International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 02 | Feb -2017 www.irjet.net p-ISSN: 2395-0072

Efficiently Detecting and Analyzing Spam Reviews Using Live Data Feed

Jyoti N. Nandimath[1], Bhavesh S. Katkar[2], Vikram U. Ghadge[3], Arjun N. Garad[4]

[1]*Professor, Dept. of Computer Engineering, SKNCOE Pune, Maharashtra, India*
[2]*Student, Dept. of Computer Engineering, SKNCOE Pune, Maharashtra, India*
[3]*Student, Dept. of Computer Engineering, SKNCOE Pune, Maharashtra, India*
[4]*Student, Dept. of Computer Engineering, SKNCOE Pune, Maharashtra, India*

Abstract - *In recent year, online reviews have become the most important resource of customer opinion. Existing research has been focused used on extraction, classification and summarization of opinion from reviews in websites, forums and blogs. Now-a-days consumer can obtain information for products and service from online review resources, which can help them make decision. The social tools provided by the content sharing applications allow online user to interact, to express their opinions and to read opinions from other users. But the spammers provide comments which are written intentionally to mislead users by redirecting them to web sites to increase their rating and to promote products less known on the market. Reading spam comments is a bad experience and a waste of time for most of the online users but can also be harming and cause damage to the reader. Several researchers in this field focused on only fake comments. But, our goal is to detect fake comments which are likely to represent spam considering some indicators like a discontinuous own of text, inadequate and vulgar language or not related to the specific context will helps in giving correct feedback of various customers reviews about given product ,Mainly we have observed that previous work is focused on extraction, classification and summarization of opinion and checking of spam and non-spam. But, proposed system aims to Evaluate genuine result of filter comments ,so that business analyst can make the decision for their organization.*

***Key Words*: Distributed Computing, Cloud Computing, Server, Sentiment Analysis and Python.**

1.INTRODUCTION

Nowadays, the Internet contains a vast amount of text messages, and these messages need to be deeply analyzed and well estimated. Opinion mining now newly became one of the most heated areas in computer science. At the same time, electronic commerce, also known as e-commerce, is shooting up, which leads to a fast growth in the amount of users comments. And, the users comments do influence other buyers final choice. Therefore, making a good use of the comments will actualize their practical use. The process of opinion mining could be on the level of the texts, and the sentences as well. Opinion mining and sentiment analysis involve opinion integration algorithm, connecting opinion analyzing algorithm, etc. This paper focuses on opinion integration algorithm, and actualizes a comment fake as well as spam detection system, based on evidence classifier. As it is said before, the development of the Internet promotes the development of economy and technology; online shopping is getting more and more popular. While making their final decisions, users tend to rely on the online comments. However, some information, which is posted on purpose, not according to the fact, is useless for users. So these comments should be regarded as spam. If they are not detected and deleted on time, they may waste the users precious time of making their decisions. A nice way to solve this problem is to establish an opinion spam detection system.

2. LITERATURE REVIEW

The increase in the data rates generated on the digital universe is escalating exponentially. With a view in employing current tools and technologies to analyze and store, a massive volume of data are not up to the mark, since they are unable to extract required sample data sets. Therefore, we must design an architectural platform for analyzing both remote access real time and offline data. When a business enterprise can pull-out all the useful information obtainable in the Big Data rather than a sample of its data set, in that case, it has an influential benefit over the market competitors. Big Data analytics helps us to gain insight and make better decisions. To support our motivations, we have described some areas where Big Data can play an important role.

In healthcare scenarios, medical practitioners gather massive volume of data about patients, medical history, medications, and other details. The above-mentioned data are accumulated in drug-manufacturing companies. The nature of these data is very complex, and sometimes the practitioners are unable to show a relationship with other information, which results in missing of important

information. With a view in employing advance analytic techniques for organizing and extracting useful information from Big Data results in personalized medication, the advance Big Data analytic techniques give insight into hereditarily causes of the disease.

In the Same way data is also generated for the reviews of the product across various services but Sometimes we have to differentiate between fake reviews and Genuine Reviews for the input of our decision making process in Business.

3. PROPOSED SYSTEM

In our proposed system for analyzing real time as well as offline data for real-time applications using term Big Data we have divided real time Big Data processing architecture into three parts, i.e., 1) Data Acquisition Unit 2) Data Processing Unit and 3) Data Analysis and Decision Unit. In these three unit various algorithms or techniques will be implied on data for its analysis. The functionalities and working of three units is as explained and shown in diagram below:

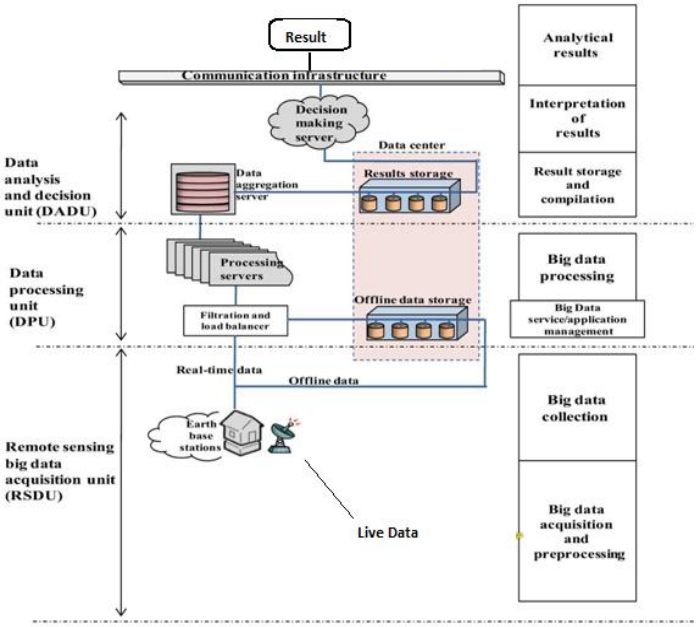

Fig -1: Architecture

3. 1 Data Acquisition Unit:-

The need for parallel processing of the massive volume of data was required, which could efficiently analyze the Big Data. For that reason, the proposed unit is introduced in the real time Big Data processing framework that gathers the massive volume of data from various available data gathering unit around the world. We assume that the data capturing unit can correct the erroneous data. For effective data analysis, the Base 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. Some relational data pre-processing techniques are data integration, data cleaning, and redundancy elimination. The data must be corrected in different methods to remove distortions caused due to the motion of the platform. We divided the data processing procedure into two steps, such as real-time Big Data processing and offline Big Data processing. In the case of offline data processing, the Base System transmits the data to the data centre for storage. This data is then used for future analyses. However, in real-time data processing, the data are directly transmitted to the filtration and load balancer server, since storing of incoming real-time data degrades the performance of real-time processing.

Algorithm I. Filtration and Load Balancing Algorithm

Input: Live Data Feed process data set

Output: filtered data in fixed size block and send each block to processing Mechanism

Steps:

1. Filter related data i.e. Processed data. All other unnecessary data will be discarded.

2. Divide the Data into Appropriate Key Value Pair.

3. Transmit Unprocessed data directly to aggregation step without processing.

4. Assign and transmit each distinct data block of Processed data to

various processing steps in Data Processing Unit.

Description: This algorithm takes live data and then filters and divides them into segments and performs load-balancing algorithm.

In step 1, related data is filtered out.

In step 2, filtered data are the association of different key value pairs and each pair is different numbers of sample, which results in forming a data block. In Next steps , these blocks are forwarded to processed by Data Processing Unit.

3.2 Data Processing Unit:-

In data processing unit, has two basic functionalities filtration and load balancer. Filtration mainly involves filtration of data and load balancing of processing power. Filtration makes process of filtering data which is useful to us for analysis and blocks other data. It will surely help to improve performance of system as we are only dealing with the useful data. Load balancer is part of server it will provide a provision of dividing data into parts and assign them to the various processing servers. The filtration and load-balancing algorithm varies from analysis to analysis. Each processing server has its own algorithm implementation for processing incoming segment of data from load balancer. Each processing server makes statistical calculations, any measurements, and performs other mathematical or logical tasks to generate intermediate results against each segment of data. These tasks are performed parallel and independently such that the performance of system is increased at an extent and result segments are generated in real time. The results generated by each server are then sent to the aggregation server for compilation, organization, and storing for further processing.

Algorithm II. Processing and Calculation Algorithm
Input: Filtered Data
Output: Normalized Disrupted data for Fake Review Calculation.
Steps:
1. For each event data or for the Product data, Categorical Data like G for good, A for average is extracted.
2. Normalize the disrupted data for all the live feed.
3. persist the data into data store and forward it.

Description: The processing algorithm calculates results for different parameters against each incoming filtered data and sends them to the next level.
In step 1, the calculation of Good and Average along with trend Furthermore, in the next step, the results are transmitted to the aggregation mechanism.

3.3 Data Analysis and Decision :-

This unit contains three major functions, such as aggregation and compilation server, results storage server, and decision making server. When results are to be send to the compilation the data is not in aggregated form. So it is necessary to make the given data in aggregated form for proper storage and processing. In this unit many aggregation algorithms are implied so that organized results are stored into the storage. The aggregation server also sends the same copy of that result to the decision-making server to process that result for making decision. The aggregation server also sends the same copy of that result to the decision-making server to process that result for making decision.

Algorithm III. Multi Modal Summarization Algorithm for Multiple Fake Reviews

Input: Normalized Disrupted Data of all Fake Reviews.

Output: Final result summary

1. Gather the data from data store in normalized format.

2. Apply Summarization for Individual modal pie from the total fake review data capture.

3. persist the final summary into data store.

Description: here the data is collected and the results from each modal is processed against all and then combines, organizes, and stores these results in NoSQL database.

4. ADVANTAGES:-

It provides product manufacturers information on their customers likes and dislikes, as well as the positive and negative comments on their products whenever available, giving them better knowledge of their products.

It also provides potential customers with useful and fair information on the products and/or services to aid in their purchase decision making process.

It involves some innovative modifications to enhance the efficiency and accuracy of the classifier.

It provides an accurate credibility on the analysis of users product reviews.

5. DISADVANTAGES:-

Security Concerns:-Just managing a complex application such as Stock market can be challenging. A classic example can be seen in the security model, which is disabled by default due to sheer complexity. If whosever managing the platform lacks the knowhow to enable it, your data could be at huge risk. This technique is also missing encryption at the storage and network levels, which is a major selling point for government agencies and others that prefer to keep their data under wraps.

Vulnerable By Nature:-Speaking of security, the very makeup of market makes running it a risky proposition. The framework is written almost entirely in programming, one of the most widely used yet controversial programming languages in existence. There has been heavily exploited by cybercriminals and as a result, implicated in numerous security breaches. For this reason, several experts have suggested dumping it in favour of safer, more efficient alternatives.

Not Fit for Small Data:-While big data isn't exclusively made for big businesses, not all big data platforms are

suited for small data needs. Unfortunately, Cloud happens to be one of them. Due to its high capacity design, lacks the ability to efficiently support the random reading of small les. As a result, it is not recommended for organizations with small quantities of data.

6. CONCLUSION AND FUTUREWORK

In this paper, we have presented a Spam as well as Fake review detection system. Our system is able to identify any products or services or trends features and opinions that are related either directly or indirectly. The extracted feature- opinion pairs along with the source documents are modelled and reliability score is generated. Currently. Handling informal texts that are very common with review documents is also one of our future works. We can also build a classifier which can process data like Multiple Language or Regional language.

7. REFERENCES

[1] Jindal N, Liu B,"Opinion Spam and Analysis," Proceeding of International Conference on Web Search and Web Data Mining. NY USA: ACM, 2008:210-229.

[2] J. Ge, Y. Qiu, C. Wu, and G. Pu, "Summary of genetic algorithms research," Application Research of Computers, vol. 25, pp. 2911-2916, 2008.

[3] Zhisong Pan, Bin Chen, "The Research on One-Class classifier," Electronic, vol. 3, p. 87, 2009.

[4] Shichuan Li, " Solution for Chinese Disorderly Codes " Network Administrator World, vol. 3, p. 2012.

[5] Xiao Li, Shengchun Ding, " Identification of waste product review information" Library and information technology,2013:63-68.

[6] Zhenqing Tian, Yue Zhou," Basic properties of entropy" Journal of Inner Mongolia Normal University, vol. 4, p.56, 2012.

[7] M.S.Patil, A.M.Bagade, "Review on Brand Spam Detection Using Feature Selection" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 9, September 2013.

[8] Chaitanya Kale, Dadasaheb Jadhav, "Spam Reviews Detection Using Natural Language Processing Techniques" International Journal Of Innovations In Engineering Research And Technology [IJIERT] ISSN: 2394-3696 Volume 3, Issue 1,JAN.-2016