RESEARCH ARTICLE OPEN ACCESS

Sentiment Analysis Using Ensemble Learners and Gini Index

Furqan Iqbal
Computer Science and Engineering, Lovely professional University

Abstract:

Sentiment analysis is a part of natural language processing for evaluation of opinions from data. In this paper sentiment analysis has been performed on movie reviews. This paper explores the sentiment evaluation using gini index feature selection method. Furthermore the ensemble learners method has been used for classification. Finally the results have been obtained from the confusion matrix.

Keywords — **opinion mining, gini index, ensemble learners, classification.**

I. INTRODUCTION

Since the explosion of digital era has taken place a huge amount of data has been collected. This data can bein many forms like audio, video and text. This data can be used for analysis of a huge number of things such as E-commerce, banking, healthcare and even prevention and detection of fraudulent activities. A large amount of this data is raw and it cannot be used to analyse or make any significant contribution. To use this elephantine amount of data analysis has to be performed so that useful information can be obtained from the data. Many billion dollar industries in world are using the data collected from internet users to make money. So from a financial point of view to obtain and analyse data and get results from it can improve profits and market value of any industry. Many companies also sell user data to other companies and making huge profits in the transaction.

A. Sentiment Analysis

Sentiment analysis or opinion mining is a term used when human emotions have to be obtained from data. When sentiment analysis is done the polarity of the data is checked. This polarity can be positive, negative or neutral based upon the opinion of the user when the data was recorded. Sentiment analysis helps in determining the emotion a user projects towards a particular targeted entity. Sentiment analysis can has become a very important factor in today's world because a lot of transactions are made online. Since users can write their views or opinions
about a particular subject, the data collected from their feedback can be analysed and relevant results can be extracted which evaluates opinions of users towards a particular subject. This can be an opportunity for different industries to capitalize on this information to improve their profits, reputation or brand. Many users write their feedback after shopping online which the other users can use as a reference for the quality of the goods and services available online. So sentiment analysis can be used if any company wants to increase its customer base and improve its products and services by taking advice directly from its customers.

B. Factors to be considered for sentiment analysis

One of the biggest concerns in sentiment analysis is language itself. Since there are many languages in the world sentiment analysis have to be done using tools developed to work for that certain language. If the data is translated into other language like English for which a variety of sentiment analysis tools have already been created, it might result in a lot of error since a large number of words lose their pristine when translated. Sarcasm is another thing to be concerning sentiment analysis, since a witty sentence can mean a very different thing than what might be perceived. If analysed improperly the results will have errors. The algorithms used in sentiment analysis can show different results for different data sets. This happens because there is no algorithm which is perfect. Algorithms have to be chosen which gives optimum results. Today not only words show emotions but also emoticons are there to express emotions.

Emoticons also have to be analysed for sentiment information.

C. Applications of sentiment analysis

1) ***Buying goods online:***When we buy anything from an e-commerce site whether it is a Mobile phones or cloths, the previous consumers who have already bought those things can post their review online. This review can be on different aspects of the product that the user wants to buy. Sentiment mining can help the upcoming customer to evaluate the reviews of the customers so that in future all the good and bad qualities of the product is well understood by the analysis of those reviews.

2) ***Betterment of goods and services:***Using sentiment mining companies can get direct feedback from the users and they can evaluate this feedback so that they can improve their services and products.

3) ***Politics:***With the help of sentiment mining the politicians can take the opinion of citizens into consideration. The election process largely depends on the sentiments of people. If the data is used to evaluate these sentiments and analyse the public perception and hence political parties can use it to make better political strategies.

4) ***Detect crime:***Many frauds and spurious claims take place on the internet. These crimes include stealing credit cards and giving spam emails and reviews. Sentiment analysis can be used to detect them.

5) ***Movie review:***Many people read a movie review before going to watch the movie. These reviews can be used to analyse the sentiments that are displayed by the reviewers after watching the movie.

II. LITERATURE SURVEY

Among the various methods for sentiment Cambria, Schuller, Xia and Havasi [1] have analyzed different techniques which are currently being applied in sentiment analysis. They discussed how research in sentiment analysis is becoming fine grained. Valdivia, Luzón, and Herrera [2] have used trip advisor to analyze the applications of sentiment analysis. Geetha, Singha and Sinha [3] discussed the connection between customer sentiments and ranking received on the internet. Akhtar, Gupta, Ekbal and Bhattacharyya [4] have discussed and given an insight into aspect level analysis. If an individual algorithm is used it is very quick to train but using multiple algorithms can increase accuracy. Catal and Nangir [5] have used an approach of using multiple algorithms for better results. Similarly Araque, Corcuera-Platas, Sánchez-Rada and Iglesias [6] used multiple algorithms in an ensemble learning fashion for better results. Among many algorithms used Nigam, Lafferty and McCullum [7] proposes the use of Maximum Entropy algorithm. Maximum Entropy can give better results than Naïve Bayes but sometimes the results are even worse than naïve Bayes. In sentiment analysis every word hold accountability for the performance. Tripathy, Aggarwal and Rath [8]propose the use of N-grams. This takes more than one word into consideration for analysis. Pang, Lee and Vaithyanathan [9] analyzed the machine learning techniques. They proposed to use Naïve Bayes, SVM and Maximum Entropy. They compared under the umbrella of features selected on the basis of n-grams. There are different ways which help in understanding how features for opinion mining are selected. Manek, Shenoy, Mohan and Venugopal [10]took Movie Reviews in consideration. Here gini index technique was used for selecting possibly beneficiary features.

III. RESEARCH METHODOLOGY

The sentiment analysis is done by using following steps.

A. Data Source selection:

The dataset is selected for sentiment analysis. There are different kinds of datasets available online and sentiment information can be retrieved from them.

B. Data pre-processing:

The pre-processing phase has many steps which have to be taken in order to make data ready for classification process. The dataset goes through tokenization process where data is divided into tokens. The data also has many words which are not helpful for the analysis. These words have to be removed by using a stop word list. The dataset also goes through stemming process where the words in the dataset are reduces to some root form. The bi-grams from the words are used to select more than one word for the process.

C. Representation of features:

$$\text{GiniIndex}(x) = 1 - \sum_{i=1}^{m} P_i^2 \quad (1)$$

$$\text{GiniIndex}(x) = \sum_{i=1}^{m} P_i^2 \quad (2)$$

Gini Index based feature selection method. The gini index is used to select the features which will be helpful in a better analysis of data. If the value of gini index is nearing to zero, it means that most helpful data is being acquired. On the other hand if a high value of gini index is obtained it shows the lack of quality of data and in this case least helpful data is being acquired.

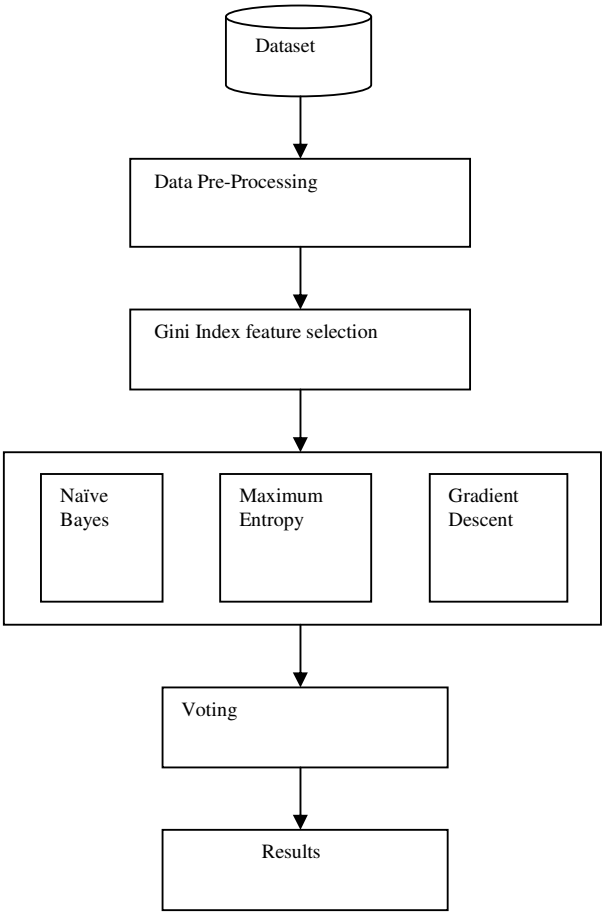

Fig.1. Proposed Methodology

*D. **Classification:***

In this process the ensemble learner's technique is used on the filtered data. This is done using Naïve Bayes, Maximum Entropy and Gradient Descent algorithms. There is no algorithm out there which can give 100% accuracy. Every algorithm has its shortcoming due to the inherent nature of how these algorithms work. Ensemble learners are used to reduce these shortcoming by using more than one algorithm. In the end voting is done to choose the best performance.

IV. RESULTS

To analyze the results that have been obtained by using a gini index based feature selection a confusion matrix is used. The study compares the results of the proposed design to the existing based on Accuracy, Precision, Recall, F-measure.

Table 1: Comparison of Various Techniques based on Correctly Classified Instances

Technique	Correctly Classified Instances (%)
Gradient Descent	70.10%
Maximum Entropy	69.44%
Naïve Bayes	65.24%
Proposed Technique	71.87%

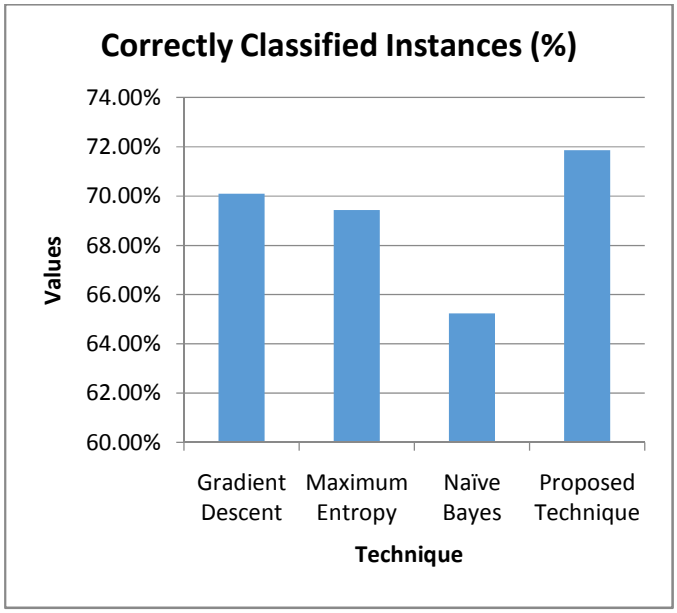

Fig.2. Showing Comparison of Various Techniques on the basis of Correctly Classified Instances

Table 2: Comparison of Various Techniques based on TP Rate

Technique	TP Rate
Gradient Descent	0.701
Maximum Entropy	0.694
Naïve Bayes	0.652
Proposed Technique	0.719

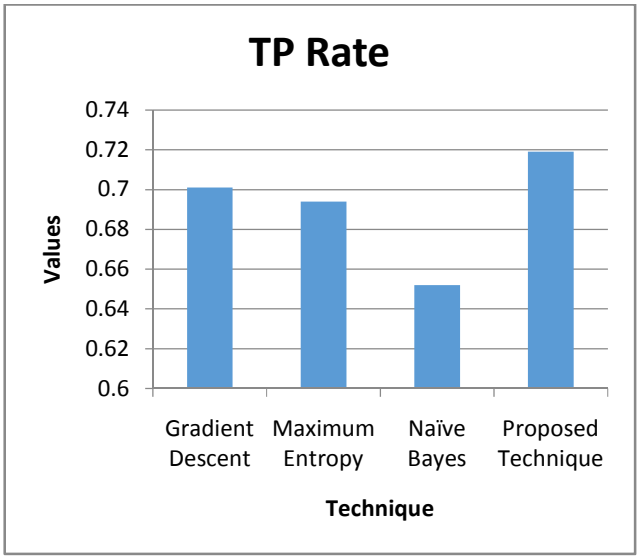

Fig.3. Showing Comparison of Various Techniques on the basis of TP Rate

Table 3: Comparison of Various Techniques based on FP Rate

Technique	FP Rate
Gradient Descent	0.298
Maximum Entropy	0.306
Naïve Bayes	0.348
Proposed Technique	0.283

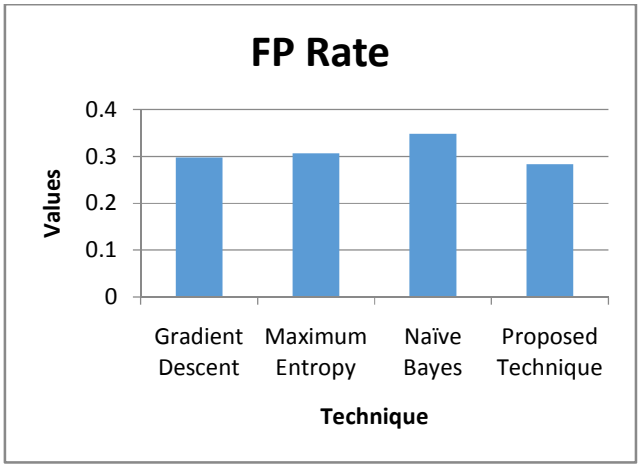

Fig.4. Showing Comparison of Various Techniques on the basis of FP Rate

Table 4: Comparison of Various Techniques based on Precision

Technique	Precision
Gradient Descent	0.702
Maximum Entropy	0.694
Naïve Bayes	0.652
Proposed Technique	0.719

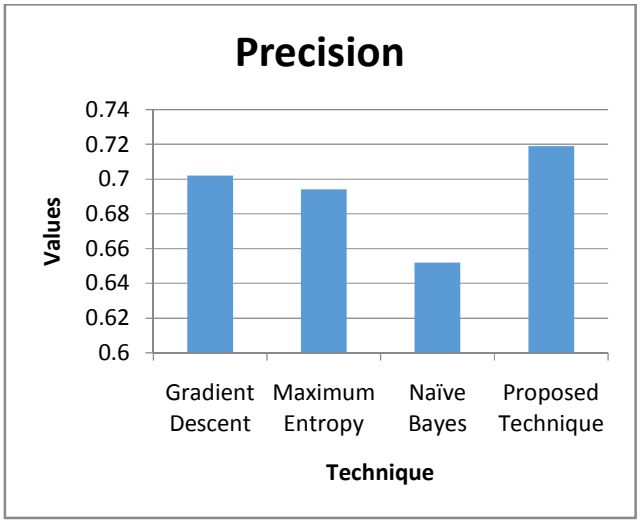

Fig.5. Showing Comparison of Various Techniques on the basis of Precision

Table 5: Comparison of Various Techniques based on Recall

Technique	Recall
Gradient Descent	0.701
Maximum Entropy	0.694
Naïve Bayes	0.652
Proposed Technique	0.719

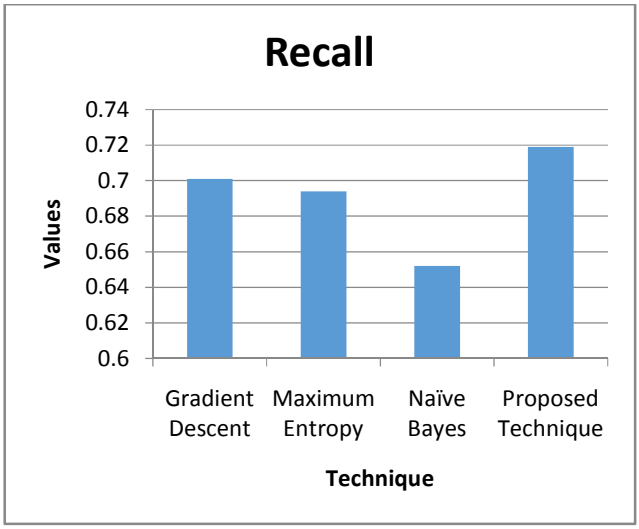

Fig.6. Showing Comparison of Various Techniques on the basis of Recall

Table 6: Comparison of Various Techniques based on F-Measure

Technique	F-Measure
Gradient Descent	0.701
Maximum Entropy	0.694
Naïve Bayes	0.652
Proposed Technique	0.719

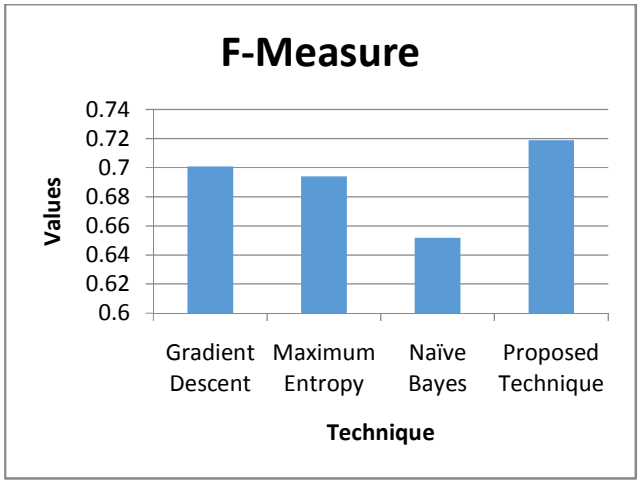

Fig.7. Showing Comparison of Various Techniques on the basis of F-Measure

Table 7: Comparison of Various Techniques based on ROC Area

Technique	ROC Area
Gradient Descent	0.701
Maximum Entropy	0.762
Naïve Bayes	0.719
Proposed Technique	0.776

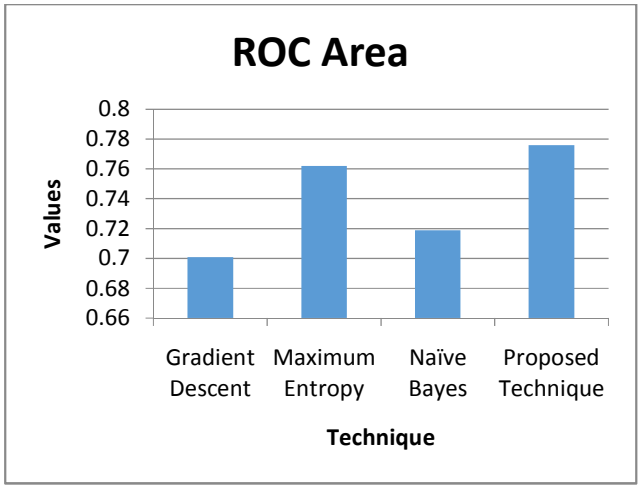

Fig.8. Showing Comparison of Various Techniques on the basis of ROC Area

V. CONCLUSION

In this study a gini index feature selection and ensemble leaning approach was used. This improved the results as compared to using only individual algorithms for analysis. There are still a lot of problems and challenges for sentiment analysis regarding accuracy which can be improved. Apart from accuracy other issues like emoticon analysis, sarcasm analysis and mash up language approach need to be addressed for the improvement of sentiment analysis.

REFERENCES

[1] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis," IEEE Intell. Syst., vol. 28, no. 2, pp. 15–21, 2013.

[2] A. Valdivia, M. V. Luzón, and F. Herrera, "Sentiment Analysis in TripAdvisor," IEEE Intell. Syst., vol. 32, no. 4, pp. 72–77, 2017.

[3] M. Geetha, P. Singha, and S. Sinha, "Relationship between customer sentiment and online customer ratings for hotels - An empirical analysis," Tour. Manag., vol. 61, pp. 43–54, 2017.

[4] M. S. Akhtar, D. Gupta, A. Ekbal, and P. Bhattacharyya, "Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis," Knowledge-Based Syst., vol. 125, pp. 116–135, 2017.

[5] C. Catal and M. Nangir, "A sentiment classification model based on multiple classifiers," Appl. Soft Comput. J., vol. 50, pp. 135–141, 2017.

[6] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," Expert Syst. Appl., vol. 77, pp. 236–246, 2017.

[7] K. Nigam, J. Lafferty, and A. Mccallum, "Using Maximum Entropy for Text Classification," IJCAI-99 Work. Mach. Learn. Inf. Filter., pp. 61–67, 1999.

[8] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," Expert Systems with Applications, vol. 57. Elsevier Ltd, pp. 117–126, 2016.

[9] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," in Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - EMNLP '02, 2002, vol. 10, pp. 79–86.

[10] A. S. Manek, P. D. Shenoy, M. C. Mohan, and K. R. Venugopal, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier," World Wide Web, vol. 20, no. 2, pp. 135–154, 2017.