Sentiment Classification of Customer Reviews Using Bayesian Classifier

Muhammad Rizwan Rashid Rana Muhammad Aun Akbar Tauqeer Ahmad

***Abstract*—Sentiment Classification is the process of exploring sentiments, emotions, opinions and facts in the sentences which are expressed by the people. Machine learning techniques and Lexicon based techniques are mostly used in sentiment classification to predicate sentiments from customers reviews and comments. In this paper, we aim to classify movie reviews in positive and negative classes by using Naive Bayes classifier (Machine Learning algorithm). Experimental results shows that accuracy of Naive Bayes classifier on large movie reviews dataset is 86.8%.**

***Index Terms*—Machine Learning; Nave Bayes classifier; Parts of speech (POS) tagging; sentiment classification**

I. Introduction

The expeditious development of Web 2.0 has caused the explosion of the amount of data generated during recent years. Micro blogging (i.e facebook, twitter) became the major outlet in increasing the data. Peoples share their ideas, opinions, feelings and many more on social media sites. Over 300,000 posts are posted on facebook and 3000 photos are uploaded from Flicker in 60 seconds [1]. Daily more than 300 million tweets are shared on twitter [2]. In addition 153 million blogs are active and posted daily. There are many market sites are working, millions of reviews are posted per day on these sites. More than 25 thousands items are purchased per day in Amazon. As a result of Micro blogging, blogs, markets sites millions of thoughts, opinions, comments, statuses spread over the internet[3].

Sentiment Classification is the process of judging the sentiments and emotions from reviews. Using these indentified sentiments and emotions system automatically determine the sentiment category such as positive, negative and neutral [4]. Sentiment categories are binary, ternary and multiple categories. Binary category includes positive and negative, ternary includes positive, negative and neutral, multiple categories include Happy, Sad, Anger etc [5]. There are many benefits of sentiment classification, It increase the decision making power on any individual person and company. Sentiment classification points out the positive and negative feature of any product. For example if a person wants to buy a laptop, he can surf the internet and read customer reviews about the product and at the end he will take decision on the basis of customer comments. Reading comments from hundreds of sites is a difficult task. Sentiment classification solves this problem using classify the positive and negative reviews in two categories which helps more effective decision making.

Sentiment classification can be divided in three different levels. Document level, Sentence level and Aspect level [6]. Classification of the whole document in positive and negative class based on the sentiment and emotions present on document is known as document level classification. Although Document level sentiment classification is fast but It hide the useful information from user. Sentence level sentiment classification works on sentence level. It classifies the sentences in positive or negative classes. Aspect based sentiment classification uses and aspects and option words in order to classify. Aspect level sentiment classification based on idea that In every sentence there is a opinion word which could be positive, negative or neutral. Aspect level sentiment classification is the most effective types provides better results from other two types of sentiment classification.

Machine learning and Lexicon based methods are widely used for sentiment classification. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data [7]. Machine learning algorithms are further divided in two types that are supervised learning and unsupervised learning. In supervised learning training is necessary. Output dataset is provided to train the algorithm and get the desire results [8] On the other hand unsupervised learning cannot train the algorithm and data is clustered in different clusters [9]. Some Supervised learning algorithms are Naive bayes, Support Vector Machine (SVM) etc and Unsupervised learning algorithms are K-Means, K-Median etc Lexicon based techniques are simple and more fast. These techniques are divided in two types, Dictionary based approaches and corpus-based approach. Sentiworldnet and Wordnet are most popular lexicon based approaches. Types of Machine learning and Lexicon based approaches are shown in Figure 1. We apply a Naive Bayes classifier on movies reviews to judge the sentiments of movie reviews. A Naive Bayes classifier is one of the most efficient algorithm comes from supervised machine learning techniques [10][11].

The rest of paper is organized as follows. Section II introduces the background study. Section III illustrates the proposed system description for Sentiment classification. System includes four major steps Dataset, Preprocessing, Feature extraction and classifier. Section IV shows the results. Conclusions are addressed in section V and then Section VI

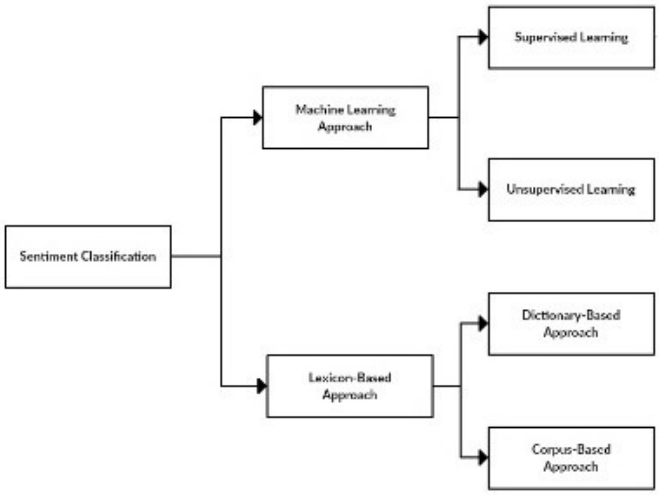

Fig. 1: Sentiment Classification Approaches

is for references.

II. BACKGROUND

In the last decade, many works has been released in sentiment classification. Implementation of sentiment classification (judging tone of text) has been carried out for a variety of applications over a wide range of classification algorithms and for varying data size. There exist many possible variants; some of them are discussed in following section.

Hangya and Farkas (2013) introduce an approach for analyzing the textual content of tweets and classify it in positive, negative and neutral classes using Logistic Regression classifier [12]. Paper uses the SentiWordNet sentiment lexicon for calculating the polarity of a word. An experimental result shows that paper achieves the accuracy of 69% on large data set of twitter tweets. Another paper uses neural network with Naive Bayes classifier because in many complex real world situations Naive Bayes cannot work well [13]. An experimental result shows that, when we are using Naive Bayes Classifier then Accuracy will be 62.35. So for getting better results Paper uses the Neural Network with Naive Bayes Classifier. Results show that accuracy of sentimental analysis increased up to 80.65 by combining Neural Network with Naive Bayes Classifier. In 2010 Ting-Chun Peng and Chia-Chun Shih purposed the hybrid unsupervised learning technique for extracting sentiments from reviews by rules pf parts of speech (POS) patterns [14]. K- nearest neighbor (KNN) is a typical classifier produces a good results in many cases. Songho tan uses the KNN classifier for classification the document in positive, negative or neutral [15]. KNN works well be calculating the weights of sentiments. Liu et al (2013) argued the sentiment classification system that uses Nave Bayes Classifier and Map Reduce framework [16]. Paper uses machine leaning algorithm Nave Bayes Classifier to classify the sentiments in two classes positive and negative and results shows that paper gets the accuracy of 83%.

III. SYSTEM DESCRIPTION

System description includes the components which are used in system setup. We design the system based on our needs that classify the movie reviews in positive or negative class. It also calculates the polarity of sentiments in percentage. There are basically four components of system. Dataset, Preprocessing, Feature Extraction and classifier. These components are shown in Fig 1.

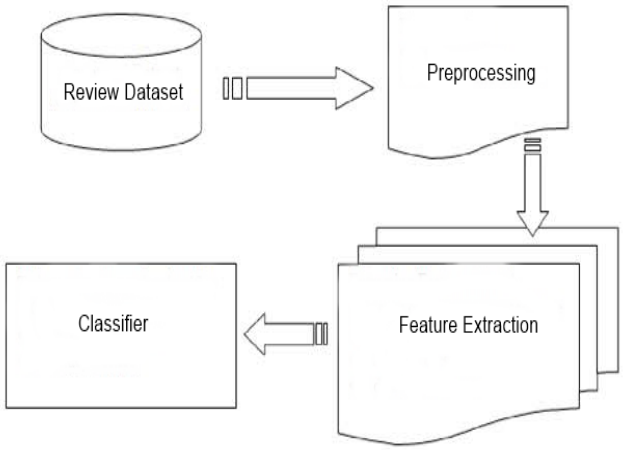

Fig. 2: Proposed Model

A. Dataset

In our experiment, we use the standard data set which is collected by research communities and used by many researchers. This dataset is the Cornell University movie review dataset. It contains 700 positive and 700 negative movie reviews. Our technique runs on this dataset and results is calculated using this dataset.

B. Preprocessing

Preprocessing is the essential step of any experiment. Without efficient preprocessing we cannot achieve our desire goals. Preprocessing converts all dataset in common format [17]. Preprocessing contains four different steps. First step is the removal of unnecessary data like special characters. The second step is to normalize the long words that may have some letters redundant [18]. Normalization also removes the multiple words in single word.

C. Feature Extraction

Machine learning provides many algorithms that work for classification, but the challenge of finding the sentiment in a text is determined the best feature to be used. Before using a classifier on the data, we need to represent the text in a format suitable for the classifier to deal with it. In Natural Language Processing, the popular model is the vector model or feature model. The text, either document or sentence, will be converted into the form of the features model before the training process of the classifier starts. This model should preserve essential information about the text. Each row of the model represents one of the data set records (either document or sentence). Each column displays the features that are chosen to build the vector model. The intersection of each row with each column contains a value that represents the relation of that feature in that data record. We break the sentence into its basic parts of speech (POS tagging).

POS tagging is a mature technology that identifies all the structural elements of a document or sentence, including verbs, nouns, adjectives, adverbs, etc. [19] it is important to find the adjectives, as these are good indicators conveying the sentiment orientation in the text. The POS tag feature will help to determine the correct meaning of the word. Turney (2002) also used the POS feature for adjectives and adverbs in order to obtain the sentiment orientation at document level [20].

D. Classification

There are many machine leaning classification techniques but Naive Bayes algorithm is the most efficient and simple algorithm. Naive Bayes algorithm is the simple probabilistic classifier which works on probability. The Naive Bayes classifier has been used in the document classification problem for decades. The Bayes Rule is the integral part of all Bayesian Models. This rule can be calculated as follows:

$$P(H/E) = P(E/H)P(H)/P(E) \quad (1)$$

where P(H—E) is the posterior probability of the hypothesis, P(H) is the prior probability of hypothesis, P(E) is the prior probability of Evidence, and P(E—H) is the conditional probability of Evidence given Hypothesis (likelihood). In order to obtain the value of word polarity, a lexical resource, such as SentiWordNet, is needed. This lexicon is constructed from the perspective of WordNet to which each synset, which is a set of one group of synonyms, is assigned three sentiment scores: positivity, negativity, and objectivity.

After classifying the movie reviews in positive and negative class we then calculate the value of positive and negative sentiments in the movie review. Value of sentiments is calculated in percentage. Many researchers named this technique as Polarity calculations. Sentiment polarity classification is presumably the most well studied subtask in sentiment analysis. Nonetheless, it is difficult to provide a single, clear-cut definition. Problem settings which are subsumed under the umbrella term ”polarity classification” are quite heterogeneous. Depending on the application domain and even more on the examined dataset, researchers and practitioners set out varying goals. The primary factor is that sentiment polarity often cannot be determined without considering the context. We know that sentiment shifters may alter the sentiment status of individual expressions, entire sentences, or even the whole document. After classification positive and negative sentiments are calculated of each sentence in percentage.

IV. RESULTS

This section illustrates the experiments that are done to investigate and test the features and performance of the Naive Bayes classifier on movie reviews. The accuracy of the classifier is measured as the number of instances correctly classified over total number of instances. Nave Bayes classifier correctly classifies 1216 movie reviews out of 1400 which is 86.8%. On close look Naive Bayes classifier classify 612 out of 700 which is 87.42 and True negatives of nave Bayes are 604 out of which are 86.2. These results are also showing in tabular form is table 1 and table 2

TABLE I. POSTIVE REVIEWS

Reviews			
Total Positive	*Correctly classify*	*Remaining*	*Percentage*
700	612	88	87.62

TABLE II. NEGATIVE REVIEWS

Reviews			
Total Negative	*Correctly classify*	*Remaining*	*Percentage*
700	604	96	86.2

We plot a movie reviews on Y-axis and total positive and negative reviews on X-axis. The graph is showing in Fig 3.

Accuracy table is showing in Table III. Total reviews, correctly classified sentiments, Incorrectly classified sentiments and its accuracy is showing

Accuracy graph is shown in Fig 4. Where Accuracy is plotted on Y-axis and correctly classified and incorrectly classified movie reviews are plot on Y-axis.

V. CONCULSION

Sentiment Classification emerges as a challenging field with lots of obstacles as it involves natural language processing and hidden emotions. It has a wide variety of applications that could benefit from its results, such as movie reviews,

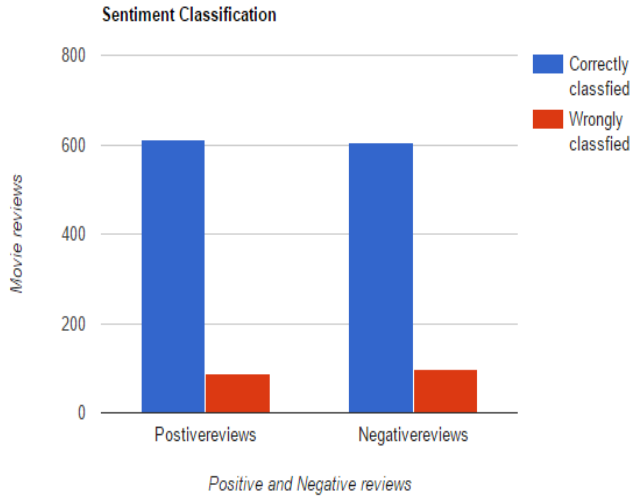

Fig. 3

TABLE III. ACCURACY TABLE

Reviews			
Total Reviews	*Correctly classify*	*Incorrectly classify*	*Percentage*
1400	1216	184	86.8%

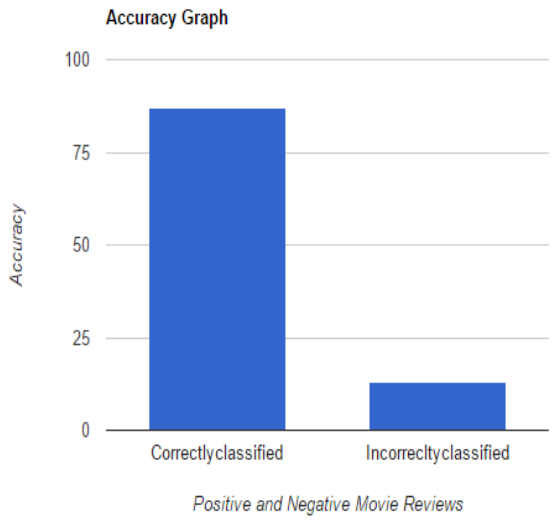

Fig. 4

product reviews, news analytics, and marketing, question answering, knowledge bases and so on. We successfully implement the Naive Bayes classifier on movie reviews and also we calculate polarity of each movie reviews. Binary classification is done in two classes: positive and negative. An experimental result shows that our work gets a accuracy of 86.8%.

In future, more work is needed on further improving the accuracy and performance measures. Todays sentiment classification involving positive and negative classes, a natural extension to this work is subjectivity detection whereby a piece of text is classified as objective or subjective class.

REFERENCES

[1] Flacy. M. "Nearly 300,000 status updates are posted to Facebook every minute." Digital Trends, 2011.
[2] I. Guellil, K. Boukhalfa, Social Big Data Mining : A Survey Focused on Opinion Mining and Sentiments Analysis." 12th International Symposium on Programming and Systems (ISPS), 2015.
[3] H. Sinha, A. Kaur. " A Detailed Survey and Comparative Study of Sentiment Analysis Algorithms." IEEE 2nd International Conference on Communication, Control and Intelligent Systems (CCIS) ,2016.
[4] G. Patil,V. Galande , V. Kekan and K. Dange, . Sentiment Analysis Using Support Vector Machine International Journal of Innovative Research in Computer and Communication Engineering, 2014, pp. 1021-1026..
[5] J. Steinberger, T.Brychcin and M. Konkol. Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. 2014.
[6] Manek, A., P. Deepa ,C. Mohan and K.Venugopal. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier, Springer,2016.
[7] K.T.Devendra and S. K. Yadav, Fast retrieval approach of sentimental analysis with implementation of bloom filter on Hadoop, International Conference on Computational Techniques in Information and Communication Technologies, 2016, pp. 529-551.
[8] O. Frunza, D. Inkpen and T. Tran, "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts", IEEE Transactions on Knowledge and Data Engineering, 2010, pp: 801-814.
[9] P. Chandrasekar, K. Qian, "The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier", IEEE 40th Annual Computer Software and Applications Conference, 2016, pp: 618-619.
[10] C. Troussas ,A. Krouska and M. Virvou, " Evaluation of Ensemble-based Sentiment Classifiers for Twitter Data", 7th International Conference on Information, Intelligence, Systems & Applications (IISA), 2016.
[11] A. S. Tewari; T. S. Ansar and A. G. Barman, "Opinion based book recommendation using Naive Bayes classifier", Contemporary Computing and Informatics (IC3I), 2014, pp: 139-144 .
[12] V. Hangya and R. Farkas, " Target-Oriented Opinion Mining from Tweets", 4th IEEE International Conference on Cognitive Info communications, 2013, pp: 251-254.
[13] L. L. Dhande, and G. K. Patnaik, Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), vol 3, 2014, pp. 313-320.
[14] T. C. Peng and C. C. Shih, An Unsupervised Snippet-based Sentiment Classification Method for Chinese Unknown Phrases without using Reference Word Pairs, IEEE/WIC/ACM International Conference on Web Intelligence and ntelligent Agent Technology, 2010.
[15] Songbo Tan and Jin Zhang, An empirical study of sentiment analysis for chinese documents , Expert Systems with Applications, 2008, pp: 26222629.

[16] Liu, B., E.Blasch, Y.Chen, D.Shen and G.Chen. Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier. IEEE International Conference on Big Data,2013, pp: 99-104.

[17] Banage T. G. S. Kumara, Incheon Paik, Jia Zhang, T. H. A. S. Siriweera and Koswatte R. C. Koswatte, Ontology-Based Workflow Generation for Intelligent Big Data Analytics, IEEE International Conference on Web Services, 2015.

[18] Viktor Hangya, Richard Farkas, Target-Oriented Opinion Mining from Tweets, 4th IEEE International Conference on Cognitive Infocommunications, 2013.

[19] Zhang Youzhi, "Research and implementation of part-of-speech tagging based on Hidden Markov Model",Computational Intelligence and Industrial Applications, 2009.

[20] Peter D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews", 40th Annual Meeting on Association for Computational Linguistics, 2002, pp: 417-424.

AUTHORS

Muhammad Rizwan Rashid Rana,
rizwanrana315@gmail.com
University Institute of Information Technology (UIIT)
PMAS-Arid Agriculture University
Rawalpindi, Pakistan

Muhammad Aun Akbar, aunakbar@gmail.com
University Institute of Information Technology (UIIT)
PMAS-Arid Agriculture University
Rawalpindi, Pakistan

Tauqeer Ahmad, Tauqeer.ahmaed443@gmail.com
University Institute of Information Technology (UIIT)
PMAS-Arid Agriculture University
Rawalpindi, Pakistan