Mining The Social Web – Twitter, Multiclass Tweet Categorization Using Naïve Bayesian Classifier

[1]Shobha Patel, [2]Khushboo Kumari, [3]Manjit Jaiswal
[1]Student, [2]Student, [3]Assistant Professor
Department Of Computer Science & Engineering
Guru Ghasidas Vishwavidhayalaya, Bilaspur, India

Abstract : Big Data refers to collection of large datasets containing bulky amount of data. Big Data is generated from various sources such as social networking sites like Facebook, Twitter, Instagram etc. and the data that is generated can be in various formats like structured, semi-structured or unstructured format. Social media monitoring is growing day by day, therefore, analysis of social data plays a vital role in knowing user behavior. These behavior of users country wise helps in getting information about various current trends and can be used further in deciding usefulness of some tasks, products, and themes. Therefore, it is important and necessary to classify these trending topics into various categories with high accuracy for better information retrieval. To address this problem, this paper aims to classify Tweets into general categories first such as sports, politics, technology and more categories later on based on Naive Bayes algorithm. Tweets are available in JSON format which is to be converted into a structured data. By classification all the tweets about a particular topic we would give the output of how data is classified for a particular topic with a good accuracy.

Index Terms **- Classification, Trending Topics, Tweets, Twitter, Naive Bayes Algorithm.**

I. INTRODUCTION

Twitter is a social networking site launched on July 2006. Twitter is an online news and social networking service where users post and interact with messages, known as "tweets", restricted to 280 characters at [1]. Registered users can post the tweets, but those who are unregistered can only read those tweets. Users access Twitter through its website interface, Short Message Service (SMS) or mobile device application software at [2]. Many people make use of social networking sites like Twitter to share their emotions, sentiments as well as providing the latest information which is evident from the reactions of people on the events encompassing the Egyptian revolution of 2011 at [3].

There is a lot of drivel on Twitter. But at the same time, there is a growing base of really useful Information and knowledge content on Twitter. Here the important thing is that which content is worth following there. If a user wants to search tweets about a particular topic then the categorization of tweets must be done first to make any sense out of the vast amounts of tweets in Twitter. For this purpose we propose the use of Naive Bayes supervised learning classifier for the categorization of tweets. Before going into the details, the necessary information related to twitter in general are given as follows.

TWEET – A tweet is simply a posting a message of up to 280 characters. People post messages about their various daily activities via these tweets. News are also been posted by the news channels, tweeting them via Twitter to alert the users. These messages also include URLs to web pages or hash tags to relate tweets of similar topics together.

Characteristics of Tweets: Twitter messages have many unique attributes, which differentiates our research from previous research:

A. **Length:** The maximum length of a Twitter message is 280 characters. From our training data set, we calculate that the average length of a tweet is 28 words or 156 characters.
B. **Data availability:** Another difference is the magnitude of data available. With the Twitter API, it is very easy to collect millions of tweets for training. In previous research, tests only consisted of thousands of training items.
C. **Language model:** Twitter users post messages from many different media, including their cell phones, laptops etc. The frequency of misspellings, repetitions and slang in tweets is much higher than in other domains.
D. **Domain:** Twitter users post short messages about a variety of topics unlike other sites which are tailored to a specific topic. This differs from a large percentage of past research, which focused on specific domains such as politics reviews.

@USERNAME - Using the '@' symbol people can address their tweets to some person who they want to communicate with. A person can also address to multiple users for example, @aaronpaul @bryancranston.

#HASHTAG – The popular or the trending topics on Twitter are highlighted using the hashtags. People can also contribute towards the trending topics by attaching their personal message to them for example, #rooney 'How much is he really worth after signing the new deal?'

RT - A retweet is a popular feature of Twitter using which a user can forward a tweet of another person like a celebrity to his own followers. Before diving deep into all the research work currently being done related to tweet text classification or categorization it is imperative to have some knowledge of classification methods or algorithms being used for the particular research work. The remainder of the paper is structured as follows. Section 2 deals with all the different research work done related to classification or categorization based on twitter content. Section 3 deals with the collection followed by the pre-processing of tweets. Section 4 deals with the use of the Nave Bayes algorithm along with the Map Reduce paradigm for categorization of tweets.

II. RELATED WORK

Text classification is a well traversed area of machine learning, thanks to its potential for wide-reaching impact. [4] and [5] present good overviews of the methods used so far and their relative strengths and weaknesses. In [4], Yang and Liu go over the theory behind classifier methods such as Support Vector Machines, Neural Nets, k-Nearest Neighbour and Naive Bayes approaches. Their results do not offer much by way of quantitative outcomes of using these methods to classify short text snippets as we have. Lee makes some fascinating observations about how humans perform (long form) document classification in [5], providing a Bayesian model that fits how people seem to reason about this problem. His approach seems specific to the limited-domain topic classification problem though, that will be relevant for the second dataset in our paper.

1. "Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier" In Big Data, 2013 IEEE International Conference on, pp. 99-104.IEEE, 2013. [6] Because of their ability to “learn” from the training dataset to predict or support decision making with relatively high accuracy the machine learning technologies are widely used in sentiment classification.
2. In [7] categorization of Twitter messages based on their content. Their results show some fascinating observation such as, 80% of the tweets belong to user to user communication whereas the rest 20% has news characters.
3. In [8] ,proposes a system called 'Twitter Stand' for capturing Tweets regarding worldwide breaking news. For this the categorization is done into two classes namely 'news' and 'junk'. All of these approaches provide a valuable insight into Tweet classification but none proposes an approach to handle large number of tweets.

III. PROPOSED WORK

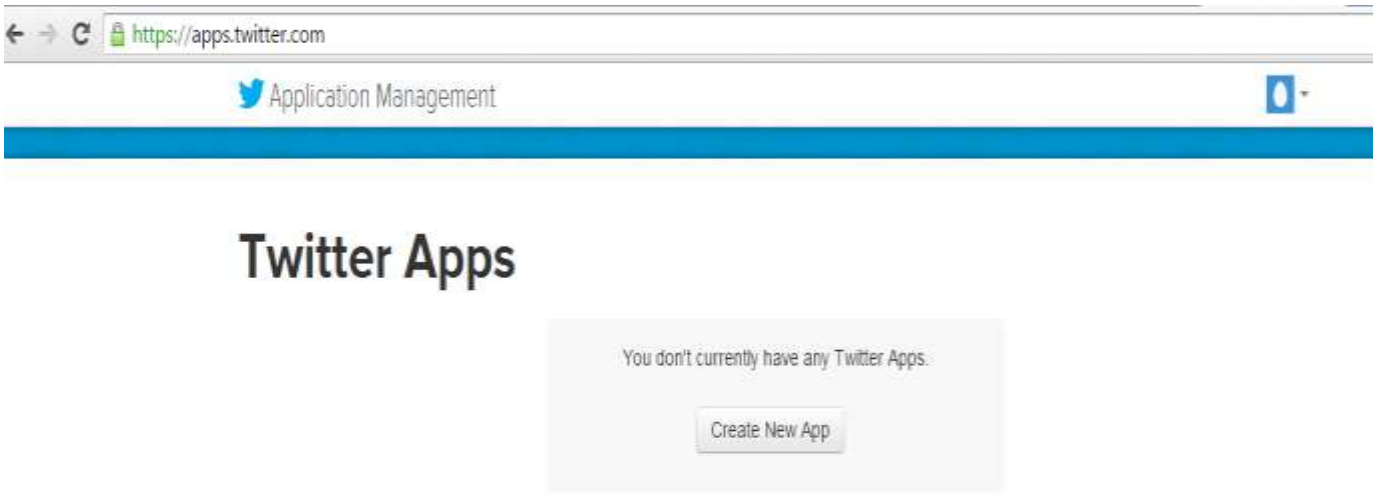
A. COLLECTION OF TWEETS:

The preliminary step deals with collection of tweets for categories like sports, elections and technology. Twitter provides two API's for gathering tweets namely the Twitter Streaming API and the Twitter REST API. We make use of the Twitter Streaming API to gather our tweets. We have used the Twitter4j library to gather tweets which internally uses twitter Streaming API. The Twitter4j library requires OAuth support to access the API. Twitter uses OAuth to provide authorized access to its API. We have used the Application Only Authentication where the application makes API requests on its own behalf, without a user context. API calls are still rate limited per API method, but the pool each method draws from belongs to your entire application at large, rather than from a per-user limit. We have generated OAuth settings by using a twitter account, as we cannot use the settings unless you do not have a registered account. For gathering the tweets some important steps are given as follows :

Step 1: Getting Twitter API keys and access token

In order to access Twitter Streaming API, we need to get 4 pieces of information from Twitter: API key, API secret, Access token and Access token secret. Follow the steps below to get all 4 elements:

- Create a twitter account if you do not already have one.
- Go to https://apps.twitter.com/ and log in with your twitter credentials.
- Click on "Create New App"



- Fill up the form, agree to the terms, and click on the "Create your Twitter application"
- In the next page, click on “API keys and access token” tab, and copy your "consumer key" and "consumer secret key ".
- Scroll down and click on "Create my access token", and copy your "Access token" and "Access token secret".

Step 2: Connecting to Twitter Streaming API and downloading data

- We are using an editor i.e. eclipse to connect to Twitter Streaming API and downloading the data.
- Next create, a java file. Make sure to enter your credentials into `access token`, `access token secret`, `consumer key`, and `consumer secret key`.

The output data looks like:

```
SimpleStream [Java Application] /usr/lib/jvm/java-8-openjdk-amd64/bin/java (29-Oct-2017, 8:38:59 PM)
[Sun Oct 29 20:39:00 IST 2017]Establishing connection.
[Sun Oct 29 20:39:06 IST 2017]Connection established.
[Sun Oct 29 20:39:06 IST 2017]Receiving status stream.
ndtv
India
924654328232775680
RT @CricketNDTVLive: 3rd ODI: Henry Nicholls hits Kedar Jadhav for a 4!  252/4 (41.1 Ov) #INDvNZ https://t.co/hDV6nWrVZp https://t.co/JbBsz…

real_Ali_XaTti
Islamabad; Pakistan
924654328015738881
RT @Saj_PakPassion: Mohammad Hafeez bowling to a left-hander.....you know what happens next #Cricket #PakvSL

NawazsTiger
null
924654329348476929
RT @Parri777: Just not International Cricket that is Coming home Today, Its Our Pride and Passion, Welcome Sri Lanka to Pakistan…

GulSWaqas
Lahore,Pakistan
924654331491766272
RT @IzhaarEMuzamat: Whole Pak lost int'national cricket bcoz of Lahore attack and this is 9th consecutive cricket match in Lahore with none…

WisdenScores
null
924654331621605377
MILESTONE! NZ 252/4 (41.1 ovs) https://t.co/P6UZGDKE53 #INDvsNZ #cricket 3rd ODI

WisdenScores
null
924654333425172480
FOUR! K Jadhav to H Nicholls, NZ 252/4 (41.1 ovs) Target: 338 https://t.co/P6UZGDKE53 #INDvsNZ #cricket 3rd ODI

SalmanS6
London,UK
924654338978574336
RT @Saj_PakPassion: Waqar Younis "Sri Lanka coming to Pakistan and showing their support for Pakistan cricket will never be forgotten" #Cri…

MajidAli66639
الرياض, المملكة العربية السعودية
924654340898359296
RT @defencepk: Sri Lankan team is always welcome !#PAKvSL https://t.co/CnKH2oMkRD
```

B. PRE-PROCESSING:

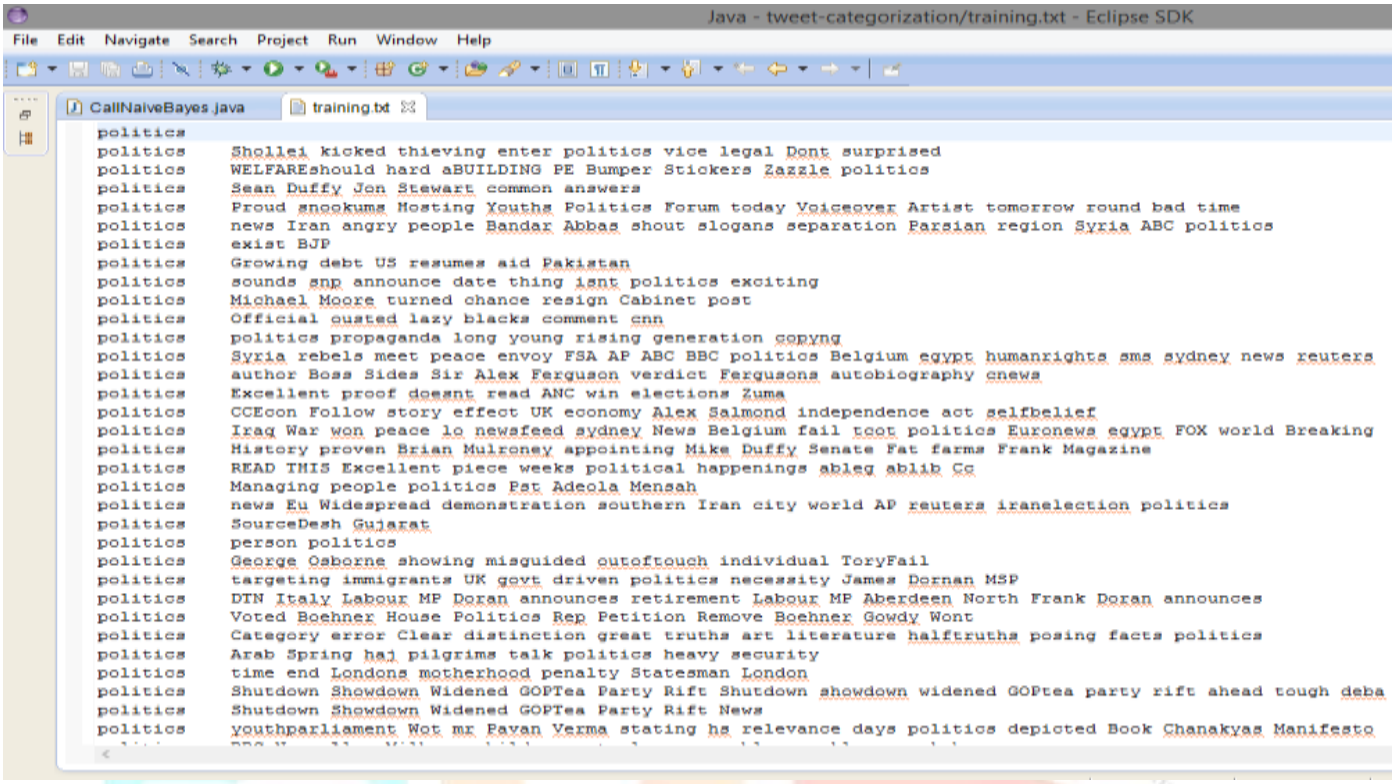
Pre-processing involves a series of techniques which should improve the next phases of classification, in order to achieve better performances. Before using the tweets collected from Twitter as training data set, pre-processing of the tweets is done to remove unnecessary information or disturbing elements and redundant word for the next phases of analysis and in the normalization of some misspelled words. The pre-processing steps include as at [9]:

1. **Removing links (URLS):** The directions to URL may be included by these tweets. The embedded URL is usually used to give the source for the detailed description of the content include in the tweets. But in our gathered tweets the description of the web content is not taken to classification of tweets. Hence all such web directions are removed from the gathered tweets.
2. **Removing usernames:** Twitter lets the users send private messages to other person. This is usually start with a '@' symbol followed by a user like @username. The user mentions doesn't indicate any important information to us in the text classification process. So we remove the @Username.
3. **Removing special symbols:** Special symbols like #, . , ^, $ which are unnecessary. If we talk about the hash tag, the hash tags are another entity associated with tweet to name or tag a topic and usually start with a # symbol. The hash tag doesn't provide significant information. So these special symbols are removed since found no significance in our classification approach.
4. **Removing emoticons:** This step removes the various emoticons in the gathered tweets.
5. **Removing Stop words:** This step removes the various stop words that constitutes the grammar of the sentence as they do not give us significant information (like the, for, who etc.).For this we gathered a corpus of stop words which is used for removing the stop words from the tweets.
6. **Removing of retweets:** Retweets looks like normal Tweets(on Twitter a message posted by another user) with the author's name and username next to it, but are distinguished by the Retweet icon and the name of the user who retweeted the Tweet. Retweets contain the copy of the original tweet, so it adds to redundancy of information. So in this step we remove the retweets so that training data contains the unique tweets only.

C. CONVERT DATA INTO TSV FORMAT:

Tab -Separated Values (TSV) is a text format that is used to store data in a tabular structure. It is very similar to the CSV format, but the delimiter is a tab rather than a comma.

The TSV format is thus a type of the more general delimiter-separated values format is a useful alternative to the CSV format if your data contains commas. Commas are very common in text data and they are used in European number formats.



D. LABELING:

After pre-processing the final step of tweets is the labeling of tweets based on categories namely politics, sports and technology etc.

Table 1: Distribution of dataset as per the category

Name of Category	Percent of total tweets in Data Set
Politics	34 %
Technology	29 %
Sports	37 %

E. TRAINED THE NAÏVE BAYES CLASSIFIER WITH TRAINING DATA SET:

1. **Naive Bayes Classifier:**

Naive Bayes is a probabilistic classifier that is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high with strong (naive) independence assumptions.

In machine Learning , Naive Bayes classifier is based on the assumptions that the presence or absence of a particular feature, given the class variable, is unrelated to the presence or absence of any other feature. We have used the Naïve Bayes classifier for the categorization of tweets due to its simplicity. The probability model for a Naïve Bayes classifier is a conditional model. Now P(C|F1, ……, Fn) over a dependent class variable C with a small number of outcomes or classes, is conditional on several feature variables F1 through Fn. Using Bayes' theorem, this can be written as:

$$P(C/F_1,F_2,\text{------},Fn) = \frac{P(C)*P(F_1,F_2\text{-------},Fn/C)}{\qquad\qquad} \qquad (1)$$

$$P(F_1, F_2 \text{------} Fn)$$

In general language, we can represent it as:

$$\text{Posterior} = (\text{Prior} * \text{Likelihood}) / \text{Evidence}$$

In simple terms, Bayes's rule says that if you have a hypothesis H and evidence E that bears on that hypothesis, then [8]:

$$Pr[H/E] = \frac{Pr[E/H] * Pr[H]}{Pr[E]}$$

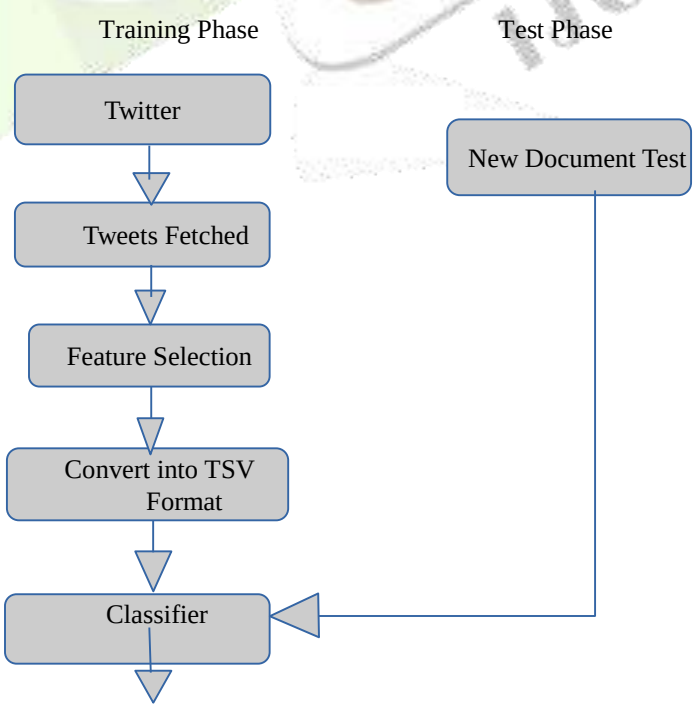
In case of Text classification Naïve Bayes classifier would model a document on basis of presence or absence of words on that document. We have employed a Multinomial Naïve Bayes classifier which considers frequency of words. In our case it can be denoted as:

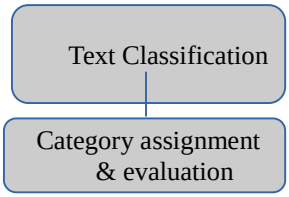
$$P(C/T) \propto P(C) * \Pi P(Wk/C) \quad \text{——————————} \quad (2)$$

Where, P (C / T) is the probability of tweet T being in category C, P(C) is the prior probability of category C (obtained from the training data) P (Wk | C) is the posterior probability of word.

ALGORITHM:

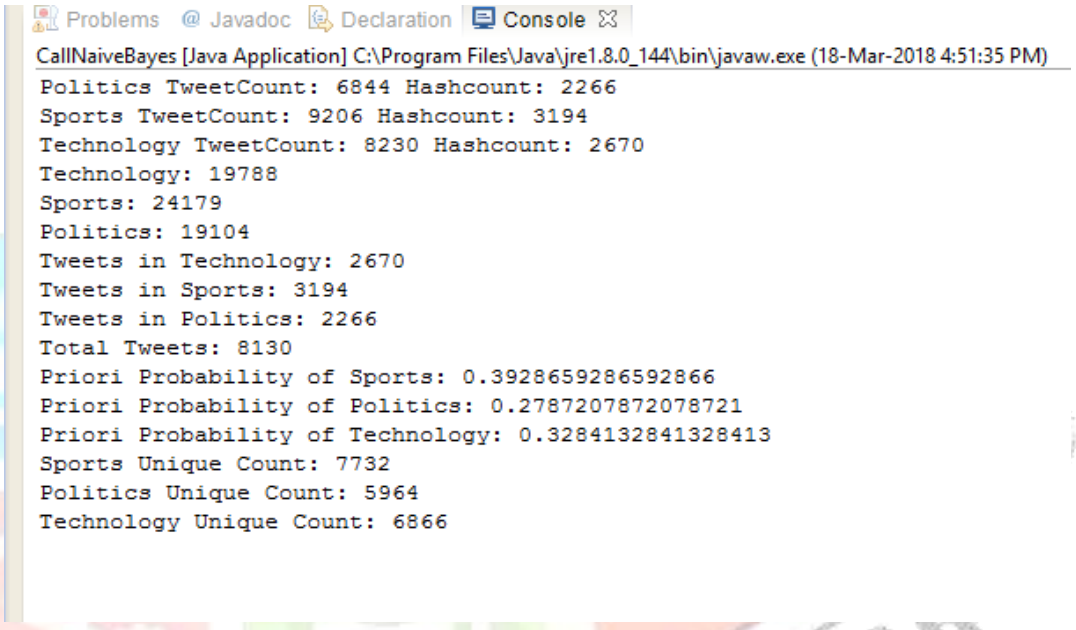
Step 1: Prepare the document-term matrix according to term frequencies.
Step 2: Perform feature selection.
Step 2-a: Compute probability for each term as per Equation (1).
Step 2-b: Select 'k' highest ranking features.
Step 2-c: Trim the document-term matrix for both training and test datasets so that only the selected features are included.
Step 3: Train the Modified Multinomial NB Classifier
Step 4: Test the sample and predict the class as per Equation (2)



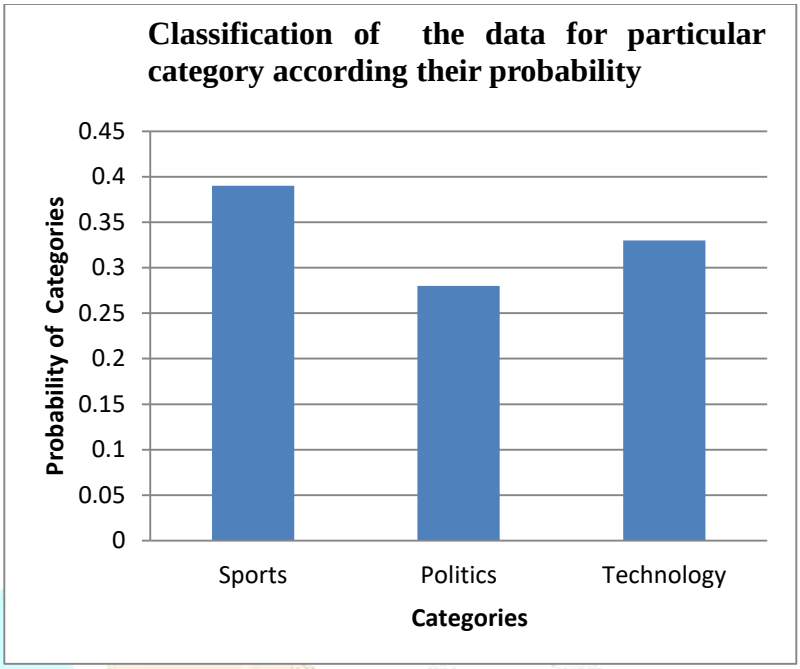


IV. RESULT

Despite of the assumption that features in the text being classified are independent of each other, we observe that this classifier performs reasonably well on our data set and also is significantly faster. We observed that the accuracy of the classifier is about 75% for the test data set, consisting of 100 sample test tweets, which we collected from Twitter.



Graphical representation of textual data using text categorization given as below :



V. CONCLUSIONS AND FUTURE SCOPE

Social networking sites research so far has become an exciting area for research projects related to machine learning and data mining. The huge amounts of data being generated on social networking sites such as Twitter can be very helpful to big multinational companies or politicians in making important strategic decisions based on user classification, sentiment or geographical distribution. For analyzing the user behaviour, first of all twitter data is extracted using Java through Twitter Streaming API. The extracted data is available in unstructured (JSON) format which is converted into structured format. Data needs to be filtered before analyzing. Data is cleaned by removing stop words. This paper thus helps for tweet classification with the help of existing Naïve Bayes classifier. This will help to analyze large dataset much easily using multiple node clusters.

In future, the data can be from multiple sources at the same time. Map-Reduce can be applied to the existing Naïve Bayes Classifier to improve the efficiency and accuracy. This will help to analyze large dataset much easily using multiple node clusters.

VI. REFERENCES

[1]"Tweeting Made Easier" Retrieved November 7, 2017.
[2]"Twitter via SMS FAQ" Retrieved April 13, 2012.
[3] Choudhary, Alok, William Hendrix, Kathy Lee, Diana Palsetia, and Wei-Keng Liao. "Social media evolution of the Egyptian revolution." Communications of the ACM 55, no. 5 (2012): 74-80.
[4]YimingYang and Xin Liu. A Re-examination of Text Categorization Methods http: //www.inf.ufes.br/~claudine/courses/ct08/ Artigos/yang_sigir99.pdf
[5]Michael D. Lee. Fast Text Classification Using Sequential Sampling Processes http://lvk.cs.msu.su/~bruzz/articles/ classification/Fast%20Text%20Classification% 20Using%20Sequential.pdf
[6] L. Bingwei, E. Blasch, Y. Chen, D. Shen and G. Chen, "Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier", In Big Data, 2013 IEEE International Conference on, IEEE, (2013), pp. 99-104.
[7] Naaman, Mor, Jeffrey Boase, and Chih-Hui Lai. "Is it really about me?: message content in social awareness streams." In Proceedings of the 2010 ACM conference on Computer supported cooperative work, pp. 189-192. ACM, 2010.
[8] Sankaranarayanan, Jagan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. "Twitterstand: news in tweets." In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 42-51. ACM, 2009.
[9] http://shodhganga.inflibnet.ac.in/bitstream/10603/57890/8/08_chapter3.pdf