Sentiment Analysis of Twitter Data Using Machine Learning Approach

|  |  |
| --- | --- |
| *Dr. Adnan Amin,*  *Data Scientist Machine Learning Practitioner*  *Institute of Management Sciences*  *Hayatabad Peshawar, Pakistan*  [*geoamins@yahoo.com*](mailto:geoamins@yahoo.com) | *Ghulam Mujtaba Adil*  *Undergraduate student BS Data science 6th-semester*  *Institute of Management Sciences*  *Hayatabad Peshawar Pakistan*  [*ghulam.mujtabadil001@gmail.com*](mailto:ghulam.mujtabadil001@gmail.com) |
| *Mohammad Shamsher,*  *Undergraduate student BS Data science 4th semester*  *Institute of Management Sciences Hayatabad Peshawar, Pakistan* [*mohammad.shamsher682@gmail.com*](mailto:mohammad.shamsher682@gmail.com) | *Irfan Shah*  *Undergraduate student BS Data science 4th-semester*  *Institute of Management Sciences*  *Hayatabad Peshawar, Pakistan* [*Irfanshah4156@gmail.com*](mailto:Irfanshah4156@gmail.com) |

***Abstract—***With the emergence of Web 3.0 and the development of Social Media platforms people are attracted and inclined to post their views and opinions on social media freely, Twitter is one of the hot social media platforms on which users can post their opinions. Twitter Sentiment Analysis is difficult in comparison to other general sentiment Analyses because of the presence of slang words, misspellings, etc. which make processing tough. Analyzing such a large number of unstructured Tweets and comments is a challenge for the world of Natural Language Processing. This Paper introduces Sentiment Analysis techniques, the methodology used and the comparison of the results obtained from different machine learning technologies, including the Naive Bayes Method, Gradient Boosting Classifier, and Support Vector Machine. Finally, the Model Classifier performance was evaluated in terms of accuracy, recall, precision, and F1 score. In term of these parameters the model classifier Support Vector Machine(SVM) gives the best performance.

***Terms—Sentiment Analysis, Machine Learning, Balanced data, Twitter, Opinion Mining, Classification.***

1. **INTRODUCTION**

Twitter is a well-known microblogging administration in which users post status messages, called” tweets”, with not more than 140 characters. In most cases, users post their tweets with much lesser words than the limit described. Twitter addresses one of the biggest and most powerful data sets of users created where around 500 million tweets are done daily and this number is rapidly increasing day by day. Tweets can offer opinions on various topics, which can assist with coordinating marketing campaigns in order to impart to consumer’s insights concerning brands and items, occasions that produce insecurity, polarity forecast in political and sports conversations, and acknowledgment or dismissal of politicians, all in an electronic information exchange way. In such application areas, one deals with huge text corpora and most frequently” Formal language”. No less than two explicit issues should be addressed in any type of computer-based tweet analysis: right off the bat, the recurrence of incorrect spellings and slang words in tweets is a lot higher than that in different domains. Besides, Twitter clients post tweets on a variety of subjects, in contrast to blogs, news, and other sites, which are tailored to specific topics. Enormous challenges can be looked at in Twitter sentiment analysis neutral tweets are much more common than positive and negative ones. This is not quite the same as other sentiment analysis areas like product reviews which will generally be overwhelmingly positive or negative, there are linguistic representational challenges, similar to those that emerge from feature engineering issues and tweets are extremely short and frequently show limited opinion cues. The machine learning strategy in sentiment classification relies upon the famous data processing machine learning techniques. Machine learning strategies are well known these days, and the accuracy is somewhat high. Due to the variety of semantic articulation, dictionary matching has an enormous blunder, yet the machine learning techniques won’t be such an issue. The scenes of machine learning techniques are more different, they can complete both positive and negative or neutral emotion classification and don’t have to go into the words, sentences, or grammar level, as well as the word reference, and coordination. This text will carry on the machine learning technique for the feeling analysis, and introduce the concrete contents and the assessment strategies of various machine learning methods. In this paper, we used a data set of 1000 tweets extracted from Twitter using Twitter API on the current ongoing political topic “Imported Government not accepted” in Pakistan.

The paper distribution is as follows, Section II emphasizes Literature Review, Section III emphasizes Background Study, Section IV emphasizes Empirical Setup, Section V emphasizes Results and Discussions, and finally Section VI concludes the study.

1. **LITERATURE REVIEW**

Before you begin to format your paper, first write and save a lot of study has been done in the topic of” Sentiment analysis” in recent years by a number of researchers. Work in the field actually began at the turn of the century. It was originally designed for binary categorization, which assigns bipolar classes such as good or negative to comments or reviews.

Paper [1] adds to the sentiment analysis for customer review classification, which is useful for analyzing data in the form of the number of tweets where opinions are extremely unstructured, either and are either positive or negative or somewhere in between. They did this by first pre-processing the dataset, then extracting the adjective from the dataset that has some meaning (called feature vector), selecting the feature vector list, and applying machine learning-based classification algorithms such as Naive Bayes, Maximum Entropy, and SVM, as well as the Semantic Orientation-based WordNet, which extracts synonyms and similarity for the content feature. They evaluated the classifier’s recall, precision, and accuracy.

In paper [2] they wanted to evaluate a few articles on Twitter sentiment analysis research, outlining the methodology and models used, as well as describing a generalized Python-based approach. They wanted to look at some studies in this area and studied it. How to use Python to perform sentiment analysis on Twitter data The scope of this work is limited to machine learning models, and was demonstrate how the efficiency of different models compares to one another. Following the preprocessing and feature extraction procedures, they concentrate on training and assessing the model’s performance as they go. The data were separated into two groups: training and testing. The training set is used to train the classifier (machine-learned model), whereas the testing set is employed to conduct the experiment. Social network data is useful in a variety of domains, including prediction, evaluating the general public’s sentiment on a particular social problem, and so on.

The paper [3] entails analyzing the public mood in response to a specific piece of news via Twitter tweets. The main goal of the article was to improve accuracy of classification by incorporating Natural Language Processing (NLP) techniques, particularly semantics and Word Sense Disambiguation. To examine the sentiment, the mined textual information is subjected to Ensemble classification. On a given classification problem, ensemble classification entails integrating the effects of multiple separate classifiers. Experiments have shown that ensemble classifiers surpass typical machine learning classifications by 3 to 5.

Similarly, in [14], Tokenization is used to convert the input string to a word vector, stemming is used to get the root of the words, feature selection is used to extract the essential terms, and classification is used to categorize the reviews as positive or negative. A model is presented that incorporates all of the methods stated earlier. On eight distinct classifiers, the model is assessed and compared. A real-world dataset is used to test the model. Five distinct assessment metrics are used to compare the eight different classifiers. Random Forest outperforms the other classifiers, according to the results. Furthermore, based to the results obtained from the evaluation metrics, Ripper Rule Learning fared the worst on the dataset.

In the recent work in [13], They used data from the microblogging website Twitter to gather information about farmers' protests in order to better comprehend public sentiment on a global scale. Based on a collection of about 20,000 tweets about the protest, they built models to categories and analyses the sentiments. They compared the performance of Bag of Words and TF-IDF and found that Bag of Words outperformed TF-IDF. They also used Naive Bayes, Decision Trees, Random Forests, and Support Vector Machines, and found that the Random Forest had the best classification accuracy.

While comprehensive reading of the mentioned paper and their methodology, it was found that they were training their training their classifiers on an imbalanced data set. So, we proposed smote technique to make the data set balanced, so that, accuracy, precision, recall, and F1 score of the target variables improves.

1. **Background Study**

In our approach, we extracted tweets from Twitter about the hottest political trend going on in Pakistan Imported Government Not Accepted. We used different Prepossessing techniques to clean and transform the raw tweets and make them appropriate for Analysis. Further, the Polarity and subjectivity of the Tweets Were Analyzed and on the basis of Polarity the tweets were classified to be Neutral, Positive, or Negative and different Supervised Machine Learning Algorithms were used to train the transformed data set.

A supervised machine learning approach is used for classification problems. And in our work, we used different Supervised Machine Learning Algorithms to obtain the desired results for our work. In the next few paragraphs, we have briefly discussed the three different Algorithms we used for our Classification Problem i.e., Naïve Bayes, Support Vector Classifier, and the ensemble approach Gradient Boosting Classifier.

* **Naive Bayes**

It has been used due to its effortlessness in both the preparing and Classifying stage. It is a probabilistic classifier and can learn the pattern with the example of analyzing a bunch of records that have been categorized. It compares the contents with the list of words to classify the reports to their right class.

***Here,***

***P(c/x) shows the posterior probability***

***P(x/c) shows the likelihood***

***P(c) represents class prior probability***

***P(x) represents the predictor prior probability***

Class c\* is assigned to tweet d, where, f represents a feature and ni(d) represents the count of feature fi found in tweet d. There is a total of m features. Parameters P(c) and P (f/c) are obtained through maximum likelihood estimates which are incremented by one for smoothing. Pre-processed data along with extracted features is provided as input for training the classifier using naïve Bayes. Once the training is complete, a classification provides the polarity of the sentiments. For example, the review comment “I am happy’ provides Positive polarity as result.

Class c\* is appointed to tweet d, where, f addresses an element and ni(d) address the inclusion of component fi found in tweet d. There is a sum of m features. Boundaries P(c) and P (f/c) are gotten through the most extreme probability estimates which are increased by one for smoothing. Pre-processed balanced data alongside extracted features are given as a contribution to preparing the classifier utilizing Naive Bayes.

* **Support vector machine**

Generally, a Support Vector Machine (SVM) is a two-class arrangement model. The meaning of the essential model is a linear classifier with an interval maximum in feature space. Its learning procedure is the most extreme distance, which can be changed into settling arched quadratic programming. Text data is ideally appropriate for SVM classification on the grounds that of the meager idea of the text, wherein few highlights are unessential, however they will more often than not be associated with each other also, for the most part coordinated into directly divisible classes. The element space is huge for any text characterization task in this way, a straight bit is normally utilized. The classification mathematical function is shown in the following:

Where ai can be obtained by the Support Vector Machine formula, K(Xi, X) is the kernel function. When the value of the ai is not 0, the sample corresponds to the support vector. The ideal discriminant mathematical function.

where

b0 = 1 – W0TXs

* **Gradient Boosting Classifier**

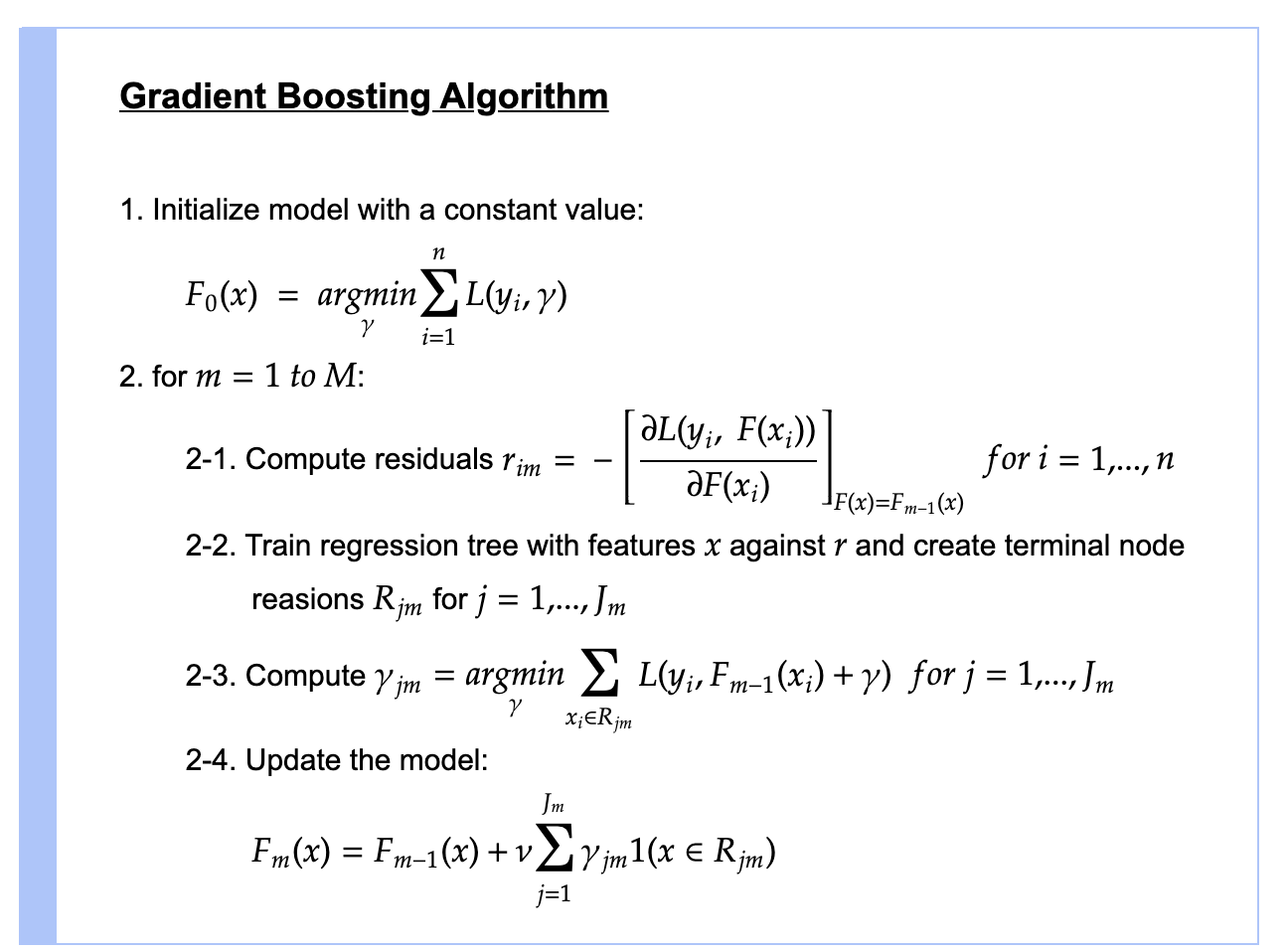
The Gradient Boosting Algorithm starts via preparing a decision tree in which every observation is assigned an equivalent weight. In the wake of assessing the first tree, we increment the loads of those observations that are hard to classify and bring down the loads for those that are not difficult to classify. The subsequent tree is accordingly developed on this weighted information. Here, the thought is to refine the predictions of the first tree. Our new model is in this manner Tree 1 + Tree 2. We then figure the classification error from this new 2-tree troupe model and grow a third tree to predict the reconsidered residuals. We rehash this cycle for a predefined amount of emphasis. Ensuing trees assist us with classifying observations that are not very much grouped by the past trees. Predictions of the last troupe model is along these lines the weighted amount of the predictions made by the past three models.

***Gradient Boost Regression Details***

***Gradient Booster Classification Detail***

***-1***

Gradient Boosting trains many models in a continuous, added substance and successive way. The Gradient Boosting Algorithm model distinguishes the deficiencies by utilizing high weight pieces of information and involving gradients in the misfortune work

 , e needs a unique notice as it is the error term). The misfortune work is an action showing how great are model’s coefficients are at fitting the hidden information. A legitimate comprehension of misfortune capacity would rely upon what we are attempting to advance. For instance, on the off chance that we are attempting to predict the deals costs by utilizing a regression, then the misfortune capacity would be based off the error among valid and predicted house costs. Likewise, on the off chance that we want to classify credit defaults, the misfortune capacity would be a proportion of how great our predictive model is at classifying awful advances. Perhaps the greatest inspiration of utilizing gradient boosting is that it permits one to upgrade a client determined cost work, rather than a misfortune work that generally offers less control and doesn’t basically compare with true applications.

1. **Empirical Setup**

The Twitter data set was used and examined. Using the trigram feature extraction technique, this study tagged data sets. We employed a system in which a preprocessor is applied to raw sentences to make them more understandable. Furthermore, different machine learning approaches use feature vectors to train the data set, and semantic analysis uses a huge number of synonyms and similarities to determine the content’s polarity. In the next subsections, a detailed description of the technique is given, as well as a graphic representation of the approach’s block diagram.

World to vectors

Data Balancing

Analysis

Training The Classifier

Feature Extraction

Subjectivity and Polarity

Preprocessing

Data Set

Figure 1 Block Diagram

Model Classification

1. **Collection of Tweets**

Collecting relevant tweets regarding a specific topic is known as tweet collection. The tweets are collected for the specified time period of research using Twitter’s streaming API. The format of the obtained text is changed to suit your needs. The dataset gathered is critical to the model’s efficiency. The model’s efficiency is also determined by how the dataset is divided into training and testing sets. The training set is the most important factor in determining the outcome.

1. **Pre-Processing The Data Set**

The data preparation phase is crucial since it determines the efficiency of the subsequent steps. It entails the desired syntactical adjustment of tweets. To eliminate ambiguity in feature extraction, the methods involved should strive to make the data more machine readable. The processes for preprocessing tweets are listed below.

• **Removing of the Re-Tweets**

• **Converting to Single Format (all upper to all lower)**

If we use case sensitive analysis, we can consider two occurrences of the same word to be different because of their sentence case. It is critical for a good analysis to avoid providing the model such reservations.

• **Removing The Stops Words Stop words** (such as and, or, yet, etc.) that do not change the text of the tweet are eliminated. For this, we employ the TFIDF machine learning software, which checks each word in the text against a dictionary.

• **Removing Images and icons**

• **Removing features With No Importance** User names and URLs are irrelevant in terms of future processing; thus, their presence is pointless. All usernames and URLs are erased or transformed to generic tags.

• **lemmatization** Terms are replaced with their origins, minimizing the number of words with comparable meanings. This helps to reduce the feature set’s dimensionality. The words are converted to their original root words so that the model can process the features in an efficient manner and enhance the subjectivity and polarity.

• **Removing Special Characters and Digits Digits** and special characters have no emotional value. They are sometimes confused with nouns, therefore removing them can assist in associating two terms that were previously regarded distinct.

• **Remove undesired words and punctuation marks from the text.**

1. **Feature Extraction**

A feature is a piece of data that can be used as a characteristic to help in issue solving. The quantity and quality of characteristics are critical for the outcomes given by the chosen model.

*• Uni-gram features One word at a time is analyzed to see if it has the potential to be a feature.*

*• Bi-gram features Two words at a time is analyzed to see if it has the potential to be a feature.*

*• N-gram features at any given time, more than one term is being considered.*

Frequency analysis is a technique for gathering features with the highest frequencies. They also deleted some of them due to the prevalence of terms with comparable sentiments (for example, happy, joy, euphoric, and so on) and grouped them together. This is accompanied by an affinity analysis that focuses on higher order n-grams in tweet feature representation. To find the weight of a certain feature in a text and hence filter it, we used unigrams and bigrams and applied Term Frequency Inverse Document

Frequency (TF-IDF). qualities with the most weight the TF-IDF is a highly useful tool. It’s a time-saving method that’s extensively utilized in text classification, and data analysis.

1. **Encoding The Analysis**

As we have been analyzed according to the subjectivity and polarity and predicted the analysis of the data set as a positive, negative, and neutral. Hence to make the predicted analysis understandable by machine we encoded as shown in Table 1.

Table 1 Encoded Analysis

|  |  |
| --- | --- |
| Positive | 1 |
| negative | -1 |
| Neutral | 0 |

1. **Smote method to Balance data**

As the extracted tweets data set contains an UN-balanced features volume which would be creating the problem of OVER FITTING OR UNDER FITTING, and hence our model would not be efficient enough to predict the data set effectively and proficient. In order to regularize it we balanced the data set in to three equal parts as; Positive, Negative, and Neutral. Such as when we will be having data set with 1000 tuples, we can divide the data set by 3 hence we can get three equal parts, and also the model would predict the data set efficiently.

1. **Splitting Data into Train and Test**

We have been using the train-split method to split the data set in to training and testing phase. The data set is split of about 20.

1. **Constant Features Removal**
2. **Removing Duplicate Features**
3. **Results and Discussions**

In this section we discuss the results obtained through Naïve Bayes, Support Vector Machine, and Gradient Boosting Classifier and compared their relative performances on three parameters Precision, Recall and F1 Score for all three categories predicted by the model. The Figures below shows the performance measures of Naive Bayes, Support Vector Machine, and Gradient Boosting based classifiers respectively in terms of Precision, Recall, and F1-Score. Similarly,

In **Table 2** shows the result of the analysis of the classifier Multiple Naive Bayes, in **Table 3** shows the analysis and performance of the Support Vector Machine, and **Table 4** show the accuracy and performance analysis of the Gradient Booster Classifier, and **Table 5**, show the performance measures of the classifiers in terms of Model Accuracy of Naive Bayes, Support Vector Machine, and Gradient Booster classifier.

Table 2 Multiple Naive Based Classifier

|  |  |  |  |
| --- | --- | --- | --- |
| **Sentiment** | **Precision** | **Recall** | **F1-Score** |
| Negative | 92 | 99 | 95 |
| Neutral | 94 | 83 | 88 |
| Positive | 90 | 94 | 92 |

Table 3 Support vector Machine

|  |  |  |  |
| --- | --- | --- | --- |
| **Sentiment** | **Precision** | **Recall** | **F1-Score** |
| Negative | 99 | 98 | 98 |
| Neutral | 91 | 97 | 94 |
| Positive | 98 | 91 | 95 |

Table 4 Gradient Booster Classifier

|  |  |  |  |
| --- | --- | --- | --- |
| **Sentiment** | **Precision** | **Recall** | **F1-Score** |
| Negative | 98 | 91 | 94 |
| Neutral | 87 | 97 | 92 |
| Positive | 98 | 94 | 96 |

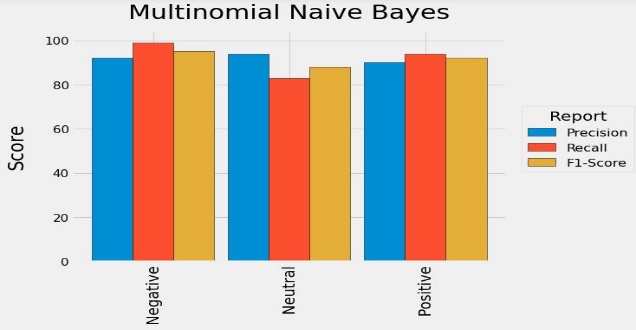
Table 5 Overall Precision and Accuracy

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Multiple naïve based | 92 |
| Support vector Machine | 96 |
| Gradient Booster Classifier | 94 |

The analysis of the above-mentioned tables we will be discussing the analysis graphically and get the clear results of precision and accuracy.

**Figure 2** represents the Precision, Recall and the F1\_Score of the Multiple naïve Based Classifier, **Figure 3** represents the Precision, Recall and the F1\_Score of the Support vector Machine, and **Figure 4** represents the Precision, Recall and the F1\_Score of the Gradient Booster classifier.

Similarly, the overall model accuracy can be analyzed in **Figure 5**

Figure 2 Multiple Naive Based Classifier

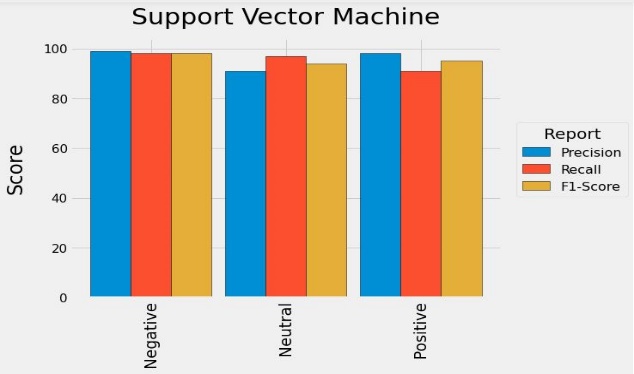
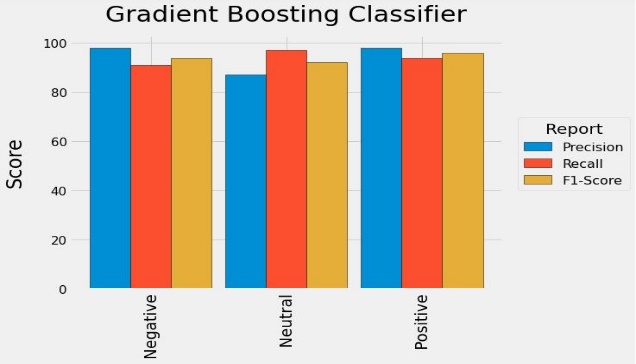
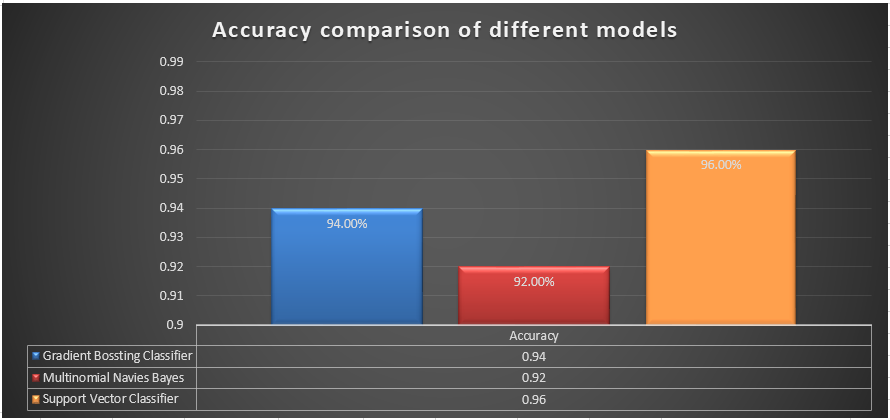
Figure 3 Support vector Machine

Figure 4 Gradient Booster Classifier

Figure 5 Models Accuracy

1. **Conclusion**

In this paper, we proposed a set of techniques of machine learning with semantic analysis for classifying the tweets of the hot trending political topic “Imported Government Not Accepted”. The key aim was to analyze the large number of tweets and was to classify them Using Natural Language Processing techniques. Three different Classifiers were used and was concluded that Support Vector Machine (SVM) applied on Trigram Analysis as Positive, Negative and Neutral gives the highest Accuracy of 96% then the other comparative models that gived 62% and 94% respectively.

1. **References**

[1] G. Gautam and D. Yadav,” Sentiment Analysis of Twitter Data Using  
Machine,” p. 6, 2014.

[2] B. Gupta, M. Negi, K. Vishwakarma, G. Rawat and P. Badhani,” Study  
of Twitter Sentiment Analysis using Machine,” International Journal of  
Computer Applications (0975 – 8887), vol. 165, p. 7, 2017.

[3] M. Kanakaraj and R. M. R. Guddeti,” Performance Analysis of En-  
semble Methods on Twitter Sentiment Analysis using NLP Techniques,”  
Proceedings of the 2015 IEEE 9th International Conference on Semantic  
Computing (IEEE ICSC 2015), p. 2, 2015.

[4] B. Le and. H. Nguyen,” Twitter Sentiment Analysis Using Machine  
Learning Techniques,” 2015.

[5] P. Yang and Y. Chen,” A Survey on Sentiment Analysis by using  
Machine,” 2017.  
[6] N. M S and R. R,” Sentiment Analysis in Twitter using Machine,” 2013.

[7] I. G. P. S. Varma and A.,” Sentiment Analysis Tool using Machine  
Learning Algorithms,” 2013.

[8] M. Ahmad, S. Aftab, S. S. Muhammad and S. Ahmad,” Machine Learn-ing Techniques for Sentiment Analysis,” INTERNATIONAL JOURNAL  
OF MULTIDISCIPLINARY SCIENCES AND ENGINEERING, vol. 8,  
2017.

[9] R. S. Jagdale, V. S. Shirsat and S. N,” Sentiment Analysis on Product  
Reviews Using Machine Learning Techniques,” vol. 768, 2018.

[10] C. Dhaoui, C. M. Webster and L. P. Tan, ”Social media sentiment anal-  
ysis: lexicon versus machine learning,” Journal of consumer marketing,  
vol. 34, 2017.

|  |  |
| --- | --- |
| [11] | S. Bhut, A. Doshi , U. Doshi and M. Narvekar , "A Review of Techniques for Sentiment Analysis," 2014. |
| [12] | S. Liao, . J. Wang, R. Yu, K. Sato and . Z. Cheng, "CNN for situations understanding based on sentiment analysis of twitter data," *8th International Conference on Advances in Information Technology,* 2016. |
| [13] | A. S. Neogi, K. A. Garg, R. K. Mishra and Y. . K. Dwivedi, "Sentiment analysis and classification of Indian farmers’ protest using twitter data," *International Journal of Information Management Data Insights 1 (2021) 100019,* p. 11, 4 June 2021. |
| [14] | M. Yasen and . S. Tedmori , "Movies Reviews Sentiment Analysis and Classification," p. 6, 2019. |