Andrej Kurusiov

How users are reacting to #karabakh OR #artsakh Twitter hashtag in the beginning of October 2020?

CS-EJ5621 - Social Media Analytics D

TABLE OF CONTENTS

OVERVIEW AND MOTIVATION

The project is centered on Twitter data analysis since it's the most affordable platform for educational research, and there are several handy tools available for analysis. Major side note is that Twitter historical data as such is virtually not available with free tools, therefore only a relatively short time period can be analysed.

Since there's no possibility to access historical data, I focus on current events, and one of the most prominent ones is a military conflict between Armenia and Azerbaijan over disputed territories, denoted by the "**#karabakh OR #artsakh**" hashtags on Twitter.

Without any doubts, the ongoing pandemic is the news number one, together with US elections and such, but I would like to analyze something not that much paid attention to in the media, though very important in terms of impact on peoples' lives. Besides personal interest in the matter, I suppose the outcomes would be interesting for the general public, and the results would potentially increase awareness about the issue of wars and/or ethnic conflicts.

DATA COLLECTION METHOD, TIMEFRAME, AND TOOLS

The project collects the data from Twitter using the "#karabakh OR #artsakh" hashtags, since they are the ones of the most used names (by both sides of the conflict) of the disputed region in question.

I use TAGS software for data collection, and since over a week historical data is not available for the platform, nor there is much time for data collection during the course, the data is gathered during 6-12 October (Tue-Mon) 2020 automatically. It shall be noted that there's no guarantee nor information which amount of data and which data samples are released by Twitter for data collection during this procedure. This way the initial amount of data is around 187000 records (tweets). For educational purposes it suffices, and further data collection is not feasible in terms of data cleansing and analysis difficulty.

In order to fight with automated ("bot") posts and to filter them out right away, the "*Follower count filter*" parameter of TAGS software is set to "10". Also, there are specialised tools for bot detection (see References), but due to lack of time, free API usage limits and technical difficulties, this possibility was not realised in the project.

Also, in connection with TAGS software and processing in Python, Google Sheets/ Google Drive and Google Colab was used for initial data collection and transfer to local machine for further analysis.

Other than the free TAGS tool used for data collection, tools (libraries) implemented in Python environment are used (see References) for data cleaning, analysis and producing a WordCloud since they are free, relatively easy to use, fast, and after all - Python is the most familiar language for me.

ANALYSIS MILESTONES, METHODS AND TOOLS

Due to the limited experience, tools availability and limited data timeframe, the analysis is primarily focused on statistical analysis, geographical distribution of tweets, most prominent keywords ("WordCloud"), and sentiment analysis related to geographical and time distribution of tweets (for example, sentiments of one or another side of the conflict and tweets in favor or not of observed time developments). The latter point is the weakest technically, as the topic is very narrow and the nature of tweets (especially in this topic context) shows, that a lot of tweet length is comprised of #hashtags. Since #hashtags per se are not taken into sentiment analysis, quite a substantial amount of data and meaning is discarded from analysis.

Getting the data out from GoogleSheets

Since the original GoogleSheets document produced by TAGS software consisted of 189618 rows/ records, the processing (and even viewing of) data could not be done efficiently in the GoogleSheets itself. Therefore I had to come up with the way to transfer the data to a local machine for further analysis. Direct export from GoogleSheets to xlsx format did not work well since upon opening in MS Excel, there were some errors reported.

Therefore I used Google Colab with Jupyter Notebooks in Python in order to open the GoogleSheets document, remove irrelevant for analysis columns, remove empty and duplicate rows (if any) and save as a regular (and error-free from the Excel point of view) xlsx file, which could be them downloaded to a local machine and analysed further. For reading and saving the data I have used google.colab.auth and pandas libraries.

After dropping 2465 duplicate rows and empty rows (most probably, the artifacts of TAGS data collection process), the resulting data frame became 187152 rows. The data was saved in the xlsx format of Google Drive and downloaded to a local machine (81,2 Mb).

Resulting data columns:

- id_str
- from_user
- text
- created_at
- time

favorite_count
retweet_count
retweeted_status
lang
user_location
user_followers_count
user_friends_count
entities_str

Data reading, parsing and cleaning on local machine

Processing and analysing on a local machine is continued in Python Jupyter Notebooks in VS Code IDE environment. The main libraries used are pandas, nltk, and matplotlib.

After analysing available data, the column 'created_at' had to be parsed as pandas datetime column. Also, the columns 'entities_str', 'from_user' were dropped as well since there is no real value in them.

A data sample:

id_str	text	created_at	favorite_count	retweet_count	retweeted_status	lang	user_location	user_followers_count	user_friends_count
1314999661393051904	RT @KirianSev: To my dear followers, those that retweet, liked or comment on recent tweets, a warm thank you! Especially those highlighting plight of #NagornoKarabakh #Artsakh that's under attack ...	2020-10-10 19:42:05	<NA>	5	{extended_entities={media=[Ljava.lang.Object;@30c0ee53}, metadata={result_type=recent, iso_language_code=en}...	en	NaN	396	626
1314938855380996096	RT @atatoyan: Since Sept. 27 Azerbaijani mil. attacks on Armenia & #Artsakh & peaceful population r w/ massive #hatespeech towards #ethnic Armenians. Social & mass media continuously s...	2020-10-10 15:40:28	<NA>	1340	{metadata={result_type=recent, iso_language_code=en}, in_reply_to_status_id_str=null, in_reply_to_status_id=null, ...	en	NaN	68	63

Retweets

Retweets are playing the role of indicating agreement, endorsement or trust toward the source or the author and can provide information about networks and opinions, but without the added comments in particular it is not user-generated information. Also, although someone shares something, we don't necessarily know why, unless it was accompanied by a comment. It is difficult to decide unambiguously what a person thinks of the re-tweeted matter. Therefore retweets were not analysed

and had to be removed. I used a 'retweeted_status' column (= empty) to drop the records with retweets. The resulting data consisted of 33800 records.

User language

User language is defined in the 'lang' column and when present, indicates a BCP 47 language identifier corresponding to the machine-detected language of the Tweet text, or 'und' if no language could be detected. Since in further analysis due to technical limitations and tools availability only tweets in English ('en' code) are analysed, only general statistics on languages used is presented

Top-10 languages used:

Language code	Records
en	22777
und*	6156
fr	1138
tr	993
es	572
hy	485
ru	385
de	348
in	208
it	141

* undefined

User location

User location is essentially a free text entered by a user on his/ her account. That leads to a situation, that a) some users do not have any location data, b) some users enter wrong/ incomprehensible information and c) some users have similar locations denoted differently (e.g. "USA", "United States").

Initial information on user locations:

```
count      20033
unique      2047
top      Armenia
freq        3187
```

Top user locations, unprocessed:

```
Armenia           3187
Azerbaijan        1463
Los Angeles, CA   1330
```

Yerevan, Armenia 1031
United States 442
City of Angels 394
Yerevan 381
Baku, Azerbaijan 355
Paris, France 220
France 193
Los Angeles 185
Armenia, Yerevan 179
Armenia 🇦🇲 177
Baku 174

As one can see, there are many instances where data on location shall be harmonised. Therefore the most frequent locations were corrected using manual mapping, like 'Yerevan, Armenia' -> 'Armenia', 'Los Angeles, CA' -> 'United States' etc.

The resulting top-10 user locations are presented below:

Location	Records
Armenia	4955
USA	3072
Azerbaijan	2069
France	413
Turkey	232
Canada	187
UK	129
Argentina	110
Paris	85
München	81

Locations per user language

After harmonisation of top user locations, we can see the following data:

language code	count	unique	top	frequency
en	13406	1493	Armenia	3777
und*	3423	463	USA	1047
fr	839	165	France	172
tr	639	163	Azerbaijan	214
es	436	53	Argentina	101
hy**	352	46	Armenia	277
ru	255	48	Armenia	107
de	195	58	München	79
it	96	34	Italia	28
in	63	32	Azerbaijan	14

* undefined
** hy - Armenian

The majority of tweets are written in English, and the top location is Armenia. It's noteworthy to see that Azerbajani language ('az' code) was not present in the top-10 list at all.
Data rows left after non-English rows removed: 22777.

Top locations for English tweets:

location	count
Armenia	3777
USA	1936
Azerbaijan	1283
France	205
Canada	145
UK	116
Turkey	100

Sentiment analysis and text preprocessing

Tweet text pre-processing was done similar to the techniques presented during the course.
The text tokenizer, lemmatizer, and stop words were used from the `nltk` package. Stop-words were extended by '*Artsakh, Karabakh, Armenia, Azerbaijan*' words. The pandas dataframe was added a 'text_cleaned' generated column with cleaned tweet text.

Choosing sentiment analysis tool

There are several sentiment analysis tools available on the internet (see References), both free and paid versions. Free online sentiment analysers are usually limited to web form with manual input or with a limited usage API which use has to be coded in the project. Therefore I have decided to turn to the local libraries available in Python which are free to use.
In practise, I have been choosing between *NLTK Sentiment Analysis Package* (in particular, VADER - Valence Aware Dictionary and sEntiment Reasoner) and the *TextBlob* library. Since both of them have similar capabilities for educational purposes, I have chosen VADER from the `nltk` package, since the package is already used in the project.

There are several features of VADER tool that make it great for starting with the sentiment analysis topic:

- It is case sensitive. The sentence 'This is great' has a different score than the sentence 'This is GREAT'.
- The punctuation matters. The exclamation marks for example have a positive score
- The emojis also have a score and actually very strong sentiments. E.g. the <3, :) , :p and :(
- Words after @ and # have a neutral score.

Therefore when using the VADER tool, there's presumably no need to clean the text first.

Sentiment Analysis using VADER

VADER tool can return the score (in the range [-1, 1]) for positive, negative and neutral sentiment of the text. It also has the '*compound*' score, which is a very useful metric in case we want a single measure of sentiment.
Typical threshold values are the following:

- positive: compound score > =0.05
- neutral: compound score between -0.05 and 0.05
- negative: compound score <= -0.05

First, I had to test the hypothesis about the need of cleaning a tweet text before analysing it with VADER. It was run on both original and cleaned tweet text with the following results:

Same score sign (positive or negative):	21547
Opposing score sign (positive or negative):	1230

An example of the results:

vader_original	vader_cleaned	Original text	Cleaned text
-0.4215	-0.4215	#StopAliyev because he broke a ceasefire agreement to bomb Armenian civilians in #Artsakh during the #COVID19 pandemic	broke, ceasefire, agreement, bomb, armenian, civilian, pandemic
-0.6124	-0.6124	1200-year-old #Ganja, #Azerbaijan's 2nd largest city. 100km far from #Karabakh. Entire residential neighborhood wiped out by #missiles launched by #Armenia. At least 9 #Azerbaijani civilians kille...	largest, city, far, entire, residential, neighborhood, wiped, launched, least, civilian, killed, wounded, incl
-0.4226	0.1027	@AFP Since when historically Inhabited territories are called "Occupied"? In the door of collapse of USSR people in #Artsakh like people in #Azerbaijan claimed for independence. This is internatio...	since, historically, inhabited, territory, called, occupied, door, collapse, ussr, people, like, people, claimed, independence, internationally, acknowledged, right, internationally, accepted, she...

As the vast majority of scores were of the same sign, I continued with the original tweet text analysis with VADER (as suggested in the documentation).

Next, VADER score has to be re-coded into 'positive, negative, neutral' code for better comprehension. Contrary to customary practise I used the following threshold values since the data has very much noise (hashtags and very short tweets):

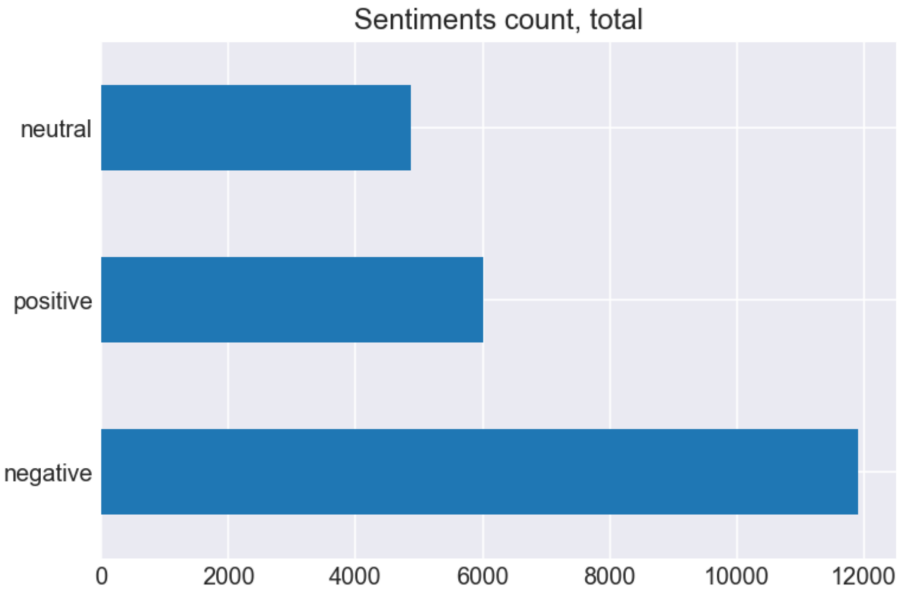
- positive: compound score >= 0.1
- neutral: compound score between -0.1 and 0.1
- negative: compound score <= -0.1

The examples of mapped compound scores with original and cleaned twitter texts:

vader score	sentiment	Original text	Cleaned text
0.0000	neutral	Bro this guys still barking #artsakhstrong #defendarmenia #stopaliyev #stoperdogan #stopazerbaijaniaggression #stopazeriaggression #peaceforarmenians #peaceforarmenia #peaceforartsakh #karabakh #ż...	bro, guy, still, barking
-0.9381	negative	Tonight at 2 AM, Armenian side launched missile attack on #Ganja, #Azerbaijan. \n\nAs a result of the attack on civilian infrastructure, a residential building completely destroyed, at least 3 civ...	tonight, armenian, side, launched, missile, attack, result, attack, civilian, infrastructure, residential, building, completely, destroyed, least, civilian, died, injured
0.4809	positive	Its probably too soon to say something about the substance of negotiations that took 11 hours behind the doors, but I know one thing for sure Azerbaijani Nation will never be the same & will n...	probably, soon, say, something, substance, negotiation, took, hour, behind, door, know, one, thing, sure, azerbaijani, nation, never, never, accept, half, baked, solution

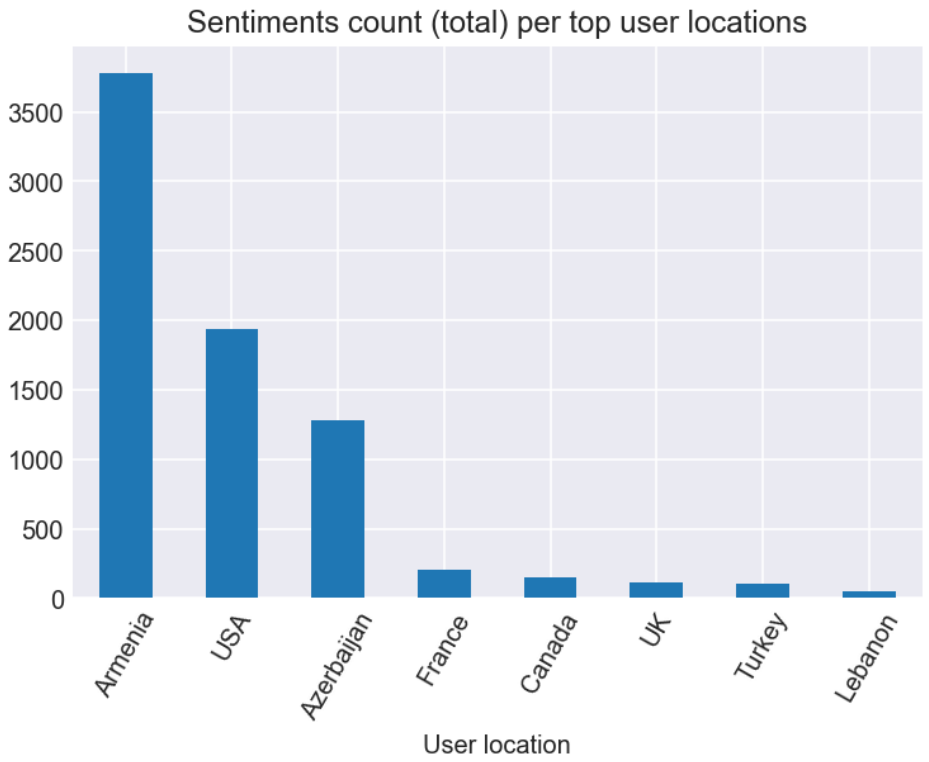
Statistics related to sentiments

The overall sentiments count in English (here and further on) tweets:



As the topic suggests, the vast majority is negative sentiments whereas it's actually surprising that VADER has detected some positive ones. I strongly believe it's due to the noise in the texts (#hashtagged words amount) and very specific topic in question.

Sentiments related to user location



Location	count	top	frequency
Armenia	3777	negative	2032
USA	1936	negative	993
Azerbaijan	1283	negative	670
France	205	negative	103
Canada	145	negative	74
UK	116	negative	68
Turkey	100	negative	47
Lebanon	52	negative	29

As observed before for all the used languages, there's still a strong representation of English language tweets in countries which are not directly involved in the conflict (e.g. USA and France).

Sentiments related to the length of the tweet

Generally (mean value) negative tweets are longer (242 char), positive tweets are shorter (222 char), and neutral ones are the shortest (188 char):

Sentiment	**Length of tweet**
negative	242
neutral	188

positive	222

Sentiments related to friends count

There might be some relations, but given that negative tweets represent the majority of the data set, it's hard to draw any concrete conclusions:

Sentiment	Number of friends (total)
negative	480
neutral	588
positive	571

Sentiments related to favourites count

There's no useful information in the obtained data, as there are very low favourites counts (mean = 7).

Sentiments related to retweet count

Retweets of the original tweets represent similar frequency as the original tweets' sentiment count with a slight skew to positive and neutral retweets:

Sentiment	Number of retweets (total)
negative	3891
neutral	1580
positive	1927

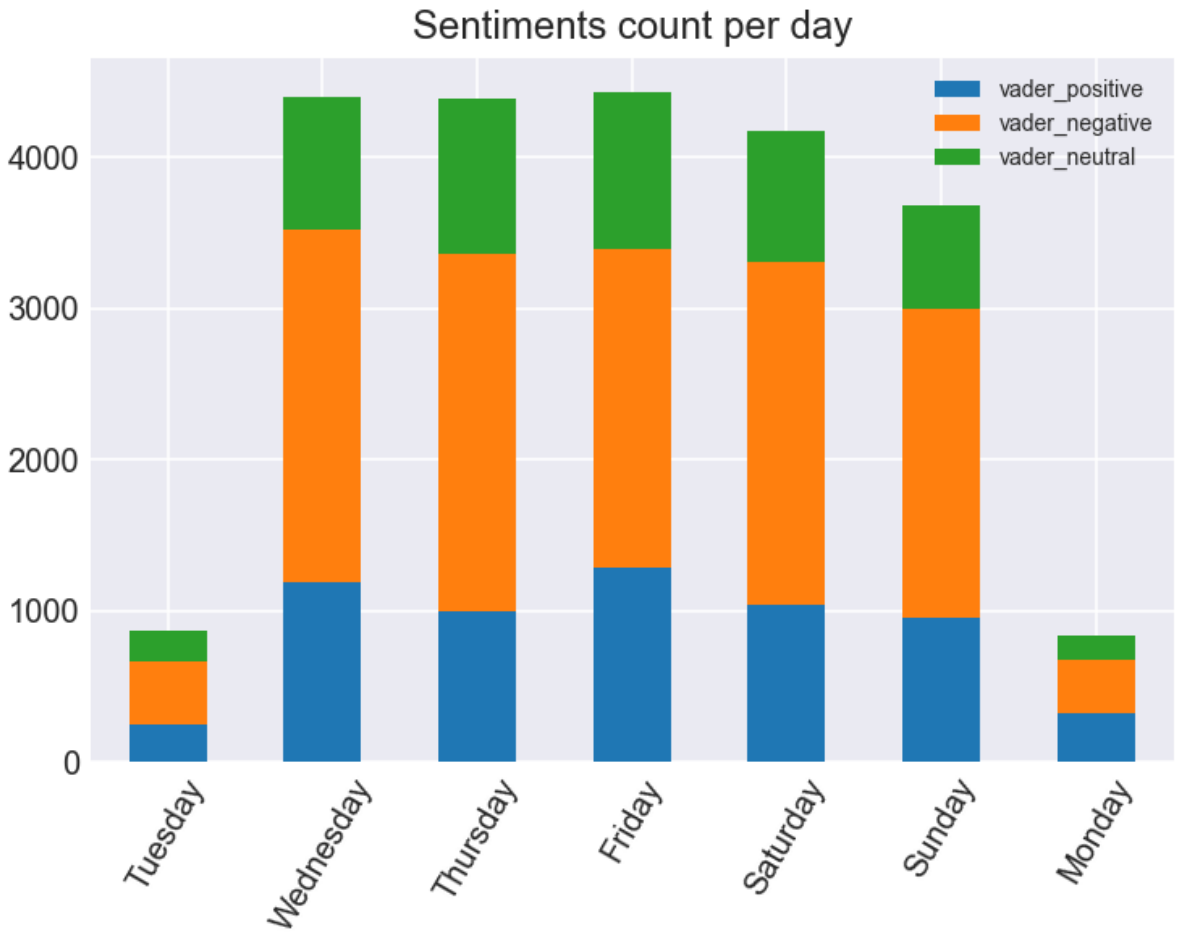
Sentiments related to user followers count

The same trend exists as with the sentiments related to retweet count. Total numbers are a bit hard to interpret:

Sentiment	Number of followers (total)
negative	11902
neutral	4865
positive	6010

Sentiments dynamic by observation day

We do not have many consecutive observation days, only a week, but still there is a certain sentiment dynamic observed. It can probably be related to the real events on those days. Dates were recoded to week days (Tuesday 6 Oct - Monday 12 Oct 2020):


Sentiments count per day
vader_positive
vader_negative
vader_neutral
4000
3000
2000
1000
0
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Monday

WordCloud

I use a Python library for WordCloud visualisation, and despite quite a substantial amount of tweet records, it does the job very well (run time is just a few seconds). As the data contains a lot of noise and #hashtagged words, I used cleaned tweets' text for WordCloud generation. Also, the topic being Twitter data analysis, I use a twitter logo image in order to represent a WordCloud.

Only 300 words maximum were set to be included in the resulting WordCloud in order for it to remain readable. I used the same stop-words as in the text cleaning procedure with an added *'armenian'* stop-word, as it naturally repeats very often.

RESULTS AND FINDINGS

First of all, the initially anticipated amount of data rows was much bigger in the end. That might be related to a) the amount of data generated by users and/ or b) the way TAGS software collects data. Anyhow, it brought an additional requirement for fast and reliable data processing.

Language wise, the majority of captured data was produced in English, which tells either a) that Twitter is mostly being used by English-speaking users (and locations) or b) the purpose of posting the tweets on the topic is mostly drawing international attention; also bot activity may be involved.

Majority os English tweets was surprisingly distributed among Armenia, USA, Azerbaijan and France, which tells that most likely many immigrants in those countries not directly involved in the conflict are of Armenian / Azerbaijani origin.

LESSONS LEARNED / FURTHER DEVELOPMENT

Data collection from open/ free sources is not an easy task. In the case of Twitter, one can never be sure which portion of data is actually captured by TAGS software either.

The technical difficulty of even viewing this amount of data is substantial. It shall be anticipated and thought of the whole data processing and transferring pipe, depending on available data processing capabilities and skills.

Further development would naturally involve more fine-tuned analysis of sentiments related to other factors, not just looking at the totals. More attention could be brought to cleaning the initial data set, especially removing *bot-generated* data, as well as following more holistic approach on filtering/ not out retweets and dealing with #hashtagged words in the sentences (as it sometimes drastically changes the meaning).

Also, some other sentiment analysis libraries (e.g. TextBlob) could have been used too in order to compare their suitability for this certain data set and topic.

Data visualisation would be another big step which requires more technical and substantial understanding of the topic.

REFERENCES, TOOLS AND RESOURCES

1. TAGS: https://tags.hawksey.info/
2. Twitter: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview/tweet-object
3. Data cleaning: https://stackoverflow.com/questions/2304632/regex-for-twitter-username
4. WordCloud:
    a. https://voyant-tools.org/
    b. https://www.danielsoper.com/wordcloud/
    c. https://github.com/amueller/word_cloud
5. Sentiment analysis:
    a. https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/
    b. https://sentigem.com/
    c. https://www.meaningcloud.com/
    d. http://sentistrength.wlv.ac.uk/
    e. www.minemytext.com
    f. https://netlytic.org/
6. Bot detection:
    a. https://github.com/IUNetSci/botometer-python → https://rapidapi.com/microsoft-azure-org-microsoft-cognitive-services/api/microsoft-text-analytics1/endpoints
    b. https://www.cs.unm.edu/~chavoshi/debot/api.html
7. Tweet Sentiment Visualization: https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/