International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Comparative Study of Classification Algorithms for Sentiment Analysis on Twitter Data

Swapnil Patil[1], Nimish Shinde[2], Nikhil Pagar[3], Rohit Yadav[4], Dr. Pravin Futane[5]

[1,2,3,4,5]*Pimpri Chinchwad College of Engineering, Pune*

Abstract - Sentiment Analysis, also called as opinion mining grows out of human need extract relevant sources, related sentences with opinions reading them, summarizing them and organizing them into usable forms. The need for textual mining or sentimental analysis was felt or increased suddenly due the outbursts of world wide web(www) and various social media platforms being available for public to express their views or opinions. Different Classifiers and their algorithms with comparative study is done.

Keywords — Sentiment Analysis, Classification Algorithms, Naïve Bayes, Random Forest, Support Vector Machine.

INTRODUCTION

A comparative study of most commonly used algorithms for sentimental analysis is performed. The task of classification is a very vital task in any system that performs sentiment analysis. This study presents the study of algorithms viz. 1. Naïve Bayes, 2. Random Forest, 3 Support Vector Machine. A basic theory behind the algorithms, when they are generally used and their pros and cons etc. are discussed. The reason behind selecting only the above mentioned algorithms are the extensive use in various tasks of sentiment analysis. Sentiment analysis of reviews is very common, the method of taking reviews has evolved over a period of time. The scope of expressing a person's thoughts is often restricted when people have to give reviews about a product in form of score / star ratings. But when a person is allowed to express reviews in form of open text he can be very precise about what aspects about the product are good and what are not. Sentiment analysis engine parses through this textual reviews and generate output in form of polarities i.e. – Positive, Negative or Neutral. This helps in finding the reasons behind crucial fluctuations in sales of products and they can be rectified accordingly.

A. Sentiment Analysis

Sentiment analysis is a task that involves information extraction from customer feedback and other authentic sources like survey agencies. As the word suggests it includes detecting sentiments of any individual from the text that is writes in digital format. There are wide applications of this concept. This concept became centre of attention since industry got revolutionized with the change in paradigm of "Sellers' Market" to "Buyers' Market" in order to capture market share.

Major steps in Sentiment analysis are:

• Text Extraction – This step involves extracting words from text that influence the outcome of the result.

• Text Refinement – This step involves refining text in form of relevant phrases, words etc.

• Text Classification – This step includes classification of text into its class (positive/negative)

• Score Aggregation – This step collects total scores from classifier and then aggregates it further to produce the total sentiment score

The approach followed in the processing of the Twitter data for sentiment analysis includes various steps such as Tweet downloader, tokenizer, pre- processing, feature classifier, SVM classifier and prediction. The fig. 2 presents the flow of operations followed considering this study.

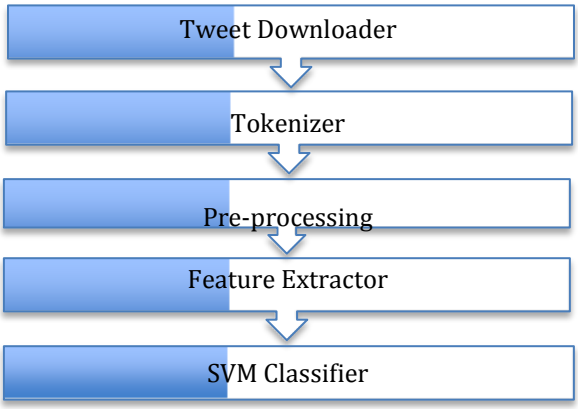

Fig. 1: Approach for sentiment identification

B. Twitter Data Collection Methods

The two possible ways to collect Tweets for research are as follows:

• APIs: Twitter provides two types of APIs such as search API and stream API. Search

API is used to collect Twitter data on the basis of hashtags and stream API is used to stream real time data from Twitter.

• Automated tools that are further classified into premium tools such as Radian6, Sysmos, Simplify360, Lithium and non-premium tools such as Keyhole, Topsy, Tagboard and SocialMention.

Twitter a social media platform consist of various kind of tweets with different content. These kind of tweets need to be taken care of in the process of sentiment analysis. The fig. 2 is an example of the random tweet available in Twitter

Fig. 2: Sample tweet

C. Data Pre-processing

Mining of Twitter data is a challenging task. The collected data is raw data. In order to apply classifier, it is essential to pre-process or clean the raw data. The pre- processing task involves uniform casing, removal of hashtags and other Twitter notations (@, RT), emoticons, URLs, stop words, decompression of slang words.

D. Feature Extraction

The pre-processed dataset has various discrete properties. In feature extraction methods, we extract different aspects such as adjectives, verbs and nouns and later these aspects are identified as positive or negative to detect the polarity of the whole sentence. Followings are the widely used Feature Extraction methods.

• Terms Frequency and Term Presence: These features denote individual and distinct words and their occurrence counts.

• Negative Phrases: The presence of negative words can change the meaning or orientation of the opinion. So it is evident to

• Parts Of Speech (POS): Finding nouns, verbs, adjectives etc. as they are significant gauges of opinions

E. Classification

Classification is a stage in sentiment analysis that can be described as a process in which we predict qualitative response, or in this case we classify the document into its polarity. Predicting a qualitative response of a document can be referred to as classifying the document since it involves assigning an observation to a category or class. There are many possible classification techniques, or classifiers that one might use for to predict the qualitative response or class of a document. In sentiment analysis some widely used classification techniques are as follows:

- Naïve Bayes Classifier
- Random Forest Classifier
- Support Vector Machine

I. NAÏVE BAYES CLASSIFIER[2]:

Naïve Bayes classifier is based on Bayes theorem. It's a baseline classification algorithm. Naïve Bayes classifier assumes that the classes for classification are independent. Though this is rarely true Bayesian classification has shown that there are some theoretical reasons for this apparent unreasonable efficiency. There are various proofs that show that even though the probability estimates of Naïve Bayes classification are low it delivers quite good results in real life examples.

In Text classification we tokenize the document in order to classify it in its appropriate class. By using the "Max Posterior Probability" Decision rule we get the following classifier:

$P(c/w)=[P(w/c)P(c)]/P(w)$

$c^*=\text{argmax}_c P(c/w)$

This means that in order to find in which class we should classify a new document, we must estimate the product of the probability of each word of the document given a particular class (likelihood), multiplied by the probability of the particular class (prior). After calculating the above for all the classes of set C, we select the one with the highest probability.

It's advised that this classifier should be used when Training time is a crucial factor in the system. Naïve Bayes is the baseline algorithm for researches in decision level classification problem. Some features of Naïve Bayes are:

- Accuracy
- Consistency – This algorithm shows consistency in results and if priors are used results also improve over a period of time.
- Performance / Efficiency - Can handle huge amounts of data
- Flexibility - The algorithm is flexible of having many different typed of data in a unified platform and classify it accordingly.
- Scalability

II. RANDOM FOREST CLASSIFIER[3]

Random forests are an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. It produces multi- altitude decision trees at inputting phase and output is generated in the form of multiple decision trees. The predictions are made by aggregating the predictions of various ensemble data sets.

MDist = $\sum$ (a-b)

Applications and real life of examples Random Forests are widespread. There is no single type for RF data sets. They can vary from any kind of applications like medical as well as general data sets. RF is a parallelized and multi-core friendly algorithm. So simultaneous running of different trees is also a support feature.

The popularity of this machine increased with practical machine learning research and their related algorithms. Performance of Random Forest for Opinion mining and have found impressive accuracy in classification of their data sets

III. SUPPORT VECTOR MACHINE [2]

Support vector machine (SVM) solves the traditional text categorization problem effectively; generally outperforming Naïve Bayes as it supports the concept of maximum margin. The main principle of SVMs is to determine a linear separator that separates different classes in the search space with maximum distance i.e. with maximum margin If we represent the tweet using t, the hyper plane using h, and classes using a set Cj € {l, -1} into which the tweet has to be classified, the solution is written as follows equivalent to the sentiment of the tweet.

The idea of SVM is to determine a boundary or boundaries that separate distinct clusters or groups of data. SVM performs this task constructing a set of points and separating those points using mathematical formulas.

F. Comparative Study

This study relates the various algorithms for the sentiment analysis to be performed on Twitter data. Considering the parameters of complexity, accuracy, training, efficiency this study concludes that the most suitable algorithm is SVM as compared to other two Naïve Bayes and random forest for the Twitter data, and the comparative study is shown in Table 1.

Table 1. Evaluation of algorithms

Algorithm	SVM	Naïve Bayes	Random forest
Understanding complexity	High	Very Less	Moderate
Theoretical accuracy	High	Low	High
Training speed	High	High	Low
Performance for small data	High	Less	Moderate

G. CONCLUSION

Study makes it pretty evident that every kind of classification model has its own benefits and drawbacks. Selection of classification models can be decided on the Twitter data and required accuracy. This study ultimately conclude the best algorithm for Twitter data sentiment analysis is SVM.

G. REFERENCES

[1]A. Celikyilmaz, D. Hakkani-Tur and Junlan Feng, "Probabilistic model-based sentiment analysis of twitter messages", IEEE Spoken Language Technology Workshop (SLT), pp. 79-84. (2010)

[2] J. Khairnar and M. Kinikar, "Machine Learning Algorithms for Opinion Mining and Sentiment Classification", in International Journal of Scientific and Research Publications, vol. 3, no. 6, (2013)

[3]Xiaowen Ding, Bing Liu, Philip S. Yu, "A holistic lexicon-based approach to opinion mining", Proceedings of the 2008 International Conference on Web Search and Data Mining. (2008)

[4] Y. Singh, P. K. Bhatia, and O.P. Sangwan, "A Review of Studies on Machine Learning Techniques," International Journal of Computer Science and Security, Volume (1) : Issue (1), pp. 70-84,(2007)

[5] Borges, J. L. "The analytical language of JohnWilkins", University of Texas Press. Trans. Ruth L. C. Simms(1964).

[6] Chinchor, N., Hirschman, L., and Lewis, D. L. "Evaluating Message Understanding systems" An analysis of the third Message Understanding Conference. Computational Linguistics, 19(3), 409–449, (1993).

[7] Heckerman, D., Horvitz, E., Sahami, M., and Dumais, S. T. "A bayesian approach to filtering junk e-mail." In Proceeding of AAAI-98 Workshop on Learning for Text Categorization, pp. 55–62 (1998)