IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-5, Issue-7 E-ISSN: 2347-2693

Sentiment Analysis using Naïve bayes Algorithm

Pramod M. Mathapati[1*], A.S. Shahapurkar[2], K.D.Hanabaratti[3]

[1*] Dept. of Computer Science, Gogte Institute of Technology, VTU, Belagavi, India
[2] Dept of Computer Science, Gogte Institute of Technology, VTU, Belagavi, India
[3],Dept of Computer Science, Gogte Institute of Technology, VTU, Belagavi, India
Corresponding Author: pramodkle064@gmail.com

Available online at: www.ijcseonline.org

Received: 26/Jun/2017, Revised: 07/Jul/2017, Accepted: 20/Jul/2017, Published: 30/Jul/2017

Abstract— Sentiment analysis is trending topic of research which works on data which is got from review websites, social networks. Today users having common platforms like Blogs, micro blogs, review sites, twitter and other social networks through which they can post their feedbacks. Organizations use Sentiment Analysis to understand user's reviews and feedbacks about the product which they have released. In this project development of a Sentiment analysis using a generic method which can be applied for sentiment analysis as well as Emotional Analysis, product reviews is done based on Naïve Bayes classifier method. Naive bayes Classifier is the better choice for Sentiment Analysis as it is more efficient and gives Quick results compared to other techniques such as Support Vector Machine and Maximum Entropy.

Keywords—Sentiment Analysis, Modified k means, NLP, Opinion Mining,

I. INTRODUCTION

People and Organizations express their views and opinions on internet as Sentiments influence human actions, activities. Sentiments are expressed by people every day through internet as people so much dependent on it. There is need of analyzing these sentiments hence opinions or views of user have got more importance. Today users having common platforms like Face book, Instagram, review sites, twitter and other social networks through which they can post their feedback about the product they are using and they can even post their comments whether they are happy with the service which they got or not [1,2,3]. Likewise even organizations can also share their experiences. Sentiment analysis is trending topic of research which works on data which is got from review websites, social networks [4,5,6].

In this Work development of a Sentiment analysis using a generic method which can be applied for sentiment analysis as well as Emotional Analysis, product reviews is done based on Naïve Bayes classifier method. Researchers have got new opportunities as well as challenges while working with Sentiment Analysis which can also be used for natural language processing [2].

II. RELATED WORK

Today people are using Twitter and Face book as the main social networks and even they are known as Micro Blogs as they allow user to broadcast short text messages, images, audio or video to other users. Messages which are exchanged by these networks can be used as data for sentiment analysis. Today users are generating thousands of reviews for product and services they are using through websites which are present on the internet [4,7,]. Decision making for new user will be decided by these data. Apart from Twitter and face book we have different social networks as well which also works as data source for sentiment analysis. There are many websites available where they can provide data for Sentiment Analysis such as www.amazon.com (product reviews), www.reviewcentre.com (product reviews), www.fonearena.com (mobile reviews), where millions of reviews are posted from the customers [1,8,9,10]. Even we can have movie review dataset which is used to identify distinguished features, method of selection of feature and best supervised learning algorithm. Cornell Movie Review Dataset is the classical dataset which consists of thousand negative reviews and thousands of positive reviews. People can easily access this dataset.

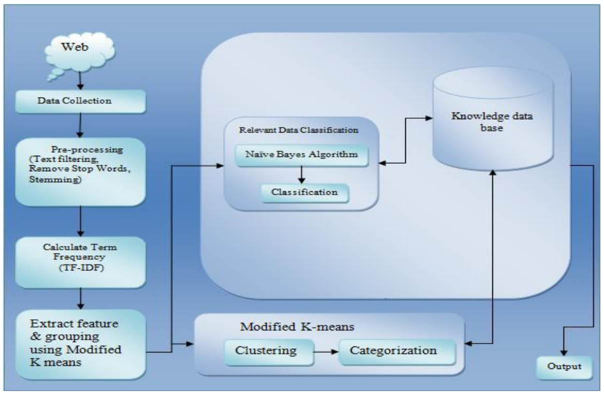

Fig1. Proposed Architecture

III. METHODOLOGY

In this Paper five modules have been used.

1) User interface Module

In this module development of a web portal which will allow any user to sign up and achieve sentimental analysis on any kind of data regardless of any type like twitter data, movies reviews, product review, etc. Even the portal will keep track of all recent activities the user has done which will kept as reference for his future work. User has an option to supply input as a text file or just a sentence as an input which will be further processed by the web portal. The web module will also have a admin section where he can get all user list with extra options, even a section where he can update the training data set of the sentimental analysis.By managing unrelated data and increases accuracy with Naïve Bayes Classification algorithm.

2) Training the Classifiers

Decision making is very important for organizations as they have to understand how people accepting their product when it is newly launched in the market. That can be understood by the reviews, compliments or opinions given by the people so that organizations can decide about their future decisions.

3) Pre-processing Module

Preprocess data

1) @username - elimination of "@username" via regex matching or replace it with generic word AT_USER is possible.
2) #hashtag - hash tags can give some useful information, so it is useful to replace them with the exact same word without the hash. E.g. #nike replaced with 'nike'.
3) Lower Case – tweets will be converted to lower case.
4) URLs - No need to follow the short urls and determine the content of the site, so we can remove all of these URLs via regular expression matching or replace with generic word URL.
5) Punctuations and additional white spaces - remove punctuation at the start and ending of the tweets. Ex: ' the evening is wonderful! 'Replaced with 'the evening is wonderful'. It is also helpful to replace multiple whitespaces with a single whitespace.

4) Feature Extraction

Selecting a subset of the words appearing in the training Database, feature extraction can be achieved. In text categorization only this subset should be taken as feature.

Two main causes for feature extraction:

1. Amount of useful words is decreased so that training and applying a classifier method can be more capable.
2. Classification accuracy can be improved by removing noise features and which also enhance feature extraction.

5) Naïve Bayes Classification.

Document can be classified by applying Naïve Bayes theorem

Naive Bayes (NB)

To get exact and proficient results for linearly separable cases Naïve bayes is the best choice and even it can be used for non-linearly separable cases also. Because of its simple interpretation and effective computation naïve bayes is used in different applications. When probability of instance which has already occurred is known then naïve bayes is used to calculate the probability of current instance. We can write like P (D | C) P(C) in Mathematics

P (D): Probability of document D occurring

P(C | D) = Where, P (C | D): Probability of Document D being in Class C

P(C): Probability of occurrence of Class C,

P (D) P (D | C): Probability of generating Document D given Class C.

Naïve Bayes Classifier

- Simple classification of words based on 'Bayes theorem'.
- It is a 'Bag of words' (text represented as collection of its words, discarding grammar and order of words but keeping multiplicity) approach for subjective analysis of content.
- Better in terms of CPU and Memory utilization

Probabilistic Analysis of Naïve Bayes

For a document d and class c, By Bayes theorem

p(c/d)=p(d/c)p(c) / p(d)

Naïve Bayes Classifier will be:

C* = arg maxc p(c/d)

Multinomial Naïve Bayes Classifier

Accuracy – around 75%

Algorithm - :

- ✓ *Dictionary Generation*

  Making a dictionary of some most frequent words can be achieved by counting occurrences of all words in given dataset.
- ✓ *Generation of Feature Set*

Representation of all documents as a feature vector over the space of dictionary words.

It is required to keep track of dictionary words with their number of occurrences in the document. .

IV. RESULTS AND DISCUSSION

Naive Bayes algorithm has given quick and accurate results compared to other methods like SVM(Support Vector Machine) and Maximum Entropy. Time efficiency is also high in case of Naive Bayes as SVM and Maximum Entropy takes more time to classify the input data and process. Naive bayes saves time for computing as it don't need and preprocessing mechanism to classify the given input data hence it saves time.

V. CONCLUSION AND FUTURE SCOPE

In this Paper, proposal of a method using naïve bayes has given more quick and accurate results compared to other techniques of Sentiment Analysis such as Support Vector Machine and Maximum Entropy. More time is needed to SVM and Maximum entropy to give optimum results. Naive bayes classifier can be used for both Sentiment and Emotion Analysis. High accuracy is achieved for sentiment analysis and as well as for product reviews, movie reviews. Even for text based classification and for social interpretation it gives better results.

In future focus will be on to find out how this method can be used as more generic method when applied to customer reviews, product reviews, human sentiments etc. Focus will be on improvement and more accurate results for sentiment analysis. Efforts will be made to add some additional features in Naïve Bayes classifier algorithm so that it can be used to develop a more generic method. Different Social networks can use this method as their API.

REFERENCES

[1]. O. Coban and B. O. and G. T. Ozyer, "*A Comparison of Similarity Metrics for Sentiment Analysis on Turkish Twitter Feeds*," 2015 IEEE International Conference on Smart City /SocialCom /SustainCom (SmartCity), Dec. 2015.

[2]. Pablo Gamallo, CITIUS, Univ. de Santiago de Compostela 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, 2014

[3]. Shakhy P S and Vidya K S, "*A Survey on KASR for Big Data Applications*", International Journal of Computer Sciences and Engineering, Vol.3, Issue.4, pp.85-89, 2015.

[4]. Ashish Shukla, Rahul Mishra, "*Sentiment Classification and Analysis using Modified K-means and naïve Bayes Algorithm*" International journal of Advanced Research in computer science and Software Engineering. Volume 5, Issue 8, August 2015.

[5]. J. Kaur, S.S. Sehra, S.K. Sehra, "*A Systematic Literature Review of Sentiment Analysis Techniques*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.4, pp.22-28, 2017.

[6]. U. Aggarwal, G. Aggarwal, "*Sentiment Analysis : A Survey*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.5, pp.222-225, 2017.

[7]. D.M. Blei, A.Y. Ng, and M.I. Jordan, "*Latent Dirichlet Allocation*," J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[8]. C.P. Robert and G. Casella, "*Monte Carlo Statistical Methods*, second ed. Springer Publisher 2005.

[9]. S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fuku-shinna, "*Mining Product Reputations on the Web*," Proc. Eighth ACM SIGKDDInt'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02), pp. 341-349, 2002.

[10]. [10] M. Hu and B. Liu, "*Mining and Summarizing Customer Reviews*," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '04), pp. 168-177, 2004.