Data Science - Mini Project III

Md Saleh Ibtasham

Åbo Akademi University

November 8, 2022

This is a report on analyzing the popular social media platform Twitter. This report is written on the works for the analysis of the ongoing debates and issues that arise because of immigrants and immigration policies. It aims to clarify the classification of the sentiments of prominent Twitter Users and their sentiments on the issue. By the end of this report, a clear way to scrutinize the tweets on the basis of their sentiments and a visualization of the users representing those sentiments as well as the most representative words of those sentiments would be explained and explored.

1 Introduction

Nowadays countries are adopting the policy of immigration of several third-world countries in order to cope with the aftermath of various reasons such as pandemics, manpower crises, birth rate, climate change, etc. This project aims to shed some insights on the sentiments of popular **Twitter Users** on the notion of adopting immigration policies and regulations. People from different walks of life have a different points of view about these policies, both positive and negative. As there are people from both spectrums, their individual thoughts are expressed on social media sites, like, Twitter, Facebook, etc.

This project aims to shed some light on the Tweets that people make and determine whether the Tweet is positively supportive or negatively rebuking immigration policies. This would also clarify the important features of the Tweet, such as the subject, and the important keywords that the tweets are being posted about.

So the main contributions of this project are -

1. The collection of Tweets related to **Immigrants and Immigration Policies** and perform **Exploratory Data Analysis** and **Pre-processing** of the collected data.
2. Analysis of the sentiments of the collected Tweets and the exploration of machine-learning models to accurately identify the sentiment of the Tweets.
3. **Network Analysis** of the collected Tweets with respect to the **Twitter Users** and their specific choice of **Words**.

And naturally, the problem statement that this project is trying to address is as follows:

How can the most prominent tweeters and topics be found in the context of immigration?

2 Overview of the Project

This project has been inspired by many other projects that have been sought out for the completion of the **Network Analysis**. There are a series of steps in order to find the underlying network of the sentiments. The work here has been divided into 4 parts which have been explored in the later parts of this report. The 4 parts of this project have been divided into 5 sections.

Firstly, the **Data Collection** section describes the tools and API that have been used for collecting the data. **The Twitter API** in python called **Tweepy** and a generic library for scraping social network data called **SNScrape** have been used for querying and retrieving data from Twitter.

Secondly, the **Sentiment Analysis** section will shed light upon how the sentiment analysis was used in this work and prepared for the following sections.

Thirdly, the **Classification** section will explain the way the data was pre-processed into workable data. 2 machine learning approaches of classification were used in this part and will be explained in detail.

Fourthly, the **Word Cloud Generation** section will focus further on identifying the important phrases and

words used in the collection of tweets and also generate **WordClouds** to visualize the themes of the tweets.

Lastly, the **Network Analysis** section will provide an explanation of how the networks of the users and the most prominent words have been made. It will also clarify the underlying finding of the networks as well.

3 Data Collection

The data was collected from Twitter. As the tweets included in the data has to be related to the immigrants and immigration policy, the search strings below were fed into the tools that were used in the collection process.

- **Query 1:** #immigrant OR #immigration -filter:retweets
- **Query 2:** (#immigrant) -is=retweet -is:nullcast lang:en

These search strings were used as the general opinions and policies related to the issue of immigration has the hashtags **immigrant** or **immigration**. These search strings yielded different results for the data collection tools. So, for this project 2 tools were used in tandem to get a sufficient amount of data needed for classification and as well as forming a general network analysis. To that end, **Tweepy** the python **Twitter API** library was used at first for data collection. The *Academic Access* of the Twitter API was used for retrieving the data. As historical data from the **Twitter Archive** was accessible but not quite explanatory, the project ended up using the normal search API that is available for public use. The **Twitter Archive** API could not be used as the full-text body of the Tweet was not accessible without paid access to the API. But using the normal search posed another problem, i.e., it only provided the Tweets 7 days prior to the search date which didn't provide much data. Roughly **5000** Tweets could be collected with this method.

So to solve the issue of limited data, this project then used a very efficient scraper library custom-made for social media, i.e., **SNScrape**. SNScrape could be used to collect tweets as early as 2010. So with the use of both the queries mentioned above this library managed to scrape about ***190000 Tweets*** related to the issues of immigration.

The collected data is saved using the CSV format and provided at this link[1].

4 Sentiment Analysis

Sentiment Analysis is generally used for exploring the positive and negative aspects of a given text. The sentiment in this project means what are the positive and negative opinions of the Tweets with regard to immigration laws and immigrants in general. The sentiments also provide a strong understanding of the moral standings of the corresponding users. Through the analysis, this project aims to identify the strong word representing the sentiments from both ends of the spectrum and the representative users behind the strong feelings about the policies.

Table 1: *Sentiment Types of the Tweets*

Sentiments Types	
Sentiment	Range(Compound Score)
Overly Negative	-1 – -0.60
Slightly Negative	-0.60 – 0
Neutral	0 – +0.20
Slightly Positive	+0.20 – +0.60
Overly Positive	+0.60 – +1.00

The sentiments have been analyzed using *NLTK* library's *VADER* sentiment analyzer. **VADER** was used in this case because VADER accurately takes into account the linguistic aspects of social media. As public opinions in social media heavily use literary terms a sentiment analyzer fine-tuned for social media use is by far the most plausible option to take.

The scores of each tweet in the sentiment analyzer are from a compound perspective of the positive, neutral, and negative aspects of the tweet itself. So the score represents the polarity of the sentiment of the tweet. Taking in the aspect of polarity and the compound score the sentiments have been divided into 5 categories, i.e.

Here the lower values of the range are inclusive in the tweets. The range of the *slightly negative* sentiments is pushed to 0. The reason behind this is that negative sentiments are scarce in the dataset and this scarcity will pose a class imbalance to the classification of the sentiments. To provide a proper balance to all the sentiment classes of the dataset the range here is increased by **.20** which is still logical as the compound score of the sentiments is still negative.

5 Classification Models

The classification module of the project required 2 parts in the process. The ***Text Vectorization*** & the **Classification**.

5.1 Text Vectorization

The tweet data collected with the API is not ready for use for the classifiers. The underlying textual data has to be quantified in some way to feed the classification models. The ***Tf-Idf*** text vectorizing algorithm is used here in the project to quantify the importance of each tweet in the data. **Tf-Idf** is used because it accurately obtains the weights of each corresponding word in the document with respect to the document and as well as

Table 2: *Classification Accuracy*

Classifiers	
Classifier Type	Accuracy(%)
Linear Regression	73.42
Artificial Neural Network (ANN)	74.89

the whole database itself. The number of features used for the vectorizing is **4000**. The text vectorization part was inspired by this article [2].

5.2 Text Classification

The text classification was done to provide classifiers for the sentiments of tweets. It is true that the ***VADER*** sentiment analyzer could be used for sentiment analysis, but these classifiers provide the classification of tweets specifically related to immigration rather than focusing on generic topics. So the classifiers were fine-tuned for tweets related to immigration topics solely. 2 Classifiers were built into the project. They are-

1. The Linear Regression Classifier
2. The Artificial Neural Network (ANN) Classifier

As the **Tf-Idf** vectorization of the tweets carries useful information about the data, the accuracy of the classifiers yielded a good result. The classification report of the classifiers is given above. As can be seen in the table above, both the classifiers have accuracies well beyond 70% which is a good score considering the textual data from Twitter is sometimes difficult to understand.

6 Word Cloud Generation

Word Clouds can easily make the visualization of the underlying topics in a vast amount of text quite easy. The most used words in the text can be highlighted in a word cloud and the most prominent of those words would be highlighted with a bigger font and bold colors.

The word cloud generation as well as the coding of the word clouds followed the concepts of this article [3] and the python code mentioned here[4].

In order to filter the stopwords, this project used the stopwords from **NLTK** and used some curated stopwords specific to the problem in question. Some common words that need to be omitted from the tweets include, *immigration, now, immigrant* etc. As these are tweets on the issues of immigration, naturally those words would appear more often than others. The word cloud generated from all the tweets can be seen below.

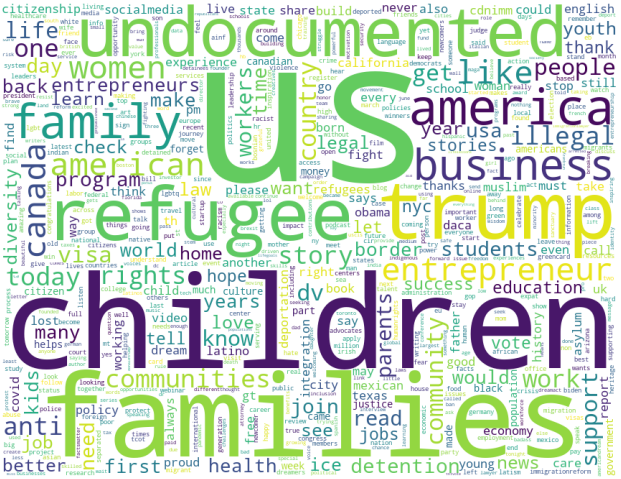

Figure 1: *Most prominent words from all of the tweets*

Naturally, the words, such as *america, us, children, refugee* can be seen in the cloud as these locations and issues are most prominent in regards to immigration.

In order to further view the words most used from a positive point of view the positive tweets had been separated and have been made into a word cloud to visualize the positive aspects of the topic.

Figure 2: *Most prominent words from the positive tweets*

Words like *entrepreneur, support, like* are most prominent as the immigrants often tend to tweet when they achieve something on foreign soil. The negative end of the spectrum was equally graphic in the word cloud generated from only the negative tweets.

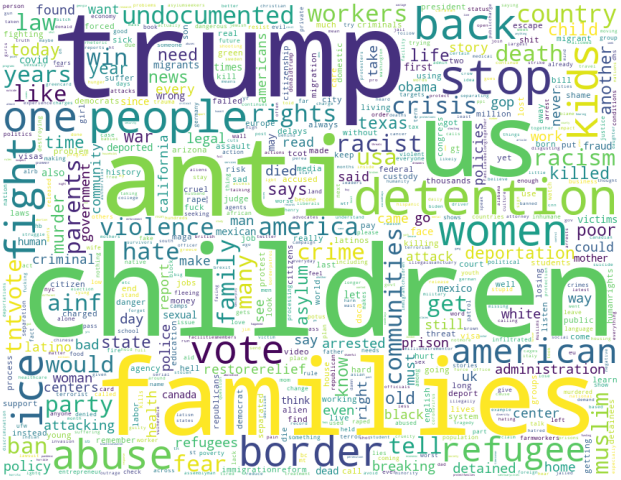

Figure 3: *Most prominent words from the negative tweets*

Words like *trump, fight, anti, ice, detention* are most prominent among the users who generally do not like the idea of foreigners coming to their country and converting into an immigrant.

7 Network Analysis

7.1 Data Preparation

For the Network Analysis to be fruitful, the tweets had to be vectorized again. But this time another vectorizing algorithm was used, i.e. ***Doc2Vec***. **Doc2Vec** vectorization takes the whole text in question into account to vectorize the texts while maintaining the inherent meaning of the texts. This is especially useful as the **Network Analysis** would include clustering the users into their respective sentiment groups of the network. The works of this GitHub resource[5] were especially useful in quantifying the tweets into their corresponding vectors.

7.2 Clustering with Scatter Plot

The vectors created from the previous vectorization technique were useful for making the scatter plots. The 2D and 3D views of the scatter plots are as follows:

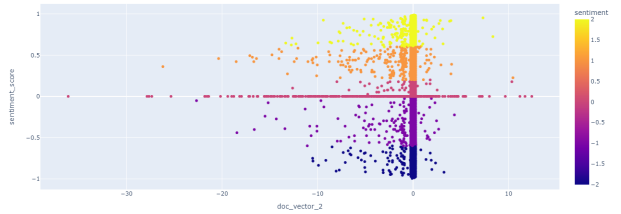

Figure 4: *2D view of the different types of Tweets*

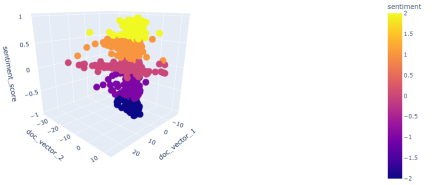

Figure 5: *3D view of the different types of Tweets*

In the scatter plots, the most yellow dots refer to the most positive, and likewise, the bluest dots refer to the most negative sentiments. The dots themselves represent each individual user in the vector space. The vector space has been reduced to 2 dimension vectors for the tweets to be able to plot in the graphs and make clusters.

7.3 Network of Twitter Users

For generating the network of Twitter Users, the unique users had to be first filtered from the main dataset. Then after filtering the unique users their frequency of tweets had to be measured in order to give an order of magnitude of how prominent a user is in the given context. The styles and generation of the networks are inspired by the works of this article[6] and this resource[7]. Below are 2 networks with and without the usernames of the users that share positive and negative sentiments regarding immigration issues.

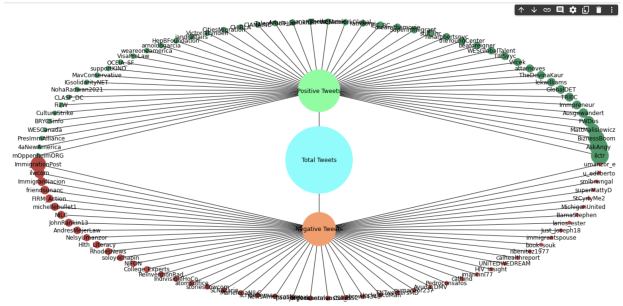

Figure 6: *Network of most prominent users from both views (with username)*

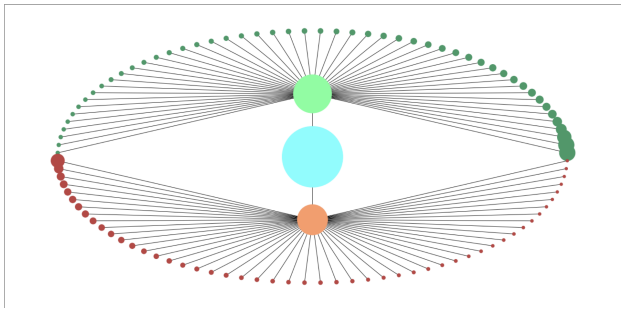

Figure 7: *Network of most prominent users from both views (without username)*

Here we can see that the most positively and negatively viewed users are both at opposite parts of the networks. This shows that their frequency of tweets is both larger and farther than that of the less positively or negatively viewed users.

7.4 Network of Prominent Words

In order to generate the network of prominent words, this project employs the most frequently used words that have been generated from the word clouds previously mentioned in the **Word Cloud Generation** section. The most prominent words from each of the views have been taken as representatives of each view. The 2 networks both with and without the words prominent words are stated below.

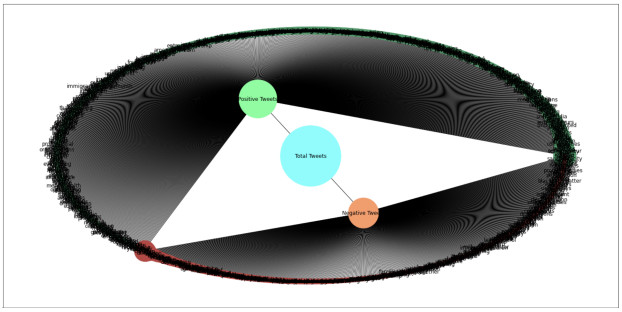

Figure 8: *Network of most prominent words from both views (with words)*

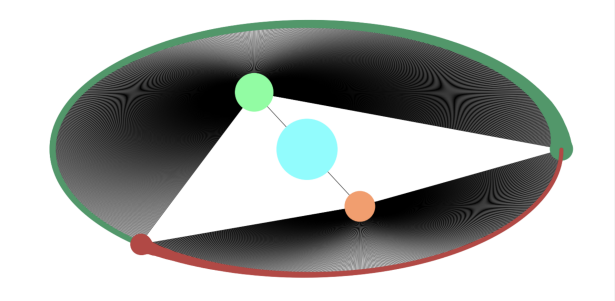

Figure 9: *Network of most prominent words from both views (without words)*

The challenge here was that some words were present in both the positive and the negatively viewed articles and they were both prominent. To solve this problem the union of both sets of words was taken and a reduction was made so that both views represent unique words. The reduction scheme was to take the common prominent words and afterward include them in the view where it is used the most.

8 Conclusion

This project aimed to satisfy the problem statement "**How can the most prominent tweeters and topics be found in the context of immigration?**". In order to find the most prominent users and topics regarding immigration, classifiers were built and word clouds have been generated for visualizing the most prominent sentiments. Afterward, the clustering of the users and generation of a network that uses the magnitude of the tweets made by each user and the frequency to which the words appear in the tweets were made to accurately identify the Twitter users and topics that have played the most significant role in the issues of immigration.

References

1. *Twitter Data* `https://drive.google.com/file/d/1QpCAZDiegeeCA1NLxgk9ystEmTnZUd7B/view?usp=sharing`. Accessed: 2022-11-08.
2. *Twitter Sentiment Analysis* `https://towardsdatascience.com/a-step-by-step-tutorial-for-conducting-sentiment-analysis-a7190a444366`. Accessed: 2022-11-08.
3. *Word Cloud Generation* `https://towardsdatascience.com/generate-meaningful-word-clouds-in-python-5b85f5668eeb`. Accessed: 2022-11-08.
4. *Word Cloud Code Resource* `https://github.com/bryan-md/wordcloud`. Accessed: 2022-11-08.
5. *Document Vectorization Resource* `https://github.com/LouiseLilyJohn/NLP-Twitter-hate-speech-detection`. Accessed: 2022-11-08.
6. *NetworkX Customizing* `https://towardsdatascience.com/customizing-networkx-graphs-f80b4e69bedf`. Accessed: 2022-11-08.
7. *Networkx Code Resource* `https://github.com/sepinouda/Intro_to_Data_Science`. Accessed: 2022-11-08.