IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

(Volume2, Issue1)

Streaming Analytics

[1] Aditya Shirke, [2] Ankur Singh, [3] Umesh Sapar, [4] Aditya Sengupta, [5] Prof Leena Deshpande

aditya.shirke@viit.ac.in[1], ankur.singh@viit.ac.in[2] , saparumesh@gmail.com[3] , adityasengupta2008@gmail.com[4] , leena.deshpande@viit.ac.in[5]

Department of Computer Engineering

Vishwakarma Institute of Information Technology, Pune.

Abstract: Nowadays Sentiment Analysis play an important Role in each field such as Stock market, product reviews, news article, political debates which help us to determining current trend in the market regarding specific product, event, issues. Here we are apply sentiment analysis on microblogging platforms such as twitter, Facebook which is used by different people to express their opinion with respect to different kind of foods in the field of home'schef. This paper explain different methods of text preprocessing and applies them with a naive Bayes classifier in a big data, distributed computing platform with the goal of creating a scalable sentiment analysis solution that can classify text into positive or negative categories. We apply negation handling, word *n*-grams, stemming, and feature selection to evaluate how different combinations of these pre-processing methods affect performance and efficiency.

Keywords: Text mining, Natural language processing, Sentiment analysis, Features Extraction

I. INTRODUCTION

Streaming analytics is related to a sentiment analysis, where we are going to extract the data, comment given by the people about specific food, meals or restaurant, from social media sites such as Facebook, twitter. Develop the problem on analytics which analyzes the different kind of sentiments, emotions, opinion, given by the customers from the Social media with respect to Home'schef and predicts and Recommend the online Review and Show the Result.

Our main challenge is to develop the web site related to the Home'schef where different people can order different types of foods from different hotels and restaurant at any time. For that, each user must register by giving their personal information such as name, date of birth, area, etc., to create their own account to our website so that he can order as much as food. And Restaurant also Register to our website by giving their personal information such as Hotel address, specialty, Menu card, etc.

Finally we are apply Naive Bayes classifier algorithm for classifying and generate the result according to area wise, different age, season of different food and gender.

This result will be helpful for restaurant as “What people think about your food”, and also he will get know “latest trends in markets of different kind of foods” for increasing their business.

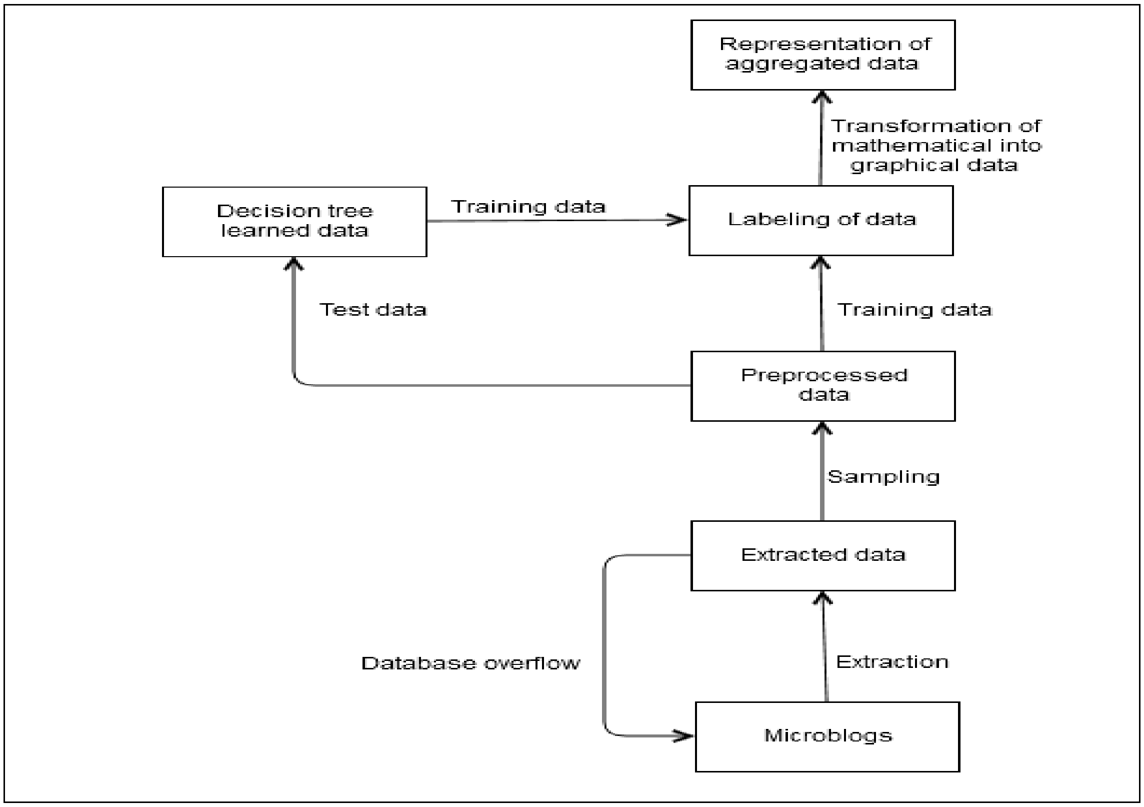

Fig 1.Flowchart

II. LITERATURE SURVEY

A) Some of the early and recent results on sentiment analysis of Twitter data are by A. Go et al(2009) [8], (Bermingham and Smeaton, Classifying sentiment in microblogs, 2010, pages 1833-1836) and Pak and Paroubek (2010) [5].

B) By Haseena Rahmath , Tanvir Ahmad [1], “Sentiment Analysis Techniques - A Comparative Study”. In this paper they have aimed on comparative study of different Sentiment analysis techniques such as Supervised Machine learning based techniques, Lexicon Based Method, Hybrid Techniques. In Hybrid Techniques, many researcher have accomplish the task by combining supervised machine learning and lexicon based approaches. They got into conclusion that more researchers are needed in this field to achieve better performance in sentiment classification.

C) A. Agarwal et al.(2009) [6], in their paper Contextual phrase-level polarity analysis using lexical affect scoring and syntactic ngrams; experimented with three types of models: unigram model, a feature based model and designed the tree kernel based model on tweets for sentiment analysis.

D) A. Go et al. (2009) [8], in their paper “Twitter Sentiment Classification using Distant Supervision” use distant learning to acquire sentiment data. They use tweets ending in positive emoticons like “:)” “:-)” as positive and negative emoticons like “:(” “:-(” as negative.

E) A. Pak and P. Paroubek (2010) [5], in their paper “Twitter as a Corpus for Sentiment Analysis and Opinion Mining” collect data following a similar distant learning paradigm. They perform a different classification task though: subjective versus objective. For subjective data they collect the tweets ending with emoticons in the same manner as Go et al. (2009). For objective data they crawl twitter accounts of popular newspapers like “New York Times”, “Washington Posts” etc.

F) A. Celikyilmaz et al.(IEEE, 2010) [3] in their paper “Probabilistic model-based sentiment analysis of twitter messages” developed a pronunciation based word clustering method for normalizing noisy tweets.

G) Zhen Niu et al. [2] introduced a new model in which efficient approaches are used for feature selection, weight computation and classification.

H) Wu et al. [9] proposed a influence probability model for twitter sentiment analysis.

I) Pak et al. [5] created a twitter corpus by automatically collecting tweets using Twitter API and automatically annotating those using emoticons. Using that corpus, they built a sentiment classifier based on the multinomial Naive Bayes classifier that uses N-gram and POS-tags as features.

J) Barbosa and Feng (2010) [10]. They use polarity predictions from three websites as noisy labels to train a model and use 1000 manually labeled tweets for tuning and another 1000 manually labeled tweets for testing. They however do not mention how they collect their test data. They propose the use of syntax features of tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words.

K) Michael Gamon (2004) [11] perform sentiment analysis on feedback data from Global Support Services survey. One aim of their paper is to analyze the role of linguistic features like POS tags. They perform extensive feature analysis and feature selection and demonstrate that abstract linguistic analysis features contributes to the classifier accuracy.

III. PROPOSED METHODOLOGY

In this paper we have studied different methodologies which can be useful to complete the given problem.

A. Feature Selection:

Feature selection method can be used to identify and remove unneeded and Redundant attributes such as punctuation, infrequent word, and stop word from data that do not contribute to the accuracy of a predictive model. We eliminate **stop words** like: a, about, above, across, anywhere, also, as, back, become, been, do, does, cases, certain, end, ended, find, group, known, myself, place, such, why, yes, work, toward, this, there, etc. We also remove punctuations like: { } , : ; ( ) . Removing unneeded and redundant attributes such as punctuation will reduce feature set such that it will need less memory size to process and it will be easier for next steps to manipulate the feature set.

B. Preprocessing of Text:

In order to reduce the dimensionally of the document words, special method such as filtering and Stemming are applied. Generally we use term Noise cancellation for this process. Filtering is a process in which those words gets left out which does not help in getting sentiment precisely. Also one major advantage of using filtering is that it will reduce the database size tremendously. Stemming is a process which reduces words to their morphological roots. For e.g. doing, done, did may be represented as ‘Do’.
Another method is negation handling, where we negate the positive sentiment of words if they are negated in the sentence

C. Feature Extraction:

Features extraction starts from an initial set of measured data and builds derived values. When the input data to an algorithm is too large to be processed and it is redundant then it can be transformed into a reduced set of features. This process is called feature extraction. The extracted features are expected to contain the relevant information from the input data. Extracted features are also small in size comparable to the size before extraction.

D. Dictionary Based Algorithm:

We use dictionary based algorithm for counting purpose. Dictionary based algorithm gives number of count of specific keywords we want to find and it also count total number of words. Suppose, there are four keyword namely : A,B,C,D and total occurrences of these four keywords are 32 in which A occurred 5 times, B occurred 10 times, C occurred 8 times and D occurred 9 times. So, using dictionary based algorithm we found out that A occurred 5/32 times. Similarly B occurred 10/32 times, C occurred 1/4 times, D occurred 9/32 times. We are then using this information in Naïve Bayes classifier.

D. Naïve Bayes classifier:

Naïve Bayes classifier is a conditional probability model which is based on Bayes' theorem which gives probability of an event, based on conditions which might be related to that specific event. We use Naïve Bayes classifier to classify the feature set into two set of class: positive and negative class. In a document which is training dataset for Naïve Bayes classifier, we gave class value to sample dataset which would be given in string. The classifier then create probability of each word given in those training sample data along with total positive and also total negative probability of given data samples. We will use classifier to give minimum possible probability to those words which are never been into the sample. Such that, if any new word came in test dataset, classifier will give minimum possible probability to those words which are never been in the sampled data. Also, Naïve Bayes will give probability on the basis of respective sentence class. The term sentence class is defined here which means that the class assigned to that overall sentence. So, if a new word came in test dataset, then according to the overall sentence the probability of that word to be in positive or negative class is defined. We use higher will be the label class priority which is used worldwide in Naïve Bayes classifier. Like, if the given stream of data in test dataset is having higher negative probability than the positive probability then that stream of data is most probably would be in negative class. So, here we used Naïve Bayes classifier so that new sentences would get appropriate label n there would be no error.

VI. CONCLUSION

In this paper, we demonstrated a simple implementation of a naive Bayes classifier that shows accuracy comparable with the state-of-the-art on the TWITTER dataset, yet is scalable and efficient, due to the minimal pre-processing and the strong independence assumption of the naive Bayes classifier. We also demonstrated how different pre and post-processing methods change accuracy, specifically how choosing the right combination of n-grams and the most contributing features improves accuracy. We also implemented sentiment analysis on a distributed computing platform, Apache Spark that showed high accuracy and efficiency with the Twitter dataset. We also suggest that effort should be put into realizing more context-aware systems, as that is where the field is currently lacking.

VII. FUTURE SCOPE

In future this paper can be expanded to provide more accurate results by providing recommendations based on review and ratings obtained from the customers. We tentatively conclude that sentiment analysis for microblogs' data is not that different from sentiment analysis for other genres. In future work, we will explore even richer linguistic analysis, for example, parsing, semantic analysis and topic modeling. Also, by using Hybrid Technique, we were able to achieve an accuracy upto 100%. So, more researchers are needed in this field.[1]. Set of current experiments indicates that there are number of interesting directions for future work.

ACKNOWLEDGEMENT

An endeavor is successful only when it is carried out under proper guidance & blessings. We hereby are thankful to Prof. Mr. S.R. Sakhare, Head of Department, Computer, Vishwakarma Institute of Information Technology, Pune. It gives us great pleasure in acknowledging the support of Prof. (Mrs.) L.A. Deshpande and Mr. A. A. Bhosale and for their productive & hopeful suggestions. We would also like to thank Mr. Sachin Deshpande CEO, InnoBytes Technologies Pvt. Ltd for his insightful ideas and suggestions. We also thank all Teaching and Non-teaching staff of

VIIT, Pune for their kind of co-operation during the course. The blessings and support of family and friends has always given me success, we are extremely thankful for their love.

REFERENCES

[1] Haseena Rahmath , Tanvir Ahmad, “Sentiment Analysis Techniques - A Comparative Study” in IJCEM International Journal of Computational Engineering & Management, Vol. 17 Issue 4, July 2014

[2] Z. Niu, Z. Yin, and X. Kong, “Sentiment classification for microblog by machine learning,” in *Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on*, pp. 286–289, IEEE, 2012.

[3] A. Celikyilmaz, D. Hakkani-Tur, and J. Feng, “Probabilistic model-based sentiment analysis of twitter messages,” in *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pp. 79–84, IEEE, 2010.

[4] L. Barbosa and J. Feng, “Robust sentiment detection on twitter from biased and noisy data,” in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 36–44, Association for Computational Linguistics, 2010.

[5] Alexander Pak and Patrick Paroubek. 2010. “Twitter as a corpus for sentiment analysis and opinion mining”, Proceedings of LREC.

[6] Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. “Contextual phrase-level polarity analysis using lexical affect scoring and syntactic ngrams”. Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 24–32, March.

[7] Adam Bermingham and Alan Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage is brevity an advantage? ACM, pages 1833–1836

[8] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford.

[9] Y. Wu and F. Ren, 'Learning sentimental influence in twitter', Future Computer Sciences and Application (ICFCSA), 2011 International Conference, IEEE, vol. 119122, 2011.

[10] L. Barbosa and J. Feng, “Robust sentiment detection on twitter from biased and noisy data,” in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 36–44, Association for Computational Linguistics, 2010

[11] Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of the 20th international conference on Computational Linguistics.

[12] Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In Proceedings of the 17th European Conference on Machine Learning.