May-June 2020
ISSN: 0193-4120 Page No. 5171-5177

Tweets Categorization and Comparison of Results using Machine Learning Models

Dr Revathy P[1], Tejashree B[2], Varshini S[3], Vibhaa S[4]

[1]Associate Professor
[1,2,3,4]Computer Science and Engineering Rajalakshmi Engineering College Chennai
[1]revathy.p@rajalakshmi.edu.in, [2]tejashreebalan@gmail.com, [3]varshini.gsk69@gmail.com, [4]vibhaacv10@gmail.com

Article Info
Volume 83
Page Number: 5171-5177
Publication Issue:
May - June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *16 May 2020*

Abstract
Various social media platforms are used on the daily basis to spread trending news and topics. Among those, twitter is widely used all over the world by the people to put forward and spread about their opinions. This creates a very powerful impact on various topics and people do this to show their support or contradiction. On considering analysis of sentiments using twitter, five main approaches are used which include Logistic Regression, Decision Tree, and XGBOOST which comes under Machine Learning and TF-IDF and Bag of Words which comes under Natural Language Processing. Initially data pre-processing happens, followed by visualization of words, then comes feature extraction, and last step is generating comparison of various Machine Learning models.

Keywords**:** *machine learning, data pre-processing, visualization, feature extraction*

1. Introduction

Sentiment analysis also known as opinion mining refers to the use of natural language processing text analysis, computational linguistics and biometrics to systematically quantify the information. Social media is where the abundance of opinion information are available and there would be positive, negative and neutral opinions. The sentiment analysis is useful in predicting the reviews about anything and everything. The Term- Frequency-Inverse- Documentation-Frequency algorithm is used for analyzing how frequently a term occurs by weighing the words used and it is also useful in information retrieval and text mining. The Bag- Of-Words is a traditional algorithm used in natural language processing, and also analyzing frequency of the word. The XGBOOST is a decision tree-based ensemble machine learning algorithm used for predicting problems involving unstructured data. It creates a vocabulary of all the unique words. Logical regression is a regression model used to predict the probability of the given data entry belonging to the category "1". Decision-Tree algorithm is used to create training model that is used to predict the class or the value which the data entry belongs to. Above mentioned models are used to analyze the sentiments using Twitter Handles of various users.

2. Literature Survey

May-June 2020
ISSN: 0193-4120 Page No. 5171-5177

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

May-June 2020
ISSN: 0193-4120 Page No. 5171-5177

		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. System Architecture Diagram

An architecture diagram is a graphical representation of a set of functions or procedures which has principles, elements and components. The system architecture diagram explains the flow of functionality from one module to another and gives an overall view of the entire system.

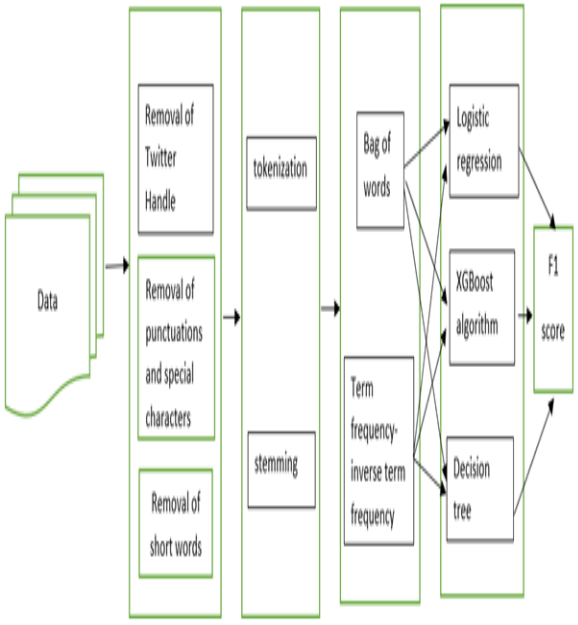

Figure 1: Architecture Diagram of Sentiment Analysis using Twitter.

The Figure 1 above is the architectural diagram graphically representing the actions like data collection, feature extraction and machine learning model classification for the sentiment analysis using twitter.

4. Data Collection

The process of collection of data for the purpose of training and testing the machine learning models is called data collection. The data collected here is the positive and negative tweets from twitter. The data set is collected from Analytics Vidhya website.

ID: The id associated with the tweets in the given dataset.

Tweets: The tweets collected from various sources and having either positive or negative sentiments associated with it

Label: A tweet with label '0' is of positive sentiment while a tweet with label '1' is of negative sentiment

5. Data Pre-Processing

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing. In sentiment analysis, the collected data is processed in order to remove unnecessary data which are not required for further processes.

A. Remove Usernames, Special Characters and Short Words

This is the module where the initial step includes removing or discarding the user handles (i.e., @user). Followed by the first step, removal of special characters except hashtags which also include punctuations, numbers, and short words.

May-June 2020
ISSN: 0193-4120 Page No. 5171-5177

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...

Table I contains the cleaned tweets after the removal of short words, special characters except hashtags and usernames.

B. Tokenization

Tokens are the terms or words, and the process of separating the terms to tokens is called as tokenization. Here, all the cleaned tweets in the dataset are tokenized.

0 [thanks, #lyft, credit, cause, they, offer, wh...
1 [bihday, your, majesty]
2 [#model, love, take, with, time]
3 [factsguide, society, #motivation]

The output of the step tokenization is mentioned above where the cleaned tweets are tokenized.

C. Stemming

Stemming is the process of removing the suffixes of words like ly, ing and so on. This process is basically done in order to remove tenses in the data.

0 [thank, #lyft, credit, caus, they, offer, whee...
1 [bihday, your, majesti]
2 [#model, love, take, with, time]
3 [factsguid, societi, #motiv]

The output of the step stemming is mentioned above where the suffixes of the words are removed.

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

The table II contains the processed tweets after performing removal of short words, special characters, punctuations, tokenization and stemming.

6. Visualization

A. Visualization of Tweets using Word Cloud

Visualization of all the tweets dynamically using word cloud is done where words from the tweets which include both positive and negative words are taken into consideration. The words those frequency is high appear large and words with lower frequencies appear small.

Figure 2: Words from Tweets using Word Cloud.

The figure 2 depicts words generated from tweets using word cloud. Here, the positive words are taken into consideration. Similarly, word cloud will be generated for negative words.

May-June 2020
ISSN: 0193-4120 Page No. 5171-5177

B. Visualization of Tweets using Bar Graph

This is another visualization technique where the frequency of the words (i.e., hashtags) in the tweets is plotted as graph.

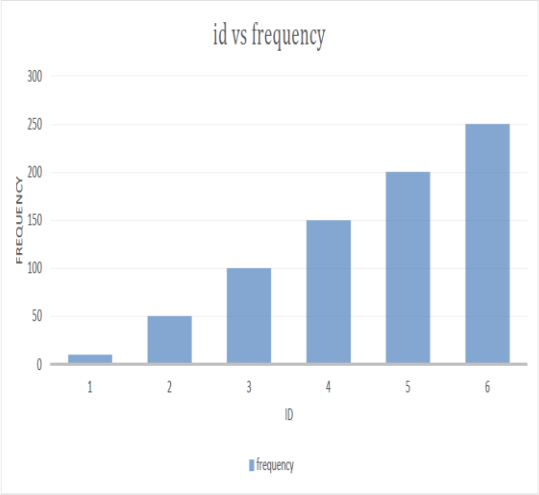

Figure 3: Visualization of Tweets by Plotting Bar Graph.

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

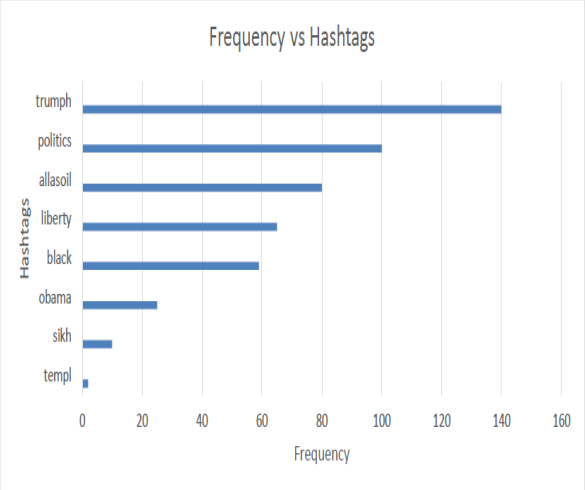

Figure 4: A Graph which shows the Frequency of the Negative Hashtags.

The figure 4 depicts the bar graph plotted which shows the frequency of negative words in the hashtags.

7. Extracting Features

Machine learning algorithms cannot work with raw text directly; we need to convert the text into vectors of numbers. This is called **feature extraction**. These features can be used for training machine learning algorithms.

A. BAG-OF-WORDS

A bag-of-words model is a way of extracting features from text document. A bag-of-words represents all unique words in a particular tweet. This is called bag-of-words approach since the number of occurrence of each word matters rather than sequence or order of words.
Example: Consider the following tweets:

1. I like this movie.
2. I hate it.

The output after applying the bag-of-words is as follows:

It creates a matrix where the attributes or columns represent the different words and each row or tuple represent a tweet.

Each cell has a value of '0' or '1'.

If the particular word is present in the tweet the cell will have the value of '1' else it will have '0'.

Output:

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

B. TF-IDF

Term Frequency–Inverse Document Frequency, reflects how important a word is in a document among a set of documents. Text mining and information retrieval uses this method largely. The TF-IDF value is proportional to the number of times a word appears in the document and is inversely proportional to the number of documents in the set of documents that contain the word. It helps to adjust for the fact that some words appear more frequently in general.

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

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

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

May-June 2020
ISSN: 0193-4120 Page No. 5171-5177

8. Machine Learning Models

The problem that is going to be solved comes under Supervised Learning. In Supervised learning one has input variables (x) and an output variable (Y) and the mapping function from the input to the output is learnt through an algorithm. The goal is to find the mapping function so well that when you have new **input data (x)** that you can predict the **output variables (Y)** for that data. Y=f(X).

Steps:

- Fit the various models on Bag-of-words and TF-IDF.
- Predict the probabilities
- Calculate the F1 score.
- Finally compare the results of various models and use the best model to predict the result.

A. Logistic Regression

When the dependent variable is dichotomous (binary) use Logistic Regression. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to show the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variable. It is implemented as follows:

Log_Reg = LogisticRegression(random_state=0,solver='lbfgs').

The random state is basically used to split the training and test sample in the same way. If it is not specified, each time the code is executed, different values are generated as output.

The lbfgs solver stands for Limited-memory Broyden–Fletcher–Goldfarb–Shanno. It is used in multi-class problems and is used as the default solver. It saves only last few updates so as to save memory. It does not work fast with large data sets.

B. XG Boost

XGBOOST is a decision-tree-based Machine Learning algorithm that uses a gradient boosting framework. It is used widely because it uses optimization techniques to yield superior results in the shortest amount of time using less resources. It is implemented as follows:

model_xg = XGBClassifier(random_state=22,learning_rate=0.9)

The random state is basically used to split the training and test sample in the same way. If it is not specified, each time the code is executed, different values are generated as output.

The learning rate is known as shrinkage. It is used to lessen the effect of branch in the model. It is used to prevent overfitting by reducing the value of learning rate.

C. Decision Tree

Decision tree is the most used tool for classification and prediction. A Decision tree is a tree like structure, where each internal node represents a test or a condition on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) depicts a class label after prediction. It is implemented as follows:

dct = DecisionTreeClassifier(criterion='entropy', random_state=1)

The criterion attribute is used to choose between gini or information gain.

The random state is basically used to split the training and test sample in the same way. If it is not specified, each time the code is executed ,different values are generated as output.

9. Model Comparison

The final output is a line graph where the machine learning models are compared. Prediction of classes for the tweets is done. The graph consists of two parameters viz., score (y-axis) and model (x-axis). The F1 score for the various machine learning models on Natural Language Processing algorithms are plotted as a graph. This metric is used instead of accuracy to remove cases of false positives.

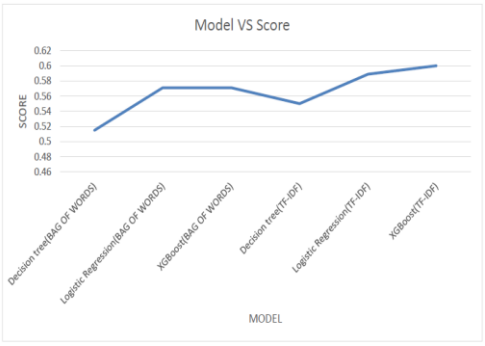

Figure 5: The Comparison of Model and Score.

The figure 5 represents the F1 scores of various models under 2 different NLP algorithms. On considering BAG-OF-WORDS, logical regression and XGBOOST have higher F1 score where decision tree has very low F1 score. While TF-IDF is concerned, logical regression has a good F1 score followed by XGBOOST and decision tree.

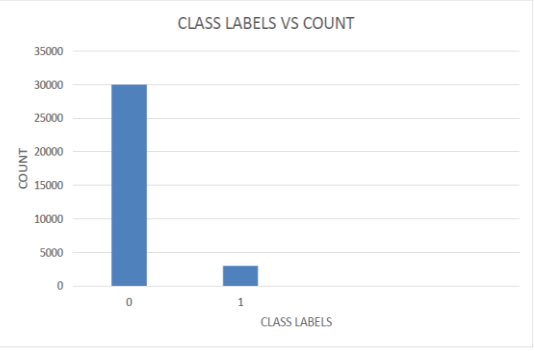

Figure 6: The Graph of Class Labels vs Count.

May-June 2020
ISSN: 0193-4120 Page No. 5171-5177

The figure 6 represents the graph which labels class labels and count. The above graph validates why F1 score is used instead of accuracy. A large fluctuation is seen in the above graph which shows there are more false positive values. This is why F1 score is used.

10. Conclusion

Categorization of sentiments is an important and promising domain in machine learning. The project focuses on comparing the various machine learning models applied on features extracted from tweets in order to find their efficiencies using the F1 score metric. The complexity of language makes it difficult to accurately predict the sentiments. Further improvements can be made by extending the sentiment analysis to various languages. The project works on textual data and categorizes the tweets into positive and negative sentiments but more fine-grain sentiment analysis and analysis of non-textual data can be a potential research path.

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'.