IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

LIVE TWITTER DATA ANALYSIS AND VISUALIZATION

[1]Dr.Vinayak A Bharadi, [2]Janhavi Lele, [3]Antara Phadnis , [4]Asawari Sawant , [5]Aakanksha Birje

[1]HOD and Professor , [2]Student, [3]Student, [4]Student, [5]Student

Department of Information Technology

Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India

Abstract: In this paper, we explain the process of storing, preparing, and analyzing Twitter streaming data, then we examine the methods and tools available in python to visualize the analyzed data. We believe that using social networks and microblogs to efficiently analyze massive real-time data about the product, services, or global events is crucial for better decisions. The popularity of Twitter as an information source has led to the development of applications and research in a wide range of fields. Humanitarian Assistance and Disaster Relief is one domain where information from Twitter is used to provide situational awareness of a crisis situation. Managerial decisions, stock prediction, and improving traffic prediction are some examples of the use cases getting favorites among researchers.

I. Introduction

Social networks and microblogging sites have now become an unparalleled source of unstructured data. This data is immense 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 such as Twitter on which people post real-time messages about their opinions on a variety of topics, discuss current issues and affairs or complain and express their sentiments about products they use in daily life. In order to understand the general sentiment about their product or service, many firms analyze massive amounts of information. It is quite common for proactive companies to monitor user reactions on social microblogs and respond to the user.

This process provides on the spot solution to make the user experience better but at a large scale, it's a very tedious and time-consuming task. Therefore, the challenge here is to build solutions that analyze data sources from various microblogging and social networks to make important decisions about long-term product design and service implementation. In this paper, we take one such social network called Twitter to analyze and visualize various important metrics related to an event, product, or service.

I.I LITERATURE SURVEY

In the world of social analytics on the Web, sentiment detection and classification have become the latest fads. It is not trivial to distill the voice of the public to gain insight into targeted information and reviews for healthcare, finance, media, consumer markets, and government. Due to the increased size, subjectivity, and diversity of social web-data, the information has become more vague, uncertain, and imprecise. Soft computing techniques have been used to handle such fuzziness in practical applications. This work is a study to understand the feasibility, relevance, and scope of this alliance of using Soft computing techniques for sentiment analysis on Twitter. Our systematic literature review identifies research gaps defining the future prospects of this coupling by analyzing and exploring the efforts and trends in a well-structured manner. The contribution of this paper is significant because the primary focus is to study and evaluate the use of soft computing techniques for sentiment analysis on Twitter, secondly, as compared to the previous reviews we adopt a systematic approach to identify, gather empirical evidence, interpret results, critically analyze, and integrate the findings of all relevant high-quality studies to address specific research questions related to the defined research domain.

2. DATA COLLECTION

For programmatic access to Twitter data, we have to register an app on Twitter developers website for authentication, and then we can use Twitter API to access the data.

2.1. APP REGISTRATION

Creating a new Twitter app at https://apps.twitter.com/ is the first step to registering the Twitter app. We will receive the consumer_key and consumer_secret_key upon registration. Next, we will get access_token and access_token_secret from the configuration page of the app, which will be used to access Twitter on behalf of our application. These authentication tokens must be kept private as they can be misused. Ideally, these tokens should be stored in a separate config file.

2.2. DATA ACCESS

Twitter provides REST APIs for connecting to their service. We will use one python library to access the Twitter REST APIs called Tweepy. It provides the wrapper methods to easily access Twitter REST API.
To install Tweepy we can use the below command.

```
pip install tweepy
```

We need to use the OAuth interface to authorize our app to access Twitter on our behalf. The below code will use tweepy OAuthHandler method and our configuration tokens to provide access to Twitter.

2.3. STORING DATA

The next step is to access all tweet data from the personal profile and store it in a JSON file for use in our analysis. Using Tweepy, you can iterate over all tweets and save them as JSON files with a simple cursor interface.

3. PREPARING DATA

Prior to analyzing Twitter data, it is important to understand the tweet’s structure and pre-process it by removing non-useful terms known as stopwords. Data Pre-processing is a very important step in data analytics. By definition, Preprocessing means taking in the data and preparing the data for optimal output considering our requirements. To preprocess the tweets we need to analyze their different parts and understand the structure of the single tweet.

3.1.THE ANATOMY OF THE TWEET

Twitter APIs return Tweets encoded in JavaScript Object Notation (JSON). With named attributes and associated values, JSON is based on key-value pairs. These attributes with their state are used to describe objects.

Tweets and Users are served as JSON at Twitter. Attributes that describe an object are encapsulated in these objects. Tweets have an author, a message, a unique ID, a timestamp, and sometimes geo metadata shared by the user. Twitter users have names, IDs, followers, and bios, most often.

With each Tweet, "entity" objects are generated, which are arrays of common Tweet contents such as hashtags, mentions, media, and links. The JSON payload can also contain metadata such as the URL and title of the webpage if there are links.

A single tweet contains a lot of information related to users, the text, created date of the tweet, the location of the tweet, and many more fields.

We will use some of this fields to complete the analysis. The key fields of a single tweet are as follows:
text: text of the tweet,
lang: acronym of the tweet language like ‘en’,
created_date: date of creation of the tweet,
favorite_count: number of favorites of the tweet,
retweet_count: retweets of the tweet,
place, geolocation: location information, if available,
user: the full profile of user,
entities: list of entities like url’s, @mentions, #hashtags

Using these data, we can look at who is most favorited/retweeted, who is discussing with who, and what are the most popular hashtags. Our main interest is the tweet's content, which is represented by the field text.

3.2.REMOVING STOP-WORDS

During the pre-processing stages, Stop-word removal is an important step. These are most popular and common words of any language. In spite of their importance in the language, they don't convey a particular meaning when taken out of context. Articles, conjunctions, adverbs, and some adverbs are commonly referred to as stop words. Some libraries provide default stop-words for different languages. NLTK library provides default stop-words for English language.

An array of custom stop words included in this analysis are as below:
[‘The', 'what', 'What', 'You', 'Your', 'A', 'new', 'https', 'Hi', 'We', 'My', 'Now', 'please', ‘get’]
This list of stop words does not add any information in data analysis where we are looking for term frequencies. These words are heavily used in English language and we might consider them most frequent terms.

We should also be careful with all the punctuation marks and with terms like RT (used for retweets) and via (used for mentioning the original author of an article or a retweet), which are not in the default stop-word list.

4. ANALYZING THE DATA

After preprocessing the text data we can now proceed with different analysis objectives.

4.1.TERM FREQUENCIES

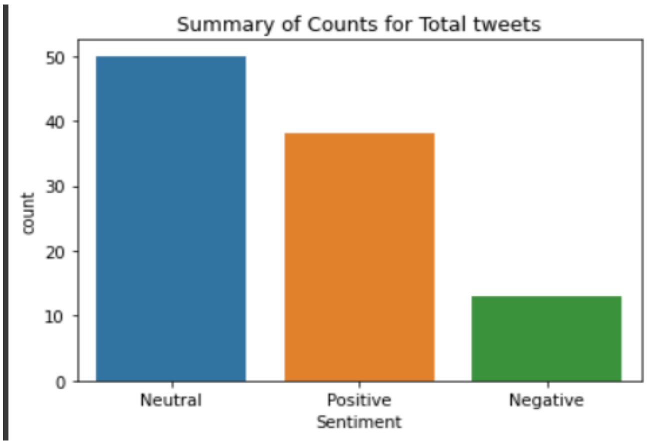
The simplest step in data analysis is Counting the frequencies of a term. By this, we can analyze for a particular user what he frequently tweets about. One use case of term frequencies is that advertisement companies can provide targeted ads based on the user's term frequencies. It will have more possible that user clicks or visit the promoted website. The below code can be an example of counting all the frequencies of all the terms.
terms_only = [term for term in preprocess (tweet['text']) if term not in stop and not term.startswith(('@', ‘#'))]
In the above code snippet we are listing out all the terms which are in preprocessed tweet text if that term is not in stop-words array stop it doesn’t starts with @ or #. To count the occurrences of the terms, we can use python's counter() collection.

5. ANALYZING STREAMING DATA

Analyzing the streaming data is very important as it allows us to make real-time decisions on the basis of real-time data. Streaming data analysis can be large in terms of data being generated so analysis of streaming data requires heavy computation resources. In streaming data analysis we will use the different analysis use cases for data which is being generated live.

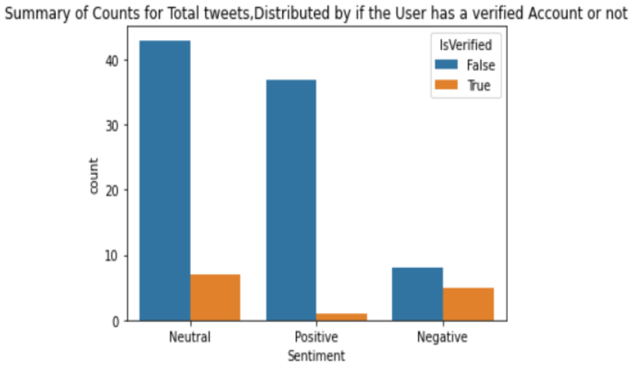
6. VISUALIZATIONS

Visualizations helps represent the analyzed data to make decisions effectively.
As a result of our focus on sentiment analysis of the tweets, we are going to display bar graphs that explain the sentiment behind the tweets.

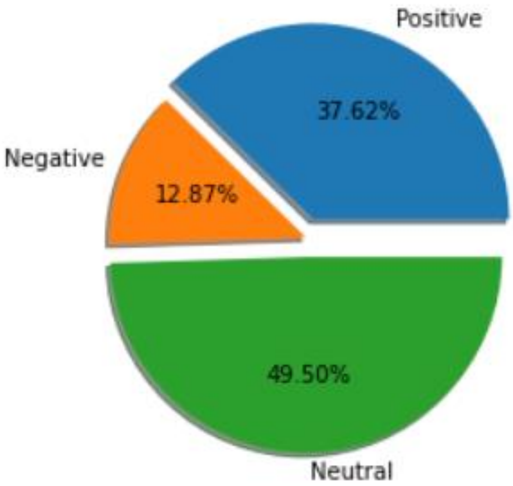


As we can see sentiment analysis is done in three categories: Positive, Negative, Neutral; Neutral has the highest range.

A summary is also provided based on whether the user's account is verified or not:



The collected data indicates that most accounts are not verified.
Additionally, we have visualized the data using a Pie-chart:



7. CONCLUSION

The purpose of this paper is to introduce the very basics of Twitter data analysis. We explained how to authenticate Twitter apps using OAuth and Tweepy. Then, we discussed how to collect historical and streaming data. We then preprocessed the data. In the final step, we tried to execute a number of use cases to analyze the stored data. We represented the results of analyzing the sentiments of the Tweets. Then we created bar graphs showing the intensity of sentiments. Since this is an introductory paper, the focus is on using Python to analyze Twitter data sentiment. Our future work will focus on representing more advanced data analysis on trending topics considering hashtags and decision making more accurately

8. REFERENCES

1.https://www.researchgate.net/profile/Vivek-Wisdom/publication/308371781_An_introduction_to_Twitter_Data_Analysis_in_Python/links/57e24cf708ae1f0b4d95b409/An-introduction-to-Twitter-Data-Analysis-in-Python.pdf
2.https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.5107
3.https://blog.devgenius.io/twitter-sentiment-analysis-with-traditional-machine-learning-algorithms-vs-deep-learning-b5fb7a4d8b00
4. A. Agarwal, B. Xie, I Vovsha, O. Rambow, R. Passonneau “Sentiment Analysis of Twitter Data” In the proceedings of Workshop on Language in Social Media, ACL, 2011