TWITTER DATA COLLECTION AND DASHBOARD

Report

ABSTRACT

To collect user data from twitter and create a basic dashboard for the client to showcase his activities.

Kalpesh Bhurabhai Odedra

Data Mining and Techniques

Table of Contents

Prepared By:
Kalpesh Odedra

Business Problem and Objective

Being a data analyst in a firm/company is very similar to being a data solutions consultant. Regularly, we must meet proper authorities and individuals who are facing a business issue and create solutions that can help them solve their problems. One of the best ways to reach the root cause of a problem was developed by Toyota Production System. Under this method repeatedly asking "Why" to a request leads to the root cause of the problem (Serrat, 2017).

For my research project, I chose to act as a consultant and my client is one of my friends VirenKumar Lathiya, a social media marketer. In his day to day, Viren has to stay active on social media especially Twitter. He has to post relevant tweets regularly to create a portfolio for his future interviews where his social media activity will be analyzed to determine the validity of his claims and his ability during interviews.

Although twitter analytics is a very powerful tool, Viren would like to collect all of his data for future use and create a dashboard that tracks his activity.

The 5 Why technique

Me: Why do you want a Twitter dashboard?

Client: Because I would like to showcase a more data-driven presentation during my interview.

Me: Why not use Twitter analytics?

Client: Though Twitter analytics is very powerful, I would like to complement it with a tool where I have more control over the type of analysis.

Me: Why would you need additional analysis?

Client: On top of usual metrics, I would like to show my hourly activity, the tone, and the feeling I use in my tweets to further showcase my ability to microblog.

Me: Why Twitter more than any other platform?

Client: In my line of business Twitter is a big platform and a person has to be active on Twitter every day.

Me: Why would you like to showcase your activity in an interview?

Client: I feel like this would be an excellent way to stand out during interviews.

Objective

- To collect twitter data going as far as possible for the client: VirenKumar Lathiya, username: @viren_lathiya.
- Create a basic dashboard of Twitter data.
- Add sentiment analysis of past tweets to the dashboard.
- Provide recommendations for improvement.

Research Plan

Step 1: Find the best tool that can collect Twitter data.

Twitter has its API where people can connect to it and collect Twitter data. One needs authorization in the form of auth key and consumer key to access tweeter data. Although tweepy is a good tool it has its limitations. Free access only gives data ranging as far as 7 days (tweepy, 2021).

For this activity, we need a tool that can give us data from as far back as the client started using Twitter. Therefore, for this activity scraping Twitter data would be the best course of action.

One of the most popular twitter scraping tools is "Twint". "Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles. Twint utilizes Twitter's search operators to let you scrape Tweets from specific users, scrape Tweets relating to certain topics, hashtags & trends" (Project Twint, 2021)

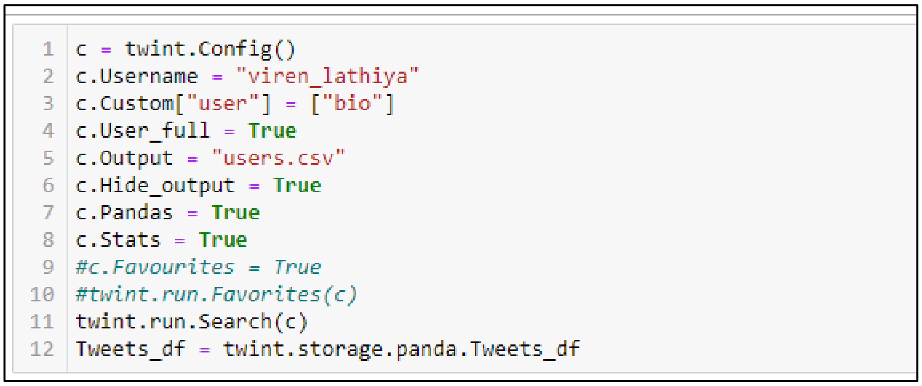
Step 2: Find methods to analyze the tone of the tweets

To analyze the tone of the tweets, sentiment analysis is a good approach. Textblob is a popular python library for text analysis. It provides functions for Tokenization and Sentiment analysis (Venkateswarlu Bonta, 2019).

The functions of sentiment analysis can create scores based on the words used in the tweet. This score can be used to analyze the sentiment of the tweet.

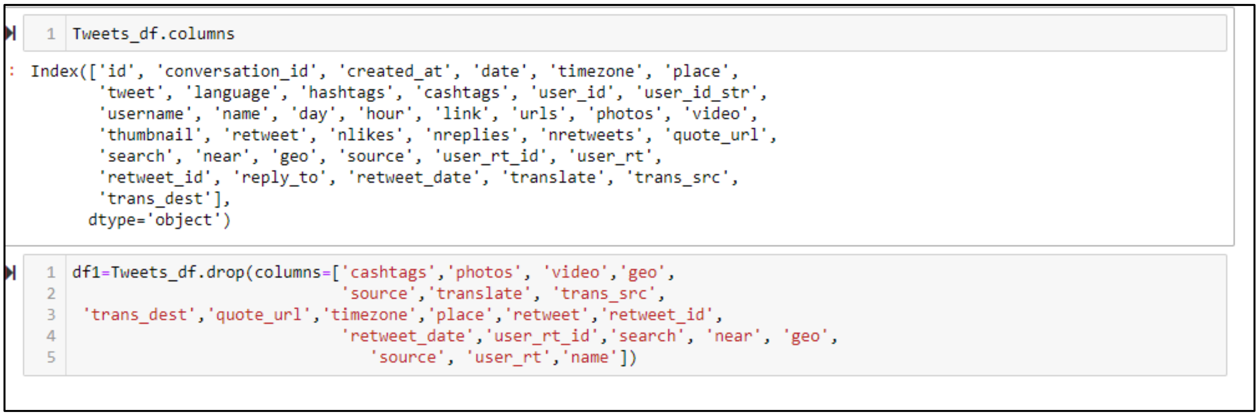
Apply data collection Techniques

Using Twint configure to add the profile information.



```python
c = twint.Config()
c.Username = "viren_lathiya"
c.Custom["user"] = ["bio"]
c.User_full = True
c.Output = "users.csv"
c.Hide_output = True
c.Pandas = True
c.Stats = True
#c.Favourites = True
#twint.run.Favorites(c)
twint.run.Search(c)
Tweets_df = twint.storage.panda.Tweets_df
```

And storing the collected data in a dataframe and dropping irrelevant columns

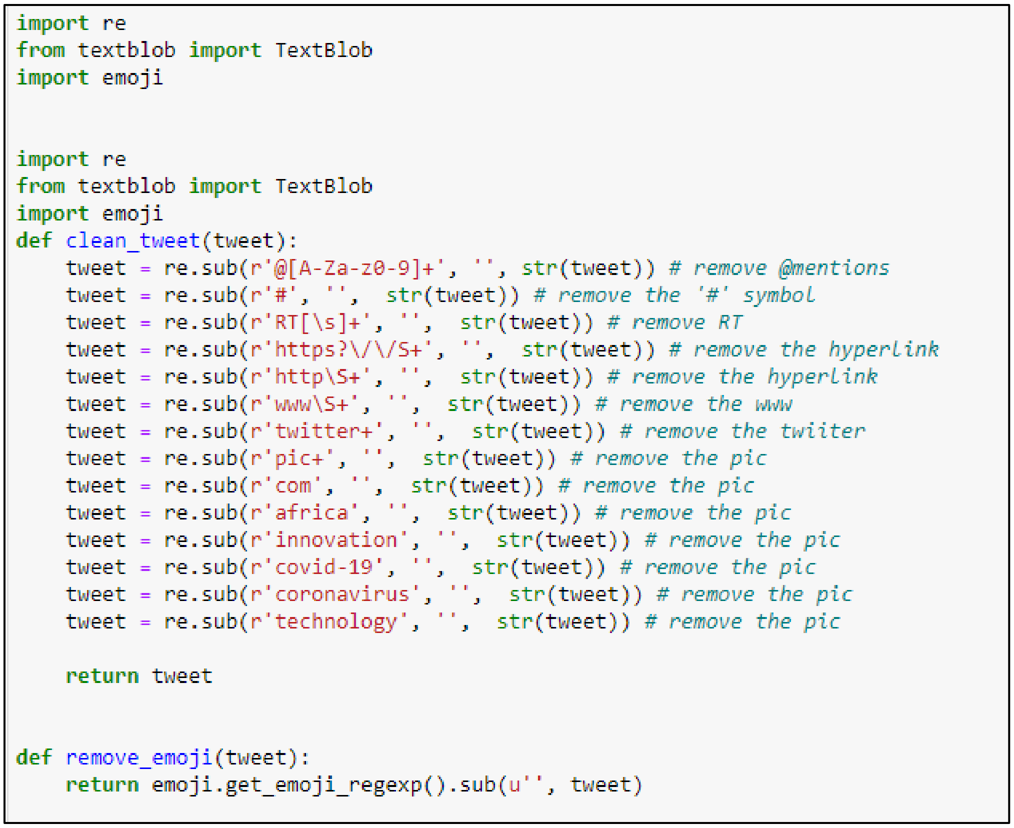


```python
Tweets_df.columns
```

```
Index(['id', 'conversation_id', 'created_at', 'date', 'timezone', 'place',
       'tweet', 'language', 'hashtags', 'cashtags', 'user_id', 'user_id_str',
       'username', 'name', 'day', 'hour', 'link', 'urls', 'photos', 'video',
       'thumbnail', 'retweet', 'nlikes', 'nreplies', 'nretweets', 'quote_url',
       'search', 'near', 'geo', 'source', 'user_rt_id', 'user_rt',
       'retweet_id', 'reply_to', 'retweet_date', 'translate', 'trans_src',
       'trans_dest'],
      dtype='object')
```

```python
df1=Tweets_df.drop(columns=['cashtags','photos', 'video','geo',
                            'source','translate', 'trans_src',
 'trans_dest','quote_url','timezone','place','retweet','retweet_id',
                            'retweet_date','user_rt_id','search', 'near', 'geo',
                               'source', 'user_rt','name'])
```

Cleaning dataFrame and text data by removing stop words.



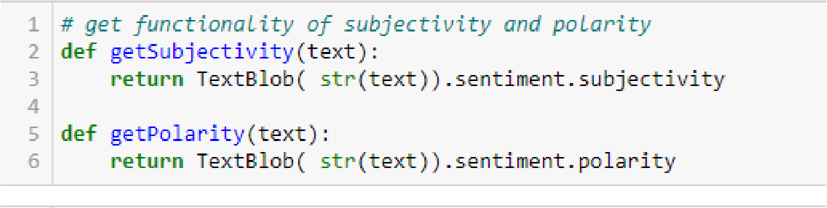
```
import re
from textblob import TextBlob
import emoji


import re
from textblob import TextBlob
import emoji
def clean_tweet(tweet):
    tweet = re.sub(r'@[A-Za-z0-9]+', '', str(tweet)) # remove @mentions
    tweet = re.sub(r'#', '',  str(tweet)) # remove the '#' symbol
    tweet = re.sub(r'RT[\s]+', '',  str(tweet)) # remove RT
    tweet = re.sub(r'https?\/\/S+', '',  str(tweet)) # remove the hyperlink
    tweet = re.sub(r'http\S+', '',  str(tweet)) # remove the hyperlink
    tweet = re.sub(r'www\S+', '',  str(tweet)) # remove the www
    tweet = re.sub(r'twitter+', '',  str(tweet)) # remove the twiiter
    tweet = re.sub(r'pic+', '',  str(tweet)) # remove the pic
    tweet = re.sub(r'com', '',  str(tweet)) # remove the pic
    tweet = re.sub(r'africa', '',  str(tweet)) # remove the pic
    tweet = re.sub(r'innovation', '',  str(tweet)) # remove the pic
    tweet = re.sub(r'covid-19', '',  str(tweet)) # remove the pic
    tweet = re.sub(r'coronavirus', '',  str(tweet)) # remove the pic
    tweet = re.sub(r'technology', '',  str(tweet)) # remove the pic

    return tweet


def remove_emoji(tweet):
    return emoji.get_emoji_regexp().sub(u'', tweet)
```

Applying sentiment analysis on the cleaned text.



```
# get functionality of subjectivity and polarity
def getSubjectivity(text):
    return TextBlob( str(text)).sentiment.subjectivity

def getPolarity(text):
    return TextBlob( str(text)).sentiment.polarity
```

After applying textblob’s sentiment analysis functions we get fields Polarity and Subjectivity.

Polarity ranges from -1 to 1. Negative values indicate negative sentiment and positive values indicate positive sentiment.

Subjectivity ranges from 0 to 1. With 0 being most factual and 1 being most subjective.

Exporting data to CSV and using Tableau to create the dashboard.

Results:

Link to Dashboard:

https://public.tableau.com/app/profile/kalpesh.odedra/viz/Twitter_dashboard/Overall

- The client has tweeted a total of 567 times with 519 unique posts.
- He has received 569 likes and 51 tweets during his lifetime on this platform.
- His tweet activity has been most during 8 am to 2 pm which is also the duration where his engagement is the most.
- His distribution of positive and negative tweets appears mostly equal.
- When his tweets are factual they are also very neutral however, with high absolute values of polarities the value of subjectivity also appears to increase.
- Recently his Twitter activity has increase which has also resulted in increased engagements. Also, recently a lot of his tweets have been positive.
- Wednesday and Thursday appear to be the most popular days for him for high engagement. Both of these days appear to have higher than average engagement scores for the recent duration.

Recommendations:

The client should increase his Twitter activity more with additional positive and factual tweets to balance out his scores.

He should also try to use Wednesday and Thursday from 8 am to 2 pm mostly to post his tweets. The end of the week appears to be the least in terms of engagement, therefore he might want to keep his casual tweets during those.

The ratio of total tweets to total posts indicates that he posts a lot but he doesn’t reply or engage with other users much. To increase his engagement rate he should engage with other users more.

Prepared By:
Kalpesh Odedra

He might want to use this dashboard for an additional purpose. Since he has the access to his data, this might allow him to do A/B testing with a lot of accuracies to showcase his skills for future interviews.

References

Project Twint. (2021, mar 2). *twintproject*. Retrieved from github: https://github.com/twintproject/twint

Serrat, O. (2017, May). *The Five Whys Technique*. doi:10.1007/978-981-10-0983-9_32

tweepy. (2021). *Tweepy Documentation*. Retrieved from docs.tweepy.org: https://docs.tweepy.org/en/latest/

Venkateswarlu Bonta, N. K. (2019). *A Comprehensive Study on Lexicon Based Approaches for*. Retrieved from researchgate: https://www.researchgate.net/profile/N-Janardhan/publication/333602124_A_Comprehensive_Study_on_Lexicon_Based_Approaches_for_Sentiment_Analysis/links/5d13452ca6fdcc2462a688ed/A-Comprehensive-Study-on-Lexicon-Based-Approaches-for-Sentiment-Analysis.pdf

Prepared By:
Kalpesh Odedra