BRAND MONITORING BASED ON SENTIMENT ANALYSIS USING NAÏVE BAYES ALGORITHM

M. Revathi[1], S.Jayanthi[2]

Assistant Professor, Department of CSE, Agni College of Technology, Chennai.

Abstract: In this modern era, everyone tends to express their opinions on social media.These opinions reflect the views of people taking that certain product.These views help the people to know about a certain product before buying it.The view point of the users in form of tweets significantly gives a customer perception of the product.It is a difficult task to manually read the tweets and rate them. This problem can be solved by creating an automated app in which we can analyze the reviews and rate the product.This analysis also helps the manufacturers to know how the product has been received by the customers.In this project we will try to build an automated web application to monitor a brand's product using a classifier called Naïve Bayes.

Keywords—Brand monitoring;Sentiment analysis;Data mining; product review;Naïve Bayes.

I. Introduction

Social media has turned out to be the most popular medium or platform for people to express their personal views on anything. It is easily accessible by anyone and they are able to post their views on social networking sites anywhere and anytime. Product based opinions of users are common in social media. It is a tedious job to read each and every review for the users and come to a conclusion whether to buy the product or not. Hence, an automated application is created, which analyses the text written by a reviewer and identify whether it is positive or negative.

In this application, we have used a machine learning approach called Naïve Bayes. This method gives accurate results when compared to other algorithms. There are two types of datasets used here, which are training sets and data sets. The reviews are classified into positive and negative based on certain keywords mentioned in the training sets.

The rest of this paper is organized as follows: Section II presents the literature survey; Section III presents the system architecture and details of the proposed system and we conclude this paper in Section IV.

II. Literature Survey

Hari Krishna et al.[1] introduced a machine learningapproach called supervised classification. This method tends to be moreaccurate than other methods as they have trained the classifierusing real world dataset. In this work we have developed a procedure for 'feature based sentiment analysis'by using a classifier called Support Vector Machine.

There should be two types ofdatasets. One is called training set and it is created manually.They are basically used to train the classifier. Theyhave used test set to determine the accuracy of the system.

ShwetaRana et al.[2] proposed a classifier called Support Vector Machine(SVM).The work of this is it can extract and analyze the users' perception of the product.They have performed the sentiment analysis on thereviews using the algorithms like Naive Bayes, Linear SVM and synthetic words.

Their experimental results indicate that the Linear SVM has provided the best accuracy which is followed by the Synthetic words approach.But the problem with this research is it doesn't perform well with large data sets as it requires more training time.

In [3], Ziegler et al.monitored the reputation of a brand as well as its competitors.It also analyses the web data for such purposes.It uses semantic perspectives for analysing the web data.They have presented a platform for analyzing Webdata for such purposes, adopting different semantic perspectivesand providing the market analyst with a flexiblesuite of instruments.

They havefocused on two of these tools andoutline their particular utility for research and exploration.But it is a complex process as web data can be from various sources of the internet.

Huang et al.[4] considered Naïve Bayes classification algorithm based on Poisson distribution model.This model keeps high classification accuracy even in small data sets.So it reduces the requirement for storage and computer resources for classification work.

Naive Bayes in the large data set showed satisfactory speed and accuracy, but the effect is noticeably poor in the case of small data set. However, it is often difficult to have a large data training set in reality. In addition, too large data scale will result in very high complexity of running time and space.

III. Proposed System

The proposed system is an improvement to the technique introduced in [1].Fig. 1 shows the proposed system architecture.

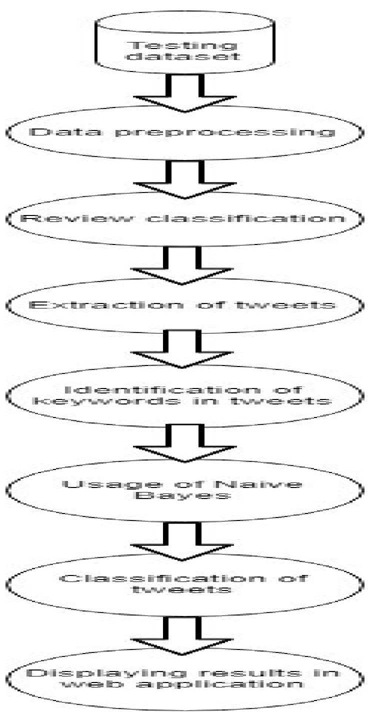

Fig.1. Proposed System Architecture

Testing Datasets

Two types of datasets are used here, one for training sets and other for testing. The datasets are created manually. The data in the testing set are manually tagged as positive and negative, based on reviews from E-commerce sites. After this process, the system will be able to analyse which is a positive review and which is a negative one. We intend to test the system bygiving reviews from test set whose polarity is already known to us.

Data Preprocessing

The main preprocessing methods performed are stemming, error correction and stop word removal. Stemming is the basic task of identifying the root of a word. The purpose of this method is to remove various suffixes. We need to incorporate error correction as the reviewers will not be following the exact grammatical rules, punctuation and spellings. Some stop words like "it" should not be removed as it may effect coreference resolution.

Review Classification

Review Classification refers to the collection of the reviews posted in social networking sites based on individual products. The tweets are classified by considering their product keywords and specific brands. In some cases, reviews are identified by the specific product hashtags.

Extraction of Tweets

Users post tweets based on the products on Twitter to express their views about the product. Product related tweets are extracted from Twitter using RapidMiner software. They can be extracted by using the relevant product keywords, brand or hashtags.

Identification of keywords in tweets

The extracted tweets are stored in the database and they are analysed, based on certain keywords, whether they are positive or negative reviews. The data in the training sets are used to identify if the review is positive or negative.

Usage of Naïve Bayes

We have used the Naïve Bayes algorithm for classification of reviews. Naive Bayes algorithm is one of the most effective methods in the field of text classification, but only in the large training sample set can it get a more accurate result. It evaluates the probability value for positive and negative reviews and provides the result.

Classification of Tweets

The result obtained from Naïve Bayes algorithm is analysed and it is classified into positive and negative and stored in the database. The total number of reviews for each product is calculated, and the number of positive and negative reviews are also recorded.

Displaying results in web application

The results which are computed are displayed in a web application, which shows the total number of reviews about the product, positive reviews, negative reviews and the neutral reviews. The users are required to enter the product name and the related result will be displayed.

IV. CONCLUSION

In this paper we have proposed a combination of Naïve Bayes algorithm to provide an accurate result when compared to other algorithms. The end application displays the number of positive and negative reviews and also the total number of reviews for the particular product the user has searched for.This application makes it easier for users to come to a conclusion whether to buy the product or not. It saves time for the user by not reading each and every tweet related to the product.

References

[1]M. Hari Krishna, K. Rahamathulla, Ali Akbar "A feature based approach for sentiment analysis using SVM ,coreference resolution" IEEE Access, pp. 397-399,July 2017.

[2] ShwetaRana, Archana Singh, "Comparative analysis of sentiment using SVM and Naïve Bayes techniques" IEEE Access, pp. 106-111,2016.

[3] Cai-Nicolas Ziegler, Michal Skubacz, "Towards Automated Reputation and Brand Monitoring on the Web" IEEE Access pp. 1066-1072, 2012.

[4] Yuguang Huang, Lei Li, "Naive Bayes classification algorithm based onsmall sample set", IEEE Access, pp. 34-39,2011.

[5]. Y. Li, H.-X. Chen, and Y. Zhao, "A new method of data hiding based on H.264 encoded video sequences," in Proc. IEEE Int. Conf. Signal Processing, ICSP, Oct. 2010, pp. 1833–1836.

[6]. D.-Y. Fang and L.-W.Chang, "Data hiding for digital video with phase of motion vector," in Proc. IEEE Int. Symp. Circuits Systems, ISCAS, Sep. 2006.

[7]. C. Xu, X. Ping, and T. Zhang, "Steganography in compressed video stream," in Proc. Int. Conf. Innovative Computing, Information and Control, ICICIC‟06, 2006, vol. II, pp. 803–806.

[8]. K. Wong, K. Tanaka, K. Takagi, and Y.Nakajima, "Complete video quality-preserving data hiding," IEEE Trans. Circuits Syst. Video Technol., vol. 19, no. 10, Oct. 2009.

[9]. K. Solanki, U. Madhow, B. S. Manjunath, S. Chandrasekaran, and I. El-Khalil, "‟Print and Scan‟ resilient data hiding in images," IEEE Trans. Inform. Forensics Security, vol. 1, no. 4, pp. 464–478, Dec.2006.

[10]. X.-P. Zhang, K. Li, and X. Wang, "A novel look-up table design method for data hiding with reduced distortion," IEEE Trans. Circuits Syst. Video Technol., vol. 8, no. 6, pp. 769–776, Jun. 2008.

[11] W. J. Lu, A. Varna, and M. Wu, "Secure video processing: Problems and challenges," in Proc. IEEE Int. Conf. Acoust.,Speech, Signal Processing,Prague, Czech Republic, May 2011, pp. 5856–5859.

[12] B. Zhao, W. D. Kou, and H. Li, "Effective watermarking scheme in the encrypted domain for buyer-seller watermarking protocol," Inf. Sci.,vol. 180, no. 23, pp. 4672–4684, 2010.

[13] P. J. Zheng and J. W. Huang, "Walsh-Hadamard transform in the homomorphic encrypted domain and its application in image watermarking,"in Proc. 14th Inf. Hiding Conf., Berkeley, CA,USA, 2012, pp. 115.

[14] W. Puech, M. Chaumont, and O. Strauss, "A reversible data
hiding method for encrypted images," Proc. SPIE, vol. 6819,pp. 68191E-1–68191E-9, Jan. 2008.

[15]. X. P. Zhang, "Reversible data hiding in encrypted image," IEEE Signal Process. Lett., vol.18, no. 4, pp. 255–258, Apr. 2011.

[16]. W. Hong, T. S. Chen, and H. Y. Wu, "An improved reversible data hiding in encrypted images using side match," IEEE Signal Process.Lett., vol.19,no. 4, pp. 199–202, Apr. 2012.

[17] .X. P. Zhang, “Separable reversible data hiding in encrypted image,” IEEE Trans. Inf. Forensics Security, vol. 7, no. 2,pp. 826–832, Apr. 2012.

[18] . K. D. Ma, W. M. Zhang, X. F. Zhao, N. Yu, and F. Li, “Reversible data hiding in encrypted images by reserving room before encryption,” IEEE Trans. Inf. Forensics Security, vol. 8, no. 3, pp. 553–562, Mar. 2013.