Sentiment Analysis of Movie Reviews Based on Feature Selection And Ranking Method

Aruna Raikwar[1], Nagendra Kumar[2]
[1] Dept of CSE
[2] Assistant Professor, Dept of CSE
[1, 2] Shri Ram Institute of Science & Technology, Jabalpur, Madhya Pradesh, India.

Abstract- *Data classification is highly significant in data mining which leads to a number of studies in machine learning with preprocessing and algorithmic technique. Class imbalance is a problem in data classification wherein a class of data will outnumber another data class. Sentiment Analysis is an evaluation of written and spoken language which determines a person's expressions, sentiments, emotions and attitudes and is commonly used as dataset in machine learning.*

Twitter is an emerging platform to express the opinion on various issues. Plenty of approaches like machine learning, information retrieval and NLP have been exercised to figure out the sentiment of the tweets. We have used movie reviews as our data set for training as well as testing and merged the naive bayes and adjective analysis for finding the polarity of the ambiguous tweets. Experimental outputs reveal that the overall accuracy of the process is improved using this model. In this work we have focused on two areas like: Feature Selection and Ranking and second using machine learning techniques. We use "Twitter" movie review dataset. We also use accuracy comparison framework for comparing algorithms based on execution time.

Keywords- Sentiment analysis, twitter, adjective analysis, naive bayes, ranking method.

I. INTRODUCTION

The meaning of opinion itself is still very broad. Sentiment analysis and opinion mining mainly focuses on opinions which express or imply positive or negative sentiments. To do an analysis, classification plays a key role in opinion mining. A Classification Algorithm is a procedure for selecting a hypothesis from a set of alternatives that best fits a set of observations. Opinions are central to almost all human activities because they are key influencers of our behaviours. Whenever there is a need to make a decision, others' opinions are required. In the real world, businesses and organizations always want to find consumer or public opinions about their products and services. Individual consumers also want to know the opinions of existing users of a product before purchasing it, and others' opinions about political candidates before making a voting decision in a political election. In the past, when an individual needed opinions, he/she asked friends and family. When an organization or a business needed public or consumer opinions, it conducted surveys, opinion polls, and focus groups. Acquiring public and consumer opinions has long been a huge business itself for marketing, public relations, and political campaign companies. Opinion summarization summarizes opinions of articles by telling sentiment polarities, degree and the correlated events. With opinion summarization, a customer can easily see how the existing customers feel about a product, and the product manufacturer can get the reason why different stands people like it or what they complain about. A seller's job can be quite complicated or it can be quite easy. The two contradictory terms define the selling experience, based on the fact as how seller interprets the consumer interests. Unless one is a psychic or knows how to get into others mind the actual demand of the customer's and the product can't be collaborated. Having a right product is important and equally important is to present it before the right customer (one who actually needs it or is interested in it). The product should put on positive feeling of ownership among the individuals. And such feelings are clearly expressed in opinion mining polls.

The internet technology has not only brought people together by connecting them on social-networks but has also played an important role in the expansion of e-commerce. Amazon, Snapdeal, Taobao, Eopinion, etc. are one of those e-commerce websites which not only sell the products online, but also provide a platform where the customers are allowed to post the reviews about the purchased products[1]. Research shows that online customer product reviews not only have significant impact on customers' online purchase decisions but are also helpful for the manufacturers to improve the product design and quality and for the online retailers to improve their services. Lengthy reviews make it hard for the online customers to read full reviews in order make a decision on whether to purchase the product or not. On the other hand, reading incomplete reviews might give a prejudiced view to the customers. Another problem that is frequently quoted in many studies, is regarding the customer preferences for

different product features. This leads to a finale that a particular review, though may be descriptive but may not be helpful to a customer who is looking for the features not mentioned in that review. There are a very few online review platforms which care about organizing the reviews in manner that is feature oriented and customer friendly. Many researchers are working in the field of opinion mining and sentiment analysis to extract product specific features.
In general, feature based opinion mining involves three subtasks viz.

(i) To correctly identify the opinionated and product specific features,
(ii) To identify the review sentences attributing positive/ negative opinions to the extracted features and
(iii) To generate a feature based summary from the information extracted.

The aim is to improve the accuracy and simplify the task of mining the opinions of customer reviews with respect to the features extracted.

A recent study focused on cost-effective values of online reviews and provides deep understanding between product reviews and their sales performance. People tend to read online reviews understanding the opinions and sentiments and trust them as much as they are recommended by their friends or families. Twitter, a social networking service plays significant role in social networking research. Tweets give rich information about movie, product, or service.

II. REVIEW OF LITERATURE

The previous study of sentiment analysis has been conducted by Pang et al.[2] using machine learning algorithms to classify a movie review. This research deals with the classification of related sentiment movie review whether a movie is rated positively or negatively. In the preprocessing stage, neither were performed stemming and elimination of stop words. The learning algorithms used were Naive Bayes, Maximum Entropy (MAXENT), and Support Vector Machine (SVM).

Later, Franky and Manurung[3] replicated the study with the data from a movie review on Pang et al.[2] by doing translating film reviews from English into Indonesian. Feature selection is made with some features include 2000 features unigram with the highest frequency of occurrence, 5000 unigram features with the highest frequency of occurrence, all unigram features, and 25 features unigram at the end of a movie review.

Recent research on Indonesian tweet sentiment analysis has been done in[4] with data collected from emoticons and national media accounts data. Other research on sentiment analysis is tweets about television shows[5], "Kurikulum 2013" (*2013 Curriculum)*, the presidential candidate. These researches employ various machine learning algorithms including SVM, Naive Bayes, and C4.5 with various features. Vohra and Teraiya [6] explained that in sentiment analysis, also, to apply machine learning algorithms such as Naive Bayes, Support Vector Machine, and C4.5 there were known other methods such as lexicon method for identifying the polarity of sentiment.

Taboada et al.[7] applied the lexicon method by providing polarity value of the features of a particular word based on the type of position words (Part of Speech) in a sentence. The development of research related to sentiment analysis, a hybrid method or the incorporation of machine learning methods and lexicon method had also been developed as the study of Zhang et al.[8]. Chen and Goodman previously studied smoothing methods for language modelling, including additive smoothing (Laplace/Lidstone), Good-Turing, Jelinek-Mercer, Katz smoothing, Witten-Bell, Absolute Discounting, and Kneser- Ney. For comparison evaluation, they use the measure cross entropy which is a common metric for evaluating language models. The best smoothing method according to this study is their modification of Kneser Ney smoothing. Smoothing methods have also been studied for language models applied to information retrieval, which investigated Jelinek-Mercer, Dirichlet, and Absolute Discounting smoothing methods. The JM method gave the best average performance in query retrieval. This paper also proposes a two-stage smoothing method which gives better performance than single smoothing. Research on smoothing methods has also been done for text classification with Naive Bayes.

Sentiment Analysis is the thorough research of how opinions and perspectives can be relate to ones emotion and attitude shows in natural language respect to an event. Recent events show that the sentiment analysis has reached up-to great achievement which can surpass the positive vs negative and deal with whole arena of behavior and emotions for different communities and topics. In the field of sentiment analysis using different techniques good amount of research has been carried out for prediction of social opinions. Pang and lee (2002) proposed the system where an opinion can be positive or negative was found out by ratio of positive words to total words. Later in 2008 the author developed methodology in which tweet outcome can be decided by term in the tweet. Jiang (2011) and Tan (2011) have applied maximum entropy (Max-Ent), Naïve Bayes (NB) and support

vector machines (SVM) as supervised classifiers. Chen (2011) employed the feed-forward BPN network and uses sentiment orientation to calculate the results at each neuron.

Malhar and Ram (2014) employed supervised machine learning techniques and artificial neural networks to classify twitter data along with case study of Presidential and Assembly elections which results SVM outperforms all other classifiers. Anton and Andrey reviewed the existing techniques and developed a model for automatic sentiment analysis of twitter messages using unigram, bigram and jointly i.e. hybrid feature. Pak and Paroubek (2010) perform linguistic analysis and build a sentiment classifier to determine positive, negative and neutral sentiments for a document. Tang, Tan and Cheng exchanges views on main approaches and issues to problems like word sentiment classification, opinion extraction, subjectivity classification and document sentiment classification. Sentiment classifier can be prevented from probably misguiding or irrelevant text by subjective classification. Kopel and Schler explain that it is very important to use neutral messages to get good knowledge of polarity. The authors also states that positive and negative messages alone will not give proper understanding about neutral messages.

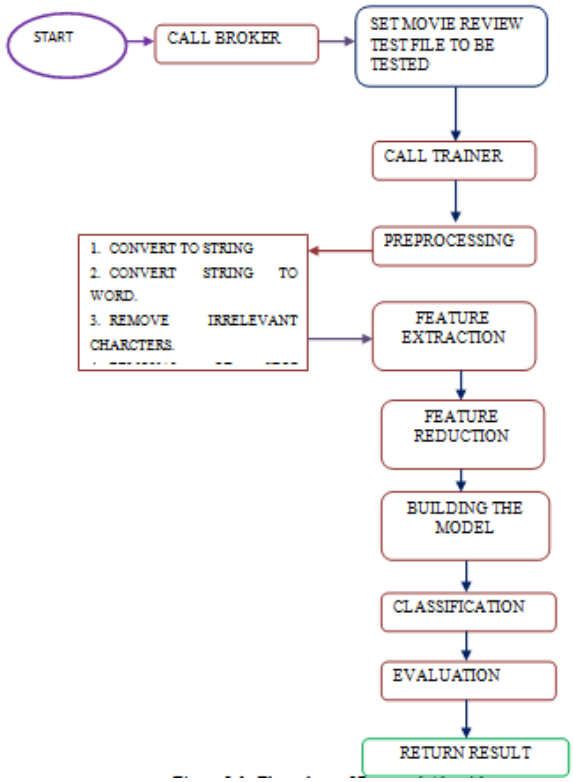
Sourav Das et. al. present a simple and robust work to gather, analyze and graphically represent people's opinion about India's new taxation system using Naive Bayes algorithm. Naive Bayes Classifier primarily works on the conditional probability theory. It offers the assumption of a particular feature from a class of features. But not necessarily, it will come out as the accurate one. It only work if the probability that is already recorded for a particular class.

Huma Parveen, Prof. Shikha Pandey, represents a supervised learning method as well as a statistical method for classification. It is probabilistic model and it permits us to capture uncertainty about the model in a principled way by determining probabilities. It helps to solve diagnostic and predictive problems.

III. PROPOSED SYSTEM

3.1 Proposed System

This section gives the description of the steps followed for the movie dataset mining for sentiment analysis. In this work we have focused on two areas like: Feature Selection and Ranking and second using machine learning techniques. We use "Twitter" movie review dataset. The flowchart as shown in Figure 3.1 explains the overall methodology.



3.2 Proposed Algorithm

Input: dataset.

Output: classified output.

1. Take a data set as input.
2. If that set has more features then apply the feature selection technique as pre-processing technique.
3. Apply parallelism from step 4 to step 6.
4. Evaluate the entropy value and information gain ratio.
5. Construct the models separately using proposed algorithm based on entropies.
6. Find the accuracy and execution time of each model and store the value in array.
7. Find a model that has maximum Accuracy.
8. If two have maximum accuracy then
9. Find a minimum execution time of the model that has maximum accuracy.
10. Classify by that model which has minimum execution time.
11. Else classification done by the model which has maximum accuracy.
12. End

3.3 Feature Extraction

In the process of feature extraction, movie features are extracted from every sentence. For finding the polarity of text document, it is necessary to understand the sentiment score with its usage as well as their relationship with all the nearby words. Following are some features that affect the polarity of the document.

For primary feature extraction, we have used N-gram tokenizer which tokenizes the input tweet into word *n*-grams such as unigram, bigram etc. Frequency of an n-gram feature in a tweet is considered as the feature value. This is a collection of positive, negative and neutral words along with their broad part-of-speech categories. For our defined feature, the feature value is calculated based on how many polarity words of particular type are contained in a tweet. For example, if there are three sentiment classes such as positive, negative and neutral, we consider three SentiWordnet features i.e. how many positive words found in the SentiWordnet are also found in the tweet, how many negative words found in the SentiWordnet are also found in the tweet and how many neutral words found in the SentiWordnet are also found in the tweet. If there are *m* number of n-gram features, our feature set contains a total of *m+3* features where 3 is for the SentiWordnet features. We transform each tweet into vector presentation of length *m+3* and the vector is labeled with the class of the training tweet under consideration.

3.4 Stop words removal

Stop words are words which do not add into much of a meaning to the topic and yet appear most frequently in the documents like articles or prepositions. We maintain a stop words list. During training phase, we skip the word if it is a stop word, and do not consider it in the Bayes probability calculation. Below are steps to find spam or ham using this method:

Step 1:- Let S be a tweet message. Take two variables Positive and Negative for counting, initialized to 0. Convert all words to lower case.
Step 2:- perform preprocessing on S.
Step 3:- For each word w_i in S
If w_i found in stopword list Then
w_i is removed from S.
End If.
End For
Step 4:-For each word w_j in S
If w_j found in Positive word dataset then
Increment count of Positive
End if
Adjust probability of S.
End For
Step 5:-For each word w_j in S
If w_j found in Negative word dataset then
Increment count of Negative
End if
Adjust probability of S.
End For
Step 6:- If Positive count >Negative Count Then
S will be identified as Positive.
Else
S will be Negative.
End if.

IV. IMPLEMENTATION AND EVALUATION

4.1 Technical Specifications

Following are minimum specifications for development of the system:

TABLE 1: Technical Specifications.

Hardware Configuration	At least 2 GB free memory on storage disk, 512 MB RAM, Intel Pentium-4 Processor, Android Mobile
Operating System	Windows 7 32-bit OS recommended
Programming Language	Java
Development Tool	Netbeans IDE

4.2 Third Party Tools

Third party tools we are using are:

1. *Weka Tool*
2. *LibSVM*

4.3 Results & Evaluation

The classification performance can be evaluated in three terms: accuracy, recall and precision as defined below. Accuracy explains correctly classified instances. Precision and Recall are in weighted average for positive and negative terms.

Classifier	Accuracy (in %)	Precision	Recall
Naive Bayes	60.7042	0.607	0.607
Logistic Regression	70.7042	0.708	0.707
SVM	69.5775	0.731	0.696
Proposed	**78.7324**	**0.788**	**0.787**

Table: Performance evaluation.

Chart below represent comparison of accuracy between different algorithms.

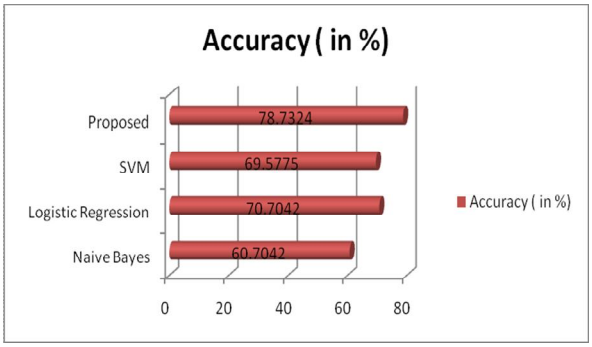


Fig 5.9: Evaluation of accuracy.

V.CONCLUSION

Opinion mining has become popular research area due to the increasing number of internet users, social media etc. In this work, we extracted new features that have a strong impact on finding the polarity of the movie reviews. We then perform the feature impact analysis by estimating the information gain for each feature in the feature set and used it to derive a reduced feature set. The main goal of this work is to classify the sentences according to its sentiment by using Decision Tree classification technique. This process of extracting the text having sentiment deals with finding the sentiment feature set from the sentences.

REFERENCE

[1] Sivarajah, Uthayasankar, Zahir Irani, and Vishanth Weerakkody, "Evaluating The Use And Impact of Web 2.0 Technologies in Local Government," Government Information Quarterly. Elsevier, pp. 473–487, 2015.

[2] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up?", Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02, 2002.

[3] Franky and R. Manurung. Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews. In Proceedings of the International Conference on Advanced Computational Intelligence and Its Application, 2008.

[4] Aliandu, Paulina, Sentiment Analysis on Indonesian Tweet. The Proceedings of The 7th ICTS, Bali, May 15th-16th, 2013.

[5] Sentiaji, Aditia R and Adam M Bachtiar. Analisis Sentiment Terhadap Acara Televisi Berdasarkan Opini Public. Jurnal Ilmiah Computer Dan Informatika (KOMPUTA), ISSN: 2089-9033, 1-6, 2013.

[6] S. M. Vohra and J. B. Teraiya. A Comparative Study of Sentiment Analysis Techniques. Journal of Information, Knowledge, and Research in Computer Engineering, Volume 02, Issue 02, 313-317, 2013.

[7] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, "Lexicon-Based Methods for Sentiment Analysis", Computational Linguistics, vol. 37, no. 2, pp. 267-307, 2011.

[8] Khan, Aamera ZH, Mohammad Atique, and V. M. Thakare. "Combining lexicon-based and learning-based methods for Twitter sentiment analysis." International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE) (2015): 89.