INTERNET SERVICE PROVIDER COMPARISON USING MODIFIED K- MEANS CLUSTERING AND NAÏVE BAYES CLASSIFIER

AASHIMA SINGH, PROF. ELA KUMAR

Abstract: Opininon of customers play a very important role in daily life. Text based data is generated by the users in the form of comments, reviews, blogs, feedbacks on various social networking and review sites. For extracting useful information from textual data, a combination of NLP technique and text analytics is applied in this study to estimate Indian internet customer satisfaction based on the quality of service provided by various internet service providers. Data for this study is collected from Mouthshut.com which is a user generated content and consumer review platform on the internet. The goal of this study is to perform Sentiment analysis on online internet service providers reviews by combining Modified K-Means clustering algorithm with Naïve Bayes Classifier.

Keywords: Modified K-Means Clustering; Naïve Bayes Classifier; Sentiment Analysis; Mouthshut.com

Introduction: Sentiment analysis or opinion mining is a process to identify the polarity of opinion by applying Natural language processing and text analysis. This is done by filtering the sentences that do not contribute to the polarity and then extracting subjective information from the remaining text. The decision of a customer about the products and services of a certain entity is largely affected by the online reviews made by the existing customers. These reviews are very important and must be taken into account by the companies in order to improve the existing services, generating new services according to customer needs and to extract meaningful information from reviews. As a result focus shifts to the development of systems that can automatically summarize opinions from a set of reviews and display them in an easy to process manner. In this system, we are going to focus on three major ISPs, namely Bharti Airtel , Vodafone and Idea Cellular. As of 2016, Bharti Airtel has 365 million total subscribers followed by Vodafone with 200.47 million subscribers and Idea cellular with 191 million total subscribers. With the increasing growth of internet access, online user reviews are becoming the de-facto standard for measuring the quality of products and services. Many Indian internet customers express their sentiments about QOS, pricing, bandwidth, usage and speed of internet access provided by various ISPs through mouthshut.com. The decision of future customers is largely affected by the online reviews and comments about services. Therefore these sentiments are very important for ISPs in improving their services in any particular location. For example, in marketing reviews posted by the users can help in judging success of a new product or an ad campaign and to determine the popularity of product or service. There can be many challenges associated with sentiment analysis as an opinion word that is considered positive in one situation may be considered negative in another situation. This is because people don't react in a similar way in same kind of situation. Reviews posted online contain a combination of positive and negative views which is understandable by a human but difficult to process by a computer system. So to solve this problem, proposed study will combine modified K-Means clustering and Naïve Bayes classifier. K-Means clustering will create clusters of similar type of reviews and Naïve Bayes will further classify the review as positive or negative.

Related Literature: Pang and Lee[1] performed a survey on sentiment analysis and opinion mining. Various issues including opinion oriented information access, challenges, opinion classification and summarization has been explained. Mikalai Tsytsarau, Themis Palpanas[2] also presented a survey on opinion mining. In the survey they explained about opinion mining and its aggregation along with its subjectivity analysis. There has been several sentiment analysis performed on different topics such as Movies[3], Products[4][5], Restraunts[6] and Travel[7] e.t.c. Most of the sentiment analysis has been done on data available on social networking sites[8]. Several researchers worked on different machine learning methods for sentiment analysis[9][10][11] that involved training classifier on datasets and using trained model for new document classification. S. Gunelius[15] focused their study to use sentiment analysis in the context of tracking of blog comment reactions of Filipino customer satisfaction with regard of the services provided by their respective ISPs. Malay K. Pakhira[14] presented a modified version of the k-means algorithm that efficiently eliminates the empty cluter problem and demonstrates the proposed algorithm.

Proposed Approach: The proposed framework has four modules namely, data extraction, preprocessing, clustering and classification. Various steps in this approach are used to conceptualize, design and perform sentiment analysis on ISPs reviews. The goal

can be achieved by combining modified K-means clustering algorithm and Naïve Bayes Classification[12]. Following are the steps to perform sentiment analysis of ISPs reviews posted on mouthshut.com:

1. Data extraction: First extract the data which is to be analyzed. Here we have taken the data from mouthshut.com.
2. Training dataset: For easy preprocessing, we have created a training dataset for positive and negative sentiments. And another dataset for stopwords.
3. Preprocessing: This step includes removal of words which does not show any sentiment or opinion.

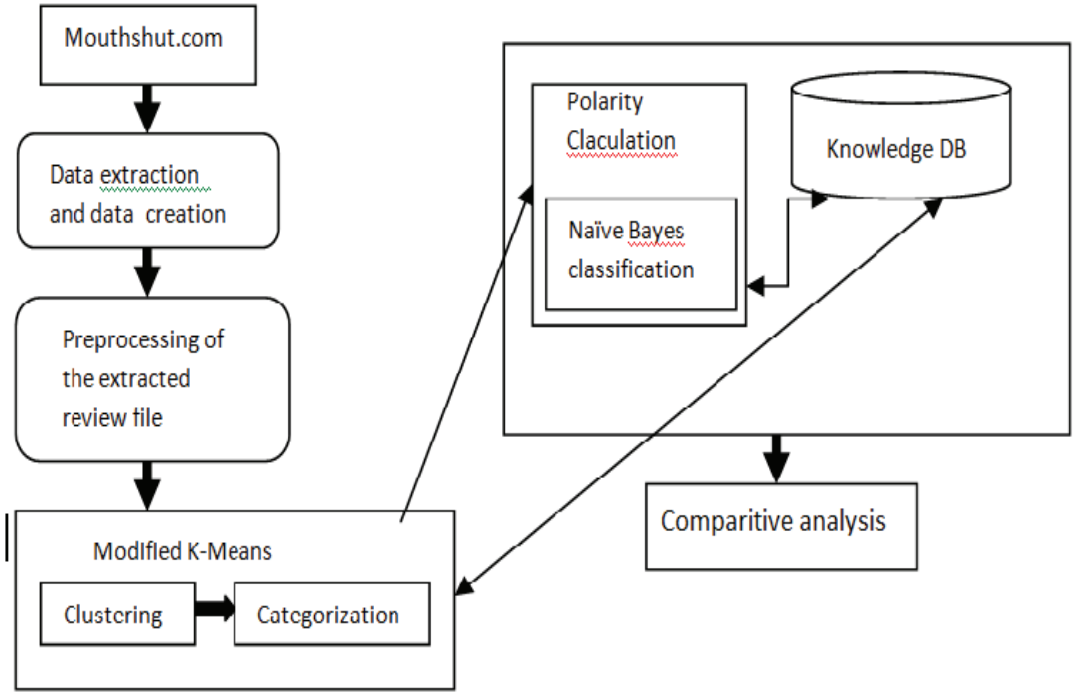

Fig 1. Proposed Framework

Stopwords are frequently occurring and insignificant words that construct sentences but does not represent any content of the document such as articles, prepositions, conjuctions and some pronouns. For example, a, an, the, about, are, for, from, these, who, where, etc.

- Stemming: A process of reducing words to their stems or roots. It includes remoing suffix or stripping. For example, "compute", "computing" and "computer" are reduced to "comput"[13].
- Digits: All the numbers and terms that contain digits ar removed except dates, times, etc.
- Hyphens: Hyphens are removedcin order to deal with inconsistency of usage. For example, "state-of-the-art" is replaced with "stateoftheart" or "state of the art".
- Case of letters: All the letters are either converted in uppercase or lowercase.
- Synonym expansion: It is the task of replacing a certain word in the given context with another suitable word having same meaning.
- Tokenization: Chopping up of a given stream of text or character sequence into words, phrases, symbols is called tokenization and the chopped word is called a token.

1. **Term frequency:** the frequency of a word in a database i.e. number of times the word occurs is known as the term frequency. To calculate term frequency vector space model is used.
2. **Polarity Calculation:** Here we have the total count of positive, negative and neutral sentiment words in the entered data which will be further used by clustering process or for creating clusters.
3. **Clustering:** Clustering is a technology for organizing data instances into similarity groups called clusters i.e. the data instances in a cluster are similar to each other but different from data instances of other clusters.

Clustering needs a similarity function to measure how similar two data points are ora distance function to measure distance between two data points.

Modified K-Means clustering algorithm: One of the most popular clustering algorithm is K-Means algorithm and it has been used in applications such as data mining, image segmentation, bioinformatics and many other fields. This algorithm works well with small datasets. Here we have all the basic

characters of K-Means algorithm in modified version[14]. The new and modified version is better as it has improved classification accuracy and decreasing number of iteration steps. The modified algorithm has a different procedure for updating center vector and at the same time it eliminates all the empty clusters. In modified K-Means algorithm only the center computation part is changed and the remaining part is same as K-means algorithm.

Classification: Naïve Bayes classification [12] is used for classifying the clusters formed by clustering. Naïve Bayes is a supervised machine learning algorithm. It determines whether a sample belongs to a particular class or not by computing the probabilities of the outcomes. Result is computed by Naïve Bayes classification algorithm based on whether the review has positive of negative sentiment.

Implementation Details: To compare various ISPs we have used modified K-Means clustering algorithm followed by Naïve Bayes classification for generating improved results, as compared to using original K-Means clustering or Naïve Bayes classification implemented alone. For creating front end of the system, Visual Studio 2010 has been used in programming. Whereas SQL server Administration Studios 2008 is used for development of database. This system will allow a user to register itself by providing name and location else it also allow user to search directly about ISPs in a city. Once the user is registered, the system will ask for the city in which the user wants to know about best ISP provider. In turn the system will generate the best ISP in that location as a result.

Conclusion and Result: The study aims at examining the service level of various ISPs serving in a particular area. The ISPs of three major cities i.e. Delhi, Mumbai and Kolkata are taken into consideration. The modified K-Means approach gives more accurate results than original K-Means as it requires less time to classify the message. This approach has less number of iterations low misclassification rates and also avoids empty clusters.

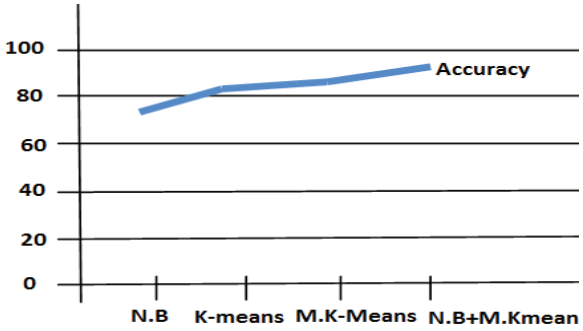

Fig 2. Comparison of accuracy between different Algorithms.

References:

1. Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval Vol. 2, Nos.1–2 (2008)
2. Mikalai Tsytsarau, Themis Palpanas "Survey on mining subjective data on the web", Data Mining Knowledge Discovery, Springer 2012, pp.478-514.
3. Pang B, Lee L, Vaithyanathan S. "Thumbs up? Sentiment classification using machine learning techniques". Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2002.
4. Dave K, Lawrence S, Pennock D. "Mining the peanut gallery: opinion extraction and semantic classification of product reviews". Proceedings of the 12th international conference on World Wide Web, ACM, New York, NY, USA, WWW'03.
5. Hu M, Liu B. "Mining and summarizing customer reviews". Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, NY, USA, KDD 2004.
6. Jingjing Liu, Stephanie Seneff, and Victor Zue, "Harvesting and Summarizing User-Generated Content for Advanced Speech-Based HCI", IEEE Journal of Selected Topics in Signal Processing, Vol. 6, No. 8, Dec 2012, pp.982-992.
7. Aditya Joshi, Balamurali A. R., Pushpak Bhattacharyya "A Fallback Strategy for Sentiment Analysis in Hindi a Case Study" Proceedings of ICON 2010: 8th International Conference on Natural Language Processing, Macmillan Publishers, India.
8. Aditya Joshi, Balamurali A R, Pushpak Bhattacharyya, Rajat Mohanty, "C-Feel-It: A Sentiment Analyzer for Micro-blogs", Proceedings of the ACL-HLT 2011, pp.127-132.
9. Alvaro Ortigosa, José M. Martín, Rosa M. Carro, "Sentiment analysis in Facebook and its application to e-learning", Computers in Human Behavior Journal Elsevier 2013.
10. Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou "Movie Rating and Review Summarization in Mobile Environment", IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 42, No. 3, May 2012, pp.397-406.
11. Alexandra Trilla, Francesc Alias "Sentence-Based Sentiment Analysis for Expressive Text-to-Speech", IEEE Transactions on Audio, Speech,

and Language Processing, Vol. 21, No. 2, February 2013, pp.223-233.
12. I. Rish. An Emperical Study of Naïve Bayes Classifier. In Proceedings of IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 2001.
13. M. F. Porter. An algorithm for Suffix Stripping. Program, 14(3), pp 130-137, 1980.
14. Malay K. Pakhira, " A Modified K-Means Algorithm to Avoid Empty Clusters," International journal of Recent Trends in Engineering, vol.1, no.1, pp. 220-226, Issue, May2009.
15. S. Gunelius, What are Blog comments? The importance of Blog Comments to Bloggers, http://weblogs.about.com/od/partsofablog/qt/BlogComments.htm.

Aashima Singh, PG Scholar, Ela Kumar, H.O.D,
Department of CSE, Indra Gandhi Delhi Technical University, Kashmere Gate, Delhi.