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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

User Behaviour Analysis using KNN/SVM vide Tweets on Big Data

Pushpa, Gaurav Garg

M. Tech(pursuing), Dept. of CSE, Advanced Institute of Technology & Management, Palwal, Haryana under the Affiliation of Maharshi Dayanand University at Rohtak, Haryana, India

Assistant Professor, Dept. of CSE, Advanced Institute of Technology & Management, Palwal, Haryana under the Affiliation of Maharshi Dayanand University at Rohtak, Haryana, India

ABSTRACT: An importanta part of our information-gathering behavior has invariably been to search out what others suppose. With the growing accessibility and recognition of opinion-rich resources like on-line review sites and private blogs, new opportunities and challenges arise as individuals currently will, and do, actively use data technologies to hunt out and perceive the opinions of others. The unexpected eruption of activity within the space of opinion mining and sentiment analysis, that deals with the procedure treatment of opinion, sentiment, and subjectiveness in dataset, has so occurred a minimum of partially as an immediate response to the surge of interest in new systems that deal directly with opinions as a superior object.This paper covers techniques and approaches that promise to directly change opinion-oriented information-seeking systems based on behavior analysis using KNN/SVM. Our focus is on ways that get to deal with the new challenges raised by sentiment-aware applications, as compared to people with certain (weight and measures) opinions which is already gifted in additional ancient fact-based analysis. We tend to embody material on summarization of appraising text and on broader problems relating to privacy, manipulation, and economic impact that the event of opinion-oriented information-access services provides rise to. To facilitate future work, a discussion of obtainable resources, benchmark datasets, and analysis campaigns is analyzed using twitter via Big Data using Hadoop.

KEYWORDS: K-Nearest Neighbour (KNN), Support Vector Model (SVM), Map Reduce, Hadoop.

I. Introduction

The current scenario with behavioural analysis would like of extracting emotional and sentimental data from the social arena which is rich with opinions, although for the most part exists in world wide web , was countered some way with bigdata solutions using machine learning. However, hadoop revolutionized the whole setup. Bigdata hadoop has been principally instrumental in adding an extra valency to the social information gathered from social sites like Facebook, Twitter, Pinterest, Instagram, whereas at the same time amalgamating technology with business pursuits for mutual profit and cooperation. With machine learning and alternative tools like listening tools and sentiment analysis, bigdata hadoop has been line of work with success to the business world, whereby it digs out unstructured data from countless Facebook Posts, Twitter Tweets, and Pinterest Pins. Enterprises sites use bigdata hadoop for storing, reporting, and process data like “how many folks checked-in province throughout twelvemonth celebrations?” Not solely effective measure the business homes, hotels, and also the aviation business creating best use of this mined information, however conjointly the social sites like Facebook. One will gauge the worth of bigdata in social media analytics if one goes through the “feeling list” of Facebook – the list with variety of emotions: positive or negative, created calculative mathematical for sentimental analysis, since the computing language fails to require into thought jumbled-up human emotions. although the linguistics wide accepts binary statements like “the flight was snug however didn't just like the food served on-board,” Facebook has crisped its list of emotions felt so as to scale back quality to minimum, thereby increasing the effectuality of information collected, that has been reportable as efficient, authentic and agile for certain effective measures .

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:*

- Establishing Connection Twitter Authorization using FLUME or Twitter4J
- Storing and Preserving Data (Tweets) which is in JSON in HDFS along-with HBASE
- Creating Meta Structures and Tables in HIVE
- Integrating and Mapping JSON data with HIVE meta Structures
- Extracting data using HIVEQL and Map Reduce
- Pre-Processing on extracted Data
- Forming KNN/SVM classifiers for results.

B. *Description of the Proposed Algorithm vide KNN and SVM classifiers:*

Step 1 : KNN linear correlation

KNN linear correlation quantifies the strength of a linear relationship between two variables. When there is no correlation between two variables, there is no tendency for the values of one quantity to increase or decrease with the values of the second quantity.

$$r = \frac{Covar(x,y)}{\sqrt{Var(x)Var(y)}}$$

$$Covar(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n}$$

r : *Linear Correlation*

$Covar$: *Covariance*

Var : *Variance*

$$Var(x) = \frac{\sum(x-\bar{x})^2}{n}$$

$$Var(y) = \frac{\sum(y-\bar{y})^2}{n}$$

Equation 1: KNN linear correlation

r will only measures the strength of a linear relationship and is always between -1 and 1 where -1 means perfect negative linear correlation and +1 means perfect positive linear correlation and zero means no linear correlation.

Step 2: SVM Categorization:

For reasons of both efficiency and efficacy, feature selection is widely used when applying machine learning methods to text categorization. To reduce the number of features, we first remove features based on overall frequency counts, and then select a small number of features based on their fit to categories as under:

1. Say, opinion type = w and drifting outcome = P.
2. P(w,p) is the joint probabilities and P(p) and P(w) are the marginals.
   P(w,p) = P(w|p) * P(p) = P(p|w) * P(w).
3. From the center cells we have P(w,p) and from the side/bottom we get P(p) and P(w).
4. Depending on what you need to calculate, it follows that:
   (1): P(w|p) = P(w,p) / P(p)

(2:)P(p|w) = P(w,p) / P(w), which is what you did with P(opinion, yes) = 3/14 and P(w) = 5/14, yielding (3/14)   (14/5), with the 14's cancelling out.

IV. PSEUDO CODE

Step 1: Extraction data via Flume from twitter or Twitter4J

```
/usr/hdp/2.2.4.2-2/flume/bin/flume-ag agent –conf ./conf/ -f conf/flume.conf –name TwitterAgent –
Dflume.root.logger=DEBUG,console –n TwitterAgent
```

Step 2: Hive Script

```
CREATE EXTERNAL TABLE Mytweets_raw (
   id BIGINT,created_at STRING,source STRING, favorited BOOLEAN, retweet_count INT,   retweeted_status
  STRUCT<text:STRING,user:STRUCT<screen_name:STRING,name:STRING>>,
  entities STRUCT< urls:ARRAY<STRUCT<expanded_url:STRING>>,
  user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,
  hashtags:ARRAY<STRUCT<text:STRING>>>,
  text STRING, user STRUCT<screen_name:STRING,  name:STRING, friends_count:INT,  followers_count:INT,
  statuses_count:INT, verified:BOOLEAN,  utc_offset:INT,  time_zone:STRING>,in_reply_to_screen_name
 STRING)
  ROW FORMAT SERDE 'org.apache.hive.hcatalog.data.JsonSerDe'
LOCATION '/data/tweets_raw';
```

Step 3: KNN model

```
kNN (dataset, sample){
   1. Go through each item in my dataset, and calculate the "distance"
from that data item to my specific sample.
2. Classify the sample as the majority class between K samples in
the dataset having minimum distance to the sample.
  3. Compute dataset containing indices for the K smallest distances $d(X_{i},x)$.
  4. return majority label for ($Y_i$ where $i \in I$)
}
```

Step 4: SVM Model.

```
1.  initialize $y_i$ = YI r $i \in I$
2.  REPEAT
3.  compute SVM solution w, b for data set with imputed labels
4.  compute outputs fi = hw, xii + b for all xi in positive bags
5.  set $y_i = \text{sgn}(f_i)$ for every $i \in I$, $Y_I = 1$
6.  FOR (every positive bag $B_I$)
7.  IF (P $_{i\in I}$(1 + yi)/2 == 0)
    compute i ∗ = arg maxi∈I fi
    set $y_i*$ = 1 END
    END WHILE (imputed labels have changed)
    OUTPUT (w, b)
.
```

V. SIMULATION RESULTS

User opinion analysis system that predicts the opinion of user whether the user is in drifting mode, positive or negative on the basis of the tweet_id of user on live social twitter data. Also to predict the general opinion of users in different locations in particular time stamp in a certain context and depicted in graph form.

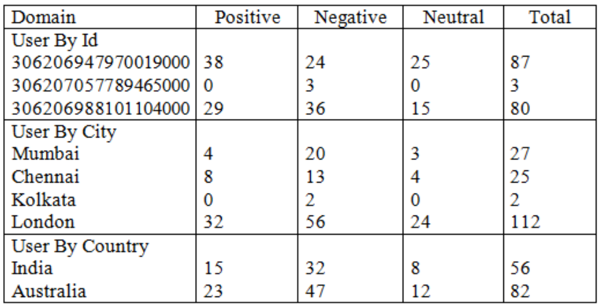
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017



Domain	Positive	Negative	Neutral	Total
User By Id				
306206947970019000	38	24	25	87
306207057789465000	0	3	0	3
306206988101104000	29	36	15	80
User By City				
Mumbai	4	20	3	27
Chennai	8	13	4	25
Kolkata	0	2	0	2
London	32	56	24	112
User By Country				
India	15	32	8	56
Australia	23	47	12	82

Figure 5.1: Statistical Data

Figure 5,1 shows the statistical data of the tweets used to analyse opinion of user corresponding to user_id, city and country.

VI. Conclusion and Future Work

For analyzing the user opinion, first of all twitter data is extracted using flume. The data extracted is available is in unstructured (JSON) format. The data is integrated with Hadoop. Using hive it is given a tabular form i.e. a structured form of data is obtained. Maven framework is used to get the executable jar to integrate eclipse and Hadoop. Data needs to be filtered before analyzing. Data is cleaned by removing stop words. For classification, KNN/SVM has been used.

For using naïve KNN/SVM technique, we have used a dictionary which stores a list of words that are positive, negative and neutral. Lastly, data is imported to excel to give a graphical form and to get the results. In the scheme, we can identify the user opinion with the help of user_id whether the user is positive, negative or in drifting mode. Also, the system tells the general behavior of users country-wise as well as city-wise for a particular topic. The system is 70-80% accurate.In future, the data can be from multiple sources at the same time. Also various different tools like R, tablue can be integrated, also we can continue with ontology in it. Finally, multiple topics also can be taken into consideration. Further works can be done to improve the efficiency and accuracy.

References

1. Aditya Bhardwaj and Ankit kumar, “Big Data Emerging Technologies: A Case Study with Analyzing Twitter Data using Apache Hive”, IEEE 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. A. 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. 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.