International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

YOUTUBE DATA SENSITIVITY AND ANALYSIS USING HADOOP FRAMEWORK

Devika Harikumar[1], Dolly Kapoor[2], Prof. Swapnil Waghmare [3]

[1]Student of Computer Engineering, Pillai HOC College of Engineering and Technology
[2]Student of Computer Engineering, Pillai HOC College of Engineering and Technology
[3]Professor of Computer Engineering, Pillai HOC College of Engineering and Technology

---

Abstract— We live in a digitized world today. An enormous amount of data is generated from every digital service we use which is called Big Data. Big data is the huge collection of data on internet which can be structured, semi-structured or unstructured which can't be processed at a single system. YouTube is one of the best examples of services which produces a huge quantity of data in a very short period. Since YouTube is an easily available platform, many people tend to misuse it to harass or threaten certain group of people or particular individuals using comment section on YouTube. Since there are millions of users using YouTube on a daily basis, different marketing companies use this platform for advertising. Therefore, before publishing harsh comments we need to filter the sensitive content and then upload it on YouTube. We thus propose a system to filter and analyze the content on YouTube to remove the sensitive content from comments. The proposed technique uses NLP (natural language processing) parsers for identifying the sensitive features. The sensitive content here are nouns and verbs because in most of the cases the identity of person or places are used for threatening and verbs describes the action of harm they intend to cause. Linear search technique is used to improve time consumption. The proposed technique is implemented using Hadoop, MapReduce and YouTube API. Hadoop is a system which provides a reliable shared storage of such huge datasets on the cloud and also provides an analysis system. The storage is provided by HDFS (Hadoop Distributed File System) and analysis is done by MapReduce. We are also focusing on doing analysis for YouTube data.

Keywords: Hadoop Distributed File System, Natural language processing, Hadoop, MapReduce, Big Data

1. INTRODUCTION

In the era of digitization the internet companies like Amazon, YouTube, Yahoo, Google and the internet addicted population in today's world are generating data in very large volume with great velocity and in structured, semi structured, unstructured formats including tweets, images, videos, blogs, and many more different sources. This huge generated data has given a birth to data called as Big data which is semi-structured/unstructured and also which is unpredictable in nature. This type of data is generated in real time from social media websites which is increasing exponentially on a regular basis. "Big Data is a word for data sets that are so huge and complex that data processing applications are insufficient to deal with them. Analysis of this data set can find new correlations to spot business sales, prevent diseases, preventing crime and so on." With billions of users are using Twitter to tweet about their most recent product buying experience or hundreds of thousands of check-ins on Facebook, millions of people talking about the recent activities done around the world on Facebook and millions of views generated by YouTube for a recently released movie trailer/videos being uploaded, we are in the world wherein we are heading into a social media data explosion. Major Companies in the world are already facing challenges getting useful information from the transactional data from their customers (for e.g. data captured by the e-commerce companies for increasing the sales of products based on the activity of the users). This type of data is structured or semi structured in nature and still manageable. However, social media data is primarily semi-structured and unstructured in nature. The unstructured nature of the data makes it very hard to analyze and very interesting at the same time. Whereas RDBMS are designed to handle structured data and that to only certain limits of data descriptions, Relational databases fails to handle this kind of semi structured, unstructured and huge amount of data called Big Data. Some of the key concepts used in Big Data Analysis are

1. Data Mining: Data mining is incorporation of quantitative methods. Using powerful mathematical techniques applied to analyse data and how to process that data. It is used to extract data and find actionable information which is used to increase productivity and efficiency.

2. Data Warehousing: A data warehouse is a database as the name implies. It is a kind of central repository for collecting relevant information. It has centralized logic which reduces the need for manual data integration.

3. MapReduce: MapReduce is a data processing paradigm for condensing large volumes of data into useful aggregated results. Suppose we have a large volume of data for particular users or employees etc. to handle. For that, we need Map Reduce function to get the aggregated result as per the query.

4. Hadoop: Anyone holding a web application would be aware of the problem of storing and retrieving data every minute. The adaptive solution created for the same was the use of Hadoop including Hadoop Distributed File System or

HDFS for performing operations of storing and retrieving data. Hadoop framework has a scalable and highly accessible architecture.

YouTube is one of the most popular and engaging social media tool for uploading, viewing videos and an amazing platform that reveals the users response through comments for published videos, number of likes, dislikes, number of subscribers for a particular channel. YouTube collects a wide variety of traditional data points including View Counts, Likes, and Comments. The analysis of the above listed data points makes a very interesting data source to extract implicit knowledge about users, videos, categories and community interests.

Most of the companies are uploading their product launch on YouTube and they anxiously await their subscribers reviews and comments. Major production based companies launch movie trailers and people provide their first reaction and reviews about the trailers. This further creates an excitement about the product. Hence the above listed data points become very critical for the companies so that they can do the data analysis and understand the customers' sentiments about their product and services.

1. This project will help user in understanding how to fetch a specific channel's YouTube data using YouTube API.

2. This project requires access to Google Developers Console and generates a unique access key. That unique key is required to fetch YouTube public channel data. With the help of the unique access key, the required data is fetched from YouTube using a Java application.

3. The extracted data is stored in HDFS file and then the data that is stored in HDFS is passed to mapper for finding key and final value which will be passed to Shuffling, sorting and then finally reducer will aggregate the values.

Big data is a largest buzz phrases in domain of IT, new technologies of personal communication driving the big data new trend and internet population grew day by day but it never reach by 100%. The need of big data generated from the large companies like facebook, yahoo, Google, YouTube etc for the purpose of analysis of enormous amount of data which is in unstructured form or even in structured form. Google contains the large amount of information. So; there is the need of Big Data Analytics that is the processing of the complex and massive datasets This data is different from structured data in terms of five parameters –variety, volume, value, veracity and velocity (5V's).

The five V's (volume, variety, velocity, value, veracity) are the challenges of big data management are:

1. Volume: Data is ever-growing day by day of all types ever MB, PB, YB, ZB, KB, TB of information. The data results into large files. Excessive volume of data is main issue of storage. This main issue is resolved by reducing storage cost. Data volumes are expected to grow 50 times by 2020.

2. Variety: Data sources are extremely heterogeneous. The files comes in various formats and of any type, it may be structured or unstructured such as text, audio, videos, log files and more. The varieties are endless, and the data enters the network without having been quantified or qualified in any way.

3. Velocity: The data comes at high speed. Sometimes 1 minute is too late so big data is time sensitive. Some organizations data velocity is main challenge. The social media messages and credit card transactions done in millisecond and data generated by this putting in to databases.

4. Value: It is a most important v in big data. Value is main buzz for big data because it is important for businesses, IT infrastructure system to store large amount of values in database.

5. Veracity: The increase in the range of values typical of a large data set. When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dirty data. Big data and analytics technologies work with these types of data.

ADVANTAGES

- YouTube is one of the most amazing platform that help to reveal the community feedback through comments for published videos.
- Number of likes, dislikes, number of subscribers for a particular channel can be found.
- In business for marketing new product .It is based on users reviews.

2. LITERATURE SURVEY

Hadoop Map Reduce is a large scale, open source software framework dedicated to scalable, distributed, data intensive computing. The framework breaks up large data into smaller parallelizable chunks and handles scheduling ▫ Maps each piece to an intermediate value ▫ Reduces intermediate values to a solution ▫ User-specified partition and combiner options Fault tolerant, reliable, and supports thousands of nodes and petabytes of data • If you can rewrite algorithms into MapReduces, and your problem can be broken up into small pieces solvable in parallel, then Hadoop's Map Reduce is the way to go for a distributed problem solving approach to large datasets • Tried and tested in production • Many implementation options. We can present the design and evaluation of a data aware cache framework that requires minimum change to the original Map Reduce programming model for provisioning incremental processing for Big Data applications using the Map Reduce model [4].

Authors Amogh Pramod Kulkarni and Mahesh Khandewal [2] stated the importance of some of the technologies that handle Big Data like Hadoop, HDFS and Map Reduce. The author suggested about various schedulers used in Hadoop

and about the technical aspects of Hadoop. The author also focuses on the importance of YARN which overcomes the limitations of Map Reduce.

Authors Sagiroglu and S.Sinanc [3] surveyed various technologies to handle the big data and there architectures. In this paper we have also discussed the challenges of Big data (volume, variety, velocity, value, veracity) and various advantages and a disadvantage of these technologies. This paper discussed an architecture using Hadoop HDFS distributed data storage, real-time NoSQL databases, and MapReduce distributed data processing over a cluster of commodity servers. The main goal of our paper was to make a survey of various big data handling techniques those handle a massive amount of data from different sources and improves overall performance of systems.

Authors Sagiroglu and S.Sinanc continue with the Big Data definition and enhance the definition given in [3] that includes the 5V Big Data properties: Volume, Variety, Velocity, Value, Veracity, and suggest other dimensions for Big Data analysis and taxonomy, in particular comparing and contrasting Big Data technologies in e-Science, industry, business, social media, healthcare. With a long tradition of working with constantly increasing volume of data, modern e-Science can offer industry the scientific analysis methods, while industry can bring advanced and fast developing Big Data technologies and tools to science and wider public.[1]

Author Kyuseok Shim [6] stated the need to process enormous quantities of data has never been greater. Not only are terabyte - and petabyte scale datasets rapidly becoming commonplace, but there is

consensus that great value lies buried in them, waiting to be unlocked by the right computational tools. In the commercial sphere, business intelligence, driven by the ability to gather data from a dizzying array of sources. Big Data analysis tools like Map Reduce over Hadoop and HDFS, promises to help organizations better understand their customers and the marketplace, hopefully leading to better business decisions and competitive advantages [3].

Author Margaret Rouse [5] stated there is a need to maximize returns on BI investments and to overcome difficulties. Problems and new trends mentioned in this article and finding solutions by combination of advanced tools, techniques and methods would help readers in BI projects and implementations. BI vendors are struggling and doing continuous effort to bring technical capabilities and to provide complete out of the box solution with set of tools and techniques. In 2014, due to rapid change in BI maturity, BI teams are facing tough time to have infrastructure with less skilled resources. Consolidation and convergence is going on, market is coming up with wide range of new technologies. Still the ground is immature and in a state of rapid evolution.

The authors Tekiner F. and Keane J.A [8] have given some important emerging framework model design for Big Data Analytics and a 3-tier architecture model for Big Data in Data Mining. In the proposed 3-tier architecture model is more scalable in working with different environment and also benefits to overcome with the main issue in Big Data Analytics for storing, Analyzing, and visualization. The framework model given for Hadoop HDFS distributed data storage, real-time Nosql databases, and MapReduce distributed data processing over a cluster of commodity servers.

Big data framework needs to consider complex relationships between samples, models and data sources along with their evolving changes with time and other possible factors. To support Big data mining high performance computing platforms are required.

With Big data technologies [3] we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time.

There are lots of scheduling technique are available to improve job performance but all the techniques have some limitation so any one technique cannot overcome that particular parameter in which they effecting the performance to whole system. Like data locality, fairness, load balance, straggler problem and deadline constrains. All the technique has advantages over any other technique so if we combined or interchange some technique then the result will be even much better than the individual scheduling technique [10].

The authors Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta and Kumar N [9] describes the concept of Big Data along with 3 Vs, Volume, Velocity and variety of Big Data. The paper also focuses on Big Data processing problems. These technical challenges must be addressed for efficient and fast processing of Big Data. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error -handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost -effective to address in the context of one domain alone. The paper describes Hadoop which is an open source software used for processing of Big Data.

The author Jimmy Lin [12] proposed a system based on implementation of Online Aggregation of Map Reduce in Hadoop for ancient big data processing. Traditional MapReduce implementations materialize the intermediate results of mappers and do not allow pipelining between the map and the reduce phases. This approach has the advantage of simple recovery in the case of failures, however, reducers cannot start executing tasks before all mappers have finished. As the Map Reduce Online is a model version of Hadoop Map Reduce, it supports Online Aggregation and stream processing, while also improving utilization and reducing response time.

The authors Aditya B. Patel, Manashvi Birla and Ushma Nair [11] stated learning from the application studies, we explore the design space for supporting data intensive and compute-intensive applications on large data-centre-scale computer systems. Traditional data processing and storage approaches are facing many challenges in meeting the continuously increasing computing demands of Big Data. This work focused on Map Reduce, one of the key enabling approaches for meeting Big Data demands by means of highly parallel processing on a large number of commodity nodes.

3. PROBLEM STATEMENT

The popularity of social media in today's world has led to a huge rise in the number of people using them. Those people create a huge amount of data called the Big Data which can be used for various useful purposes. The people who tend to exploit social media have risen at an alarming rate. They tend to misuse these platforms to threaten or harass people online. YouTube is the best example of a social media which creates a huge amount of big data every few minutes. The big data created can be utilized to analyse the trends for marketing companies to advertise their products as well as for YouTubers to upload popular type of content. Also the comments should be free of any sensitive content as the comments would be filtered for any sensitive content and would be published only after its removal.

4. EXISTING TECHNIQUE

In the existing technique, there have been systems to analyze the big data from YouTube for likes, types of comments and views for analyzing the popularity of different types of videos. This analyses the trends and finds which kind of content on YouTube is the most popular among the users. Although this analyses trends, there is no system available to filter out the sensitive content from the comments. There have also been sensitive content removal for social Medias like Twitter to remove sensitivity from tweets, but none have been developed for YouTube.

DRAWBACKS OF EXISTING TECHNIQUE

The existing technique uses Twitter API to remove the sensitive content from the tweets or posts but it does not provide clear cut results. It doesn't have any specific analysis for trending or sensitive content. It is restricted to analyze likes and views and doesn't analyze the trends to understand the popular type of content.

In our approach we focused more on the speed of performing data analysis than its approach i.e. performing data analysis on YouTube statistics analysis using Hadoop framework by splitting the various modules of data in following steps and collaborating with Map reduce programming. Delve into on Hadoop application development: HDFS (Hadoop Distributed File System) is the core component popularly known as the backbone of Apache Hadoop framework. HDFS is the one, which makes it possible to store different types of large data sets such as structured, unstructured and semi structured data. Hadoop Distributed File System has two core components, namely Data Node and Name Node. The Data Node stores actual data, whereas Name node contains metadata. Map Reduce is a programming model of Hadoop framework which helps in writing applications that processes large data sets using distributed and parallel algorithms inside Hadoop environment. In a Map Reduce program, Map () and Reduce () are two functions.

In order to deal with the sensitive data contents in an organization an effective data model using DLD (data leakage detection). This data model is considered for an organization to client communication for prevention of data leakage. But in this presented work this concept is extended for improving the data sensitivity in a public domain i.e. social media. The concept is to analyze the YouTube data and identify the possible sensitive contents form available post to be published. Therefore, first the data is pre-processed and then the NLP parser is applied to identify the NOUN and VERBSs in different posted strings. The noun content may be any person's name or the target and verbs are the action of hurt they intend to cause. We then analyze it to and compare it to the sensitive database we have to reduce the sensitivity of contents. This section provides the basic overview of the proposed sensitive data exposer technique in further the proposed system architecture is discussed

5. IMPLEMENTED SYSTEM

The implemented system does sensitive content filtering as well as popular content analysis. The system architecture explains how big data from YouTube is analyzed for sensitive content and for marketing purposes. The system is divided into 2 parts first is sensitive content filtering and second is analysis for marketing purposes. In the first part the comments are analyzed and filtered for harsh comments and then after filtering we get the sensitive words. In the second part the big data is analyzed for popular content and the trending content are reviewed and published.

6. SYSTEM ARCHITECTURE OF IMPLEMENTED SYSTEM

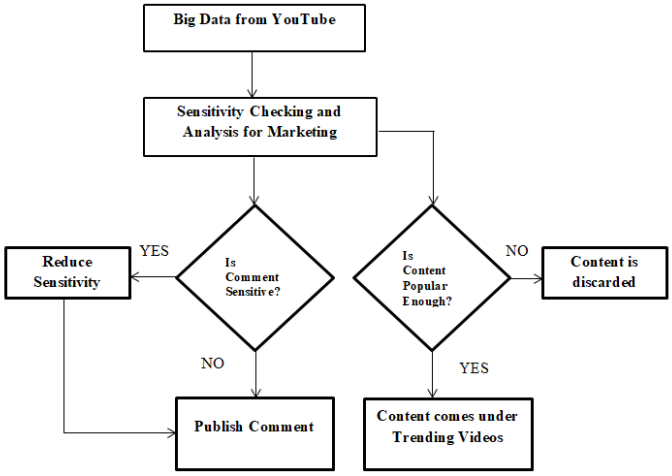

Fig. 1. System Architecture.

The above figure i.e Fig. 1 shows the system architecture of the implemented system.

7. RESULTS AND SNAPSHOTS

The following screenshots shows our results of the implemented system.

Fig. 2. Comment Fetching.

The comments are being fetched from YouTube API in this section for being analysed for sensitivity in them.

Fig. 3. Pre-processing of comments.

This removes the noise from the data that is the hashtags and other special characters from the comments so as to filter the comments in pure English sentence form. This preprocessed data is then sent to Hadoop for Processing and then the processed data is sent as output to the next step.

Fig. 4. Part Of Speech Tagger.

It is the part of speech tagger section. It divides the English sentence into parts of speech for example noun, pronoun, verbs etc out of which we only require the noun and verb because in most of the cases the sensitive words are found from the nouns and verbs.

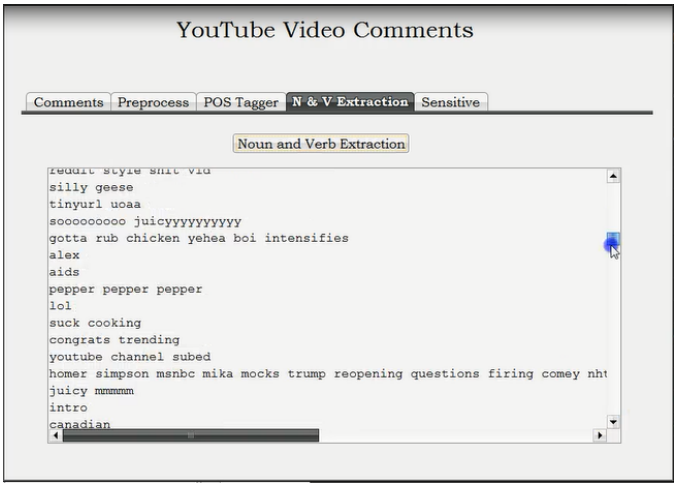

Fig. 5. Noun and Verb Extraction.

Here the nouns that is any person's or specific group's name which can be a target for threatening purposes and verbs which are the action or means of harm they intend to cause are extracted.

Fig. 6 Sensitive words.

The sensitive database is compared and sensitive words are found out in the last step.

The second part consists of analysis of the trending data sets.

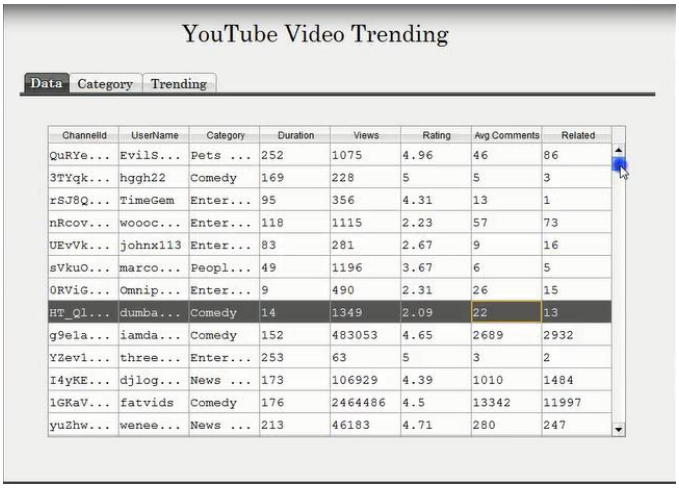

Fig. 7. Data for Trend Analysis.

Here the data for trend analysis are found and tabulated giving various names for various attributes of a particular video.

Fig. 8. Categories of Data.

Here the data are categorized according to the channel and type of content in their YouTube video that they publish.

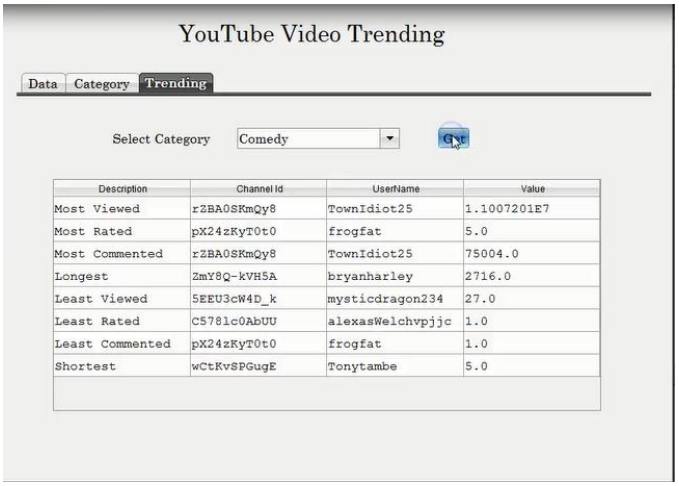

Fig. 9. Trends.

Now, here finally we can see which type of data do the users prefer watching according to the views, rating and no of comments on the video.

8. CONCLUSION

This project is analyzes the YouTube Big Data and has come up with a solution for sensitive content posted in the YouTube comments and analyze the trends to find the most popular content for both YouTubers to upload and marketing companies to publish advertisements. The aim is to implement an efficient technique for detection of sensitive text in the comments from YouTube videos along with identifying the most popular type of content. This is basically done in two modules: Collection and Handling of data. The data collected is scanned for sensitive content and if any found they are substituted. Hence this avoids the misuse of social media for threatening or illegitimate purposes. The data is also analyzed for the marketing companies to advertise their products and YouTubers to make such content which could earn them Ad Revenue.

9. FUTURE SCOPE

Future work may include analyzing the videos for sensitive content by looking into the language used in the videos, whether they are abusive or not. We can also have methods to improve the speed of analysis even better.

REFERENCES

[1] Yuri Demchenko ―The Big Data Architecture Framework (BDAF)‖ Outcome of the Brainstorming Session at the University of Amsterdam 17 July 2013.
[2] Amogh Pramod Kulkarni, Mahesh Khandewal, ―Survey on Hadoop and Introduction to YARN, International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014).
[3] Sagiroglu, S.Sinanc, D.,‖Big Data: A Review‖,2013, 20-24.
[4] Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule, ―Survey Paper On Big Data‖ International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014.
[5] Margaret Rouse, April 2010―unstructured data‖.
[6] Kyuseok Shim, MapReduce Algorithms for Big Data Analysis, DNIS 2013, LNCS 7813, pp. 44–48, 2013. [7] Dong, X.L.; Srivastava, D. Data Engineering (ICDE),‖ Big data integration― IEEE International Conference on , 29(2013) 1245–1248.
[8] Tekiner F. and Keane J.A., Systems, Man and Cybernetics (SMC), ―Big Data Framework‖ 2013 IEEE International Conference on 13–16 Oct. 2013, 1494–1499.
[9] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N ―Analysis of Big Data using Apache Hadoop and Map Reduce‖ Volume 4, Issue 5, May 2014‖.
[10] Suman Arora, Dr.Madhu Goel, ―Survey Paper on Scheduling in Hadoop‖ International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.
[11] Aditya B. Patel, Manashvi Birla and Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce," in Proc. 2012 Nirma University International Conference On Engineering.
[12] Jimmy Lin ―Map Reduce Is Good Enough?‖ The control project, IEEE Computer 32 (2013).