Journal of Network Communications and Emerging Technologies (JNCET) www.jncet.org
Volume 5, Special Issue 2, December (2015)

LEXICON BASED OPINION MINING SYSTEM USING HADOOP

Dr. Neelam Duhan
Department of Computer Engineering
YMCA University of Science & Technology Faridabad, India

Amrita Kaur
Department of Computer Engineering
YMCA University of Science & Technology Faridabad, India

Sangeeta
Department of Computer Engineering
YMCA University of Science & Technology Faridabad, India

Abstract – With the advent of web 2.0, huge dimensions of opinionated text is now available on web. To extract feeling of the mass about an object from this huge web, automated opinion mining system i.e. algorithmic method of analysis of large number of reviews is thus needed. This paper carefully addresses the significant challenges and presents a system for analyzing sentiments which aims at resolving many of listed issues. The proposed system uses a lexicon based sentiment analysis approach to categorize the text as positive, negative and neutral in a fast and accurate manner. In this work, sentiment analysis is performed on a large data set of tweets using Hadoop and the performance of the technique is measured in form of the speed and accuracy.

Index Terms – Sentiment Analysis, Opinion Mining, Hadoop, Hive, Flume, Lexicon.

This paper is presented at International Conference on Recent Trends in Computer and information Technology Research on 25th& 26th September (2015) conducted by B. S. Anangpuria Instituteof Technology & Management, Village-Alampur, Ballabgarh-Sohna Road,Faridabad.

1. INTRODUCTION

With the advent of web 2.0, the web content has drastically changed to also contain user posted reviews from just containing facts and figures. As more and more people are becoming familiar with web and using e-commerce, social media websites, writing blogs, the content containing sentiments and opinions about a product, service, event, and person has increased and so it is increasing in an exponential manner. Therefore, in today's web world, textual information which is present on the web can be chiefly categorized into two broad categories, facts and opinions. Facts present the objective statements about the events and objects, which are not subject to change from person to person, whereas opinions or sentiments are the subjective statements that reflects a view or judgment formed by a person about something, not necessarily based on fact or knowledge.

Most of the existing research on text information processing has been exclusively focused on mining and retrieval of factual information, e.g., information retrieval, Web search, and many other text mining and natural language processing tasks. A Little research work has been done on the handling of opinions until only recently. However, opinions are so important that whenever one has to make a decision, one wants to hear others' opinions. This is not only true for individuals but also for organizations.

Unlike previously, when for making a decision one used to ask his family and friends and also organizations used to conduct expensive surveys and focused groups, now, if one wants to buy a product or a service, and for organizations to know user review about its products and services, it is no longer necessary to ask one's friends and families and to conduct surveys and hire external consultants because there are abundant product reviews available on the Web which gives the opinions of the existing users of the product. Therefore, sentiments and opinions is of great importance in making decisions and so is sentiment analysis.

Sentiment data is unstructured data that represents opinions, emotions, and attitudes contained in sources such as social media posts, blogs, online product reviews, and customer support interactions. While, Sentiment Analysis is the technique to extract subjective information from text and determining the overall contextual polarity or opinion of the writer of the text [1]. In short, Sentiment Analysis is the process of identifying the contextual polarity of text i.e., it determines whether a piece of writing is positive, negative or neutral. An alternative term is opinion mining, as the process determines the opinion, or the attitude of a speaker about a topic from a given piece of text. Therefore, in this paper both 'sentiment analysis' and 'opinion mining' has been used alternatively.

Sentiment Analysis has applications in numerous areas. For example, in marketing it helps in determining the success or failure of a new product launch or any new commercial promotion or determining which version of a product is liked more in which part of the world. Various companies can use this data to determine their future strategies regarding a particular product or service.

Sentiment analysis can be performed on a document, sentence or a phrase. In document based sentimental analysis, sentiment of the complete document is calculated as a whole and summarized according to its polarity. In sentence based sentiment analysis, each individual sentence is classified as positive, negative or neutral [2] whereas, in phrase based sentimental analysis, polarity is assigned to each individual phrase contained in a sentence.

The process of sentiment analysis starts with determining the subject towards which the opinion is expressed. After that the sentiment is classified as positive (which denotes satisfaction or gladness on behalf of user), negative (which shows rejection or dissatisfaction) or neutral (which denotes no strong sentiment involved). Then the sentiment can be given score which denotes the degree of positive or negative response from the user.

The done literature survey has showed two major techniques for sentiment analysis: machine learning techniques and lexicon based techniques. However hybrid of the two has also been tried and tested to obtain the best of both the worlds [3] [4] [5]. The machine learning methods for sentiment analysis often rely on supervised classification methods, in which, labeled data is used to train classifiers. Several research papers have been published which exhibits machine learning approach for analyzing sentiments [1]. In lexicon based approach, a domain dependent opinion lexicon is maintained, which is precisely nothing but a piece of text or document containing some opinion words or phrases and idioms. The opinion lexicon is used for classifying the sentiment of each word in the piece of text under analysis. The lexicon based approach is based on the assumption that the sum of the sentiment orientation of each word or phrase gives the total contextual sentiment orientation. The literature survey depicts the implementation of this approach in [6] and [7].

2. CHALLENGES INVOLVED WITH SENTIMENT ANALYSIS

Sentiment Analysis problem is essentially a text based analysis problem, but the challenges it involves, makes it challenging as compared to traditional text based analysis. This section lists the challenges found from the extensive literature survey in performing sentiment analysis. The challenges broadly include challenges to handle huge data, and challenges related to Natural Language Processing (NLP). Following are the challenges in performing sentiment analysis on a large scale:

2.1. Big Data

The sentiment data collected from web is of large volume and is considered as Big Data. Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications [3]. The challenges include how to capture, store, search, share, transfer, analyze and visualize enormous amount of varied data. Because sentiment data is collection of opinioned data from varied sources amounting to a vast collection of data, the sentiment data hence is now considered as big data which requires tools that are capable of handling such huge amounts of data efficiently.

2.2. Unstructured data

Sentiment Data is retrieved from various varied sources such as social networking websites, blog posts, product review sites, forums hence it is unstructured data, which makes it difficult to process and analyze. The first requirement is to refine it and convert it in to a structured data form like tables. Also most of the previous work to structure the unstructured data doesn't scale to large data sets efficiently.

2.3. Negations

Negation plays a vital role in altering the polarity of the accompanying adjective and hence the polarity of the complete text. Negation words in general, includes not, neither, nor. One possible solution to handle negations is to reverse the polarity of the adjective happening after a negation word, for instance, "The restaurant is good" (should be

classified "positive"), and whereas, “The restaurant is not good” should be classified “negative”.

However, this solution fails to entertain the cases like "No wonder the restaurant is good" and "Not only the food was yummy, the interior and service was also excellent". The use of pure language processing techniques or pure use of mathematical models fail to tackle negations completely.

2.4. Domain Generalization problem

Certain words exhibit different polarities when used in different domains.

For example, "The movie was inspired from a Tollywood movie" (negative orientation), and whereas, "I got inspired from the novel"(positive orientation)

Here, the word 'inspired' displays two different polarities for two different contexts.

Sentiment Analysis is usually carried out aiming a specific domain and this has even depicted very good outcomes of accuracy. But a generalized sentiment analyzer still remains a challenge because of difference in the meaning of a word/sentence in varied domains.

2.5. Language Generalization

The survey presents a language problem which says to perform sentiment classification, a separate dictionary is required for each different language and each varied domain. So far, sentiment analyzers have been implemented for a certain language only. A language general sentiment analyzer is, however, a fruitful idea as it gives a broad view of opinion towards a product.

2.6. Mapping Slangs

Slangs are usually the informal short forms of original words often used in online texting. For instance, ‘gr8’ is a slang used for opinionated word ‘great’, ‘5n’ or ‘fyn’ for ‘fine’.

These words are not a part of traditional language dictionaries but are found extensively in online texts. If slangs can be mapped to original words, the performance results of sentiment analysis can be improved.

2.7. Emoticons

The literature survey shows that emoticons are extensively used in web texts by the reviewer or blogger to express their emotions. Text can in fact be handled, however, handling emoticons is a more difficult challenge. Since the emoticons express sentiments and so are of great importance. For example, “I loved the food at ‘xyz’ restaurant :)”

Here, ‘:)‘expresses positive sentiment and so this must be considered as opinionated word in the process of sentiment analysis.

2.8. Noisy Data

The literature survey has revealed that the sentiment data available on the web contains vast quantity of irrelevant stuff and spams like URLs to other page links, tagging other users, misspellings, etc. These kind of content brings complexity and lack of accuracy to the sentiment analysis system. For example, “ @happyharry , The food was awsom today”: should be positive with score one(if ‘owsom’ has been successfully classified as positive), but the system would fail to quantify it correctly and incorrectly scores the sentence with two as it also takes ‘happy’ in the tagged name as a positive word. Hence, gives inaccurate results.

The identified challenges motivates to bring up a solution to all the problems stated in the above section. Following are the objectives of the proposed approach:

- To implement an algorithm for automatic classification of text into positive, negative or neutral.
- Graphical representation of the sentiments
- Deal with the huge volume of data in a speedy fashion
- Handle misspellings and noisy data accurately
- A method that can transform unstructured data to structured format

The focus of this paper is to put forward an approach that can perform sentiment analysis quicker because there is vast amount of data which is required to be analyzed. Also, it has to make sure that accuracy is not compromised too much while focusing on speed.

The proposed approach is a lexicon based technique i.e. a dictionary of sentiment bearing words along with their polarities was used to classify the text into positive, negative or neutral opinion. The subsequent section contains the elaborated view of the proposed sentiment analysis system.

3. SOLUTION TO ISSUES OF BIG DATA SETS

For huge Datasets, it is impossible to mine them on a single machine. Therefore, to mine and then analyze such huge data sets, here, twitter data, some parallel and distributed computing has to be used as a solution. But this solution itself comes with a set of problems along with it, which are as follows:

- Load balancing

- Data partition and distribution
- Jobs assignment and monitoring
- Parameter passing between nodes
- Handling failed nodes and their assigned work
- Synchronization among multiple machines

Therefore, a complete solution is needed which has the power of distributed and parallel computing along with fault tolerance. Apache Hadoop is one such solution to the challenges in Big Data.

Apache Hadoop is a software framework that allows for the distributed processing of large data sets across clusters of commodity computers using a simple programming model. Apache Hadoop is an open-source software framework for distributed storage and distributed processing of Big Data on clusters of commodity hardware [8], to efficiently process large volumes of information in parallel. Apache Hadoop framework comprises of various tools to load, examine, process large data sets in a fault tolerant fashion.

4. PROPOSED OPINION MINING SYSTEM

The proposed opinion mining system is based on lexicon based approach for classifying sentiments. English lexicon is built and used for movie domain. The data to be analyzed for the experimental purpose is collected from twitter. Only the unigram adverbs, adjectives, nouns, emoticons and slangs are taken into account as features to sentiment analysis process. The collected data is in semi structured form, which is then converted to structured form and then analyzed. Sentiment Analysis is then done on the structured data, which results into the classification of tweets into positive, negative and neutral categories. Analysis from the results then can be done as per the requirements and domain used. For instance, high volume of positive tweets about a movie may indicate success of a movie. The analysis can be done in a better way from the visualization of the resultant data. The visualization can be in the form of graphs, pie-charts, maps or any other form as per the domain of the data and required analysis type.

4.1. The Proposed System Architecture

The proposed opinion mining system takes into account the existing challenges in sentiment analysis and tries to overcome them. The movie domain is considered here for analyzing sentiments and Twitter data is used for analysis i.e. to classify a tweet regarding a movie as positive, negative or neutral sentiment. As mentioned above, the best web source for collecting opinions or sentiments about a movie from audiences is Twitter because of its large number of active users, which are available in diverse variety from a common man to a celebrity and also for the most recent and updated opinions. However, if Twitter is used as a data source then the data is termed as Big Data, because of the huge number of active users and tweets made by them. So, to deal with the Big Data issues, Hadoop and its ecosystem is used. Apache Flume, Apache HDFS, and Apache.

Hive are used to scheme an end-to-end data pipeline that enables to analyze Twitter data. For getting raw data from the Twitter, Apache Flume is used which is a Hadoop online streaming tool. In this tool only the data to be fetched from Twitter is configured. For this the configuration is needed to be made in the flume configuration file and also what information is needed to get form Twitter is also defined. The fetched data will be saved onto HDFS (Hadoop Distributed File System) in the prescribed format. The raw data is firstly preprocessed to get a cleaned data i.e. data free from misspellings, URLs, tagged names. From this cleaned raw data a table is created using Hive QL (Hive Query Language). Once the data is structured, the Sentiment Analysis is performed by using some UDF's (User Defined Functions), which is done by taking MPQA Subjectivity Lexicon as the data dictionary and manually adding emoticons, slangs to it to form the complete data lexicon. The lexicon is taken into account so as to decide sentiment type and score for each word. The total sentiment score of a tweet is the sum of the sentiment score of each opinion word in the tweet. Since, Big Data is more valuable when visualized and analyzed. Finally, the computed sentiments are visualized so as to analyze the Big Data in a fruitful and clearer manner. The following figure, Fig. 3.1 clearly shows the architecture view of the proposed system by taking the above steps into consideration:

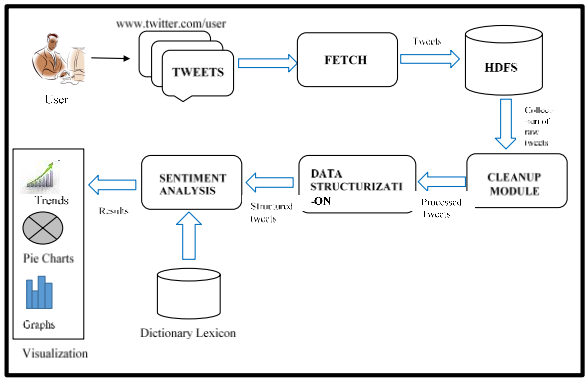

Figure1 Proposed Framework for Opinion Mining System

In the following sub sections, the detailed description of the functional components of the proposed architecture is given.

4.2. Functional Components of the Proposed Architecture

To overcome the limitations that are present in the so far existing work, the proposed architecture is presented in the above section. The detailed description of each of the involved modules is given in this section.

Twitter is an online social networking service that enables users to send and read short 140-character messages called "tweets". In general, the users registered on Twitter posts Tweets regarding any topic of their interest. Registered users can read and post tweets, but the tweets can be only read by unregistered users. Twitter has more than 500 million registered users, out of which more than 284 million users are active users. Also, more than 500 million tweets are generated every day [9]. The tweets were set to a largely constrictive 140-character limit for compatibility with SMS messaging, introducing the shorthand notation and slangs commonly used in SMS messages. Tweets are frequently used to express a tweeter's emotion on a particular subject.Twitter audience varies from common man to celebrities. Users often discuss current affairs and share personal views on various subjects. The advantage tweets provide is that they are small in length and hence unambiguous. There are firms which poll twitter for analysing sentiment on a particular topic.The challenge is to gather all such relevant data, detect and summarize the overall sentiment on a topic. Using Twitter for Sentiment Analysis proposes several challenges such as:

• Tweets are highly unstructured and also non-grammatical:

The tweets are short length texts , which encourages the use of sentences which lack the correct grammatical rules. Moreover, there is no particular format in which all the users write tweets. The proposed opinion mining system handles the unstructurization using Hive QL to structure the tweets into a tabular form.

• Out of vocabulary words:

The out of vocabulary words are the slangs which have not been added to the basic dictionary. To handle out of dictionary words, the proposed approach handles the most usually used and popular slang words by adding them in the dictionary lexicon.

• Lexical variations:

A word belonging to Engish vocalbulary may be written in varying ways by different users. To handle the misspellings, spell correction module is introduced before the tweets are set to analysed.

• Extensive usage of acronyms like asap, lol, rofl, etc:

Acronyms such as asap, lol, rofl are higly used in the tweets. These acronyms must be handled in order to add to the overall accuracy of the system. These are handled in the proposed system by adding the a wide list of acronyms into the dictionary lexicon.

Now, the proposed approach is to be implemented on these tweets using the following functional components:

1. Fetch Module

2. HDFS

3. Cleanup Module

4. Data Structurization

5. Sentiment Analysis

6. Visualization

A data structure called Lexicon Dictionary is also employed in the proposed work.

The components and data structures are explained below in detail:

4.2.1. Fetch Module

This module deals with the fetching function of the system, in which the tweets related to a certain mentioned movie (topic) are fetched from the www.twitter.com. Data analysis is only half the battle; bringing the data into a Hadoop cluster is the first and essential step in any Big Data deployment.

The twitter data is fetched into the Hadoop cluster using Apache Flume which uses an elegant design to make data loading easy and efficient. The Fetch module has two sub-parts: The Twitter Application and the Flume. So to get the Twitter data, a Twitter Developer account is needed to be created and a new application is to be created. After creating a new application, the access tokens are created which are fed to the flume's configuration file so that the need to authenticate oneself to the twitter while downloading data is eliminated. Also, a pair of consumer key and consumer key secret is provided which is required to access and get the Twitter data. These details are fed into the flume configuration file called 'flume.conf'. Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS) [10]. It has a flexible yet simple architecture based on streaming data flows; and is fault tolerant and robust with reliability mechanisms for failover and recovery.

4.2.2. HDFS-Store

The fetched tweets are stored in HDFS (Hadoop Distributed File System) so as to gain the advantages of distributed and reliable storage, which is needed for processing the Big Data. HDFS is aimed to reliably store very large files across machines in a big cluster. It is inspired from the Google File System. In HDFS, each file is stored as a sequence of blocks, wherein, the size of all blocks in a file are of same size except the last block. However, the default block size is 64MB. As the file is uploaded to HDFS, on one node in the cluster, each file block is stored. Blocks belonging to a file are then replicated to attain fault tolerance. The block size and replication factor are both configurable per file [14].

HDFS cluster is formed by a cluster of Data Nodes and a single Name Node. A Name Node performs a number of file system operation such as close, open rename etc. Block replication is also managed by Name Node. However, the read and write operations that file system clients require are run on Data Nodes. A file is divided into a number of blocks that is stored in a different Data Node for each different block and HDFS is responsible for determining mapping between Data Nodes and blocks [11].

4.2.3. Cleanup Module

Now once the twitter data has been fetched and stored in the directory on HDFS, before starting with the analysis, the data must have been cleaned in order to remove unnecessary stuff and make the data less noisy. The raw data that is stored and downloaded from Twitter is in the JSON format. JSON is a light weight data interchange format. Also, it allows to overcome the cross domain issue. JSON is a text format which is totally language independent.

This module converts noisy data to less noisy one. Noisy Data is the data which has words written in shorthand notations and may also have URLs' to some other web page and username tags. The cleaning of data is done in 4 steps:

Step 1: POS tagging

Step 2: URLs are removed

Step 3: Tagged names are removed

Step 4: Work for misspellings using edit distance

4.2.4. Data structuring

The data which is fetched, stored and cleaned so far is in semi-structured form. Analysis of such form of data is bit difficult because to analyze a data, the data must first be quantified and to quantify, the data must have a defined structure. Therefore, to quantify and analyze the Twitter data, it is firstly converted to a structured form. The unstructured tweet data is converted to structured form using Hive. Hive offers a database query interface to Apache Hadoop. It provides a SQL-like language called HiveQL for efficiently performing SQL-style queries on huge data sets which is not efficiently done using relational databases [12].

Since, the map-reduce programming model is very low level model which necessitates developers to write custom programs for each application, which are hard to maintain and reuse. Therefore, Hive comes with a capability of automatically converting the queries written in HiveQL to the corresponding map reduce jobs, thus providing with reusability and easy maintenance and customizations [13].

4.2.5. Sentiment Dictionary

Sentiment Dictionary is the dictionary of sentiment bearing words in the English language. Sentiment bearing or opinionated words are the words which portrait some sign of either positivity or negativity. The dictionary lexicon plays a significant role in the proposed approach. Unlike the relatively small sized dictionaries, which are used in machine learning approaches, the dictionary lexicon used in the proposed approach contains a huge set of opinionated words along with their polarity and corresponding polarity strength.

The computing capability of Hadoop enables us to use a huge set of words, which is precisely used for comparing each word in the tweet to this pool of words, hence eliminating the need training data, which saves a lot of time. Also, the dictionary lexicon is kept in the primary memory of the machine i.e. local to the program so that the time is not lost in searching a word in the secondary memory.

The proposed dictionary lexicon contains several different adjectives, nouns, negation words, emoticons etc.

The Dictionary Lexicon used in the proposed approach exhibits the following properties:

1.The dictionary is domain specific i.e. the polarities of the words in the dictionary are set according to a specific domain e.g. book reviews, political blogs etc. The dictionary used in this approach is made for movie review domain. The dictionary lexicon used is always domain specific because same word in different domains can have different meanings.

2.The dictionary contains all forms of a word i.e. every word is stored along with its various verb forms e.g. applause, applauding, applauded, applauds. Thus, eliminating the need for stemming each word, which saves more time.

3.Emoticons are generally and extensively used by people to present emotions. Hence, it can be concluded that they

possess very useful sentiment information in them. The dictionary used in the proposed approach contains more than 30 different emoticons along with their polarities and strength.

4.The Dictionary also contains the strength of the polarity of every word. Some word portrays stronger emotions than others. For instance, 'good' and 'great' are both positive words but 'great' depicts a much stronger emotion.

5.Negation and blind negation are very important in quantifying the sentiments, as their presence can alter the polarity of the sentence. Negation words are the words which inverse the polarity of the sentiment involved in the text. For example 'the movie was not good'. Although the word 'good' depicts a positive sentiment the negation – 'not' reverses its polarity. In the proposed approach whenever a negation word is encountered in a tweet, the polarity of the computed tweet is reversed.

Blind negation words are also maintained in this dictionary lexicon. Blind negation words are the words which operate on the sentence level and points out a feature that is desired in a product or service. For example in the sentence 'the acting needed to be better', 'better' depicts a positive sentiment but the presence of the blind negation word- 'needed' suggests that this sentence is actually depicting negative sentiment. In the proposed approach whenever a blind negation word occurs in a sentence, which is identified using the defined dictionary lexicon, its polarity is immediately labelled as negative.

4.2.6. Sentiment Analysis

Since the data is now structured, analysis of the data can then be done. A UDF (User Defined Function) is used for performing the sentiment analysis on the tables that are created by using Hive. Sentiment calculation is done for every tweet and a polarity score is given to each tweet. If the score is greater than 0 then it is measured to be positive sentiment on behalf of the user, if less than 0 then negative else neutral.

The sentiment calculation process starts with comparing every word of the sentence to every word in the dictionary. Then, whenever a word in text is found that also exist in the dictionary, its polarity is checked. If the polarity is blind negation then the sentence is immediately labelled as negative sentiment tweet. Else if the polarity of the word is positive or negative accordingly its sentiment score is added to the sentiment score of the sentence. Strong and weak positive words have a sentiment score of 2 and 1 respectively. Similarly strong and weak negative words have sentiment score of -2 and -1 respectively. Also, it was concluded from the patterns of the tweets that users use capitalization to emphasize their emotion. So, if positive word has been capitalized, 0.5 is added to the sentiment score. Otherwise, if the word is of negative polarity and has been capitalized, then 0.5 is deducted from the sentiment score of the tweet. After this step, if a negation word is found in the sentence then, whatever the polarity of the sentence is, the polarity is reversed.

In the final step if the sentiment score of the tweet is greater than 0, it is considered to be positive, if it is less than 0 than the sentence is considered as negative otherwise as neutral. 'Positive' polarity is returned when tweet contains positive sentiment about the object under consideration, 'negative' polarity is returned when the user is unhappy about the said object and 'neutral' polarity is returned when either the tweet is an objective sentence or when the user posting the tweet is equally happy and dissatisfied about the said object.

The following table 4.1 illustrates few example tweets and its corresponding computed sentiment score and final contextual polarity computed according the procedure and method explained above:

Tweet	Sentiment Score	Polarity
"movie was awesome , we all loved it"	3(awesome=2)+(loved=1)	positive
"Storyline was good but acting was hell poor"	-1(good=1)+(hell=-1)+(poor=-1)	negative
"Movie is gr8"	1 (gr8=1)	positive
"Movie was fine but I LIKED the hero"	2.5 (fine=1)+(like=1)+(0.5)	positive
"I loved Iron man 3 movie ☺"	2 (loved=1)+(☺=1)	positive

Table 1 Example Tweets And Sentiment Score

4.2.7. Visualization of Computed Sentiments

Data visualization is the exhibition of data in a pictorial or graphical format. Because of the manner the human brain processes information, it is faster for people to understand the meaning of many data points when they are presented in

charts and graphs rather than showing in piles of spreadsheets or if had to read pages and pages of reports.

On similar grounds, Big Data is more valuable and usable when visualized and analyzed. Here, in the proposed system Powerview feature of MS Excel is used to visualize the results. The Data structures from HDFS are imported to the Excel's environment using Hive ODBC tool. The data then can be seen in the form of bar graphs, pie charts, geo maps, etc. Hence, adding simplicity to analyze the results of the sentiment computation.

5. PERFORMANCE OF PROPOSED SYSTEM

In the proposed approach, the focus is on the speed of performing sentiment analysis without compromising much on the accuracy aspect. This is achieved by performing sentiment analysis on big data by filtering and structuring the data and collaborating with Hadoop for mapping it on different machines. The proposed system starting from the tool to fetch and gather data comprises of modules to filter the data and structure it. The proposed system is capable of:

- Getting Twitter data without having to code.
- Store the huge amount of downloaded data in distributed and reliable fashion
- Cleansing the data by removing URLs, tagged names
- Handling misspellings
- Converting data to a structured format
- Giving attention to emphasized words such as capitalized words e.g. 'HAPPY'
- Handling emoticons
- Handling slangs
- Speedy sentiment computation
- Handling Big Data
- Handling negations and blind negations
- Handling cases like 'cooool', 'goooooooooood'

With the proposed approach, we are able to store, evaluate, and analyze such huge data in a speedy and efficient way.

6. CONCLUSION AND FUTURE WORK

With the rapid growth of e-commerce and more and more common users becoming comfortable with the Web, an growing number of people are writing reviews, which is why, Sentiment analysis is becoming more and more prominent and a desirable field to work on.

In the proposed approach, Hadoop, a tool used to handle big data. Developing an approach that can scale to huge data sets was needed because of the large volume of sentiment data that is available today. A lexicon based method for Sentiment Analysis on Twitter data has been presented in this paper which needs no training of data and practical approaches to identify and extract sentiments using emoticons, slangs and acronyms has been used. The use of dictionary lexicon and its storage local to the program makes the approach all together speedier.

The proposed system also promises to solve various issues related to big data, negation handling, handling emoticons, slangs etc.

form.

The proposed system can be further enhanced in future works by considering retweet as a factor for determining the sentiment score and filtering all non-English tweets, which also adds on the accuracy to the system. Also, the proposed system is applicable only on a single node cluster, a multi node level configuration is yet to be designed and presented.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques" In Proceedings of the Empirical Methods on Natural Language Processing, pp. 79-86, Pennsylvania, 2002.
[2] NouraFarra, ElieChallita, RawadAbouAssi, Hazem Hajj, "Sentence-level and Document level Sentiment Mining for Arabic Texts", IEEE International Conference on Data Mining Workshops, 2010.
[3] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", Technical report, HP Laboratories, 2011.
[4] A. Mudinas, D. Zhang, M. Levene, "Combining lexicon and learning based approaches for concept level sentiment analysis", Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, New York, NY, USA, Article 5, pp. 1-8, 2012.
[5] Ji Fang and Bi Chen, "Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classification", In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), pages 94–100, 2011.
[6] A. Khan, B. Baharudin, K. Khan; "Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure" ICSECS 2011: 2nd International Conference on Software Engineering and Computer Systems, Springer, pp. 317-331, 2011.
[7] Moreo A, Romero M, Castro JL, Zurita JM, "Lexicon-based comments-oriented news sentiment analyzer system", Expert SystAppl ;39:9166–80, 2012.
[8] http://hortonworks.com
[9] https://en.wikipedia.org/wiki/Twitter
[10] https://flume.apache.org/
[11] http://www-01.ibm.com/software/data/infosphere/hadoop/
[12] https://developer.yahoo.com/hadoop/tutorial
[13] Introduction To Hadoop by Prof. Jeffery Owens
[14] https://hadoop.apache.org/