International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 5 (2017), pp. 1083-1091

http://www.ripublication.com

Sentiment Analysis of Customers using Product Feedback Data under Hadoop Framework

Ms. Soumya P.C[1] and Dr. Venkatramana Bhat P[2]

[#]*Department of computer Science and Engineering, Mangalore institute of Technology and Engineering, Moodabidri, Karnataka, India.*

Abstract

Sentiment Analysis (SA) also called opinion mining is the recent trend in a day. 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 this paper, under hadoop frame work, sentiment analysis of product review using dataset from amazon.com was performed. The dataset contains reviews composed of both text and emoticon (smiley). Generally, sentiment analysis considers only text. Here we are considering both text and emoticon and found that it gives more accuracy than sentiment analysis with only text.

Keywords: Sentiment Analysis, HDFS, Emoticon analysis, polarity.

1. INTRODUCTION

Bigdata is the concept evolved when the technology is advanced and the data generated in each second is enormous. The data generated is of large variety, volume and velocity it can't be handled by using normal RDBMS. In order to handle this bigdata we are using hadoop frame which is capable of handle data in a distributed computing manner. Hadoop Distributed File System(HDFS), in hadoop frame is there to store 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. Data access from HDFS is not that flexible, so everyone is

going for Apache hive or pig. Apache Hive is a data warehouse to process structured data in Hadoop. Hive is installed on top of Hadoop to consolidate Big Data to make querying and analysing easy. 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. Reviews are normally a combination of texts and smiley. Reviews composed of many words which convey an opinion and smiley's which express the feelings in visual format. This smiley's which gives clues about the customer's emotion on product. From this huge amount of opinion it is difficult for a customer or the manufacturer to evaluate the feedback. This gives the scope to the researcher in the field of sentiment analysis. Many researchers mentioned that the sentiment analysis of data that exploiting the emotions give more accuracy than the sentiment analysis of data without considering emotions in that. Sentimental analysis (SA), also known as opinion mining, is the analysis of feelings behind the words using the method of natural language processing. SA involves classifying the review in to three categories like positive, negative and neutral. Sentiment analysis can be done in different methods, either it can be document level sentiment analysis, sentence level and aspect or entity level. In document level sentiment analysis, the document should concentrate on single object only. And classify the document based on polarity. Sentence level sentiment analysis is the case in which each sentence is taken as separate unit and finds opinion of each sentence. It first classifies the sentence as either it contains fact or opinion. If it is opinion then find it and check the polarity. In case of aspect level we need to find different features of product and sentiment about that aspect.

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. Fig 1.1 shows the different commonly used sentiment analysis techniques.

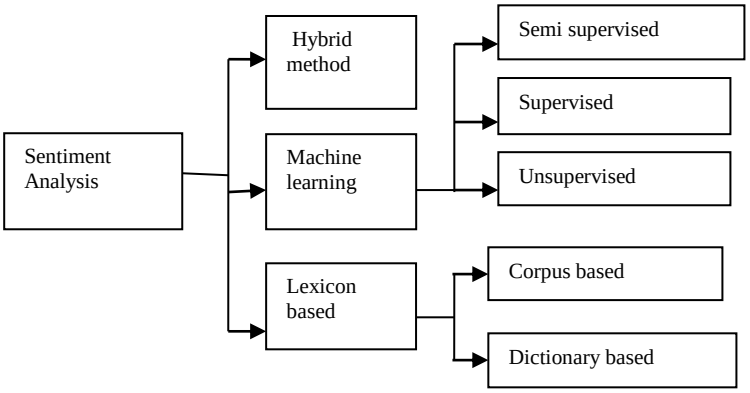

Fig 1.1 Sentiment Analysis techniques

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. In the paper titled Sentiment Expression via Emoticons on Social Media[4], made an analysis of the effect of emoticons in the sentiment analysis. And the analysis found that :-) and :-(is the most frequently used emoticon in the social media. When the sentiment analysis of data with and without emoticon is performed, it came to found that accuracy of sentiment is more for data with emoticons. Xing Fang and Justin Zhan[5] performed sentiment analysis on product review dataset collected from amazon.com. The detailed description about the processes in sentiment analysis is given. Both sentence level and review level categorization has performed here. The classification methods used are SVM, Random Forest and Naive Bayesian. Xia Hu et al.[6] proposed Emotional Signals for unsupervised Sentiment Analysis (ESSA), a frame work to perform sentiment analysis on data with emoticons in an unsupervised manner. For that they made study on emotion indication and emotion correlation. The paper tittled Emoticon Analysis with Dynamic Text based Opinion Mining [7] introduced a new algorithm which automatically add new words present in the text data while going for a dynamic sentiment data analysis , so that the dictionary itself is dynamic. The polarity of newly added words depends on the neighboring words polarity. Finite State Machine (FSM) is used to find the emoticons polarity. Michał Skuza[8] designed a system , that estimate and predict the future stock price in hadoop distributed environment. The datasets was the twitter micro blogging. Based on the users comments about the stock price the future will be predicted like whether it is going to raise or goes down. Lada Banic et al.[9] which analyze the hotel reviews and gives sentiment about hotel as the outcome. They used a tool called KNIME software.

III. PROPOSED SYSTEM

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. The proposed method is to find the sentiment analysis of customer review by exploiting both text and emoticon under hadoop frame work. The data is normally stored in hadoop distributed file system, but here we are using Apache hive as the data warehouse, because of its ease of accessibility.

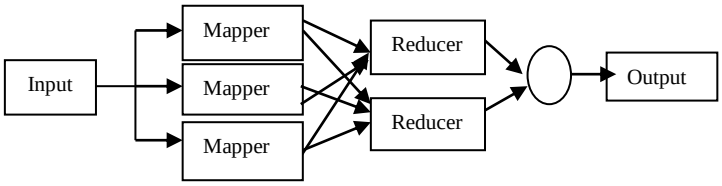

Fig 2.1: Mapreduce Programming generic flow

The diagrammatic representation of mapreduce programming is given in fig 2.1. The program is divided into mapper and reducer part. All the mapper parts execute first at once and the output of mapper is given as input to the reducer part.

Fig 2.2 shows the flow diagram of sentiment analysis.

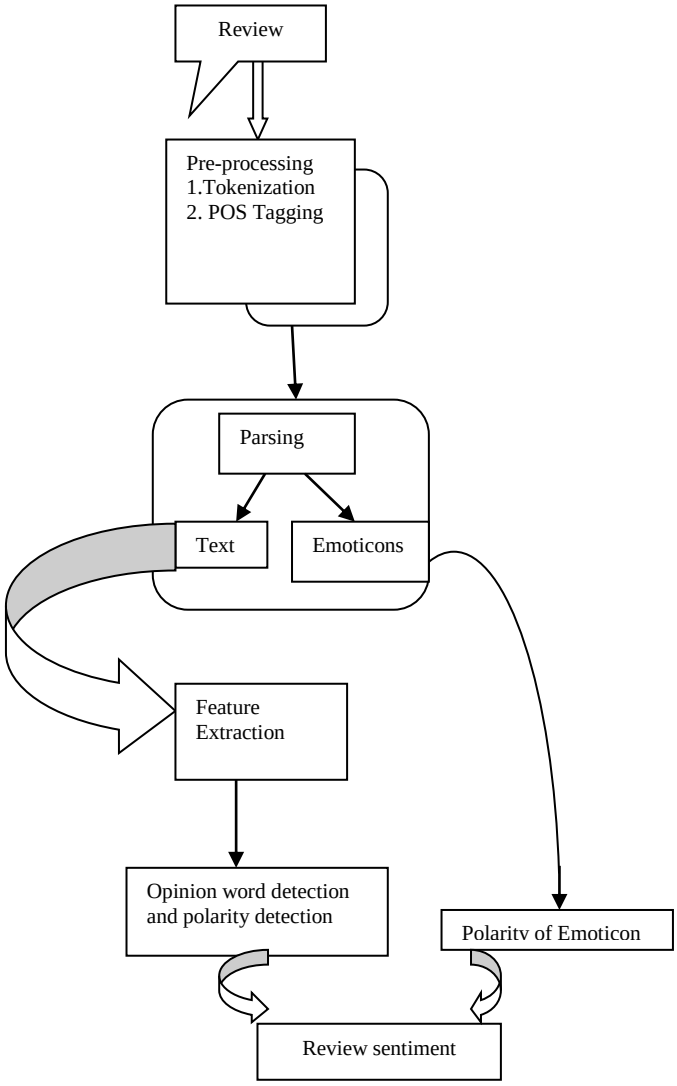

Fig 2.2: Flow diagram of proposed system

A. Dataset

The data used is the review data of mobiles collected from Amazon website which is in json format. Sentiment analysis on multiple domains normally fails in some situations, so selected the specific domain as mobile. The review dataset is composed of both text and emoticons.

B. Data Preprocessing

Data pre-processing consist of two steps tokenization and pos tagging. Tokenization is the process of segmenting the sentences in to chunks such as punctuation, words, numbers, etc. after removing stop words and white space. Tokenization is done by locating the word boundaries. Pos tagging is the process of tagging each words as part of speech elements like adjective, noun, verb, adverb etc. In sentiment analysis pos tagging is important because this helps to identify sentiment words and feature element of product. Normally nouns, pronouns became features and adverbs, adjective contain sentiments. Parsing of this processed data is used to separate the emoticons from it. Emoticons are classified in to two types like positive emoticons and negative. We are considering only limited emoticons from the vast variety.

Table 1 show the common type of positive and negative smiley's used in the customer review.

Table 1. Examples for emoticons

Positive emoticons	:) :D :] :* ;) ;] 8) <3 ;P =P =) B) :-> :-D XD =3 ;> ;-) :‟) =>
Negative emoticons	:(:‟(:-(=[:-@ :- :-[:O :-/ :-\ :/ :\ :”(:-O :[:<

C. Feature Extraction

Feature extraction is the process of extracting features of the product from the review. Since the dataset selected is mobile we are extracting the features of the mobile. During pos tagging the tokens are tagged with their corresponding pos, a product feature is a noun or noun phrase which is appeared in review sentences. It can be inspired that the nouns with high frequency can most likely be considered as feature words. Frequent pattern mining techniques tend to determine multiple occurrence of the same item. So we are using Apriori algorithm to find the features. Using the POS associated with words during the data generation phase, we created a new view of data where each sentence in the review was considered as a bag of words. The words chosen to represent a sentence were those that were marked as nouns (NN/NNS). We chose to ignore proper nouns, which we believe cannot be features associated with a product. Furthermore, as we observed in the data, some phrases that represent features were made of two classes of words, nouns and adjectives. So as to be able to detect such features, we also included words marked as adjectives (JJ/JJR/JJS) in our bag of words model. The next step was to extract frequent features from the candidate

feature words. We used an implementation based on the APRIORI algorithm for identifying frequently occurring word/word pairs from the bag of words data model

D. Polarity Detection

Polarity of text and polarity of emoticon comes under this category. In case of text polarity categorization, the POS associated with the words as adjectives, adverbs, and verbs are words that mainly convey sentiment. Dictionary method in which each positive and negative word in the dictionary is given with a sentiment weight ranging from -2 to 2. Based on the frequency of these positive and negative words in the review, we calculate the sentiment score of the text. For emoticon also we will provide weight of -2 for negative emoticon and 2 for positive emoticon. The case in which positive emoticon and negative text sentiment means and vice versa we will take it as negative. In more logical way we can tell it's as we are doing logical AND operation.

Table.2.2: How sentiment is calculated when text and emoticon give opposite sentiment

Review	Text sentiment	Emoticon sentiment	Total sentiment
I love my phone :-)	+	+	+
I hate the screen size :-)	-	+	-
My phone is not hanging :(	+	-	-
Battery Draining fast :D	_	+	_

IV. EXPERIMENTAL SETUP

The sentiment analysis of review data by exploiting both text and emoticon is evaluated by means of set of experiments. The experiment is conducted using a dataset of 6ooo reviews about phone accessories. The computer with minimum 4 GB Ram with Linux operating system installed on it is the prerequisite. The system is installed with hadoop 2.6 version. For the hadoop installation sort of steps are followed referring the link www.bogotobogo.com/Hadoop. On top hadoop, hive installation is done.

The first step of this sentiment analysis is the data preprocessing which includes the tokenization and pos tagging. The tokenizer used for experimental setup is string tokenizer. Using Stanford pos tagger the tokenized output of raw data is tagged according to their corresponding pos. Since the dataset contain both text and emoticon

the data to be parsed to separate the emoticon from it. Inorder to find the features of the product from review, an Apriori algorithm is used to find the frequent noun that is mostly used. Nouns are word without any sentiments, they are normally features. So from the output of Apriori we can identify the most important features about which the customers are talking. The method here adopted to find the review polarity is dictionary method. The dictionary includes more than 2500 positive and negative words. Based on the occurrence of these positive and negative words in the review, the text is classified in to positive or negative. Similarly the dictionary of frequently appearing smiley is used to find smiley's polarity. The work regarding this experiment is going on.

A. *Experiment result*

The final output screenshot is given in the fig: 2.3 in which one map task failed to run. Now am working on the error to rectify it. By future will make the code run to give the exact output of sentiment analysis as positive, negative and neutral.

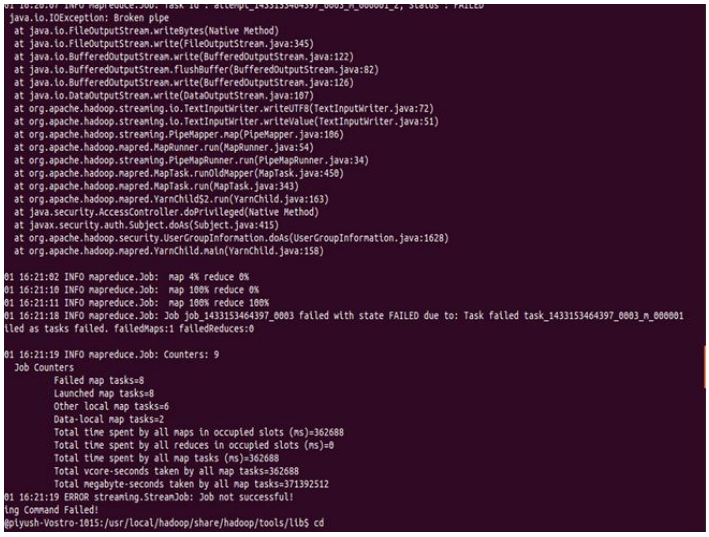

Fig: 2.3 Output screenshot of sentiment analysis in which one task failed.

V. CONCLUSION

Sentiment Analysis is analysis of feelings behind the words using the method of natural language processing. Using the dictionary method SA is done but it was not successful, it failed to run one map task. Am working on the project to make it run properly and get the expected output.

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 Frasincar :"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 UsingMachine 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] Xing Fang and Justin Zhan :"Sentiment analysis using product review data" ,journal of big data Springer 2015.

[6] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu :"Unsupervised Sentiment Analysis with Emotional Signals " ACM 978-1-4503-2035

[7] Sanam Kadge Saba Panchbhai :"Emoticon Analysis with Dynamic Text based Opinion Mining" International Journal of Computer Applications (0975 – 8887)

[8] Michał Skuza Andrzej Romanowski :"Sentiment Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction"

[9] Lada Banic ,Ana mihanovic:-"Using Big Data and Sentiment Analysis in Product Evaluation" IEEE 2013

[10] M. Edison , A. Aloysius :"Concepts and Methods of Sentiment Analysis on Big Data" IJIRSET.2016.