Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, No. 1, August : 2023

TWITTER ANALYSIS ON REAL TIME DATA

Pratik S Patil Department of Computer Engineering SKNSITS Lonavala,Pune Lonavala,Pune
pratik.chalisgaon@gmail.com
Suraj Ahire Department of Computer Engineering SKNSITS Lonavala,Pune Lonavala,Pune
surajahire.sknsits.comp@gmail.com
Prathamesh P Bhamanage Department of Computer Engineering SKNSITS Lonavala,Pune
prathameshbhamanage07@gmail.com
Durgesh P Jadhav Department of Computer Engineering SKNSITS Lonavala,Pune
durgeshjadhav.sknsits.comp@gmail.com

Abstract
In this era of growing social media users, Twitter has significantly large number of daily users who post their opinions in the form of tweets. This paper presents an idea of extracting sentiments out of the tweet and an approach towards classifying a tweet into positive, negative or neutral. This approach can be in many ways useful to any organization, who gets mentioned or tagged in a tweet. Generally the tweets being unstructured in format, first of all the tweet needs to be converted into the structured format. In this paper, tweets are resolved using pre-processing phase and access of tweets has been accomplished via libraries using Twitter API. The datasets need to be trained using algorithms in a way, such that, it becomes capable of testing the tweets and it releases the required sentiments out of the feeded tweets.

Keywords—Sentiment Analysis, Social Media, Tweets, Tweepy, Textblob, Pandas, Dataset, KNN, Naive Bayes.

I. INTRODUCTION

Twitter is a trending micro online journal service wherein users can justify their sentiments in the form of "tweets". These tweets sometimes express opinions about different topics. Sentiment analysis is the prediction of emotions in a word, sentences or corpus of documents. It is intended to serve as an application to understand the attitudes, opinions and emotions expressed within an online mention. Precisely, it is a paradigm of categorizing conversations into positive, negative or neutral labels. The greater part of the information that is accessible in interpersonal organizations is unstructured. Such unstructured information is basically some amount of the information present everywhere throughout the world. This makes it hard to examine and increase important judgment from such information. Notion examination or assessment mining is the significant system, which help in recognizing conclusions of individuals via web-based networking media information.

The objective of feeling investigation is recognizing content assumption extremity. Assumption examination could be taken as an order issue. According to [1-2], sentiment analysis is a process that isolates the content into positive, negative or neutral conclusion. Profound neural system and the Gaussian blend model is one of the hearty models for normal preparing language.

Conviction analysis is useful for consumers who are trying to research a product or service, or marketers researching public opinion of their company/product. However, doing the analysis of tweets that express human emotions isn't an easy job. A lot of challenges are involved in terms of tonality, polarity, lexicon and grammar of the Twitter sentiment analysis using bag of words tweets. They tend to be highly unstructured and non-grammatical and therefore it get's difficult to interpret their meanings.

II. SENTIMENT ANALYSIS

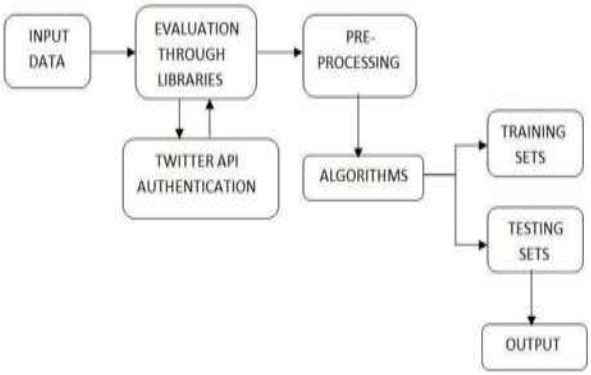
The region of concentrate that deciphers individuals' feelings, against a specific point, about any occasion and so on in content mining it is known as conclusion mining or assumption examination. It creates a huge issue zone. There are additionally different names and having various errands, e.g.,

notion investigation, conclusion extraction, feeling mining, assumption mining, influence examination, subjectivity examination, survey mining, and so on.
Twitter goes about as stubborn data bank with enormous measure of information accessible, utilized for conclusion analysis. Twitter is very convenient for research in light of the fact that there are enormous quantities of messages, many of which are freely accessible, and acquiring them is actually basic contrasted with scarping sites from the web.
Twitter information is gathered for investigation utilizing Twitter API. Two broadly utilized methodologies utilized for the equivalent are Machine Learning and Dictionary Based methodology. We are utilizing Dictionary Based methodology for dissecting the notions of information posted by various clients. At that point extremity arrangement of this information is done. For example Tweets gathered after examinations are grouped into three classes as Positive, Negative and Neutral.
A. Tools Available for Sentiment Analysis
1) Tweepy
Tweepy is used for accessing the twitter api. It is basically a python library, which is magnificient for the generation of automations and twitters bots.The StreamListener object in Tweepy monitors the tweets in real time and catches them.
2) Textblob
Textblob is an effective NLP (Natural Language Processing) library for python. It is built upon NLTK (Natural Language Toolkit) and can be used to perform various tasks ranging from conviction analysis to part-of-speech tagging and text- hierarchies to language translation.
3) Pandas
It is one of the data frames in python. Pandas offer ground- breaking, expressive and adaptable information structures that make information control and examination simple, among numerous different things.

III. TWITTER :

One of the most popular social networking websites, Twitter, came into existence on 21st of March, 2006. On this website users can read as well as send tweets. Tweets are basically a post on twitter with a limited size of character block. Twitter is a website that determines the confinement of substance of the assessment communicated on it. A tweet isn't just a basic instant message however it is a blend of content information and Meta information related with the tweet. These qualities are the highlights of tweets.
They communicates the substance of the tweet or what is that tweet about. The Metadata can be used to discover the area of the tweet. The Metadata of tweet are a few substances and spots. These substances incorporate client specifies hashtags, URLs, and media Users, Twitter user ID. RT represents retweet, '@' trailed by a client identifier report the client, and '#' trailed by a word portrays a hashtag.
A. Conviction Analysis of Twitter Data :
The prime purpose of twitter conviction's analysis is to classify various tweets into different sentiments category. Various approaches in this field regarding twitter emerge by training up a model and then checking its efficiency.
Challenging the tweets is not that easy as it seems to be.
Reasons behind this are as follows [4]:
1) Varying Language Origins
Different people from different cultures tweet using some words of their cultural origin, this words may be promotionary, slang etc.
2) Limited Character Block Size
With just 280 characters in scope, the amount of data content that can be recognized is very limited.
3) Use of hashtags

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, No. 1, August : 2023
Twitter provides hashtags to mention emotions, event etc, which requires separate processing than the actual word based tweet.
Conviction analysis is mainly composed of steps as shown in fig 1:



- Preprocessing

In [1, 4], this step involves the cleaning up of unstructured data from the tweets such as userid, blank spaces, numbers, hashtags etc. After this step only the main twitted text becomes available for analysis.

- Algorithms

Various algorithms such as KNN (K-nearest neighbors), Naive Bayes, SVM (Support Vector Machine) etc. can be used to train and test the datasets generated as a result of pre- processing stage. One among the above mentioned algorithm is explained below -

- K –nearest neighbors :

K Nearest Neighbor (KNN) is a basic, straightforward, flexible and one of the highest AI calculations. KNN calculation utilized for both order and relapse issues. KNN calculation is dependent on highlight similitude approach. The model structure determined from the dataset. This will be very helpful in practice where most of the real datasets do not follow mathematical theoretical assumptions. All training data used in the testing phase. This makes training faster and testing phase slower and costlier. Costly testing phase means time and memory. In the worst case, KNN needs more time to scan all data points and scanning all data points will require more memory for storing training data.
KNN is known to be basic and simple. KNN is a model based learning gathering. This calculation is likewise one of the languid learning strategies. KNN is done via scanning for the gathering of K questions in the nearest preparing information (comparative) to objects in new information or information testing For the most part the Euclidean separation equation is utilized to characterize the separation between the two i.e. preparing items and testing . Mechanism of KNN is as follows -

a) Initialize the K –

A little estimation of k implies that clamor will impact the outcome and a huge worth make it computationally costly.
Typically pick as an odd number if the quantity of classes is 2 and another basic way to deal with select k is to set k=sqrt (n).
Fig. 1.

- Accept input

Take any of the tweets. This tweet may be a combination of various sentiments, tags and hashtags.

- Evaluation through libraries

Libraries such as Tweepy, can be used to perform twitter API authentication and to generate the access tokens for testing the functionality and sentiments of tweets.

b) Compute the distance between input sample & trained sets –

Based on the number of nearest neighbor's i.e. on the value of k distance, has to be computed between the input data and each and every trained set of neighbor, wherein we further sort that distance so that we can generate the nearest distance between the sample and the neighbor.

c) Take the nearest sample –

Depending upon the distances estimated between the input and trained sets, select the nearest sample i.e k-nearest neighbor.

d) Apply the majority –

Apply the majority of such input samples with minimum distances to obtain the maximum efficiency of KNN algorithm.

e) Output –

According to [2], accuracy of prediction for KNN is 99.6456%.

• Naive Bayes :

The Naive Bayes classifier [8] is the least complex and most usually utilized classifier as shown in fig.2. Guileless Bayes characterization model processes the back likelihood of a category, in view of the conveyance of the words in the report. It depends on very basic portrayal of archive as Bag of words. Naive Bayes classifier is effectively utilized in different applications, for example, spam sifting, content characterization, supposition, examination and recommender frameworks. It utilizes Bayes hypothesis of likelihood for forecast of unknown categories. It utilizes Bayes Theorem to anticipate the likelihood that a given list of capabilities has a place with a specific name. For twitter opinion examination bigrams from the twitter information are utilized as highlights on Naive Bayes. It Classifies tweets into positive and negative marks.

It's established formulation is as follows [7]: P (h/D) = P (D/h) P (h) / P (D)

Where,

P (h/D) = The afterwards probability of h given the data D. P (D/h) = The afterwards probability of D given the data h. P (h) = The before probability of event h being true.

P (D) = The before probability of event D being true.

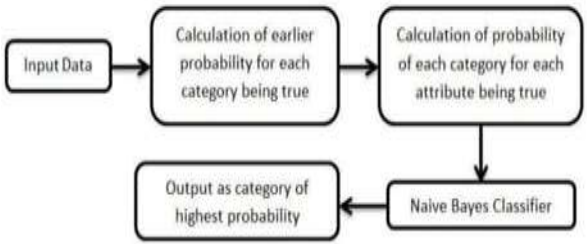


Fig. 2.

• Training and Testing the Dataset

Training of datasets means to create a model and train or fit that models within the parameters .Testing of datasets is used to check the quality and performance. The analysis of tweet is done by splitting into words. Intensity of words is determined by the algorithms or libraries. Depending on the intensity, positive and negative words get separated. If the intensity of positive words is high then, the tweet is positive. Sometimes the tweet may be neutral. In that case tweet is neither positive nor negative.

Hence the desired conviction analysis for any of the tweet can be performed.

IV. CONCLUSION

Twitter is a huge platform and source of improperly structured and sentiment datasets that can be analyzed to produce trending emotions and many more. In Twitter sentiment analysis we inspect or mine each and every element of the tweet. This paper explains various steps involved in analysis of twitter sentiments along with the various tools that are used to perform twitter sentiment analysis. Amongst the various algorithms available, KNN algorithm is used to increase the efficiency of sentiment analysis whereas Naive Bayes for simple and efficient sentiment analysis by classifying the tweets as either positive, negative or zero. Whenever a tweet is fed for sentiment analysis, it goes through various phases of sentiment analysis. For analyzing a tweet it is very necessary to know the

morph and elements of the tweet. Each of these components and phases of sentiment analysis are briefly described in this review paper.

REFERENCES

1) A. Sarlan, C. Nadam and S. Basri, "Twitter sentiment analysis," Proceedings of the 6th International Conference on Information Technology and Multimedia, 2014.
2) C. Kariya and P. Khodke, "Twitter Sentiment Analysis," 2020 International Conference for Emerging Technology (INCET), 2020..
3) S. A. El Rahman, F. A. AlOtaibi and W. A. AlShehri, "Sentiment Analysis of Twitter Data," 2019 International Conference on Computer and Information Sciences (ICCIS), 2019.
4) V. Pandya, A. Somthankar, S. S. Shrivastava and M. Patil, "Twitter Sentiment Analysis using Machine Learning and Deep Learning Techniques," 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4), 2021.
5) A. Ikram, M. Kumar and G. Munjal, "Twitter Sentiment Analysis using Machine Learning," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2022.
6) J. F. Raisa, M. Ulfat, A. Al Mueed and S. M. S. Reza, "A Review on Twitter Sentiment Analysis Approaches," 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 2021.
7) A. Roy and M. Ojha, "Twitter sentiment analysis using deep learning models," 2020 IEEE 17th India Council International Conference (INDICON), 2020.
8) R. Wagh and P. Punde, "Survey on Sentiment Analysis using Twitter Dataset," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018.
9) V. Prakruthi, D. Sindhu and D. S. Anupama Kumar, "Real Time Sentiment Analysis Of Twitter Posts," 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), 2018.