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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

Extract, Refine and Visualise User Behaviour Analysis using Twitter by SVM in Big Data

Seema Fagna[1], Shweta Gambhir[2]

P.G. Student (MTech), Department of Computer Science & Engineering, NGF College of Engineering and Technology at Palwal, Haryana, India[1]

Assistant Professor, P.G. Student (MTech), Department of Computer Science & Engineering, NGF College of Engineering and Technology at Palwal, Haryana, India[2]

ABSTRACT: --- In this scheme we look at how we can mine and use the raw data from social media i.e. crowd sourcing platform i.e. twitter to develop and provide useful and valuable insights from as termed behaviour We shall take a look into how to evaluate and capture the different features, resources and information of the language used in micro blogging. To classify tweets into positive, negative and neutral sets using Support Vector Machine (Machine Learning Model under Supervised Machine Learning Tweets) and Map Reduce under Big Data will formed as amalgamated solution by which we can analyse behaviour of the user, city, country (groups of people) for any contextual subjective or objective matter thereof.

KEYWORDS: Support Vector Machine, Map Reduce, Big Data, Machine Learning, Natural Language Processing.

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. User Behaviour 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. User Behaviour 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 User Behaviour data is the key to gathering these insights as User Behaviour 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. Subsequently, the machine learning technique under the supervised machine learning mode i.e. SVM can be amalgamated with map reduce and classification sets of data can derived using hyper plane to categorised the drifting behaviour of the users for better analytical structure to come to best to conclude.

Machine Learning, Natural Language Processing Information and Text Mining Techniques 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. Support Vector Machine
      ii. Naïve Bayes Model
      iii. Decision Tress

   b. **Unsupervised Machine Learning Techniques**

      When classification is done without the help of a training data set. Some examples of these techniques are:

      i. Point wise Mutual Information (PMI)
      ii. Semantic Orientation.
      iii. Clustering

2. **Text Mining Techniques**

   Text mining process has four stages:

   a. Texts Collection
   b. Pre-processing
   c. Analysis
   d. Validation

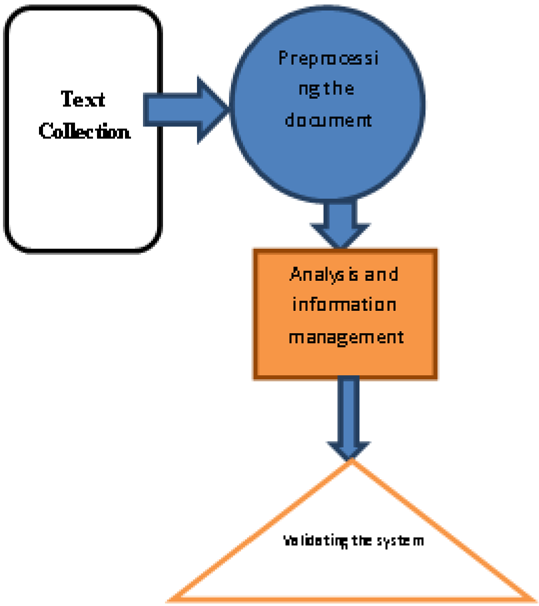

Text Collection
Preprocessing the document
Analysis and information management
Validating the system

Figure 1: Generic Technique for Text Mining Model enabling Extract, Espionage and Conclude the Context.

3. 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.

4. Hybrid Approaches

To perform the sentiment analysis according to our needs, we can use a combination of any of the above approaches. The combination of any of the two or more techniques what we proposed in our scheme i.e Map Reduce with SVM Classification 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.

II. Related work

Aditya Bhardwaj and Ankit kumar(2015)[1] have discussed on big data analysis. According to them, Big Data refers to the volume of data beyond the traditional database technology capacity to store, access, manage and compute efficiently. They said by analyzing this large amount of data companies can predict the customer behavior, improved marketing strategy, and get competitive advantages in the market. According to them hadoop is a flexible and open source implementation for analyzing large datasets using Map Reduce. They focused various emerging technologies such as Apache Pig, Hive, Sqoop, HBase, Zookeeper, and Flume that can be used to improve the performance of basic Hadoop Map Reduce framework. They said Apache Pig is a scripting language that can be used to reduce development time of Map Reduce program because it requires less number of lines of code and provides nested data types that are missing from Map Reduce. Hive provides easy to use platform for the developers who are comfortable in SQL language for Map Reduce programming, HDFS has the inability of random read/write to Big Data that can be provided by HBase. Theytransferred data between Hadoop and RDBS system using Sqoop, Zookeeper can be used for synchronization of Hadoop cluster and finally Flume can be used for moving streaming web log data to HDFS. Their paper also discussed fetching and executing Twitter tweets by using Hive query on HDInsight cluster and results shows that as we increase number of nodes in the cluster, then Map Reduce slot time increase but overall total time taken for executing Hive query decease.

Raj Kumar Verma and RituTiwari(2016) [2] have focused on social networking websites which is a source of various kind of information. They said this is because of the nature of these websites on which peoples comments and post their opinions on different types of topics i.e. they express positive or negative sentiments about any product that they use in daily life, complains and current issues etc. They said the sentiments help in getting information about various current trends and can be used further in deciding usefulness of some tasks, products and themes. Also social web data like twitter has a large amount of data that people post so it's become important to work on efficient intelligent systems that can do data refinement, analysis of tasks intelligently and efficiently. DhirajGurkhe and NirajPal(2014) [3] have discussed the effective Sentiment Analysis of Social Media Datasets Using Naive Bayesian Classification. The process involves extraction of subjective information from textual data. A normal human can easily understand the sentiment of a document written in natural language based on its knowledge of understanding the polarity of words (unigram, bigram and n-grams) and in some cases the general semantics used to describe the subject. The paper aims to make the machine extract the polarity (positive, negative or neutral) of social media dataset with respect to the queried keyword. The paper introduced an approach for automatically classifying the sentiment of social media data by using the following procedure: First the training data is fed to the Sentiment Analysis Engine for learning by using machine learning algorithm. After the learning is complete with qualified accuracy, the machine starts accepting individual social data with respect to keyword that it analyze and interprets, and then classifies it as positive, negative or neutral with respect to the query term.

Laurie Butgereit(2015) [4] has focussed on the event held on 1 November in South Africa, 2014 in which a coal silo collapsed at Eskom's newest power station, Majuba. The paper focused on the damage forced Eskom to implement rolling block-outs(called load-shedding) throughout the country. The paper investigated if it was possible to quantify the relative anger against Eskom as expressed in pairs of posts on Twitter (called tweets). The paper proposed an algorithm was developed that measured certain characteristics of the tweets such as swear words, emoticons, emojis, uppercase letters, and certain punctuation marks. The results were evaluated against results provided by two independent people acting as coders. These two people also evaluated the same tweets. The results show that as the polarity(or difference) in anger in two tweets increases, the algorithm is nearly as accurate as two human coders. A. K. Santra and S. Jayasudha(2012) [5] have focused on behavior of the interested users instead of spending time in overall behavior. The existing model used enhanced version of decision tree algorithm C4.5. In the paper, they use the Naive Bayesian Classification algorithm for classifying the interested users and also they presented a comparison study of using enhanced version of decision tree algorithm C4.5 and Naive Bayesian Classification algorithm for identifying interested users. The performance of this algorithm is measured for web log data with session based timing, page visits, repeated user profiling, and page depth to the site length.

III. PROPOSED ALGORITHM

A. *Design Considerations:*

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. 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 analysed using Hadoop and SVM as tool.

The process has 4 steps:

1. Stream, store and extract data from twitter through Twitter API and data sets.
2. Preprocessing in Hadoop Map Reduce algorithm.
3. Classify the processed data by scoring using Support Vector Machine
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 hortonworks/cloudera 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 the below figure depicts the scenario used under scheme.

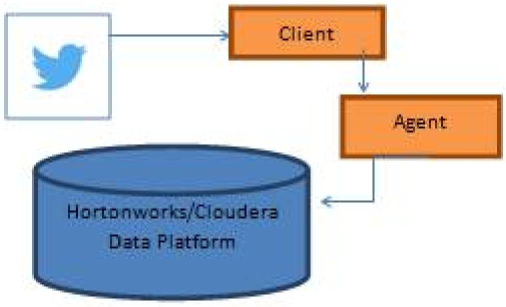


Figure 2: Flume to stream and store the data from Twitter on Hortonworks/Cloudera Data Platform

2. Preprocessing in Hadoop using Map Reduce

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 using Support Vector Machine

We shall create a dictionary of our own for our closed domain. In this dictionary termed to known a polarity, 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 using Support Vector Machine.

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

IV. PSEUDO CODE

Pseudo code of MapReduce:

```
class Mapper
method Map(nid n; node N)
    p←N:Tweets/|N.AdjacencyList|
    Emit(nid n, N)  ► Pass along graph structure
    for all nodeid m N.AdjacencyList do
    Emit(nid m, p) ► Pass  Score  to neighbors nodes for key pair value

class Reducer
method Reduce(nid |m, [p1, p2, …])
    M ←
    for all p counts [p1, p2, …] do
                if IsNode(p) then
            M←p ► Recover graph structure
            else
    s←s + p ► Sum incoming Score via tweets  contributions
    M.Score←s
    Emit(nid m, node M)
```

Pseudo code of SVM:

In this we present the user behavior formulation for SVM classification This is a representation proposed under the scheme.

SV classification:

$$\min_{f,\xi_i} \|f\|_K^2 + C\sum_{i=1}^{l}\xi_i \quad y_i f(x_i) \ge 1 - \xi_i, \text{ for all i} \quad \xi_i \ge 0$$

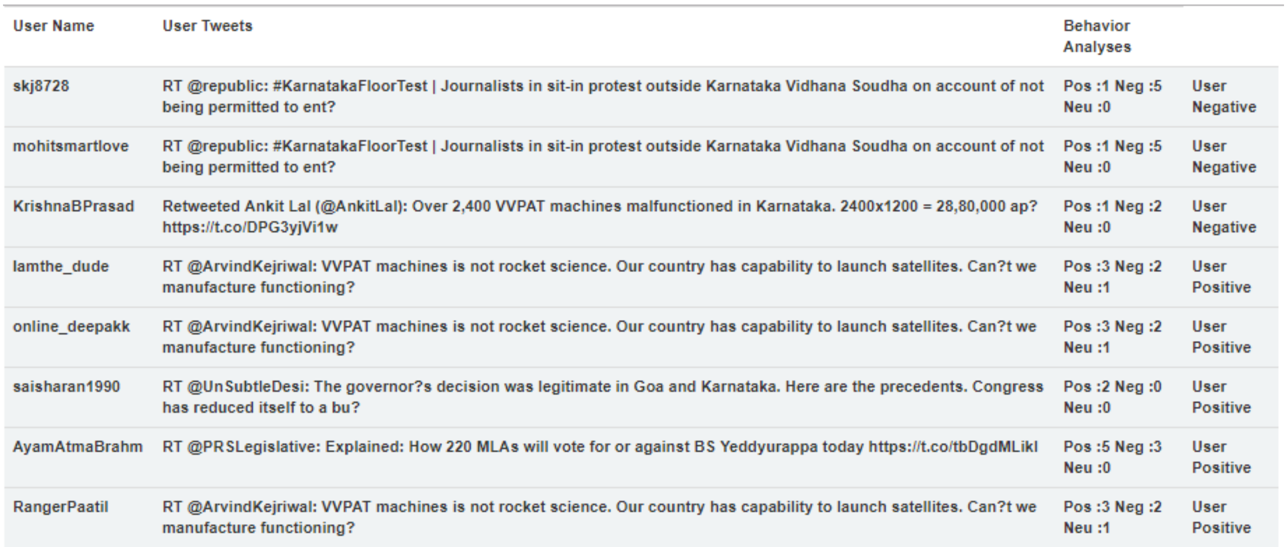
SVM classification, Dual formulation:

$$\min_{\alpha_i} \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) \quad 0 \le \alpha_i \le C, \text{ for all i;} \quad \sum_{i=1}^{l}\alpha_i y_i = 0$$

Variables ξ_i are called slack variables and they measure the error made at point (x_i,y_i). However, SVM becomes quite accurate when the number of training points is large. A number of methods for fast SVM training have been proposed as min $L = ½\, w'w - \sum \lambda_k\, (\, y_k\, (w'x_k + b) + s_k - 1) + \alpha \sum s_k$ to achieve the user behaviour under the drifting scenarios.

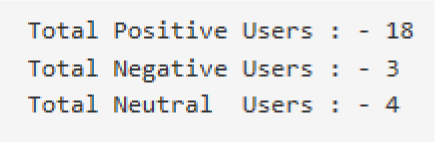
V. Simulation Results

Both MapReduce and Support Vector Machine are synchronous computation models. However, both support iterative algorithms and are used for parallel Computing. MapReduce does support iterative algorithms with the help of extensions and is data parallel. In this paper we reviewed three frameworks with respect to SVM based problems by extracting, analysing and depicting the behaviour analysed using least computation resources over Big data comparing their mechanisms for solving the same, with SVM and MapReduce the amalgamated model can achieve the best and accurate results therein.



User Name	User Tweets	Behavior Analyses	
skj8728	RT @republic: #KarnatakaFloorTest | Journalists in sit-in protest outside Karnataka Vidhana Soudha on account of not being permitted to ent?	Pos :1 Neg :5 Neu :0	User Negative
mohitsmartlove	RT @republic: #KarnatakaFloorTest | Journalists in sit-in protest outside Karnataka Vidhana Soudha on account of not being permitted to ent?	Pos :1 Neg :5 Neu :0	User Negative
KrishnaBPrasad	Retweeted Ankit Lal (@AnkitLal): Over 2,400 VVPAT machines malfunctioned in Karnataka. 2400x1200 = 28,80,000 ap? https://t.co/DPG3yjVi1w	Pos :1 Neg :2 Neu :0	User Negative
Iamthe_dude	RT @ArvindKejriwal: VVPAT machines is not rocket science. Our country has capability to launch satellites. Can?t we manufacture functioning?	Pos :3 Neg :2 Neu :1	User Positive
online_deepakk	RT @ArvindKejriwal: VVPAT machines is not rocket science. Our country has capability to launch satellites. Can?t we manufacture functioning?	Pos :3 Neg :2 Neu :1	User Positive
saisharan1990	RT @UnSubtleDesi: The governor?s decision was legitimate in Goa and Karnataka. Here are the precedents. Congress has reduced itself to a bu?	Pos :2 Neg :0 Neu :0	User Positive
AyamAtmaBrahm	RT @PRSLegislative: Explained: How 220 MLAs will vote for or against BS Yeddyurappa today https://t.co/tbDgdMLikl	Pos :5 Neg :3 Neu :0	User Positive
RangerPaatil	RT @ArvindKejriwal: VVPAT machines is not rocket science. Our country has capability to launch satellites. Can?t we manufacture functioning?	Pos :3 Neg :2 Neu :1	User Positive

Figure 3: Results achieved using SVM and Map Reduce under proposed scheme however hyper-plane classifying the sentiments based on polarity and the composition of words either positive, negative or neutral in respect to contextual query raised on twetter.



```
Total Positive Users : - 18
Total Negative Users : - 3
Total Neutral  Users : - 4
```

Figure 4: Behaviour Analyses via counts based on positive, negative and neutral using SVM & Map Reduce under proposed scheme.

DOI: 10.15680/IJIRCCE.2018.0604105

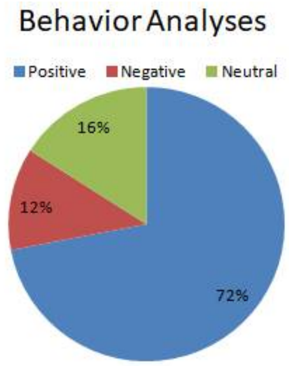


Figure 4: Graph Representation of Proposed Scheme whereas depicting the total numbers of tweets in contextual model vide behaviour based on positive, negative or neutral.

VI. Conclusion and Future Work

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 return over investment of different organizations. 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.

REFERENCES

1. Aditya Bhardwaj and Ankit kumar, "Big Data Emerging Technologies: A Case Study with Analyzing Twitter Data using Apache Hive, IEE Proceedings of 2015 RAECS UIET Panjab University Chandigarh, 21-22nd December 2015.
2. Dhiraj Gurkhe and Niraj Pal, "Effective Sentiment Analysis of Social Media Datasets using Naive Bayesian Classification , International Journal of Computer Applications (0975 8887) ,Volume 99, No. 13, August 2014.
3. Laurie Butgereit , "An Algorithm for measuring anger at Eskom during Load-Shedding using Twitter, IEEE, 978-1-4799-7498-6/15.
4. K. Santra and S. Jayasudha, "Classification of Web Log Data to Identify Interested Users Using Naïve Bayesian Classification, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012.
5. Sagiroglu, S., & Sinanc, D, "Big data: A review, IEEE International Conference on Collaboration Technologies and Systems (CTS), pp 42- 47, 2013.
6. 7.Pal, A., & Agrawal, S "An experimental approach towards big data for analyzing memory utilization on a Hadoop cluster using HDFS and MapReduce, IEEE, First International Conference on Networks & Soft Computing (ICNSC), pp.442-447, August 2014.
7. Bedi,P.,Jindal,V., & Gautam, A,"Beginning with Big Data Simplified, IEEE International Conference on Data Mining and Intelligent Computing (ICDMIC), pp.442-447, 2014.
8. Hassan. S., Yulan. H., and Alani. H., "Semantic sentiment analysis of Twitter." The Semantic Web– ISWC. Springer, pp. 508-524, 2012.
9. Abdul-Mageed. M., Diab. M., and Korayem M., "Subjectivity and sentiment analysis of modern standard Arabic." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Vol. 2. 2011
10. Almas Y., and Ahmad K., "A note on extracting sentiments in financial news in English, Arabic & Urdu." The Second Workshop on Computational Approaches to Arabic Script-based Languages. 2007.
11. Abdul-Mageed M., and Diab M., "AWATIF: A multi-genre corpus for Modern Standard Arabic subjectivity and sentiment analysis." Proceedings of LREC, Istanbul, Turkey, 2012.
12. Elhawary M. and Elfeky M., "Mining Arabic Business Reviews." Data Mining Workshops (ICDMW), P. 1108-1113, 2010.
13. Pang B., and Lee L. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." Proceedings of the 42nd annual meeting on Association for Computational Linguistics. 2004.

DOI: 10.15680/IJIRCCE.2018.0604105