User Analytics-Based System Using Social Media Stream Data

Sarika N. Raut

Abstract:
Twitter, is one of the largest social media site that receives tweets in millions of data in each and every day in range of Zettabyte per year. This huge amount of raw data can be used for industrial or business. It provides a way of analyzing of big data such as twitter data using Apache Hadoop. It will analyze the tweets on the Hadoop clusters. In this paper, I am going to talk how effectively sentiment analysis is done on the tweets which is collected from the Twitter using tweepry. Twitter is an online web application which contains large amount of data that can be a structured, semi-structured and unstructured data. Productive E-commerce web-sites, Facebook, Twitter are larger social media sites. That sites are receives millions of comments, tweets or customer reviews every day in the range of terabyte or petabytes per day. Ideas and opinions of people are influenced by the opinions of other people. A Lot of research is done and going on the analysis of reviews given by people. I can collect the data from the social media site by using BIGDATA and Hadoop eco-system using online streaming tool Tweepy. Here I used specific algorithm for sentiment analysis on mapreduce. I have categorized this sentiment analysis into 3 groups like comments that are having positive, neutral and negative comments. Twitter is also difficult due to language that is used for comments.

Keywords: Hadoop, Big Data, Map Reduce, HDFS, Sentimental Analysis, Flume, tweets, twitter.

INTRODUCTION
Now a day’s BIG data is a name used everywhere in distributed paradigm on web. BIG data is the collection of sets of very large amount of data in above terabytes. The upcoming of online social media and mobile communication technologies has triggered a fastly increased in the flow of user generated content of various forms. People are express their reactions, fancies and emotions through social media by textual fragment of epigrammatic nature rather than writing long text. We made friends through the traditional fashion as G-friends, which stand for geographical location-based friends.

In social media, most of people adds their daily activities, habits, comments etc. on that data (habits, activities, comments) we found that who is most matched to another person... we analyze that person to user by calculating polarity.

One challenge in existing social networking services that large data is stored in hdfs means it is more scalable... Millions of tweets are generated in each and every day on multifarious issues. We propose an unsupervised and domain-independent approach by using the polarity scores on three lexical resources-SentiWordNet 3.0, SenticNet 2 and SentiislangNet. SentiWordNet contains polarity scores of unigrams. for that we express positive or negative opinion.

PROBLEM STATEMENT
The problem in sentiment analysis is classifying the polarity of a given text at the document, sentence or feature/aspect level. Whether the expressed opinion in the document, in the sentence or in an entity feature/aspect is positive, negative or neutral. Given a message, classify whether the message is the positive, or negative, or neutral sentiment. For messages conveying both the positive and negative sentiment, which is the strongest sentiment should be chosen.

SYSTEM ARCHITECTURE
In our Analytics paper evaluate existing system extends on large-scale Experiments. We proposed evaluate Friendbook system on large scale experiments using social media data. Recently analysis is worked for few Twitter tweets analysis but here we are to doing work for user behavior analysis using Facebook Comments or any social media data reviews, user likes, interests So here we are going to use Hadoop and its Ecosystems, for getting raw data from the Facebook or any social media sites we are using Hadoop online streaming tool using Apache Flume [6]. Using Flume tool only we configure everything that we want to get data from the Facebook.[7] For analysis we want to set the configuration and also want to define what information that we want to get form Facebook All these will be saved into our HDFS (Hadoop Distributed File System)[8] in our prescribed format. In this proposed system, a method to calculate Analysis of reviews or comments given by the customers or user is proposed and implemented in Java on Hadoop. The method works in two phases: Mapper phase and Reducer phase. We are use a positive and negative word dictionary to identify positive and negative words [9] [10]. Stop word dictionary is used to identify and remove stop words from the reviewed product [11]. The focus of our project is to assign the polarity to each comment i.e. whether the author express positive or negative opinion.[12]

Map-Reduce
Map Reduce is the software programming framework in the Hadoop stack that simplifies processing of big data sets. A Map Reduce job consists of at least a map function and a reduce function, called Mapper and reducer respectively. The Mapper takes as input a pair of key-value and it produces a set of key-value pairs. Each reducer receives a key and a set of values which have the same key. This makes Map Reduce an excellent tool for computations that need Sorting or counting. The map and reduce are the functions which are left to the user for implement their desired functionalities to process each key-value pair. Hadoop Map-Reduce (Hadoop Map/Reduce) is the software framework for distributed processing of big data sets on compute clusters of commodity hardware. According to The Apache Software Foundation, the primary objective of Map-Reduce function is to split the input
data set into independent chunks that are processed in a completely parallel manner. The Hadoop Map Reduce framework sorts the outputs of the maps, which are then input to the reduce tasks.

![System Architecture](image)

**Fig 1. System architecture**

The above figure shows clearly that architecture view for the proposed system by this we can understand how our project is effectively using the Hadoop ecosystems and how the data is going to store from the Tweepy, and also how it is going to create tables using Twitter Dumper. Also how the analysis is going to perform.

First, the raw data of Twitter accessed by Twitter dumper and stored in hdfs. Secondly, Hdfs can give the data that is stored in it to Hadoop Framework i.e. the map reduce, hive, pig. Hadoop Frameworks are analyze twitter data using the specific algorithms. That specific algorithms are the first is Maximum entropy Algorithm and second is naive bayes algorithm for sentiment analysis. After that result of analyzed Twitter data stores in hdfs, and show result in the form of percentage and graph.

**RESULTS**

*In Percentage*

![Naive Bayes Classifier Result](image)

**In Graph**

![Naive Bayes Classifier Result Plot](image)

**CONCLUSION**

Data volume increasing rapidly nowadays, it is required to process on data speedily. Analytics is used to tell the writers/producers sentiment about their thoughts. There are several ways to define and analyze the social media data such as facebook, Twitter etc. Here anyone can perform different operations queries in these type of data. But the problem is arisen when dealing with BIGDATA. In this paper I have try to execute problem statement and solving it in the BIGDATA with the help of Hadoop and its Eco-systems. And finally I will try to done User analytics based on user Twitter comments or reviews, likes, interests. Here it is solving by using Hadoop and its packages. And we have trying to done some User analysis on their comments and the most number of comment ids.

**FUTURE WORK**

Using this application or project we can recommended friend to user. And also recommended product to user....

**REFERENCES**

1. Zhibo Wang, Jilong Liao, Qing Cao, Hairong Qi, and Zhi Wang, Member, IEEE,” Friendbook: A Semantic-based Friend Recommendation System for Social Networks” IEEE TRANSACTIONS ON MOBILE COMPUTING (Volume:14, Issue:3)


8. Bo Pang, Lillian Lee, Opinion Mining and Sentiment Analysis.


