IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Sentiment Analysis of Twitter Data Using Machine Learning Algorithms

Nishanth Vaidya

Student

Department of CSE,

Sambhram Institute of Technology, Bangalore, India

***Abstract*—**In the current era, opinions and ideologies of people are often expressed through the internet via sites like social media sites, micro blogging sites, etc. Sentimental analysis, a part of text mining is used classify these opinions or sentiments in a positive, negative or neutral manner. For this particular work, the sentiments of the tweets or reviews published in the twitter is identified by specific keywords in tweets. They tweets are then classified as to whether they have a positive or a negative outlook. The sentiments of the tweets are based on a specific selection of tweets. In this paper, we will check the accuracy of different fields such as precision, accuracy, etc, with three machine learning classifiers, namely, Random Forest, Naive Bayes and Support Vector Machine(SVM).

Keywords-Sentiment Analysis; Twitter analysis; Naïve Bayes classifier; Support Vector Machines

1. INTRODUCTION

The use of social media platforms such as Twitter and Facebook have skyrocketed in the past few years. People have been enabled to express their views and post their memories and experiences that they want to share with the world. Emotions play a major role while expressing their views. Sentiment analysis, also referred to as opinion mining is a technique that is used to study people's behavioural quirks. It is the combination of natural language processing and machine learning methods, the area under text mining that is gaining traction [1]. The tweets that are posted by people online, can be extracted, identified and evaluated and proves to be a vital source for decision making. The amount of posts, tweets, etc, that are being shared are scaling exponentially with every passing day. Statistics show that over 1400 Exabytes of data is transferred yearly via varying sources, out if which 80% is unstructured [2]. Sentimental Analysis proves to be a vital source for business intelligence for various uses such as marketing, prediction, ideas, etc. Hence Sentiment Analysis can prove to be a vital need for our generation and it can be achieved through various algorithms that will be discussed below. Usually, sentiment analysis is employed to figure out the attitude of the person based on the specific topics. Sentiment analysis can be applied to social media platforms such as Twitter, Facebook, Instagram, etc. For any particular topic, every individual will have their own opinion and it is of utmost importance that all of their opinions are gathered to find out the exact prevalent sentiment.

The basic task of sentiment analysis is emotion recognition as well as polarity detection. Sentimental analysis has been extensively used in applications such as emotion or sentiment mining for analyzing reviews as well as opinion about products and also in trending topics such as political analysis, entertainment, etc [4], [5]. In this paper, we make use of an analytical model. These algorithms will work on data continuously. The machine learning approaches are employed in diverse fields like analytics, IoT, Cyber Security and so on [6].

In this work following tasks are carried out:

- Tweets are downloaded from the twitter and then data pre-processing is one to the dataset taken.
- The pre-processed datasets are trained.
- Finally, the behaviour of Naive Bayes is analzyed with different topics to obtain the results for sentiment analysis.

- Naive Bayes classifier assumes that the presence of a particular article in a class is unrelated to the existence to any other feature of the said class.

We use naive bayes mainly due to the facts that:

- It is simple and efficient in dealing with larger datasets.
- It can outperform sophisticate sophisticated methods of classification.
- This approach helps the user to obtain the summarized report of the twitter data.
- Naive Bayes classifier is one of the machine learning algorithms that uses Bayes algorithm with the assumption of having independent features.

2. MACHINE LEARNING METHODS

Machine learning is one of the hottest topics in research and industry, with new methodologies being developed all the time. The speed and complexity of the field makes keeping up with new techniques difficult for experts and beginners alike.

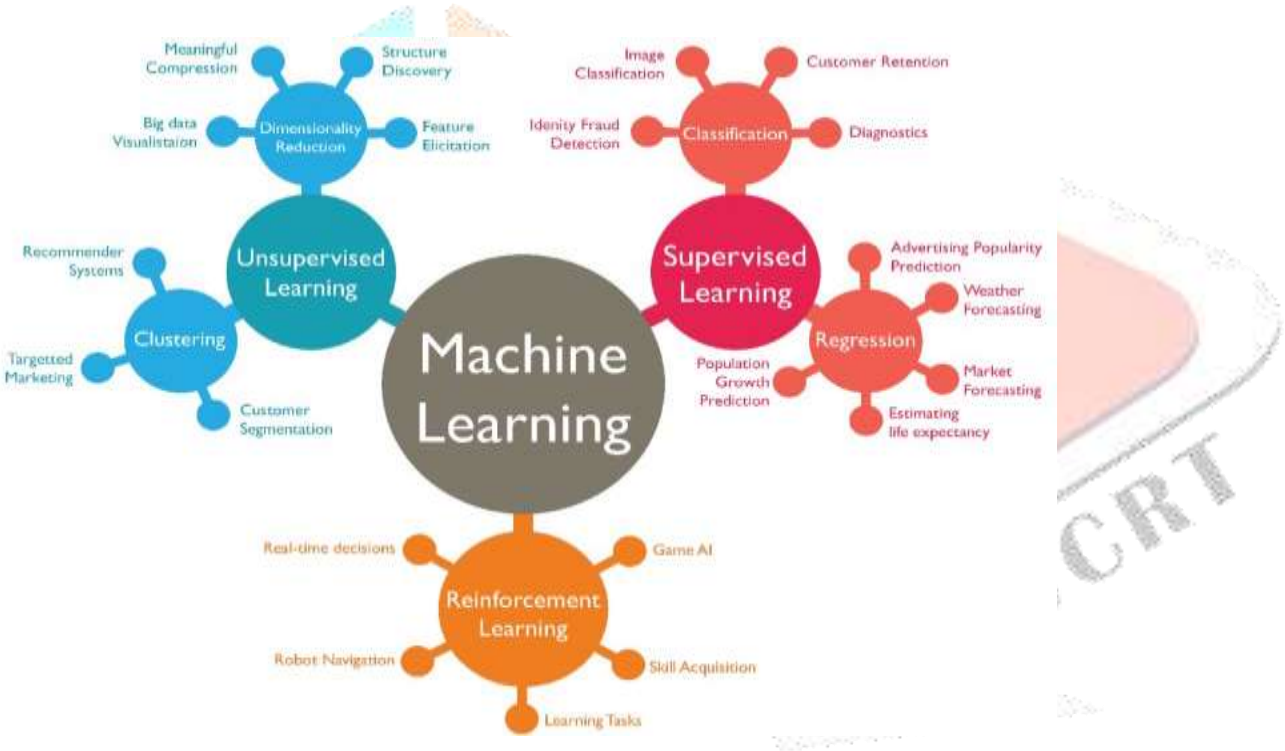

Figure 1 – Machine Learning Methods

A) Supervised Learning - Supervised learning is a type of machine learning task which learns a function that maps an input to an output based on example or training input-output pairs. It includes algorithms such as Support Vector Machines (SVM), k-Nearest Neighbours (k-NN) and Artificial Neural Networks (ANN), Genetic algorithms and Decision Trees (DT).

- **Naïve Bayes (NB) -** These algorithms entail probabilistic classifiers that make the prior assumption that the features of the input data are independent of each other. They are scalable and only require small to medium training datasets to produce appreciable results.
- **Support Vector Machines (SVM) - Support Vector Machines** or SVM for short, are models with associated learning algorithms which analyze the given datasets for both, classification as well as regression. Consider examples consisting of training data, with each example being marked as belonging to either of two categories. For these examples, the SVM algorithm builds a model that separates new examples to either category. Doing so makes it a non-probabilistic classifier. An SVM is a represents its examples as varying points in space, mapped so that the examples of the separate categories are clearly distinct.

B) Unsupervised Learning – Unsupervised Learning is a type of machine learning algorithm that makes decisions from datasets consisting of input data without any labelled outputs. Unsupervised Learning consists of two tasks being Association and Clustering.

- **Clustering** – Clustering consists of grouping data points that present alike characteristics. Well known approaches typically include algorithms such as k-means and hierarchical clustering. Clustering methods are scalable but the scalability is limited in a sense, but they represent a feasible solution that is used as a phase before adopting a supervised algorithm or for other detection purposes.
- **Association**. The aim of association is to identify unknown patterns across data, making them suitable for the purpose of prediction. However, they tend to produce a large output of sometimes invalid rules, hence they require a human overseer to function.

3. LITERATURE SURVEY

1. Shamantha Rai B, Sweekriti M Shetty "Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance", IEEE 4th International Conference on Computer and Communication Systems 2019.
2. Saumya Chaturvedi, Nitin Mishra, Vimal Mishra "Sentiment Analysis using Machine Learning for Buisness Intelligence", IEEE 2017
3. Y. Wan and Q. Gao, "An ensemble sentiment classification system of twitter data for airline services analysis," in Data Mining Workshop (ICDMW), 2015 IEEE International Conference on, pp. 1318–1325, IEEE, 2015.
4. D. Vilares, C. Go´ mez-Rodr´ıguez, and M. A. Alonso, "Universal, unsupervised (rule-based), uncovered sentiment analysis," Knowledge-Based Systems, vol. 118, pp. 45–55, 2017.

4. CLASSIFIERS FOR SENTIMENT ANALYSIS

1. Random Forests Algorithm - The pre-processed tweets are first stored in a csv file and used as a training set. Out of the three classifiers are being considered, the Random Forest (RF) classifier is below, and it has many sub classes. While we have considered 100 estimators, it can be used for exponential number of estimators. Initially, the classifier has to be set with the trained classes. After which, predicting class is given as test to the classifier that matches the predicted class with an accuracy score module of the classifier and provides results. The stored csv file is forwarded to the next classifier i.e. Support Vector Machine (SVM).

Random Forest Algorithm

1: for Test Class do

2: Initialize the forest to a training class 1 and training class 2

3: testing class is given with test class to predict

4: Import the module of accuracy score

5: accuracy score and matrix of test class prediction given

6: end for

2. Support Vector Machines - A support vector machine is a supervised model that uses classification algorithms for group classification problems. It primarily compromises of Support Vector Classification (SVC) and Support Vector Regression (SVR). The SVC supports both binary as well as multiclass classifications. The support vector is the closest point to the separation hyperplane. In the classification process, the mapping input vectors located on the separation hyperplane side of the feature space all eventually fall into one class, and the positions fall into the other class on the other side of the plane. In case of data points that cannot be linearly separated, the SVM uses appropriate kernel functions to map them into higher dimensional spaces. In this method, initially the collected tweets are forwarded as the training set to SVM. It further requires arguments like class weight and the number of classes which has to fit with the training set. Finally, it is compared with imported accuracy score of SVM. This in turn gives accuracy score and metrics as the output.

SVM algorithm for Sentiment Analysis

1: for Test Class do

2: training the classifier Naive Bayes

3: linear SVC=SVM.linear SVC(C,class weight,dual).fit(training classes)

4: linear SVC.predict(test class)

5: accuracy score and matrix of test class prediction given

6: import the module of accuracy score

7: accuracy score and matrix of test class prediction given

8: end for

3. Naïve Bayes Classifier – It is the final classifier we consider in this paper. It is a probabilistic classifier and it assumes that the features used are independent to each. In this algorithm, the stored csv file is used as the training set for classifier. The algorithm incorporates Gaussian Naive Bayes to deal with real time inputs like tweets. This will distribute the class univariate distribution for the given data. Finally, prediction is done for given test and accuracy score is predicted. The output of this algorithm will result in the accuracy score.

Naïve Bayes Algorithm

1: for Test Class do

2: training the classifier Naive Bayes

3: gaussian nb=gaussianNB ().fit (training classes)

4: predicting the target variable test class clf.predict(test class)

5: import the models of accuracy score

6: accuracy score and matrix of test prediction given

7: end for

After using the aforementioned algorithms, It was observed that for most tweets the sentiment analysis turns out to be neutral. The actual values obtained are 87.7% of neutral sentiment, 9.3% of positive sentiment and merely 3% of negative sentiments.

5. CONCLUSION AND FUTURE ANALYSIS

Sentiment analysis is used to find the sentiment of author with regards to their tweets. In this work, tweets are extracted using a string search method and these tweets are subjected to sentiment analysis using RF, SVM and NB classifiers. They are done so in order to classify them into positive, neutral and negative. As a part of analysis, the machine learning classifiers RF, SVM are considered and their accuracy is estimated considering three features and accordingly increasing the number of tweets considered. Also, the precision of RF, SVM and NB is also calculated by increasing the number of tweets. As a part of future work, some more features will be added which will in turn improve the accuracy of the prediction. Also, we have considered only positive, negative and neutral polarity to label the tweets but other different labels can also be incorporated.

6. REFERENCES

[1]. Shamantha Rai B, Sweekriti M Shetty "Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance", IEEE 4th International Conference on Computer and Communication Systems 2019.

[2]. Saumya Chaturvedi, Nitin Mishra, Vimal Mishra "Sentiment Analysis using Machine Learning for Buisness Intelligence", IEEE 2017

[3] S. Rosenthal, P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, and V. Stoyanov, "Semeval-2015 task 10: Sentiment analysis in twitter," in Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pp. 451–463, 2015.

[4] Y. Wan and Q. Gao, "An ensemble sentiment classification system of twitter data for airline services analysis," in Data Mining Workshop (ICDMW), 2015 IEEE International Conference on, pp. 1318–1325, IEEE, 2015.

[5] Z. Jianqiang and G. Xiaolin, "Comparison research on text preprocessing methods on twitter sentiment analysis," IEEE Access, vol. 5, pp. 2870–2879, 2017.

[6] S. Tokle, S. R. Bellipady, R. Ranjan, and S. Varma, "Energy-efficient wireless sensor networks using learning techniques," Case Studies in Intelligent Computing: Achievements and Trends, pp. 407–426, 2014.

[7] M. Bouazizi and T. Ohtsuki, "A pattern-based approach for multiclass sentiment analysis in twitter," IEEE Access, vol. 5, pp. 20617–20639, 2017.

[8] W. Hua, Z. Wang, H. Wang, K. Zheng, and X. Zhou, "Understand short texts by harvesting and analyzing semantic knowledge," IEEE transactions on Knowledge and data Engineering, vol. 29, no. 3, pp. 499–512, 2017.

[9] D. Vilares, C. Go´ mez-Rodr´ıguez, and M. A. Alonso, "Universal, unsupervised (rule-based), uncovered sentiment analysis," Knowledge-Based Systems, vol. 118, pp. 45–55, 2017.

[10].V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in Proc. 8th Conf. Eur. Chap. Assoc. Comput.Linguist., Morristown, NJ: Assoc. Comput. Linguist, 1997, pp.174–181

[11]. A. Esuli and F. Sebastiani, "Determining the semantic orientation of terms through gloss classification," in Proc. 14th ACM Int. Conf. Inf. Knowl.Manage., 2005, pp. 617–624.