Big Data Sentiment Analysis using Hadoop

Ramesh R
Assistant Professor
Department of Information Technology & Computer Science Engineering
ASIET, Kalady

Divya G
Assistant Professor
Department of Information Technology & Computer Science Engineering
ASIET, Kalady

Divya D
Assistant Professor
Department of Information Technology & Computer Science Engineering
ASIET, Kalady

Merin K Kurian
B. Tech Student
Department of Computer Science & Engineering
ASIET, Kalady

Vishnuprabha V
B. Tech Student
Department of Computer Science & Engineering
ASIET, Kalady

Abstract

Social media gives users a platform to communicate effectively with friends, family, and colleagues, and also gives them a platform to talk about their favourite (and least favourite brands). This "unstructured" conversation can give businesses valuable insight into how consumers perceive their brand, and allow them to actively make business decisions to maintain their image. Rapid increase in the volume of sentiment rich social media on the web has resulted in an increased interest among researchers regarding Sentimental Analysis and Opinion Mining. However, with so much social media available on the web, Sentiment Analysis is now considered as a Big Data task. The main focus of the research was to find such a technique that can efficiently perform Sentiment Analysis on Big Data sets. In this paper Sentiment Analysis was performed on a large data set of tweets using Hadoop and the performance of the technique was measured in form of speed and accuracy. The experimental result shows that the technique exhibits very good efficiency in handling big sentiment data sets.

Keywords: Big Data; Hadoop; Lexicon; Machine learning; Negation; NLP; Sentiment Analysis

I. INTRODUCTION

Big Data is trending research area in Computer Science and Sentiment Analysis is one of the most important part of this research area. Big Data is considered as very large amount of data which can be found easily on web, Social media, remote sensing data and medical records etc. in form of structured, semi-structured or unstructured data and we can use these data for Sentiment Analysis.

Sentimental Analysis is all about to get the real voice of people towards specific product, services, organization, movies, news, events, issues and their attributes[1]. Sentiment Analysis includes branches of Computer Science like Natural Language Processing, Machine Learning, Text Mining and Information Theory and Coding. By using approaches, methods, techniques and models of defined branches, we can categorize our unstructured data which may be in the form of news articles, blogs, tweets, movie reviews, product reviews etc. into positive, negative or neutral sentiment according to the sentiment expressed in them.
Sentiment Analysis is done on three levels [1].

- Document level
- Sentence level
- Aspect or Entity level

1) Document Level Sentiment Analysis is performed for the whole document and then decide whether the document express positive or negative sentiment [1].
2) Entity or Aspect Level Sentiment Analysis performs fine-grained analysis. The goal of entity or aspect level Sentiment Analysis is to find sentiment on entities and/or aspect of those entities. For example consider a statement "My HTC Wildfire S phone has good picture quality but it has low phone memory storage." so sentiment on HTC's camera and display quality is positive but the sentiment on its phone memory storage is negative.
3) Sentence level Sentiment Analysis is related to find sentiment form sentences whether each sentence expressed a positive, negative or neutral sentiment Sentence level Sentiment Analysis is closely related to subjectivity classification. Many of the statements about entities are factual in nature and yet they still carry sentiment. Current

Sentiment Analysis approaches express the sentiment of subjective statements and neglect such objective statements that carry sentiment [1].

II. EXISTING METHODS

Lexicon Based techniques work on an assumption that the collective polarity of a document or sentence is the sum of polarities of the individual words or phrases. Some of the significant worksdone using this technique are:

Kamps [18] used a simple technique based on lexical relations to perform classification of text.

Andrea [19] used word net to classify the text using an assumption that words with similar polarity have similar orientation.

Ting-Chun [20] used an algorithm based on pos (part of speech) patter. A text phrase was used as a query for a search engine and the results were used to classify the text.

Prabhu [21] which used a simple lexicon based technique to extract sentiments from twitter data.

Turney [22] used semantic orientation on user reviews to identify the underlying sentiments.

Taboada [23] used lexicon based approach to extract sentiments from micro blogs.

Sentiment analysis for micro blogs is more challenging because of problems like use of short length status message, informal words, word shortening, spelling variation and emoticons. Twitter data was used for sentiment analysis by [24].

Negation word can reverse the polarity of any sentence. Taboada [23] performed sentiment analysis while handling negation and intensifying words. Role of negation was surveyed by [25]. Minquing [26] classified the text using a simple lexicon based approach with feature detection. It was observed that most of these existing techniques doesn't scale to big data sets efficiently. While various machine learning methodologies exhibits better accuracy than lexicon based techniques, they take more time in training the algorithm and hence are not suitable for big data sets. In this paper, lexicon based approach is used to classify the text according to polarity.

III. PROPOSED METHOD

The focus of this research was to device an approach that can perform Sentiment Analysis quicker because vast amount of data needed to be analyzed. Also, it had to be made sure that accuracy is not compromised too much while focusing on speed. Sentiment Analysis on Big Data is achieved by collaborating Big Data with hadoop.

Table -1:
Sample Data Dictionary and Its Polarity

Strength	*Word*	*polarity*
weaksubj	*abandoned*	*negative*
weaksubj	*abandonment*	*negative*
weaksubj	*abandon*	*negative*
strongsubj	*needed*	*Blind negation*

The proposed approach is a dictionary based technique i.e. a dictionary of sentiment bearing words was used to classify the text into positive, negative or neutral opinion. Machine learning techniques [12] are not used because although they are more accurate than the dictionary based approaches, they take far too much time performing Sentiment Analysis as they have to be trained first and hence are not efficient in handling big sentiment data.

A. Real Time Data and Features:

1) *Length;*

The maximum length of a tweet is about 140 characters. This is very different from the previous sentiment classification research that focused on classifying longer bodies of work, such as movie reviews.

2) *Data Availability:*

Another difference is the magnitude of data available. With the Twitter API, twitter4j [13], it is very easy to collect millions of tweets for training which allows the developer an access to 1% of tweets tweeted at that time basis on the particular keyword..

3) *Language Model:*

Twitter users post messages from many different media, including their cell phones. The frequency of misspellings and slang in tweets is much higher than in other domains.

4) *Domain:*

Twitter users post short messages about a variety of topics unlike other sites which are tailored to a specific topic. This differs from a large percentage of past research, which focused on specific domains such as movie reviews.

B. Sentiment Dictionary:

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. Hence eliminating the need for stemming each word which saves more time. The Dictionary also contains the strength of the polarity of every word. Some word depicts stronger emotions than others. For example good and great are both positive words but great depict a much stronger emotion.

Negation and blind negation are very important in identifying the sentiments, as their presence can reverse the polarity of the sentence. The dictionary used here also contains various negation and blind negation words so that they can be identified in the sentence.

C. Handling Negation and Blind Negation:

Negation words are the words which reverse 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, its polarity is reversed [12, 15, and 16].

Blind negation words are the words which operates 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 its polarity is immediately labelled as negative.

D. Sentiment Calculation Algorithm:

Sentiment calculation is done for every tweet and a polarity score is given to it. If the score is greater than 0 then it is considered to a positive sentiment on behalf of the user, if less than 0 then negative else neutral. The polarity score is calculated by using algorithm 1 by using mapreduce programming model.

1) *Algorithm 1: ALGO_SENTICAL*
   - Input: Tweets, SentiWord_Dictionary
   - Output: Sentiment (positive, negative or neutral)

BEGIN
1) For each tweet T_i *do* the following
2) Initialize SentiScore = 0;
3) For each word W_j in T_i that exists in Sentiword_Dictionary.
   - If polarity[W_j] = blind negation then Return negative.
   - Else
   - b.1. If polarity[W_j] = positive && strength[W_j] = Strongsubj then increment seniscore by 1.
   - b.2 Else If polarity[W_j] = positive && strength[W_j] = Weaksubj then add 0.5 to sentiscore.
   - b.3. Else If polarity[W_j] = negative && strength[W_j] = Strongsubj then decrement sentiscore by 1.
   - b.4 .Else If polarity[W_j] = negative && strength[W_j] = Weaksubj then substract 0.5 from sentiscore.
   - b.5. If polarity[W_j] = negation multiply sentiscore by -1.
   - If Sentiscore of T_i >0 then Sentiment = positive.
   - Else If Sentiscore of T_i<0 then Sentiment = negative.
   - Else Sentiment = neutral
4) Return Sentiment
5) END

IV. PERFORMANCE EVALUATION

A. Experimental Setup:

The proposed algorithm was implemented using a 1.x version of Apache Hadoop. Hadoop is designed to work in a multimode environment but for research purposes often a single node virtual environment is used that creates an illusion of several nodes which are situated at different locations and are working together. An Intel Core i5-3210M CPU@2.50GHz processor with 6 GB memory was used to simulate the Hadoop Environment. Data was imported from Twitter using Flume, a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS).

B. Results and Evaluation:

As explained earlier the purpose of this research was to device a method that can quickly compute the sentiments of huge data sets without compromising too much with accuracy. The proposed approach has performed very well in terms of speed.
We will evaluate our experiment results by using following Information Retrieval matrices [20].

- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)
- F-measure = 2*Precision*recall/(Precision + recall)
- Accuracy = TP + TN /(TP + TN + FP + FN)
- The proposed method has an accuracy of 75% when worked with 50 comments of hp laptops.

Table -2:
Number of True Positive, True Negative, False Positive and False Negative from 50 Comments

True Positive	*True Negative*	*False Positive*	*False Negative*
22	*8*	*11*	*0*

Table -3:
Experimental Results

Precision	*Recall*	*F-measure*	*Accuracy*
66.666%	*100%*	*79.95%*	*75%*

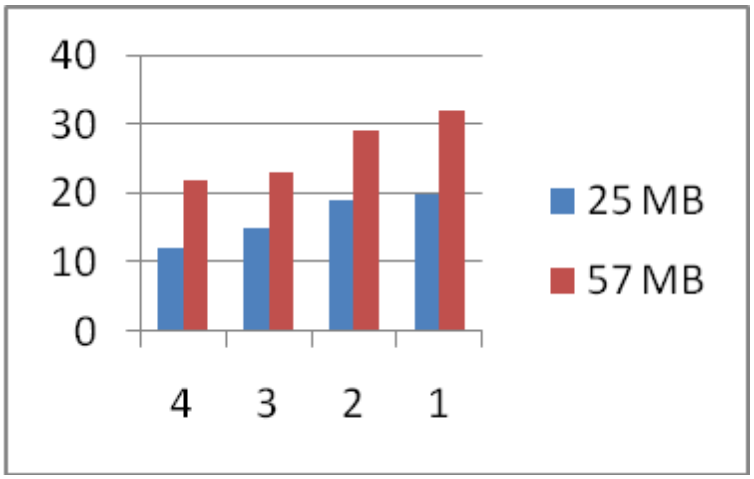

Fig. 1: Execution time comparison over single node and multiple nodes(Upto 4 nodes)
X-axis: Nodes
Y-axis: Time

V. CONCLUSION AND FUTURE WORK

Sentiment Analysis is being used for different applications and can be used for several others in future. It is evident that its applications will definitely expand to more areas and will continue to encourage more and more researches in the field. We have done an overview of some state-of the-art solutions applied to sentiment classification and provided a new approach that scales to Big Data sets efficiently. A scalable and practical lexicon based approach for extracting sentiments using emoticons and hash tags is introduced. Hadoop was used to classify Twitter data without need for any kind of training. Our approach performed extremely well in terms of both speed and accuracy while showing signs that it can be further scaled to much bigger data sets with similar, in fact better performance.

In this research, main focus was on performing Sentiment Analysis quickly so that Big Data sets can be handled efficiently. The work can be further expanded by introducing techniques that increase the accuracy by tackling problems like thwarted expressions and implicit sentiments which still needs to be resolved properly. Also as explained earlier, this work was implemented on a single node configuration and although it is expected that it will perform much better in a multimode enterprise level configuration, it is desirable to check its performance in such environment in future.

REFERENCES

[1] Bing Liu, Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers, May 2012.p.18-19,27-28,44-45,47,90-101.
[2] Nitin Indurkhya, Fred J. Damerau , Handbook of Natural Language Processing, Second Edition, CRC Press, 2010.
[3] Ronen Feldman, James Sanger,The Text Mining Handbook-Advance Approaches in Analyzing Unstructured Data, Cambridge University Press,2007.
[4] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publications, 2006.
[5] Chihil Hung and Hao-kai Lin, "Using Objective Word in SentiWordNet to Improve Word-of-Mouth Sentiment Classification", IEEE Computer Society, P.47- 54, March-April 2013.
[6] Wikipedia article on supervised machine learning http://en.m.wikipedia.org/wiki/Supevised_learning
[7] Jintao Mao and Jian Zhu, "Sentiment Classification based on Random Process", IEEE Computer Society, International Conference on Computer Science and Electronics Engineering, p.473-476, 2012.
[8] Sang-Hyun Cho and Hang-Bong Kang, "Text Sentiment Classification for SNS-based Marketing Using Domain Sentiment Dictionary", IEEE International Conference on Conference on consumer Electronics (ICCE), p.717-718, 2012.
[9] Gautam Shroff, Lipika Dey and Puneet Agrawal, "Social Business Intelligence Using Big Data",CSI Communications, April 2013,p.11-16.
[10] Sentiment Analysis: Capturing Favorability Using Natural Language Processing Tetsuya Nasukawa IBM Research, Tokyo Research Laboratory Jeonghee Yi IBM Research, Almaden Research Center.
[11] ZHU Jian , XU Chen, WANG Han-shi, "" Sentiment classification using the theory of ANNs", The Journal of China Universities of Posts and Telecommunications, July 2010, 17(Suppl.): 58–62 .[16] Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li, "Sentiment classification of Internet restaurant reviews written in Cantonese", Expert Systems with Applications xxx (2011)
[12] Long-Sheng Chen, Cheng-Hsiang Liu, Hui -Ju Chiu, "A neural network based approach for sentiment classification in the blogosphere", Journal of Informetrics 5 (2011) 313–322.
[13] http://twitter4j.org/en/index.html
[14] Building Machine Learning Algorithms on Hadoop for Bigdata Asha T, Shravanthi U.M, Nagashree N, Monika M International Journal of Engineering and Technology Volume 3 No. 2, February, 2013
[15] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll and Manfred Stede. 2011. Lexicon based methods for Sentiment Analysis. Computational linguistics, volume 37, number2, 267–307, MIT Press.
[16] Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, Andr´es Montoyo. 2010. Asurvey on the role of negation in Sentiment Analysis. Proceedings of the workshop on negation speculation in natural language processing 60–68, Association for Computational Linguistics.
[17] Twitter Sentiment Classification using Distant Supervision Alec Go Stanford University Stanford, CA 94305 alecmgo@stanford.edu Richa Bhayani Stanford University Stanford, CA 94305 rbhayani@stanford.edu Lei Huang Stanford University.
[18] Kamps, Maarten Marx, Robert J. Mokken and Maarten De Rijke, "Using wordnet to measure semantic orientation of adjectives", Proceedings of 4th International Conference on Language Resources and Evaluation, pp. 1115-1118, Lisbon, Portugal, 2004.
[19] Andrea Esuli and Fabrizio Sebastiani, "Determining the semantic orientation of terms through gloss classification", Proceedings of 14th ACM International Conference on Information and Knowledge Management,pp. 617-624, Bremen, Germany, 2005.
[20] Ting-Chun Peng and Chia-Chun Shih , "An Unsupervised Snippet-based Sentiment Classification Method for Chinese Unknown Phrases without using Reference Word Pairs", 2010 IEEE/WIC/ACM International Conference on Web Intelligence and intelligent Agent Technology JOURNAL OF COMPUTING, VOLUME 2, ISSUE 8, AUGUST 2010, ISSN 2151-9617 .
[21] Prabu Palanisamy, Vineet Yadav, Harsha Elchuri, "Serendio: Simple and Practical lexicon based approach to Sentiment Analysis", Serendio Software Pvt Ltd, 2013.
[22] Peter Turney and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21(4):315–346.
[23] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. Computational linguistics, volume 37, number2, 267–307, MIT Press.
[24] Albert Bifet and Eibe Frank. 2010. Sentiment knowledge discovery in twitter streaming data, Discovery Science 1–14, Springer.
[25] Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, Andr´es Montoyo. 2010. A survey on the role of negation in sentiment analysis. Proceedings of the workshop on negation and speculation in natural language processing 60–68, Association for Computational Linguistics.
[26] Minqing Hu, Bing Liu. Mining and Summarizing Customer Reviews, Department of Computer Science, University of Illinois at Chicago, Research Track Paper.