ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

Sentiment Analysis using Maximum Entropy Algorithm in Big Data

Durgesh Patel [1], Sakshi Saxena [2], Toran Verma[3]

P.G. Student, Department of Computer Engineering, Rungta College of Engineering & Technology, Bhilai, India[1]

Assistant Professor, Department of Computer Engineering, Rungta College of Engineering & Technology, Bhilai, India[2]

M. Tech Coordinator, Department of Computer Engineering, Rungta College of Engineering & Technology, Bhilai, India[3]

ABSTRACT: As years are passing people are getting more and more attracted towards Social Media Sites and they had started attaching there life with these electronic gadgets. As this things are happening the data being created and shared by millions of people that is increasing per second and these data are not easy to store and manage. As we know that everything in this world has some meaning full existence in this world in that similar way this Big data can also be used for some application and for that the analysis done is called Sentiment Analysis. But Analyzing this large dataset is not so easy , and same way Sentiment Analysis is lacking for Negation Statement case. This problem is reducing the accuracy of Sentiment Analysis which can be verified easily with the proposed methodology in this paper.

KEYWORDS: Negation, Sentiment Analysis, Big data

I. INTRODUCTION

B ig Data is the most emerging word during this period of time and it is the most major topic to be concerned for. Social Media has taken over a very large space in human life style that is why Big data is also became very big problem to handle and use it simply[4].
And this data from social media is very important and can be used for various analysis like Sentiment analysis. With the innovation of digital Technologies and smart devices , a huge amount of digital data is being generated everyday. This data is very hard to manage using conventional techniques like Relational Database Management System[7] , these system can not able to store this huge data and impossible to analyze this data fastly and accurately.
Due to the failure of conventional technique of storage of data for big Data a new system is invented termed as Hadoop , Hadoop is a open source framework which has ability to store the data and also to analyze the data accurately , precisely and fastly also.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is mainly designed to maintain thousands of machine through a single server. Hadoop has various tools which supports different languages for its processing , example, HDFS , Hive , Map-Reduce , Pig , Oozie , flume etc [13]. This paper will mainly need to get in touch with Flume, Hive , HDFS, Pig. Flume supports for the collection of dataset from the developers site of Twitter (i.e http://dev.twitter.com) and hive mainly uses HSQL language which is for the data handling as in tables, views etc. HDFS is mainly for the storage purpose. And in this project Pig is used for making a user defined function for different procedure.
As various studies has been conducted in this area but the accuracy of sentiment analysis is not up to the mark till now. And in case of Big Data the accuracy is not good because to handle large amount of data is not an easy task. Our main purpose is not only to store large amount of data but also to analyze it for the benefit of organizations and also for the peoples welfare.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

This Paper contains 5 section which are discussed in detail -
Section 1 -It contains the Introduction of the concept which is used for the project.
Section 2 - Preliminary Definition of some required Terms
is discussed here.
Section 3 -Experimental Setup & Results , these section will give the basic requirement to perform this task and what are the preliminary steps that is to be followed need to be discussed here.
Section 4 -Proposed Model This section mainly specify the basic need of the process to be followed for the overall sentiment analysis with improved accuracy..
Section 5 - Conclusion and Future Work, this section mainly discuss the result what we have got from the analysis and how much the algorithm is suitable for the dataset is need to be discussed here. And what can be done further to explore more this topic is also suggested.

II. PRELIMINARY DEFINITION

The raw Data that need to be processed for opinion mining or sentiment analysis can be of Two type -

1. Structure Data - Structured data is information, usually text files, displayed in titled columns and rows which can easily be ordered and processed by data mining tools. This could be visualized as a perfectly organized filing cabinet where everything is identified, labeled and easy to access.

2. Unstructured Data -Unstructured data is that which has no identifiable internal structure. It can be visualized as a room with unorganized objects .Some examples of unstructured data are E-mails, PDF files , Digital Images , Video, Audio etc.

For This type of Unstructured Data [5] the Traditional Database Management System will not work , RDBMS can not able to handle large and unstructured Data that is why it need to have other Big Data handling system which store and process it in faster speed.

2.1. Lexical Analysis

Sentiment Analysis can be widely divided into various methods like Lexical analysis
Lexical analysis mainly use the analysis from the reference of any well defined corpus. This corpus contains English words with its meaning and the parts of speech given for individual words.
In the case of sentiment analysis, a simple lexicon based approach would be to classify the sentiments of tweets based on number of 'positive' and 'negative' terms contained in the tweets and choose the label with the most contained terms.[12]. The flowchart for this is shown in figure 2.1.

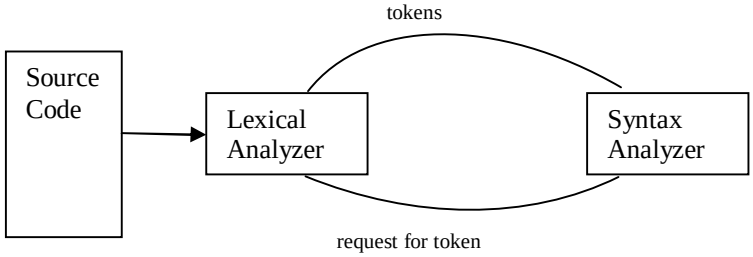

Figure 2.1. Block diagram showing process involved in Lexical Analysis

But lexical Analysis also fails in some cases that is why we can not take only lexical analysis for our study , for improved accuracy and precision it should have followed some other method or path to overcome the disadvantage of the lexical Analysis.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

2.2. Machine Learning Algorithm

Machine learning algorithm mainly divided into two parts i.e. Supervised Approach and Unsupervised Approach.
There are various Machine Learning algorithm till date which can be used and compared which can give the better accuracy.
But as nothing is perfect Machine Learning algorithm also not give the perfect accuracy and precision for example this machine learning algorithm [12] needs training and test data set in large amount to get perfect accuracy.
That is why by studying both the perspective for this paper the combination of Lexical Analysis and Machine Learning Algorithm both have been considered to overcome the disadvantage of negation problem from lexical analysis Machine learning Algorithm suits the best and for providing the large amount of training dataset Lexical Analysis provide it.

2.3. Natural Language Processing

Natural language processing (NLP) is the ability of a computer program to understand human speech as it is spoken. NLP [9] is a component of artificial intelligence (AI).
Today NLP is a necessary tool since human has to communicate to the computers and computer understand only machine language for that Naturally converting processing is required. Current approaches of Natural Language Processing is through the use of Machine Learning Algorithm .
NLP can mainly performs various functions (shown in Figure 2.2) like :

- Tokenization
- Part of Speech Tagging
- Parsing
- Named Entity Recognition
- Chunking

There are various NLP toolkits available like :

- Stanford NLP
- Natural Language Toolkit(NLTK)
- Apache OpenNLP
- Apache Lucene & Solr

For this paper the Apache OpenNLP have been considered better which is open source with the use of Maximum Entropy Algorithm as the Machine Learning Algorithm.

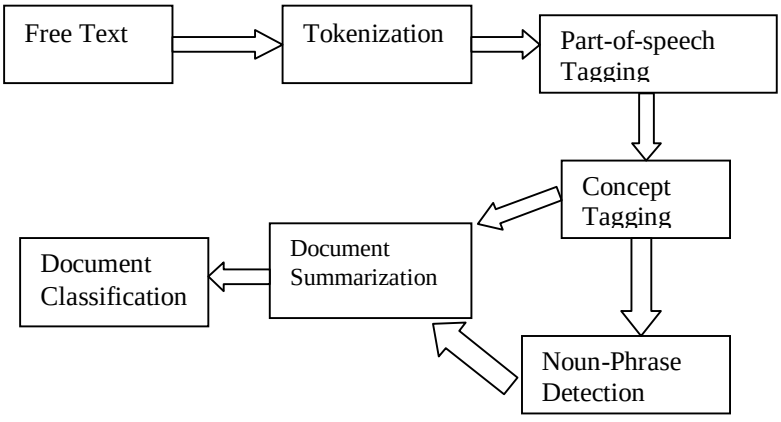

Figure 2.2. Block diagram showing process of NLP

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

2.4. Maximum Entropy Algorithm

Maximum Entropy Algorithm is a Machine Learning Algorithm .The Max Entropy classifier is a probabilistic classifier which belongs to the class of exponential models. Our target is to use the contextual information of the document (unigrams, bigrams, other characteristics within the text) in order to categorize it to a given class (positive/neutral/negative, objective/subjective etc).
Maximum Entropy main Principle is Higher the entropy higher is the uniformity. The maximum entropy principle is based on selecting the most uniform distribution which is to be known by the one having maximum entropy.
To define Log Likelihood of the model p with respect to the empirical distribution p , which is derived below in the formulae :-

$$L_{\tilde{p}}(p) = \log \prod_{x \in \mathcal{X}} p(x)^{\tilde{p}(x)} = \sum_{x \in \mathcal{X}} \tilde{p}(x) \log p(x).$$

And in this case as it is using the maximum Entropy Algorithm to the Training Dataset obtained by the Lexical Analysis of the Data using Hadoop and as the output it will give the Log-likelihood in 100 iterations which will show us the accuracy of the analysis which have been done using Lexical Analysis after removing the Negation Problem from it.

III. PROPOSED MODEL

For improving the accuracy of Sentiment analysis the system in which these model has to be implemented should have RAM of capacity minimum of 8GB and for Big Data we require a framework called Hadoop in the System . With the use of Flume we can collect the dataset (http://dev.twitter.com) which is to be processed.
As the method followed is Lexical Analysis first it need some corpus of English words which will Analyze the sentence and check with the corpus to be created.
After the collection of dataset from twitter it can further proceed to get the analysis done and the output will be in the form of Positive , Negative and Neutral cases .In this paper three cases have been assumed that is of Positive , Negative and Neutral only. In this Model all the steps of data pre processing is done in Hive tool and after that it must create a User Defined Function which is written in Java using Pig tool of hadoop. As shown in Figure 4.1, it represents the overall methods to be followed during this process of Sentiment Analysis. In above procedure for solving the negation removal problem it is easy to write a User Defined Function in Java Language which combines the negative adverb word with its neighboring adjective positive word and make its polarity to -1 i.e. negative.
And as discussed earlier about Machine Learning Algorithm and out of various machine learning Techniques Maximum Entropy Algorithm is taken up, and applied with the base of OpenNLP toolkit. As Natural Language Processing tool kit is the tool which is capable of Understanding the Naturally using Human language to machine till some extent.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

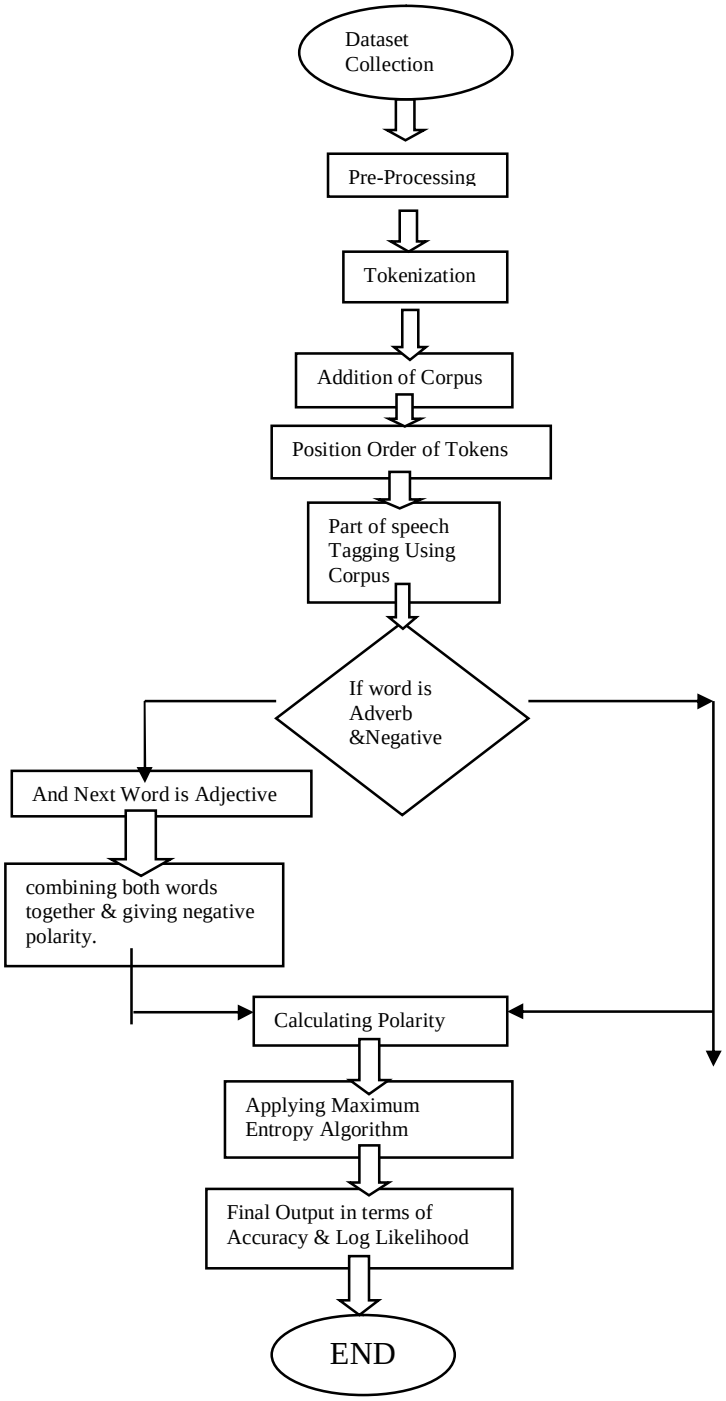

Figure 4.1. Proposed Algorithm for Sentiment Analysis

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

IV. RESULTS & DISCUSSION

Through this Paper, an efficient algorithm for Sentiment Analysis has been checked and proved to be efficient. This study proves that Sentiment Analysis gets better accuracy and Precision by using the Lexical and Machine Learning Algorithm together. Figure 5,1 shown below give the output of the Maximum Entropy Algorithm using OpenNLP tool.

This Maximum Entropy Algorithm Log-likelihood data predicts the graphical output on the basis of its iterations which is shown in below figure 5.2 :-

```
...done.
Computing model parameters...
Performing 100 iterations.
  1:  .. loglikelihood=-2169608.899571239       0.0
  2:  .. loglikelihood=-2040859.7804328476      0.7494770108406036
  3:  .. loglikelihood=-1940176.6629931435      0.7561765115568784
  4:  .. loglikelihood=-1859007.6408229321      0.7617041587382319
  5:  .. loglikelihood=-1791755.667663587       0.7662062743364075
  6:  .. loglikelihood=-1734750.9520797003      0.7697997881207022
  7:  .. loglikelihood=-1685528.3119008588      0.7728501854902297
  8:  .. loglikelihood=-1642380.4030299599      0.7754188066518343
  9:  .. loglikelihood=-1604088.1371733875      0.7774660360552624
 10:  .. loglikelihood=-1569756.204198629       0.7790966632205397
 11:  .. loglikelihood=-1538710.337909146       0.7805062100569825
 12:  .. loglikelihood=-1510431.375628575       0.7817988271241283
 13:  .. loglikelihood=-1484511.7540273038      0.7827936886038841
 14:  .. loglikelihood=-1460626.042972584       0.7836703423933671
 15:  .. loglikelihood=-1438510.4936582488      0.7844818222130779
 16:  .. loglikelihood=-1417948.5217154683      0.7852549644035112
 17:  .. loglikelihood=-1398760.1910357862      0.7858926469704967
 18:  .. loglikelihood=-1380794.4540989576      0.7865443866682172
 19:  .. loglikelihood=-1363923.3302048233      0.7870613057029779
 20:  .. loglikelihood=-1348037.4714930023      0.7874759910596648
 21:  .. loglikelihood=-1333042.7396964652      0.7878683127992732
 22:  .. loglikelihood=-1318857.5304275514      0.7882452994871703
 23:  .. loglikelihood=-1305410.6581469618      0.7886529562784896
 24:  .. loglikelihood=-1292639.6670965056      0.789060613069809
```

Figure 5. 1 Output of Maximum Entropy Algorithm

V. CONCLUSION

This Proposed Method is suitable for English Language only. In future it can be tested for other Language data Set also and it can be more accurate and precise by understanding each line and sentence in any language like as a human being understanding. And also solve different problems like Increment - decrement statements and also solve the emoji cases to be analyzed. One further more work is Possible to increase the speed of the working of whole analysis. In this paper data is taken from Twitter only but we can take data from any of the social media sites. This project is mainly handling the big Data which need to be handled better with connection of parallel distributed systems and connected to a single server this can be done in future work.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

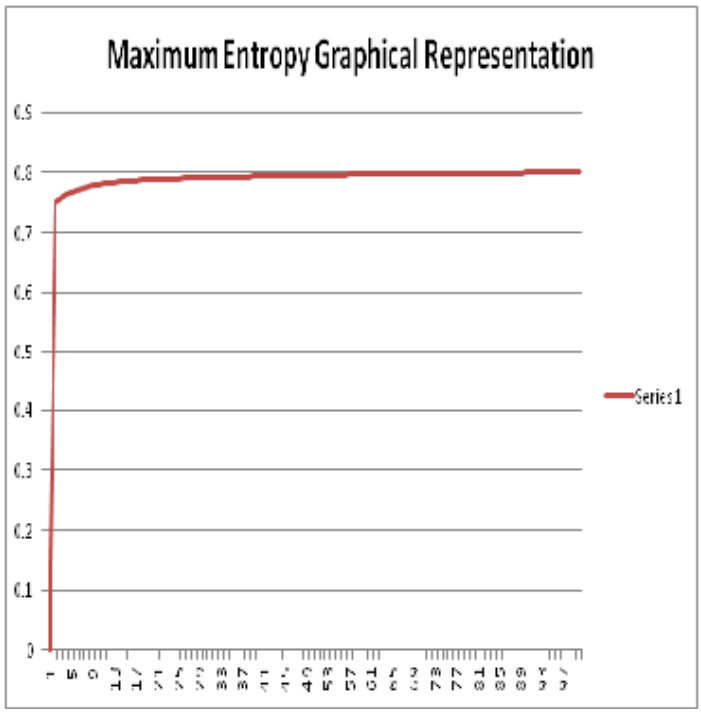

Figure 5.2 Graphical representation of accuracy of the Algorithm

REFERENCES

[1] Ramesh R, Divya G , Divya D , Merin K Kurian , Vishnuprabha V, " Big Data Sentiment Analysis using Hadoop" in Volume I April 2015, IJIRST, ISSN(online):2349-6010.
[2] Andrea Esuli and Fabrizio Sebastiani ,"SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining
[3] Apoorv Agrawal , Jasneet Singh Sabarwal , "End to End Sentiment Analysis of Twitter data " Proceeding of the workshop on Information Extraction & Entity Analysis on Social media data, COLING 2012, Mumbai , December 2012 .
[4] Efthymois Kouloumpis , Theresa Wilson , Johanan Moore , "Twitter Sentiment Analysis : The Good the Bad and the OMG! "fifth International AAAI Conference on Weblogs and Social media.
[5] Milagros Fernandez-Gavilanes , Tavara Alvarez-Lopez , Jonathan Juncal-Martinez , Enrique Costa-Montenegro , Francisco Javier Gonzalez-Castano , "An Unsupervised Approach for Sentiment Analysis in Twitter", Proceedings of the International Workshop on Semantic Evaluation ,june 4-5 2015.
[6] Manisha Shinde-Pawar , "Formation of smart Sentiment analysis Technique for Big Data "Vol.2 ,Issue 12 ,December 2014IJIRCCE.
[7] M.Vasuki , J.Arthi, K. Kayalvizhi , "Decision Making using Sentiment Analysis from twitter" Vol.2 ,Issue 12 ,December 2014 , IJIRCCE(an ISO 3297:2007 Certified organization).
[8] Chetan Kaushik and Atul Mishra , "A Scalable , Lexicon Based Technique for Sentiment Analysis" Vol.4 ,No. 5 September 2014, IJFCST.
[9] Sunil B.Mane , yashwant Sawant , Saif Kazi , Vaibhav Shinde , "Real Time Sentiment Analaysis of Twitter Data Using Hadoop " Vol 5(3) , 2014 , IJCSIT.
[10] Mahalaxhmi R. , Suseela S, "Big-SoSA: Social sentiment Analysis and Data Visulaization on Big Data " Vol 4, Issue 4, April 2015 , IJARCCE
[11] Penchalaiah.C , Murali.G , Suresh Babu.A , "Effective Sentiment Analysis on Twitter Data Using: Apache Flume and Hive"IJISET , Vol. I Issue 8,October 2014.
[12] Christopher Johnson , Parul Shukla , Shipla Shukla , "On classifying the Political Sentiment of Tweets"
[13] Tom White , "Hadoop: The Definitive Guide " Text Book.
[14] Shlomo Argamon-Engelson and Moshe Koppel and Galit Avneri , " Style-based Text Categorization: What Newspaper Am I Reading? " AAAI Technical Report WS-98-05. Compilation copyright © 1998, AAAI (www.aaai.org).
[15] Adam L.Berger, Stephen A.Della Pietra and Vincent J.Deila Pietra 1996. "A maximum Entropy approach to natural Language processing " Computational Linguistic. [16] Deepali Arora, Kin Fun Li and Stephen W. Neville, " Deepali Arora, Kin Fun Li and Stephen W. Neville" , 2015 IEEE 29th International Conference on Advanced Information Networking and Applications.
[17] Vinh Ngoc Khuc, Chaitanya Shivade, Rajiv Ramnath, Jay Ramanathan ," Towards Building Large-Scale Distributed Systems for Twitter Sentiment Analysis " . [18]John Dodd, "Twitter Sentiment Analysis",National College of Ireland, Final Thesis Report.