iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.43233

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Opinion Mining from Social Media

Mirza Ayaan Beg[1], Dushyant Jadon[2]
[1,2]Department of Computer Science and Engineering, Raj Kumar Goel Institute of Technology, Ghaziabad, India

***Abstract*— Twitter is a platform extensively utilized by humans to express their reviews and display opinions on one-of-a-kind activities. Opinion analysis is an approach to investigate statistics and retrieve opinion that it embodies. Twitter opinion analysis is ansoftware of opinion evaluation on information from Twitter (tweets), which will extract opinions conveyed by way of the consumer. In the past many years, the research on this field has always grown. The reason behind that is the difficult layout of the tweets which makes the processing difficult. The tweet format may be very small which generates a whole new measurement of problems like use of slang, abbreviations etc. In this paper, we intention to study some papers concerning research in opinion analysis on Twitter, describing the methodologies followed and fashions carried out, at the side of describing a generalized Python based method.**

Keywords*— Opinion analysis, Machine Learning, Natural Language Processing, Python

I. INTRODUCTION

Twitter has emerged as a main micro-running a blog internet site, having over a hundred million customers generating over 500 million tweets every day. With such massive target audience, Twitter has continuously attracted customers to carry their reviews and attitude approximately any problem, brand, employer or any other subject matter of hobby. Due to this purpose, Twitter is used as an informative source by many agencies, institutions and groups.

On Twitter, users are allowed to percentage their reviews in the shape of tweets, the usage of best a hundred and forty characters. This results in human beings compacting their statements by the usage of slang, abbreviations, emoticons, brief bureaucracy etc. Along with this, humans deliver their critiques by way of the use of sarcasm and polysemy. Hence it's miles justified to time period the Twitter language as unstructured. In order to extract opinion from tweets, opinion analysis is used. The outcomes from this may be used in many regions like reading and tracking adjustments of opinion with an occasion, opinions concerning a specific logo or release of a precise product, analyzing public view of government policies and so forth. A lot of studies has been carried out on Twitter records that allows you to classify the tweets and analyze the outcomes. In this paper we aim to check of some researches in this area and examine how to carry out opinion analysis on Twitter statistics the use of Python. The scope of this paper is limited to that of the device mastering fashions and we display the evaluation of efficiencies of those models with each other.

II. ABOUT OPINION ANALYSIS

Opinion evaluation is a method of deriving opinion of a unique declaration or sentence. It's a class technique which derives opinion from the tweets and formulates an opinion and on the idea of which, opinion category is done. Opinions are subjective to the topic of interest. We are required to formulate that what sort of functions will determine for the opinion it embodies. In the programming model, opinion we seek advice from, is class of entities that the person performing opinion evaluation desires to locate within the tweets. The size of the opinion elegance is critical aspect in determining the efficiency of the model. For example, we are able to have -magnificence tweet opinion class (superb and terrible) or three elegance tweet opinion category (fine, poor and neutral). Opinion analysis approaches may be widely categorized in two lessons – lexicon based and device studying primarily based. Lexicon primarily based approach is unsupervised because it proposes to carry out analysis the usage of lexicons and a scoring method to evaluate opinions. Whereas system getting to know method includes use of characteristic extraction and education the model the usage of feature set and a few dataset. The simple steps for performing opinion evaluation consists of statistics series, pre-processing of facts, function extraction, selecting baseline features, opinion detection and performing class both the usage of easy computation or else device studying techniques.

The intention whilst appearing opinion evaluation on tweets is basically to categorise the tweets in special opinion instructions appropriately. In this subject of research, numerous processes have developed, which endorse strategies to educate a model after which test it to test its efficiency. Performing opinion evaluation is challenging on Twitter facts, as we stated earlier.

Here we define the motives for this:

- Limited tweet size: with just a hundred and forty characters in hand, compact statements are generated, which outcomes sparse set of features.
- Use of slang: those phrases are distinctive from English phrases and it is able to make a method old due to the evolutionary use of slangs.

- Twitter capabilities: it lets in the usage of hashtags, user reference and URLs. These require specific processing than other words.
- User variety: the customers specific their opinions in a type of approaches, a few the use of one of a kind language in among, even as others the use of repeated words or symbols to convey an emotion All these troubles are required to be faced in the preprocessing phase.

Apart from these, we are facing problems in function extraction with much less features in hand and decreasing the dimensionality of functions.

III. METHODOLOGY

In order to perform opinion analysis, we are required to gather information from the favored supply (here Twitter). This information undergoes various steps of pre-processing which makes it extra gadget practical than its preceding form.

A. Tweet Collection

Tweet collection entails collecting applicable tweets about the precise region of interest. The tweets are accrued the usage of Twitter's streaming API [1], [3], or any other mining device (for example WEKA [2]), for the preferred time period of evaluation. The format of the retrieved textual content is converted as according to comfort (for instance JSON in case of [3], [5]). The dataset gathered is imperative for the performance of the model. The division of dataset into education and testing sets is also a finding out element for the efficiency of the model. The education set is the principle factor upon which the outcomes relies upon.

B. Pre-processing of Tweets

The preprocessing of the information is a very essential step because it decides the performance of the opposite steps down in line. It entails syntactical correction of the tweets as preferred. The steps worried should goal for making the records extra machine readable with a view to lessen ambiguity in function extraction.

Below are a few steps used for pre-processing of tweets -

1) Removal of re-tweets.
2) *Converting Top Case to Decrease Case:* In case we're using case touchy analysis, we'd take two occurrence of equal words as distinct because of their sentence case. It important for an powerful evaluation no longer to provide such misgivings to the version.
3) *Stop Word Removal:* Stop words that don't affect the that means of the tweet are eliminated (for example and, or, still and many others.). [3] uses WEKA gadget getting to know package for this reason, which exams each word from the textual content towards a dictionary ([3], [5]).
4) *Twitter Feature Removal:* User names and URLs are not essential from the angle of destiny processing, therefore their presence is futile. All usernames and URLs are transformed to well-known tags [3] or removed [5]
5) *Stemming:* Replacing words with their roots, lowering exceptional forms of words with comparable meanings [3]. This facilitates in reducing the dimensionality of the function set.
6) *Special Person and Digit Removal:* Digits and unique characters don't carry any opinion. Sometimes they're combined with phrases, hence their removal can assist in associating two phrases that were in any other case taken into consideration one of a kind.
7) Creating a dictionary to eliminate unwanted words and punctuation marks from the text [5].
8) Expansion of slangs and abbreviations [5].
9) Spelling correction [5].
10) Generating a dictionary for phrases which are vital [7] or for emoticons [2].
11) *Part of Speech (POS) Tagging:* It assigns tag to every phrase in text and classifies a word to a particular class like noun, verb, adjective and so on. POS taggers are green for specific function extraction.

C. Three Feature Extraction

A function is a bit of facts that can be used as a function that can assist in fixing a trouble (like prediction [11]). The excellent and amount of functions could be very critical as they're crucial for the outcomes generated by means of the selected model. Selection of beneficial words from tweets is feature extraction.

1) *Unigram Capabilities:* one phrase is considered at a time and determined whether or not it is capable of being a function.
2) *N-gram Functions :* more than one word is taken into consideration at a time.
3) *External Lexicon* : use of listing of words with predefined advantageous or terrible opinion.

Frequency evaluation is a way to acquire capabilities with highest frequencies utilized in [1]. Further, they removed a few of them due to the presence of words with comparable opinion (for example satisfied, joy, ecstatic and so forth.) and created a collection of these phrases. Along with this affinity analysis is done,
which focuses on higher order n-grams in tweet function illustration. Barnaghi et al [3], use unigrams and bigrams and practice Term Frequency Inverse Document Frequency (TF-IDF) to locate the weight of a specific characteristic in a text and subsequently filter the features having the maximum weight. The TF-IDF is a very green method and is widely utilized in text class and statistics mining.
Bouazizi et al [4], suggest an technique had been they don't just rely on vocabulary used but additionally the expressions and sentence shape used in special conditions. They labeled features into 4 instructions: opinion based totally functions, punctuation and syntax based capabilities, unigram based capabilities and sample based totally features.
The paintings of [5] is a bit extraordinary as they don't cognizance on a particular subject matter or occasion however endorse to find trending topics in a place. The capabilities extracted are divided in classes:
Common Features and Tweet Specific Features. The former is aggregate of not unusual opinion words even as the later includes @-community features, user opinion functions and emoticons. Based at the publish time of each user, characteristic vector is built.

IV. TWITTER OPINION ANALYSIS WITH PYTHON

A. *Python:* Python is a high level, interpreted programming language, created by way of Guido van Rossum. The language is very famous for its code readability and compact line of codes. It uses white area inundation to delimit blocks. Python affords a large popular library which may be used for diverse packages as an example herbal language processing, machine learning, data evaluation and so forth.
B. It is desired for complicated tasks, due to its simplicity, diverse range of features and its dynamic nature.
C. *Natural Language Processing (NLTK):* Natural Language toolkit (NLTK) is a library in python, which affords the bottom for textual content processing and classification. Operations along with tokenization, tagging, filtering, textual content manipulation may be executed with using NLTK. The NLTK library also embodies various trainable classifiers (instance – Naïve Bayes Classifier).
D. NLTK library is used for growing a bag-of phrases version, which is a sort of unigram version for text. In this model, the number of occurrences of every word is counted. The facts acquired can be used for schooling classifier models. The opinion of the complete tweets is computed by using assigning subjectivity score to every word the use of a opinion lexicon.
E. *SCIKIT-LEARN:* The Scikit-research mission began as scikits.Learn, a Google Summer Code task by using David Cournapeau. It is a effective library that gives many system studying class algorithms, efficient tools for statistics mining and facts analysis.

Below are numerous functions that may be done using this library:
- *Classification:* Identifying the category to which a particular item belongs.
- *Regression:* Predicting a continuous-valued attribute related to an object.
- *Clustering:* Automatic grouping of comparable objects into sets.
- *Dimension Reduction:* Reducing the range of random variables underneath consideration.
- *Model Selection:* Comparing, validating and deciding on parameters and fashions.
- *Preprocessing:* Feature extraction and normalization with a view to remodel enter facts for use with system mastering algorithm.

In order to ork with scikit-analyze, we are required to install NumPy on the gadget.

F. *NumPy:* NumPy is the fundamental bundle for scientific computing with Python. It offers a high-overall performance multidimensional array item, and equipment for working with those arrays.

It carries amongst different things:
- A effective N-dimensional array item
- Sophisticated (broadcasting) capabilities
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier remodel, and random range capabilities.

Setting Up Environment for Opinion Analysis Using Python the following additives are required to be downloaded and mounted well:
- Download and set up Python 2.6 or above in a favored place.
- Download and installation NumPy.

- Download and install NLTK library.
- Download and install Scikit-study library

G. Data Collection

We have two alternatives to collect records for opinion evaluation.
First is to use Tweepy - client for Twitter Application Programming Interface (API).
It may be mounted using pip command: pip install tweepy To fetch tweets from the Twitter API one needs to register an App through their Twitter account.
After that the subsequent steps are accomplished:

- Open https://apps.Twitter.Com/ and click on button – 'Create New App'.
- Fill the details requested.
- When the App is created, the page will be robotically loaded
- Open the 'Keys and Access Tokens' tab. Copy 'Consumer Key', 'Consumer Secret', 'Access
- token' and 'Access Token Secret'.

The keys copied are then inserted into the code, which enables in dynamic series of tweets every time we run it.The other choice is to acquire information non-dynamically using the current records provided via websites (like kaggle.Com) and shop the information into anything layout we require (as an example JSON, csv and many others.).
The former technique is slow in nature as it performs tweet collection every time we begin this system. The latter technique won't offer us with the first-rate of tweets we require.
To clear up this we are able to placed the code for tweet series in specific module in a way that it doesn't perform each time we run the venture

H. Pre-Processing in Python

The pre-processing in Python is simple to perform due to capabilities supplied by means of the standard library. Some of the stairs are given under:

- Converting all upper case letters to decrease case.
- Removing URLs: Filtering of URLs may be executed with the assist of everydayftp)://[a-zA-Z0-9./]+.
- Removing Handles (User Reference): Handles can be removed the usage of regular expression - @(w+) .
- Removing hashtags: Hashtags can be removed using everyday expression - #(w+).
- Removing emoticons: We can use emoticon dictionary to clear out the emoticons or to store the occurrence of them in a unique record.
- Removing repeated characters

I. Feature Extraction

Various methodologies for extracting capabilities are to be had in the cutting-edge. Term frequency-Inverse Document frequency is an effective method. TF-IDF is a numerical statistic that displays the value of a phrase for the complete file (right here,tweet). Scikit-learn gives vectorizers that translate input documents into vectors of functions. We can use library function TfidfVectorizer(), the use of which we will offer parameters for the kind of features we need to keep by using bringing up the minimal frequency of appropriate functions.

V. APPLICATIONS

A. Commerce:

Companies can employ this research for collecting public opinion related to their logo and merchandise. From the business enterprise's attitude the survey of target audience is vital for making out the scores of their products. Hence Twitter can serve as a terrific platform for facts series and analysis to determine purchaser delight.

B. Politics:

Majority of tweets on Twitter are related to politics. Due to Twitter's large use, many politicians are also aiming to connect to people thru it. People put up their help or confrontation in the direction of authorities guidelines, actions, elections, debates etc. Hence studying statistics from it could help is in figuring out public view.

C. Sports Events:

Sports contain many events, championships, gatherings and some controversies too. Many humans are enthusiastic sports fans and follow their favorite players present on Twitter.

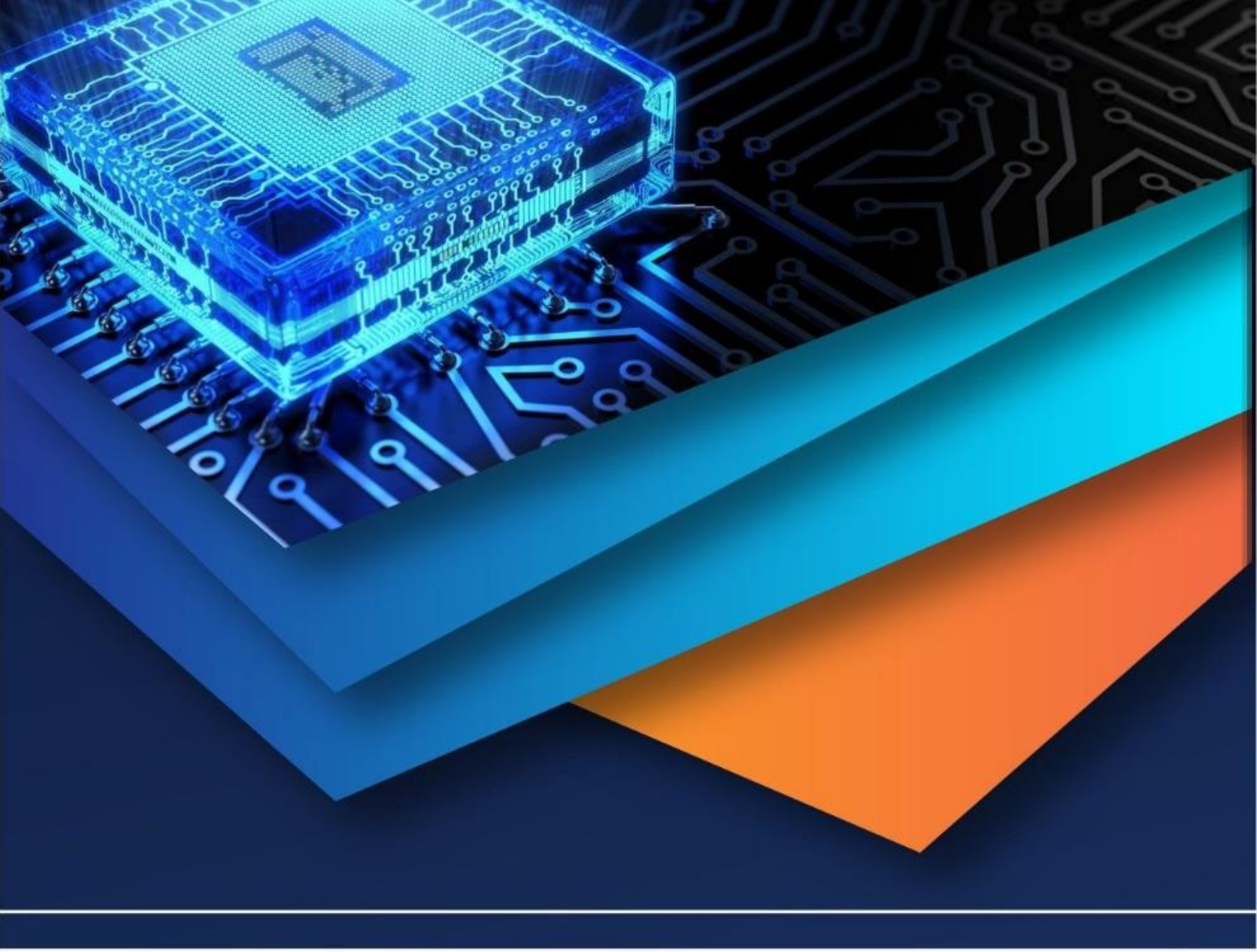
D. These humans regularly tweet approximately distinctive sports activities related activities.We can use the statistics to acquire public view of a participant's motion, crew's overall performance, authentic decisions and so on.

VI. CONCLUSION

Twitter opinion analysis comes beneath the class of textual content and opinion mining. It specializes in studying the emotions of the tweets and feeding the facts to a device gaining knowledge of version in order to teach it after which test its accuracy, in order that we will use this model for destiny use consistent with the outcomes. It incorporates of steps like information series, text pre-processing, opinion detection, opinion class, schooling and trying out the version. This studies subject matter has advanced during the last decade with fashions attaining the performance of just about 85%-90%. But it still lacks the dimension of variety in the facts. Along with this it has a whole lot of application issues with the slang used and the short sorts of words. Many analyzers don't carry out well while the wide variety of lessons are accelerated. Also it's still no longer tested that how accurate the model might be for subjects different than the one in attention. Hence opinion evaluation has a very shiny scope of development in future.

REFERENCES

[1] David Zimbra, M. Ghiassi and Sean Lee, "*Brand-Related Twitter Sentiment Analysis using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks*", IEEE 1530-1605, 2016
[2] Varsha Sahayak, Vijaya Shete and Apashabi Pathan, "*Sentiment Analysis on Twitter Data*", (IJIRAE) ISSN: 2349-2163, January 2015.
[3] Peiman Barnaghi, John G. Breslin and Parsa Ghaffari, "*Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment*", 2016 IEEE Second International Conference on Big Data Computing Service and Applications.
[4] Mondher Bouazizi and Tomoaki Ohtsuki, "*Sentiment Analysis: from Binary to Multi-Class Classification*", IEEE ICC 2016 SAC Social Networking, ISBN 978-1- 4799-6664-6.
[5] Nehal Mamgain, Ekta Mehta, Ankush Mittal and Gaurav Bhatt, "Sentiment Analysis of Top Colleges in India Using Twitter Data", (IEEE) ISBN -978-1-5090-0082-1, 2016.
[6] Halima Banu S and S Chitrakala, "*Trending Topic Analysis Using Novel Sub Topic Detection Model*", (IEEE) ISBN- 978-1-4673-9745-2, 2016.
[7] Shi Yuan, Junjie Wu, Lihong Wang and Qing Wang, "*A Hybrid Method for Multi-class Sentiment Analysis of Micro-blogs*", ISBN- 978-1-5090-2842-9, 2016.
[8] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau, "*Sentiment Analysis of Twitter Data*" Proceedings of the Workshop on Language in Social Media (LSM 2011), 2011
[9] Neethu M S and Rajasree R, "*Sentiment Analysis in Twitter using Machine Learning Techniques",* IEEE – 31661, 4th ICCCNT 2013.
[10] Aliza Sarlan, Chayanit Nadam and Shuib Basri, "*Twitter Sentiment Analysis*", 2014 International Conference on Information Technology and Multimedia (ICIMU), Putrajaya, Malaysia November 18 – 20, 2014.
[11] Feature engineering, Wikipedia 2017, https://en.wikipedia.org/wiki/Feature_engineering.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY