ISSN 2395-1621

# Social Media Data Controlling and Processing using Big Data

Pooja Shelke, Neha Abhang, Ankita Shete, Neha Sonawane,
Mr. R. N. Devikar

poojashelke144@gmail.com , nehaabhang15@gmail.com, ankitashete2216@gmail.com,
sonawane94neha@gmail.com , rohit.devikar89@gmail.com

Department of Information Technology,
AVCOE, Sangamner, Savitribai Phule Pune University, Maharashtra

## ABSTRACT

**Twitter is an online social networking service with more than 300 million users, generating a huge amount of information every day. Twitter's most important characteristic is its ability for users to tweet about events, situations, feelings, opinions, or even something totally new, in real time. For one thing, there's no system of accuracy or reliability in place: Anyone can say just about anything. It can be a temptingly easy way to attack your detractors or for them to attack you igniting the sort of "Twitter war". This project will attempt to develop an analytical framework with the ability of in-memory processing to extract and analyse structured and unstructured Twitter data. The proposed framework includes twitter data controlling, stream processing, and data visualization components that is used to perform data ingestion task.**

**Keywords – Twitter, Big Data, Twitter Data Controlling, Twitter Data Processing, Anomaly Detection, Map Reduce**

ARTICLE INFO

**Article History**

Received: 24th June 2020

Received in revised form :

24th June 2020

Accepted: 26th June 2020

**Published online :**

**27th June 2020**

## I. INTRODUCTION

In social media, posts control has always been considered as the most challenging task for twitter analyst/data scientist. Due to the rapid development of digital technology, there is an opportunity for anomaly detection, Map Reduce, Big Data technology to be used in the field of social media network research which could help the researcher to solve a complex problem. One of the best things about Twitter indeed, perhaps its greatest appeal is in its accessibility. It's easy to use both for sharing information and for collecting it. The project motivation is to show the strength and the importance of real-time data analytics on social media streaming information. The continuous, large volume of unstructured information is called Big Data. Social Media plays very important roles in today's life, social Media are web-based online tools that enable people discover and learn new information, share ideas, interact with new people and organizations. Twitter has become a popular micro-blogging platform where millions of users express their opinions on a wide range of topics on a daily basis via tweets, producing large amounts of data every second that can be modelled as time- series data streams and analysed for anomalies. Twitter allows real-time collection of streams of tweets related to any specified topic keywords, hash tags, Social Media data controlling and processing using Big data or user names through their public streams service. In social media, posts control has always been considered as the most challenging task for twitter analyst/data scientist.

Data is the raw material for information before sorting, arranging and processing. It cannot be used in its primary form prior to processing. Information represents data after processing and analysis. The technology has been developed and used in all aspects of life, increasing the demand for storing and processing more data. As a result, several systems have been developed including cloud computing that support big data. While big data is responsible for data storage and processing, the cloud provides a reliable, accessible, and scalable environment for big data systems to function. Big data is defined as the quantity of digital data produced from different sources of technology [3].

This days the data produce from many sources such as Social networks, website and sensor network. Also the total of data volume is expanding Continuity .however; big data refers essentially to the following data types; Traditional enterprise data such as Customer ınformation in Data Base, the transactions websites companies. Machine generated and Sensors data such as smart meter, manufacturing sensors etc. and Social data such as social network and application platforms like Facebook, LinkedIn, what's app, Twitter and

YouTube. According to a recent report the most of data unstructured or semi structured and the size of data exists now is doubling in every two years. So between 2013 and 2020 it will go to 44 trillion GB from 4.4 trillion GB. Moreover the huge amount of data recorded mostly in nonstandard forms which cannot be analysed using traditional data models and methods. Big Data today have a wide range of challenges but the opportunities are also exists the right decision making, marketing strategies and improved customer relations, better public services and so on [6].

The concept of big data became a major force of innovation across both academics and corporations. The paradigm is viewed as an effort to understand and get proper insights from big datasets (big data analytics), providing summarized information over huge data loads. As such, this paradigm is regarded by corporations as a tool to understand their clients, to get closer to them, find patterns and predict trends. Furthermore, big data is viewed by scientists as a mean to store and process huge scientific datasets. This concept is a hot topic and is expected to continue to grow in popularity in the coming years.

Although big data is mostly associated with the storage of huge loads of data it also concerns ways to process and extract knowledge from it (Hashem et al., 2014). The five different aspects used to describe big data (commonly referred to as the five “V”s) are Volume, Variety, Velocity, Value and Veracity (Sakr & Gaber, 2014):

1. Volume describes the size of datasets that a big data system deals with. Processing and storing big volumes of data is rather difficult, since it concerns: scalability so that the system can grow; availability, which guarantees access to data and ways to perform operations over it; and bandwidth and performance.
2. Variety concerns the different types of data from various sources that big data frameworks have to deal with.
3. Velocity concerns the different rates at which data streams may get in or out the system and provides an abstraction layer so that big data systems can store data independently of the incoming or outgoing rate.
4. Value concerns the true value of data (i.e., the potential value of the data regarding the information they contain). Huge amounts of data are worthless unless they provide value.
5. Veracity refers to the trustworthiness of the data, addressing data confidentiality, integrity, and availability. Organizations need to ensure that data as well as the analyses performed on the data are correct [2].

## II. LITERATURE SURVEY

Several customers access the data in a cloud environment. Security of such data for storing, processing, retrieving, and updating in the cloud environment is critical. If the data is sensitive, access level for every step takes primary responsibility of cloud provider. Currently, the communication of internet-enabled devices (IED) is processing big data and require cloud environment due to their limited storage, processing, and energy capabilities (battery). Encryption is the primary security for transmitting the data from IED’s to the cloud, but there is no access control for the data stored in the cloud [1]. The term big data arose under the explosive increase of global data as a technology that is able to store and process big and varied volumes of data, providing both enterprises and science with deep insights over its clients/experiments. Cloud computing provides a reliable, fault-tolerant, available and scalable environment to harbour big data distributed management systems. Although big data solves much of our current problems it still presents some gaps and issues that raise concern and need improvement. Security, privacy, scalability, data governance policies, data heterogeneity, disaster recovery mechanisms, and other challenges are yet to be addressed [2].

Big data processing represents a new challenge in computing, especially in cloud computing. Data processing involves data acquisition, storage and analysis [3]. Big Data and cloud computing are two important issues in the recent years, enables computing resources to be provided as Information Technology services with high efficiency and effectiveness. Now a day’s big data is one of the most problems that researchers try to solve it and focusing their researches over it to get ride the problem of how big data could be handling in the recent systems and managed with the cloud of computing, and the one of the most important issue is how to gain a perfect security for big data in cloud computing, our paper reviews a Survey of big data with clouds computing security and the mechanisms that used to protect and secure also have a privacy for big data with an available clouds. [4] Disaster Recovery (DR) plays a vital role in restoring the org anization’s data in the case of emergency and hazardous accidents. While many papers in security focus on privacy and security technologies, few address the DR process, particularly for a Big Data system. However, all these studies that have investigated DR methods belong to the “single-basket” approach, which means there is only one destination from which to secure the restored data, and mostly use only one type of technology implementation [5].

Cloud storage is a widely utilized service for both personal and enterprise demands. However, despite its advantages, many potential users with enormous amounts of sensitive data (big data) refrain from fully utilizing the cloud storage service due to valid concerns about data privacy. An established solution to the cloud data privacy problem is to perform encryption on the client-end. This approach, however, restricts data processing capabilities. Accordingly, the research problem we investigate is how to enable real-time searching over the encrypted big data in the cloud. In particular, semantic search is of interest to clients dealing with big data [6]. Mobile Cloud Computing (MCC) is a recent technology used by various people worldwide MCC is a combination of mobile computing and cloud computing that presents various challenges like network access, elasticity, management, availability, security, privacy etc. [7].

Anomaly detection is an important problem with multiple applications, and thus has been studied for decades in various research domains. In the past decade there has been a growing interest in anomaly detection in data represented as networks, or graphs, largely because of their robust expressiveness and their natural ability to represent complex relationships. Originally, techniques focused on anomaly detection in static graphs, which do not change and are capable of representing only a single snapshot of data. As real-world networks are constantly changing, there has been a shift in focus to dynamic graphs, which evolve over time [8].

## III. PROPOSED SYSTEM

We have proposed a framework for real-time analysis of Twitter data. This Social Media data Controlling and Processing using Big data frame-work is designed to collect, filter, analyse, feature selection, training data, Map Reduce, anomaly detector, context switching, and summarization of twitter data.

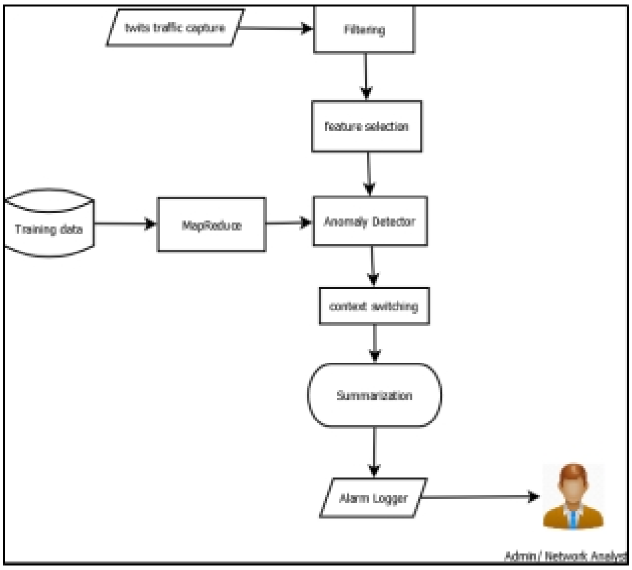


Fig 1. System Architecture

As we see the Fig, this project aims to show the strength and the importance of real-time data analytics on social and identification of twitter post. It is rather simple to use and exhibits the same performance level as a classical manual approach. Our goal is to detect, identify, controlling and processing the twitter post as early as possible and reduce the crime.

System provides a simple, efficient and fast solution in detecting, identifying, controlling and processing twitter data using anomaly detection technique. The prototype system proved reliable for rapid detection and identification of twitter post. It is rather simple to use and exhibits the same performance level as a classical manual approach. Our goal is to detect, identify, controlling and processing the twitter post as early as possible and reduce the crime.

## IV. TOOLS

1. JDK: The Java Development Kit (JDK) is a software development environment used for developing Java applications and applets. It includes the Java Runtime Environment (JRE), an interpreter/loader (java), a compiler (javac), an archiver (jar), a documentation generator (javadoc) and other tools needed in Java development. There are different version of JDKs available and also support various platforms. It is supported by platforms like Windows, Linux or Solaris and many more.

2. Net beans: Net Beans IDE is an open source or free tool. Integrated development environment or IDE is software which enables us to develop desktop, mobile and web applications in easy way with many build in features. The IDE supports application development in various languages, including Java, HTML5.

3 MySQL Query Browser: The MySQL Query Browser is an extension of MySQL which is basically a graphical tool provided by MySQL AB for creating and executing or optimizing queries in a graphical environment.

## V. HARDWARE AND SOFTWARE REQUIREMENTS

### 1. Software Requirements

OS- Microsoft Windows 7 or Above Programming
Language- JAVA
Database- MySQL
Tools- Net bean, MySQL Query Browser

### 2. Hardware Requirements

Processor- Core Intel 3 or Above
RAM- 2GB or Higher
Hard Disk- 100GB (min)

## VI. APPLICATION

1. We have proposed a framework for real-time analysis of Twitter data.

2. Complex system and abnormal situation in Social Media and Its Networks.

## VII.ADVANTAGES

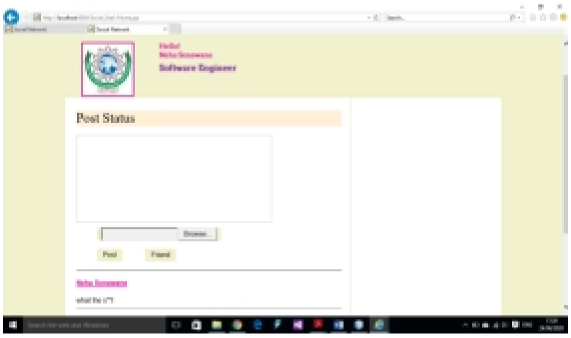
1. Ability to ‘move processing to data’ rather than ‘moving the data to computer resources.

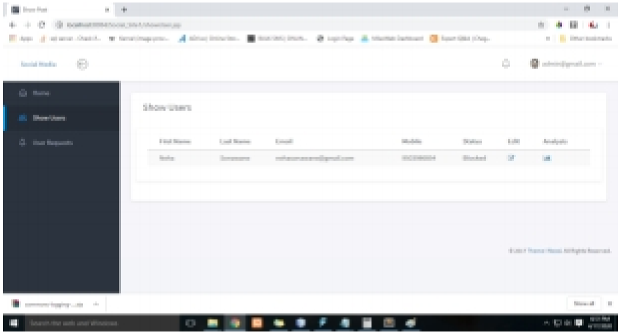
2. Designing a system that can detect, identify the post using anomaly detection technique.

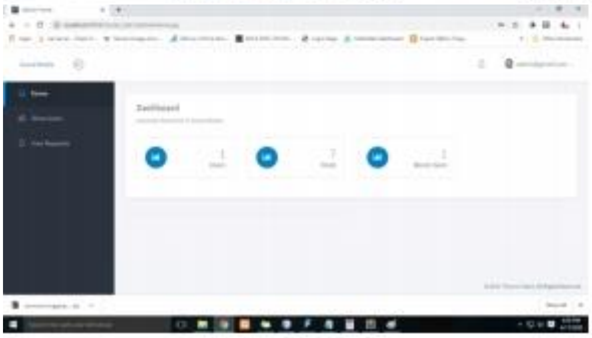
## VIII. RESULT

For User
Step 1: Registration
Step 2: Login and enter in account. User can update profile, Search Friend as well as Accept and Reject friend request.
Step 3: Post massage. If massage contain abuse word then system shows “Your massage contains anomaly content” and convert that abuse word into invisible format.

For Admin
Step 1: Registration
Step 2: Login and enter account where admin can see user performance, user request and can do updating like Block and If user doing misbehave regularly then Admin can block that user. For again activation of account users need to send request to admin to do active account.







## IX. CONCLUSION

Anomaly detection and Map Reduce plays an important role in the controlling, processing, detection and identifying twitter data/post. Our first objective is to detect twitter war. We propose a novel approach for early controlling, detection and identification of posts using anomaly detection. To detect bad twitter posts we use anomaly detection and Map Reduce. So without disturbing the posts we are able to convert negative posts into positive posts. It illustrates the collaboration of complementary disciplines and techniques, which led to an automated, robust and versatile system.

## X. REFERENCE

[1] Yenumula B Reddy “Big Data Processing and Access Controls in cloud Environment” Department of Computer Science, Grambling State University Grambling, USA, 4th IEEE International Conference on Big Data Security on Cloud, 2018

[2] Pedro Caldeira Neves, Bradley Schmerl, Jorge Bernardino, and Javier Cámara., “Big Data in Cloud Computing: features and issues”, International Conference on Internet of Things and Big Data, Rome, Athens, April 2016

[3] Nabeel Zanoon, Abdullah Al-Haj, Sufian M Khwaldeh., “ Cloud Computing and Big Data is there a Relation between the Two: A Study”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 17 (2017) pp. 6970-6982

[4]Elmustafa Sayed Ali Ahmed and Rashid A.Saeed., “A Survey of Big Data Cloud Computing Security”, International Journal of Computer Science and Software Engineering (IJCSSE), Volume 3, Issue 1, December 2014, pp.: 78-85

[5]Chang, V., 2015. Towards a big data system disaster recovery in a Private cloud. Ad Hoc Networks, 000, pp.1 – 18.

[6] K. Gai and M. Qiu., "Blend Arithmetic Operations on Tensor-based Fully Homomorphic Encryption over Real Numbers", IEEE Transactions on Industrial Informatics, December 2017. 33

[7] Survey on Security Issues in Mobile Cloud Computing and Preventive Measures Rahul Neware 1Computer Science & Engineering Department, GHRCE Nagpur, Maharashtra, India neware_rahul.ghrcemtechcse @raisoni.net, 2019

[8] Anomaly detection in dynamic networks: a survey Stephen Ranshous, 1, 2 Shitian Shen, 1, 2 Danai Koutra,3 Steve Harenberg,1,2 Christos Faloutsos3 and Nagiza F. Samatova1,2, Dec 6,2019