Using Apache Spark for Analysing the Sentiments of Unstructured Data with Logistic Regression Algorithm

Chetan Balaji
Student, Department of Information Science and Engineering, R N S Institute of Technology, Bengaluru, India

Abstract-- Sentiment analysis has become an interesting field for both research and industrial domains. The expression sentiment refers to the feelings or thought of the person across some certain issues. Besides, it is additionally viewed as an immediate application for feeling mining. The tremendous measure of unstructured information has been the wellspring of printed information and one of the most fundamental information volumes; subsequently, this information has various points, for example, business, modern or social points as indicated by the information necessity and required preparing.

As a matter of fact, the measure of information, which is huge, develops quickly every second and this is called large information which requires unique preparing methods and high computational force so as to play out the necessary mining errands. Here we propose an idea to perform a sentiment analysis with the help of Apache Spark framework, which is considered an open source distributed data processing platform which utilizes distributed memory abstraction.

The goal of using Apache Spark's Machine learning library (MLIB) is to handle an extraordinary amount of data effectively. We recommend some Pre-processing and Machine learning text feature extraction steps for getting greater results in Sentiment Analysis classification. The effectiveness of our proposed approach is proved against other approaches achieving better classification results when using Naïve Bayes, and Decision trees classification algorithms. Finally, our solution estimates the performance of Apache Spark concerning its scalability

I. INTRODUCTION

Information Mining is an explanatory cycle intended to investigate information (normally a lot of information - commonly business or market related - otherwise called "large information") looking for steady examples or potentially deliberate connections among factors, and afterward to approve the discoveries by applying the distinguished examples to new subsets of information. A definitive objective of information mining is forecast - and prescient information mining is the most widely recognized kind of information mining and one that has the most immediate business applications. Text mining, likewise alluded to as text information mining, generally proportional to message investigation, alludes to the way toward getting excellent data from text. Excellent data is regularly inferred through the contriving of examples and patterns through methods, for example, measurable example learning. Text Mining is to handle unstructured data, remove significant numeric lists from the content, and, consequently, make the data contained in the content available to the different information mining (statistical and machine learning) algorithms. Data can be removed to infer rundowns for the words contained in the records or to process outlines for the archives dependent on the words contained in them. Subsequently, you can break down words, bunches of words utilized in reports, and so on, or you could investigate records and decide similitude between them or how they are identified with different factors of enthusiasm for the information mining venture. In the most broad terms, text mining will "transform text into numbers" (meaningful indices), which would then be able to be consolidated in different examinations, for example, prescient information mining ventures, the utilization of solo learning techniques (clustering).

Parsing only verifies that the program consists of token s arranged in a syntactically valid combination. Now we'll move forward to semantic analysis, where we delve even deeper to check whether they form a sensible set o f instructions in the programming language. A large part of semantic analysis consists of tracking variable/function declarations and type checking. In many languages, identifiers have to be declared before they're used. As t he compiler encounters a new declaration, it records the type information assigned to that identifier. Then, as it continues examining the rest of the program, it verifies that the type of an identifier is respected in terms of the operations being performed [3].In syntactic analysis, parse trees are used to show the structure of the sentence, but they often contain redundant information due to implicit definitions (e.g., an assignment always has an assignment operator in it, so we can imply that), so syntax trees, which are compact representations are used instead. Trees are recursive structures, which complement CFGs nicely, as these are also recursive (unlike regular expressions)[4].

Lexical analysis is the extraction of individual words or lexemes from an input stream of symbols and passing corresponding tokens back to the parser. If we consider a statement in a programming language, we need to be able to recognise the small syntactic units (tokens) and pass this information to the parser

II. LITERATURE SURVEY

In the previous years, studies of Sentiment Analysis and emotional models had a wide attention. The reason for that is basically due to the recent enlargement of data which exists on the social networks, particularly of those that describe people's point of view, thoughts and comments.

Walaa Medhat et al. presented and discussed in brief details different types of sentiment analysis and its applications. Algorithms and their originating references of various SA techniques are categorized and shortly explained. (Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. Ain Shams Engineering Journal 5.4 (2014): 1093-1113**).**Yang et al. introduced the common sentiment analysis methods from the perspective of machine learning technologies, which encompass Naive Bayes technique, Maximum Entropy method, Support Vector Machine technique, and Artificial Neural Network method and performance assessment and difficulties.(Yang, Peng, and Yunfang Chen, IEEE, 2017.) Pang et al. were the first to apply Machine Learning for sentiment mining on movie reviews corpus, many classification algorithms were used, whereas unigram and bag of words are utilized for obtaining features. The ratio of accuracy differs according to what they applied for example it was 82.9% by applying Support Vector Machines, while it was78.7% by applying Naive Bayes classifier. (Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan, Association for Computational Linguistics, 2002).Wang et al. used training dataset which contains 17000 Tweets to come up with a real time Twitter Sentiment Analysis System regarding to U.S. voting Presidential Election Cycle in 2012. (Wang, Hao, Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012**).** The authors introduced a new method combining the help of SentiWordNet alongside with an implementation of Naive Bayes; therefore more accuracy can be achieved. One of the possible techniques to get more accuracy of classification of tweets is applying SentiWordNet and Naive Bayes that give positive, negative and objective degree of the words exist in tweets. (Goel, Ankur, Jyoti Gautam, and Sitesh Kumar, IEEE, 2016). Bindalet al. proposed a twostep system can be applied for sentiment classification of the tweet. During the initial step, sentiment lexicons are used to classify tweets, while the polarity of each tweet is also assigned by aggregating the scores of each token. During the next step, the SVM classifier receives all the tweets with low absolute scores to strengthen the whole accuracy. (Bindal, Nimit, and Niladri Chatterjee,IEEE,2016). Here, authors suggested a real-time solution using spark framework, for processing sentiment analysis Saudi dialect in twitter based on lexicon-based algorithm. (Assiri, Adel, Ahmed Emam, and Hmood Al-Dossari, IEEE, 2016). Here, authors recommended an efficient sentiment prediction technique in Big Data, using Spark. The outcomes got from the suggested work were subject to analysis to demonstrate high levels of scalability in relation to accuracy and time. It was noted that even with the growth of data volume, the processing time indicated very less variance. (Nirmal, V. Jude, and DI George Amalarethinam,IEEE,2017). Here, authors recommended an efficient sentiment prediction technique in Big Data, using Spark. The outcomes got from the suggested work were subject to analysis to demonstrate high levels of scalability in relation to accuracy and time. It was noted that even with the growth of data volume, the processing time indicated very less variance. (Nirmal, V. Jude, and DI George Amalarethinam,IEEE,2017)

III. PROBLEM IDENTIFICATION

Presently informal communities locales make the whole world a little town, where clients can share their perspectives, emotions, encounters, guidance through those destinations with the goal that others can find support from these. Since a large number of us utilize web-based media every day, a gigantic amount of remarks, feeling, article have been made. finding a programmed way for examining and characterizing clients' feelings in informal organizations could be very basic.

This is mainly because it is considered a great tool for getting direct notes or information from users. The method of classifying texts or documents in keeping with their polarity is referred to as Sentiment Analysis (SA). Sentiment analysis can be described as a major branch of Natural Language Processing (NLP), its aim to identify the meaning from a document in order to discover the polarity of the text. For the sentiment analysis, we focus our attention in the direction of the Twitter, a microblogging social networking website, where users can communicate with each other or share their opinions in short blogs

Huge diverse number of text posts exists on twitter which builds each day, the fast colossal information development make the current data sets unfit to deal with a broad measure of information in a brief timeframe. Additionally, these information bases type intended to handle organized information however there is an impediment on it when managing enormous information. So the traditional arrangements are not useful for associations to oversee and handle unstructured or huge information.

Frameworks, such as Hadoop, Apache Spark, Apache flume and distributed data storages like Hadoop Distributed File System (HDFS), Cassandra and HBase are being very widespread, as they are designed in a manner which facilitates the process of huge amounts of big data and makes it almost effortless. One of the most effective manners is the parallel computing techniques when dealing with big data, which include multicore processors, distributed computing, and etc.

Partitioning issues into many sub-issues, to be handled utilizing strings and machines in the group is the principle highlight of Parallel registering techniques. One of the most significant focal points of equal processing is upgrading the presentation of the calculations to be quicker than the sequential habits.

IV. PROPOSED WORK

Here we propose an idea of performing a sentiment analysis with the help of Apache Spark framework, which is considered an open source distributed data processing platform which utilizes distributed memory abstraction. The goal of using Apache Spark’s Machine learning library (MLIB) is to handle an extraordinary amount of data effectively. We recommend some Pre-processing and Machine learning text feature extraction steps for getting greater results in Sentiment Analysis classification..

There are two significant issues experiences in web-based media information handling.

- Initial one is equivocalness of information which makes information sudden.
- Second one is information in online media isn't composed in an organized way, it is fundamentally unstructured.

An expression whose meaning cannot be determined from its contexts is ambiguity. The interpretation of data gets difficult when the data is obscure in nature and data in most of the social media is indistinct. Due to this vagueness, ambiguity arises. So the interpretation of data is not up to the mark due to this ambiguity.

In order to resolve the identified issues in the social networking data analysis a new model is proposed in this system. The proposed data model is based on the hybrid concept of graph theory and data mining techniques.

The entire data model development is performed in two major modules first training and then testing. In training first the system accept the training data samples on which the pre-processing is performed during pre-processing the stop words are removed from the input text and the significant text is remain from the training set. The word probability is majored in this measured from the text remain using the below given formula.

$$word\ probebility = \sum_{i=1}^{N} \frac{word_i}{N}$$

Where the N is number of total words in text documents, in the similar ways the sentence formation probability is estimated from the text using the following formula.

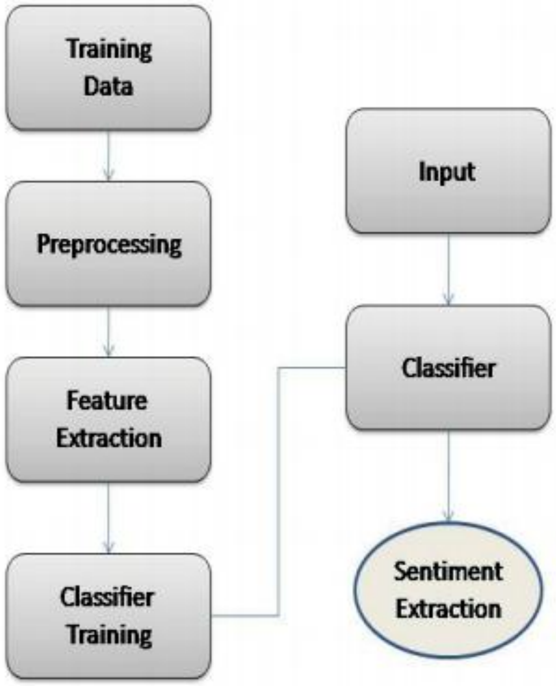
$$senatance\ probebility = \sum_{i=1}^{M} \frac{word_i}{M}$$

Where the M is the numbers of sentences in the text document after that the weights for the weighted graph is computed from both the probability.

Weight=sentence probability*word probability.

Using the computed weights and the user feedback the correlation graph is developed in further steps and using these weighted graphs the classification rules are developed. These rules sets are consumed to identify the sentiments of the words involved in the micro-blog text.

The effectiveness of our proposed approach is proved against other approaches achieving better classification results when using Naïve Bayes, and Decision trees classification algorithms. Finally, our solution estimates the performance of Apache Spark concerning its scalability



V. REQUIREMENTS SPECIFICATION

A. Hardware Requirements

Processor	Intel Core i5 or AMD FX 8 core series with clock speed of 2.4 GHz or above
RAM	6GB or above
Hard disk	40 GB or above
Input device	Keyboard or mouse or compatible pointing devices
Display	XGA (1024*768 pixels) or higher resolution monitor with 32 bit color settings
Miscellaneous	USB Interface, Power adapter, etc

B. Software Requirements

Operating System	Windows and Raspberry PI
Programming Language – Backend	Python and Apache spark
Programming language - Frontend	Bootstrap Framework, HTML, CSS, JavaScript, Ajax, JQuery
Development environment	Eclipse Oxygen IDE, PyDev Eclipse plugin
Application Server	Apache Tomcat v9.0
Database	MySQL

CONCLUSION

In this paper, we perform a sentiment analysis with the help of Apache Spark framework, which is considered an open source distributed data processing platform which utilizes distributed memory abstraction.

The goal of using Apache Spark's Machine learning library (MLIB) is to handle an extraordinary amount of data effectively. We recommend some Pre-processing and Machine learning text feature extraction steps for getting greater results in Sentiment Analysis classification.

The effectiveness of our proposed approach is proved against other approaches achieving better classification results when using Naïve Bayes, and Decision trees classification algorithms.

Finally, our solution estimates the performance of Apache Spark concerning its scalability

References

[1] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams Engineering Journal 5.4 (2014): 1093-1113.

[2] Yang, Peng, and Yunfang Chen. "A survey on sentiment analysis by using machine learning methods." Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2017 IEEE 2nd Information. IEEE, 2017.

[3] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002

[4] Wang, Hao, et al. "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle." Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012

[5] Goel, Ankur, Jyoti Gautam, and Sitesh Kumar. "Real time sentiment analysis of tweets using Naive Bayes." Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on. IEEE, 2016.

[6] Bindal, Nimit, and Niladri Chatterjee. "A Two-Step Method for Sentiment Analysis of Tweets." 2016 International Conference on Information Technology (ICIT). IEEE, 2016

[7] Nodarakis, Nikolaos, et al. "Large Scale Sentiment Analysis on Twitter with Spark." EDBT/ICDT Workshops. 2016.

[8] Assiri, Adel, Ahmed Emam, and Hmood Al-Dossari. "Real-time sentiment analysis of Saudi dialect tweets using SPARK." Big Data (Big Data), 2016 IEEE International Conference on. IEEE, 2016

[9] Nirmal, V. Jude, and DI George Amalarethinam. "Real-Time Sentiment Prediction on Streaming Social Network Data Using InMemory Processing." Computing and Communication Technologies (WCCCT), 2017 World Congress on. IEEE, 2017.

[10] Chikersal, Prerna, Soujanya Poria, and Erik Cambria. "SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning." Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). 2015.

[11] Parveen, Huma, and Shikha Pandey. "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm." Applied and Theoretical Computing and Communication Technology (iCATccT), 2016 2nd International Conference on. IEEE, 2016

[12] Yan, Bo, et al. "Microblog Sentiment Classification Using Parallel SVM in Apache Spark." Big Data (BigData Congress), 2017 IEEE International Congress on. IEEE, 2017.