Review of Classifiers for Automated Opinion Mining

Shina
Computer Science and Engineering
Assistant Professor, Chitkara University,
Punjab, India

Navpreet Rupal
Computer Science and Engineering
Assistant Professor, SUSCET,
Punjab, India

ABSTRACT
Opinion Mining has been a field of great interest and use as it helps the producers to know about the reviews of the product so as to enhance their sales. At the same time it helps the consumers as well to judge which product suits their requirements and whether the overall feedback from other consumers who have used it is positive or negative. This research proposes a system which aims to help the customers deal with bulk of reviews in an easier manner. The research also comprehends the optimum classifier to judge the contextual polarity of the review.

Keywords
Opinion Mining, classifier, contextual polarity

1. INTRODUCTION
Data Mining has always enabled to discover the unknown patterns that have helped to formulate policies to help take better decisions. This field has numerous applications in the real world that have made the judgment of scenarios straightforward. One of the most common uses of data mining is the classification used for spam filtering, movie recommendation systems, firewall implementation etc. One of the most studied fields is opinion mining where the textual material is classified as positive or negative. Opinion Mining deals with the classification of opinions in textual form into the two categories i.e. "positive" or "negative". The basic applications of sentiment analysis application would be to track the word of mouth or the voice of the customer like what they have to say about a particular service or commodity like salons, cameras or movies. The richest content and knowledge nowadays is available through blogs and forums. Each day a humongous number of blogs are added where the people discuss the topics of common interest. Opinion Mining is used to analyze the attitude of the writer or to the service or commodity written about. Nowadays it is a vital ingredient of social media analytics, it processes the text written by the several users to recognize and figure out the feelings of the consumer based on online conversations that include: blogging sites, discussion forums, Facebook statuses, tweets etc. Opinion Mining classifies the polarity (positive or negative) of a post, comment, or statement which determines whether sentiment around a topic is positive, negative. To judge the polarity it employs both supervised and unsupervised approaches. But the data over which the mining is to be employed is not easy to fetch. But the biggest setback lies in the structure of these blogs. The data on these blogs is unstructured and cannot be used when required for help. The enormous quantity of data becomes futile if it cannot be used for generation of some constructive information. The data over the websites is generally in the unstructured format. The layout of one website may be different from another so a web scraper is needed to extract data from the multiple sources. Here the field of web crawling comes into play. To extract the useful part of the information on blogs and discard the unsolicited one can be done by crawling the web and scraping the required data. Then the raw facts and figures are processed to generate the valuable information. The obtained information is further classified by the classifier and outcome regarding the polarity of the review is inferred.

The remaining of the paper is organized as follows. In Section II, we discuss the previous works done. Section III discusses the proposed system. The conclusion is given in Section IV.

2. RELEVANT WORK
Several works have been done in the field of opinion mining which have offered us with an outsized amount of information. Mureşan et al. (2013) [1] evaluated two approaches for predicting the sentiment polarity of an utterance. The first method was based on a 3-dimensional model which takes into account text expressiveness in terms of valence, arousal and dominance. The second one determined the word's semantic orientation according to Chi-square and Relevance factor statistic metrics. Chaovalit and Zhou (2005) [2] investigated movie review mining using two approaches: machine learning and semantic orientation. The approaches were adapted to movie review domain for comparison. The results show that the research results were comparable to or even better than previous findings. A Baloglu et al. (2010) [3] introduced an architecture, implementation, and evaluation of a Web blog mining application, called the BlogMiner, which extracts and classifies people's opinions and emotions (or sentiment) from the contents of weblogs about movie reviews. The Sentiment analyzer calculates the sentiment scores for a movie for different keywords by mining the comments from blog pages. Singh et al. (2013) [4] implemented SentiWordNet approach with different variations of linguistic features, scoring schemes and aggregation thresholds. They used two pre-existing large datasets of Movie Reviews and two Blog post datasets on revolutionary changes in Libya and Tunisia. Banic et al. (2013) [5] proposed a system that collected opinions about hotels from the web, evaluated them, aggregated evaluations and offers cumulative, easy-to-understand information. Four different categories were introduced: *location*, *service*, *atmosphere* and *general.* Every term was categorized before it was graded with marks in the range from 1 to 5. For every review all four categories were specified separately. The total grade for every hotel was determined as average grade of all reviews aggregated by category on the hotel level. Dehkharghani and Yilmaz (2013) [6] studied the application of sentiment analysis on extracting the quality attributes of a software product based on the opinions of end users that have been stated in micro blogs such as Twitter. Their findings obtain advantageous techniques such as document frequency of words in a large number of tweets. Pang and Lillian (2008) [7] surveyed the techniques and approaches that promise to directly enable opinion-oriented information-seeking systems. Their focus was on methods that seek to address the new challenges raised by sentiment-aware applications, as compared to those that were already present in more traditional fact-based analysis. Jiang et al. (2010) [8] presented an approach based on tree kernels for opinion mining

of online product reviews. In the research, they defined several tree kernels for sentiment expression extraction and sentiment classification, which are subtasks of opinion mining. Experimental results on a benchmark data set indicated that tree kernels can significantly improve the performance of both sentiment expression extraction and sentiment classification. Liu (2012) [9] proposed a novel approach based on latent semantic analysis (LSA) to identify product features. Furthermore, they found away to reduce the size of summary based on the product features obtained from LSA. Mouthami (2013) [10] automated the task of classifying a single topic textual review; document-level sentiment classification is used for expressing a positive or negative sentiment. So analyzing sentiment using Multi-theme document was very difficult and the accuracy in the classification was less. The document level classification approximately classifies the sentiment using Bag of words in Support Vector Machine (SVM) algorithm.

3. PROPOSED SYSTEM

The proposed architecture consists of five steps that will be used to generate an automated opinion mining system.

3.1 OBJECTIVES

The rise of social media (such as online web forums and social networking sites) has attracted interests to mining and analyzing opinions available on the web. The online opinion has become the object of studies in many research areas; especially that called "Opinion Mining and Sentiment Analysis".

1. The foremost objective of the research is to accumulate the data from several web sources like movie portals and other movie review websites like rottentomatoes.com by web crawling the websites. Just the collection of data will not be sufficient; the quality of the data has to be improved by cleaning the data and modifying i.e. formatting the data as per the requirements of the classifier.

2. After the data has been collected, the data shall be used to train the several classifiers used to classify the input as a binary value. The records used to train the classifier will create a database or a repository of data which will be used to accurately classify the data by the means of its artificially generated intelligence.

3. The last and the most important objective will be to study and analysis of the algorithms that are used for the classification. The classified data's accuracy in terms of several performance parameters will be analyzed and eventually the optimum methodology for deducing the output or classification of the data is comprehended.

3.2 METHODOLOGY

The proposed system comprises of tasks like data collection, data cleaning, data formatting, designing the workflow of elements, training, setting threshold parameters, testing of classifiers and comparing and analyzing the several classifying techniques. Hence the entire work been sub divided in to five subtasks on the basis of the similarity of tasks. For the ease of understanding the below given figure 1 has been portrayed to give a detailed idea of each of the step.

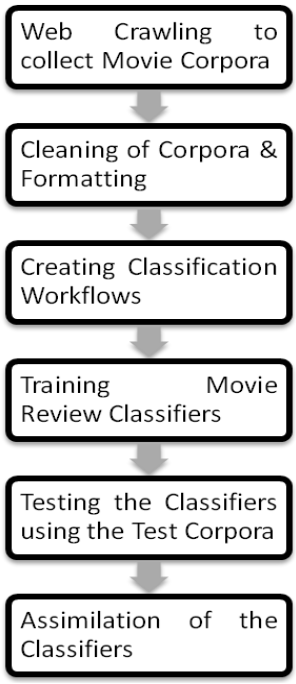

Figure 1

The first phase deals with the **collection of the data** which will be used for training the classifiers as well as a different set of data that would be used for the testing of the trained classifier. The data shall be collected from the several discussion forums, movie groups and social blogs.

The second phase is the **Preprocessing Stage**, where the collected data will be filtered and set in the format as required in the process in order to train the classifiers. The data filtering is a very essential phase because the irrelevant data may give inaccurate results while mining the data. We will use Review Corpora which are different in many aspects.

The third phase will be the **training of the classifiers** individually with the help of the training dataset and the manipulations of the several parameters in order to get the best results from each of the classifier independently. These threshold values play a vital role in the research because unless the appropriate value is set the results of the classifiers may not be the best it can deliver.

The fourth phase deals with **testing of the classifiers** by feeding in to them the Test Dataset and obtain the results from the classifiers as per the training given to the classifiers. The data that was fed in to the classifier as training data has created a repository of words along with its weight. So when a data set with textual content is entered the result generated by the classifier is purely on the basis of the training given to the classifier.

The last phase is the analyses and **comparison** of all the classifiers on the basis of the performance metric like Accuracy, Recall, Precision and F-Measure. We investigate some settings to identify those that allow achieving the best results.

3.3 CLASSIFIERS TO BE USED

In data mining in order to classify a test case there are several classifiers available, but some of the most popular algorithms are Naïve Bayesian Classifier, Support Vector Machine Classifier, Decision Tree Classifier and the k- Nearest Neighbor Classifier [13].

The Naïve Bayesian classifier works on the probability model that is created on the basis of the training data set. The factors like mean and variance of the attributes are calculated on the basis of which the input test class is evaluated and classified in to any one of the pre defined classes or categories.

The SVM classifier generates a clear demarcation between the categories of the output by the means of a hyper plane. The values of the attributes are assumed to be points in the space which are separated by a hyper plane and more clearly the demarcation, the better is the efficiency and accuracy of the classifier. It is a binary classifier that is non-probabilistic in nature.

The decision tree creates a tree like model based on the previous learning given to the classifier by the means of the training data set. The judgment for a test case can be done by the traversal of the tree and traversing the nodes that meet the conditions of the test case. Thus the prediction assigns the test case to the predefined class or category.

The k Nearest Neighbor algorithm uses the coherence feature to adjudge the test case to the predefined class. The features or attributes of the test case are matched to its neighboring cases or the similar cases and the test case is classified. The value of 'k' plays a vital role in the classification process because the accordingly the 'k' nearest neighbors are checked where value of 'k' is a positive integer. The class to which majority of the 'k' neighbors belong to, the test case is also put into the same class.

In the proposed research the above mentioned classifiers will be comprehended and the optimum classifier shall be adjudged.

3.4 OPINION THESAURAUS

The research aims at developing a opinion thesaurus that has the word along with its polarity and the strength of the polarity along. The score of the sentiment has been already evaluated by several researches but the limited number of words in the thesaurus leads to a limited result in terms of actual accuracy. Several opinion thesauruses are available SentiWordNet 3.0 [11], Sentic Net 2[12]. Our system shall amalgamate the available thesauruses and develop a humungous pool of words along with their polarity and sentiment score.

3.5 PERFORMANCE METRICS

In order to judge the optimum classifiers out of those used, the following metrics will be used to evaluate the performance

- True positive = Reviews that have been correctly classified as a positive review
- False positive = Reviews that have been incorrectly classified as a positive review
- True negative = Reviews that have been correctly classified as a negative review
- False negative = Reviews that have been incorrectly classified as a negative review
- Sensitivity = Ability of the classifier to classify a review correctly
- Specificity = Ability of the classifier to exclude a review correctly
- F- measure = measure of the classifier's accuracy that will consider both the precision of the classifier and the recall as well.

4. CONCLUSION

In this paper, we proposed a model that we will be implementing which will be useful to mine the opinions given in the online reviews automatically by the means of the supervised learning technique. The classifier shall be trained using a large pool of onion reviews that will be collected by web scraping and the opinion thesaurus to be used will also be a amalgamation of the individually existing thesauruses. We shall perform the research and comprehend the optimum classifier for the purpose of online reviews classification.

5. REFERENCES

[1] Muresan, I. Stan, A.; Giurgiu, M. ; Potolea, R. 2013 "Evaluation of Sentiment Polarity Prediction using a Dimensional and a Categorical Approach" 7th Conference on Speech Technology and Human - Computer Dialogue (SpeD), pp 1 – 6, DOI 10.1109/SpeD.2013.6682645

[2] Pimwadee Chaovalit; Lina Zhao 2005 "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches" Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05), pp 112.3, DOI 10.1109/HICSS.2005.445

[3] Baloglu, A.; Aktas, M.S. 2010 "BlogMiner: Web Blog Mining Application for Classification of Movie Reviews" Fifth International Conference on Internet and Web Applications and Services (ICIW), pp 77 – 84, DOI 10.1109/ICIW.2010.19

[4] Singh, V.K.; Piryani, R. ; Uddin, A. ; Waila, P. 2013 "Sentiment Analysis of Movie Reviews and Blog Posts, Evaluating SentiWordNet with different Linguistic Features and Scoring Schemes" 2013 IEEE 3rd International Advance Computing Conference (IACC), pp 893 – 898 , DOI 10.1109/IAdCC.2013.6514345

[5] Banic, L. ; Mihanovic, A. ; Brakus, M. 2013 "Using Big Data and Sentiment Analysis in Product Evaluation" 36th International Convention on Information & Communication Technology Electronics & Microelectronics (MIPRO), pp 1149 – 1154

[6] Dehkharghani, R.; Yilmaz, C. 2013 "Automatically Identifying a Software Product's Quality Attributes through Sentiment Analysis of Tweets" 1st International Workshop on Natural Language Analysis in Software Engineering (NaturaLiSE), pp 25 – 30, DOI 10.1109/NAturaLiSE.2013.6611717

[7] Pang B; Lee L 2008 "Opinion Mining and Sentiment Analysis" Foundations and trends in information retrieval, pp Pages 1-135, DOI 10.1561/1500000011

[8] Peng Jiang ; Chunxia Zhang ; Hongping Fu ; Zhendong Niu 2010 "An Approach Based on Tree Kernels for Opinion Mining of Online Product Reviews" IEEE 10th International Conference on Data Mining (ICDM) , pp 256 – 265 DOI 10.1109/ICDM.2010.104

[9] Chien-Liang Liu ; Wen-Hoar Hsaio ; Chia-Hoang Lee ; Gen-Chi Lu ; Jou, E. 2012 "Movie Rating and Review Summarization in Mobile Environment" IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, pp 397 – 407, DOI 10.1109/TSMCC.2011.2136334

[10] Mouthami, K. ; Dept. of CSE, Kongu Eng. Coll., Erode, India ; Devi, K.N. ; Bhaskaran, V.M. 2013 "Sentiment Analysis and Classification Based On Textual Reviews" International Conference on Information Communication

and Embedded Systems (ICICES), pp 271 – 276, DOI 10.1109/ICICES.2013.6508366

[11] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani 2010 "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining" Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)

[12] Cambria E.; C. Havasi C.; Hussain A. 2012 "SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis" AAAI FLAIRS, Marco Island, pp. 202-207

[13] http://www.cs.put.poznan.pl/jstefanowski/sed/DM-4-evaluatingclassifiersnew.pdf