International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

A Novel on Classification Techniques of News with Help of Sentiment Detection

Shailja Joshi, Mayank Patel, Manish Tiwari

Abstract: ***The problem of data classification is an important topic in the field of machine learning and information retrieval. This has been widely studied and has been applied in many fields. There are multiple models which are proposed for the classification, like tree-structured classifiers, genetic algorithms, Bayesian classification, neural networks etc. These have a large range of applications in different areas like, fraud/spam detection, Customer Segmentation, Medical Diagnosis, Credit approval, weather prediction etc. This project tries to aim at a particular subclass of classification, namely sentiment analysis. Hybrid techniques should be applied in this field of study as each of the existing models have brought about some new expertise and their improvements need to be combined to give higher performance and accuracy. The sentiment analysis problem requires to take as input a block of text and correctly predict the sentiment of the writer or the speaker of the text. We have sufficient data to build a system that uses hybrid techniques like Naïve Bayes and combines the existing models to perform sentiment analysis on a dataset and study its results. The hybrid approach using Naïve Bayes to this problem gives promising results.***

Keywords: ***Naïve Bayes Classifier, Machine learning techniques, Sentiment Analysis***

I. INTRODUCTION

As the name suggests sentiment analysis stands for detecting, extracting, characterizing information from emotions, opinions, attitudes or feeling present in the set of data by the use of Statistics, Natural language processing (NLP) or Machine Learning. SA is a method of measuring opinions of people whether single or in groups, for instance data from a particular brand's customer care.

SA, according to a points mechanism monitors the discussion and further evaluates voice & language reflections to emotions, opinions, attitudes related to a service, product or a business. Hence, it has got multiple names like sentiment mining, opinion mining, subjectivity analysis, opinion extraction etc.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.

Shailja Joshi*, Student, Post Graduate, Department of Computer Science & Engineering, Rajasthan Technical University, India.

Mayank Patel, Associate Professor, Department of Computer Science and Engineering, Geetanjali Institute of Technical Studies, Udaipur, Rajasthan

Manish Tiwari, Associate Professor, Department of Computer Science and Engineering of Geetanjali Institute of Technical Studies, Udaipur.

The data getting uploaded on the internet is rising exponentially, there is a need to organize the data present over the internet which is moreover unorganized or unstructured. We can classify data broadly in three types namely, Experiences, opinions or facts. Polarity of these data can be categorized in multiple ways.

In general, we assume a binary opposition like good or bad, for or against, like or dislike etc. or in form of negative, positive and neutral. There could be various levels of polarity

Classification ranging from simple to Advanced. For instance, a simple would involve whether an emotion is positive or negative whereas a complex one would be rating the positivity/negativity present in the data. Each of the above present data can be further classified in multiple ways. For instance, opinions can be further classified as implicit, explicit and they can be even further classified too. Basically, data can be categorized in uncountably multiple types. Sentiment Analysis can be implemented at multiple levels like sub sentence level, sentence level, document level or database level too.

II. RELATED WORK

Balakrishnan et al. [1] proposed an approach in which a centralized block of tweets taken away through the Twitter, preprocessed and later analyzed to irrelevant, positive and negative categories depending on their proportion of emotions. Further, the manuscript also analyses, and collates the production of multiple classifying methods for the motive of analyzing sentiments on Twitter database. This manuscript is directed towards the identifying of admissible classifiers depending on the degree of achievement, which are aimed to partition data as neutral, polar or spam/ unrelated depending on the sentiment indicated by the tweet. Later, the sub-partitioning of the polar class into the negative and positive classes takes place. In this work, the issues with the analyzing of sentiments and that of opinion classification are examined. This approach is totally contrasting from the other approaches used in mining of opinions on structured and elaborated communications.

For the performance of analyzing sentiments, Saeideh et al. [2] gave an easy and effective resolution. At first, the information was gathered for the twitter oeuvre as polar tweets, which included positive and negative data. Later, a descent classifier for sentiments was assembled through the support of the Naïve Bayes algorithm to adjudge and analyze the sentiments of a tweet. Further, the system is adjudged upon the five fields of user tweets, that include finance, sport, jobs, news and movies. To categorize a new tweet for certain fields, the aftermath predicts that it is achievable to use the Twitter oeuvre alone.

The timing constraint of the process was considerably minimized by the use of this system. The use of domain selection approach method on the Naïve Bayes Classifier is much more fruitful than the approach used traditionally, could be proved by this study. An approach is proposed by Liang et al. [3], in which a machine could be automated to analyze the sentiments of short texts like tweets or news. Further, combining this approach with physically labelled tweets' dataset to process
with the mining of opinions. By the aid of filtering away the unopinionated data and later adjudging the opinionated class for their sentiments, whether positive or negative, this technique was assembled in such a way that it automates the system to know exactly how to extract the tweets which contained polar opinions. A mixture of opinion extraction and the supervised learning is assembled in this study to categorize the tweets accordingly.

A way to use the opinion mining to categorize the particular types of emotions is given by Kumar et al. [4]. A bi-step method is being suggested, in which we at first categorize the sentiments, pull the words depicting opinion (a mixture of the adjectives along with the verbs and adverbs) in the dataset of the tweets. Later, by the usage of Novel classifier finding the values of emotions of pulled opinion words. It can be concluded from the foremost outcomes that it is a terrific technique, and will surely have its potential in multiple applications like policy making, business intelligence etc. To evaluate a particular set of emotions like disgust, fear, sadness, anger, happiness etc. in a dataset using opinion mining is suggested in the paper. Linear equation is used to calculate the cumulative value of the emotions.

Movie review can be used to identify sentiments, reach, rating and popularity of a movie. Use of movie reviews as the training and testing set is done in research done by Mertiya et al. [5]. The work suggests a method which coalition the Naïve Bayes algorithm and the adjective analysis to categorize the sentiments of twitter data. Two clusters of tweets (ambiguous and Polarized tweets) is obtained by applying Naïve Bayes on the data set. Further, on the cluster of ambiguous data, we apply adjective analysis in which the cluster is re-divided in two sets of contradictory and non-contradictory by the aid of polarity adjectives and adverbs. They also classified the polarized data into true and false polarized by the aid of naïve Bayes classifier. The falsely contradictory set was additionally processed by the use of adjective analysis to decide the tweet polarity. Thus, finally clustering the polar data into positive/ negative classes. As more and more adjectives and adverbs get added to the oeuvre, the performance of the approach increases simultaneously.

Dengel et al. [6] proposes an exclusive technique for analyzing of informal and short content with the paper mainly focus on tweets. The main focus of this work was on the analyzing of informal statements. It is common to use the emoticons or emojis in the tweets and other forms of communication. The use of these features of sentiments, aids in finding out the text's sentiments. To assign values to features like emojis of sentiments, they used a sentiment feature generator module. Also, these features of sentiments were given more weightage. For the performance of analyzing sentiments, a hybrid method was suggested and this approach is further equated with the three generally used classifiers which are Support vector machines (SVMs), Naïve Bayes Classifier and Maximum entropy. It could be concluded from this work that use of conventional text analysis is a traditional technique with less accuracy, and using the features to analyze sentiments could give higher accuracy and better results.

From the above discussed work, we can say that machine learning algorithm like Naïve Bayes classifier is perfect to use for the classification of sentiments of short text like News. For the aim of testing the accuracies of multiple algorithms of machine learning, almost all the works have uses tools like WEKA (a data mining tool). On the contrast, we aim to check a generalized method to classify and analyze sentiments by the use of hybrid approach using multiple classifiers in a one go to generate better results.

III. METHODOLOGY

Model of machine learning is nothing more than a code snippet with an ability to learn. It is made Intelligent by training with data by a data scientist. So, we will only get garbage results which is false/wrong predictions, if we insert garbage dataset to the training model. The steps involved in the algorithm involve:

Collecting Data: This step is the first and the foundation step of the upcoming learning. The data could be in any form, be it in text files, excel, access or something else. The type of project we opt for decides the type of data to be collected. The source of the dataset could be from any place like sensors, databases, files or multiple sources, although the dataset gathered could not be applied straight forward to perform the sentiment analysis as the dataset could be having multiple unorganized or noisy text data, extremely large values or bunches of missing data. The main aim is to get a data with diversity. By diversity, it means with more volume, density and variety of data for the betterment of the machine learning.

Preprocessing Data: The raw data received is to be cleaned. The process of cleaning it is called data pre-processing i.e. the dataset is transformed to a cleansed dataset from the one gathered from multiple sources. It could also be said that the data collected from multiple sources is in raw format and therefore, it is not possible to proceed with analysis with such data. So, it is necessary to convert this raw data to a clean and small data set with a few steps, this is known as pre-processing.

Pre-processing could be said as a process to clean the raw data in the clean data, to use it to train our model. Therefore, it is a must to pre-process the data in order to achieve better results. The data in the real world is mostly a mess. A few types of these data are:

A. Missing data: It might be possible that due to issues with technique or due to discontinuity, missing data could be there in a dataset.
B. Noisy data: Due to technical issues through the device or by human errors, this type of data might get created. It is also known as outliners.

C. Inconsistent data: Data could be duplicate or might be created due to human errors, this type of data might get accumulated.

Suitable Model: After the pre-processing is done, it is time to check for a suitable match of machine learning model according to the data & type of requirement. It can be chosen from three different kinds of models. Namely, Supervised Learning, Unsupervised Learning or Reinforcement Algorithm.

Model Training: The data cleaned and pre-processed is divided into three clusters namely test data, validation data and the training dataset (proportion depending on the prerequisites). The classifier is trained using the training data set. Later, the parameters are tuned by the use of Validation dataset. Further, to test the accuracy of the model, we use the test dataset. It should be noted that while the classifier is trained, only the training and Validation data is to be used & is available. The test data is meant for future (i.e. for classifier testing) use and shouldn't be used while Classifier's training.

Model Evaluation: Evaluation of Model is done with the aid of Classification Metrics and other parameters like accuracy, precision, F-measure, recall etc.

There are multiple techniques available in this World to do Sentiment Analysis. These could be seen as follows:

A. Naïve Bayes Algorithm (NBC): A part of supervised learning algorithms, Naive Bayes method is established on pertaining theorem of Bayes including a naïve supposition of no relation in all the pairs of features. Although naïve assumptions are used in naïve Bayes algorithms, the classifier has played a remarkable role in daily situations in real world mostly spam filtering and document classification.

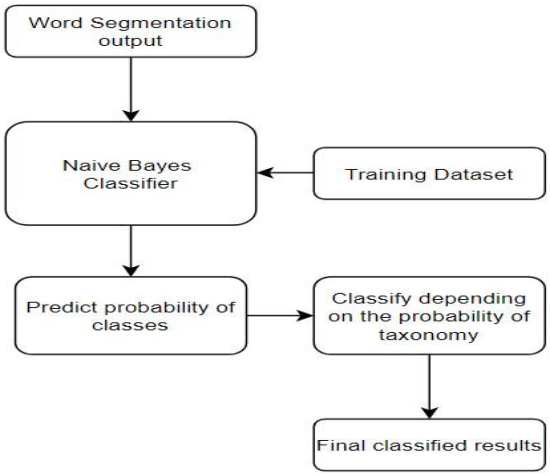

Fig. 2 Naïve Bayes classifier model

B. Support Vector Machine (SVM): A cluster of supervised learning methods which are used in outlier's detection, regression and classification, support vector machines play a crucial role in detecting sentiments.

C. Decision Tree (DTs): A non-parametric supervised learning algorithm which is used in regression and classification. The main aim is to assemble a method which detects the value of the selected variable by the ability to learn easy decision rules through the features of data.

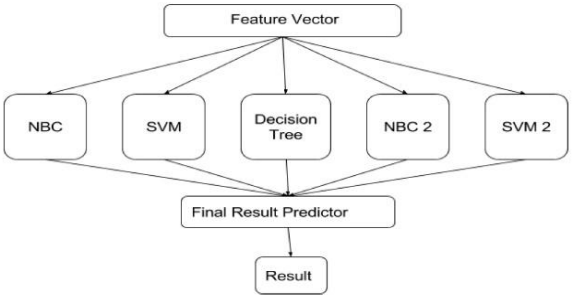

Fig. 1 Sentiment Analysis model

D. Hybrid Approach: With the help of multiple classifiers, the data can be classified easily with high accuracy. Various approaches have been used such as Machine Learning, Information Retrieval, Rule Based Analysis etc. Machine learning approaches like SVM, Naive Bayes, Neural Networks etc. have been the ones with the highest accuracy. However, it is clear from the study that there is strong need to develop a model to better improvise the performance of sentiment analysis by improvising and using modern hybrid optimization techniques. Different approaches, even the ones with not so high an accuracy provide different and important method to find the polarity of a News. Emoticons, Adverbs, Adjectives and various lexical and non-lexical features can be used to analyze the sentiment of a News and therefore, approaches which perform better with each one of these should be collaborated to develop an even powerful model for sentiment analysis.

NBC1

Classification is carried out with the help of Gaussian Analyzer, which is the best to be used here.

SVM1

With the remarkable repo of handling textual data, linear SVC would create aggressive results.

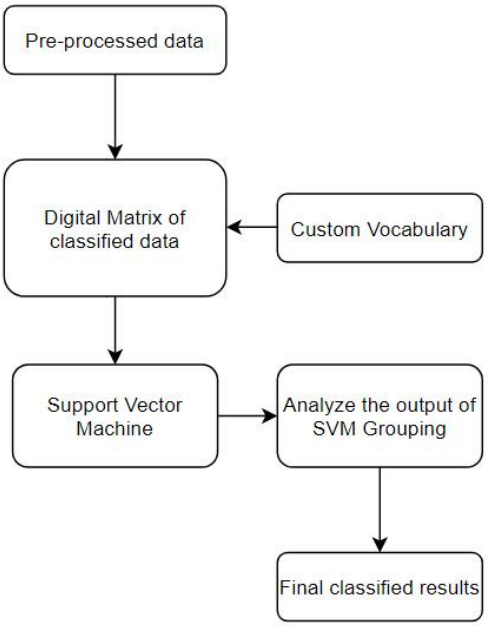

Fig. 3 Support vector machine model

Decision Tree

They could play a great role in the classification of Sentiment Analysis and improve the accuracy to a great extent.

NBC2

In the NBC2 block, we have implemented another Naive Bayes Classifier which works on a different feature set than the previous one. It builds up a vocabulary from all the tweets it initially had. For this block and the next, each tweet is represented as a vector with a dimension equivalent to the length of the vocabulary. This high dimensional feature vector along with an NBC classifier is a more traditional way of performing sentiment analysis.

SVM2

As SVMs are good classifiers for data of sparse nature. So, implementing a second SVM classifier for our analysis as well with the same feature vector as in NBC2.

IV. CALCULATED RESULTS

This section will be discussing the results and calculated performance of the project.

These methods all try to tackle the problem of sentiment analysis in their own unique way and each have their own advantages and disadvantages. In the next block which takes as input, the output of all these blocks.

Table 1. Calculated accuracies of Classifiers.

Classifier	Accuracy
Naïve Bayes	70-80%
Support Vector Machine	75-80%
Decision Trees	73-75%
Hybrid Model	83-88%

This model then tries to combine the benefits of all these blocks and gives a final optimized output. This adds to the model a hybrid optimization technique of performing analysis.

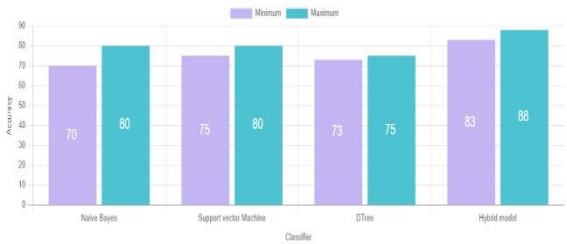

Fig. 4 Performance scores

With the help of hybrid model which is made by combining 2 naïve Bayes classifier, 2 support vector machines and a decision tree, it could be concluded that the calculated accuracy of hybrid model could be nearly 83-88% in a single run.

V. CONCLUSIONS

With the above evaluations and calculated accuracy, it could be concluded that the results of this project were quite a significant improvement on the existing models both in terms of accuracy (by atleast 10% in our tests), removing bias and efficiency. They clearly show the advantage of using Hybrid models over traditional ones. The model developed as a part of this project have been developed while keeping modularity and code reusability in mind. As a result of which it can be further developed and increase in accuracy by simply plugging in more modules or feeding a larger amount of test data.

REFERENCES

1. Priyanthan, Ragavan, Prasath, Perera. “Opinion Mining and Sentiment Analysis on a Twitter Data Stream.” The International Conference on Advances in ICT for Emerging Regions, 2012, pp. 182-188.
2. Saeideh Shahheidari, Hai Dong, Md Nor Ridzuan Bin Daud. “Twitter sentiment mining: A multi domain analysis”, Seventh International Conference on Complex, Intelligent, and Software Intensive Systems, 2013, pp. 144 – 149.
3. Po-Wei Liang, Bi-Ru Dai. “Opinion Mining on Social Media Data”, IEEE 14th International Conference on Mobile Data Management, 2013, Vol. 2 pp. 91 – 96.
4. Akshi Kumar, Prakhar Dogra, Vikrant Dabas. “Emotion Analysis of Twitter using Opinion Mining”, Eighth International Conference on Contemporary Computing (IC3), 2015, pp. 285 – 290.
5. Mohit Mertiya, Ashima Singh. “Combining Naive Bayes and Adjective Analysis for Sentiment Detection on Twitter” International Conference on Inventive Computation Technologies (ICICT), 2016, Vol. 2 pp. 1 – 6.
6. Seyed-Ali Bahrainian, Andreas Dengel. “Sentiment Analysis using Sentiment Features”, IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT), 2013, Vol. 3 pp. 26 – 29.
7. Alec Go, Richa Bhayani, Lei Huang. “FOR ACADEMICS - SENTIMENT140 - A TWITTER SENTIMENT ANALYSIS TOOL”. Internet: http://help.sentiment140.com/for-students.html [Accesesed Online: March. 29, 2017.

AUTHORS PROFILE

Shailja Joshi is currently a Post Graduate student in the Department of Computer Science & Engineering from Rajasthan Technical University, India. She has completed her B.Tech in Computer Science Engineering in year 2016 from Rajasthan Technical University, India. Her research topic is on sentiment analysis of different kind of data using multiple algorithms.

Mayank Patel has completed his Ph.D. in the domain of Multimedia Services over Wireless LAN from College of Technology and Engineering (MPUAT, Udaipur). He is working as an Associate Professor in the Department of Computer Science and Engineering at Geetanjali Institute of Technical Studies, Udaipur, Rajasthan. His area of interest includes Programming in C, Data Structure and Algorithms, OOPS, Programming in JAVA, Web Application Development through J2EE, Web Application Development through .Net Framework, Computer Networks and Wireless Networks, Principles of Programming Languages and many more. He also publishes various textbooks in the programming domain

Manish Tiwari is currently serving as Associate Professor department of Computer Science and Engineering of Geetanjali Institute of Technical Studies, Udaipur. He obtained Ph.D. (Engg) degree from Mewar University in 2019, MTech. from SOIT, RGPV, Bhopal (Govt. University) in Computer Technology and Application in 2009 and B.Tech in Information Technology in 2005. He has about 24 papers in national and international journals and conferences in his credit and 9 filed Indian patents.