SENTIMENTAL ANALYSIS OF TWITTER THROUGH BIG DATA

Dr. H Venugopal
Professor, Dept. Of MTech CSE
SSIT, Tumkuru

Ayesha Azeeza
PG Student, Dept. Of MTech CSE
SSIT, Tumkuru

***Abstract*— The use of social networking sites is one of the approaches for putting views of users. Social media has gained a lot of light in past few years. Twitter is a widely used social media for posting comments and short statuses. Sentiment analysis is used to analyze the different opinions of users across the world. Sentiment analysis is done on the data that is collected from internet and various social media platforms. In this paper we present a system which collects tweets from social networking sites, we'll be able to do social analysis on those tweets and thus provide some prediction on business analysis. Results of sentiment analysis will be display as tweets from different sections in terms of positive, negative and neutral from information resources.**

Keywords—*sentiment analysis, twitter, social media.*

1. INTRODUCTION

Big Data is the trending research area in the computer science and Sentiment analysis is one of the most important part of this research. Big data is considered as very large amount of data which can be found easily on web, social media, remote sensing data and medical records etc. in the form of structured, unstructured or semi-structured data and we can use this data for sentiment analysis.

Apache's Hadoop is a leading Big Data idea used by IT giants (such as Yahoo, Face book & Google). 'Big Data' is also a **data** but with a huge size. In short, Big Data is a data which is so big and complex that none of the traditional data management tools are capable to store it or process it efficiently.

Sentiment analysis is all about to get the real voice of people towards specific product, services, organizations, movies, events, issues and their attributes. If we take twitter as our example nearly 1TB of text data is generating every week in the form of tweets. So, by this it understood clearly how social media is changing humans living styles and thoughts. Among these tweets can be categorized by the hash value tags for which they are commenting and posting their tweets.

Many companies and organizations are using this Sentiment analysis process to predict the reviews and ratings of users. But to calculate the different views is difficult in a normal way by taking these heavy data that are going to generate day by day.

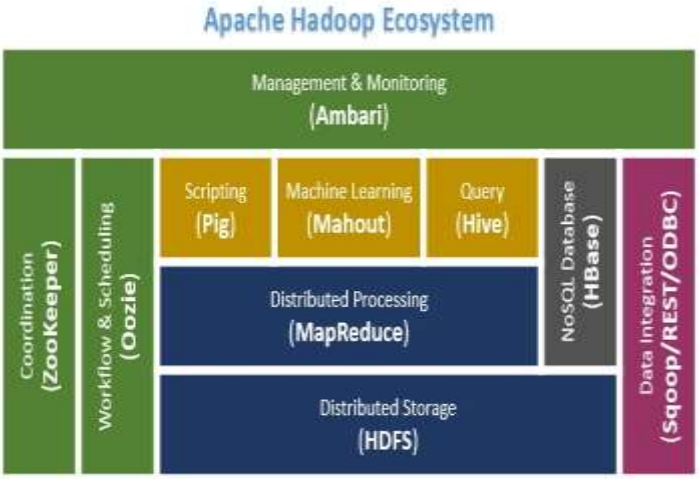

Fig 1.1: Early Apache Hadoop Ecosystem

The above figure clearly describes the different types of ecosystems that are available on Hadoop. So, now this problem can be resolved by using Big data.

1.1 Categories of Big data

Big data is categorized in three forms:

1. Structured Data
2. Unstructured Data
3. Semi-structured Data

1. *Structured Data :*Every data that can be stored, accessed and processed in the form of unchanging format is termed as a 'structured' data. Examples of Structured Data**:** A table in a database is an example.

2. *Unstructured Data:* Every data with unfamiliar form or the structured form is classified as unstructured data. In adding together to the size being huge, un-structured data pose numerous

3. *Semi-structured Data***:** Combination of structured and unstructured form of data is classified as Semi-structured Data. It's a form of data in relational DBMS which gives table description is defined

properly in the ordered form. Example of Semi-structured data is data within the XML file.

1.2 Sentiment Analysis

To know whether the slice of text is positive, negative or neutral is determined by using the concept of Sentiment Analysis. It's used to get the opinion or attitude of a speaker it's similar as opinion mining. To discover what's the feeling of a people on a particular topic is identified by this sentimental analysis technology. If we want to know the feedback of an food Masala Dose by twitter whether the food is good or bad. We can get the answer for this question by Twitter sentiment analysis. By extracting the words that shows why the food is liked or didn't liked by people we come to know why people had given the opinion that whether the food is good or bad. From this we come to know why the people are happy and not happy. This is one of the way to conduct a on the particular interested topics without using huge amount of budget and also man power to conduct a survey.

1.3 Advantages of Sentiment Analysis

- In the business purposes, it provides a good benefit.
- It helps the people to identify about products good or bad feature they want to buy.
- It helps the organization or company to know about their product limitations or bad features to improve.
- It is beneficial for the competitive party so that they can know the weakness of the opposition party.
- It decreases the effort of human being to evaluate the product as good or bad.
- It is beneficial for social media analysis.

2. PROBLEM DEFINITION

The project focuses on using Twitter, the most popular micro blogging platform for the task of Sentiment analysis. The Tweets are important for analysis because data arrive at a high frequency and algorithms that process them must do so under very strict constraints of storage and time. It will be shown how to automatically collect a corpus for Sentiment analysis and Opinion mining and then perform Linguistic analysis for the collected corpus. All the public Tweets posted on the twitter are freely available through a set of APIs provided by a twitter. Using a corpus, a sentiment classifier, is constructed that is able to determine positive, negative and neutral sentiments.

3. EXISTING SYSTEM

The present system used for sentiment analysis is a standalone system on local native machine. This system performs analysis on the base of Hadoop itself. It uses normal databases for storing of data from twitter server, also it lacks performance as calculating the sentiment on a normal machine without hadoop may utilize a lot of system resources and may result in the system failure.

The existing system uses only java for processing the data and calculating the sentiments of the particular tweet, as java processes all the data on a single node and the processing of data takes a lot of time.

4. PROPOSED SYSTEM

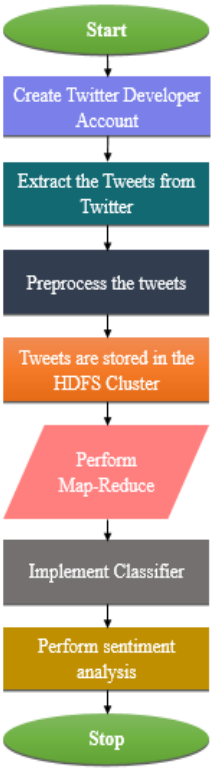

Fig 4.1: Proposed system methodology

Twitter is one of the popular social media has gained popularity dues to availability of valuable information for governments, business across the world.

The objective of this research is to develop a methodology to perform sentimental analysis of live twitter data and categorize the tweets into positive, negative and neutral opinions.

Following are steps to be followed for the above proposed methodology:

Step 1: Create an application and retrieve live data from Twitter

- Create a Twitter Dev App
- Authenticate the App and generate consumer and authentication keys.
- Provide permission like Read or Read/Write to access the App.
- Implement the code to connect to Twitter App and get the required tweets for the analysis based on some specific topic.

Step 2: Store the Tweets

- Configure HDFS
- Implement the code to connect to Mongo DB server.
- Implement the code to connect to twitter app from HDFS to get and store and process the tweets.

Step 3: Pre-processing

- This step involves removing of unwanted words and symbols from the tweets and get the data

points to generate the cluster as an input to the algorithms.

Step 4: Create the HDFS Clusters map-reduce

- These clusters will be input for the sentiment analysis step.
- The document file containing tweets is pushed into HDFS cluster using the command

Step 5: Comparing the Effectiveness of Algorithms.

- Comparing the time taken to create large number of clusters by these machine learning algorithms. This will provide the effectiveness of algorithms.
- Using the Naïve Bayes classification algorithm analysis of the percentage of the negative and positive tweets of the particular topic is determined.

Step 6: Create Classifier

- Classifier is created to use with different clusters to do the sentiment analysis. This will provide the base of sentiment analysis.

5. SYSTEM DESIGN

5.1 System Model

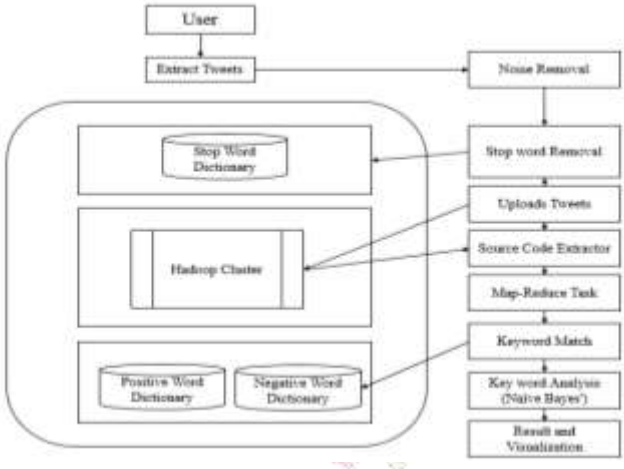

Fig 5.1: System Architecture

Analyzing Twitter sentiments with Big Data works in two phases. First phase is creating the twitter developer account and extracting the tweets by giving the authentication and consumer key and later extract the tweets and remove the noise like white space, null, special characters and stop word from the tweets by comparing the words with keywords present in the stop word dictionary list which are created by using the English grammar.

Second phase is the most important and main phase in this project first the tweets extracted into the hadoop cluster and later the tweets are read by line by line and the source code extractor is formed. Map task is performed and the set of words are outputted with the value, from this task and in the reduce task these set of words are given as input after the reduce task we are going to get the similar words with the counts. These words are matched with the key words lists in the positive and negative word dictionaries. The dictionaries are formed by taking the positive and negative words from the English grammar.

Later Naïve Bayes algorithm is applied to get the probability of positive and negative tweets feedback. The proposed system model consists of the following features: Tweets Extraction, Tweets Loader, Source Code Extractor, Noise Removal, Stop Word Cleaning, Positive Word Cleaning, Negative Word Cleaning, Map-Task, Reduce-Task, Keyword Matching, Sentimental Analysis.

Give the document file containing tweets as input and remove the noise and stop word later load the file into hadoop cluster and perform the map-reduce task and compare the set of similar words with the positive and negative dictionary and find the probability of the positive words and negative words based on the topics like politics, war, entertainment, sports, technology.

5.2 Technologies

Software Requirements

Operating System: Window 6, window 8, window 10.
Language : Java
IDE : Net Beans 6.3.1 and Eclipse,
Cross-platform : Oracle VM Virtual Box
Cloudera : Apache Hadoop based software

Hardware Requirements

Processor : Dual core
Speed : 1.1 GHz
RAM : 8 GB
Hard disk : 40 GB

5.2.1 Apache Hadoop ecosystem

For parallel, distributed data processing over several clusters an open source framework was developed that is called as Apache Hadoop Architecture. This architecture design helps to scale efficiently from pair of servers to hundreds by means of additional commodity hardware.
Hence, for computing given dataset the processing time is condensed by assigning the jobs into
different several nodes in the cluster and also distributed storage is offered by this platform.
Within the Apache ecosystem there are many technologies allowed by this computational capabilities. Key technologies that are essential of Apache Hadoop are Hadoop Distributed File System (HDFS), Hadoop YARN and Hadoop MapReduce.

5.2.2 HDFS

HDFS system is developed with the skills to identify and overcome the errors in the data processing process and avoid hardware failure, which might links to electricity blackout also.

The huge data which we going to get from the social media like twitter it's usually of more than terabytes volume this data to be stored and processed using distributed feature but traditional system lack this feature.

But HDFS to overcome the problem of data loss it offers consistent data storage and also provides data duplication. HDFS permits to store data in different formats like structured or semi-structured or unstructured format.

Within HDFS name space there will be a checksum algorithm to prove the data integrity in secured way. In a distributed file system the data will be stored in a block of fixed size. Each block size will be of 64 MB, within the cluster on different nodes the data blocks will be store. The location and the type of the data stored and all keep tracked by the Name Node.

5.2.3 Apache MapReduce

To attain parallel batch processing of huge data over cluster is done by a framework called Apache MapReduce. Algorithm implementation is done but the other development within Apache Hadoop ecosystem is done by using different tools such as (Apache Hive or Apache Pig) to accomplish the similar goal.

Benefit of MapReduce is primary data set is further divided into blocks because dataset could go closer to data for the computation. It needs applying two functions to continue further progress with MapReduce.

The map function will start to read the input dataset and later it start to process the dataset and finally it yields key and value pairs later sorting process will start it will yields output file for reduce function by sorting creating a key pairs.

MapReduce architecture stands mainly established on a single master node called Job Tracker, MapReduce tasks job scheduling and monitoring and keeping track of the failed task and also rescheduling of the task all these tasks are done by Job Tracker. All master tasks are done by the Task Tracker which is included in this architecture.

MapReduce task one output is can be given input for other task and formerly the ultimate outcomes are warehoused on distributed file system.

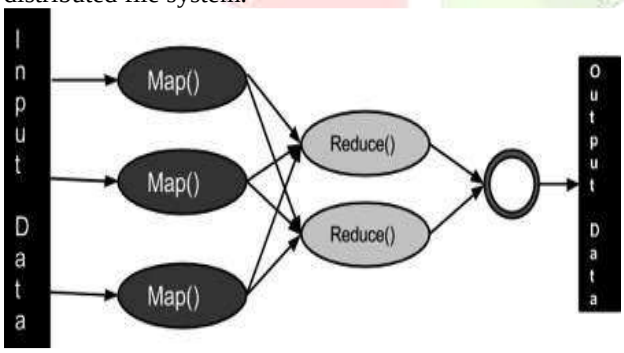

Fig 5.2: Map-Reduce Methodology

5.3 Algorithm

5.3.1 Naïve Bayes Theorem

Naïve Bayes algorithm is used to predict the result of unlabeled data and it's based on the probability of Bayes theorem. Bayes theorem accepts the individuality between predictors (features) and class using the classification technique. There are several models in Naïve Bayes models they are further they are divided based on the feature they will handle. Boolean feature vector is used by the Bernoulli model it defines that each model should have binary variables it doesn't bother about the occurrence of the individual word in the document. To classify the dataset with high volume is easy to implement because it works efficiently by using Naïve Bayes classification algorithm.

$$P(c \mid x) = [P(x \mid c) * P(c)] / P(x)$$

- $P(c \mid x)$ = Posterior Probability
- $P(x \mid c)$ = Likelihood
- $P(c)$ = Class Prior Probability
- $P(x)$ = Predictor Prior Probability

$$P(c \mid X) = P(x1 \mid c) \times P(x2 \mid c) \times \ldots \times P(xn \mid c) \times P(c)$$

For the given information probability of outcome is calculated by Posterior probability. In P (c | x), c denotes class then x is predictor. Within the class Likelihood probability of predictor is the class and class and predictor are belongs to other two probabilistic values.

Directive to know the working of Naïve Bayes theorem, to show probability prediction of a tweets occurrence based on the topics War(W), Entertainment(E), politics(P), Technology(T), Sport(S) here is an example.

5.3.2 Naïve Bayes Twitter Classification

Training set will consist of topic and sentiment, which will indicate whether an event occurs. An event will occur, variable indicate Y (Yes), otherwise N(No) as shown in figure.

Topics	P	W	P	S	P	E	P	E	E	T	T	W	T	T	P	S	W	S	W
Sentiment	Y	Y	Y	N	N	Y	Y	N	Y	Y	Y	N	N	Y	Y	Y	N	N	Y

Frequency distribution in Fig 4 shows number of event occurrences by particular topics. Moreover, the total sum for each possible event outcome is calculated.

Topics	Yes	No
Politics	4	1
War	2	2
Sports	2	2
Technology	3	1
Entertainment	2	2
Sum	13	8

Fig 5.3: Frequency distribution

Probability for each row and column is calculated. Total sum of topics occurrences is 21, thus probability for the topic Politics is as shown in Fig 4.5:

P (Politics) = [Y (4) + N (1)] / 21=~ 0.23 %. Hence, the probability for the topic Politics is 23%.

Topics	Yes	No	Probability
Politics	4	1	~0.23
War	2	2	~0.19
Sports	2	2	~0.19
Technology	3	1	~0.19
Entertainment	2	2	~0.19
Sum	13	8	Sum=21
Probability	~0.61	~0.38	

Fig 5.4: Probabilities calculations

With Naïve Bayes theorem, we can now calculate posterior probability that event will occur on the topic Politics:
P (Politics | Yes) = 4 / 13 = ~ 0.30
P (Yes | Politics) = P (Politics | Yes) * P (Yes) / P (Politics)
P (Yes | Politics) = 0.30 * 0.61 / 0.23 = ~ 0.7956
Calculation results show that probability is 79.56% for event to occur on the topic politics.
As Naïve Bayes is machine-learning technique, be applied for Twitter and sentiment analyses.

6. IMPLEMENTATION

Implementation of Sentiment analysis has the following important steps.

Fig 5.5: Training set

1) **Tweets Extraction**
Tweet extraction will take the input as authentication and consumer key given by the user and check whether the twitter developer account is a valid account it will extracts the tweets otherwise display the errors.

2) **Tweets Loader**
Tweets loader will take the input as document file containing tweets and load the file into Hadoop cluster display of the document file present in the cluster.

3) **Source Code Extractor**
For source code extractor user uploads the file into hadoop cluster if the file is existed in the hadoop cluster the tweets are read line by line from the file.

4) **Noise removal**
Noise is defined as the white space or backward spaces or special characters. The tweets containing noise will cause the problem to further classification so the noise is removed in this phase by using regular expression and then tweets without noise is displayed.

5) **Stop Word Cleaning**
Initially we create the stop word dictionary using the stop word English grammar. Later we start to read the tweets line by line and tokenize the tweets into words and start to compare the words with the stop word dictionary list. If the word matches we will discard the word displayed.

6) **Map-Task**
In this map-task method tweets are extracted from the source code extractor are given as input to the map method. It will read all the words and performs the map task and gives the set of sorted words with count displayed.

7) **Reduce Task**
Reduce task will take input from the output of the map task and it will start to merge the words with the count displayed.

8) **Positive Word Cleaning**
In the positive word cleaning, initially we create the positive word dictionary using English grammar. Later we start to read the tweets line by line and tokenize the tweets into words and start to compare the words with positive word dictionary list. If the word miss matches we will discard the word displayed.

9) **Negative word cleaning**
In the negative word cleaning, initially we create the negative word dictionary using English grammar. Later we start to read the tweets line by line and tokenize the tweets into words and start to compare the words with positive word dictionary list. If the word miss matches we will discard the word displayed.

10) **Keyword matching**
Keyword matching method will take the input from the output of the reduce task and start to match the reduce task similar words with the positive and negative dictionary word lists. If the words are matched it will increase the counts displayed.

11) **Sentimental analysis**
Based on Bayes Theorem which defines this classifier uses simple method - In what way the conditional probability of individual set of possible reasons for a given observed outcome can be figured from knowledge of the probability of each reason and the conditional probability of the outcome of each reason. We can determine the accuracy of classification using Naïve Bayes classifier. Commonly, for efficient algorithm the accuracy turns out 80% is given sentiments and distributed according to the polarity-positive, negative and neutral.

7. RESULT AND ANALYSIS

I. Output screen

Fig 7.1: Output screen

By clicking the extract, the tweets from twitter button Fig 7.2 window will open and enter the public account holder name to extract the tweets.

Fig 7.2: Extracting the tweets from public account

The tweets will be stored in the text file.

Fig 7.3: Tweets are stored in the text file

By clicking the button tweets we can load the tweets into GUI from the text file.

Fig 7.4: Loading the tweets into GUI to view

View of the tweets in GUI

Fig 7.5: View of the tweets in GUI

By clicking Transfer the tweets into Hadoop cluster in fig 7.1. The fig 7.6 window will open.

Fig 7.6 Selecting the particular topics in the tweets to perform sentimental analysis

By clicking MapReduce button particular topic sentimental analysis is done.

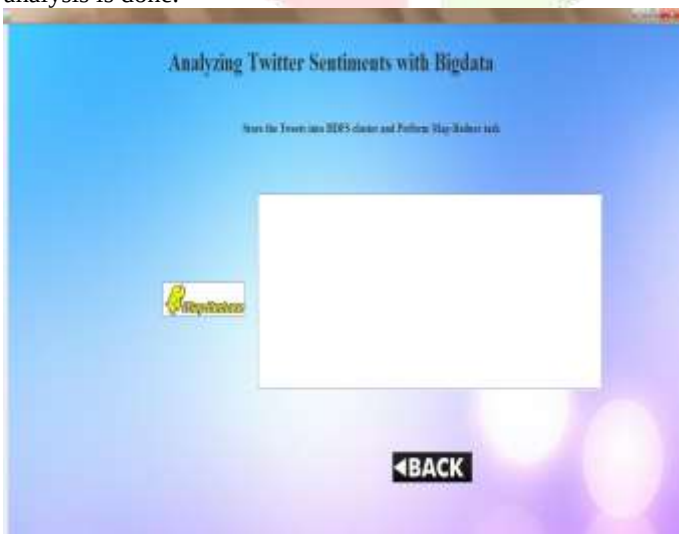

Fig 7.7: Selecting the particular topic in the tweets to perform sentimental analysis

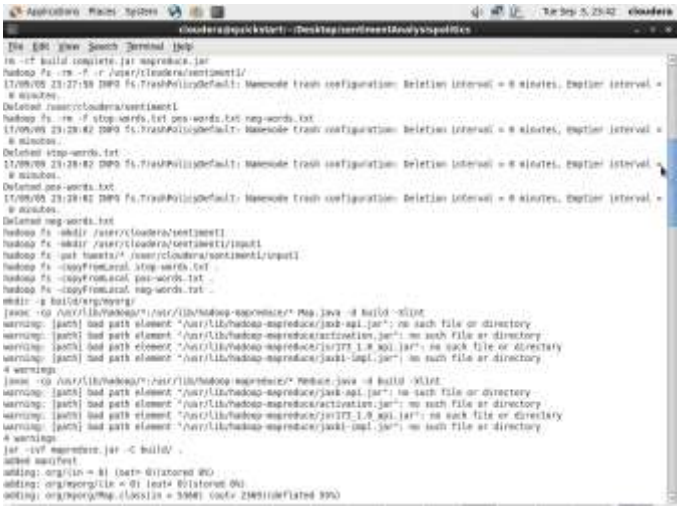

Fig 7.8: The files in the sentiment directory is deleted

The tweet file and other positive, negative and stop word dictionary files are copied from local file system to HDFS file system.

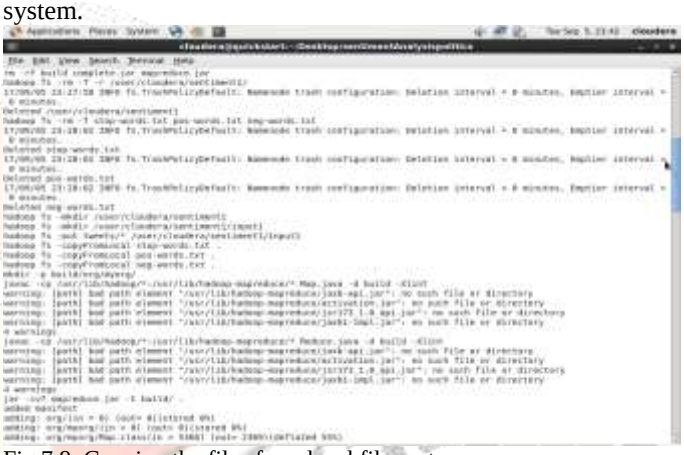

Fig 7.9: Copying the files from local file system

Fig 7.10: Map-Reduce task is performed

Fig 7.11: Sentiment analysis is performed and Naïve Baye's algorithm is done

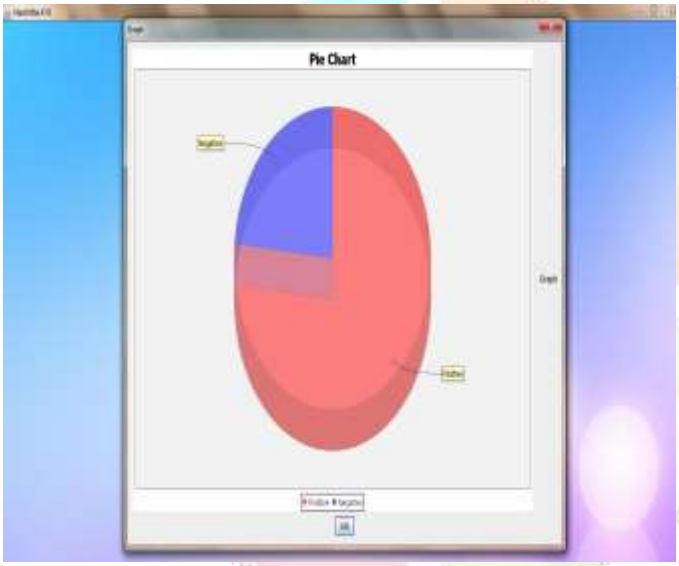

Fig 7.12: Pie chart of positive and negative scores of a particular topic

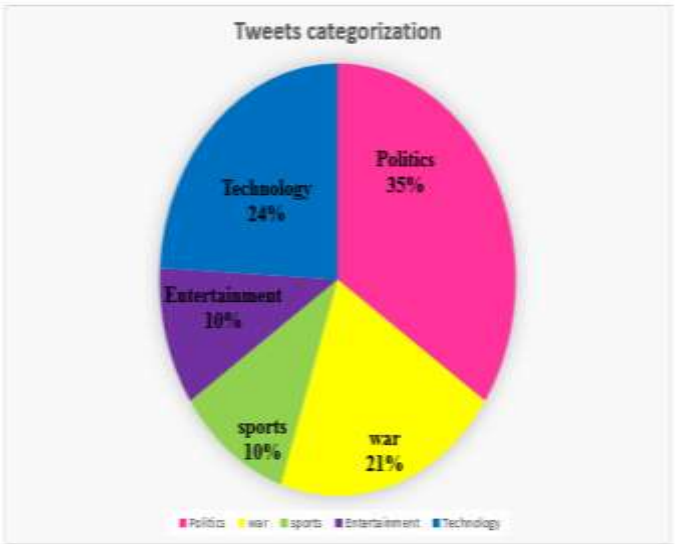

Fig 7.12: Pie chart graph visualization of different topics in tweets categorization with positive scores

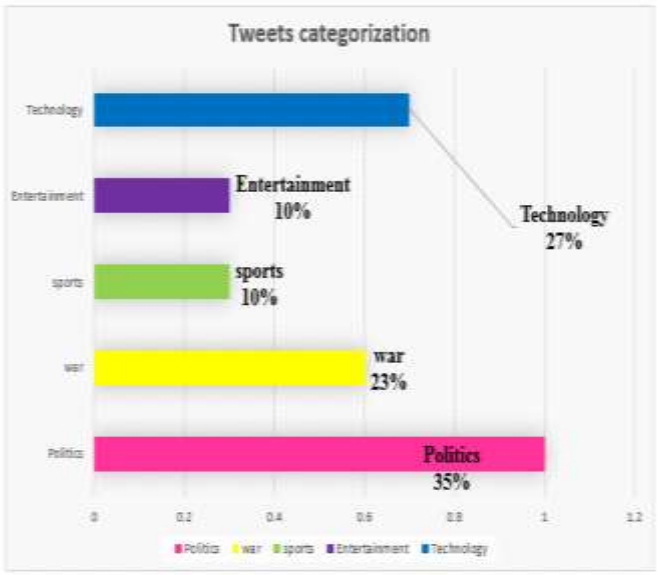

Fig 7.13: Bar chart graph visualization of different topics in tweets categorization with positive scores

ACKNOWLEDGEMENT

Firstly, I express my gratitude to my institution **Sri Siddhartha Institute of Technology**, Tumkuru. I express my deep and sincere gratitude to **Department of Computer Science and Engineering** which provided me an opportunity to fulfill the desire of reaching the goal.

I like to extend my sincere thanks to **Dr. M K Veeraiah,** Principal, SSIT, Tumkuru for providing facilities required to complete this project.

I owe immense debt of gratitude to **Dr. M Siddappa**, Prof and Head, Dept. of CSE (PG), for his valuable guidance and constant support.

I extend my sincere thanks to my project guide **Dr. H Venugopal**, professor, SSIT, Tumkuru for his valuable guidance and constant support and for helping me throughout the project work. I also thank her for her immense support, valuable guidance, inspiration and ideas which have helped me in the completion of my project.

I take this opportunity to thank the entire teaching and non-teaching staff of Dept. of CSE (PG), SSIT, Tumkuru.

My heartfelt thanks to my family and friends who have contributed for the accomplishment of this project by their constant moral and material support.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up Sentiment Classification using Machine Learning Techniques". In Proceedings of the Empirical Methods on Natural Language Processing, Pennsylvania, 2002, pp. 69-86.

[2] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: A survey," International Journal, vol. 2, no. 6, 2012.

[3] Y. Mejova, "Sentiment analysis: An overview," Comprehensive exam paper,

http://www.csuioedu/˜ymejova/publications/CompsYelenaMejova. pdf
[2010-03], 2009.
[4] [Rui Xia, Chengqing Zong, Shoushan Li, "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences 181 (2011) 1138–1162.
[6] H. Cui, V. Mittal, and M. Datar, "Comparative Experiments on Sentiment Classification for Online Product Reviews." In Proceedings of AAAI-06, 2006, pp.1266-1260.
[6] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," Machine Learning, vol. 29, no. 2-3, pp. 103–130, 1996
[6] Z. Niu, Z. Yin, and X. Kong, "Sentiment classification for microblog by machine learning," in Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pp. 286–289, IEEE, 2012.
[8] Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proceedings of LREC, vol. 2010, 2010.
[9] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: An International Journal, vol. 181, no. 6, pp. 1138–1162, 2011.
[10] Neethu M S and Rajasree R, "Sentiment Analysis in Twitter using Machine Learning Techniques" 4th ICCCNT 2013 July 4 - 6, 2013, Tiruchengode, India IEEE – 31661.
[11] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, "Using wordnet to measure semantic orientations of adjectives," 2004.
[12] C. Fellbaum, "Wordnet: An electronic lexical database (language, speech, and communication)," 1998.
[13] P. D. Turney, "Thumbs up or thumbs down: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 416–424, Association for Computational
Linguistics, 2002.
[14] Y. Xia et al, "Word polarity disambiguation using Bayesian model and opinion level features" Cognitive Computation, vol. 6, no.3,2016.
[16] M. Dragoni, A.G. Tettamanzi and C. da Costa Pereira, "Combined system for concept-level sentiment analysis" Semantic web evaluation challenge, springer,2014, pp,21-26.
[16] M. Araujo "iFeel: A System that compares and combines sentiment analysis methods," Proc.23 International Conf. World Wide Web,2014, pp.66-68.
[16] J.M Chenlo and D.E. Losada, "An Empirical Study of Sentence Features for Subjectivity and Polarity Classification", Information Sciences, vol.280,2014, pp.266- 288.
[18] J. K, C. Chung, C.E. Wu, and R.T.H. Tsai, "Improve Polarity Detection of Online Reviews with bag of Sentiment Concepts" Proc,11 European Semantic Web Conferrence,2014.
[19] F. Bravo-Marquez, M. Mendoza, and B. Poblete, "Meta-Level Sentiment Models for Big Social Data Analysis," Knowledge-Based Systems, vol.69,2014, pp,86-99.
[20] G, Gezici, "SU-Sentilab: A Classification System for Sentiment Analysis in Twitter," Proc. International Workshop Semantic Evaluation,2013, pp.461-466.