Journal of Engineering Sciences
Vol 11, Issue 6,June/ 2020
ISSN NO:0377-9254

SENTIMENT DETECTION ON NEWS DATA USING HYBRID APPROACH

Shailja Joshi[1], Mayank Patel[2]
[1]Masters Scholar, Computer Science and Engineering, Geetanjali Institute of Technical Studies, Rajasthan, India.
[2]Head of Department, Computer Science and Engineering, Geetanjali Institute of Technical Studies, Rajasthan, India.

***Abstract* - In the modern era, an important and significant area of focus in the area of information retrieval and machine learning is Data Classification. Because of the mammoth potential and growth, the area is researched extensively and is being implemented in numerous fields. Hence, this has resulted to numerous algorithms being suggested for this classification. For instance, neural networks, Bayesian classification, genetic classifiers, tree-structured methods etc. These algorithms have a huge base for their applications in multiple fields, like predicting weather, approval of credit, medical diagnosis, segmentation of customers, Spam/ fraud detection etc.**

This research is aimed towards a specific section of classification, which is the Sentiment Analysis. All the potential and ongoing algorithms/classifiers have shown an ecstatic proficiency, and now is the time to incorporate their outcomes to stretch the accuracy and target better performance by the use of hybrid techniques in the area of Sentiment analysis. It is essential to detect the sentiments or the nature of data given by the writer, speaker of the text efficiently to solve the problem of Sentiment Analysis. The data presently available is adequate only to develop a method which takes hybrid techniques like the Naïve-Bayes and also consolidate the available methods to analyze the sentiments on a set of data and examine its outcomes. It can be seen from the outcomes that this approach results in optimistic outcomes.

Index Terms—Naïve Bayes Classifier, Machine learning, Logistic regression, Random forest, Sentiment Analysis

1. INTRODUCTION

Sentiment Analysis is the technique used for detection, extraction and characterization of information in the form of attitudes, opinions or emotions which are existing in the given dataset by using modern techniques like Machine Learning, NLP or stats. The main use of SA is to measure the ideology of the individuals, be it alone or in bulk, like a dataset of news articles. By the use of the points procedure, SA supervises the conversation of speech and dialect reflections to attitudes and also judges the speech and dialects, opinions and emotions required for some business or product at a later stage. Thus, it can be given multiple names like opinion extraction, opinion mining, sentiment mining, subjectivity analysis etc.

Uploading of data over the internet has been growing exponentially which is completely unorganized and it is necessary to systematize it. The data available can be distributed into 3 major types: facts, opinions and experiences, and the further signaling of this data could be bifurcated in different ways. Generally, a pair of mutually opposite keywords is taken, for instance, like-dislike, good-bad, for-against, neutral-positive-negative etc. Similarly, there can even be deeper levels of mutually opposite distributions which could range from being basic to complex. For example, a basic one would be if a fact is true or false, while the advanced one would include checking for multiple emotions in the given set of data. The data classification mentioned above can also be later branched in different ways, like the facts can be classified as true or false, and also in different categories like sports, media, politics etc. So, at the onset, we can say that the data can be distributed in countless different genres. Also, SA could be achieved at different stages, inclusive of database, dataset, file, sentence etc.

In this research work, categories considered include the positive and negative sentiments, and the news data for the observations. In this work, by the use of binary classification, we try to predict whether the news article or the heading is positive or negative.

2. LITRETURE SURVEY

Sentiment Analysis and Opinion mining has been an intensively researched field since roughly a decade now and has recently been reviewed and compared

thoroughly by Soonh Taj in his recent paper "Sentiment Analysis of News articles: A Lexicon based approach" [1]. Many techniques have been suggested to analyze the sentiments of Twitter data by Supun R. Muthutantrige and A.R. Weerasinghe [2] like Support Vector Machines, Logistic regression, Bayes Net and ANN, and also a few ensemble approach algorithms such as Voting classifiers, random forest classifiers, ADA boost, Bagging etc. on a very particular use of SemEval. The previously done research in the field of analyzing the sentiments of Twitter data by Pak and Paroubek[3] used a naive Bayes classifier. Gupta, Kumar and Bhasker [4] used a rough set-based approach in this field. The capability of analyzing the sentiments of the Twitter dataset is studied as well by A. Tumasjan and Philipp Sandner[5] and by Shrikanth Narayanan[6]. In both these works, the contributors used the twitter data to survey the opinion of public in politics, and by using keyword-based characteristics to obtain the sentiments of public during elections. Both the studies got quite an accurate prediction of the election results which emphasizes the importance of mining sentiment of microblog data. A recent research work by Aman Ahuja [7] uses both words and hashtags to provide a topic-based sentiment analysis model MSTM on microblog data which outperformed many existing techniques.

There have been several advancements made by hybrid optimization Techniques recently. They prove to be quite useful for handling machine learning problems as well. Trying to solve sentiment analysis problems using these techniques also gives inspiring results. Work done in this field by Han [8] focused mainly on features present in Chinese microblogs. Xie [9] suggested a grading structure-based mixed and combined method to analyze the sentiments, and also analyzed the participation of multiple traits in the SVM. Both these approaches performed better than the techniques which used traditional methods and indicate the hidden potential of hybrid approaches for improving the Sentiment prediction on microblog data.

Some of the key works are discussed in detail below:

2.1 Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis [10]

This is a very detailed study of different text preprocessing methods which have been involved in the sentiment analysis of microblogs. It compares and suggests the best practices which should be followed to get the highest accuracy and speed while performing sentiment analysis of microblogs. It uses four different methods to compare the sensitivity of different text preprocessing methods: Naive Bayes, SVM, Random Forest and Logistic Regression.

2.2 Combining Naive Bayes and Adjective Analysis for Sentiment Detection on Twitter [11]

In this work, the authors have used a movie review-based dataset of tweets and have used Naive Bayes method of ML to find the polarity of a tweet. However, this model may produce false positives and false negatives, and to deal with this they have used adjective and adverb polarities in the tweet to get the final sentiment. They achieved a final accuracy of 88.5% over their dataset which was higher than the SVM or Markov Model.

2.3 Microblog sentiment analysis algorithm research and implementation based on classification [12]

Here the authors built Naive Bayes and SVM classifier models for sentiment analysis of mainly Chinese microblogs and compared their performance.

3 PROPOSED APPROACH

A basic code snippet which can learn could be considered as machine learning algorithm. By training the program or the snippet by a data scientist fed with the data, the code is made smarter. So, it is simple give and take. If the code is trained with garbage data, only garbage outcome could be expected, which would be inaccurate predictions. The flow of the algorithm goes as follows:

3.1 Collecting Data:

Collecting data is considered as the base step for the learning as the data decides how the algorithm would perform. The data could be collected in multiple forms depending on the requirement of the project, like collecting it via access, excel, documents or text files etc. and it could be collected from multiple sources like files, databases, websites, surveys, sensors and many others. Although, the data collected couldn't directly be used for the framework as it would be including bundles of missing data, exponentially high values, noisy or unorganized text data. The goal is to obtain diversified data which means data with good amount of variety, density and volume. More diversified is the data, more accurate would be the prediction by the algorithm.

3.2 Preprocessing Data:

The data obtained needs to be cleaned for errors. Pre-processing the data is the term used for the process of cleaning the data, i.e., the transformation of the obtained dataset from multiple sources to a cleansed data-set. As for the betterment of the model, the raw format of data should not be used, which makes it a compulsion to process the obtained raw data into a smaller and cleaner data-set with some steps, and that is what is referred to as pre-processing.

To achieve tangible results, the data needs to be pre-processed as the data available in reality is always messy. Types of data which cannot be used for obtaining good results are:

3.2.1. Missing data: By discontinuous data or by issues in the ongoing techniques, there are chances of receiving missing data in the obtained data set.

3.2.2. Noisy data: Also known by the name of outliners. This type of data could be created by issues in techniques, devices or human errors.

3.2.3. Inconsistent data: Gathering of data from multiple sources or because of human errors, duplicate data could be accumulated in the dataset.

In this research work, a ready to use dataset is taken from GitHub and used, which already has labelled data as per the suitable headlines

3.3 Suitable Model:

Once the collection and pre-processing of the data is done, X there could be a need to probe for a machine learning algorithm which would fit the best depending on the requirement. The choice for this could be made from multiple algorithms like reinforcement algorithms, supervised or unsupervised algorithms. As the aim of this research is to implement the sentiment detection in real-time, multiple classifiers like Multinomial NB, Bernoulli NB, Random forests, logistic regression classifiers and their combined results are used to predict the final outcome.

3.4 Model Training:

By this stage, the collection and pre-processing of data would have been done. Also, a suitable model would have been chosen. Here, the data is distributed into three clusters, which are training, validation and the test dataset, the size of which is governed by the pre-requisites. The training of the classifier is done using the trained dataset. Further, the parameters and model tuning are done by the use of validation block. At last, the test data is used to check if the model is giving considerable results and could be released. In general, the test data is not at all used in the training of model and is solely meant for future testing. In this research, the use of k-fold Cross validation algorithm is performed to create subsets of the dataset and grouping them in three subsets of training, validation and test subset.

3.5 Model Evaluation:

Finally, it is time to check out the performance of the model. Depending on various factors which could help in proving the real-time impacts of the system, the evaluation metrics used are accuracy, F1 scores and confusion metrics.

4 IMPLEMENTATION

Depending on the data present in the training dataset, we detect the class of an unseen and new observation, known as classification. With the help of classification, the data is separated in multiple groups which are discrete. With the help of a few parameters provided as the input, the dataset is classified under various labels and this is how we give labels to the data.

Imbalanced Dataset:

To work with a dataset which is imbalanced is the most common problem faced in machine learning. In the dataset of news, there are more negatives than the positives which make it an imbalanced dataset.

Some issues, like detection of fraud, the news dataset is not imbalanced. There could be times when the datasets could contain only a fraction of 1% of positives and rest of them as negatives.

Working with an imbalanced dataset, there is a need to be more cautious. We may get accuracy close to 90% while using this classifier, which is generally referred to as the Accuracy Paradox.

The reason being that our model would be analyzing the dataset and showing it to be always negative. This would ensure it would higher accuracy, of nearly 90% towards the negative, hence resulting in higher accuracy of nearly 90%.

This problem could be resolved with the help of various methods, such as:

- Collection of more data: This could be very helpful by adding multiple minor class examples and balancing the dataset.

- Change of metric: Various measures could be used to predict the ability of the system, like the recall, F1 score, precision, or the Confusion Matrix.
- Oversampling of data: To help creating extra 'fake' data, there is a need to arbitrarily sample the attributes from examples in the minority class.
- Penalized model: It gives an extra cost on the given model for classifying mistakes while training. Minority class are biased by the model through these penalties.
- In the dataset used, there are a few positive examples compared to the negative ones, so will explore both the different metrics altogether by utilizing an oversampling technique known as SMOTE.

Dataset used: The dataset used in the project is taken from the GitHub open source which is used by the user Rahulkg007 in their research on media broadcasting.

Further, the data should be checked through Cross validation to form multiple sets of training and test data to be passed and verified on the training model.

4.1 The Multinomial Naive Bayes' Classifier

An exclusive version of the Naïve-Bayes algorithm, Multinomial Naïve-Bayes is sketched specifically to check the documents with text. When a basic Naïve-Bayes could highlight a doc to mark the existence and non-existence of specific words, multinomial version of NB model explicitly models the number of word poll and regulates the elementary computation to tackle within. Classification of text is the main mechanism behind the analysis of sentiments.

Naïve Bayes classifier is a division of the supervised learning algorithms which is made on the foundation of Bayes Theorem, which is inclusive of simple and wide assumptions. A considerable role is played by this classifier in regular real time situations, which includes classification of documents and filtering of spams.

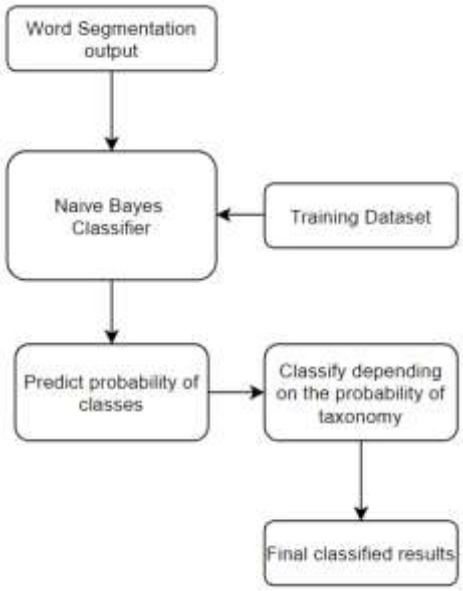


Fig. 1 Naïve Bayes classifier model

4.2 Bernoulli Naïve Bayes Classifier

The only part which separates Bernoulli naïve Bayes and multinomial naïve Bayes is that the Bernoulli classifier takes only binary values. In our work, we have only 2 values which include whether or not a word is present in a document, which is a very untangled model. Nevertheless, it could be concluded that Bernoulli Naïve Bayes is more efficient when the word frequency is not essential.

4.3 Random Forests Classifier

A type of supervised learning algorithm, Random Forests could be used for both classification and regression. It is one of the simplest algorithms to be used. A bundle of trees makes a forest. For the robustness of a forest, there should be more trees. Decision trees are created from Random Forests on anonymously chosen data specimen by getting probability from each tree and further choosing the best possible resolution by the use of voting. It also acts as a pretty good indicator.

4.4 Logistic Regression Algorithm

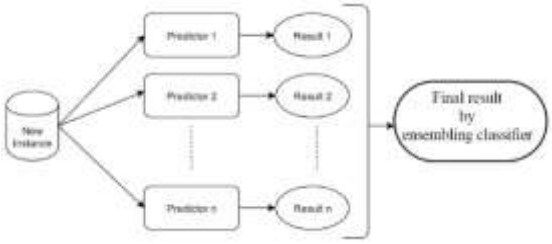
A classification algorithm which is utilized to allocate observations to a discrete dataset, Logistic regression plays an important role in detection of whether an online transaction is fraudulent or not, Email is spam or not, checking of tumor in Human body etc. With the aid of logistic sigmoid function, Logistic regression changes its output to produce a probability value.

4.5 Ensemble Vote Classifier

A meta-classifier to combine similar or conceptually contrasting machine learning classifiers for categorizing via seniority or plurality voting, the Ensemble Vote Classifier executes both the "hard" and "soft" voting.

Fig. 2 Ensembling Classifier

The final label is taken as the class label which is



predicted mostly by the various classifiers, and that is hard voting. While, when the final class label is taken by calculating the average of the class probabilities, it is soft voting.

The implementation of 5 methods (Multinomial NB, Bernoulli NB, Logistic Regression, Random Forest and Ensemble Voting) is done with the help of pre-defined functions in Python 3.7 due to which it takes a bit more time in the execution. However, the execution time can be easily lowered by manually making targeted functions and going through the execution.

5 RESULTS

The success of the experiment can be seen by the F1 score and Accuracy values determined with the help of confusion matrix. Confusion Matrix acts as an algorithm to check the performance of machine learning classification problem with output being of 2 or more classes.

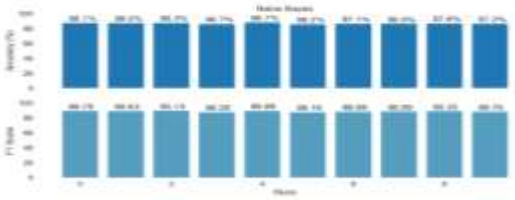


Fig. 3 Accuracy – F1 Score plot

4 parameters which are used for a confusion matrix are 'True positives', 'True Negatives', 'False Positives' and 'False Negative'. We require significant values in the True negatives and true positives to get a more accurate model. Depending on the number of classes, size of the Confusion Matrix is chosen.

5.1 Accuracy

It tells us the percentage of correct predictions to the total ones.

Accuracy = (True Positives + True Negatives)/ (Total number of Classes)

Average Accuracies for multiple classifiers are:

Table I: Accuracies of Classifiers

Classifier	Avg. Accuracy
Multinomial NB	87.36
Bernoulli NB	87.12
Logistic Regression	89.70
Random Forest	88.10
Ensemble Voting	90.93

5.2 F-Measure

Comparison of two models with high recall & low precision or vice versa is a very tedious task.

F-Measure = (2* Recall * Precision)/ (Recall + Precision)

Table 2: F-Scores of Classifiers

Classifier	Avg. F Score
Multinomial NB	89.10
Bernoulli NB	88.82
Logistic Regression	90.66
Random Forest	88.59
Ensemble Voting	92.02

So, F-Measure is used to differentiate them. F-Measure allows us to get the measures with precision

& recall them at the same time by the use of harmonic mean instead of Arithmetic Mean.

6 CONCLUSION

The results received from the multinomial Naïve Bayes and other classifiers did not differ much from each other. It can be seen from the combined result of the 4 classifiers, the ensembling voting technique gave the best results and the highest accuracy.

7 REFERENCES

[1] S. Taj, B Shaikh and A. Meghji, Sentiment Analysis of News Articles: A Lexicon based approach", 2019 Internaltional Conference of Computing Mathematics and Engineering Technologies- iCoMET 2019

[2] Supun R. Muthutantrige and A.R. Weerasinghe, "Sentiment Analysis in Twitter messages using constrained and unconstrained data categories." in IDSCE, 2016.

[3] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." In LREC, vol. 10, pp. 1320–1326, 2010.

[4] Mukul Gupta, Pradeep Kumar and Bharat Bhasker, "Clustering of users on microblogging social media: A rough set-based approach" in ICDSE, 2016.

[5] A. Tumasjan, T.O.Sprenger, P.G.Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment." ICWSM, vol. 10, pp. 178–185.

[6] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 US presidential election cycle," in Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012, pp. 115–120

[7] Aman Ahuja, Wei Wei and Kathleen M. Carley"Microblog Sentiment Topic Model." In ICDMW.2016.110 pp. 103 -1038.

[8] Ping Han; Shan Li and Yunfei Jia, "A Topic-Independent Hybrid Approach for Sentiment Analysis of Chinese Microblog" in ICIRI, vol. 17, pp 463 - 468, 2016.

[9] L. Xie, M. Zhou and M. Sun, "Hierarchical structure-based hybrid approach to sentiment analysis of chinese micro blog and its feature extraction," in Journal of Chinese Information Processing, vol. 26, no. 1, pp. 73-83, 2012

[10] Zhao Jianqiang and GuiXiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis" in IEEE Access 2016, vol. PP issue 99

[11] Mohit Mertiyaand Ashima Singh, "Combining naive bayes and adjective analysis for sentiment detection on Twitter" in 2016 International Conference on Inventive Computation Technologies (ICICT), vol. 2, 2016.

[12] Yanxia Yang and Fengli Zhou, "Microblog sentiment analysis algorithm research and implementation based on classification. " in DCABES.2015.79, pp. 288-291.

[13] Mondher Bouazizi; TomoakiOhtsuki, "Sentiment Analysis in Twitter: From Classification to Quantification of Sentiments within Tweets", in IEEE Global Communications Conference (GLOBECOM), year 2016.

[14] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani," SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining ", Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10).