International Journal of Advance Research in Science and Engineering
Vol. No. 5, Issue No. 08 , August 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

SENTIMENT ANALYSIS BY ASSOCIATING MODIFIED K MEANS WITH NAÏVE BAYES CLASSIFICATION AND SUPPORT VECTOR MACHINE

Shivani Rana

Department of Computer Science and Engineering, Hindu College of Engineering, Sonipat, Haryana, (India)

ABSTRACT

Sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining. The inception and rapid growth of the field coincide with those of the social media on the Web, e.g., reviews, forum discussions, blogs, micro blogs, Twitter, and social networks, because for the first time in human history, we have a huge volume of opinionated data recorded in digital forms. Since early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing. It is also widely studied in data mining, Web mining, and text mining. In fact, it has spread from computer science to management sciences and social sciences due to its importance to business and society as a whole. In recent years, industrial activities surrounding sentiment analysis have also thrived. Numerous startups have emerged. Many large corporations have built their own in-house capabilities. Sentiment analysis systems have found their applications in almost every business and social domain. The goal of this paper is to give an introduction to this fascinating problem and to present a framework which perform sentiment analysis on online mobile phone reviews by associating modified K means algorithm with Naïve bayes classification and Support vector machine.

Keywords: Modified K-Means Algorithm, Naïve Bayes Classification, Sentiment Analysis , Support Vector Machine.

I. INTRODUCTION

Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product or topic. Sentiment analysis, which is also called opinion mining, involves in building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets. Sentiment analysis can be useful in several ways. For example, in marketing it helps in judging the success of an ad

campaign or new product launch, determine which versions of a product or service are popular and even identify which demographics like or dislike particular features.

There are several challenges in Sentiment analysis. The first is an opinion word that is considered to be positive in one situation may be considered negative in another situation. A second challenge is that people don't always express opinions in a same way. Most traditional text processing relies on the fact that small differences between two pieces of text don't change the meaning very much. In Sentiment analysis, however, "the picture was great" is very different from "the picture was not great". People can be contradictory in their statements. Most reviews will have both positive and negative comments, which is somewhat manageable by analyzing sentences one at a time. However, in the more informal medium like twitter or blogs, the more likely people are to combine different opinions in the same sentence which is easy for a human to understand, but more difficult for a computer to parse. Sometimes even other people have difficulty understanding what someone thought based on a short piece of text because it lacks context. For example, "That movie was as good as its last movie" is entirely dependent on what the person expressing the opinion thought of the previous model. Sentiment analysis concentrates on attitudes, whereas traditional text mining focuses on the analysis of facts. There are few main fields of research predominate in Sentiment analysis: sentiment classification, feature based Sentiment classification and opinion summarization. Sentiment classification deals with classifying entire documents according to the opinions towards certain objects. Feature-based Sentiment classification on the other hand considers the opinions on features of certain objects. Opinion summarization task is different from traditional text summarization because only the features of the product are mined on which the customers have expressed their opinions. Opinion summarization does not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as in the classic text summarization. Languages that have been studied mostly are English and in Chinese .Presently, there are very few researches conducted on sentiment classification for other languages like Arabic, Italian and Thai.

II. RELATED WORK

The approaches used in sentiment analysis include TF*PDF algorithm, Support Vector Machine, F-Measure, EFS algorithm. Sentiment analysis is basically expressed as sentiment of the individuals. Currently, Sentiment Analysis concentrates for subjective statements or on subjectivity and overlook objective statements which carry sentiment(s). Jalaj et al.[1] discussed about exiting methods, approaches to do sentimental analysis for unstructured data which reside on web here, they proposed new approach to classify and handle subjective as well as objective statements for sentimental. K. Bun et al.[2] carried out work on topic extraction from news archive using TF*PDF algorithm. There exist prominent topics that were discussed frequently in many documents from many newswire sources. TF*PDF algorithm recognizes the terms that try to explain the main prominent topics. These would be the terms that appear frequently in many documents from many newswire sources concurrently. TF*PDF algorithm assigns heavy term weight to these kind of terms and thus reveal the main topics. Zhu et al.[3] have suggested R-TF-IDF algorithm which is an enhancement over TF-IDF. Here they multiplied the TF-IDF formula with an adjusting factor. This factor increase the importance of term frequency in a document and discard the terms that appear less frequently in a document whereas have relatively

higher term frequency weighting. Mukhrjee et al.[4] performed work on techniques F-measure and EFS algorithm. F-measure explores the notion of implicitness of text and is a unitary measure of text relative contextuality and formality. Contextuality and formality can be captured by certain part of speech.EFS Algorithm: EFS takes the best of both worlds. It first uses a number of feature selection criteria to rank the features following the filter model. Upon ranking, the algorithm generates some candidate feature subsets which are used to find the final feature set based on classification accuracy using the wrapper model. Chen et al.[5] developed a semantic based information retrieval model for blog. As lack of semantic for information description and semantic support for the query processing, traditional blog systems are unable to satisfying users in the performance of information organization and retrieval. In order to implement semantic description for blog information resource, they designed a blog ontology and a domain subject classification ontology. With predefined rules, they put forward a class hierarchy tree generating algorithm, expand blog classification item semantic retrieval, and finally implement semantic retrieval by SPARQL query. Savoy et al.[6] described a new and simple classification scheme based on specific vocabulary using term's Z score values (*n*-gram of characters, words or lemmas).The suggested scheme determines the values of those terms specific to a given subset compared to the entire corpus. The resulting Z score values reflect the differences between the expected occurrence frequencies and those observed. When a term has a large positive Z score, it belongs to the specific vocabulary, while a large negative Z score indicates that the term is underused. Popowich et al.[7] developed an application. The application makes use of a natural language processing (NLP) engine, together with application-specific knowledge, written in a concept specification language. Using NLP techniques, the entities and relationships that act as indicators of recoverable claims are mined from management notes, call centre logs and patient records to identify medical claims that require further investigation. Text mining techniques can then be applied to find dependencies between different entities, and to combine indicators to provide scores to individual claims. Claims are scored to determine whether they involve potential fraud or abuse, or to determine whether claims should be paid by or in conjunction with other insurers or organizations. Dependencies between claims and other records can then be combined to create cases. Issues related to the design of the application are discussed, specifically the use of rule-based techniques which provide a capability for deeper analysis than traditionally found in statistical techniques.

III. PROPOSED APPROACH

Figure 1 shows the proposed framework of our approach which has four modules: Data extraction, Pre-processing, Clustering and Classification. In this proposed approach, number of steps are used conceptualize, design and perform an effective sentiment analysis of online mobile phone reviews that is to be achieved by associating modified K means algorithm with Naïve Bayes Classification and Support Vector Machine and to evaluate which technique is more accurate for sentiment analysis.

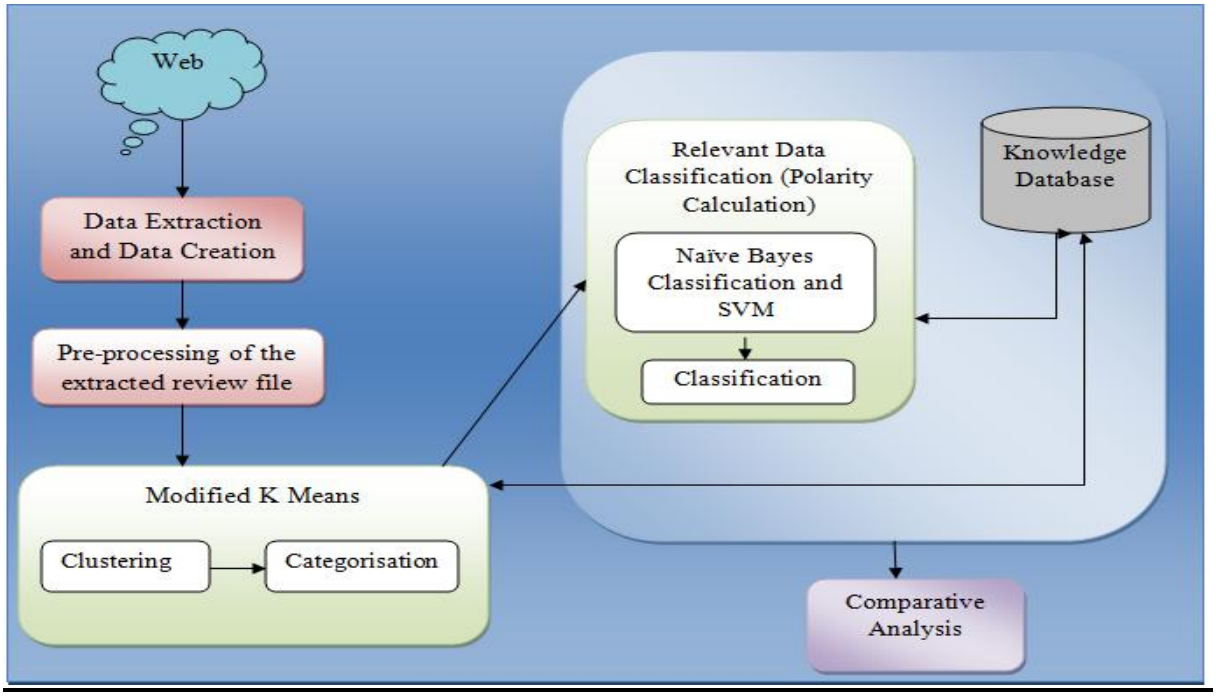

Figure 1: Proposed Framework

The procedure of sentiment analysis has following steps:

1.First step: First extract the data to be analyzed from the web. In our work we have extracted data from twitter mobile phone database.

2.Second step: For preprocessing and polarity calculation of the data we have created a training dataset for positive, negative, average sentiment words and stop words in SQLyog.

3.Third step: Preprocessing- In the pre-processing of the data the words which are not carrying any sentiments or opinion are removed from the data. Another task performed in pre-processing is stemming. It is the process of reducing derived words into their root forms e.g. word happiness is reduced into root form happy. So, after pre-processing we get only the meaningful data on which we can easily apply the techniques.

4.Fourth Step: "Calculate Polarity" provides us the count of positive, negative and average sentiment words in the entered data which is used by the techniques as an input for further processing.

5.Fifth Step: Clustering is the process of forming the clusters of the data, objects within a cluster have similar properties, but they are not similar to objects in other clusters. Modified K-Means algorithm has been used to implement clustering. Problem of empty cluster generation is avoided by the modified k means algorithm. Here, the working structure of modified k means is same as that of original k-means, keeping all its required characteristics intact. It introduces a new center vector computation theory which is different from the previous method used by original k means.

6.Sixth step: Classification is done using Naïve Bayes Classification and Support Vector Machine. Naïve Bayes Classification is based on supervised learning. It is a statistical method for classification. It computes the probabilities of the outcomes to determine whether a sample belongs to a particular class or not. It is used for both diagnostic and predictive problems.
Support Vector Machine is based on supervised learning. It has associated learning algorithms which is used for performing tasks such as data analyses, pattern recognition and is used for classification and regression analysis. SVM model represents examples as points in space, mapped so that the examples of the different categories are classified by a wide gap which is as large as possible.

The framework has two phases for Sentiment Analysis. First the training data set of various sentiment words in SQLyog 5.14. Second is an interface in NetBeans IDE 7.3.1 which is used for performing testing data, clustering and classification.

This approach uses JAVA for implementation. First a review is entered and then pre-processing of the data is done so as to remove all the meaningless data. It helps in real time sentiment analysis by reducing the noise of the data and improves the classifier performance and increase the speed of the classification process. In the second step feature extraction using modified K-means is done to make classifiers working effective; amount of data to be investigated is reduced as well as relevant features are known which are useful in classification process. In the last phase classification techniques have been applied Naïve Bayes Classification and Support Vector Machine. Both the techniques gives result in the form of a probability function.

IV. CONCLUSION

This study examined the performance of two advance machine learning techniques Naïve Bayes and SVM in case of sentiment analysis of online mobile phone reviews. Comparison of performance of both the techniques is done on five mobile phone review datasets. The results show probability function value of SVM is greater than probability function value of Naïve Bayes in all the five datasets. With the help of confusion matrix parameters TP rate, accuracy, FP rate and error rate, it is evaluated that TP rate and accuracy of SVM is greater to that of Naïve Bayes across all the five datasets. The FP rate and error rate of SVM is less than Naïve Bayes in all the five datasets. So, with this work it is concluded that Support Vector Machine performs better than Naïve Bayes in case of sentiment analysis of online mobile phone reviews.

REFERENCES

[1] Jalaj S. Modha, Prof & Head Gayatri S. Pandi and Sandip J. Modha, "Automatic Sentiment Analysis for Unstructured Data", Volume 3, Issue 12, December 2013 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[2] K. Bun and M. Ishizuka, "Topic extraction from news archive using TF*PDF algorithm" In Proceedings of Third International Conference on Web Information System Engineering.

[3] Dengya Zhu, and Jitian XIAO,"R-tfidf, a Variety of tf-idf Term Weighting Strategy in Document Categorization" published in 2011 Seventh International Conference on Semantics, Knowledge and Grids.

[4] Mukhrjee, A. and B. Liu, 2010, "Improving gender classification of weblog authors. Proceedings of Conference on Empirical Methods in Natural Language Processing", (EMNLP' 10), 10RDF Primer. W3C Recommendation . http://www.w3.org/TR/rdf-primer, 2004.

[5] Ying Chen, Wenping Guo, Xiaoming Zhao, "A semantic Based Information Retrieval Model for Blog" Third International Symposium on Electronic Commerce and Security, 2010, IEEE.

[6] Jacques Savoy, Olena Zubaryeva, "Classification Based on Specific Vocabulary" published in 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology 978-0-7695-4513-4/11 2011 IEEE .

[7] Fred Popowich, "Using Text Mining and Natural Language Processing for Health Care Claims Processing", SIGKDD Explorations. Volume 7, Issue 1 - Page 59

[8] G.Vinodhini and RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey", Volume 2, Issue 6, June 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[9] Zhongwu Zhai, Bing Liu, Hua Xu and Hua Xu, "Clustering Product Features for Opinion Mining", WSDM'11, February 9–12, 2011, Hong Kong, China. Copyright 2011 ACM 978-1-4503-0493-1/11/02...$10.00

[10] Singh and Vivek Kumar, "A clustering and opinion mining approach to socio-political analysis of the blogosphere", Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference.

[11] Bing Liu. "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, May 2012.

[12] V. S. Jagtap and Karishma Pawar, "Analysis of different approaches to Sentence-Level Sentiment Classification", International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume 2 Issue 3, PP : 164-170 1 April 2013

[13] Catherine Blake "A Comparison of Document, Sentence, and Term Event Spaces", published in IEEE 2010.

[14] Ying Chen, Wenping Guo, Xiaoming Zhao, "A semantic Based Information Retrieval Model for Blog" Third International Symposium on Electronic Commerce and Security, 2010, IEEE.

[15] Pablo Gamallo and Marcos Garcia, "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets" Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 171–175, Dublin, Ireland, August 23-24 2014.