ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse179952020.pdf
https://doi.org/10.30534/ijatcse/2020/179952020

Comparison of Classification Algorithms on Twitter Data Using Sentiment Analysis

S Anjali Devi, Prastut Sapkota, K Rohit Kumar,S Pooja , M Sai Sandeep
Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur, Andhra Pradesh, India
Email: swarnaanjalidevi@gmail.com

ABSTRACT

For the past few decades tremendous effort were dedicated for sentiment analysis. One of the uses of sentiment analysis is to get customers view of a product or company. In this paper we have explored sentiment analysis using two classification algorithms i.e. Support Vector Machine (SVM) and Naïve Bayes considering the positive and negative sentiments using the twitter dataset. Our implementation results indicate that SVM provides the best accuracy which is followed by the Naïve Bayes.

Key words: Sentiment analysis, Naïve Bayes, SVM.

1. INTRODUCTION

In the fourth industrial revolution whenever a product is released the plan of action of the industrialists depends on the customer reviews. Sentiment analysis is a machine learning approach in which machines break down and characterize the human's opinions, feelings, sentiments and so forth about some theme which are communicated as either content or discourse [3]. These reviews play a key role in deciding the future of the product. But it is not possible to go through every review and there is very limited time to take the next step. This is the point where sentiment analysis plays an important role. Sentiment analysis (SA) or opinion mining is an application of cognitive model. Sentiment analysis is basically the process of categorizing sentiments or opinions from a piece of text. In sentiment analysis there is a binary decision made like users like or dislike something, somebody is either with or in disagreement with something, or the product is either the best or worse.

In this paper we get the opinions from the twitter dataset. To perform sentiment analysis, we need to decide which algorithm classifies the given text into positive or negative sentiment efficiently. The aim of SA is to locate the expressions, determine the sentiments and classify the nature of the sentiment [2]. We are going to provide a novel method to compare and decide the best algorithm among Naïve Bayes and SVM. The comparison of result is cleanly visualized using charts.

In this paper, we analyze the following three steps given below,

Data Collection and Pre-processing: The accuracy of any mining model depends the dataset which is used for training. To increase the efficiency of the model we pre-process the data-set based on our requirements. Unnecessary words are omitted

Mining: Two main classification algorithms i.e. Support Vector Machine (SVM) and Naïve Bayes are implemented and evaluated. The focus is providing the positive negative and neutral analysis by opinion Mining [4].

Result: The accuracy of the give algorithms is retrieved from the model by performing sentiment analysis on the twitter dataset. These results are visualized using charts. In addition to academic research, sentiment analysis is now a standard part of online business intelligence software, such as Market Sentinel's Skyttle and sysomos's Map [10].

2. ALGORITHM

In this paper, we used two of the best classification algorithms, Naïve Bayes and Support Vector Machine (SVM). These algorithms were implemented in python version 3.6.

Naïve Bayes: Naive Bayes algorithm is the machine learning algorithm for classification problems. Machine learning a transpire technology which is engaged with almost all the fields, where its algorithms are more powerful that give with better faultless results [5]. It is mostly used for the text classification involving training data sets of higher dimensionalities. Naive Bayes classifiers are highly scalable and are useful when the dataset is of higher dimensionality i.e. a greater number of columns [1]. Some classification algorithms performances have been compared over a data

set. A few examples are spam filtration, sentimental analysis and classification new articles. This algorithm is very simple and effective we can build models very quick and make fast predictions by using this algorithm, Naive Bayes learns the probability of an object with certain features that belongs to a particular group in class. Navie Bayes is a famous sort calculation which grew fundamentally dependent on Bayes Theorem [6]. This algorithm is a probabilistic classifier, Naïve Bayes algorithm is called Naïve because it makes the assumption that the occurrence of a certain feature will not depend on the other features [14]. Whereas coming to the Bayes part which refers to the statistician and philosopher whose name is Thomas Bayes he named the theorem as Bayes theorem and it is the base part for the Naïve Bayes Algorithm. Naïve Bayes Classifier: This is a classification algorithm based on probabilities. This algorithm is mainly used for text classification or in any classification problem where the features are categorical [8]. Bayes Theorem gives a method to calculate the conditional probability which the probability of an event based on a previous knowledge available on the events

$$P(D/x) = \frac{p(x/D)p(D)}{p(x)}$$

$P(D/x) = p(x_1/D) * p(x_2/D) \ldots * p(x_n/D) * p(c)$

$P(D/x)$ is a posterior probability of class.

$P(x)$ is the prior probability.

$P(x/D)$ is the likelihood from which the probability is the predictor of the given class.

$P(x)$ is prior probability of the predictor.

Support Vector Machine (SVM): Support vector machine is a machine learning algorithm which is used for classification. Traditional machine learning classifiers such as SVM, naïve bayes and rule mining techniques are capable of finding essential patters [9].Each and every point or feature is mapped onto an n dimensional space and need to classify them into two different classes by using a hyper plane; whereas hyper plane is just a line for two dimensional spaces, when we take three dimensional spaces then hyper plane is used to classify the two classes separated as wide as possible.

To choose a right hyper plane we are having two rules:

Rule 1: Select the hyper plane so that which segregates the classes better.

It means we need to separate the two classes as wide as possible which in turns that there should not be any misclassification and we should be able to tell that given a new point which class it belongs to.

Rule 2: Maximize the distance between the nearest data points of either class or Margin.

Which means that the maximum distance from the data point of either class should me more and they should be maximum distance from the nearest data points of either class?

SVM [15] is very effective to the high dimensional spaces we can do any mathematic problem which is an n dimensional using SVM, as we are using only support vectors it will be very memory efficient and it need not to occupy more space. By using we will get an optimal solution.

3. FRAMEWORK

In the Sentence Based Analysis, we basically focus on three words Positive, Negative and Neutral. The positive contains people's reactions like happiness, Surprised, etc. The negative contains reactions like anger, sadness, fear etc. The neutral contains all the reactions that does not comes under positive and negative. Figure 1 describes the framework of sentiment analysis.

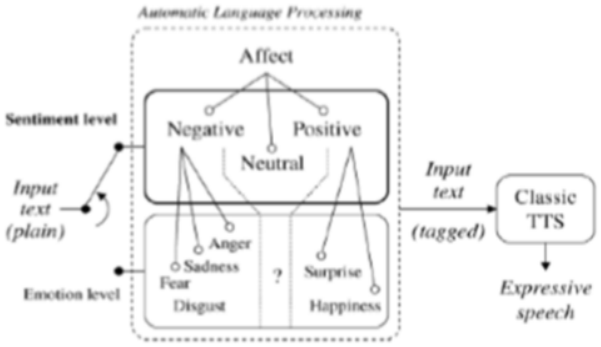

Figure 1: Framework of sentiment analysis

Let us consider some datasets where we need to identify the sentiments whether it is positive, negative or neutral. Table1 lists attributes of the considered data set. Here, there is an example of three sentences and how it is categorized into the sentiments.

Table 1: Attributes of Cleveland dataset

Sentiment	Query	Tweet
Positive	IPhone	Doug123: IPhone is the best phone ☺
Neutral	Cineplaza	Helena2: arrived Cineplaza this morning.
Negative	charge	Sammy22: This phone takes a lot of time to charge ugh!!

Here, we used two algorithms, Naïve Bayes[16] and SVM and analyzed which performs better. First, we used the same data set for both the algorithms and the outputs were analyzed depending upon their accuracy and efficiency. Figure 2 describes the flowchart of the process that we adopted.

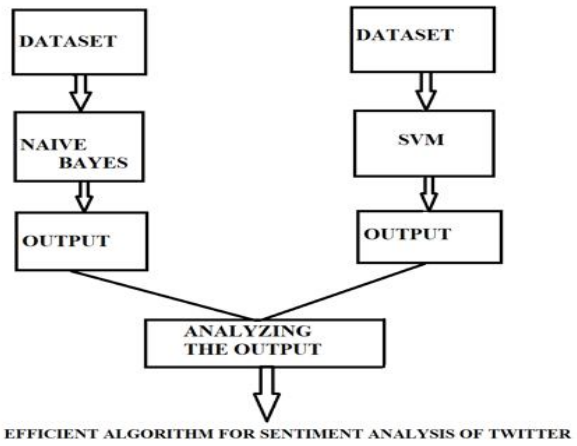

Figure 2: Flow graph of the proposed model

4. IMPLEMENTATION

The first step that comes up in implementation part is Data pre-processing. This helps in getting better results through the classification algorithms. It includes the following steps:

1. Remove blank rows if any.
2. Change all the text to lower case. This is required as python interprets 'dog' and 'DOG' differently.
3. Tokenization: In this each entry in the corpus will be broken into set of words
4. Remove Stop words, Non-Numeric and perform Word Stemming/Lamenting.

Therefore, we manually select a few seed words for each pre-defined aspect and use them as input to the aspect segmentation algorithm [11].

After the data is pre-processed we then split the model into train and test data set. Natural language tool kit (NLTK) in Python is most popular natural language processing package for English [7]. Now, we encode the target variable. This is done to transform Categorical data of string type in the data set into numerical values. After obtaining the words we vectorize the words using TF-IDF Vectorizer. There are several streams of research investigating the role of Twitter in social media, product marketing, and project management [12]. This is done to find how important a word in document is in comparison to the corpus. Now, the final step includes running the two classification algorithms to classify out data check for accuracy. We perform the following operations on both the classification algorithms:

1. fit the training dataset on the classifier
2. predict the labels on validation dataset

Use accuracy score function to get the accuracy.

5. RESULTS

In this paper, we used three different dataset of 500 tweets, 1000 tweets and 1500 tweets respectively. From this project we obtained the accuracy score of both naïve Bayes and SVM of these three datasets. The results of the dataset are shown in figure 3, figure 4 and figure 5 respectively for dataset of 500 tweets, 1000 tweets and 1500 tweets.

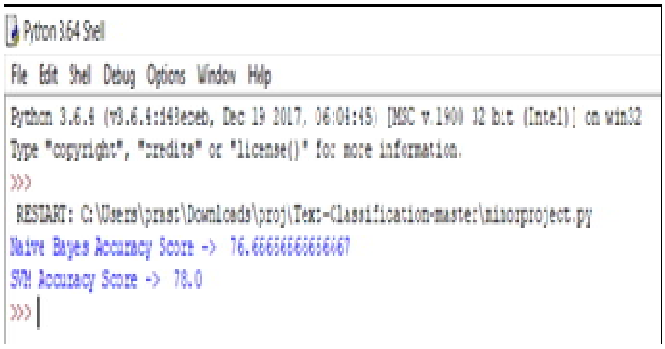

Figure 3: Result of dataset with 500 tweets

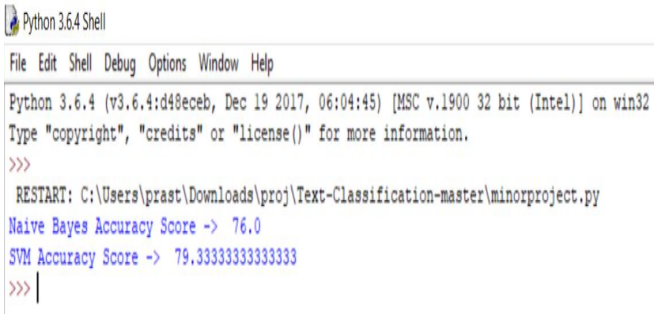

Figure 4: Result of dataset with 1000 tweets

Figure 5: Result of dataset with 1500 tweets

6. CONCLUSION

From this paper, we were able to determine that both the algorithms are the best classification algorithms. The best way to avoid mistakes in growth curves is discussed considering four different classification techniques of machine learning on the collected data set[13]. Here, SVM algorithm is a bit bitter as per the accuracy score. we obtained from the results. SVM algorithm is a bit efficient by slightest of margin then Naïve Bayes algorithm. Now, with this project developer of various software companies can develop software based on sentiment analysis through SVM as it is more efficient. Also, this project gives us the information of the social media and its huge data information about products.

It will also help people to determine the best products for their regular use.

7. FUTURE WORK

A tweet can be classified in a better way by grasping the role of nouns. Let's take an example, "Safari often crashing: Apple, what are you doing?" Here Safari is a browser and a product whereas Apple is a company. Semantic labelers can be used to achieve the semantic analysis.

REFERENCES

1. Anjali Devi, Prastut Sapkota and M.Obulesh, **Sentiment analysis on products using social media** *Jour of Advanced Research in Dynamical & Control Systems*, Vol. 10, 02-Special Issue, 2018.
2. Muthukumaran S., Suresh P., Amudhavel J. **Sentimental analysis on online product reviews using LS-SVM method** *Journal of Advanced Research in Dynamical and Control Systems* vol.9 special issue. 12, pp.1342-1352.
3. Balaji P., Nagaraju O., Haritha D. **Levels of sentiment analysis and its challenges: A literature review** Proceedings of the 2017 International Conference On Big Data Analytics and Computational Intelligence, ICBDACI 2017.
4. Vyshnavi R., Venkata Raju K., Vamsi Krishna G., Bhavya Shree Y **A survey on sentiment study in twitter data using Hadoop streaming API** *International Journal of Engineering and Technology(UAE)* vol.1 pp-618-621.
5. Uma Ramya V., Thirupathi Rao K **Sentiment analysis of movie review using machine learning techniques** *International Journal of Engineering and Technology(UAE)* vol.7 pp-676-681.
6. Rachapudi V., Vaddi S.S., Karumuri R.R., Sripurapu S. **Heart disease prediction using machine learning algorithms** *International Journal of Recent Technology and Engineering*,2019
7. Londhe D.D., Kumari A., Emmanuel M **Language identification for multilingual sentiment examination** *International Journal of Recent Technology and Engineering* vol.8 pp-3571-3576,2019
8. Paul C., Sahoo D., Bora P **Aggression in social media: Detection using machine learning algorithm** *International Journal of Scientific and Technology*,2020
9. Atmakuri, Krishna Chaitanya; Rao, Y. Venkata Raghava **An IOT based Novel approach to predict Air Quality Index (AQI) using Optimized Bayesian Networks** *Journal of Mechanics of Continua and Mathematical Sciences*,2019
10. M. Thelwall, K. Buckley, and G. Paltoglou, **Sentiment in twitter events** *J. Am. Soc. Inform. Sci. Technol*., vol. 62, no. 2, pp. 406–418, 2011.
11. Hongning Wang, Yue Lu, Chengxiang Zhai; **Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach** *ACM New York, NY*, USA 2010.
12. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. **Welpe Predicting elections with twitter: What 140 characters reveal about political sentiment** in Proc. *4th Int. AAAI Conf. Weblogs Soc. Media, 2010*, vol. 10, pp. 178–185.
13. Venubabu Rachapudi et.al. **A Comparative Analysis of Classification Algorithms for Fetal Growth** Journal of *Advanced Research in Dynamical and Control Systems*, Vol 9, issue Sp-18,pp 592-600, Dec-2017.
14. Bashir, AK; Arul, R; Basheer, S; Raja, G; Jayaraman, R; Qureshi, NMF **An optimal multitier resource allocation of cloud RAN in 5G using machine learning** *Transactions on Emerging Telecommunications Technologies*, 2019, DOI: 10.1002/ett.3627
15. Thein Yu and Khin Thandar Nwet **Myanmar News Sentiment Analyzer using Support Vector Machine Algorithm** *International Journal of Advanced Trends in Computer Science and Engineering*. Volume 8 No. 6, pp. 3520 – 3525, 2019.
16. Lakshmana Phaneendra Maguluri and R Ragupathy **A New sentiment score based improved Bayesian networks for real-time intraday stock trend classification** *International Journal of Advanced Trends in Computer Science and Engineering*. Volume 8 No.4, pp. 1045 - 1055, 2019.