# *Sentimental Analysis using ensemble of Naive Bayesian, Logistic Regression and Decision Tree Algorithm*

**Pratik Kumar Dutta**

Master of Computer Application
Vellore Institute of Technology
University in Vellore
Tamil Nadu, India.

**Sabyasachi Chakraborty**

Master of Computer Application
Vellore Institute of Technology
University in Vellore
Tamil Nadu, India

**Sahil Gandhe**

Master of Computer Application
Vellore Institute of Technology
University in Vellore
Tamil Nadu, India.

Under the guidance of

**Dr. Sumaiya Thaseen I**
(Associate Professor)
School of Information Technology and Engineering
Vellore Institute of Technology
University in Vellore
Tamil Nadu, India.

***Abstract***

***This paper describes that sentimental analysis is performed using an ensemble of Naive Bayesian Classification, Logistic Regression and Decision Tree Algorithm. In this technique, team has calculated and given the result of the polarity for the provided terms and data. System is able to get prediction automatically with this approach and approximate outcome for a big division of sentimental analyzed data and expressions by deploying efficient algorithms that are significantly better and thereby achieving results in minimal span of time.***

***Keywords-Sentimental Analysis learning, machine learning, Algorithms merging***

## I. INTRODUCTION

In current days Internet and its power to connect with the whole world in just few moments has effected and changed the lifestyle and views of thinking of people. People can communicate and express how they feel or their valuable opinion through blogs, online social media and many other survey related websites where people can share and perform their opinions, votes etc. Nowadays, Facebook or Twitter, Instagram, WhatsApp and many other social media applications are used by millions of people to share their opinions and their way of views, comments and about their daily lifestyle activity. Nowadays from online forum and the social media communities, people get interesting media contents where clients notify and inspire others through these social media forums. Online media or social media sites are creating a huge volume of supposition of rich information as tweets, notices, blog entries, remarks, audits, and so forth. Besides, online media or social media gives a chance to organizations by giving a stage to associate with their clients for promotion. Individuals generally rely on client created content over online, as it were, for basic leadership. E.g. Nowadays people are going to buy product from online shopping portal only by the reviews from social media site also they ask for feedbacks for newly launched products before confirming their transaction. The measure of substance created by clients is excessively immense for an ordinary user, making it impossible to break down. Along these lines, through technology people need to computerize online environment as sentimental analysis procedures and techniques are broadly used and utilized.

Procedure and techniques of sentimental analysis is the process of recognizing and sorting conclusions communicated in a piece of content, particularly with the end goal to decide if the, item is sure, negative or unbiased. Different application like emotion detection and Crime identification, Age prediction, Selection and Multilayer Perception can be done using: 1. Naive Bayesian Classifier, 2. Logistic Regression, 3. Decision Tree Algorithms.

Information of text and its recovery systems mostly center around handling, sorting or sequentially arranging also examining and executing down the truthful information present in it. Details have a target segment in any case, there are some other text encrypted substance which express abstract attributes. These substances are for the most part conclusions, estimations, examinations, dispositions, and emotions which shape the center of Sentiment Analysis (SA). It offers many testing chances to grow new applications, for the most part because of the tremendous development of accessible data on online sources like web journals and interpersonal organizations, Internet social media sites. For instance, proposals of components proposed by a suggested system can be anticipated by considering contemplation, for an example, positive or negative feelings about those things by making utilization of Sentimental Analysis.

***Sentimental study and process analysis***

The arrangement while performing Sentimental study and process investigation on tweets is essentially to order the tweets in various sentiment categorized classes and diverse emotional classes precisely and independently. In this ground of research, different methodologies have developed, which propose techniques to prepare a model and after that test it to check its proficiency. As mentioned, main test is to executing emotion study and process analysis on Twitter dataset here we define the reasons for this:

(1) **Restricted tweet amount:** Restricted amount of data defined and with only 150 characters close by, minimized articulations are produced, which results sparse arrangement of highlights.

(2) **Usage of abusive words:** Few words in the data set are not the same as English words and it can make a methodology obsolete as a result of the transformative utilization of abusive language and terms.

(3) **Twitter type:** This part define that it can permit the utilization of hash labels, client reference and URLs. These require unexpected preparing in comparison to different words.

(4) **Consumer range:** Clients express their suppositions in an assortment of ways, some utilizing diverse language in the middle, while others utilizing rehashed words or images to pass on a feelings or sentiments.
Each one of these issues are required to be looked in the pre-processing area. Apart from these, we face problems in feature extraction with fewer features in hand and reducing the dimensionality of features.

## II. LITERATURE REVIEW

a) The authors [1] have implemented to learn Naive Bayesian Classifier in support of implementing Rapid task of rRNA series. By the test we get that exactness level is 95%. RDP (The Ribosomal Database Project) Classifier which gives information, devices, and administrations identified with rRNA series to the research network was tested with a corpus (a collection of written text) of 23,095 RNA sequences. The authors [2] has given the outcomes from leave-one-out testing on the two corpora demonstrate that the general correctness in every stages of certainty in favour of close and full length and 400-base portions were 89% or more behind to the sort stage, also most of order mistakes show up for anomalies in the current taxonomies

b) Naive Bayesian Classifier filters are implemented by authors [3] on two datasets namely Spam and SPAMBASE and their performance was tested. The different metrics analyzed are accuracy, recall, precision and F-measure. The authors utilized WEKA for the evaluation of email spam filtering using Naive Bayesian classifier algorithm. The results illustrate the type of email and number of samples of the dataset manipulates the performance of Naive Bayesian classifier.

c) The authors [4] have implemented Logistic Regression of Spam Recognition. Spamming is the misuse of Email to send unwanted bulk message described by author [5]. It is a serious problem for an organization and different email users. Web threats like hacking, internet worms which directly damage our information, spam can damage users information indirectly. It may increase server load, decreases network performance.

d) The Authors [6] researched and provided with results of Age Prediction from Text using Linear Regression. In this paper Team framed the creator's age estimate as of content as a regression issue. We find a similar undertaking utilizing three altogether different sources: web journals, phone discussions, and online gathering posts. We utilize a strategy from area adjustment that offer us to prepare a joint model including every one of the three corpora together and in addition particularly and examine contrasts in prescient highlights crosswise over joint and corpus particular parts of the model. Compelling highlights incorporate both complex ones, (for example, POS designs) and additionally content situated ones. Utilizing linear regression demonstrate dependent on shallow text highlights, we acquire connections up to 0.74and mean supreme blunders among 4.1 and 6.8years.

e) The Authors [7] has researched and implemented Decision Tree Algorithm Feature selection and Multilayer Perceptron for sentimental study. A decision tree implementation picks out significant features. Decision tree induction creates a tree structure with internal nodes signifying a quality test. Internal nodes denotes branch representing test result and external node denotes class prediction. IMDB (Internet Movies Database) dataset is used to calculate the proposed method. Results showed that the Multilayer Perceptron (MLP) with planned feature selection improves the performance of MLP. Authors came out with significant results by this test.

## III. BACKGROUND

### *A. Naive Bayesian classifier*

We can say Naive Bayesian classifiers defines that this classifiers are considering the characterization job from a Statistical perspective. The beginning stage is that the likelihood of a class is given by the back likelihood or probability given a preparation

report. Here alludes to the majority of the content in the whole preparing set. Here is given by , where is the aspect (word) of record .

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \text{ ----------[1]}$$

*P= probability of two objects*
*A= first object*
*B= Second object*

### B. *Logistic Regression*

Spamming is the misuse of Email to send unwanted bulk message described by author [1]. It is a serious problem for a organization and different email users due to the increasing population and low cost of electronic mail. Unlike other web threats like hacking, internet worms which directly damage our information, spam also can damage our information indirectly. It may increase server load, decrease network performance. In this case we will see how team [1] can detect and prevent spamming.

$$P = \frac{e^{a+bX}}{1+e^{a+bX}} \text{ ---------- [2]}$$

*P=probability of Positive Occurrence*
*a,b= Constant*
*e=exponential*

### C. *Decision Tree*

A Decision Tree-based Feature Ranking is planned for feature selection. A decision tree induction picks out significant features. Decision tree induction creates a tree structure with internal nodes signifying a quality test with the branch representing test result and external node denotes class prediction. In this application, a hybrid algorithm based on Differential Evolution (DE)[1] and Genetic Algorithm (GA) for weight optimization algorithm to improve MLPNN is scheduled. IMDB dataset is used to calculate the proposed method. Experimental results showed that the MLP with planned feature selection improves the performance of MLP significantly.

$$Entropy(S) = p_+(-\log_2 p_+) + p_-(-\log_2 p_-) = -p_+\log_2 p_+ - p_-\log_2 p_- \text{ ---------- [3]}$$

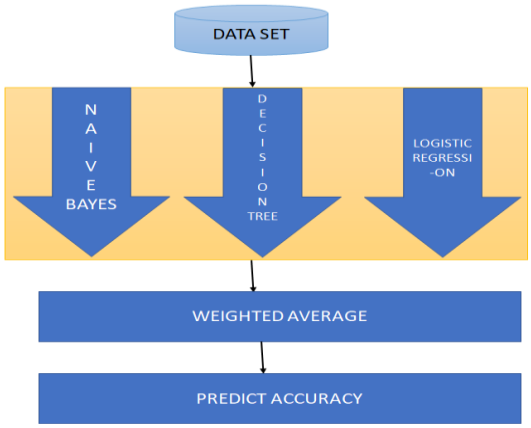
*P=probability of positive occurrence*

Can be generalized to more than two values

## IV. PROPOSED MODEL

- The model is an ensemble model of Naive Bayesian, Decision Tree and logistic regression.
- All the three classifiers are supervised which requires the class label for training and testing the data.
- The advantage of using ensemble model is to improve the accuracy prediction.
- Weighted average is used to predict the final accuracy from all the classifiers.

### A. *Architectural diagram of ensemble algorithm -*




### B. *Performance Metrics –*

- **Accuracy:** Result of calculation or specification conforms to the correct value or a standard of three algorithms. Accuracy is used as a metric for evaluating the usefulness of one classifier.

**Accuracy = TP+TN/TP+FP+FN+TN** ---------- *[4]*

- **Precision:** Refinement in three algorithms for calculation, or specification, especially as represented by the number of digits given. By definition we can elaborate that Precision can estimate the accuracy of one classifier. The higher precision goes, that implies less false positives. On other hand a lower precision implies all the high false positives. This shows regularly inconsistent comparing with recall part, as a simple method to enhance precision is to diminish recall.

**Precision = TP / TP+FP** ---------- *[5]*

- **Recall:** Recall section defines that deals with the culmination, or affect ability, of a classifier and measures for output. Higher recall implies less false negative, while lower recall review implies all the more false negatives. Enhancing recall can regularly diminish accuracy since it persuades progressively harder to be exact as the demo space increments.

**Recall = TP / TP+FN ----------** *[6]*

## V. RESULTS

*Individual outputs of three algorithms along with the Polarity of Pacifiers and the final Accuracy of three Pacifiers together:*

- *Naive Bayesian Classifier –*

```
BernoulliNB(alpha=1.0, binarize=0.1, class_prior=None, fit_prior=True)
The Accuracy using Naive Bayes is:  0.7027027027027027
[[28  0]
```

- *Logistic Regression –*

```
             precision    recall  f1-score   support

          0       0.72      1.00      0.84        23
          1       0.00      0.00      0.00         9

avg / total       0.52      0.72      0.60        32

[[23  0]
 [ 9  0]]
The Accuracy of the prediction using Logistic Regression is:  0.71875
```

- *Decision Tree –*

```
          0       0.67      1.00      0.80        28
          1       0.00      0.00      0.00        14

 avg / total       0.44      0.67      0.53        42

 The Accuracy of the prediction using Decision Tree is:  0.6666666666666666
```

- *Polarity –*

```
RT @EricTrump: Sean Hannity: If Hillary wins, you own it https://t.co/kQCwK3ttjK -------->> positive
RT @DonaldJTrumpJr: Thanks New Hampshire!!!
#NH #NewHampshire #MAGA https://t.co/JDqcyJvJpk -------->> positive
RT @detroitnews: .@IvankaTrump in Michigan: This is your movementÛª https://t.co/0Sa7huoOPl @realDonaldT
Unbelievable evening in New Hampshire - THANK YOU! Flying to Grand Rapids, Michigan now.
Watch NH rally here: https://t.co/hF88anrfgk -------->> positive
Big news to share in New Hampshire tonight! Polls looking great! See you soon. -------->> positive
```

- *Final Accuracy –*

```
After getting 3 Algorithm We Can estmate the Final Accuraccy:  0.6960397897897898

Process finished with exit code 0
```

## VI. CONCLUSION

In this paper, Sentimental Analysis is performed by collecting the tweets and individual pacifiers such as Naïve Bayesian Classifier, Decision Tree, and Logistic Regression. The individual three classifier accuracies of 70% for Naive Bayesian Classifier, 71% for Logistic Regression and 66% for Decision Tree Algorithm are obtained. Final accuracy result of 69% is obtain after merging three classifiers together however in the future, an ensemble of Naive Bayesian Classifier, Decision Tree and Logistic Regression is to be deployed for better prediction.

## VII. REFERENCES


[1] Wang, Qiong, et al. "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." *Applied and environmental microbiology* 73.16 (2007): 5261-5267.

[2] Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, *73*(16), 5261-5267.

[3] Jindal, Nitin, and Bing Liu. "Opinion spam and analysis." *Proceedings of the 2008 international conference on web search and data mining*. ACM, 2008.

[4] Jindal, Nitin, and Bing Liu. "Opinion spam and analysis." In *Proceedings of the 2008 international conference on web search and data mining*, pp. 219-230. ACM, 2008.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[5] Jindal N, Liu B. Opinion spam and analysis. InProceedings of the 2008 international conference on web search and data mining 2008 Feb 11 (pp. 219-230). ACM.

[6] Nguyen, Dong, Noah A. Smith, and Carolyn P. Rosé. "Author age prediction from text using linear regression." *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Association for Computational Linguistics, 2011.

[7] Jotheeswaran, Jeevanandam, and S. Koteeswaran. "Feature selection using random forest method for sentiment analysis." *Indian Journal of Science and Technology* 9.3 (2016).