Public Opinion Analysis Using Hadoop

Pratik Balasaheb Jharikar
Department of Information Technology
MCT's Rajiv Gandhi Institute of Technology
Mumbai, Maharashtra, India.
pratikjharikar@gmail.com

Shalvi Naresh Raut
Department of Information Technology
MCT's Rajiv Gandhi Institute of Technology
Mumbai, Maharashtra, India.
shalviraut@gmail.com

Ashwini Ratan Patil
Department of Information Technology
MCT's Rajiv Gandhi Institute of Technology
Mumbai, Maharashtra, India.
ashwini1.patil1@gmail.com

Mukul Hemant Sakharkar
Department of Information Technology
MCT's Rajiv Gandhi Institute of Technology
Mumbai, Maharashtra, India.
sakharkar.mukul@gmail.com

Abhay Patil
Assistant Professor
Department of Information Technology
MCT's Rajiv Gandhi Institute of Technology
Mumbai, Maharashtra, India.
abhay.patil@mctrgit.ac.in

Abstract—Recent technological advances in devices, computing, and social networking have revolutionized the world but have also increased the amount of data produced by humans on a large scale. If you collect this data in the form of disks, it may fill an entire football field. According to studies, 2.5 billion gigabytes of new data is generated every day and 2.5 petabytes of data is collected every hour. This rate is still growing enormously. Though all this information produced is meaningful and can be useful when processed, it gets neglected. Social media has gained massive popularity nowadays. Twitter makes it easy to engage users in expressing, sharing and discussing hot latest topics but these public expressions and views are hard to analyze due to the bigger size of the data created by Twitter. In order to perform analysis and predictions over the hot topics in society, latest technologies are needed. The most popular solution for this is Hadoop. Hadoop acts as an open-source framework for developing and executing distributed applications that process very large amounts of data. It stores and process big data in a distributed fashion on large clusters of commodity hardware. The risk, of course, in running on commodity machines is how to handle failure. Hadoop is built with the assumption that hardware will fail and as such, it can easily handle most failures. Hadoop can be used for developing and executing distributed applications that process very large amounts of data. It provides a suitable environment needed for treating or processing huge data. Our job is to extract and store data into its file system and query the data according to the desired output.

We propose to perform analysis on Public opinion expressed over Twitter regarding the trending topics of the society by using Apache Hadoop framework along with its services Apache Flume and Apache Hive.

Keywords*- Analysis; Twitter; Hadoop; BIGDATA; Hive; Flume; Knime; AFINN; Structured; JSON; HQL*

*****

I. INTRODUCTION

We live in a society where the textual data on the Internet is increasing at a rapid pace and many companies are trying to use this huge amount of data to extract people's views towards their products. Online social network platforms, with their large-scale repositories of user-created content, provide unique opportunities to gain insights into the emotional "pulse of the nation", and for the global community. A source of unstructured text information is included in social networks, where it is not easy to manually analyze such amounts of data. There is a large number of social networking websites that enable users to contribute, modify and grade the content, as well as to express their personal opinions on specific topics. Some examples include blogs, forums, product review sites, and social networks, like Twitter (http://twitter.com/). Twitter (San Francisco, CA, USA) is a micro blogging site that offers the opportunity for the analysis of public opinion.

*Micro blogging and more particularly Twitter is preferred for the following reasons:

*Micro blogging platforms are used by different people to express their opinion or views about different topics, thus it is a valuable source of public opinion.

*Twitter contains a massive number of text posts and it grows every day. The collected corpus is mostly arbitrarily large.

*Twitter's audience can be regular users to celebrities, company representatives, politicians, and country presidents too. Thus, it is possible to collect text posts of users from different social and interests groups.

*Twitter have users from many countries around the world.
We are collecting this data by using BIGDATA online streaming Eco System Tool known as Apache Flume and also the shuffling and processing ofit into structured data can be done by using Apache Hive. We would visualize and analyze the retrieved data using KNIME Analytics Platform
.

II. PROBLEM STATEMENT

2.1 Existing System

The existing work on sentiment analysis can be classified from different points of views: technique used, view of the text, level of detail of text analysis, rating level, etc.
From a technical point of view, we identified machine learning, lexicon-based, statistical and rule-based approaches.

- The machine learning method uses several learning algorithms to findthe sentiment by training on a known dataset.
- The lexicon-based approach includes calculating sentiment polarity for a review using the semantic orientation of words or sentences in the review.
- The semantic orientation is a great measure of subjectivity and opinion in text.
- The rule-based approach finds opinion words in a text and then classifies itbased on the number of positive & negative words. It counts on different rules for classification such as negation words, booster words, idioms, emoticons, mixed opinions etc.

2.2 Proposed System

We are going to overcome few drawbacks from the existing systems by using Hadoop and its services. For getting raw data from Twitter, we are using Hadoop's online streaming tool Apache Flume. In this tool, we are going to perform some configurationswhich are needed to get the data from Twitter. We would also want to define what information or keywords we want to retrieve from Twitter. The retrieved data will be saved into HDFS (Hadoop Distributed File System) in JSON format. From this raw data we are going to create tables and filter the information that is needed for us using HIVE. We are going to perform the Data Analysis using some UDF's (User Defined Functions) along with AFINN dictionary. The following figure shows an architectural view for the proposed system. Thus, we propose our project Public Opinion Analysis using Hadoop. We use Apache Flume and Apache Hive on top of Hadoop for better analysis of trending topics in today's world.

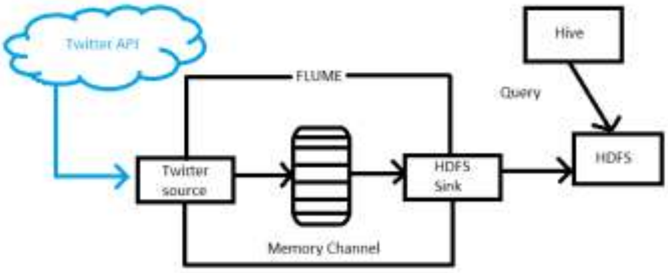


Fig. 1 Architectural Diagram for Proposed System.

III. METHODOLOGY

Our basic method includes the following three steps which would get us ready with data for further analysis:

- Creating Twitter Application.
- Retrieving data using Flume.
- Querying data using Hive Query Language (HQL)

3.1 Creating Twitter Application

First of all if we want to perform analysis on Twitter data, weneed to create an account as a Twitter developer and create an application on http://apps.twitter.com. After creating a new application, a unique access token along with consumer key gets generated. We will be having one consumer key to access the application for getting Twitter data. The following is the figure that shows how the application data looks after creating the application andthus we can see the consumer details and also the access token details. We would take these keys and token details and set in the Flume configuration file such that we can get the required data from the Twitter in the form of tweets.

Pflume_Twitter

Application Settings

[illegible]

Consumer Key (API Key) [redacted]

Consumer Secret (API Secret) [redacted]

Access Level [illegible]

Owner [illegible]

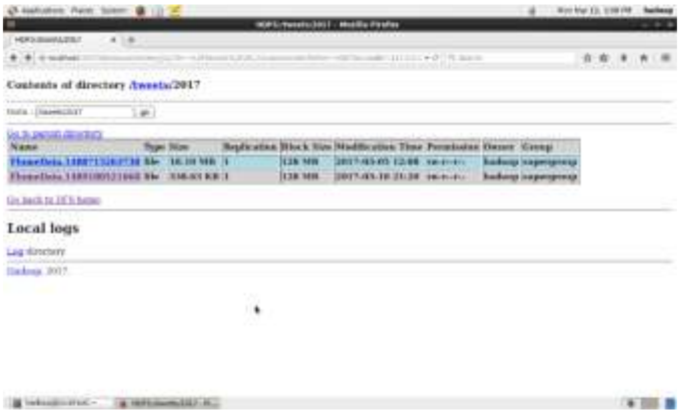
Owner ID [illegible]

Application Actions

Fig 2. Creating Twitter Application from Twitter Developer.

3.2 Retrieving data Using Flume

After creating an application on the Twitter developer site we would use the consumer key and secret key along with the access token and secret values and set them in the flume configuration file in conf folder of Flume Home. After everything is configured we would start the flume-ng and we can access Twitter and can get the live streaming data about the keywords we want. Here we will get every data in JSON format and this is stored in the HDFSprovided we give the location where to save the data in HDFS. The following is the content of HDFS directorywhere Flume Data is getting stored.



Contents of directory /tweets/2017

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
FlumeData.1488715263738	file	16.10 MB	1	128 MB	2017-03-05 12:08	rw-r--r--	hadoop	supergroup
FlumeData.[illegible]	file	336.83 KB	1	128 MB	2017-03-10 21:38	rw-r--r--	hadoop	supergroup

Local logs

Log directory

Hadoop, 2017

Fig 3. Retrieving data from Twitter into HDFS

File: /tweets/2017/FlumeData.1488715263738

Fig 4. Data is retrieved in JSON Format i.e. Semi-structured

3.3 Querying using Hive Query Language (HQL)

After running Flume, the Twitter data will automatically be saved into HDFS. The raw data is in the JSON format. From this data, first we would create a table where the semi structured data would be converted and stored in a structured format. For this we would use Hive SerDe. SerDe stands for serializerdeserializer which is inbuilt in Apache Hive. It is a simple way of converting the semi structured data to a structured data .This would give us optimum ways to perform queries on the data and retrieve desired results.

Fig 4. Convert Json format to Structured format using HQL

IV. Analysis of Sentiment on Public Topics

For analyzing the data in hive, we use a dictionary named as AFINN to sort the polarities of the words in tweets.

4.1 AFINN Dictionary

This dictionary proves to be a great tool in analyzing the data by ranking the words from positive to negative. This dictionary has more than 2400 words ranked from -5 to 5 according to their polarities where -5 stands for extremely worst and 5 stands for extremely Best.

Fig 5. Contents of AFINN Dictionary

4.2 Split words from tweets and assign Ratings

We would split the words in tweets and compare them with AFINN dictionary for performing a match. This would provide us with the exact sentiment a user has expressed.

Fig 6. Assigning rating to every splitted word

V. INTERACTIVE VISUALIZATION

KNIME Analytics Platform is a most known tool for its analyzing purposes. We can build a workflow which would provide us with detailed insight of where and how the data is analyzed.

We load the queried data in KNIME and perform a statistical analysis on it.

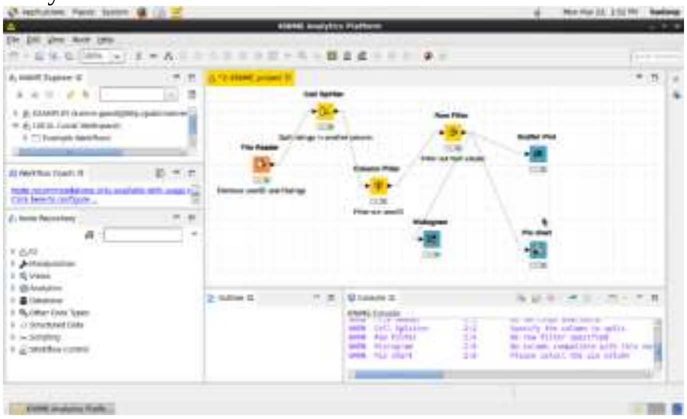


Fig 7. Workflow of KNIME

Visualization of the data can be performed using the above workflow. We provide a statistical representation of the public opinion in the form of

- Histogram
- Pie Chart
- Scatter Plot

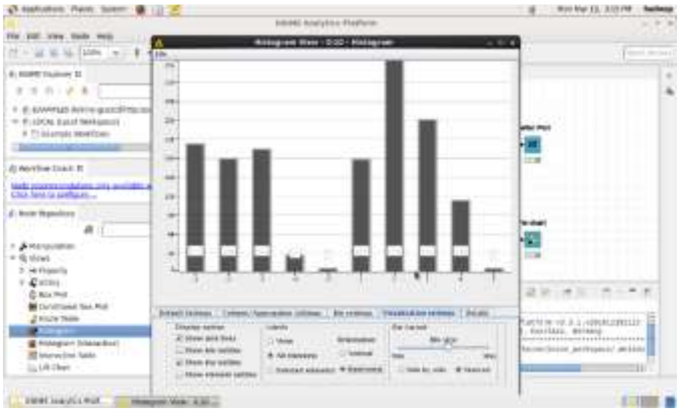


Fig 8. Analysis of Public opinion using Histogram

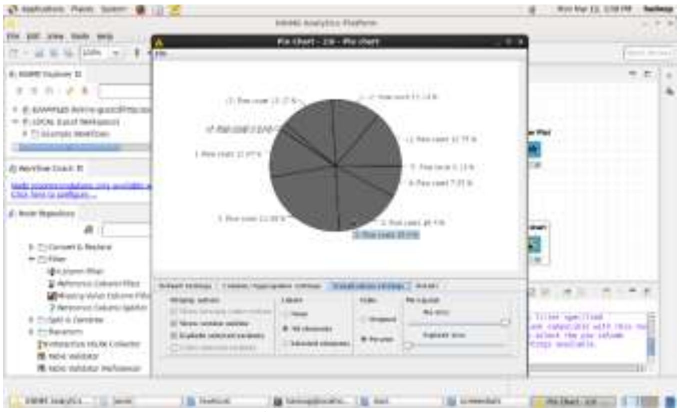


Fig 9. Analysis of Public Opinion using Pie Chart

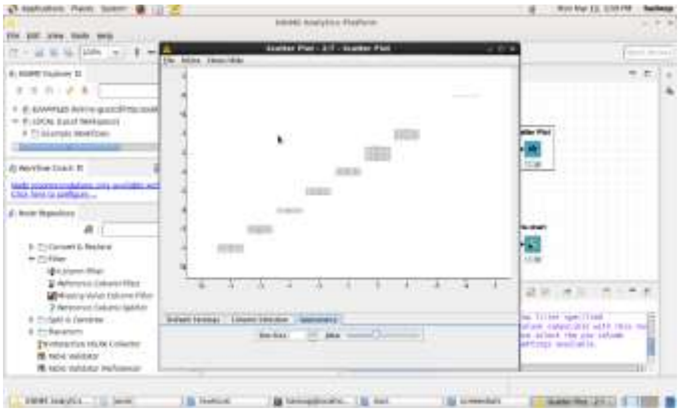


Fig 10. Analysis of Public Opinion using Scatter Plot

VI. RELATED WORK

There is a growing interest in visualizing sentiments from Web posts and related content. Chen *et al*. [1] presented a visualanalysis system using multiple coordinated views, such as decision trees and terminology variation, to help users to understandthe dynamics of conflicting opinions.Wanner *et al*. [2] described a concise visual encoding scheme to represent attributes, such as the emotional trend of each RSS news item. Both works for analyzing text contents are efficient by usingword matching methods. However, they lack semantic analysis. Draper*et al*. [3] developed an interactive visualization system toallow users to visually construct queries and view the results inreal time. For sentiment mining and analysis, Gregory *et al*. [4] proposed a user-directed sentiment analysis method to visualizeaffective document contents. Although they analyze and visualizeemotion, they only use statistical methods. To demonstrateand predict the trend for an event, we suggest that rules about theevolution of public sentiments related to the participants about hot topic types should be modeled and discovered. Collective behavior has many characteristics, such as being spontaneous, zealous, unconventional, and transient. Sentimental contagion and imitation are the main psychological mechanisms of the collective behaviors. Hoyst *et al*. [5] andSznajd-Weron *et al*. [6] proposed two different opinion dynamicsmodels using the aforementioned theory. For example, whendiscussing a debatable topic on forums, some participants' sentiments can easily be affected by others, which might result in booing or other extreme actions. In this study, we identify thechanging trend of an author's sentiment from his/her posts. Many have focused on social network visualization. Rios*et al*. [7] described how visualizations about the evolution of events on Twitter are created by presenting several case studies in recent years. Dork *et al*. [8] provided an interactive multifaceted visual overview of large-scale ongoing conversation son Twitter, including a spiral to present participants and theiractivities and an image cloud to encode the popularity of eventphotos by size.Wu *et al*. [9] presented an interactive visual system, Opinion Seer, to analyze the collection of online hotel customerreviews by augmenting scatterplots and radial visualization.

The opinion mining method can also be used in microblog sentiment mining. Baur *et al*. [10] provided an interactive visualization, LASTHISTORY, to display musical listening histories and context representing one’s past. However these methods are designed for hotel customers’ feedback and music listening histories, rather than for sentiment analysis. RadViz [11], [12] was often used to map data from an *n*-dimensional space onto a 2-D plane, to show these features in a multidimensional space. For a complex social network, these visualization models do not focus on sentiment analysis.

Several approaches focus on the visual exploration of blogs, forum posts, and Web logs. Adnan *et al*. [13] used frequent closed patterns to model and analyze data, and create a social network. They also analyzed Web logs by integrating data mining and social network techniques [14]. Indratmo*et al*. [15] visualized Web tags and comments arranged along a time axis. Dork *et al*. [16] provided faceted visualization widgets for visual query formulation according to time, place, and tags. Ong *et al*. [17] proposed an interactive Web-based tree map, News map, to represent the relative number of articles per news item. Fisher *et al*. [18] found the evolution of topical trends in social media by using line graphs indicating term trends. The aforementioned works focus on social networks, text analysis and knowledge representation of social networks to analyze microblogs and forum content without sentiment analysis. Rose *et al*. [19] represented the change of stories by clustering keywords into themes and tracking their temporal evolution. Neviarouskaya [20] presented SentiFul to automatically generate and score a new sentiment lexicon. Lin *et al*. [21] detected sentiments and topics simultaneously from text using the JST model that is based on LDA. Zhang *et al*. [22] analyzed the sentiment of restaurant reviews in Cantonese (an important dialect in some regions of Southern China) using classification. Zhang *et al*. [23] represented a sentiment analysis method on Chinese reviews. The visualization techniques exploring the idea of live-updating views include the encoding of data changes as animations [24], and representing changes in tag frequencies [25].

The aforementioned approaches mostly provide analysis and visualization on a single sentiment aspect. Sentiment analysis alone cannot discover the law of hot topics on microblogs and forum. We believe that sentiment analysis with different visualization perspectives would be more useful for users to find and understand sets of topics. Cao *et al*. [26] presented a method to show real-time information diffusion from Twitter using multiple viewing options. It traces information diffusion but ignores the relationships between different users and different hot topics, different timeframes and regions. While there have been promising advances on visualizing the development of topics over time, research in model-driven visualization of public sentiments remains an open area.

VII. APPLICATION AND SCOPE

1. Opinion Analysis: Public data is unstructured data that represents opinions, emotions, and attitudes contained in various sources such as social media posts, blogs, online product reviews, and customer support interactions. Different companies and Organizations use social media analysis to determine how the public feels about something at a particular moment in time, and also to track how these opinions change over time.
2. Text Analytics: It is the process of deriving the high quality information from the raw data such as unstructured or semi-structured data and predicting the analysis.

3. Volume Trending: The volume is estimated in terms of amount of data to process a job. Volume trending is a big issue in today’s world. Day by day it is increasing in a much higher rate in the organizations and social media sites etc.

4. Predictive Analytics: Predictive analysis gives the predictive scores to the organizations to help them in making the smart decisions and improve business solutions. It optimizes marketing campaigns and website behavior to increase customer response in business, conversions and meetings. Each customer's predictive score determines actions to be taken with that customer.

5. Social Media Data: With Hadoop, we can mine Twitter, Facebook and other social media sites for sentiment data about people and use it to make targeted, real time decisions that increase market share.

6. Web Click stream data: Hadoop makes it easy to track customers and their activities in different issues like product purchasing and viewing etc. It lets analysts know the behavior and interest of the customers and can able to visualize similar type of products to the customers.

Scope: The Proposed system can determine the most popular information about the people, organizations and can be used in the field of analytics.

Applications:

1. Determines the most number of sentiments in the social networking sites.

2. The system can be useful to track the business analysis of the organizations.

3. Allows analysts to retrieve and analyze the data easily from large datasets.

VIII. CONCLUSION

Big data is mainly a collection of data sets which are very large and complex in nature. It is very difficult to handle such data using on-hand database management tools. The main challenges with big databases are creation, duration, storage, sharing, search, analysis and visualization. So to manage these databases we need, “highly parallel software’s”. Data is retrieved from different sources such as social media, traditional enterprise data or sensor data etc. Flume can be used to retrieve data from social media network. Then, this data can be organized and stored using distributed file systems such as Google File System or Hadoop File System. These file systems are very efficient and optimal when number of reads are very high as compared to writes. At last, data can be processed & analyzed using map reducer with various Hadoop services and tools. In this project, we have queried the data acquired from twitter and have performed successful Public Opinion Analysis using Hadoop with the help of Apache Flume, Apache Hive and KNIME Analytics Platform.

ACKNOWLEDGMENT

We would like to express our special thanks of gratitude to our Guide Prof. A. E. Patil, as well as our Head of DepartmentProf. Sunil Wankhade, who gave us the golden opportunity to do this wonderful project on the topic Public Opinion Analysis using Hadoop, which also helped usin doing a lot of Research and wecame to know about so many new concepts for which we are really thankful to them.

REFERENCES

[1] Apoorv Agarwal, Jasneet Singh Sabarwal, “End to End Sentiment Analysis of Twitter Data”

[2] Real Time Sentiment Analysis of Twitter Data Using Hadoop Sunil B. Mane, YashwantSawant, SaifKazi, VaibhavShinde

[3] Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data.

[4] Semantic Sentiment Analysis of Twitter Hassan Saif, Yulan He and Harith Alani Knowledge Media Institute, The Open University, United Kingdom.

[5] Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good the bad and the omg! In: Proceedings of the ICWSM (2011).

[6] Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford (2009).