Comparative Study of Support Vector Machine and Naïve Bayes Classification Algorithm on Amazon Data

[1] Priyanka Tyagi, [2] R.C Tripathi
[1] Assistant Professor,Lingaya's Lalita Devi Institute of Management, New Delhi
[2] Professor, Lingaya'sVidyapeeth,Faridabad

Abstract:

This paper explores the comparison of support vector machine and Naïve Bayes classification algorithm on the basis of accuracyand confusion matrix parameter. This paper consider sentimental review of Amazon prime movies. This paper shows the SVM achieve substantial performance over Naïve Bayes and behave robustly over a different parameter of matrix.To extract the data Tweepy api application software is used . Tweepy is open sourced and enables python to communicate with twitter and used its API and extracted data stores in csv files.Twitter character length is extend to 280 from 140 in 2018.The most common length of tweet is 33 character only.Only 12% of tweets hit twitter's 140 character limit.Only 1% of tweets hit twitter's 240 character limit.

Keywords: *introduction, methodology, metrics, implementation, future scope*

I. Introduction

Sentiment analysis (or) opinion mining is an area of research that analyzes opinions, sentiments, evaluations, attitudes, and emotions from a written text. The feelings of others have a key effect in our day by day process. The way to deal with sentiment order neglects to fulfill execution when moving to different areas. These choices range from purchasing an item, to making interests in purchasing a property or to watch a film in a theater and so forth . prior, individuals would look for assessments on items and administrations from sources, for example, known persons, neighbours or online networking. The web has a gigantic measure of opinionated data, as web journals, surveys, the unsupervised methodologies.[10] thrive. Ideas use reviews, assessment surveys, and online networking as a device to obtain criticism on their items and services. The web is the impetus for these progressions. More than eighty percent of data on the Internet is unstructured. It so happens that more than eighty rate of information on the Internet is unstructured and is accessible from input fields in review, sites, wikis etc[8].

Since the execution is subject to the alternatives of the data, numerous studies dedicate on building capable information accessible with cautious designing. In this, we address the outline of systems, ways and means which are steady and set apart as the fundamental field in the area of Sentiment Analysis. Many pioneering research works have been proposed in the past with regards to this research area. This immense volume of information may forces potential beneficial business related data, which when separated keenly and spoke to sensibly, can be a mine of gold for administrations Research andDevelopment (RandD), attempting to add library an item in light of prominent popular opinion.Our work aims at forfeiting these researches with greater accuracy and potential to excel in both retail and corporate forums. newline

II. Methodology

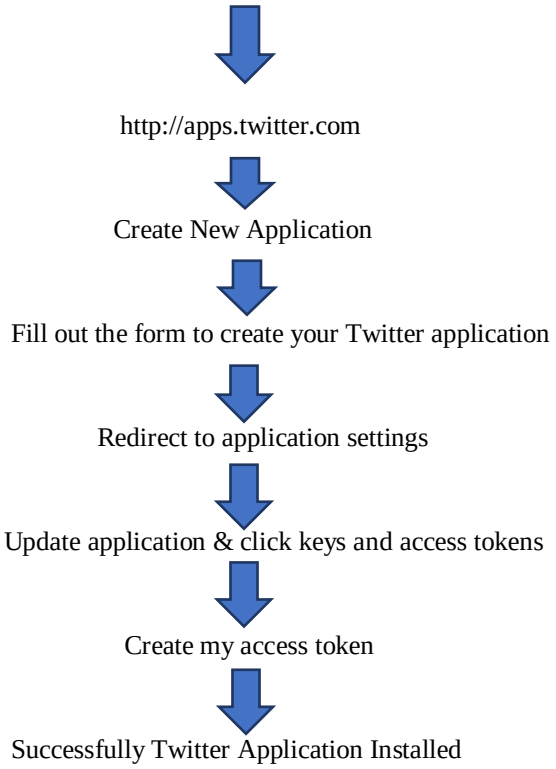
Data Extraction: To extract the data Tweepy api application software is used . Tweepy is open sourced and enables python to communicate with twitter and used its API.Tweepy is the library in order to authorize to access twitter on our behalf. Twitter allows a maximum of 3200 tweets for extraction.[6]

Pip is package management system used to install and manage software packages written in Python. Some packages have default pip program and some have to install pip to make Python program easier and faster.

To get the painless authentication we have to create Oauth handler instance in this we have to pass our consumer token and secret which we create during application setting in Tweepy application installation.[10]Twitter character length is extend to 280 from 140 in 2018.The most common length of tweet is 33 character only.Only 12% of tweets hit

twitter's 140 character limit.Only 1% of tweets hit twitter's 240 character limit.

Steps to install Tweepy:



III. Metrics to evaluate the Algorithms

Classification Accuracy:

Classification accuracy is the performance measure or it is the ratio of correctly predicted observations to the total number of input samples .It works well only if there are symmetric datasets.[5]

Accuracy = GP+GN/GP+FP+FN+GN

Genuine Positives (GP) –Genuine positives values are correctly predicted data which means that the data value of actual class is yes and the data value of predicted class is also yes. E.g. if genuine class value tell that this student will win and predicted class tells yes also.

Genuine Negatives (GN) –Genuine negative values are the correctly predicted data which means that the data value of actual class is nodatavalue of genuine class is no and data value of predicted class is also no.

Fake Positives (FP) – When the data value of genuine class is no and the data value of predicted class is yes. E.g. if genuine class says this student did not win but predicted class tells you that this student will win.

Fake Negatives (FN) – When the data value of genuine class is yes but the data value of predicted class in no. E.g. if genuine class value tells that this studentwin and predicted class tells you that the student will not win.

IV. Comparison of Naïve Bayes and SVM on the basis of Matrics

It is often easier to run everything. Usually some completely unexpected model will perform best. You could then profile the dataset based on which models worked best. Naive Bays is the generative models for classification. SVM is based on discriminant function given by y+w.x+b.[4] It tries to find a hyperplane that maximizes the margin and there is optimization, performance wise SVMs likely to perform better result as compare to Naïve Bayes . In our study model we compare Naïve Bayes and SVMs. On the basis of AMAZON PRIME MOVIES REVIEWS with the reference of [Table 1] in respect of accuracy. The total no. of movie reviews are 1964. From which we classified the data into three classes.

Total Movies Reviews = 1964

Genuine Review	Fake Reviews	Natural Reviews
960	981	21

Table 1.

Naïve Bayes: The Naïve Bayes algorithm is a machine learning algorithm for classification problems. It is primary used for text classification ,which involves high dimensional training datasets. Span filtration, sentimental analysis. It is very fast and probalistic classifier [7].

The Naïve Bayes algorithm is "Naïve" because it makes the assumptions that the occurrence of a certain features is independents of the occurrence of other features. $P(A|B) = \underline{P(A \cap B)}$

$$P(B)P$$

$$P(B|A) = \underline{P(B \cap A)}$$

$$P(A)$$

$P(B|A).P(A) = P(A|B).P(B) \rightarrow P(A|B)$

$$= \underline{P(B|A).P(A)}$$

$$P(B)$$

Multiclass evaluation is an extension of the binary case. It is a collection of Genuine versus predicted data. Here confusionmatrics are useful to generate classification report. The overall evaluation metrics are average across the classes.In Macro average each class has equal weight,compute metric within each class, average resulting matrics across classes.

In Micro average each instance has equal weight largest classes have most influlcence .outcwe calculate micro average with aggreate outcomes across all classes and compute metric with aggregate outcomes [5].

When we applied Naïve Bayes on the [Table 1.] data the accuracy score is 67.22 with the reference of [Table 2.].We can also check the accuracy graph with the reference of figure 1.

Classes	Precision	Recall	f 1 - score	Support
Fake	0.63	0.86	0.73	399
Genuine	0.76	0.49	0.63	375
Neutral	0.00	0.00	0.00	10
Micro Avg.	0.67	0.67	0.67	784
Macro Avg.	0.46	0.45	0.44	784
Weighted avg.	0.68	0.67	0.66	784

Table 2.

```
The accuracy score is 67.22%
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\metrics\classification.py:1143:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in
labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
```

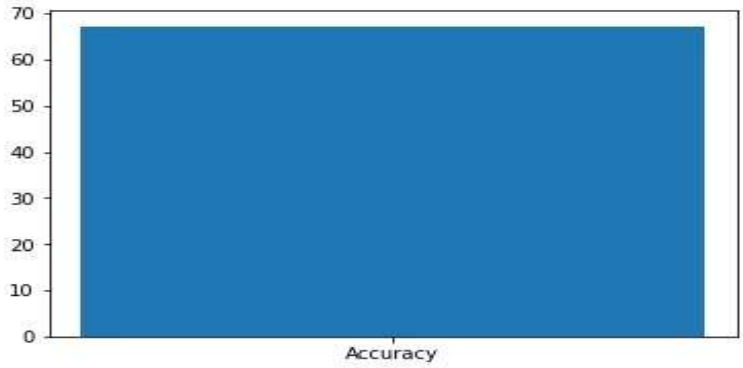
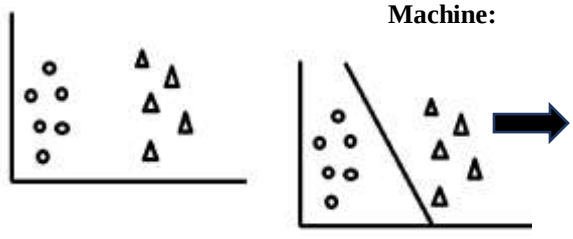


Figure 1.

Support Vector Machine:

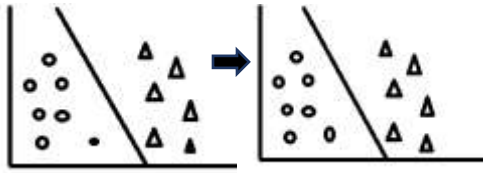
Support Vector Machine is supervised learning model learns from the past input data and makes future prediction as output. Support vector machine is a supervised learning method that's looks at data and sorts it into one of two and many categories. In SVM we divide the dataset into classes by using 1-D ,2-D and 3-D hyperplane basis on the complexity of datasets[4].

Process of Support Vector Machine:



Sample data Boundary Labeled Sample

Boundary data



DecisionPlotting Predicting the

The new unknown

Figure 2.

If the classes have the same number of samples then macro and micro average will be the same.[2] If some classes are much larger samples than others, then you can weight your metric with the largest ones and use micro averaging in other case you can weight your metric with the smallest ones use macro averaging ,if the micro average is much lower than the macro average then identify the larger classes for poor metric performance. If the macro average is much lower than the micro average then identify the smaller classes for poor metric performance. The accuracy score is 81.38.

When we applied SVM on the [Table 1.] data the accuracy score is 81.38 with the reference of [Table 3.].We can also check the accuracy graph with the reference of figure 3.

Classes	Precision	Recall	f 1 – score	Support
Fake	0.83	0.80	0.82	399
Genuine	0.79	0.85	0.82	375
Neutral	0.00	0.00	0.00	10
Micro Avg.	0.81	0.81	0.81	784
Macro Avg.	0.54	0.55	0.55	784
Weighted avg.	0.80	0.81	0.81	784

Table 3.

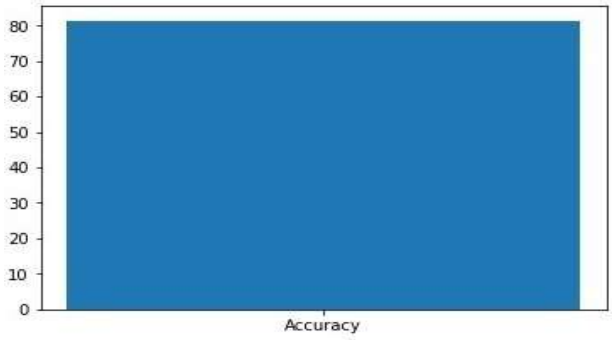


Figure 3.

V. Comparative Study

The comparison between the models on the basis of features point of view the Naïve Bayes is independent in nature ,Whereas the SVM check the interactions point between them to a certain degree but in the theoretical case it is hard to compare both of them because one is probabilistic in nature and other is geometric in nature.After the study of Naïve Bayes we conclude,properly trained Naïve Bayes classifiers are usually provide accurate result and very fast to train and noticeably faster than any classifier builder I have ever used.

VI. Conclusion and Future Scope

When you want to improve classifier prediction you can tune classifier and apply some sort of classifier combination technique, so in our research we build a hybrid classifier to increase the performance of the classifier.Microblogging Siteor Social media site, applications, including you tube, Facebook, Twitter and blogs, have become the attractive platform for youth today.In earlyworld we were used text extraction methods and factextraction methods to examine the relation between polarity classification and subjectivity detection for compress the reviews into shorter extracts, after that the researchers by using supervised machine learning on different topics in conditional sentences are positives,negatives or neutral. Sentiment analysis challenges appear in the field through reviews sites, blogs, news and discussion forum.

This paper presents the machine leaning algorithms for twitter sentiment analysis. Experimental results show the comparison between two methods and compare the result on the basis of accuracy. In today world various hybrid methods are proposed for determining the accuracy by exporting the data from microblogging and social media sites. In future, we would like to incorporate rules for handling sarcasm and more discourse relations.

VII. References

[1] Harsh Thakkar and Dhiren Patel , Approaches for Sentiment Analysis on Twitter A State-of-Art study Dept. of Computer Engineering, National Institute of Technology, Surat.

[2] Orestes Appel ; Francisco Chiclana ; Jenny Carter ; Hamido Fujita.,2016 IEEE Congress on Evolutionary Computation (CEC),2016.

[3] Dhaoui, C., Webster, C. and Tan, L. (2017). Social media sentiment analysis: lexicon versus machine learning. Journal of Consumer Marketing, 34(6):480–488.

[4] Thakare Ketan Lalji, Sachin N. Deshmukh , Twitter Sentiment Analysis Using Hybrid Approach, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 - 0056 Volume: 03 Issue: 06 | June-2016

[5] B. Keith & E. Fuentes & C. Meneses. A Hybrid Approach for Sentiment Analysis Applied to Paper Reviews KDD'17, August 2017, Halifax, Nova Scotia, Canada

[6] http://cs229.stanford.edu/proj2016/report/Tsui-PredictingStockPriceMovementUsingSocialMediaAnalysis-report.pdf

[7] Approaches for Sentiment Analysis on Twitter A State-of-Art study Harsh Thakkar and Dhiren Patel, Depa rtment of Computer Engineering, National Institute of Technology, Surat-395007, India

[8] https://pdfs.semanticscholar.org/501c/fa39b2d94443ddbf4bfb7bb82c9845fa789c.pdf

[9] Xia Liu, (2019) "A big data approach to examining social bots on Twitter", Journal of Services Marketing, https://doi.org/10.1108/JSM-02-2018-0049

[10] tyagi, Priyankaand Chakraborty, Sudeshna, Literature Review of Sentiment Analysis Techniques for Micro blogging Site(March, 2019).http://dx.doi.org/10.2139/ssrn.3403968.