IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

Interpreting the Public Sentiment Variations on Twitter using a open Source Framework

Adithya Aithal[1], Varun Pai M[2], Sriram K L[3], Ramesh[4], Sandeep B L[5]

UG Student, Department of Information Science and Engineering, M S Ramaiah Institute of Technology, Bangalore, India[1,2,3,4]

Assistant Professor, Department of Information Science and Engineering, M S Ramaiah Institute of Technology, Bangalore, India[5]

Abstract: Social Networking sites provide tremendous impetus for Big Data in mining people's opinion. Public API's catered by sites such as Twitter provides us with useful data for perusing writer's attitude with reference to a particular topic, product etc. To discern people's opinion, tweets are tagged into positive, negative or neutral indicators. This project provides an effective mechanism to perform opinion mining by designing a end to end pipeline with the help of Apache Flume, Apache HDFS, Apache Oozie and Apache Hive. To make this process near real time we study the workaround of ignoring Flume tmp files and removing default wait condition from Oozie job configuration. The underlying architecture employed here is not restricted only to opinion mining but also has a gamut of applications. This paper explores few of the use cases that can be developed into actual working models.

Keywords: Opinion mining, Big data, Hadoop.

I. INTRODUCTION

1.1 Motivation

Twitter information estimation analysis can be a magnificent wellspring of data and can give bits of knowledge that can:

1) Determine marketing strategy
2) Improve campaign success
3) Improve product messaging
4) Improve customer service
5) Generate leads

Be that as it may, on the off chance that you are settling on choices utilizing incorrect sentiment analysis data, the results can be calamitous. The investigation of assessment examination, if done legitimately, is extraordinarily unpredictable and is really a field of study, not only an element in an online networking apparatus. The multifaceted nature when you understand the quantity of times individuals misconstrue discussions or miscomprehend the composed word. Components, for example, mockery make it considerably more hard to decide importance from words.

1.2 Scope

The recent emerging area of interest is sentiment analysis of social issues. Presently a day the majority of the exploration researchers have been dealing with Twitter and YouTube remarks information set. To perform sentiment analysis the most and regular wellspring of information set are pages, social site like face book, twitter, YouTube and so on. There is an unfathomable extension for analysis researchers to build the precision level up to some degree by utilizing all around planned sentence structure. As of now, the code can deal with the investigation part with a decent exactness. Be that as it may, there are a couple of zones which have a great deal of degree in this angle. Mocking remarks are the ones which are to a great degree difficult to recognize. Tweets containing mocking remarks give decisively opposite results inferable from the state of mind of the creator. These are practically difficult to track. Additionally relying upon the context in which a word is utilized, the understanding changes. For instance, the word "unpredictable" in "unpredictable plot" in connection of an area plot is negative though "unpredictable plot" in setting of a movie's plot is sure. So it's important to relate the understanding with the connection of the tweets**.**

1.3 Objectives

- To implement an algorithm for automatic classification of text into positive, negative or neutral sets.
- To extract the meaning of an input text or tweet using natural language processing.
- To determine the attitude of the mass into various objective sets towards the subject of interest.
- To improve the accuracy of the analysis using our algorithm.
- To Obtain the Graphical representation of the sentiment in form of Pie-Chart.

II. SYSTEM ANALYSIS

Social networks have changed the path in which individuals communicate. Information available from social networks is beneficial for analysis of user opinion. Sentiment analysis is a generally new region, which

DOI 10.17148/IJARCCE.2016.5471

manages separating client supposition consequently. There are numerous routes in which social network information can be utilized to give a superior comprehension of client assessment such issues are at the heart of natural language processing (NLP) and data mining research.

We are going to get the data through twitter API which is set up by means of the streaming API. Once the tweets are gathered it will be encode in JSON.

We utilize Flume to mechanize development of the tweets from API to HDFS. Tweets are gathered and collected into units of data. Utilizing memory channel, tweets are prepared in clumps. We utilize HDFS-sink to store tweets into HDFS.

The master server architecture of HDFS having single name node helps in directing the document framework access. Data is put away as pieces in the HDFS.

Gigantic measures of tweets are gathered, put away and dissected. We utilize Map Reduce to isolate the colossal measures of information. The Job Tracker of the Map Reduce initiates map and reduce jobs. The tweets are queried using Hive’s Execution Engine.

As per our outline appeared in figure 1.1, we assemble the twitter data utilizing flume. The data gathered is put away in hadoop distributed file system.

Using our algorithm we analyze the data and forward it to map reduce technique. In the map reduce phase the tweets are mapped into objective sets of data and reduced into final result. We have arrived at the following design for its simple architecture and efficient features.

III. MODULES

1. Extracting Tweets from twitter based on the query keywords. In this module based on the twitter access key and consumer key, we are going to extract tweets based on the query keyword.

2. Storing Tweets to Hadoop. In this module extracted tweets are stored into data nodes of hadoop.

3. Based on the Hash tag Retrieve data. In this module whatever the data stored in the hadoop we are going to retrieve back in the format of j son data, based on the hash tag we will get tweets and finally store it into database.

4. Preprocessing of Tweet. Remove unnecessary words. Remove hyperlink. Remove special characters. Get filtered data.

5. Sentiment Process. In this module we have keep first initial positive, negative and neutral words ,based on this initial expansion we are going to give positive, negative and neutral count for the words in the tweet and finally we will detect the sentiment of the tweet.

IV. DESIGN

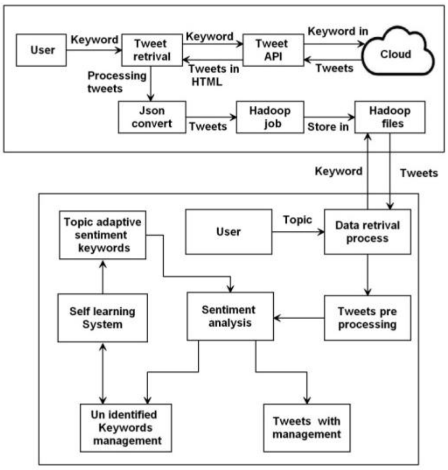

Fig 1.1 Design of the project

The important functions of the given design are as follows:

- Data from Twitter
- Data Collection with Apache Flume
- Hadoop Distributed File System
- Map Reduce

1. Data From Twitter

Twitter gives us a Streaming API which will be utilized to get a consistent stream of tweets empowering us to gather and examine user opinion. As per the fig 2.1, the Streaming API works by making a request for a particular sort of data which is filtered by keyword, a user, geographic territory and so on. When association with the Twitter API is set up by means of the Streaming API, data accumulation happens. The tweets gathered will be encoded in JavaScript Object Notation (JSON). JSON gives us an approach to encode this data. The entire tweet is viewed as a word reference comprising of different fields. The fields might be contributors (indicates users who have authored the tweet), coordinates (Represents the geographic location of the Tweet as reported by client application), favorite count (No. of times the tweet has been “favorite”), text (actual text of the tweet) and several other fields..

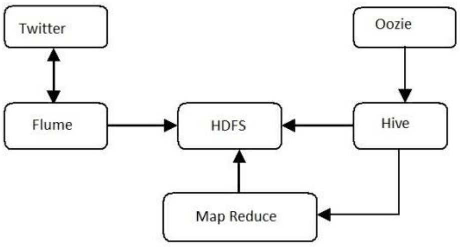

Fig 2.1 Design of the project

2. Data Collection with Apache Flume

To computerize the development of tweets from the API to HDFS, without our manual intercession, Flume is utilized. Apache Flume is a reliable and distributed framework for adequately assembling and moving a lot of data from different sources to a common storage area. Real parts of flume are source, memory channel and the sink which is

DOI 10.17148/IJARCCE.2016.5471

appeared in the figure. Twitter source is an event-driven source that uses Twitter library for getting to streaming API. As per the fig 2.2 Tweets are gathered and collected into key units of data called as an event. An event fuses a byte payload and a optional header. The coordination of event flows from the streaming API to HDFS is undertaken by Agent. The procured tweets are put away into one or more memory channels. A memory channel is an impermanent storage that uses an in-memory line to hold event until they are ingested by the sink. Utilizing memory channel, tweets are prepared in groups that can be designed to hold a consistent number of tweets. To secure tweets for a given keyword channel filter is utilized. Sink composes event to a preconfigured area. This framework makes utilization of the HDFS-sink that stores tweets into HDFS.

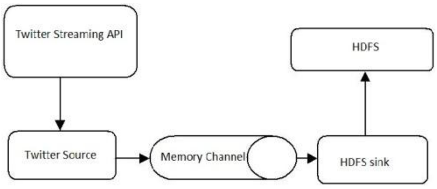

Fig 2.3 Design of the project

3. Hadoop Distributed File System

(HDFS) is a distributed file system which rests on top of the native file system and is written in java. It is highly fault tolerant and is designed for commodity hardware. HDFS has a high throughput access to application and is suitable for applications with large amount of data. In the figure we see the master-server relation or architecture. The master-server architecture of HDFS having single name node helps in regulating the file system access.

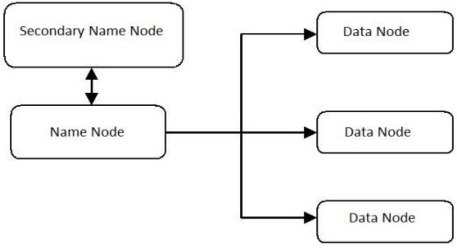

Fig 2.2 Design of the project

4. MapReduce

Map Reduce is one of the two major components of Hadoop. It is a programming paradigm intended to help processing which includes data intensive operations. As per the fig 2.3MapReduce includes two distinct jobs that Hadoop programs perform. Map Reduce jobs are controlled by a software daemon known as the Job Tracker. The Job Tracker dwells on an master node. The Job Tacker starts the map() and reduce() jobs in the data nodes where the Task Tracker daemon resides. Map Reduce requires a Java programmer. Other than extremely fundamental applications, Map Reduce requires different stages prompting bulky codes. It's clients need to rethink inbuilt functionalities, for example, joining and questioning which are given by Hive as inbuilt capacities. The tweets are queried utilizing Hive whose Execution engine as a part of turn creates Map and Reduce jobs so as to inquiry out the parts of the tweets which the user is interested in.

V. SCOPE FOR FUTURE WORK

A. Deployment in Cloud: It is an arduous task for average data sets. Hence there is a requirement of powerful machines which are made available through cloud platforms (IAAS or VPC).Deploying the whole system in cloud provides hassle-free access to it.

B. Secure e-Voting System: A secure mechanism is required for expressing people's consensus on policy initiatives and electoral procedures in countries that follow direct democracy. Direct democracy involves people's opinion in government decision making by conducting regular referendums involving people casting their votes in polling booths. A secure model having similar architecture can provide a mechanism to cast people's vote from their home.

VI. CONCLUSION

Opinion Mining is a very wide branch for research. Have covered some of its important aspects. The same architecture could be used for a variety of applications designed to look at Twitter data, such as identifying spam accounts, or identifying clusters of keywords. Taking the system even further, the general architecture can also be expanded to other social media platform usages like Facebook, movie reviews, personal blogs, etc. Evidently, taking into account all the constraints, this method is one of the most efficient ways to perform opinion mining in real-time.

REFERENCES

[1]. Alec Go, RichaBhayani, Lie Huang. "Twitter Sentiment Classification Using Distant Supervision". CS224N Project report, Stanford, Google scholar, Page1-12, 2009.

[2]. Pak, A., and Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In Proc. of LREC.

[3]. Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In Proc. of Coling.

[4]. Sunil B. Mane, YashwantSawant, SaifKazi, VaibhavShinde, " Real Time Sentiment Analysis of Twitter Data Using Hadoop", International Journal ofComputer Science and Information Technologies, (30983100),Vol.5(3),2014.

[5]. Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of Coling.

[6]. ApoorvAgarwal, FadiBiadsy, and Kathleen Mckeown 2009. Contextual phrase-level polarity analysis usinglexical affect scoring and syntactic n-grams. Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 24–32, March.

[7]. James Spencer and Gulden Uchyigit- Sentimentor: Sentiment Analysis of Twitter Data School of Computing, Engineering and Mathematics,University of Brighton,Brighton, BN2 4GJ

[8]. M.Vasuki, J.Arthi, K.Kayalvizhi- Decision Making Using Sentiment Analysis from Twitter,International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization), Vol. 2, Issue 12, December 2014

[9]. EfthymiosKouloumpis, TheresaWilson, Johanna Moore. "Twitter Sentiment Analysis:The Good the Bad and the OMG!", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media,2010.

DOI 10.17148/IJARCCE.2016.5471