ISSN 2395-1621

Facebook Post Classification on Hadoop

[#1]Sandeep Kokare,[#2]Pooja Chavre,[#3]Akshay Khobare ,[#4]Yogesh Mundhe

[1]sk4196485@gmail.com
[2]pooja.chavre@gmail.com
[3]akshaykhobare12@gmail.com
[4]yrmundhe540@gmail.com

[#1234]Computer Engineering Department SKNCOE , Pune

ABSTRACT

aWhile recent NLP(natural language processing)-based sentiment analysis has centered around Facebook and product/service reviews, we believe it is possible to more accurately classify the emotion in Facebook status messages due to their nature. Facebook status messages are more succinct than reviews, and are easier to classify because their ability to contain more characters allows for better writing and a more accurate portrayal of emotions. We are using the Facebook API itself to fetch the data from Facebook server and that data will be stored on Hadoop. And to analyze the sentiments of it we are using NLP . We classify both binary and multi-class sentiment labeling. The processing will be done on Hadoop server

Keywords— Hadoop cluster. Cloud service. Social media , Sentiment Analysis

ARTICLE INFO

Article History

Received :13th February 2016

Received in revised form :

14th February 2016

Accepted :16th February , 2016

Published online :

18 th February 2016

I. INTRODUCTION

To develop a system that classifies fb post on Hadoop (installed on cloud) with the help of NLP. Extract texts from posts, images, videos of social media as Facebook to know how people feel about different posts and do sentiment analysis, using Hadoop Many businesses large and small use cloud computing today either directly (e.g. Google or Amazon) or indirectly (e.g. facebook) instead of traditional on-site alternatives. There are a number of reasons why cloud computing is so widely used among businesses today. So using the hadoop on cloud with the help of feacebook api , data is fetch from the facebook and allowed to analysis the text post with the help of sentiment analysis i.e natural language processing (NLP) .Reduction of costs, Universal access, Choice of applications, Potential to be greener and more economical Flexibility are some of the reasons why to use cloud. Whereas,. It's a lot of data produced very quickly in many different forms.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

The benefits of this cloud based big data analytical service are user friendliness & cost as it is developed using open-source software. The system is cloud based so users have their own space in cloud where user can store there data. User can browse data, files, folders using browser and arrange datasets. User can select dataset and analyze required dataset and store result back to cloud storage.

This paper describes a Sentiment Analysis study performed on over than 1000 Facebook posts about newscasts. This study takes also in account the data provided by Audited regarding newscast audience, correlating the analysis of Social Media, of Facebook in particular, with measurable data, available to public domain.

II.MATERIAL & METHODOLGY

Login will be done to Facebook account, and request API for posts to fetch. This will give a pop up to user asking if to allow to fetch posts or not. If allowed, the user account from which login is done, all posts from that user will be fetch along with public posts of his friends. And then these posts will be dumped into Hadoop.

Posts will be classified and categorized into sentiments required, through other API. Report will be made in terms of graph using R- language. Giving overall post sentiment analysis. Report generation will be done on HadoopMethod used in the projects is sentiment analysis algorithm and k – means, natural language processing is a field of computer science, artificial intelligence and computational linguistics concern with the interactions between computers and human (natural) languages.

HADOOP

Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. Open-source software. Open-source software is created and maintained by a network of developers from around the globe. It's free to download, use and contribute to, though more and more commercial versions of Hadoop are becoming available. Framework. In this case, it means that everything you need to develop and run software applications is provided – programs, connections, etc. Massive storage. The Hadoop framework breaks big data into blocks, which are stored on clusters of commodity hardware. Processing power. Hadoop concurrently processes large amounts of data using multiple low-cost computers for fast results.

One of the top reasons that organizations turn to Hadoop is its ability to store and process huge amounts of data – any kind of data – quickly. With data volumes and varieties constantly increasing, especially from social media and the Internet of Things, that's a key consideration. Other benefits include: Computing power, Flexibility. Fault tolerance. Low cost, Scalability.

CLOUD

The term Cloud refers to a Network or Internet. In other words, we can say that Cloud is something, which is present at remote location. Cloud can provide services over network, i.e., on public networks or on private networks, i.e., WAN, LAN or VPN .

Software as a Service (SaaS) model allows to provide software application as a service to the end users. It refers to a software that is deployed on a hosted service and is accessible via Internet. There are several SaaS applications, some of them are listed below: Billing and Invoicing System, Customer Relationship Management (CRM) applications, Help Desk Applications Human Resource (HR) Solutions

Some of the SaaS applications are not customizable such as an Office Suite. But SaaS provides us Application Programming Interface (API), which allows the developer to develop a customized application.

Using SaaS has proved to be beneficial in terms of scalability, efficiency, performance and much more. Some of the benefits are

Modest Software Tools, Efficient use of Software Licenses, Centralized Management & Data, Platform responsibilities managed by provider, Multitenant solutions

III.NATURAL LANGUAGE PROCESSING (NLP)

SENTIMENT ANALYSIS

Extract subjective information usually from a set of documents, often using online reviews to determine "polarity" about specific objects. It is especially useful for identifying trends of public opinion in the social media, for the purpose of marketing.

It is basically used for analysing the human text written on social media for comparing the human behaviour.

The text-processing.com API is a simple JSON over HTTP web service for *text mining* and *natural language processing*. It is currently free and open for public use without authentication, though that may change in the future.

In this Facebook post classification the data which is fetch with the help of Facebook api , for classify the data NLP api is being used .NLP api which is used will do the further step to classify the data .

SentiWordNet (http://sentiwordnet.isti.cnr.it/) is an excellent publicly available lexicon. Technically, the resource contains Princeton WordNet data marked with polarity scores.

TOKENIZATION

Transforming a stream of characters into a stream of processing units called tokens. In this project the data which is fetch from the Facebook and the data is divided and the every block of memory and each data block will have a token and process according to the token data is processed

Stop Words Filtering: Consists in eliminating stop-words.

In this token is used to separate out the words which will affect the sentence .Dividing the full text data and then analysing is simple as compared to classify full data .

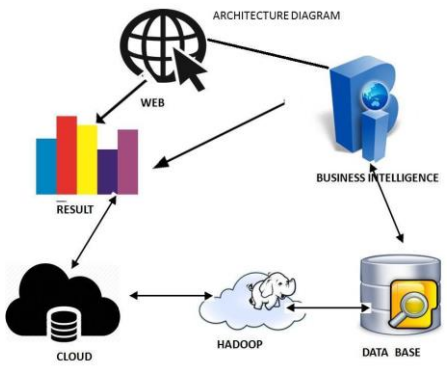
.

STEMMING

The process of reducing each word (i.e., token) to its stem or root form, by removing its suffix. Stem Filtering: Consists in reducing the number of stems of each SUM. Featuring Representation used for representing. After removing the stop words, the main words which actual have a meaning of the sentence which effect the sentence is calculated with the help of sentiment wordlist in which the each and every meaning of word is stored and the priority in integer form. And at last it just get added to check whether the sentence is positive or negative.

To stem text, do an HTTP POST to http://text-processing.com/api/stem/ with form encoded data contain the text you want to stem. You'll get back a JSON object response whose text attribute contains the stemmed text.

IV.ARCHITECTURE



The architecture diagram shows the components required for the working. Whole database required for analysis is present on cloud. The diagram shows that Hadoop is installed on cloud. The business intelligence block contains all the algorithms required for the analysis. Facebook is linked to the Hadoop, where only public data can be accessed by using Application interface (API). The result is displayed in the form of graphs and percentage

V.RESULT

The algorithm is made with the help of some concept and method and the flow is For the Facebook post classification. Initialization is done by face book page using Facebook API, it will send an acknowledgement to use Facebook. The Facebook data will be stored in a variable in JSON format.
This file variable will be processed on Hadoop which is deployed on cloud and the whole process is done on cloud
Sentiment analysis is carried forward in which NLP used. In which classification of data is done using tokenization, stemming and it will send result whether the post is positive or negative.

VI.CONCLUSION

The Hadoop system is a challenging environment to work with, but cloud deployments introduce additional levels of complexity because of the constraints (and freedoms) offered by the cloud environment. In this system we are basically public cloud in order to make available for all the customer without spending any cost. As well as we are monitoring the social media activities is a good way to measure customers' loyalty and interests, keeping track of their sentiment towards brands or products. It maps Sentiment Analysis on Social Media with observations and measurable data. The system is performing sentiment analysis in order to get to know about the reviews of individual regarding the particular product. Making availability of public cloud makes the enterprise survey about the product with less efforts and make best product available.

REFRENCES

[1]. Roger S. Barga, Jaliya Ekanayake, Wei Lu, "Project Daytona: Data Analytics as a Cloud Service", IEEE 28th International Conference on Data Engineering, 2012.

[2]. Nikolay Laptev, Kai Zeng, Carlo Zaniolo., "Very Fast Estimation for Result and Accuracy for Big Data Analytics: the EARL System", IEEE's, ICDE Conference 2013.

[3]. Vignesh Prajapati, "Big Data Analytics with R and Hadoop", Packt publication, 2013.

[4] Q. Ethan McCallum & Stephen Weston, "Parallel R", O'Reilly, 2012. [10] Josep Adler, "R IN A NUTSHELL", second edition, O'Reilly, 2012.

[5] National Institute of Standards and Technology, "The NIST Definition of Cloud Computing," Special Publication 800-145, Sep. 2011.

[6] Hadoop on OpenStackhttp://hortonworks.com/labs/openstack. (2014,July. 18).

[7] Apache Hadoop. http://hadoop.apache.org. (2014, July. 16). [8]. http://hortonworks.com/labs/openstack. (2014,July. 18).

