Sentiment Analysis On Movie Reviews Using NAÏVE BAYES Classifier

Mohana Pranadeep potti, ManneDineshKmar, Nagabhyrava Saswanth Ram,P.V.R.Sandeep

P.R.Krishna Prasad

Student, Computer Science Department, Vasireddy Venkatadri Institute of technology, Andhrapradesh, India
Student, Computer Science Department, Vasireddy Venkatadri Institute of technology, Andhrapradesh, India
Student, Computer Science Department, Vasireddy Venkatadri Institute of technology, Andhrapradesh, India
Student, Computer Science Department, Vasireddy Venkatadri Institute of technology, Andhrapradesh, India
Assoc.Professor, Computer Science Department, Vasireddy Venkatadri Institute of technology, Andhrapradesh, India

***Abstract*—Sentiment Analysis or Opinion mining on movie reviews. Sentiment is a thought, view, or attitude, especially one based mainly on emotion instead of reason For this we need to classify each and every review, this can be done through classifiers such as NAÏVE BAYES, SVM, K-NN, we preferred NAÏVE BAYES classifier, since it is a probabilistic approach we depend upon each word's probability to be positive and negative, through this we will classify the review to be one of those. After classifying each review the sentiment is drawn based on the count of the classes. Consider 100 reviews, in those reviews there will be positive as well as negative reviews of that particular movie, the review is purely based on the public opinions being expressed in the portal. After classifying the reviews we go for the max count method to predict the outcome.**

This algorithm is proposed in order to avoid the fake reviewing system followed now a days to promote the movie by the producers and crew of the movie. This gives the straight opinion of the public.

***Keywords*—SentimentAnalysis; OpinionMining; NaïveBayes; Probablistic approach.**

I. Introduction

Sentiment analysis is a series of methods, techniques, and tools used to detect and extract subjective information,such as opinion and attitudes, from language. Traditionally, sentiment analysis has been about opinion polarity, i.e., whether someone has positive, neutral, or negative opinion towards something.why sentiment analysis and opinion mining are often used as synonyms, although, we think it is more accurate to view sentiments as emotionally loaded opinions?

The interest on other's opinion is probably almost as old as verbal communication itself. Historically, leaders have been intrigued with the opinions of their subordinates to either prepare for opposition or to increase their popularity.

Sentiment analysis is a well-known task in the realm of natural language processing. The objective is to determine the polarity of that text.The sentiments can consist of different classes. In this study, we consider two cases: 1) A movie review is positive (+) or negative (-). 2) A movie review is very negative (- -), somewhat negative (-), neutral (o), somewhat positive (+), or very positive (+ +).

1) Positive Sentiment in subjective sentence: "I really love the movie BAAHUBALI"—This sentence expresses positive sentiment about the movie BAAHUBALI and we can tell that from the sentiment threshold value of word "love". So, threshold value of word "love" has positive numerical threshold value.
2) Negative sentiment in subjective sentences: "MISTER is a disaster" this sentence expresses negative sentiment about the movie named "MISTER" and we can decide that from the sentiment threshold value of word "disaster". So, threshold value of word "disaster" has negative numerical threshold value.

Sentiment Analysis is of three different types: Document level, Sentence level and Entity level.However we are studying phrase level sentiment analysis. The traditional text mining concentrates on analysis of facts whereas opinion mining deals with the attitudes. The main fields of research are sentiment classification, feature based sentiment classification and opinion mining. Now, the use of sentiment analysis in a commercial environment is growing. This is evident in the increasing number of brand tracking and marketing companies offering this service. Some services include: - Tracking users and non-users opinions and ratings

This is both advantageous to the producers and consumers. How in the sense is the producer get the exact and accurate review from the consumers and the consumers use the other consumers' reviews to use the product similarly in the movie industry it is both use full to the movie goers and the crew of the movie.

II. RELATED WORK

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 *International Journal of Computer Applications (0975 – 8887) Volume 179 – No.7, December 2017* 46 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%.

III. METHODOLOGIES

NAÏVE BAYES classifier

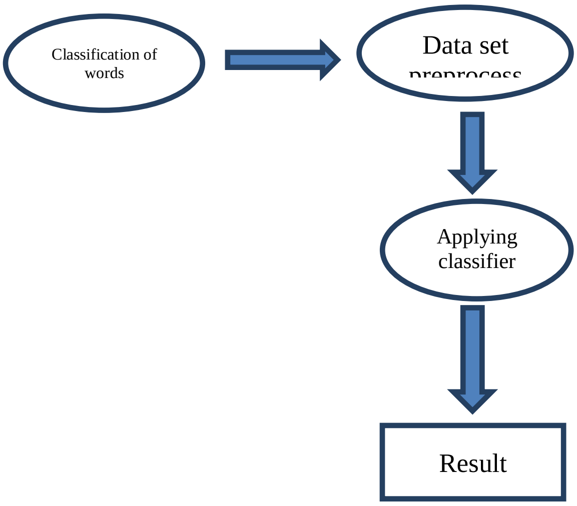
Naive Bayes has been studied extensively since the 1950s. It is a probabilistic approach.The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Though prediction of sentiment is not that easy Naïve Bayes makes it easy through its simplest approach.

BAYES RULE:

$$P\left(\frac{C}{T}\right) = \frac{P(C) * p\left(\frac{T}{C}\right)}{P(T)}$$

P(c/t): Posterior Probability
P(c)/P(t): Prior Probability
P(t/c):Conditional Probability

c - classified class
t -test/trained data



FREQUENCY TABLE

Word	+ve(count)	-ve(count)
Good	2	1
Bad	1	1
Very	1	1
Not	1	1
Excellent	1	
Superb	1	
Poor		1
Disaster		1
Flop		1

PRIOR PROBABILITY TABLE

Word	P(t/c) +Ve	P(t/c) -Ve
Good	2/7	1/7
Bad	1/7	1/7
Very	1/7	1/7
Not	1/7	1/7
Excellent	1/7	
Superb	1/7	
Poor		1/7
Disaster		1/7
Flop		1/7

For Example :

let us classify the following words

good	
Very good	+ve
Very bad	-ve
not bad	+ve
not good	-ve
Excellent	+ve
Superb	+ve
Poor	-ve
Disaster	-ve
Flop	-ve

TOTAL: 10
+ve: 5
-ve : 5

P(c):5/10=0.5(+ve)
P(c):5/10=0.5(-ve)
Let us consider a review:
"This movie is very good"

Very, good
Are the classified words

Now let us calculate the posterior probability using BAYES RULE

Very : p(t/c) =1/7(+ve)
=1/7(-ve)

Good : p(t/c) =2/7(+ve)
=1/7(-ve)

P(c) = 1/2(+ve)
= 1/2(-ve)
P(t) = very :2/14
P(t) =good:3/14

Substituting in the formula:

For Positive:
(1/2)*(1/7)*(2/7)
(2/14)*(3*14)

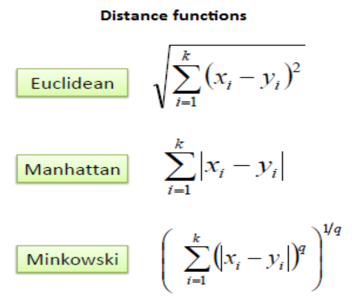
=0.6669

For negative:
(1/2)*(1/7)*(1/7)
(2/14)*(3/14)

=0.3334

K-NN:

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.



$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

$$\sum_{i=1}^{k}|x_i - y_i|$$

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

Decision Tree:

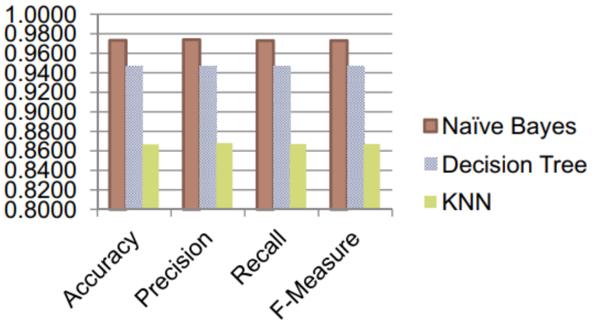
A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning, covering both classification and regression. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, its also widely used in machine learning, which will be the main focus of this article.

Lets take a closer look at cost functions used for classification and regression. In both cases the cost functions try to find most homogeneous branches, or branches having groups with similar responses. This makes sense we can be more sure that a test data input will follow a certain path.

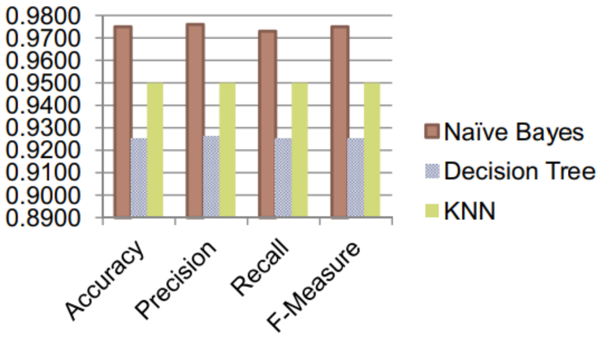
Regression : sum(y—prediction)²

Classification : G = sum(pk * (1—pk))

IV. Experimental Results

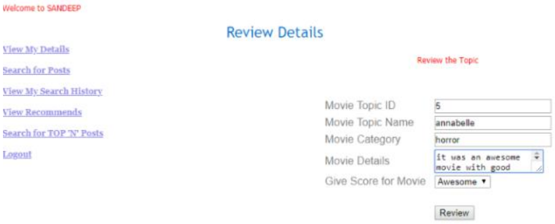


Comparison of results of three Algorithms on training dataset



Comparison of results of three Algorithms on testing dataset

V. Implementation

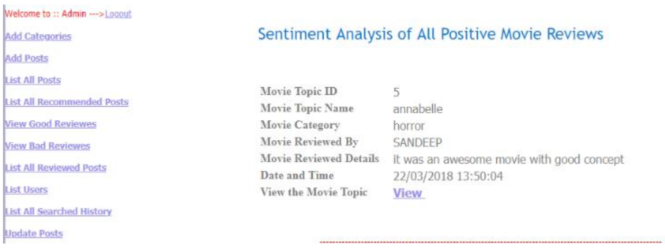


Review Details

Review the Topic

Movie Topic ID 5
Movie Topic Name annabelle
Movie Category horror
Movie Details it was an awesome movie with good
Give Score for Movie Awesome

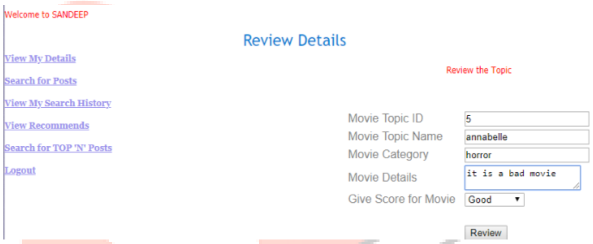
In the above picture a positive review is given and we expect that review to be stored in positive reviews data base. After entering the review the classifier compares the words in it and then it classifies the review.



Sentiment Analysis of All Positive Movie Reviews

Movie Topic ID	5
Movie Topic Name	annabelle
Movie Category	horror
Movie Reviewed By	SANDEEP
Movie Reviewed Details	it was an awesome movie with good concept
Date and Time	22/03/2018 13:50:04
View the Movie Topic	View

As expected the review is stored the positive reviews.

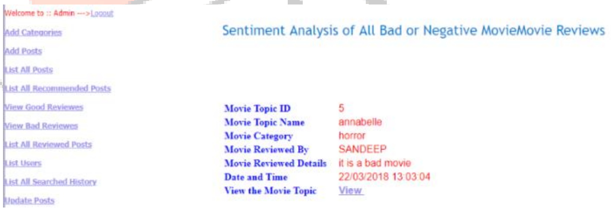


Review Details

Review the Topic

Movie Topic ID 5
Movie Topic Name annabelle
Movie Category horror
Movie Details it is a bad movie
Give Score for Movie Good

In the above picture a negative review is given and we expect that review to be stored in negative reviews data base. After entering the review the classifier compares the words in it and then it classifies the review.



Sentiment Analysis of All Bad or Negative MovieMovie Reviews

Movie Topic ID	5
Movie Topic Name	annabelle
Movie Category	horror
Movie Reviewed By	SANDEEP
Movie Reviewed Details	it is a bad movie
Date and Time	22/03/2018 13 03 04
View the Movie Topic	View

As expected the review is stored the negative reviews.

VI. Conclusion

This paper aims at the text classification and opinion mining, Naïve Bayes classifier partitions the text composed of the documents with highest probabilities. It is the optimized probabilistic technique here we drew the comparison with two another classification techniques. By comparing we conclude that Naïve Bayes out performs the other two techniques.

VII. REFERENCES

[1] Joscha Markle-Hub, 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. Hogeboom, F. Frasincar, F. de Jong, and U. Kayak,. 2015, Using Rhetorical Structure in Sentiment Analysis, Communications of the ACM, vol. 58, no. 7, pp. 69–77.

[3] Amir Hossein Cadaver, 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 Loui 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] HumiraShaniya, Kavitha, Rania Zaheer, 2015, Text Categorization of Movie Reviews for Sentiment Analysis, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Issue11.

[9] AkshayaMaloik, 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.