CECS 632 Data Mining

Project 1: Gathering and exploring immigration data from Twitter

Aliaa Elshamekh 9/23/19 Fall 2019

1. Abstract

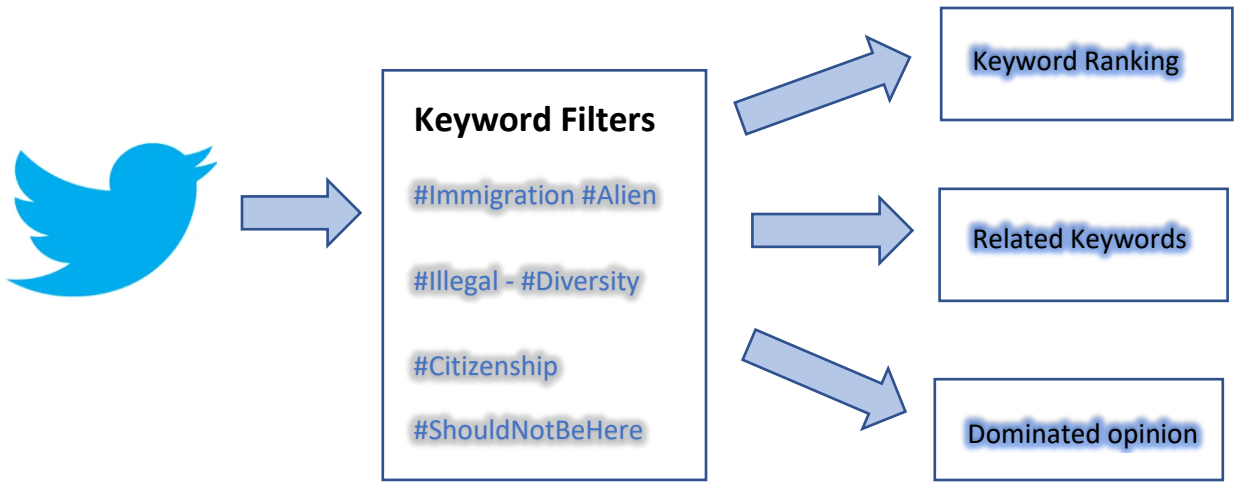
Twitter is one of the most popular social media sites and often becomes the primary source of information. Twitter has found substantial use in several settings. For example, Twitter played a major role in the 'Arab Spring' and has been adopted by many the Fortune 100. All of these and other events have led to a large database of tweets that have attracted the attention of several companies and researchers through what has become known as 'Twitter mining' (also known as 'TwitterMining'). Current hashtag studies have taken two approaches, either they have concentrated on event-oriented hashtags such as the ones used in US presidential elections or they have used a **hashtag** agnostic approach where a random extract of Twitter data is taken and analyzed

Twitter data constitutes a rich source that can be used for capturing information about any topic imaginable. This data can be used in different use cases such as finding trends related to a specific keyword, measuring brand sentiment, and gathering feedback about new products and services. Twitter provides both researchers and practitioners a free Application Programming Interface (API) which allows them to gather and analyze large data sets of tweets. Twitter data are not only tweet texts, as Twitter's API provides more information to perform interesting research studies. However, analysis of Twitter messages (tweets) is regarded as a challenging problem due to some difficulties such as large amounts of data that cannot be easily handled. Moreover, tweets are short. In addition, tweets are of an informal type of discourse that does not cover any functional topic.

This report describes process gathering information and knowledge from Twitter. I**mmigration**, keywords are used to collect tweets that support immigrants' rights and others which are against immigration ideas. Immigration hashtags used in tweets usually are driven by recent changes in the USA. For example: President trump suggestion to change the laws of immigration and lottery immigrants and building wall on Mexico borders.

In this project, we are going to study the rationale behind some of hashtags (pro and against keywords) that are related to immigration topic. We have an initial manual classification for the tweets, and we are going to discuss the steps used to preprocess the gathered data from 'Tweepy. These keywords hashtags are added to show that the tweet's content is related to the other side of immigration debate. We are going to analyze the data make visualizations that will help to show the frequency of the keywords in the dataset.

The findings are to be validated with what news says or what your real-life experiences tells us? How to make some conclusions about other keywords that are related to the immigrations topic which reflects users' opinions.



2.Gather and explore immigration data from Twitter

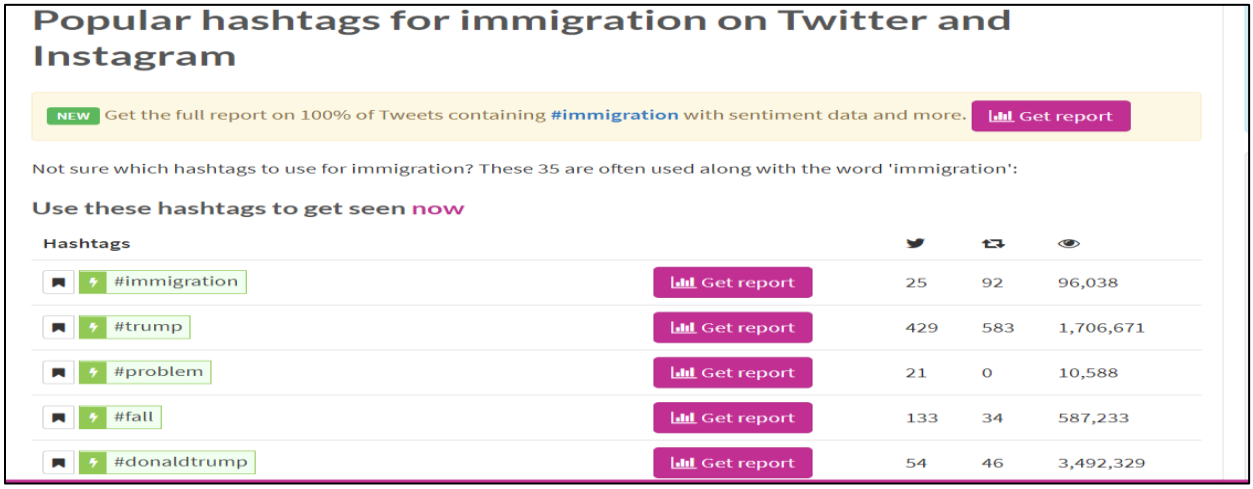
2.1 Problem description:

During this month the case of immigrants are come in the front again due to some announced changes in the laws related to immigration to USA, so people started to express their opinions about this issue on Twitter. So, some keywords and hashtags are related to immigration were used to by the people to express their opinion about these changes in the laws. Of course, there are people are against immigrants and want them to get back to their country. On the other side, others believe in diversity and citizenship and that all immigrants must have their rights in the country.

I chose some keywords that are related to both sides and start to search for tweets about them to check whether they are popular or not.

- chosen keywords for pro-immigration tweets are citizenship, diversity and EqualityForAll
- chosen keywords about against immigration tweets are: ShouldNotBeHere, Alien,Illegal,undocumented

I used the famous website https://ritetag.com/best-hashtags-for/immigration to find out what is popular immigration hashtags this days and make sure that they are related to immigration



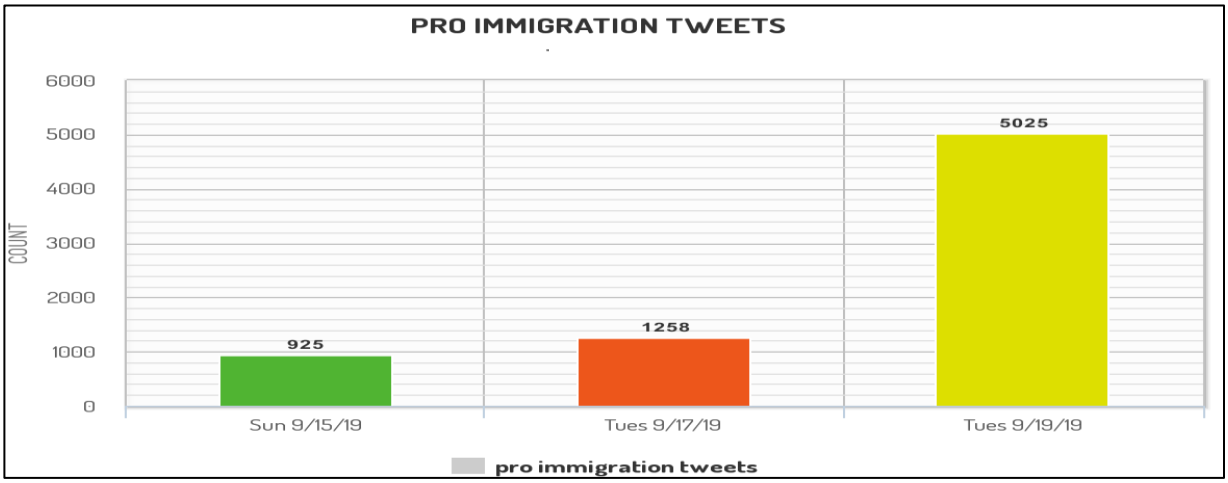
2.2 Data gathering:

In this project I used Twitter data over three days to compare the frequency of some keywords that are related to the immigration topics which are divided into two categories: pro-immigration keywords and others used for finding tweets which are against immigration.

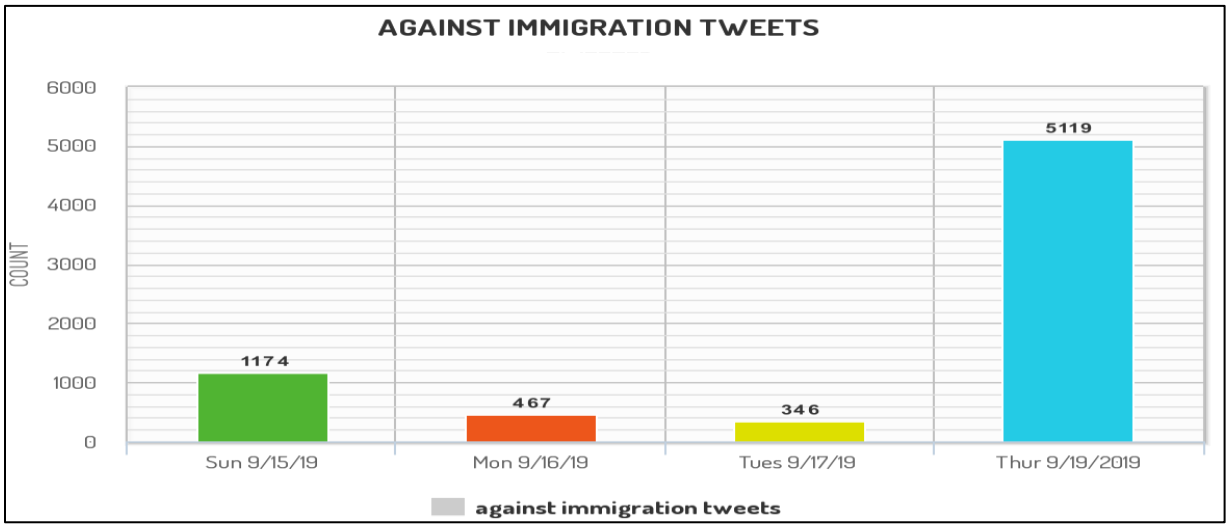
The time interval is from September 15th, 2019 to September 19th, 2019. the data is collected for 2 hours every 2 days. Data are gathered into 2 distinct categories according to the keywords used.Finally, at the end of data gathering I have two separated files of tweets. The size of each file is almost 50 MB each. The data is collected in JSON form which has one record for each tweet in the stream that created during streaming time whatever it is a retweet or quoted status. The tweet has so many attributes that I had to analyze them to see which of them would be helpful to the objective of the project. The following figure shows how one tweet is written in JSON.


```
{"created_at":"Sun Sep 15 21:07:36 +0000 2019","id":1173342625958522881,"id_str":
"1173342625958522881","text":"@kennnyyd @emrazz They're the same people who say \"if we make
abortion illegal it will stop all abortion\" and \"sens\u2026 https:\/\/t.co\/6c3bBzqdlY",
"display_text_range":[18,140],"source":"\u003ca href=\"http:\/\/twitter.com\/download\/android\"
rel=\"nofollow\"\u003eTwitter for Android\u003c\/a\u003e","truncated":true,"in_reply_to_status_id":
1173339715149291520,"in_reply_to_status_id_str":"1173339715149291520","in_reply_to_user_id":
1120704924067590146,"in_reply_to_user_id_str":"1120704924067590146","in_reply_to_screen_name":
"kennnyyd","user":{"id":21698182,"id_str":"21698182","name":"GraveyDice \ud83d\udc9a","screen_name"
:"GraveyDice","location":"New Zealand","url":"http:\/\/www.goodgravey.wordpress.com","description":
"For thy sweet love remembered such wealth brings That then I scorn to change my state with kings.
http:\/\/youtu.be\/rEOIG5uDwe0 | cis he\/him","translator_type":"none","protected":false,"verified"
:false,"followers_count":751,"friends_count":746,"listed_count":26,"favourites_count":32857,
"statuses_count":72046,"created_at":"Mon Feb 23 21:50:12 +0000 2009","utc_offset":null,"time_zone":
null,"geo_enabled":true,"lang":null,"contributors_enabled":false,"is_translator":false,
"profile_background_color":"DBD8D7","profile_background_image_url":
"http:\/\/abs.twimg.com\/images\/themes\/theme18\/bg.gif","profile_background_image_url_https":
"https:\/\/abs.twimg.com\/images\/themes\/theme18\/bg.gif","profile_background_tile":false,
"profile_link_color":"5913F0","profile_sidebar_border_color":"CEDBCE","profile_sidebar_fill_color":
"F2F0D9","profile_text_color":"080808","profile_use_background_image":true,"profile_image_url":
"http:\/\/pbs.twimg.com\/profile_images\/824348928594235394\/WypQN8JZ_normal.jpg",
"profile_image_url_https":
"https:\/\/pbs.twimg.com\/profile_images\/824348928594235394\/WypQN8JZ_normal.jpg",
"profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/21698182\/1392510975",
"default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,
"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,
"is_quote_status":false,"extended_tweet":{"full_text":"@kennnyyd @emrazz They're the same people
who say \"if we make abrtion illegal it will stop all abortion\" and \"sensible gun control won't
work because criminals will still get guns\".\n\nThey really need to make up their mind.",
"display_text_range":[18,223],"entities":{"hashtags":[],"urls":[],"user_mentions":[{"screen_name":
"kennnyyd","name":"kendra","id":1120704924067590146,"id_str":"1120704924067590146","indices":[0,9
]},{"screen_name":"emrazz","name":"feminist next door","id":30395567,"id_str":"30395567","indices"
:[10,17]}],"symbols":[]}},"quote_count":0,"reply_count":0,"retweet_count":0,"favorite_count":0,
"entities":{"hashtags":[],"urls":[{"url":"https:\/\/t.co\/6c3bBzqdlY","expanded_url":
"https:\/\/twitter.com\/i\/web\/status\/1173342625958522881","display_url":
"twitter.com\/i\/web\/status\/1\u2026","indices":[117,140]}],"user_mentions":[{"screen_name":
"kennnyyd","name":"kendra","id":1120704924067590146,"id_str":"1120704924067590146","indices":[0,9
]},{"screen_name":"emrazz","name":"feminist next door","id":30395567,"id_str":"30395567","indices"
:[10,17]}],"symbols":[]},"favorited":false,"retweeted":false,"filter_level":"low","lang":"en",
"timestamp_ms":"1568581656337"}
```


- Pro- tweets collected on each day:



- Against- tweets collected on each day:



2.3 Software tools

a) Tweepy API

I used for this project Twitter API to download twitter related to those keywords mentioned above for both pro and against immigration tweets. The 1st step is getting Twitter API keys in order to access Twitter Streaming API, we need to get 4 pieces of information from Twitter: API key, API secret, Access token and Access token secret. These were discussed in detail in the tutorial provided with project. The 2nd step is connecting to Twitter Streaming API and downloading data, I used the library provided in the tutorial which called Tweepy to connect to Twitter Streaming API and downloading the data. A file called tweets.py was provided which I just changed the credentials to mine and change the keywords I will used to search for them

b) Libraries in python

Data Gathering

JSON library the data was stored in json format as we mentioned before in data gathering section these formats make it easier to humans to read, and for the machines to parse it. I used Jason library of the python for parsing the data.

Data Manipulation

Pandas library is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language. The two primary data structures of pandas, **Series** (1-dimensional) and **DataFrame** (2-dimensional), handle the vast majority of typical use cases in finance, statistics, social science, and many areas of engineering.

Data Visualization

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and for graphical user interface toolkits.

2.4 Preprocessing

In this phase we need to preprocess the data gathered from Tweepy so that we are be able to manipulate and interpret information from this collected uncleaned data. In Pandas data frame the data will look like a table structure

	created_at	id	id_str	text
2	2019-09-15 21:15:00+00:00	1173344488187731968	1173344488187731968	Ruling party's indecision delays finalisation of the bill to amend Citizenship Act
6	2019-09-15 21:15:01+00:00	1173344492956655619	1173344492956655616	RT @animatedtext: its not forced diversity people just exist
8	2019-09-15 21:15:01+00:00	1173344494835486720	1173344494835486720	RT @She_Run4Me: Both my parents have been living in Malaysia for almost 30 years. App
9	2019-09-15 21:15:02+00:00	1173344495838121986	1173344495838121984	RT @thatstarwarsgrl: Forced Diversity?:Featuring @TheQuartering & The Sad Crying Boys.
30	2019-09-15 21:15:10+00:00	1173344531208474624	1173344531208474624	RT @JoseCanseco: If I were president I would give citizensl
34	2019-09-15 21:15:11+00:00	1173344537126801408	1173344537126801408	@Lee84379818 @SkyCricket A disgrace. So some arsewipe can tick
38	2019-09-15 21:15:13+00:00	1173344545368621057	1173344545368621056	RT @EmeraldRobinson: Diversity is our strength - if by "strength" you mean the collaps
49	2019-09-15 21:15:18+00:00	1173344566226968580	1173344566226968576	RT @NickRewind: Taina was a show that represented and reflected diversity, especially fo
61	2019-09-15 21:15:22+00:00	1173344581515141120	1173344581515141120	RT @not_lewd: 英語: "Do you want to play UNO? It came free with your Eulmore citizens
79	2019-09-15 21:15:30+00:00	1173344616415780864	1173344616415780864	Diversity leads to broader perspectives. So nature prevails in more informed deci
80	2019-09-15 21:15:31+00:00	1173344617540083712	1173344617540083712	RT @EmeraldRobinson: Diversity is our strength - if by "strength" you mean the collaps
84	2019-09-15 21:15:33+00:00	1173344627773960192	1173344627773960192	The Dangers Of Mistaking Diversity For Inclusion In The Workplace https://t
92	2019-09-15 21:15:35+00:00	1173344633650176000	1173344633650176000	RT @EmeraldRobinson: Diversity is our strength - if by "strength" you mean the collaps
102	2019-09-15 21:15:38+00:00	1173344648061960192	1173344648061960192	RT @jefffajans: How LinkedIn's HR Chief is Changing the Diversity Conversation with "Bel
103	2019-09-15 21:15:38+00:00	1173344648393363456	1173344648393363456	RT @SerbianPM: Today at the #BelgradePride. Tolerance and respect for diversity - thes
104	2019-09-15 21:15:38+00:00	1173344649445920768	1173344649445920768	RT @MarkYoungTruth: Except diversity of thought, the left will never accept anyone
106	2019-09-15 21:15:39+00:00	1173344651719446528	1173344651719446528	RT @DIGADA1: @Jorvik4 @sbd1704 @EmmaKennedy @grahamtriggs @guyverhofstadt I rathe
123	2019-09-15 21:15:45+00:00	1173344676147077120	1173344676147077120	@Czakal Diversity Macht Frei https://t.co/hK

- First step is the find out the missing values in each column (feature) and calculate the percentage or filling factor of the data, For the pro-immigration tweets the filling percentage of columns is the following, the columns are in descending order with a threshold percentage is 50%

in_reply_to_status_id	91%	in_reply_to_status_id_str	91%
in_reply_to_user_id	90%	in_reply_to_user_id_str	90%
in_reply_to_screen_name	90%	geo	99%
coordinates	99%	place	98%
contributors	100%	quoted_status_id	77%
quoted_status_id_str	77%	quoted_status	77%
quoted_status_permalink	77%	possibly_sensitive	80%
extended_entities	93%	display_text_range	90%
extended_tweet	88%	withheld_in_countries	99%

- Similarly, the same behavior for against immigration tweets have approximately the same columns that have much many empty values. In such case my decision was to discard this column from data analysis since they are almost empty in addition the most important columns in my analysis is **text** and **User** attributes which don't have any null values.
- For features that are potentially noisy the main important feature for me was **text** feature. At the beginning, I figured out that not all the tweets have my keywords in text attribute for example this tweet

Result:

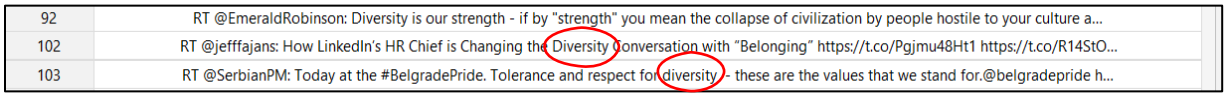
```
result = {str} 'Oh my god I've just had the maddest flashback □'
```

So clearly, I remove all the tweets that does not have any of my keywords.

- **Removal of Non-English Tweets**, I found out that there are little number of tweets that are not written in English**,** so I have removed these tweets also from data so that the tweets become more consistent. This is a list of language found in pro immigration data set:

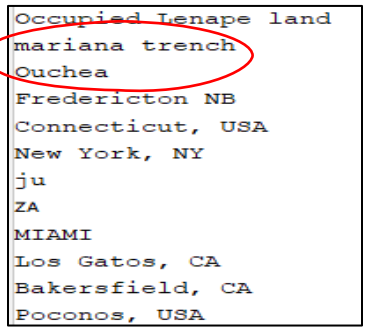
en	de	und	es	in	ja	ca	sv	it	ro	ht	tl
2798	3	8	2	12	16	1	1	1	1	1	1

- Also, when to overcome that searching for text is **a case sensitive** will affect the frequency of keywords in the data set so I transform all my keyword and text feature of the data to **lower case** so that the finding process will be more accurate. The image below illustrates the previous problem.



92	RT @EmeraldRobinson: Diversity is our strength - if by "strength" you mean the collapse of civilization by people hostile to your culture a...
102	RT @jefffajans: How LinkedIn's HR Chief is Changing the Diversity Conversation with "Belonging" https://t.co/Pgjmu48Ht1 https://t.co/R14StO...
103	RT @SerbianPM: Today at the #BelgradePride. Tolerance and respect for diversity - these are the values that we stand for.@belgradepride h...

- What also can be categorized as a noisy date is **the location of the user** who writes the tweet after some process on this feature ,I figured out that it is a free text entry ,in other words it is any fuzzy word that the user can enter .Here is some example of users' location in the dataset :



```
Occupied Lenape land
mariana trench
Ouchea
Fredericton NB
Connecticut, USA
New York, NY
ju
ZA
MIAMI
Los Gatos, CA
Bakersfield, CA
Poconos, USA
```

```
redacted
Northborough, MA
Toronto, Ontario, Canada
Los Angeles, CA
Toronto CA
USA
Tuscola,IL
Toronto, Ontario, Canada
Carolina, Puerto Rico
New York, USA
United States
Austin, TX
```

So, this format of location is **ambiguous** data that is not a good feature we can depend on. A feature like this needs advanced text processing techniques to discard fuzzy places and provide a standard name for each location

- **Remove URL (Links)** : The tweets above have some elements that you do not want in your word counts. For instance, URLs will not be analyzed in this lesson. You can remove URLs (links) using regular expressions accessed from the `re` package.re stands for `regular expressions`. Regular expressions are a special syntax that is used to identify patterns in a string. Tells the search to find all strings that look like a URL and replace it with nothing – "". It also removes other punctuation including hashtags - #

- In addition to the pre-processing done above , There were some noisy data (tweets) **that have one of the used keywords but it is irrelevant to the topic of immigrants** ,Hence we apply multiple filtering by searching for more than keyword at one time for both datasets,for example : In pro-immigration dataset ,I searched first for diversity , then diversity combined with immigration then diversity and citizenship together .In another words, **all combinations of the keywords** to be able to view the relavent tweets .

- Moreover,I found about that the feature **number of followers** related to the user who wrote the tweets needs **discrteriztion** to specify categories or classes depends on followers count and give evey user a new feature according to his number of followers : v. low level influence,low influence level,meduim level influence,high level influence and v. high level influence.

- This would help to visualize the number of users in each dataset and see if those accouts have many followers to analyze the influence of users and the distribution of each influence level among our datasets .Though , that will help to create a model for the public opition. We can observe the distribution of useres who are pro or against immigration.

2.5 Visualization patterns

2.5.1 Histogram of tweets:

- The first chart here describing the histogram of tweets over streaming days for pro and against immigration tweets we can observe that the highest number of tweets among these days was September 19th and the lowest is September 15th.
- This visualization is created by a ready function that plots histogram of data. this number displayed in the following charts are the number of tweets after preprocessing. Also, we cannot say that most tweets are against or pro-immigration since the difference between them is not a large number.
- However, on these three days always against immigration tweets number is always greater than pro-immigration tweets.

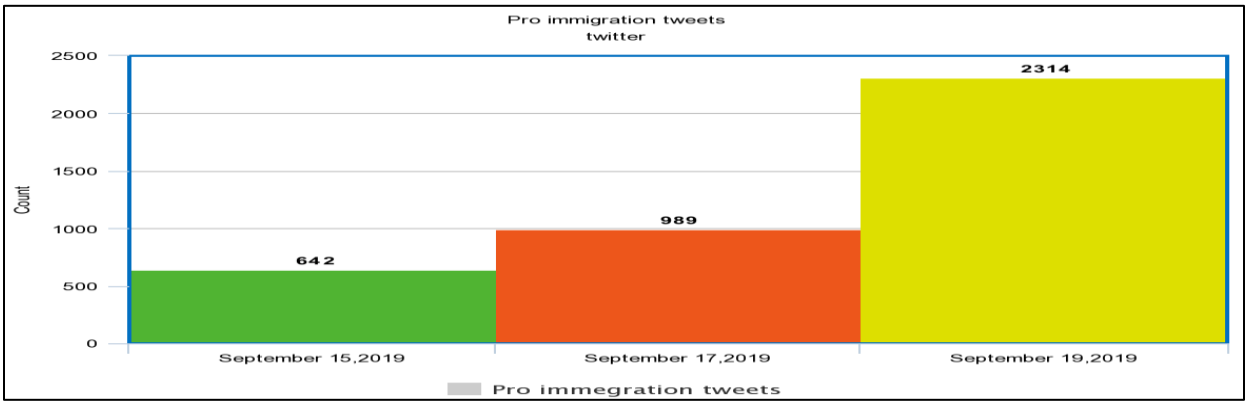


Figure 1: pro-immigration tweets histogram

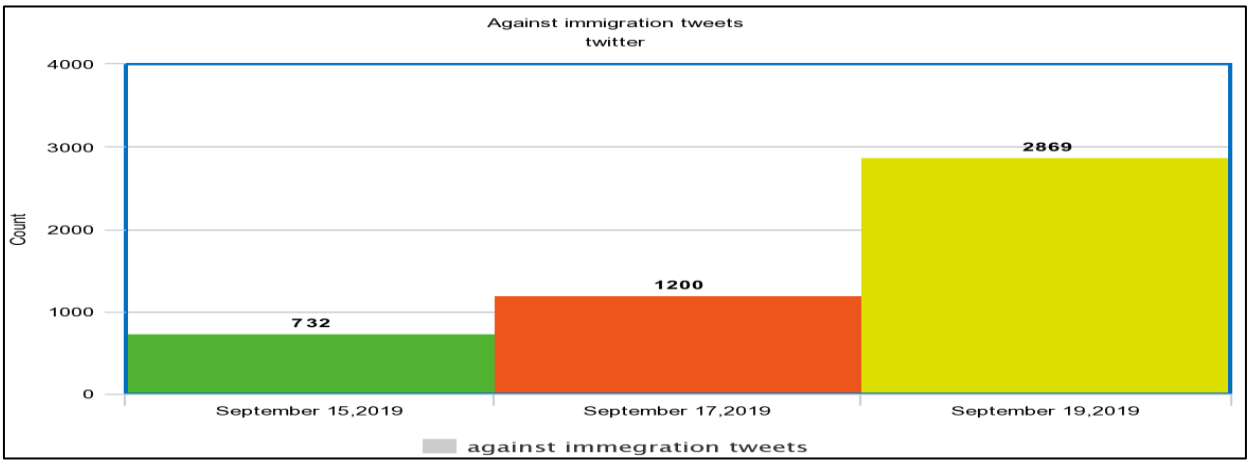


Figure 2: against immigration tweets histogram

2.5.2 Frequency of keywords

- The following pie charts show the percentage of each keyword in the data set, which keyword has the highest frequency, and which does not.
- This is a built in pie chart plot, which takes my slices as a parameters so at the beginning I have calculated the frequency of each keyword on the dataset ,the calculate their percentage from whole tweets then I used the built in function to plot the data.

- For pro-immigration tweets, the chart shows that the most frequency keyword is **diversity** then followed by **immigration** and the least one is **equalityforall** .
- So, from previous analysis, we can see that the choice of **equalityforall** as a keyword for pro-immigration tweets was not the good option.
- For against immigration tweets, the chart shows that the most frequency word is illegal**.**
- Moreover, from this chart we can figure out that selection of the keyword **shouldnotbehere** was not a good choice since it does not appear in the stream of the against immigration tweets.
- For **relevant keyword that would be helpful** when add to search combined with **immigration** keyword is **trump,** the word trump as there are 719 tweets in the against immigration data set that has the 2 words together.

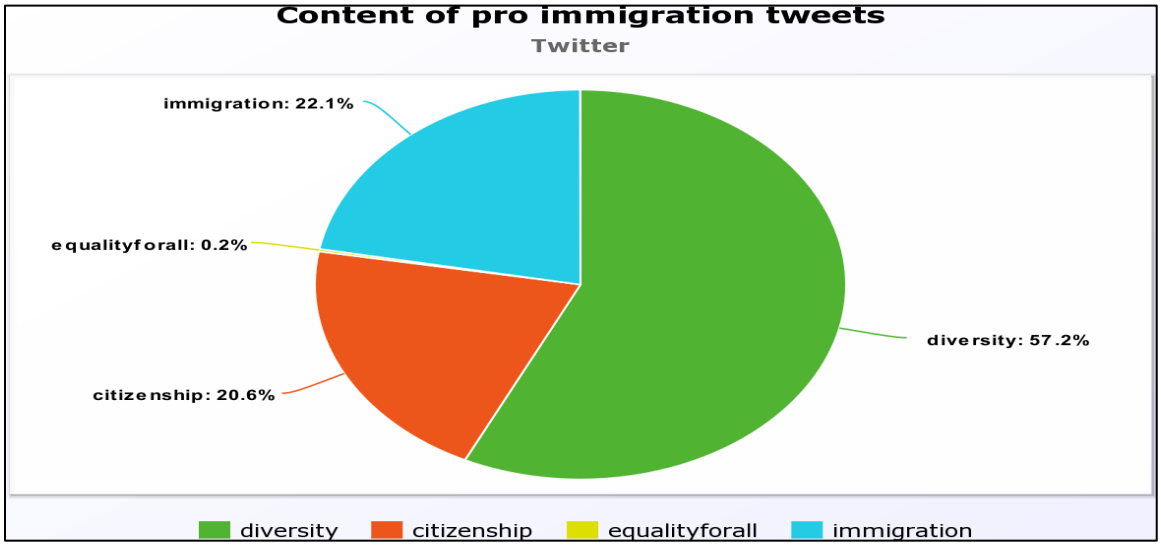


Figure 3 : Frequency of pro-immigration keywords

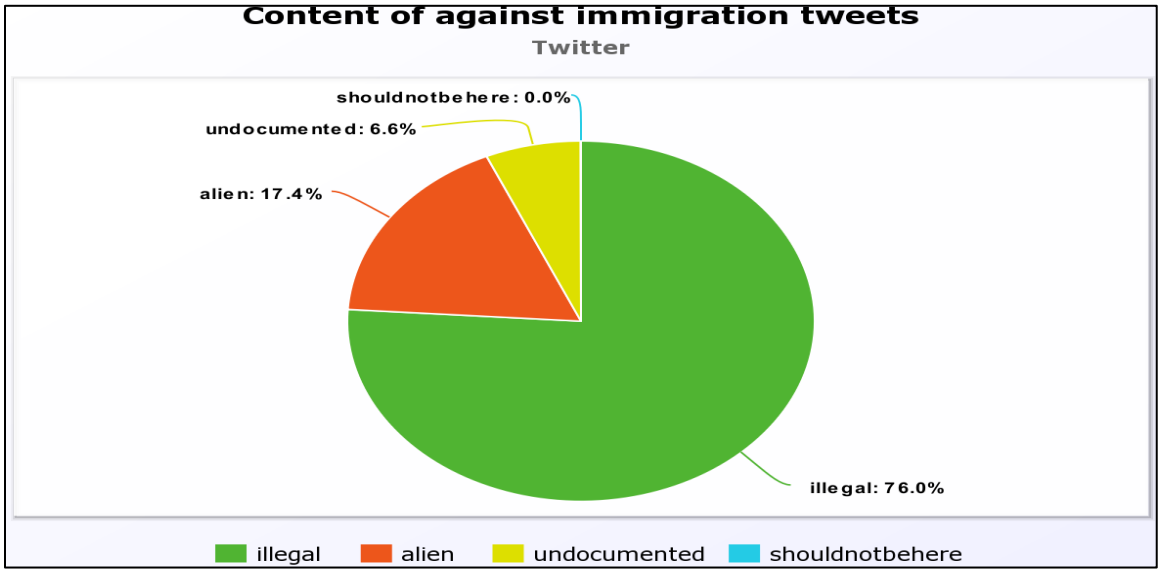


Figure 4 : Frequency of against immigration tweets

2.5.3 Number of followers of users in dataset

- For this visualization we will show a boxplot of number of followers in both data set against and pro tweets dataset. We know that the main goal of any analysis on any social media website is to know what the trend or to be able to build a model about how do people think about certain issues and the percentage of people who are against or support in any raised debate in real life.
- Therefore, this analysis is very crucial since we categorize the users who write each tweet has one of keyword according to the account's followers count which is, in my opinion, a very good indicator for how much influence does this user have in his society.
- At the beginning it was trying to scatter the data, but I find out the data is very skewed not on a normal curve which implies that the number of followers has a very large number of outliers
- In this case, **outliers are very important and very interesting to discuss** it, since this is an indicator for how much is the influence of users in each data set. In other words, the question which data set has more influencers which is a good metric to know people opinions about immigration.
- This visualization is created by build in function in python and its parameter is the number of followers for each distinct user in my data set.

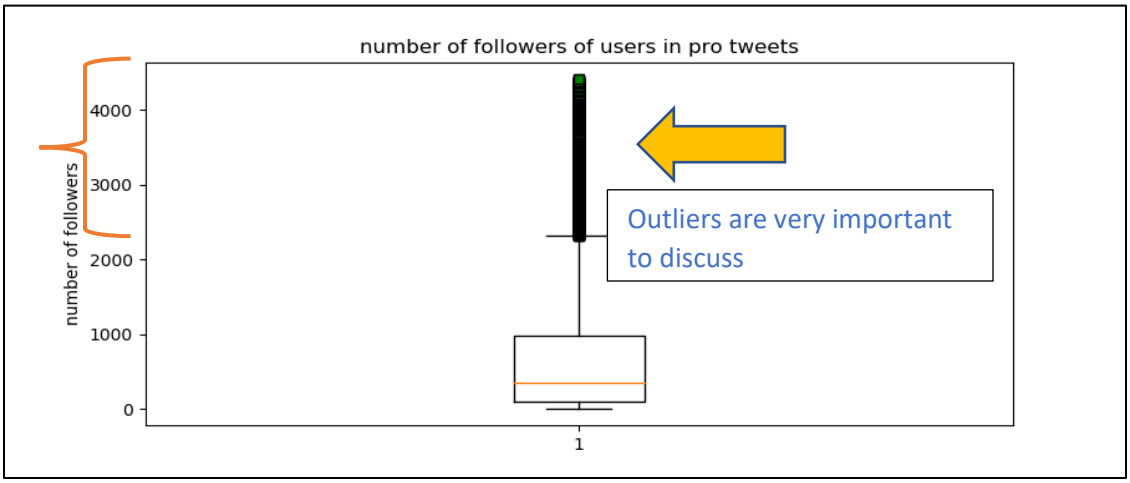


Figure 5 : boxplot of number of followers in pro immigration tweets

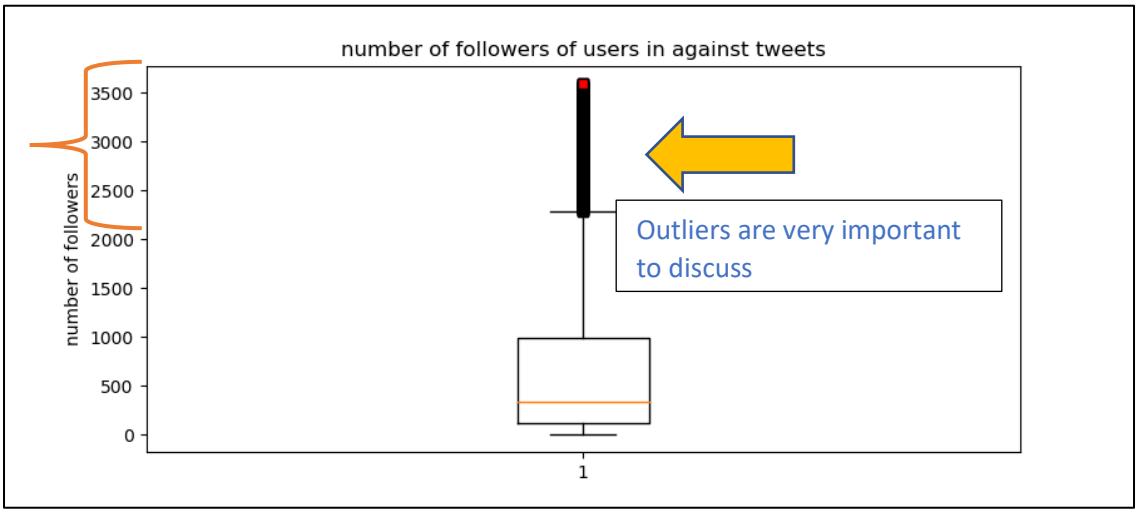


Figure 6 : boxplot of number of followers in against immigration tweets

- From the two plots above we can figure out that the two plots have almost the same number of the outliers. In other words, I can say that the influencers on both sides are almost the same percentage.
- I can say that this is true reflection for the real situation now, since there is no dominated opinion about immigration issues and people are debating till now.

2.5.4 Users influence level in dataset:

- In this part, we will see the distribution of users in each dataset according to their influence level, which is indicated by the number of followers each of the users had on their Twitter account,
- We can see that almost the distribution of each discretized value between two data sets are almost the same number (little bit difference)
- This figure is plotted using the ready function in python to make an ordinary bar plot, this shows how many users in each category of users according to the number of followers

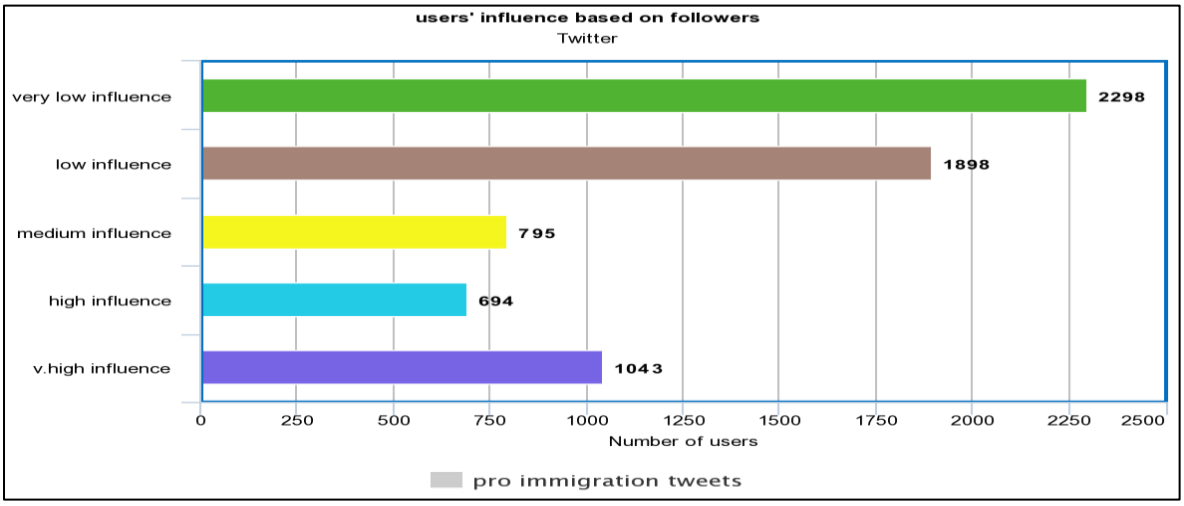


Figure 7 : users' influence in pro immigration dataset

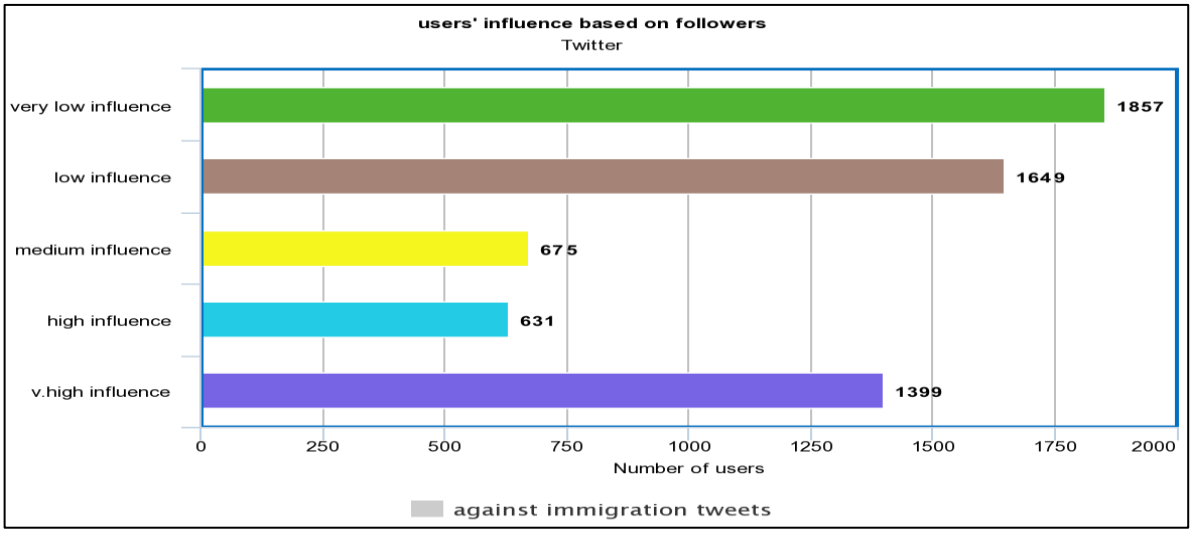


Figure 8: users' influence in against immigration dataset

- In the previous figure, we can see that there is a difference between the number of users in low influence level is little bit higher that the one in against immigration. However, the number of v. high level influence is greater in against immigration dataset and the middle three levels are almost the same with little difference.
- We can see that the interpretation from this visualization matches what is concluded earlier in 2.5.3 about the dominance of any of the point of views in the society .Cleary we can see in the figures that both dataset have almost the same number of very high level influence users.

2.6 Conclusion

- From the results obtained in this project, it is shown that not always the data extracted from real world (Twitter) with a specific search or filter criteria like here keywords will match the expectation. As we see in the preprocessing phase, the data has a lot of irrelevant tweets to be removed from the dataset since they are irrelevant to the topic.

- Another major problem when dealing with real data is the noisy data that are commonly found such as news hashtags and generic common hashtags. The noisy data are not useful in most of the cases.

- Doing the analysis on five different days showed that data are highly dynamic and when dealing with life stream data as twitter it is very hard to expect what is coming. In some cases, some event might be completely irrelevant, but it changes the coming data distribution significantly.

- building any solid information raw data needs lots of filtering and preprocessing prior to start using them. The portion of important data is small relative to what is obtained. These what led in this project to find that in most cases the most trending hashtag is less than or near 30 % of data found.

- I figured out also that the large number of tweets on September 19th which matches the events in the real life as people were tweeting about some debated happened during that time in USA which make the immigration hashtags are popular that day.

- Finally, the opinion mining on Twitter data helps us to analyze to analyze behaviors of people using social networks. The analysis of Twitter data is being done in various perspectives, the presence of words like good, bad and emoticons in the tweets can be used to infer the sentiment (not the scope of the project “advanced text mining”). The Twitter users can be classified into positive, negative and neutral users based on the followers and the followings and their behaviors can be studied based on the tweeting and retweeting activity. Tweets can also be used to analyze the influence factor in any debate and hence can be used to predict the results.