iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VI **Month of publication:** June 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Twitter Sentiment Analysis

Deepti S[1], Anusha J[2], Manasa Hegde[3], Manjunath R Y[4], Venugeetha Y[5]
1,2,3,4 (BE(CSE)) 8th Sem, Dept of CSE, Global Academy of Technology, Bengaluru
5, Associate Professor, Dept of CSE, Global Academy of Technology, Bengaluru

Abstract: ***Social media have received more attention nowadays. Public and private opinion about a wide variety of subjects are expressed and spread continually via numerous social medias. Twitter ais one of the social media that is gaining popularity. Twitter offers organization a fast and effective way to analyze customers perspectives toward the critical to success in the market place. Developing a program for sentiment analysis is an approach to be used to computationally measure customer's perception. A novel approach for automatically classifying the sentiment of Twitter messages can be obtained by using the naïve bayes algorithm. Here messages are classified as either positive or negative with respect to a query term. Our training data consists of Twitter messages with emoticons. The results so obtained is represented using a pie chart and html page.***
Keywords: component, Twitter, sentiment, social media;

I. INTRODUCTION

In the past decade, new forms of communication, such as micro-blogging and text messaging have emerged and become ubiquitous. Twitter messages are often used to share opinions and sentiments about the surrounding world, and the availability of social content generated over sites such as Twitter creates new opportunities to automatically study public opinion.

weets are short in length: a sentence or a headline rather than a document. They are unambiguous short text messages, which has a limit of up to a maximum of 140 characters. The language they use very informal, with creative spelling and punctuation, misspellings, slang, new words, and URLs and genre-specific terminology and abbreviations, e.g., RT for **re-tweet** and #hashtags. It is very challenging to automatically mine and understand the opinions as sentiments that people are communicating. Another challenge of micro-blogging is the incredible breadth of the topic that is covered. People these days tweet about anything and everything.

Twitter generates huge data which cannot be handled manually to extract some useful information and therefore, automatic classification is required to handle the Twitter data.. The Twitter interface has provisions for the user to post short messages and that can be read by another Twitter user. Because of its popularity and opinion mining Twitter is chosen as the source in this case. The existing database is not able to process the large amount of data within a specified amount of time and is limited for processing of structured data and has a limitation when dealing with a large amount of data. Hence traditional solution cannot be used to manage and process unstructured data. Big data technologies like Hadoop is the best way to solve Big Data challenges.

A. Limitations of available systems

The available systems are not sufficient to deal with the complex structure of the big data. Limitations of the existing systems are:-

1) Extensive data cleaning, data scraping and integration strategies that will ultimately increase the overhead is required by the available systems like Twitter-Monitor and Real Time Twitter Trend Mining System.
2) The available system is inefficient for the real time analytics.
3) Analyzing huge amount of data in a short period of time is very time consuming.

The proposed system can eliminate all the above mentioned draw backs

B. Hadoop and mapreduce architecture

The Hadoop and MapReduce architecture can be explained using the HDFS architecture and MapReduce.A.HDFS architecture Hadoop enables the application to work in a distributed environment. In this case, thousands of distributed components are working together to accomplish a single task. The huge log files are distributed over various clusters known as HDFS cluster (Hadoop Distributed File System). HDFS is able to store large amount of data. It helps to create the clusters of machines and perform parallel work among them. It operates cluster without losing data or interruption of any work. It helps to manage cluster by breaking incoming files into small chunk called block.

©IJRASET: All Rights are Reserved

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

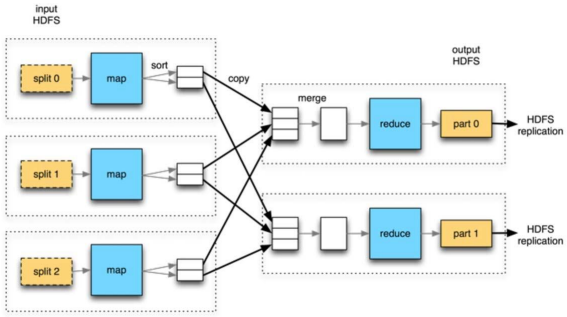


Figure 1: Architecture of the MapReduce process

Name node and data node: Name node store the information about meta data which maps to the data-node for actual data. Data node simply contains the actual data.
Data Replication: HDFS stores each files as a sequence of blocks. These blocks are replicated to various racks on HDFS for fault tolerance. The block size and replication factor can be configured from the configuration file of hadoop.
Racks: Racks are the collection of data-node. The data nodes which belong to the same network can be treated as one rack. If one of the data nodes crashes, the replica of that data node which is present on another node starts moving to the failed data node.

C. MapReduce Architecture

The Apache Hadoop is a framework that allows for the distributed processing of large data sets across clusters of computers using simple servers to thousands of machiness, each offeringlocal computation and storage.
Fig . 1 shows the architecture of MapReduce. MapReduce is aprogramming model for the processing of huge data. It is divided into two phases, the ma and reduce phase. It allows the specific application to run in parallel so that the task is accomplished in less period of time. MapReduce jobs are controlled by the JobTracker. JobTracker simply schedules the jobs submitted by the user and provide the mechanism to monitor the jobs [3].

II. THE PROPOSED SYSTEM

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.Sentiment analysis is extremely important in analyzing social media .It can be an excellent source of information and provide us great insights that can determine marketing strategy,improve campaign success, improve product messaging,improve customer service,test business KPIs ,generate leads .In a nutshell, if done properly, social media sentiment analysis can improve your bottom line.
A novel hybrid approach involving both corpus-based and dictionary-based techniques has been implemented,which will find the semantic orientation of the sentiments words in tweets.Acronyms, hash tags, emoticons, slang and special characters has also been taken into consideration ,which forms a huge part of the internet language.
Initially , the client can login the webpage either as a user or as an admin. The user and admin get different set of privileges. The user has to enter a keyword based on which the tweets are extracted and sentiment analysis is performed.The admin has access to the source code and gets the count of visitors visited the page.

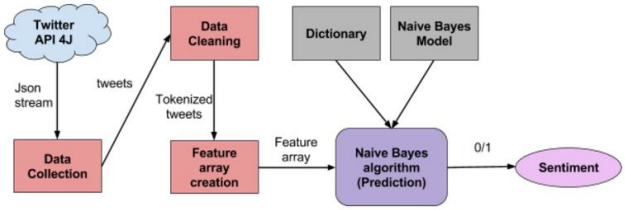


Figure 2:The System Architecture

©IJRASET: All Rights are Reserved

There are 4 important phases

A. Data Collection

Twitter allows us to access to a range of streaming APIs. The Streaming APIs offer ow latency access to flows of twitter data .For the purpose of data collection,the public stream APIs was used , which is the most appropriate method for gathering information for opinion mining purpose as it allowed access to a global stream of twitter data that could be filtered as required . The scala interface library has been installed since it was necessary for scala to interface with the twitter4J.Since twitter has numerous rules and regulations imposed ,the user has to register with user name , email id to receive the authentication details which give us access to the API.

Scala script was then created along with the authentication details to the API and initialized a streaming pprocess where data could be extracted from the twitter's web.The tweets were extracted based on the keyword entered by the user using a filter function.The downloaded data is transfered in JSON format which is less verbose than the alternative format offered by xml.A Scala script was later written to remove all unwanted content and to parse each packet and this content is stored in the main memory. Once the required content is removed from the JSON package and stored in RAM,it can now be accessed from the main memory.This information is stored as a comma separated values(CSV) file.

B. Data Cleansing

The second phase of the system will be to cleanse the data collected, this will involve removing any punctuation such as "(,.!?')" and making everything lower case. This will help in the next stage of the project especially in the "Bag of Words" approach[1]. Removing lower case words will decrease the redundancy in the database that will be used to store the words. It would be very beneficial if a spell check could be performed on the data collected but at this stage of the project it is unknown whether this would be possible using python without adding large computational overhead.

Tweets contain the user name,hashtags along with the text .The user name, hashtags are nonessential for the classification.Data cleansing is the process of removal of any unwanted content from the input tweets and the training data. The hashtags or usernames are unwanted since they are not useful for the machine learning algorithm to assign a class to that tweet. Data cleaning immensely reduces the processing cost and and simplifies the classification task for the machine learning model.

The scala script also serves to remove stop words from the tweets. Stop words are words such as 'the, which, is, and at' and they have little value for machine learning algorithms as they are contained in approximately equal measure in the positive and the negative training sets. Removing them allows more specific word features to be passed into the classification models and hugely reduces processing during the training stages. There is no definite standard for removing stop word as each application has different requirements, for this project the default stop word list was taken from the Natural Language Tool Kit (NLTK) for scala.

People use twitter not only for expressing their opinions but also for sharing information with others. Given the short maximum length of tweets, one way of sharing is using links. Tweets include various links or URLs and these do not contribute to the sentiment of the tweet. The URLs in the data used in this project are of the form http://plurk.com/p/116r50. These do not contribute to the sentiment of the tweet. Hence these were parsed and replaced by a common word, URL.

Tweets are often refer to other users and such references begin with the @ symbol.These again do not contribute to the sentiment and hence are replaced by the generic word USERNAME duplicates or repeated characters . People use a lot of acronyms on twitter .for example 'happy' can be written as 'hpy',though it means the same , the classifier considers this as 2 different words.A dictionary has been created for the same.The extracted twitter data also contain a lot conjunctions like 'and','before' etc which do not exactly contribute to the sentiment of the tweet. These words to remove in the cleansing phase to avoid using them as features.These words are removed from the data so as to avoid using them as features. The stopwords corpus was obtained from NLTK [3]. Some modifications were required to this as the corpus also had some negative words such as nor, not, neither which are important in identifying negative sentiments and should not be removed.

C. Classification

Classification of data is considered to be the most difficult part of the project.Classifying the data will entail looking at individual words or groups of words in a tweet and attempting to assign a sentiment to them.It is difficult to make computer understand the

©IJRASET: All Rights are Reserved

slang words and sarcasms.For example “This product is bad ass” , this can be wrongly classified because of the presence of “bad” but it actually means that the product is good.

The bag of words approach will involve building databases of positive, negative and neutral words.Every tweet is broken down into individual words and is compared with the words in the database.When there is a match a counter will be increamented or decremented by a fixed amount depending on a weighting assigned .This counter is used to classify the sentiment for example if the words in the tweet are largely positive the counter should be high.

D. Analysis

Once the data is classified , some kind of analysis has to be performed on it.This may include simple percentages of customer satisfaction or a more complex analysis could be performed such as comparing the customer sentiment on two similar products with the aim of finding a correlation between good sentiment and high sales of those products.

It may even be possible to look at specific features of a product such as the screen or battery life of a smart phone with the aim to find customer sentiment. This would be extremely valuable information as it would allow companies to identify perceived weaknesses in their products and allow them to improve upon them in future generations/iterations.The Description of the process in pseudo code form is given below[5]:

Input: Labeled Dataset

Output: Positive and negative polarity with synonym of words and similarity between words

Step-1 Pre-Processing the tweets:

Pre-processing ()
Remove URL:
Remove special symbols
Convert to lower:

Step-2 Get the Feature Vector List:

For w in words: Replace two or more words
Strip: If (w in stopwords)
Continue
Else: Append the file
Return feature vector

Step-3 Extract Features from Feature Vector List:

For word in feature list
Features=word in tweets_words
Return features

Step-4 Combine Pre-Processing Dataset and Feature Vector List

Pre-processed file=path name of the file
Stopwords=file path name
Feature Vector List=file path of feature vector list

Step-5 Training the step 4

Apply classifiers classes

Step-6 Find Synonym and Similarity of the Feature Vector

For every sentences in feature list
Extract feature vector in the tweets ()
For each Feature Vector: x
For each Feature Vector: y
Find the similarity(x, y)
If(similarity>threshold)
Match found Feature Vector: x= Feature Vector: y
Classify (x, y)
Print: sentiment polarity with similar feature words

©IJRASET: All Rights are Reserved

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

III. IMPLEMENTATION

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. There exists many classifiers in machine learning but we emphasis on the implementation of Naive Bayes classifier because of its simple implementation, low computational cost and it relatively high accuracy.

A.Naive Bayes

The Naive Bayes classifier is a probabilistic model which relies on the assumption of feature independent in order to classify input data.The algorithm assumes that each feature is independent of the absence or presence of any other feature in the input data, because of this assumption it is known as 'naïve'. Despite its simplicity , this algorithm is used for text classification in many opinion mining application because of it high classification accuracy when used with quality training and in specific domains .This Classification is named as Naive Bayes after Thomas Bayes, who proposed the Bayes Theorem of determining probability[3].

$P(s|M) = P(s).P(M|s)/P(M)$

Where s is the sentiment and M is the Twitter message.Since we have both positive and negative messages in a twitter message,the above equation can be simplified as:

$P(s|M) = P(M|s)/P(M)$

$P(s|M) \sim P(M|s)$.

IV.RESULTS

A visual representation of the overall sentiment contained in the input data is displayed graphically as a pie chart in figure 3.

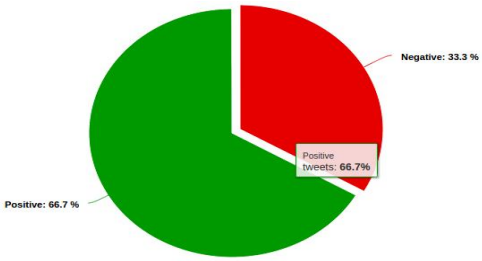


Figure 3:Sentiment Analysis

The computed result is given to a third party API (high charts) which in turn represents the results in the form of a pie chart.

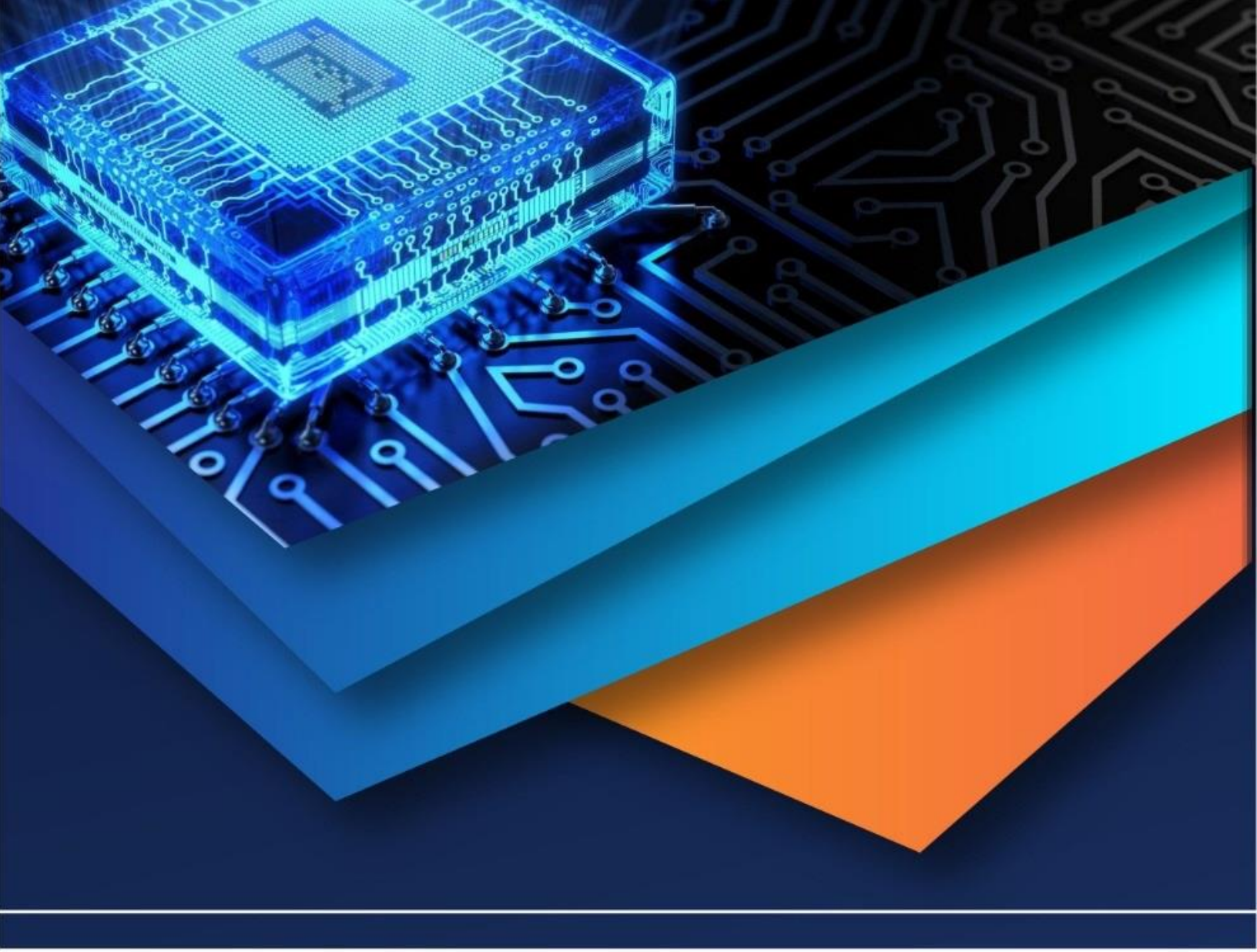
V. CONCLUSIONS & FUTURE SCOPE

Twitter Data in the form of opinion, feedback, reviews, remarks and complaint are treated as big data and it cannot be used directly. These data first convert as per requirement. Pre-processing of data is done to remove noise from the data. Sentiment analysis is implemented for data set, on Hadoop framework and analyzed with large number of tweets. The classification of positive and negative tweets saw somewhat satisfactory results if the number of tweets used for training was not too large. High degree of accuracy can be achieved using Naïve Bayes technique. This method is suitable to train and classify sentiment from twitter and other social network data. Scala programming language has been used, which adds on to the efficiency of the model. Hence, the future scope in the sentiment analysis for the other social networking websites like Facebook, Google Plus etc. Also, the database can be encrypted to prevent it from reaching wrong hands.

REFERENCES

[1] Brett Duncan and Yanqing Zhang," Neural Networks for Sentiment Analysis on Twitter", Proc. 2015111E 14th Int'! Conf. on Cognitive Informatics & Cognitive Computing IICCI"CC151,91B-1-4613-1290-91151$31.00 ©2015 IEEE
[2] Trainer Perception", 2015 IEEE Conference on e-Learning, e-Management and e-Services.
[3] Huma Parveen Prof. Shikha Pandey Sentiment Analysis on Twitter Data-set using Naïve Bayes Algorithm", 978-1-5090- 2399- 8/16/$31.00_c 2016 IEEE.
[4] Ankur Goel Jyoti Gautam Sitesh Kumar," Real Time Sentiment Analysis of Tweets Using Naive Bayes", 2016 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016.
[5] geetika Gautam Divakar yadav Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis", 978-1- 4799-5173- 4/14/$31.00 ©2014 IEEE

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)