JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

DESIGN AND DEVELOPMENT OF A FRAMEWORK TO PROCESS TWITTER DATA

E Srinivasa Raju[1], Dr. Trayambak Hiwarkar[2]

[1]Research Scholar, Dept. of Computer Science & Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Bhopal-Indore Road, MadhyaPradesh, India
[2]Research Guide, Dept. of Computer Science & Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Bhopal Indore Road,Madhya Pradesh, India

ABSTRACT: Twitter is the mainstream microblogging site where a great many individuals trade their musings every day as tweets. The characteristics of tweet is to be short and basic method of articulations. Despite the fact that this paper will focus on analysis of twitter data. The examination territory of investigation are text data mining and NLP. By utilizing distinctive supervised machine learning techniques we will play out the supposition examination on twitter data. Anyway we will zero in on techniques and kinds of feeling examination where we will perform how to separate tweets from twitter. Further we will look at changed machine learning techniques on the equivalent dataset and furthermore locate some standard measures.

Introduction

Social networks and microblogging sites have become the unparalleled source of unstructured data. This data is enormous in quantity and also in terms of the useful information they can provide if we process them effectively. This is due to the nature of microblogs on which people post real-time messages about their opinions on a variety of topics, discuss current issues, complain, and express their sentiment for products they use in daily life. In fact, many companies have started analyzing such massive amount of important data to get a sense of general sentiment for their product and/or services. Many times proactive companies study user reactions and reply to the user on social microblogs. This process provides on spot solution to make the user experience better but at large scale its very painful and time-consuming task. So the challenge here is to build solutions which analyze the sources of data coming from various microblogging and social networks to make important long-term product design and service implementation decisions.

Interpersonal organizations and microblogging sites have become the unmatched wellspring of unstructured data. This data is gigantic in amount and furthermore as far as the helpful data they can give in the event that we cycle them viably. This is because of the idea of microblogs on which individuals post continuous messages about their suppositions on an assortment of themes, examine current issues, whine, and express their items they use in every day life. Truth be told, numerous organizations have begun dissecting such gigantic measure of significant data to get a feeling of general assumption for their item as well as administrations. Ordinarily proactive organizations study client responses and answer to the client on social microblogs. This cycle gives on spot answer for make the client experience better yet everywhere scale its extremely difficult and tedious undertaking. So the test here is to fabricate arrangements which dissect the wellsprings of data originating from different microblogging and informal communities to make significant long haul item plan and administration execution choices.

Client produced substance can make numerous open doors for showcasing and publicizing cases in which the data mining techniques are utilized. Twitter is valuable for perusing and finding fascinating subjects that grabs client's eye. Individuals can find continuous news about what's going on the planet or keep in contact with companions. Then again, numerous organizations use Twitter to keep clients refreshed about offers and arrangements. Textual context of tweets has a relationship with two extra metadata that are separated into elements and spots. Tweet elements are client specifies, which speak to method of referencing different clients in own tweets by including @ sign, trailed by their username. Besides, tweet substances may contain likewise hashtags and URLs. On opposite, tweet places speak to true areas that might be coordinated to a tweet.

Wording is a significant aspect of the Twitter, since it shows clients the administration usefulness and highlights. Besides, it characterizes parts of Twitter and different prospects how to utilize it. So as to comprehend the Twitter wording, a concise diagram is introduced. Twitter use at (@) sign so as to call somebody username in the tweet or send client a message. In addition, this sign is utilized at whatever point client needs to make association with other client and connection to his Twitter profile. Username

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

extraordinarily recognizes every client and are commonly utilized with @ sign, for instance, Andy Murray is @andy_murray.

Another well known image utilized on Twitter is called hashtag. It causes clients to arrange messages. Actually, its structure accompanies the # sign followed by the significant watchword in association with tweet message. Basically, it arranges the tweets dependent on its context and empowers better query items by Twitter Search. Hashtags can be found anyplace in the tweet. At the point when clients click on it, they will be coordinated to classification that bunches all tweets from the Twitter clients inside a similar theme.

II. PROPOSED METHODOLOGY

Yet gathering the data isn't exceptionally basic task. We do need to consider endless focuses while gathering the data. So in our theory we will gather the dataset for preparing, testing and for data investigation. This paper comprise how data will gather, how data will prepared, put away and mostly how to arrange those data.

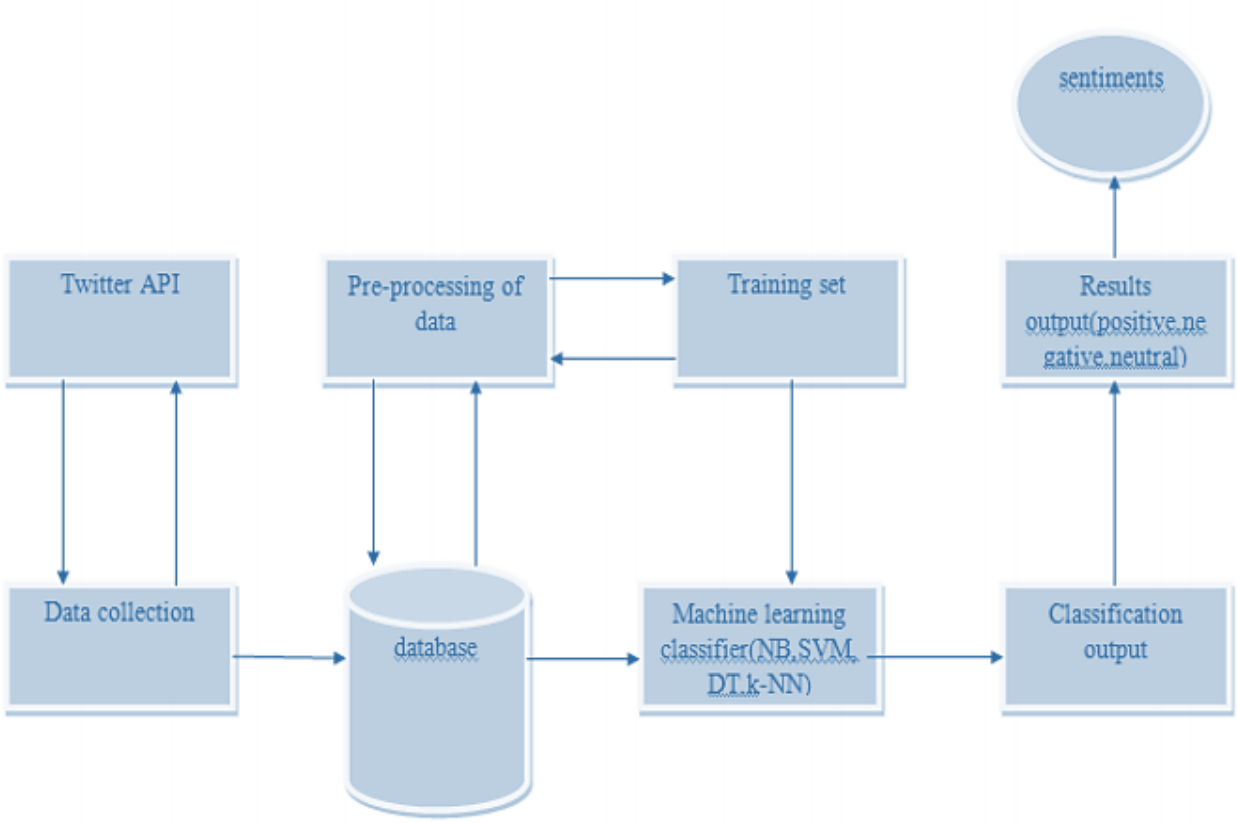

Figure 1. Architecture Diagram

Algorithm

- in this step we will extract tweets for our new classifier by using the Tweepy API library in python.
- in this step after streaming the tweets, we will do preprocessing of those tweets so that they could works well in feature extraction and mining.
- now after completing the pre-processing, we will send our data to our new build classifier

Which will classify our data into classes, for example, positive, negative and impartial and our classifier will likewise tell the precision. Since, for examining the tweets, we have took the data from twitter that will stream into our database. Consequently we will utilize twitter application.

Extraction of Twitter Data using API

With assistance of Tweepy, we can download the tweet from the twitter. Twitter API is of two kind i.e, REST API and Streaming API. Rest API represents Representational state move. So there is bit distinction between these two. Lay API takes a shot at reaction and answer premise. In which customer sends solicitation to worker and afterward worker sends back solicitation to customer while on account of streaming API it is diverse in

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

regard to reaction. At whatever point there is any update accessible on worker. At that point it began ceaselessly to customer.

Collection of Data

Twitter data so as to remove data through tweeter API, we ought to have account on twitter.com. It could without much of a stretch access by topping close down structure on twitter website. In the wake of getting fruitful enlistment on twitter they will give a substantial username and a legitimate secret key, which you can use for additional login. When you finished with this cycle, at that point you are permitted to do tweets, retweet, likes and so on any subject you need.

Twitter is a stage which give you to get to data and furthermore permit you to utilize it for self-reason. In reality before utilizing the twitter site right off the bat we should need to login on twitter then no one but we can get to the data. This website request that you give the important detail to making an application which later you could use for streaming reason. At the point when our API is created then we will get access of certain keys, for example, client keys, access token key, client mystery key, access mystery key. These key assume significant job when any client needs to get the data.

Pre-processing of twitter data

Twitter data might be in unstructured arrangement that isn't useful for separating highlight. Tweets may comprise of void spaces, stop words, slangs, uncommon characters, hashtag, emojis, time stamps, truncations, URL's and so forth for mining these data we ought to need to pre-measure the data first by the utilizing the elements of NLTK. While doing pre-preparing our first point is to separate message then we will eliminate all hashtags (#), void spaces, rehashing words, stop words, (for example, he, she, them, the and so forth) emojis and contraction will be supplanted by their relating significance, for example, :-), =D, LOL. They will be supplanted by glad, chuckle and snickering out uproariously separately. After done this thing we are prepared to give this pre-measure data to our new classifier for additional cycle so we could get our necessary outcome.

We did code in python where we define function which would be used to get processed data.

Remove quotes: give the access to user to eliminate the quotes from the tweet.

Remove @- give the option to remove the @ symbol, delete @ together with the username or replace @ and the username with a word 'AT_USER' and append to the stop words.

Remove URL’s- URL stands for uniform resource locator. offers options to remove URLs or replace them with the word 'URL' and append to stop words.

Removal of RT(Re-Tweet)- it deleted the RT word from the text.

Removal of emoticons- replace emoticons with their correct meaning

Removal of duplicates- delete all the duplicate word from the tweet.

Removal of hashtag(#)- remove hashtags from the tweet.

Removal of stopwords- delete all the stopwords from the tweet such as he, she, them because they do not convey meaning in classification.

Removal of slang- remove all slangs with their specific meaning such as loli.e laughing out loudly Etc.

We made a classifier in with the end goal that it comprises of different kind supervised learning classifier which at that point classify the tweets into twofold classes that is positive and negative. Python library utilized in building a classifier, scikit-learn. Scikit-learn is a library which is utilized to perform machine learning in python. It is an open source library and furthermore dependent on well known library, for example, numpy, scipy, matplotlib. Scikit has many tuning boundary alongside great documentation and backing. Thusly it has numerous apparatuses for grouping, perception, relapse, bunching and so forth through a basic order in python we can introduce scikit-learn in our framework, for example, 'pip introduce scikit-learn'.

There are several classifier comes under the scikit-learn. Some of them we are going to explain below:

Naïve bayes(NB)

Support vetorMachine(SVM)

Decision Tree(DT)

K-nearest neighbor(k-NN)

III. RESULT AND DISCUSSION

For the most part, to quantify the exhibition of data classification we utilize some predefined guidelines, for example, exactness, accuracy, review. Exactness is subject to two measure.

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

Precision

To get the correct estimation of precision, we partition the absolute number of appropriately grouped positive perception by the complete number of anticipated positive perception. High precision indicates that the perception arranged positive is undoubtedly sure.

Precision = tp/tp +fp

Recall

It is the proportion between the right classified positive perception to the complete number of positive perception high recall indicates that the class is properly arranged.

Recall = tp/tp + fn

Accuracy

In order to find the which model gives better result, then it is necessary to find the accuracy. Accuracy for any model can be given as:

Accuracy = tp + tn/tp+ fn +tn + tp

In this paper distinctive classifier have tried on same dataset in which some give best execution regarding precision, review and exactness. These are data which we have gotten from the twitter dataset. This data contain the 'gaganyaan' tweet. Here beneath are execution of some classifier.

Naïve bayes(NB):

```
NB accuracy:  0.6341463414634146
NB Precision:  0.6271929824561403
NB Recall:  0.89375
```

Naïve bayesclassifer is tested on our dataset. Generally it works on large datasetand it is fast. It the accuracy 63% and precision is 62% and it gives better performance in term of recall i.e, 89%.

Support vector machine (SVM)

```
SVM accuracy:  0.7896341463414634
SVM Precision:  0.8531468531468531
SVM Recall:  0.7625
```

Support vector machine classifier generally it works better in small dataset. It give accuracy of 78% and moving on to side of precision and recall 85% and 76% respectively.

Decision tree(DT):

```
DT accuracy:  0.7957317073170732
DT Precision:  0.8301886792452831
DT Recall:  0.825
```

this classifier is used to tested our dataset. It gives accuracy of 79% and it terms of precision it gives result 83% and recall 82%.

K-nearest neighbor(k-NN):

```
KNN accuracy:  0.7073170731707317
KNN Precission:  0.8557692307692307
KNN Recall:  0.55625
```

It is functions admirably in accuracy it gives result 70%. precision and recall 85% and 55% respectively. With the assistance of Tweepy API we gathered absolute of 1431 tweets from twitter and did examination on those tweets by utilizing some supervised learning classifier, for example, support vector machine, k-nearest neighbor,

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

naïve bayes, decision tree. with the help of these classifier we are able to find standard measure such as accuracy, precision and recall.

IV. CONCLUSION

In today's world, spacious amount of data is generated by various communication such as social media, organizations etc. these data may or may not be in structured form. Therefore to understand the polarity of data first we need to do the analysis of data. Opinion mining can be performed in various field such as marketing and customer feedback. large number of organizations are taking the valuable feedback of person and performing opinion mining on those data so that they could provide the better services to the customer and this data helps the organizations to enhance their future services. Furthermore, there are various scopes where we can perform the opinion mining such as sentence, paragraph, documents, sub sentences levels. In addition to this we took some classifiers such as support vector machine, naïve bayes, decision tree, K-nearest neighbor which performs best in terms of accuracy, precision, and recall. Out of these classifiers we conclude that DT performs best in finding accuracy of twitter dataset. It is best classifier on this dataset.

V. REFERENCES

[1] Gurkhe D., Pal N. and Rishit B. "Effective Sentiment Analysis of Social Media Datasets using Naïve Bayesian Classification." (2014).
[2] Bouazizi, M., Ohtsuki, T.: Multi-Class Sentiment Analysis in Twitter: What if Classification is Not the Answer. IEEE Access. 6, 64486-64502 (2018).
[3] Gautam, G., Yadav, D.: Sentiment analysis of twitter data using machine learning approaches and semantic analysis. 2014 Seventh International Conference on Contemporary Computing (IC3). (2014).
[4] Amolik, Akshay, et al. "Twitter sentiment analysis of movie reviews using machine learning techniques." International Journal of Engineering and Technology 7.6 (2016): 1- 7.
[5] Mukherjee S., Malu A., Balamurali A.R, Bhattacharyya P."TwiSent: A Multistage System for Analyzing Sentiment in Twitter".
[6] Davidov D., Tsur O., Rappoport A." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys".
[7] Neethu, M., Rajasree, R.: Sentiment analysis in twitter using machine learning techniques. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). (2013). PulkitGarg, HimanshuGarg, VirenderRanga "Sentiment Analysis of the Uri Terror Attack UsingTwitter" International Conference on Computing, Communication and Automation (ICCCA2017).
[8] Prof. SudarshanSirsat, Dr.Sujata Rao, Dr.BhartiWukkadada"Sentiment Analysis on Twitter Data forproduct evaluation" IOSR Journal of Engineering (IOSRJEN) ISSN (e): 2250-3021, ISSN (p): 2278-8719PP 22-25.(2019)
[9] HetuBhavsar, RichaManglani" Sentiment Analysis of Twitter Data using Python"International Research Journal of Engineering and Technology (IRJET) Mar 2019e-ISSN: 2395-0056 p-ISSN: 2395-0072
[10] "India announces first manned space mission". Bangalore: BBC News. ^ Press Trust of India (25 April 2012). "Spaceflight stuck due to budget: CAG". Times of India. New Delhi. Retrieved 11 June2013.
[11] Press Trust of India. "Human space flight mission off ISRO priority list". Retrieved 18 August 2013.