Volume 2, Issue 10 October 2016 ISSN NO: 2454 - 7514

North Asian International Research Journal Consortium

North Asian International Research Journal Of Science, Engineering and Information Technology

Chief Editor
Dr. Bilal Ahmad Malik

Publisher
Dr. Bilal Ahmad Malik

Associate Editor
Dr.Nagendra Mani Trapathi

IRJIF IMPACT FACTOR: 3.821

NAIRJC JOURNAL PUBLICATION

North Asian
International
Research Journal Consortium

Welcome to NAIRJC

ISSN NO: 2454 -7514

North Asian International Research Journal of Science, Engineering & Information Technology is a research journal, published monthly in English, Hindi. All research papers submitted to the journal will be double-blind peer reviewed referred by members of the editorial board. Readers will include investigator in Universities, Research Institutes Government and Industry with research interest in the general subjects

Editorial Board

Address: -North Asian International Research Journal Consortium (NAIRJC) 221 Gangoo, Pulwama, Jammu and Kashmir, India - 192301, Cell: 09086405302, 09906662570, Ph. No: 01933-212815,
Email: nairjc5@gmail.com, nairjc@nairjc.com, info@nairjc.com Website: www.nairjc.com

IRJIF IMPACT FACTOR: 3.821

DETECTION AND ELIMINATION OF FAKE REVIEW FROM REAL-TIME DATA USING CLOUD COMPUTING

YASHASHRI BONDE [1], KAJAL KHARABI [2], AMARDEEP SABALE [3] & DHANSHRI PATIL [4]
[1234][NMIET], Savitribai Phule, Pune University, Pune, India

ABSTRACT

Today massive amount of people are using Internet, for expressing their personal views on a particular issue or a person. While expressing themselves they may hurt the sentiments of specific community or a person. To avoid this Opinion Mining is used. In today's big data era opinion mining on a customer's mining has become one of the most effective ways to rigorously use the great amount of information. Opinion mining uses unstructured information format, which is mainly related to emotional analysis, subjective comments recognition. The process of opinion mining could be on the level of the texts and may be sentences also. Opinion mining and sentiments analysis involve opinion integration algorithm conflicting opinion analyzing problem. It is an important part of knowledge discovery to extract hidden information from unstructured or semi-structured data. In the era of key algorithm for opinion mining and integrating, opinion integration algorithm plays important role, to avoid the unnecessary parts from users comments or reviews. The research of opinion integration relates with its four core parts, opinion spam detection, opinion summarization, opinion visualization and opinion assessment. Opinion integration algorithm and opinion spam detection uses evidence classifier. Spam refers to fake reviews posted by users. Identification of spam comment becomes an important task for improving accuracy of opinion mining.

1. INTRODUCTION

User or customer needs secure and trustworthy opinion or review about a product or person. Opinion or review plays very important role while purchasing a product. Generally user decides quality of product after reading reviews about that particular product. Person normally decides opinion about a person after reading opinion of that person given by other person. So, for this reason opinions/reviews must be genuine. Sometimes user posted unwanted or useless comments/opinions on the particular situation, so these comments should be called as spam. So, read that comment is a time consuming process. Therefore deleting and eliminating these opinions/reviews on time, so it is very easy for user for making their decision. Companies use these opinion and feedback to improve sales. Many strategies can be planned for particular product as per feedback from user. This will develops a scope for a corrupt people to put fake review for improving the reputation of company. This will mislead a customer about the product. So, it demands for a system which will detect and omit fake reviews given for misleading the customer.

This work depends on the opinion integration [1] algorithm. This algorithm includes opinion spam detection, opinion summarization, opinion visualization and opinion assessment. This algorithm actually analyzed the user comments based on the categorical classifier. Today, Internet promotes the very fast development in economic industry and technology. The trend of online shopping is very popular. For using internet peoples can express their feeling in the form of opinions/reviews on particular community, subject/ product etc. But many times this reviews/opinions has been proven as wrong/fake.

2. RELATED WORK

Review spam detection [2] proposes to allow duplicate detection and classification to identify review spam. It focuses on feature construction and model building. Opinion spam and analysis [3] identified three types of spam. Detection of spam is done in type 1 as, by detecting duplicate review. Then detecting type 2 and type 3 review by supervised learning with manually labelled training. Cluster based one class ensemble for classification; Random subset for training phase [4] is comprised of a balanced number of outliers. The parameters of classifier are optimized on respective of training set. For identifying suspicious review alternative strategies are evaluated for aggregating the criteria into a single ranking [5]. Evaluation suggests the positive reviews which quickly follow negative review are suspicious. Feature extraction of product review investigated potential features for identifying ad reviews. By using content, statistical, and social features for classifying reviews with SVM [6], it’s possible to achieve an accuracy of 94.55% for cosmetics ad reviews vs. general articles. The best accuracy of 87.79% can be obtained in classifying ad reviews vs. non-ad comments in discussion forums. For finding unusual review pattern, the problem of identifying atypical behaviors of reviewers is studied. The problem was formulated as finding unexpected rules and rule groups [7]. A set of expectations was defined, and proposed the corresponding unexpectedness measures. Maximum Entropy Modelling for Assessing Results on Real-Valued Data, This work, proposed a well-founded approach for assessing results on real-valued rectangular databases. Introduced maximum entropy models [8] for general real-valued data that can be used to assess whether or not a discovery is the trivial result of the row and column marginal distributions in the database. We gave theory for incorporation.

3. MOTIVATION

Today opinions/reviews on the web are rapidly more used in practise by customers, administration and companies for their decision making. The posted reviews are useful only if reviews posted without any incorrect intention. To detect genuine reviews become more and more critical (difficult).

Online survey shows that around 90% of customers are satisfied after reading 10 on 10 rating and good review. Many customer decide to buy product or not only by following reviews, But when intention of a person is not good, behind giving review such opinion or review can be spam. There is needed to detect such spamming activities to make sure that opinions/reviews on the web are trustworthy source of information. Therefore, there is need to develop a system which will detect and eliminate fake reviews/opinions for avoiding people from getting mislead.

4. IMPLEMENTATION DETAILS

4.1 System architecture

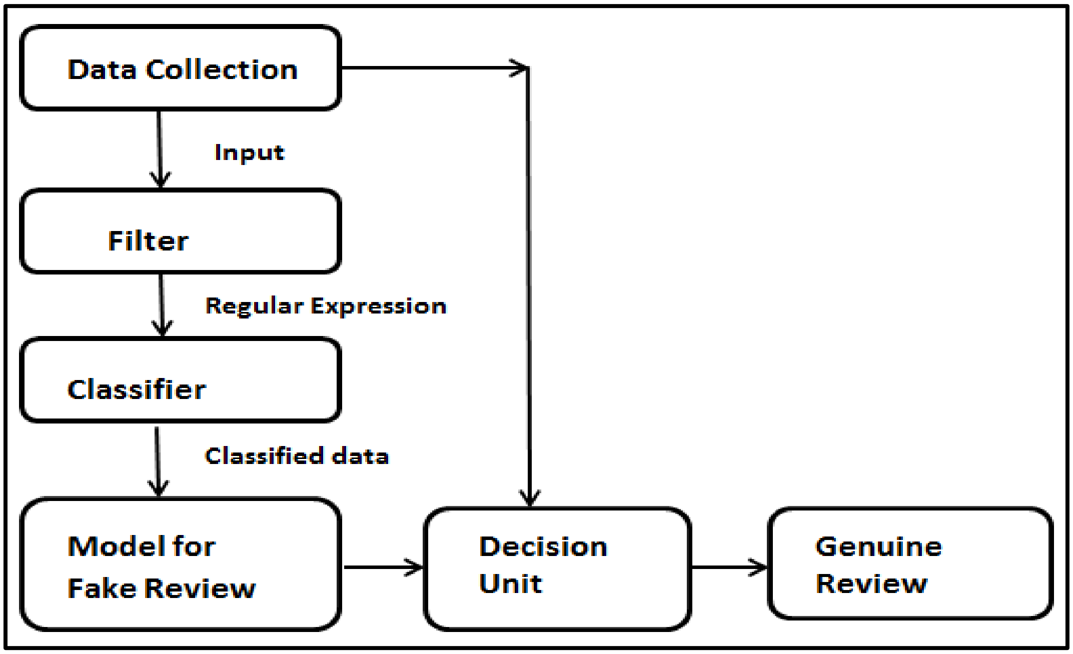

Fig.1: System Architecture

Figure 1 shows architecture for this system which consists of four main blocks. The blocks for the system are: Filter, Classifier, Model of Fake Review and Decision Unit.

Data will be gathered from particular sources this data will be in the JSON format. JSON format consist of Key-Value pair. This collected data is given as input to filter. Basic function of Filter is to filter data and convert collected data into Regular Expressions. When review or comments are fetched it may have its IP address along with it, which is not required at any stage. So, this will be eliminated in this block. User might be using smileys in review or comments to express them. This smileys are also eliminated as that is unwanted part.

Filtered data is applied to the classifier. Classifier is nothing but, a unit of classification which constructs the classification model based on training data set and that model classifies the new data. There are many classification algorithms available like Decision Tree Classifier, Select Tree Classifier, Evidence Classifier and many more. For this system we used Categorical Classifier [9].

Model of fake review is kind of database in which spam words are stored and these spam words will get compared with the new input taken in decision unit. Decision unit will simply take a input from collected data and compared it with the store spam words and after comparison decision unit will omit the fake reviews and it will generate genuine review.

4.2 Modules

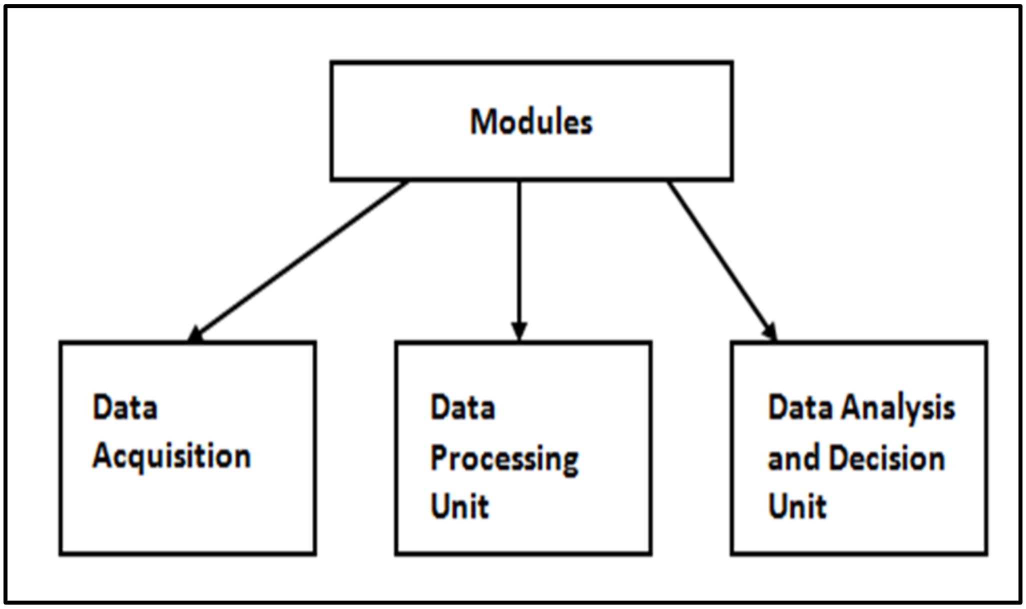

Fig.2: Modules used in System

4.2.1 Data Acquisition

Parallel Processing is needed to analyze the Big data. System pre-processes data under many situations to integrate the data from different sources, which not only decreases storage cost, but also improves analysis accuracy. The major tasks of data pre-processing techniques are data integration, data cleaning, and redundancy elimination. The data processing procedure is divided into two steps i.e. Real-time Big Data processing and Offline Big Data processing.

Offline Big data processing:

In offline data processing the base system transmits the data to the data centre for storage.

Real-time Big data processing:

In Real-time data processing the data are directly transmitted to the filtration and load balancer server. So because of this storing of incoming real-time data degrades the performance of real-time processing.

4.2.2 Data Processing Unit

Filtration and Load balancer are two basic functionalities of Data processing unit. Filtration generally includes filter the whole data and load balancing of processing power. For analysis purpose filtering is more important. This is helpful to improve the system performance. In Load balancer the data is divided into parts and then assign to different processing server. This two techniques are varies from analysis to analysis. For each processing server has its own algorithm to calculate statistical analysis. The tasks of system in parallel or independently so because of this improve the performance of system.

4.2.3 Data analysis

Aggregation and compilation server, results storage server, and decision making server are three major function of Data analysis unit. When result send to the compilation, the data not in aggregated form. So make the data in aggregated form it is important for proper storage and processing. It uses aggregation algorithm for store the organized result into storage and then send same copy of the result to the decision making server for making right decision.

4.3 Proposed Work

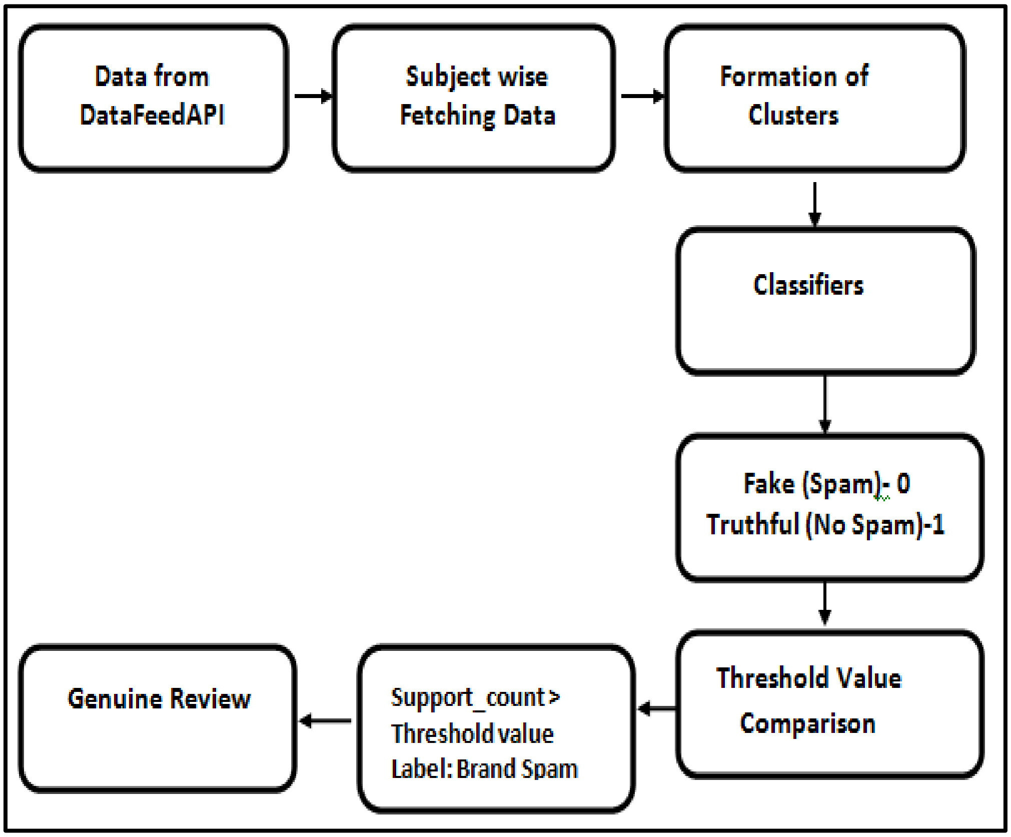

Fig.3: Problem Solving Approach

Review and comments are obtained from data feed API. API fetches review related to the subject specified suppose i want a review on PM Narendra Modi so, API will fetch all the reviews regarding with PM Narendra Modi. After that clustering will be done Clustering Algorithm is implemented for clustering the reviews in to groups. So if i fetch review on PM Narendra Modi But, i want review related to Digital India. So grouping or clustering will be done accordingly, Make in India, Jan-DhanYojna, Gram SadakYojna etc.

After completing clustering Pickle file is generated. Pickle file contains features required for detecting the original reviews. Pickle file will contain number of attributes like capital word, link, polarity etc. Pickle file is passed as input to classifier. Training and testing process is done by classifier. For solving approach categorical classifier is used.

After completing the process of classification Fake and Truthful reviews are detected. It means it classifies review in spam and no spam. 0 is assigned for spam and 1 is assigned for no spam. Meaningless or useless comments and advertisement are also marked with 0 i.e., spam. After marking 0 and 1 model will be generated and it will be considered as database for spam words.

Threshold value is prestored for spam word or phrase. Support count will be calculated for spam words in current input and it will get compared with predefined threshold value. If support count is more than threshold value it is labelled as Brand Spam[10]. Brand spam will get completely omitted and genuine review will displayed to user.

4.4 Algorithm

```
Algorithm Data_acquisition()
Input: specified subject
Output:  Data according to subject
Process:
Step1: Import library function for communication
Step 2: Authentication using OAuthHandler ().
Step 3: Capture API key and pass as parameter to handler.
Step 4: Set access token values
Step 5: Access API
```

5. EXPECTED RESULT

At end, result of spam detection is analyzed and decision will be taken on whether each review is spam or not a spam. Such result is helpful to users for making their respective decisions. System will be giving Spam free Result.

6. ACKNOWLEDGEMENT

We would like to take this opportunity to thank **Prof. Dhanshri Patil** for giving us all the help and guidance we needed. We are grateful to them for their kind support. Their valuable suggestions were very helpful.

7. CONCLUSION

The recent work related to spam detection is done with classifier, language model and Decision tree, which gives more efficiency and trustworthiness while considering reviews. The system gives convenience to administrators, flexible settings are available. It provides efficient and trust worthy opinion and feedback.

8. REFERENCES

1. Yue Lu, Chengxiang Zhai, "Opinion Integration Through Semi-supervised Topic Modeling", the International World Wide Web Conference Committee (IW3C2). WWW 2008, April 21–25, 2008, Beijing, China.ACM 978-1-60558-085-2/08/04
2. Nitin Jindal, Bing Liu ,"Review Spam Detection" , ACM Proceedings of the 16th international conference on World Wide Web, pp-1189-1190, 2007.
3. Nitin Jindal, Bing Liu, "Opinion Spam and Analysis", ACM Proceedings of the international conference on Web search and web data mining, pp.219-229,2008.
4. NedimLipka, Benno Stein, Maik Anderka, "Cluster-Based One-Class Ensemble for Classification Problems in Information Retrieval" Electronic, vol. 3, p. 87, 2009.
5. GuangyuWu, Derek Greene, Pdraig Cunningham ,"Merging multiple criteria to identify suspicious reviews", Proceedings of the fourth ACM conference on Recommender systems, pp.241- 244, 2010.
6. Jenq-Haur Wang, Yu-Hsin Liu, "Feature Extraction for Product Advertising Reviews Identification in Social Media" National Taipei University of Technology, Taipei, Taiwan. , vol 3, p.67, 2010.
7. Nitin Jindal, Bing Liu, Ee-Peng Lim "Finding unusual review pattern using unexpected rules", Proceedings of the 19th ACM international conference on Information and knowledge management, pp.1549-1552,2010. (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
8. Kleanthis-Nikolaos Kontonasios, JillesVreeken, Tijl De Bie, "Maximum Entropy Modelling for Assessing Results on Real-Valued Data"IEEE Computer Society,VOL. SE-I, No.2, 2011.
9. RisiThonangi ,"CLASSIFYING CATEGORICAL DATA", International Institute of Information Technology, Hyderabad December 2005.
10. Sushant Kokate, Bharat Tidke, "Fake Review and Brand Spam Detection using J48 Classifier" *Department of Computer Engineering, Flora Institute of Technology, Pune, India*
11. Ruxi Yin Hanshi Wang, Lizhen Liu "Research of Integrated Algorithm Establishment of a Spam Detection System", 2015 4th International Conference on Computer Science and Network Technology (ICCSNT 2015)

Publish Research Article

Dear Sir/Mam,

We invite unpublished Research Paper,Summary of Research Project,Theses,Books and Book Review for publication.

Address:- North Asian International Research Journal Consortium (NAIRJC)
221, Gangoo Pulwama - 192301
Jammu & Kashmir, India
Cell: 09086405302, 09906662570,
Ph No: 01933212815
Email:- nairjc5@gmail.com, nairjc@nairjc.com , info@nairjc.com
Website: www.nairjc.com