

Sentiment Analysis of Review Data Using Naïve Bayesian Classification Algorithm

¹Ms. Soumya P C, ²Dr. Venkatramana Bhat P

^{1,2}Department of Computer Science, Mangalore Institute of technology and Engineering , Moodabidri, Mangalore, India.

¹soumyacherianp@gmail.com, ²venkatramana@mite.ac.in

Abstract— Sentiment Analysis (SA) also called opinion mining is the recent trend now a days. Sentiment analysis is the method of finding the opinion or sentiment of particular sentence. Sentiment analysis categorizes the sentiment of review to positive, negative or neutral. Because of large size of emerging data, the sentiment analysis of review will come under big data. Data mining is the process of extracting useful information from large data warehouse. SA also comes under data mining also with Natural Language Processing. Various methods are available to find the sentiments. In this paper we used machine learning based sentiment analysis by using laptop review from amazon.com. Under Linux system we installed hadoop on top it and performed mapreduce programming model using Naïve Bayesian classification algorithm which is based on probability.

Index Terms— Sentiment analysis, hadoop, mapreduce, HDFS.

I. INTRODUCTION

Sentiment analysis is categorizing the sentiment to positive, negative or neutral. Because of large size of emerging data, the sentiment analysis of review will come under big data. In order to handle this bigdata, hadoop frame which is capable of handling data in a distributed computing manner can be used. Hadoop frame work has two major layers, they are computational layer called Mapreduce and storage layer called HDFS.

Hadoop Distributed File System (HDFS), in hadoop frame to stores large datasets and stream these data at large bandwidth according to the user application. The hadoop system is using distributed computing so that large data can be processed parallely. Mapreduce affords a brand new technique of reading statistics which is complementary to the talents provided by SQL. The system based totally on Mapreduce that may be scaled up from single servers to hundreds of excessive and occasional cease machines. Mapreduce programming model contain Mapper function and reduce function. Map function takes the input as <key, value> pair and output <key', value'> as intermediate output. Reducer which combines the intermediate output with same keys and produces the output as <key, value> which is stored in HDFS.

Humans always influenced by others opinion and thinking. So after the evolution of ecommerce websites, they are crazy in purchasing things from e-commerce websites. The customer's opinion about the product can be put in the sites so that it may help the customers to evaluate the product and

even the manufacturers to refine their product based on customers satisfaction. From this huge amount of opinion it is difficult for a customer or the manufacturer to evaluate the feedback. This leads to the concept for research called sentiment analysis. For example in case of a company who want evaluate their product can find the drawbacks from customer review by opinion mining.

Three different techniques used for sentiment analysis are machine learning based, lexicon based and hybrid technique.

Machine learning method means make the system to perform in such a way as human can interpret. This method is again classified into supervised and unsupervised. Supervised method consists of training dataset and a test dataset. Training datasets are labelled data, based on the quality and quantity of training set the performance of the algorithm or method varies. Unsupervised and semi supervised method performs less compared to machine learning algorithm. Labelled datasets are unavailable for every domain so it is an impediment for the use of supervised method. Data sparsity is the other major concern in case of supervised. Unsupervised means the dataset is unlabelled and looking on different criterion's we will classify the data. One example for this method is clustering. Different supervised machine learning algorithms are present such as Support Vector Machine (SVM), Naive Bayesian, K Nearest Neighbours (KNN), Maximum Entropy, etc. Figure 1 shows the commonly used sentiment analysis techniques.

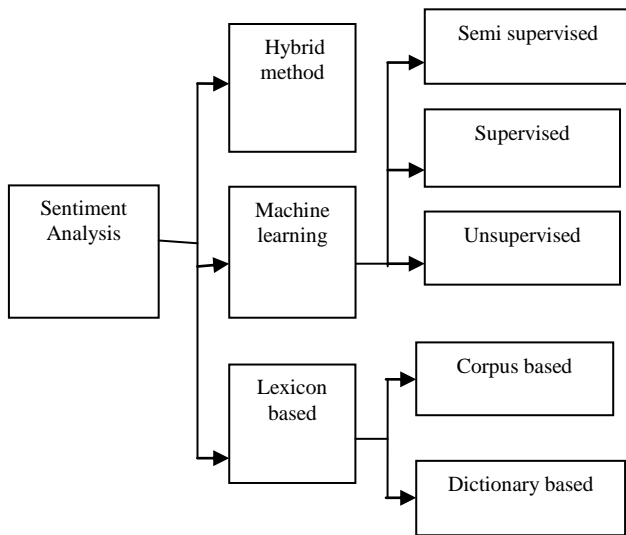


Figure1. Sentiment Analysis techniques

Lexicon based method classification sentiment lexicons are created whose sentiment weight is pre determined and compare it with the given text features. Sentiment lexicon consist of set of words and expressions ,used to show people’s subjective feelings and opinions. Dictionary method in which we are suppose to make the dictionary with lots of opinion words that might comes in an opinion sentence. Based on the frequency of the opinion words, the polarity is determined. Corpus based method is based on syntactic pattern; this can produce the result with more accuracy. Hybrid approach is the combination of machine learning and lexicon based method

II. RELATED WORK

Many works has been released about sentiment analysis in past years. Sentiment analysis is implemented for various applications using variety of datasets of various sizes and using various algorithms either supervised machine learning or unsupervised machine learning. Most of the existing sentiment analysis techniques focus only on aggregate level, classifying sentiments into positive, neutral or negative, and will loss the capabilities to perform fine-grained sentiment analysis. Zhaoxia WANG et al.[1] describes a social media analytics’s engine that employs a social adaptive and fuzzy similarity-based classification method to automatically classify text messages into sentiment categories as positive, negative, neutral and mixed, with the ability to identify their prevailing emotion categories such as e.g., satisfaction, happiness, excitement, anger, sadness, and anxiety. Normally people used to skip emoticons in case of sentiment analysis to avoid the complexity of processing, emoticons are normally classified in to 4 emotions like happy, sad, anxiety, neutral. In the paper titled Exploiting Emoticons in Sentiment Analysis [2] describes about the sentimental analysis they have performed by exploiting the emoticons. Normally these emoticons are used for intensification of emotions expressed in words or it is used to express an emotion if it

is not clearly specified in the text. Some cases emoticons are used when the sentiment associated with sentiment text is to be negated. Here the sentiment is purely based on emoticons only. Ebru Aydo et al.[3] conducted a survey on sentiment analysis and stated about two approaches used in sentiment analysis as machine learning and lexicon based. In machine learning based method machine learning algorithms are use, but in lexicon based counting and weighting of words are used. In case of machine learning algorithms out of the different algorithms SVM and Naive Bayesian is most commonly used because of its accuracy. Xiao Yang et al[4] proposed a simple Mapreduce performance model for understanding the impact of each component on overall program performance. In paper titled Scalable sentiment classification for big data analysis using Naive Bayes Classifier[5] evaluate the scalability of Nave Bayes classifier (NBC) in large datasets instead of using standard library Mahout. They found that NBC is able to scale up with increasing review data with increasing throughput.

III. SYSTEM DESCRIPTION

The development of technologies give rise to the frequent development of large amount of data thus came the concept of bigdata. To handle this huge data we are using a bigdata distributed frame work called hadoop. The hadoop system handles the program to run in a parallel manner by using mapreduce concept. Flow diagram of the system is given in figure 2.

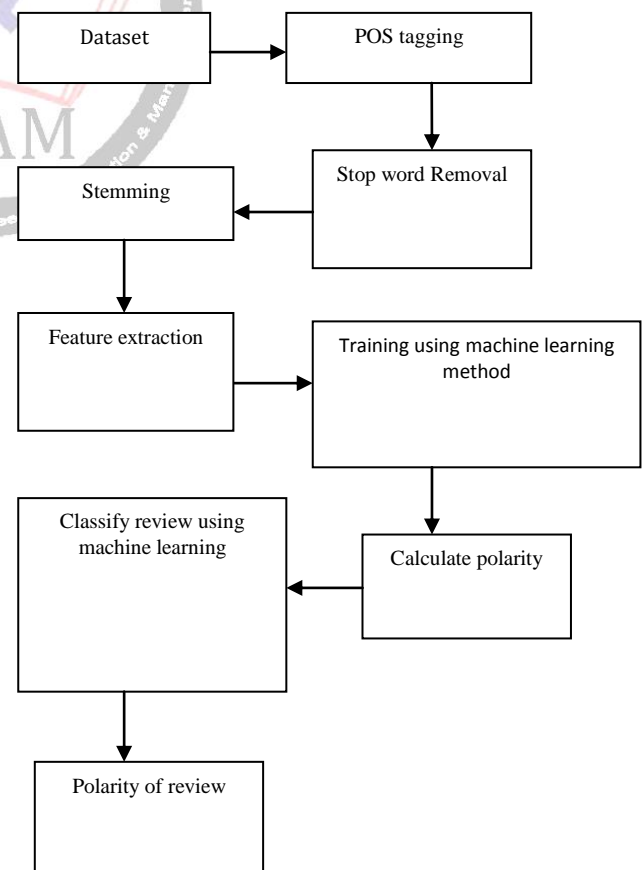


Figure 2: Flow of the sentiment analysis

A. Dataset

The dataset used here is the review data of laptops from Amazon. The dataset contains more than 10000 reviews. Due to the less availability of dataset which is suitable for machine learning algorithm I adopted this dataset for my experiment.

B. Preprocessing

This step involves pos tagging, stop words removal and stemming. The dataset normally in an unstructured manner or it may contain many unnecessary words which is not needed for our purpose. So what we first perform is tag the review sentences with part of speech such as noun, verb, adjectives etc. And stop word removal means removing the connectives in the sentences like ‘is’, ‘the’, ‘and’ etc. Word can be written in different format, so we need to normalize it in to standard form which is called stemming.

C. Feature extraction

Feature extraction is needed to extract the features about which the review is mentioning. Even though the overall review is considered for most cases, there are some people who are interested about the particular feature of the product. In such cases the polarity of the review about particular feature is considered, so that they can go for that product which they are pleased. Adjectives show the opinion about product feature. And the nouns notify the features of the product.

D. Training and Classification

Machine learning method contains a function that maps input to desired output, which is called training data or labeled data. This data are labeled by a human expert. Based on this training data the machine will classify the test data or unlabelled dataset using the method used in the algorithm. The algorithm we are using is the Naïve Bayesian classification algorithm.

Naïve Bayes Algorithm

Naïve Bayesian Classification represents a supervised machine learning method as well as a statistical method for classification. The Bayesian classifier is one of the probabilistic model works positively on text categorization and employed on Bayes rule. It is flexible in way of handling with any number of classes or attributes. For a given review d , C^* is a class variable which defines the sentiment given by

$$C^* = \arg \max_c P(c/d) \tag{1}$$

$$Pnb(c/d) = \frac{p(c) \cdot \prod_i p(f_i | c)}{P(d)} \tag{2}$$

Here f represent the feature and $n_i(d)$ represents count of specific feature f_i found in review d . $P(c)$ and $p(f/c)$ are maximum likelihood estimates.

In general we can represent as

$$\text{Posterior} = \frac{\text{Prior} * \text{Likelihood Evidence}}{\text{Evidence}}$$

While training, the algorithm counts the number of times each word appears in a review in the class and divides that by the number of words appearing in that class. Take individual word probabilities and multiply them together to determine the probability of a review given a class. Testing phase in which a new unlabelled review is given as input and based on the probability of the words in test data compared to the training data, it determines the category.

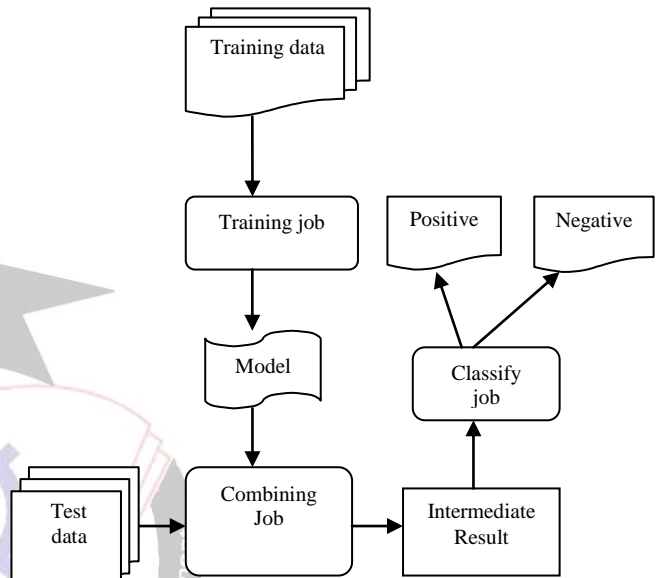
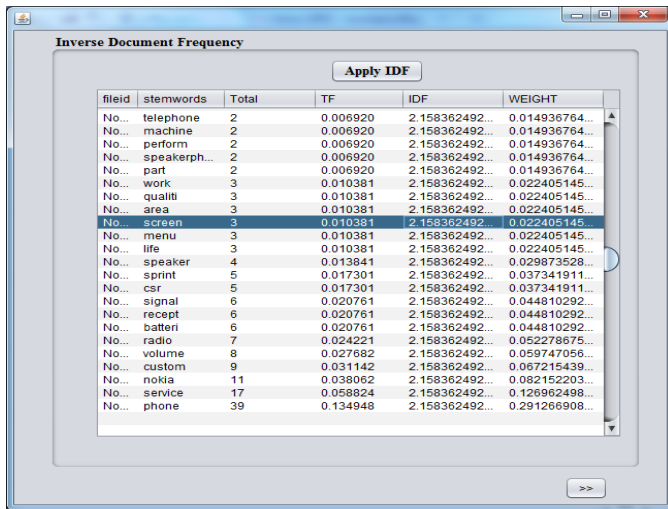


Figure 3. Naïve Bayes classifiers job sequence

Training job create a model that compute the probability of each word in the two classes. The combining job associates test data with the model, apart from the words that appear in test data but not in the model. The intermediate result produced by combining job is then given to the classify job. At the end of classify job, all reviews are classified into positive or negative classes.

IV. IMPLEMENTATION AND RESULT OF OBSERVATION

The experiment was conducted on a dataset of laptop review from amazon.com. The review dataset is of 20MB size. The dataset consist of training dataset, ie the labeled dataset and test dataset that need to be classified in to positive, negative or neutral. Hadoop framework is installed on VMware under Linux OS. And After saving the eclipse plug-in in Hadoop, code is run on eclipse. Figure 3 shows the partial output of my experiment which shows the extracted features in the review set.



fileid	stemwords	Total	TF	IDF	WEIGHT
No...	telephone	2	0.006920	2.158362492...	0.014936764...
No...	machine	2	0.006920	2.158362492...	0.014936764...
No...	perform	2	0.006920	2.158362492...	0.014936764...
No...	speakerph...	2	0.006920	2.158362492...	0.014936764...
No...	part	2	0.006920	2.158362492...	0.014936764...
No...	work	3	0.010381	2.158362492...	0.022405145...
No...	qualiti	3	0.010381	2.158362492...	0.022405145...
No...	area	3	0.010381	2.158362492...	0.022405145...
No...	screen	3	0.010381	2.158362492...	0.022405145...
No...	menu	3	0.010381	2.158362492...	0.022405145...
No...	life	3	0.010381	2.158362492...	0.022405145...
No...	speaker	4	0.013841	2.158362492...	0.029873528...
No...	sprint	5	0.017301	2.158362492...	0.037341911...
No...	csr	5	0.017301	2.158362492...	0.037341911...
No...	signal	6	0.020761	2.158362492...	0.044810292...
No...	receipt	6	0.020761	2.158362492...	0.044810292...
No...	batteri	6	0.020761	2.158362492...	0.044810292...
No...	radio	7	0.024221	2.158362492...	0.052278675...
No...	volume	8	0.027682	2.158362492...	0.059747056...
No...	custom	9	0.031142	2.158362492...	0.067215439...
No...	nokia	11	0.038062	2.158362492...	0.082152203...
No...	service	17	0.058824	2.158362492...	0.125962498...
No...	phone	39	0.134948	2.158362492...	0.291266908...

Figure 3 : Screenshot of partial output

The final output of this shows the classified review data in to 2 categories like positive and negative. Since review data contains many features about the product we will consider the features with a minimum support of 3. Work is in progress on rest of the part of sentimental analysis.

V. CONCLUSION

Sentiment Analysis deals with the function of finding the customers sentiment or opinion about the product. Sentiment analysis can be done in different ways. One such way is using machine learning algorithm. Here using the naïve Bayesian classification algorithm the dataset of laptop reviews are classified in to positive or negative. Since under hadoop frame we are implemented the input and output datasets are stored in HDFS. Output will specifies, about which feature the opinion is given. First we experimented this sentiment analysis on single system and got final output. Then extended it to perform on a 3 systems cluster. But with some errors, it is not giving expected output. Work is in progress on sentiment analysis on hadoop cluster.

REFERENCES

[1] Zhaoxia WANG, Chee Seng CHONG, Landy LAN, Yinping YANG, Seng Beng HO and Joo Chuan TONG : "Fine-Grained Sentiment Analysis of Social Media with Emotion Sensing" IEEE 2016.

[2] Alexander Hogenboom, Daniella Bal, Flavius Frasinca : "Exploiting Emoticons in Sentiment Analysis" Copyright 2013 ACM 978-1-4503-1656-9/13/03.

[3] Ebru Aydo ,M. Ali Akcayol "A Comprehensive Survey for Sentiment Analysis Tasks Using Machine Learning Techniques" IEEE 2016.

[4] Hao Wang, Jorge A. Castanon : "Sentiment Expression via Emoticons on Social Media" 2015 IEEE International Conference on Big Data (Big Data).

[5] Bingwei Liu, Erik Blasch, Yu Chen, Dan Shen and Genshe Chen: "Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier" 978-1-4799-1293-3/13/\$31.00 ©2013 IEEE.

[6] Manvee Chauhan, Divakar Yadav, " Sentimental Analysis of Product Based Reviews Using Machine Learning Approaches" ©EverScience Publications ISSN:2395-5317.

[7] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," in Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation - Volume 6, ser. OSDI'04. Berkeley, CA, USA: USENIX Association, 2004.

[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing- Volume 10, 2002, pp. 79–86.