ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Review Paper on Mine, Process and Envisage Social Media Sentiment Data

Nidhi, Renu Singla

M.Tech Student, Department of CSE, SRCEM, Palwal, MD University, Haryana, India

Assistant Professor, Department of CSE, SRCEM, Palwal, MD University, Haryana, India

ABSTRACT : In this paper we glance at however we are able to mine and use the data from social media to develop and supply helpful and valuable insights from it. we tend to shall take a glance into a way to appraise and capture the various options, resources and data of the language utilized in micro blogging. To classify tweets into positive, negative and neutral sets.

KEYWORDS: Analysis of Sentiment Data, Big Data, Natural Language Processing, Machine Learning, Text Mining.

I. INTRODUCTION

Social media is a rapidly growing medium of communication. They have changed the way and helped communication to be much simpler and easier. The amount of data obtained from these social networks can be used to analyze user opinions and emotions. The Big Data framework Hadoop and its tools are used to store and analyze the data. Sentiment analysis is a really important part of research in Big Data. Big Data is a developing aspect where we are storing huge amounts of data. Big Data could be structured, unstructured or semi structured data which can be found anywhere over the internet. Analytics on such data help us gather various kinds of insights. These could be for security purposes, for marketing purposes, and many more. Sentiment analysis is an important part of Big Data as it involves unstructured data that is gathered from different social media sources to provide useful insights. The mining of the sentiment data is the key to gathering these insights as sentiment data represents different opinions and emotions, positive or negative, in multiple sources. Hadoop is a Big Data open source framework which allows us to store data and run applications on clusters of commodity hardware. It is not an ordinary data base as it allows us to store massive amounts of any kind of data and the ability to handle many tasks and jobs on these massive data sets.

In this paper, we shall use a simple technique of gathering the data from different data sets by the help of different Hadoop tools like Flume and Sqoop. Hadoop is a Big Data open source framework which allows us to store data and run applications on clusters of commodity hardware. It is not an ordinary data base as it allows us to store massive amounts of any kind of data and the ability to handle many tasks and jobs on these massive data sets.

Natural Language Processing, Machine Learning, Information Theory and Coding, and Text mining are some of the branches of computer science that are used for sentiment analysis. These approaches, methods and techniques will help us categorize and organize and structure this unstructured data, which is in the form of tweets, into positive, negative or neutral sentiment.

Sentiment analysis can be classified into two types:

1. **Subjectivity/objectivity identification**
2. **Feature/aspect based sentiment analysis**

EXISTING TECHNIQUES

1. **Machine Learning Techniques:**

Machine learning techniques can be classified on the basis of

a. Supervised Machine Learning Techniques

This basically uses a training data set for categorization of the document or text and has two different algorithms which have achieved great success1. They are as follows :

i. **Naïve Bayes**

ii. **Support Vector Machines**

Unsupervised Machine Learning Techniques

When classification is done without the help of a training data set. Some examples of these techniques are Point wise Mutual Information (PMI) and Semantic Orientation.

Text Mining Techniques

Text mining process has four stages:

a. Texts Collection
b. Preprocessing
c. Analysis
d. Validation

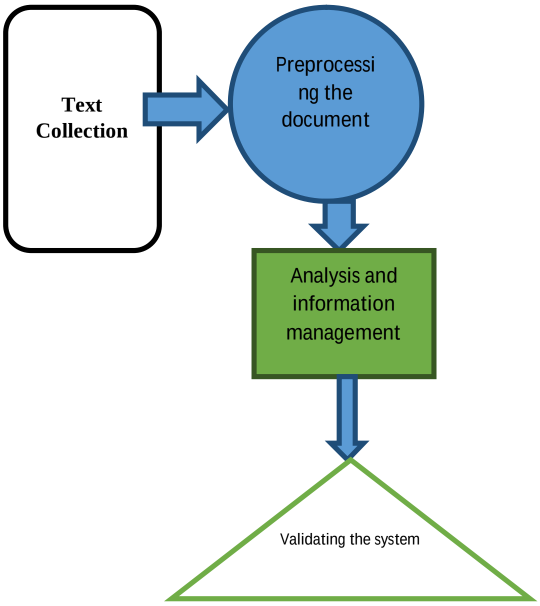
Text Collection
Preprocessing the document
Analysis and information management
Validating the system

Figure 1: Text Mining

1. **Natural Language Processing**

   The techniques or tasks of Natural Language Processing play a major role in Sentiment analysis. The different tasks like Part Of Speech tagging, Speech Recognition, N-gram algorithms, Markov model, sentiment lexicon acquisition and parsing techniques can express opinion on document level, sentence level and aspect levels.

2. **Hybrid Approaches**

   To perform the sentiment analysis according to our needs, we can use a combination of any of the above approaches.

- A combination of any of the two or more techniques mentioned above can be used for more accurate results for explicit and implicit sentiment analysis. For identifying Twitter messages, we use SVM and N-gram algorithms. .
- Generation of an implicit opinion for proper semantic orientation can be done with the combination of NLP and Machine Learning techniques with semantic approach.
- Combination of any of the NLP techniques with/without semantic approach, machine learning techniques can be made for generation of proper semantic orientation as and when needed for analyses of objective sentences that carry sentiment.

II. RELATED WORK

Sentiment Data is the representation of the different opinions, emotions and attitudes which can be found in social media posts, blogs, online product reviews, and customer support interactions. It is a data set of unstructured data. In this paper we are going to use a hybrid model of a corpus based and dictionary based approach where we can find the different orientations of the sentiment words in tweets. The proposed system focuses on a few important parts where data is extracted, processed and analyzed using Hadoop and its tools.

The process has 4 steps:

1. Stream, store and extract data from twitter through apps and data sets.
2. Preprocessing in Hadoop.
3. Classify the processes data by scoring
4. Provide visualization of the sentiment analysis.

1. Stream, store and extract data

This step will include the extraction and collection of data from the twitter apps and data sets formed. In this we build a data set from the unstructured data and store this data on a big data platform(in this case hort onworks/ clouder a Hadoop). A twitter application is used to store all the incoming and live data (tweets). This application data is then moved to the Hortonworks/Cloudera Big Data Platform using Flume, a tool used to stream data collection and aggregation system for massive volumes of data.

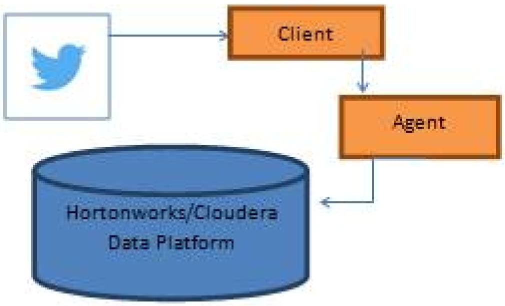

Figure 2: use of Flume to stream and store the data from Twitter on Hortonworks/Cloudera Data Platform

2. Preprocessing in Hadoop

When data is stored on the platform, it still is not structured and needs to be modified and put into tables. For this, we Run Hive Script on the data. The script will start running and a series of MapReduce Jobs will be executed on behalf of this script. Using sql querying we can then classify and convert this data into tabular format. Once the data is tabulated and assembled, we shall compare it to the dictionary file.

3. Classify the process data by scoring

We shall create a dictionary of our own for our closed domain. In this dictionary, there are going to be words and thresholds given. A comparison shall be made between the number of positive words and negative words to determine the score of the Tweet, which could be positive, negative or neutral. The value of each tweet shall be put into a new table containing the sentiment value for each Tweet.

4. Provide visualization of the sentiment analysis

Visualization of the sentiment analysis of all the data gathered shall be provided through excel sheets. Each Tweet shall be assigned a Sentiment value which will be displayed in tabular form once the sentiment analysis is performed. Excel sheets shall show all the accumulated data with their sentiment value of positive, negative, or neutral.

III. FUTURE WORK

A lot can be done in the future of Sentiment data. We have worked on a closed domain and implementing the same on an open domain is a big challenge and step ahead. Sentiment analysis is a tough process as to gather billions and trillions of data and analyze it as it is received will take a lot of storage and smart and good dictionaries to tabulate the data. The accuracy of the data on an open domain is still to be done publicly and hence this field of Sentiment Data has a lot of scope. Implementing these same techniques on an open domain is the biggest task for future work. Many techniques can be added and we shall hope to implement it on a larger scale.

IV. CONLCUSIONS

As we know, in today's world the peoples reaction and feedback to certain events that take place, products, decisions made, food items and many other situations are very fast to turn up on the internet. These reactions and feedback aren't private but they are publicly shared with the whole world through the internet. We require and automated system to consider these views, to take them into account and to work on them to make our products, decisions and opinions better. Gathering this Sentiment data in an open domain and taking all the sentiments into consideration is a needed in the world. The analysis of this sentiment data could prove very useful in predicting people's opinions, current trends, political views, events in the future. Analysis could also help in Business Intelligence applications and increasing the ROI of different organizations. Some organizations like SAS, SAP and TCS are already using sentiment analysis in their applications. Automatic Sentiment Analysis is a positive step for the future.

REFERENCES

1. Varsha Sahayak, Vijaya Shete, Apashabi Pathan, "Sentiment Analysis on Twitter Data" , International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 1, Volume 2 (January 2015)
2. Ronen Feldman, "Techniques and Application of Sentiment Analysis", Communication of ACM, April 2013, vol. 56.No.4.
3. Ana C.E.S Lima and Leandro N.de Castro, "Automatic Sentiment Analysis of Twitter Messages", IEEE Fourth International Conference on Computational Aspect .of Social Networks (CASoN), p.52-57, 2012.
4. Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P.Sheth, "Harnessing Twitter „Big Data‟ for Automatic Emotion Identification " , ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on privacy, Security, Risk and Trust,p.589-592, 2012.
5. Shichang Sun, Hongbo Liu,Hongfei Lin, Ajith Abraham, "Twitter Part of Speech Tagging Using PreClassification Hidden markov Model", IEEE International Conference on Systems, Man and Cybernetics, October 14-17,p.1118-1123, 2012.
6. Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.
7. Huising Xia, Min Tao and Yi Wang, "Sentiment Text classification of customers Reviews on the Web Based on SVM", IEEE Circuts and System Society, Sixth International Conference on Natural Computation (ICNC), p.3633-3937, 2010.
8. Bo Pang and Lillian Lee, "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval, vol.2, No1-2(2008)1-135.
9. Jalaj S. Modha, Gayatri S. Pandi, Sandip J. Modha, "Automatic Sentiment Analysis for Unstructured Data" , IJARCSSE
10. Xing Fang, Justin Zhan, "Sentiment analysis using product review data".
11. G.Vinodhini, RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey", IJARCSSE.
12. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data"
13. Efthymios Kouloumpis, Theresa Wilson, Theresa Wilson, "Twitter Sentiment Analysis: The Good the Bad and the OMG!", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.
14. Douglas R. Rice, Christopher Zorn, "Corpus-Based Dictionaries for Sentiment Analysis of Specialized Vocabularies"