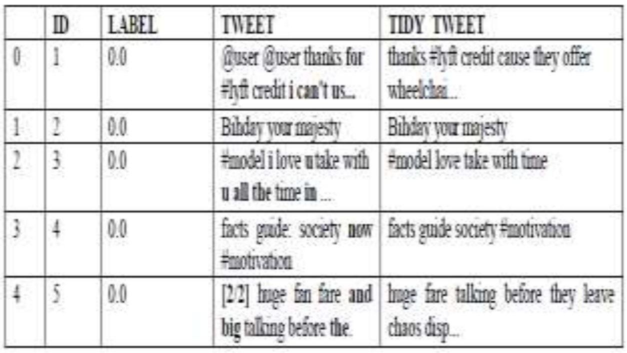
The Fifth Stage: Pre-Processing the Data

The first step in every data mining project is known as "data pre-processing," and it entails translating raw data into a usable format. Incomplete, inconsistent, or deficient in specific behaviors or patterns, and possibly including numerous mistakes, real-world data is a common problem. Pre-processing of data has been shown to be effective in addressing these kinds of problems. Through a procedure known as "pre-processing," raw data is made ready for analysis. When doing sentiment analysis, it is common practice to cleanse the data set of any extraneous information.

Take Out A. All Short Words, Special Characters, and Usernames

In the first stage of this module, the user handles (i.e., @user) are discarded or removed. The first phase involves excluding non-alphanumeric characters such as punctuation, numerals, and short sentences, with the exception of hashtags.

Table I: Cleaned Tweets after the Removal of Usernames and Short Words.



| | ID | LABEL | TWEET | TIDY TWEET |
|---|---|---|---|---|
| 0 | 1 | 0.0 | @user @user thanks for #lyft credit i can't us... | thanks #lyft credit cause they offer wheelchai... |
| 1 | 2 | 0.0 | Bihday your majesty | Bihday your majesty |
| 2 | 3 | 0.0 | #model i love u take with u all the time in ... | #model love take with time |
| 3 | 4 | 0.0 | facts guide: society now #motivation | facts guide society #motivation |
| 4 | 5 | 0.0 | [2/2] huge fan fare and big talking before the. | huge fare talking before they leave chaos disp... |

Tweets without unnecessary short words or symbols (other than hashtags and usernames) have been collected in Table I.

Tokenization, Part B

Tokenization refers to the process of reducing a set of terms to their component words. In this step, the whole dataset of cleansed tweets is tokenized.

0 [appreciate it, credit, for they give [birthday, your, majesty], 1.There are two [#model love take with time]Three [#motive, #society, #facts]As noted up above, the tokenized, cleansed tweets are the result of the preceding step's tokenization.

C. Origin

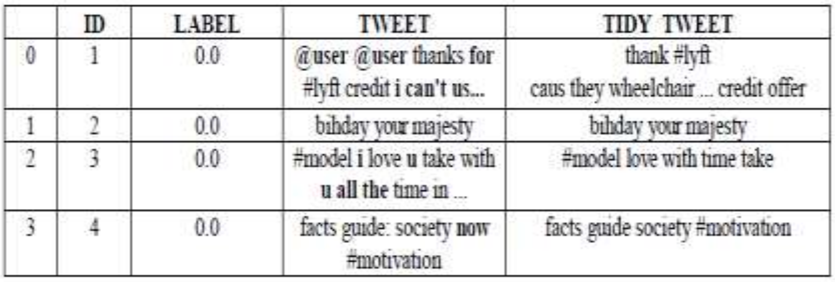
Stemming is the practice of eliminating common suffixes from words. This is done to get rid of tenses in the data.A zero [thank, #lyft credit, cause, they provide, whee...

B.1 [birthday, your, majesti]Second [#model love take with time]

Count them: 3 [factsguid, societi, #motiv]

The result of the stemming process, in which the words' suffixes are eliminated, was previously discussed.

Table II: The Processed Tweets.



|  | ID | LABEL | TWEET | TIDY TWEET |
|---|---|---|---|---|
| 0 | 1 | 0.0 | @user @user thanks for #lyft credit i can't us... | thank #lyft caus they wheelchair ... credit offer |
| 1 | 2 | 0.0 | bihday your majesty | bihday your majesty |
| 2 | 3 | 0.0 | #model i love u take with u all the time in ... | #model love with time take |
| 3 | 4 | 0.0 | facts guide: society now #motivation | facts guide society #motivation |

Table II shows the modified tweets after they've been cleaned up using tokenization, stemming, and the elimination of short words, punctuation, and other symbols.

6 - Imagination

A. Word Cloud as a Tool for Tweet Visualization

Words from the tweets, both good and negative, are considered in a dynamic word cloud visualization of all the tweets. High-frequency terms stand out more, whereas less-used ones are scaled down.

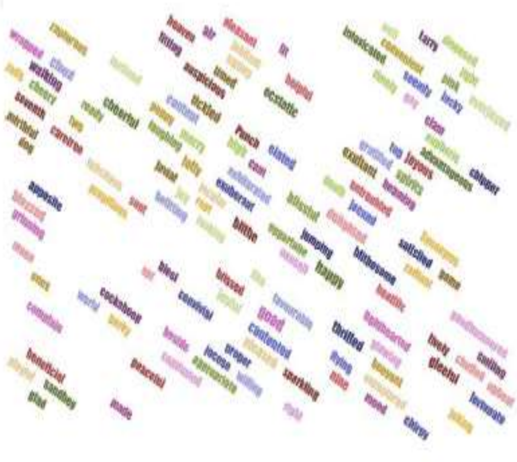


Figure 2: Words from Tweets using Word Cloud.

Tweets were used to create a word cloud, which is shown in figure 2. In this case, only the good feedback counts. Word clouds are also created for unfavorable terms.

Twitter Data Displayed as a Bar Graph

Hashtag frequency is displayed as a graph, which is another method of visualizing the content of tweets.

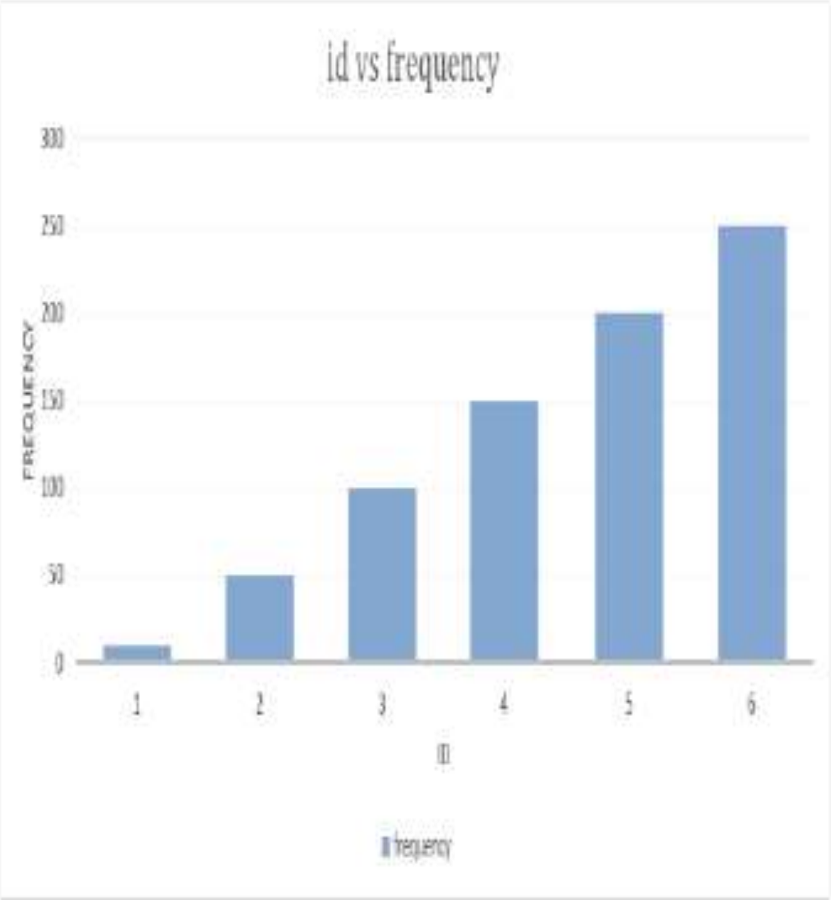


Figure 3: Visualization of Tweets by Plotting Bar Graph.

The figure 3 depicts the frequency of the words from the tweets as a graph.

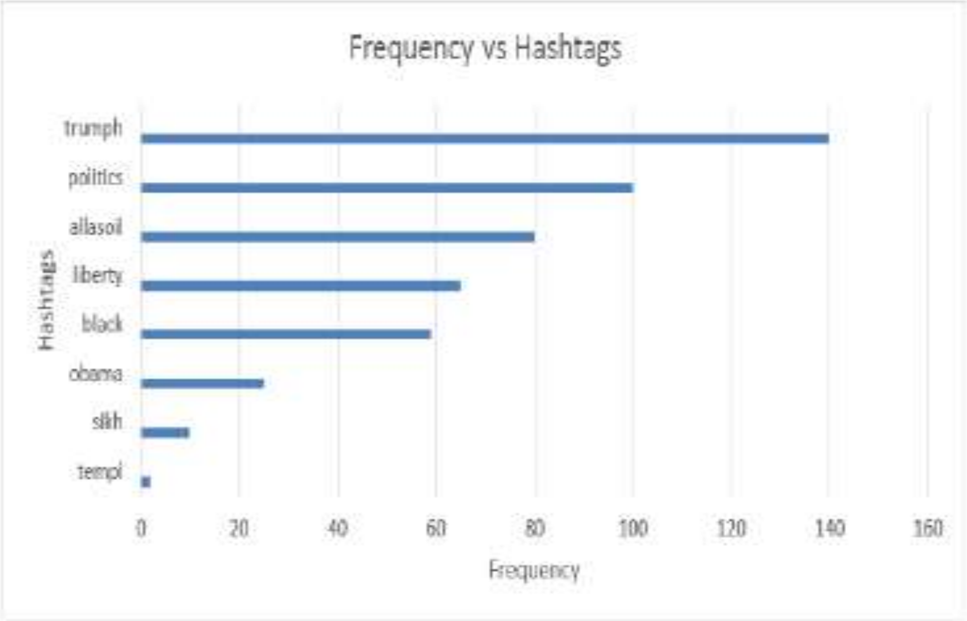


Figure 4: A Graph which shows the Frequency of the Negative Hashtags.

The figure 4 bar graph displays the distribution of hashtags containing negative terms.

Seventh, Take Out Characteristics

In order for machine learning algorithms to function, we must first transform the text into numerical vectors. Feature extraction describes this process. Using these characteristics, machine learning programs may be educated.

A. DISPOSABLE **LANGUAGE**

To achieve this, a bag-of-words model may be used. The distinct words in a tweet are represented as a bag-of-words. Bag-of-words strategy describes this method because frequency of use is more important than word order or sequence.

Example: Think about these tweets:

For one, I enjoyed the film.

And number two: I despise it.

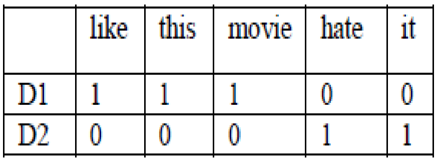
This is what comes out when the bag-of-words method is used:

It generates a matrix where each row represents a tweet and each column represents a distinct term.

Each cell may either be set to 0 or 1.

The cell will contain the value "1" if that term appears in the tweet, and "0" otherwise.

**Output**:

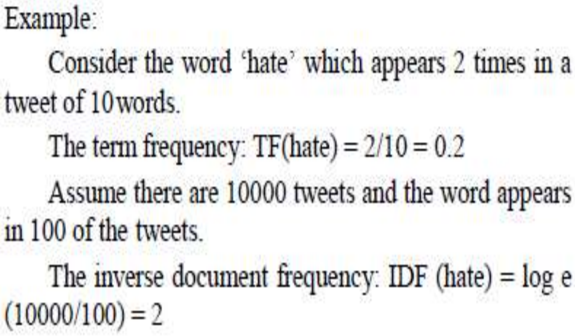


| | like | this | movie | hate | it |
|---|---|---|---|---|---|
| D1 | 1 | 1 | 1 | 0 | 0 |
| D2 | 0 | 0 | 0 | 1 | 1 |

Term frequency inverse document frequency (TF-IDF) measures the significance of a term inside a collection of documents. This technique is widely used in the fields of text mining and information retrieval. Values for TF-IDF are proportional to the frequency with which a word occurs in the document and inversely proportional to the total number of documents in the collection that include the term. This allows you to account for the fact that certain words tend to occur more often than others.

$$TF(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in document}} \quad (1)$$

$$IDF(t) = \log e \left(\frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}\right) \quad (2)$$



Example:

Consider the word 'hate' which appears 2 times in a tweet of 10words.

The term frequency: TF(hate) = 2/10 = 0.2

Assume there are 10000 tweets and the word appears in 100 of the tweets.

The inverse document frequency: IDF (hate) = log e (10000/100) = 2

Automated learning models

The approach used to resolve the issue is known as Supervised Learning. An algorithm is used to "learn" the mapping function between input variables (x) and output variables (Y) in supervised learning. The aim is to identify the mapping function such that one can reliably anticipate the values of the target variables (Y) given a set of input data (x). Y=f(X).

Steps:

Bag-of-words and TF-IDF models should be fit.

Infer the odds by considering the variables

Perform the F1 score calculation.

Finally, choose the model that performs best in a comparison of outcomes.

Regression Analysis Using Logical Data

Logistic Regression should be used when the dependant variable can only take on two possible values. The logistic regression is a kind of predictive analysis similar to other types of regression studies. A dependent binary variable and an independent variable with one or more nominal, ordinal, interval, or ratio levels may be shown by logistic regression. The procedure looks like this:

Log Reg = LogisticRegression(random state=0,solver='lbfgs') is an example of this.

The random state is utilized to divide the sample equally across the training and testing phases. Every time the code is run, it will provide a different result unless otherwise stated.

Limited-memory Broyden-Fletcher-Goldfarb-Shanno describes the solver used by this algorithm. In multi-class situations, it is often the go-to method of resolution. To save space, it just keeps the most recent few changes. When dealing with massive amounts of data, it is slow.

B. XG Boost XGBOOST is a Machine Learning method that makes use of a gradient boosting framework and is built on decision trees. Its popularity stems from the fact that it may provide high-quality outcomes with less investment of time and effort by using optimization strategies. The procedure looks like this:

Specifically: model xg = XGBClassifier(random state=22,learning rate=0.9)

The random state is utilized to divide the sample equally across the training and testing phases. Unless otherwise provided, the code will produce unique values every time it is run.

Shrinkage measures how quickly a person picks up new information. It is implemented to mitigate the model's reliance on branch. It lowers the learning rate to avoid overfitting.

**The C. Decision Tre**e

One of the most used methods for making predictions and class distinctions is the decision tree. A decision tree is a diagram that looks like a tree, but each node within the tree indicates a different test or condition that must be met before a certain outcome is determined.

property, each sidebar represents a test result, and each leaf node (terminal node) provides a predicted class label. The procedure looks like this:

It may be written as: dct = DecisionTreeClassifier(criterion='entropy', random state=1)

To decide between gini and information gain, the criteria attribute is employed.

The random state is utilized to divide the sample equally across the training and testing phases. In the absence of such a declaration, the code will produce a distinct set of values every time it is run.

Compared Models, Number 9

The comparison of the machine learning models is shown as a line graph. The tweets' predicted categories have been determined. Score (shown on the y-axis) and model (represented on the x-axis) are the two variables used to create this graph (x-axis). A scatter plot of F1 scores for ML models of NLP algorithms is shown. Rather than relying on accuracy alone, this statistic is used to rule out any false positives.

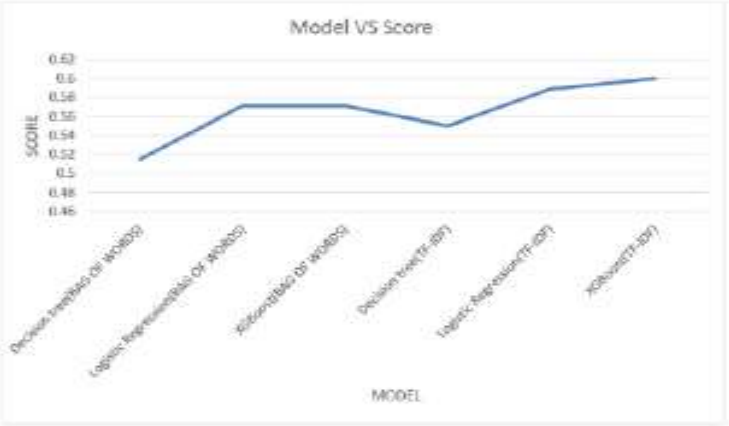


Figure 5: The Comparison of Model and Score.

Model F1 ratings for 2 distinct NLP systems are shown in Figure 5. Taking BAG-OF-WORDS into account, the F1 score is greater for logical regression and XGBOOST and much lower for decision tree. As far as TF-IDF is concerned, logical regression gets the highest F1 score, followed by XGBOOST and the decision tree.

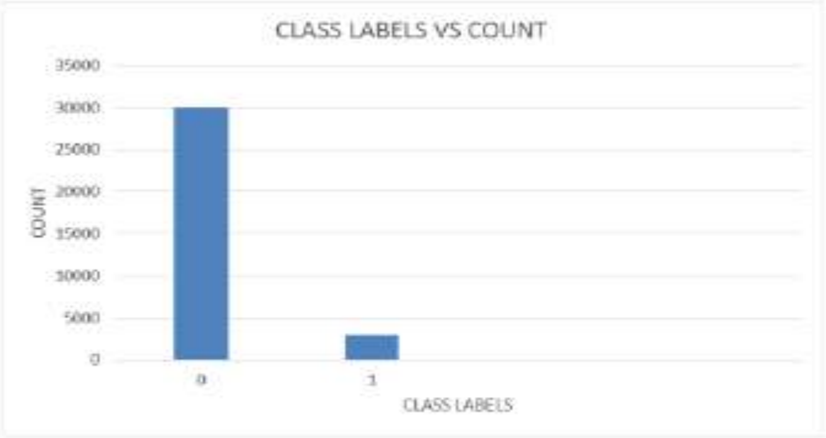


Figure 6: The Graph of Class Labels vs Count.

Figure 6 shows the class-labeling and count-showing graph. As the above graph shows, using the F1 score rather than accuracy is a sound decision. The above graph shows a lot of variation, which indicates that there are more false positives. Thus, the F1 score is implemented.

10. Concluding, the field of machine learning devoted to the classification of emotional states is an exciting and developing one. The goal of this research is to utilize the F1 score as a measure to evaluate the relative efficacy of different machine learning models that have been applied to feature extractions from tweets. Predicting someone's emotions with any degree of precision is complicated by language's inherent ambiguity. Adding other languages to the sentiment analysis will allow for even greater advancements. More nuanced sentiment analysis and analysis of non-textual data are possible future directions for study, notwithstanding the project's current focus on textual data and its categorization of tweets into positive and negative attitudes.

## References

[1] Akshi Kumar and Arunima Jaiswal (January 2019), ‘Systematic literature review of sentiment analysis using soft computing techniques’.

[2] Ankita Gupta and Jyothika Pruthi (March 2017), 'Survey on sentiment analysis for twitter'.

[3] Ankit Pradeep Patel, Ankit Vithalbhai Patel, Prashant B Sawant (March 2017), 'Sentiment analysis of twitter data using machine learning approaches'.

[4] Bhlane Savita Dattu, Prof. Deipali V.Gore (June 2015), 'A survey on sentiment analysis on twitter data using different techniques'.

[5] Hana Anbert, Akram Salah, Abd El Aziz (June 2016), 'Twitter data analysis'.

[6] Kiruthika, Sanjana Woona, Priyanka Giri (April 2016), 'Sentiment analysis of twitter data'. [7]. Kumari Bhawana and Dr. Rajesh S.L. (April 2018), 'Sentiment analysis of twitter information exploitation Hadoop framework'.

[7] Mold Ridzwan Yaakub, Muhammad Iqbal Abu Latiffi and Liyana Safra Zaabar (March 2020), 'A review on sentiment analysis techniques and applications'.

[8] Vishal. A. Kharde and S.S. Sonaware (April 2016), 'Sentiment analysis of twitter data: a survey of techniques'.