Sentiment Analysis of Movie Reviews using Machine Learning Techniques

Palak Baid
Department of Computer
Science & Engineering
Jaipur Engineering College and
Research Center
Jaipur, Rajasthan, India

Apoorva Gupta
Department of Computer
Science & Engineering
Jaipur Engineering College and
Research Center
Jaipur, Rajasthan,India

Neelam Chaplot
Department of Computer
Science & Engineering
Jaipur Engineering College and
Research Center
Jaipur, Rajasthan, India

ABSTRACT

Sentiment analysis is the analysis of emotions and opinions from any form of text. Sentiment analysis is also termed as opinion mining. Sentiment analysis of the data is very useful to express the opinion of the mass or group or any individual. This technique is used to find the sentiment of the person with respect to a given source of content. Social media and other online platforms contain a huge amount of the data in the form of tweets, blogs, and updates on the status, posts, etc. In this paper, we have analyzed the Movie reviews using various techniques like Naïve Bayes, K-Nearest Neighbour and Random Forest.

General Terms

Reviews, Classification, Weka, Artificial Intelligence, Machine Learning, Sentiments

Keywords

Sentiment Analysis, Opinion Mining, Movies Reviews, Naive Bayes, K-Nearest Neighbour, Random Forest.

1. INTRODUCTION

With the advent of Web 2.0 various platforms like Facebook, Twitter, LinkedIn, Instagram allows citizens to share their comments, views, feelings, judgements on the myriad of topics ranging from education to entertainment. These platforms contain the huge amount of the data in the form of tweets, blogs, and updates on the status, posts, etc. Sentiment Analysis aims to determine the polarity of emotions like happiness, sorrow, grief, hatred, anger and affection and opinions from the text, reviews, posts which are available online on these platforms. Opinion Mining finds the sentiment of the text with respect to a given source of content. Sentiment analysis is complicated because of the slang words, misspellings, short forms, repeated characters, use of regional language and new upcoming emoticons. So it is a significant task to identify appropriate sentiment of each word. Sentiment Analysis is one of the most active research areas and is also widely studied in data mining. Sentiment analysis is applied in almost every business and social domain because opinions are central to most human activities & behaviors.

Sentiment analysis is very popular because of its efficiency. Thousands of documents can be processed for sentiment analysis. Since it is an efficient process which provides good accuracy, therefore it has various applications:

- Purchasing Merchandise or Service: While purchasing a merchandise or service we must take a right decision which is not a difficult task anymore. By sentiment analysis, people can easily evaluate reviews and opinions of any commodity or service and can effortlessly compare the competing brands.

- Quality Improvement in Product or Service: By Opinion mining, the producers can collect the user's opinion whether favourable or not about their product or service and then they can enhance and upgrade the quality of their product or service.

- Recommendation Systems: By analyzing and categorizing the people's opinion according to their preferences and interests, the system can predict which item should be recommended and which one should not be recommended.

- Decision Making: People's sentiments, ideas, feelings are very important factor to make a decision. While buying any item be it a book or clothes or electronic items user's first to read the opinions and reviews of that particular product and those reviews have a great impact on user's mind.

- Marketing research: The result of sentiment analysis techniques can be utilized in marketing research. By this technique, the attitude of consumers about some product or services or any new government policy can be analyzed.

- Detection of flame: The monitoring of newsgroups, blogs and social media is easily possible by sentiment analysis. This technique can detect insolent, arrogant, over heated words used in tweets, posts or forums and blogs on the internet.

There are following phases of Sentiment Analysis:

Pre-Processing Phase: The data is first cleaned to reduce noise.

Feature Extraction: A token is given to the keywords and this token is now put under analysis.

Classification Phase: Based on different algorithms these keywords are put under certain category.

2. LITERATURE REVIEW

Joscha et. al, in their paper [1] devised and compared various techniques like Bag of words models, n-grams for using semantic information to improve the performance of sentiment analysis. The earlier approaches did not consider the semantic associations between sentences or documents parts. Research by A. Hogenboom et al. [2] neither compared the methodological variants nor provided a method to merge disclosure units in the most favorable manner. They aimed to improve the sentiment analysis by using Rhetoric Structure Theory (RST) as it gives a hierarchical representation at the document level. They proposed an integration of the grid

search and weighting to find out the average scores of sentiment from Rhetoric Structure Theory (RST) tree. They encoded the binary data into the random forest by using feature engineering as it greatly reduced the complexity of original RST tree. They concluded that machine learning raised the balanced accuracy and gives a high F1 score of 71.9%.

Amir Hossein Yazdavar et al. in this paper [3] provided novel understanding of sentiment analysis problem containing numerated data in drug reviews. They analyzed sentences which contained quantitative terms to classify them into opinionated or non-opinionated and also to identify the polarity expressed by using fuzzy set theory. The development of fuzzy knowledge base was done by interviewing several doctors from various medical centers. Although the number of researches has been done in this field (Bhatia, et al., [4]) these do not consider the numerical (quantitative) data contained in the reviews while recognizing the sentiment polarity. Also, the training data used has a high domain dependency and hence cannot be used in different domains. They concluded that their proposed method knowledge engineering based on fuzzy sets was much simpler, efficient and has high accuracy of over 72% F1 value.

Dhiraj Murthy in his paper [5] he identified what roles do tweets play in political elections. He pointed out that even though there were various researches and studies done to find out the political engagement of Twitter, no work was done to find out if these tweets were Predictive or Reactive. In his paper, he concluded that the tweets are more reactive than predictive. He found out that electoral success in not at all related to the success on Twitter and that various social media platforms were used to increase the popularity of a candidate by generating a buzz around them.

Ahmad Kamal in his paper [6] designed an opinion mining framework that facilitates objectivity or subjectivity analysis, feature extraction and review summarization etc. He used supervised machine learning approach for subjectivity and objectivity classification of reviews. The various techniques used by him were Naive Bayes, Decision Tree, Multilayer Perceptron and Bagging. He also improved mining performance by preventing irrelevant extraction and noise as in Kamal's paper. [7].

Humera Shaziya et al. in this paper [8] classified movie reviews for sentiment analysis using WEKA Tool. They enhanced the earlier work done in sentiment categorization which analyzes opinions which express either positive or negative sentiment. In this paper, they also considered the fact that reviews that have opinions from more than one person and a single review may express both the positive and negative sentiment. They conducted their experiment on WEKA and concluded that Naïve Bayes performs much better than SVM for movie reviews as well as text. Naive Bayes has an accuracy of 85.1%.

Akshay Amolik et. al. in his paper [9] created the dataset using twitter posts of movie reviews and related tweets about those movies. Sentence level sentiment analysis is performed on these tweets. It is done in three phases. Firstly, preprocessing is done. Then Feature vector is created using relevant features. Finally, by using different classifiers like Naïve Bayes, Support vector machine, Ensemble classifier, k-means and Artificial Neural Networks, tweets were classified into positive, negative and neutral classes. The results show that we get 75 % accuracy form SVM. He negated Wu et. al.

paper [10] which made an observation that if @username is found in a tweet, it influences an action and also helps to influence the probability. But in this paper Akshay Amolik replaced @username with AT_USER and hashtags were also removed due to which we used Support Vector Machine rather than Naive Bayes which increased the accuracy by 10%.

3. MACHINE LEARNING METHODS

3.1 Naïve Bayes

It is a technique based on Bayes' Theorem. Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. This model is easy to build and particularly useful for very large datasets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

$$P(C|X) = P(X|C) * P(C)/P(X) \qquad (1)$$

P(C|X) is posterior probability of class C

P(C) is prior probability of class C

P(X|C) is probability of predictor given the class.

P(X) is prior probability of predictor

3.2 K- Nearest Neighbour

K-NN is the simplest of all machine learning algorithms. The principle behind this method is to find a predefined number of training samples closest in distance to the new point and predict the label from these. The number of samples can be a user-defined constant or vary based on the local density of points. The distance can be any metric measure. Standard Euclidean distance is the most common choice for calculating the distance between two points. The Nearest Neighbours have been successful in a large number of classification and regression problems, including handwritten digits or satellite image processing and so on.

3.3 Random Forest

Random Forests are the learning method for classification and regression. It construct a number of decision trees at training time. To classify new case it sends the new case to each of the trees. Each tree perform classification and output a class. The output class is chosen based on majority voting that is the maximum number of similar class generated by various trees is considered as the output of the Random Forest.

Random Forests are easy to learn and use for both professionals and laypeople with little research and programming required. It can easily be used by persons that don't have a strong statistical background.
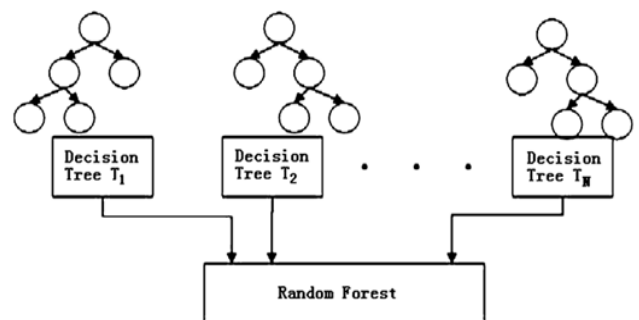

Fig 1: Random forest figure taken from techleer.com

4. EXPERIMENTAL SETUP

The research started with analyzing different research and review papers on sentiment analysis and the summary for each paper was created by reading and understanding the paper. Study of frequently used classification algorithms such as Naïve Bayes, Random Forest, k-nearest neighbour, Decision Tree Induction, Support Vector Machine was done.

Data was collected from 2000 user-created movie reviews archived on the IMDb (Internet Movie Database) web portal at http://reviews.imdb.com/Reviews and is known as "Sentiment Polarity Dataset version 2.0": 1000 positive and 1000 negative processed reviews. This data was then converted into arff format. The data was in the txt_token file which had 2 Sub folders for positive and negative. The converted data was imported into the WEKA tool using Text Directory Loader. After that text pre-processing was done on WEKA tool.

WEKA is open source software under GNU General Pubic License. The system is developed at the University of Waikato in New Zealand. "WEKA" is an abbreviation for Waikato Environment for Knowledge Analysis. The software is written in Java language. It provides an implementation of machine learning algorithms. It contains modules for data processing.

WEKA supports various data mining tasks like data pre-processing, binning, clustering, regression and feature selection. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is used for processing using WEKA. The classification problem is a supervised learning task that consists of assigning a class label to an unclassified tuple according to an already classified instance set, that is used as a training set for the algorithm.

The 3 classification algorithms that we have used are Naive Bayes and k-nearest neighbors (k = 3) and Random Forest. The quality measure that will be considered is the percentage of correctly classified instances. For the validation phase, we have used the 10 fold cross validation. Steps adopted for performing Experiments are as below:

Step 1: Importing the dataset in WEKA. The first step performed was to import the dataset into the WEKA tool. To perform this step a simple import procedure for textual datasets called TextDirectoryLoader component was used.

Step 2: After importing the dataset it is converted and saved in the ARFF format.

Step 3: After that, a relation is created by containing 2000 instances and two attributes "Text" and "Class". The figure shows the uniform distribution of the attribute Class. Blue color represents reviews of negative polarity and Red color represents reviews of positive polarity as shown in Fig 2.

Step 4: Then the StringToWordVector filter is applied.

Step 5: Then the AttributeSelection filter is applied.

Step 6: After applying the AttributeSelection filter, the results are obtained as shown in Fig 3.

Step 7: Three algorithms are performed on the data generated from above steps. The three algorithms are Naïve Bayes, K Nearest Neighbour, and Random Forest.

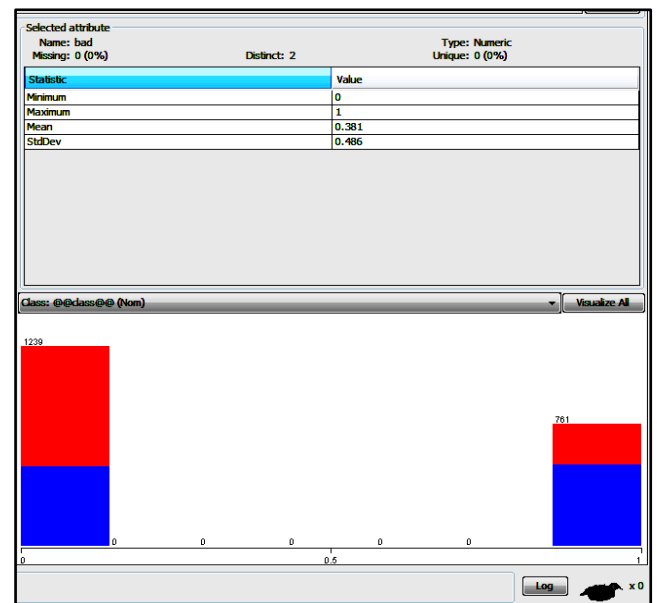**Fig 2: Weka interface after loading the data in arrf format**

Fig 3: Weka interface after AttributeSelection Filter

5. RESULTS

Navie Bayes Algorithm, K-Nearest Neighbour Algorithm and Random Forest Algorithms were performed on the Data set. Results are presented in the form of figures and tables as shown below. Best accuracy was given by the Naïve Bayes algorithm.

Table 1. Percentage Accuracy of Various Algorithms

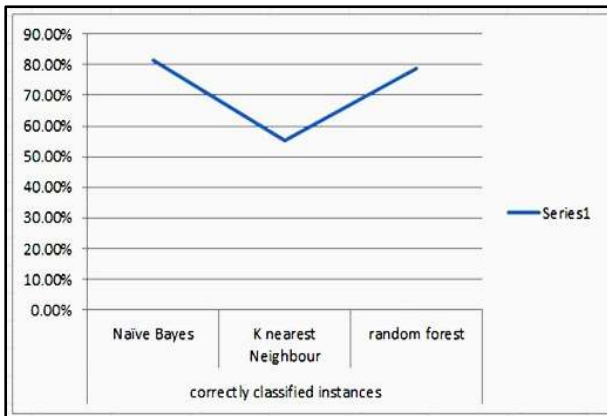| Naïve Bayes	K Nearest Neighbour	Random Forest
81.4 %	55.30 %	78.65 %

Fig 4: Graph of Accuracy of various Algorithms

Table 2. Mean Absolute Error of Various Algorithms

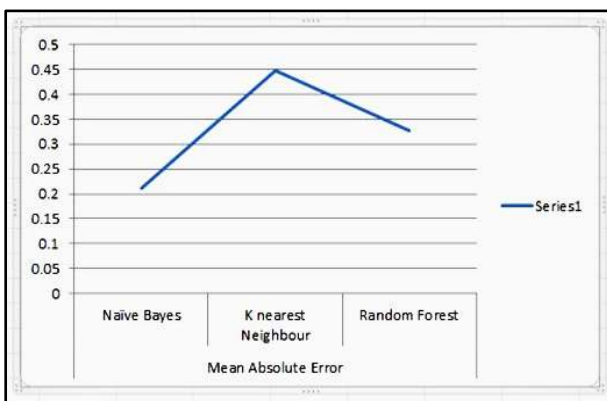| Naïve Bayes	K Nearest Neighbour	Random Forest
0.2112	0.4475	0.3278

Fig 5: Graph of Mean Absolute Error of Algorithms

Table 3. Mean Absolute Error of Various Algorithms

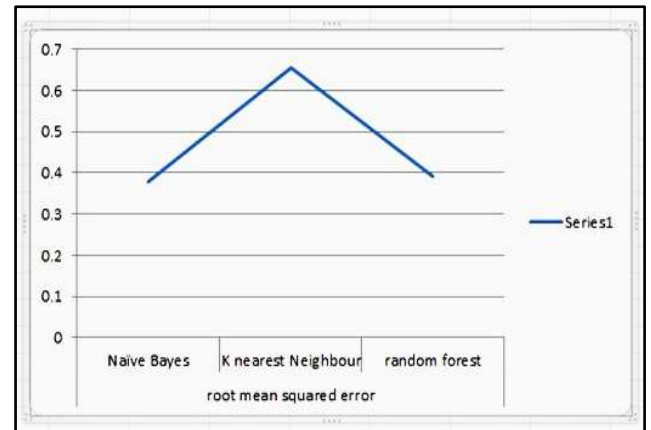| Naïve Bayes	K-Nearest Neighbour	Random Forest
0.3793	0.6554	0.3925

Fig 6: Graph of Root Mean Squared Error of Algorithms

6. CONCLUSIONS

In this research, various techniques were used to identify the polarity of the tweets. The algorithms performed were Naïve Bayes, K-Nearest Neighbour, Random Forest. The best results were given by Naïve Bayes classifier. The Naïve Bayes classifier achieved 81.45% accuracy, Random Forest classifier we achieved 78.65% accuracy, K-Nearest Neighbour classifier achieved 55.30% accuracy.

As only few algorithms were tested , it is required to test other algorithms or create hybrid methods so that accuracy of the results can be increased.

Finding the polarity of the reviews can help in various domain. Intelligent systems can be developed which can provide the users with comprehensive reviews of movies, products, services etc. without requiring the user to go through individual reviews, he can directly take decisions based on the results provided by the intelligent systems.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] Joscha Markle-Huß, Stefan Feuerriegel, Helmut Prendinger. 2017 Improving Sentiment Analysis with Document-Level Semantic Relationships from Rhetoric Discourse Structures, Proceedings of the 50th Hawaii International Conference on System Sciences.

[2] A. Hogenboom, F. Frasincar, F. de Jong, and U. Kaymak,. 2015, Using Rhetorical Structure in Sentiment Analysis, Communications of the ACM, vol. 58, no. 7, pp. 69–77.

[3] Amir Hossein Yazdavar, MonirehEbrahimi, Naomie Salim, 2016, Fuzzy Based Implicit Sentiment Analysis on Quantitative Sentences, Faculty of Computing, UniversitiTechnologi Malaysia, Johor, Malaysia, Journal of Soft Computing and Decision Support Systems vol 3:4, pp.7-18.

[4] Bhatia, R. S., Graystone, A., Davies, R. A., McClinton, S., Morin, J., & Davies, R. F. 2010, Extracting information for generating a diabetes report card from free text in physicians notes. Paper presented at the Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents.

[5] Dhiraj Murthy, Twitter and elections: are tweets, predictive, reactive, or a form of buzz?, Information, Communication & Society, 18:7, 816-831, DOI:10.1080/1369118X.2015.1006659

[6] Kamal A., 2015, Review Mining for Feature Based Opinion Summarization and Visualization.

[7] Kamal, A. 2013 Subjectivity Classification using Machine Learning Techniques for Mining Feature-Opinion Pairs from Web Opinion Sources. International Journal of Computer Science Issues 10(5), 191- 200.

[8] Humera Shaziya, G.Kavitha, Raniah Zaheer, 2015, Text Categorization of Movie Reviews for Sentiment Analysis , International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Issue11.

[9] Akshay Amolik, Niketan Jivane, Mahavir Bhandari, Dr .M. Venkatesan, Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques, School of Computer Science and Engineering, VIT University, Vellore.

[10] Y. Wu and F. Ren, 2011, Learning Sentimental influence in twitter, Future Computer Science and Application (ICFCSA), 2011, International Conference IEEE vol. 119122.