The Chirping of the Country

A broad descriptive analysis of USA Midterms 2022 through Twitter lens

Matteo Scianna

> *"Politics is the art of looking for trouble, finding it everywhere, diagnosing it incorrectly, and applying the wrong remedies."*
>
> —Groucho Marx

> *"No, that is Data Science."*
>
> —Me

Abstract

The purpose of this paper is to exploit Twitter data in order to perform sociological analysis. In particular, the frame chosen for this work are past USA midterm elections held on the 8th of November 2022. After properly collecting tweets using specific and suited techniques, text was cleaned in order to obtain a proper material for the analysis proposed. Then, most frequent words and hashtags were extracted and visualized, together with the evolution through time of some important endorsement hashtags. Additionally, a network analysis was performed in order to identify the most frequent topics, using techniques such as skipgram analysis and community detection to connect words that appeared together and identify communities of related words.

1 Introduction

USA midterm elections are a very important moment in the American socio-political sphere. Every four year, people are called to vote for half of the Congress members and it is a crucial situation to test the stability of the majority. 2022 USA midterm elections were held on November 8th and in this paper, we propose a content and topic analysis of this phenomenon through Twitter.

The first part of the paper refers to the collection and cleaning process: a sample of 200,000 tweets from November 3rd to 14th were collected using different kind of hashtags and divided into three time periods. The text of the tweets was subsequently cleaned to avoid misinterpretation, eliminate words useless for our goal and to merge words with the same meaning. Sections 2 and 3 refer to this process.

The last part of the paper regards the actual analysis (Sections 4 and 5): a quantitative analysis was performed to identify the most frequent words and hashtags, which were visualized using bar plots and wordclouds. Furthermore, a network analysis was performed using skipgrams to identify the biggest connected component and detect communities.

To conclude, this works aism to show the strong potential that Twitter data and data analysis tools have in order to provide a better comprehension of social phenomena.

2 Data Collection

The data from this study were collected using Twitter API in a time range of 14 days centered on the elections day (November the 8th). All tweets and retweets containing some specific hashtags were collected. Those hashtags followed the following criteria:

1. **General topic related tweets**: all tweets and retweets containing an hashtags related to the midterm elections event were collected. Hahstags selected here are: *#midterms*, *#midtermelections* and *#midterms2022*, as *super partes* hashtags with a direct reference to the event.

2. **Left wing endorsement tweets**: all tweets and retweets containing a direct endorsement hashtags in favour of democrats were collected. Hashtags selected here are: *#voteblue* and *#votedemocrat.*

3. **Right wing endorsement tweets**: all tweets and retweets containing a direct endorsement hashtags in favour of republicans were collected. Hashtags selected here are: *#democratshateamerica*, *#redwave* and *#votered.*

Starting from a web scratch for official hashtags of miderm elections, this selection derived from a Twitter scraping "in-field" analysis. Time was spent in order to find out which were the most popular hashtags in order to obtain a number of tweets that could embrace as much as possible the phenomena. Furthermore, in order to have a balanced number of tweets endorsing the two parties, a different number of hashtags was chosen.

In the end, all tweets were merged together resulting in a 196.466 tweets dataset, subsequently splitted in three different dataset according to time ranges. The first dataset (pre-elections) contains tweets from 1st to 6th of November (32.550 tweets); the second one (during elections) contains tweets from 7th to 9th of November (140.093 tweets); while the last one contains tweets from 10th to 14th of November (23.823 tweets). Fig 1 shows the evolution of the number of tweets per minute through time.

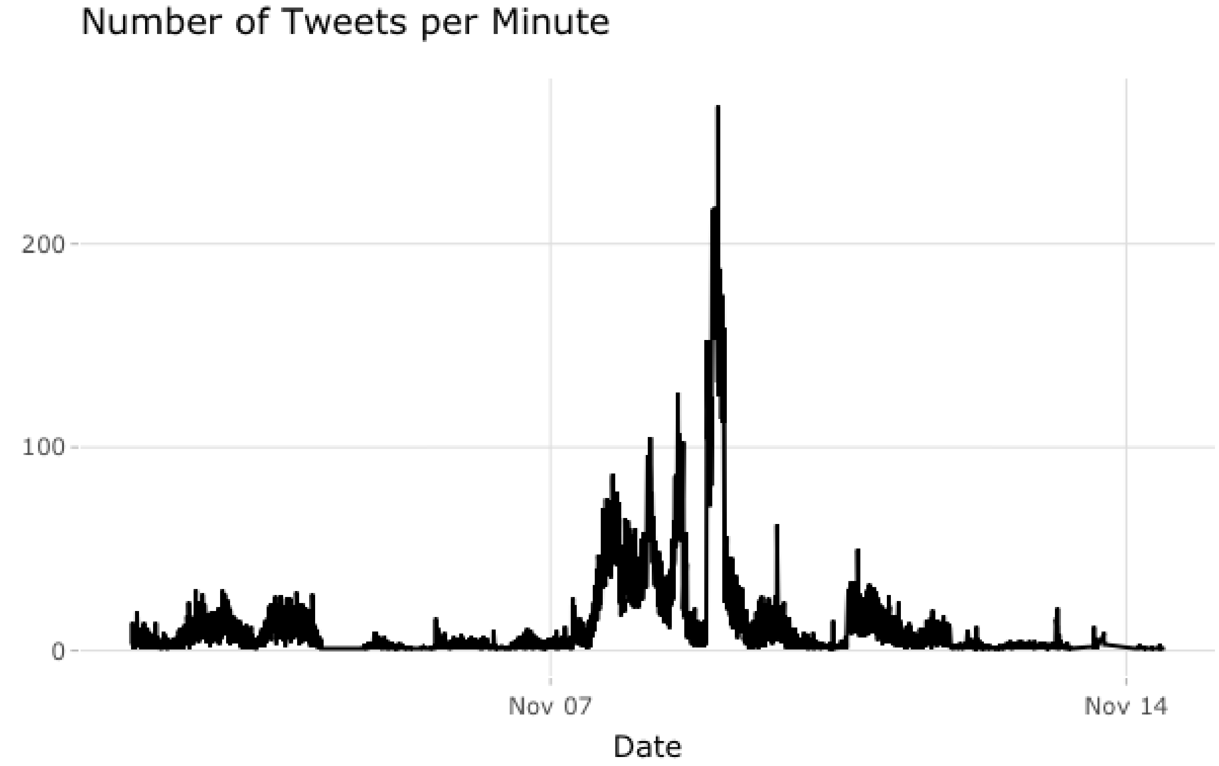


Figure 1: Frequency of Twets per minute

3 Data Cleaning

Before doing any kind of analysis, all tweets needed to be cleaned.

First of all, since tweets collected covered a quite wide time range, the `created_at` column was substitued by the time rounded by minutes.

Furthermore, a massive text cleaning was performed, in order to avoid misinterpretation and improve the analysis. Text was converted to lowercase and many different kinds of characters, like `rt`, urls, emoticons and other special characters were removed.

Then, in order to perform a first hashtag analysis, all different hashtags with the same meaning were merged together. To give an example, *#redwave2022*, *#redwave* and *#redtsunami* were all merged in *#redwave*. This was done for 45 different hashtags. In order to have also an idea of which were the most used hashtags

a part from the ones we seeked for, hashtags used for collection purposes were then removed.
Furthermore, text was cleaned from a semantic point of view, merging together different words corresponding to the same concept. This was done for different kind of words:

1. **Often mentioned people**: we wanted, for example, to refer *joe biden*, *biden*, *bidens* etc. to the same entity. This was done for circa 70 different entities.

2. **Usa and mitderm elections**: *USA*, *united states*, *america* were very frequent words referring all to the same concept, as like *elections*, *election*, *midterms*, *midterm elections*..

3. **Republicans and Democrats**: there are mani different ways in which these terms are referred to (*dems*, *blue*, *left*, *reds*..).

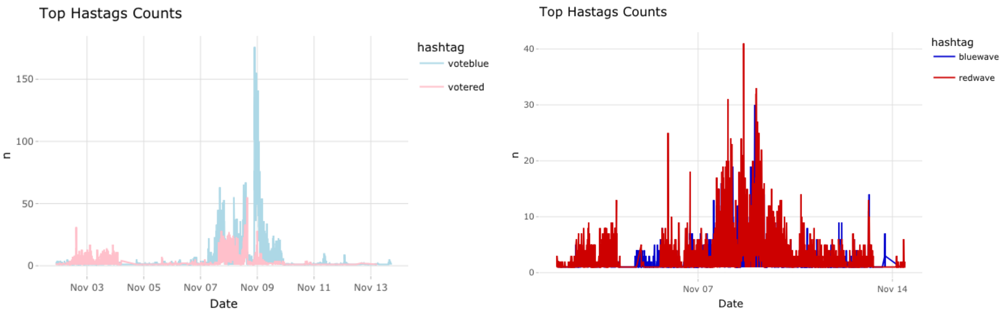
4. **Different words contribuing to a single concept**: Words like *wall street journal*, *make america great again*, *new york* etc. need to be considered as a single entity.

This process was done more or less for each word appearing more than 100 times.
Finally, stopwords were removed, together with useless white spaces, punctuation and accents, and only tweets containing at least three words were considered.

4 Quantitative analysis

In order to have a clear general idea of most frequent words and hashtags, a first quantitative analysis was performed. It is divided into three parts:

1. **Evolution of endorsement hashtags through time**: The most used endorsement hashtags for the two parties were *#redwave*, *#bluewave*, *#votered* and *#voteblue*. A plot with the evolution of the frequency of those hashtags through time is presented in Figure 2.



(a) Votered vs Voteblue

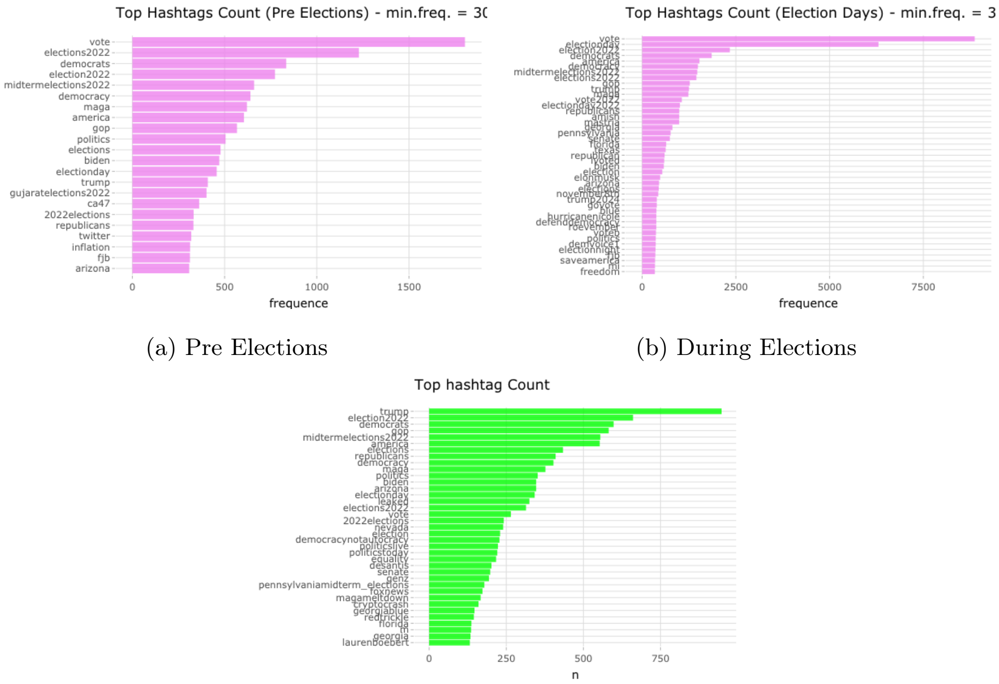
(b) Redwave vs Bluewave

Figure 2: Endorsement hashtags evolution

We see that the two plots are quite different and that might sound conterintuitive in a first look. In particular, what may sound surprising is the massive presence of *redwave* hashtags also after the elections, which we remind were not such a win for the republicans. Analyzing the tweets containing *redwave* hashtag dated right after the elections, it has been understood that lots of them are ironic tweets, mocking repbulican results. Some examples can be seen here and here.

Furthermore, while *redwave* and *bluewave* hashtags remained frequent also after the elections, *voteblue* and *votered* hashtags drastically reduced frequency, after a clear peak during election day and hours immediately following. This can be explained by noticing that these are actually "vote declaration" hashtags and so quite pointless after the elections took place.

2. **Most frequent hashtags**: for each of the three datasets, a barplot containing the most frequent hashtags is presented, as shown in Figure 3. It is important to underline that from the hashtag corpora all hashtags from which we downloaded the tweets are removed. Since these hashtags are of course the most frequent ones, this process allows us to have a general idea of the main topics without the noise of frequent endorsement or miterms hashtags.



(a) Pre Elections

(b) During Elections

(c) Post Elections

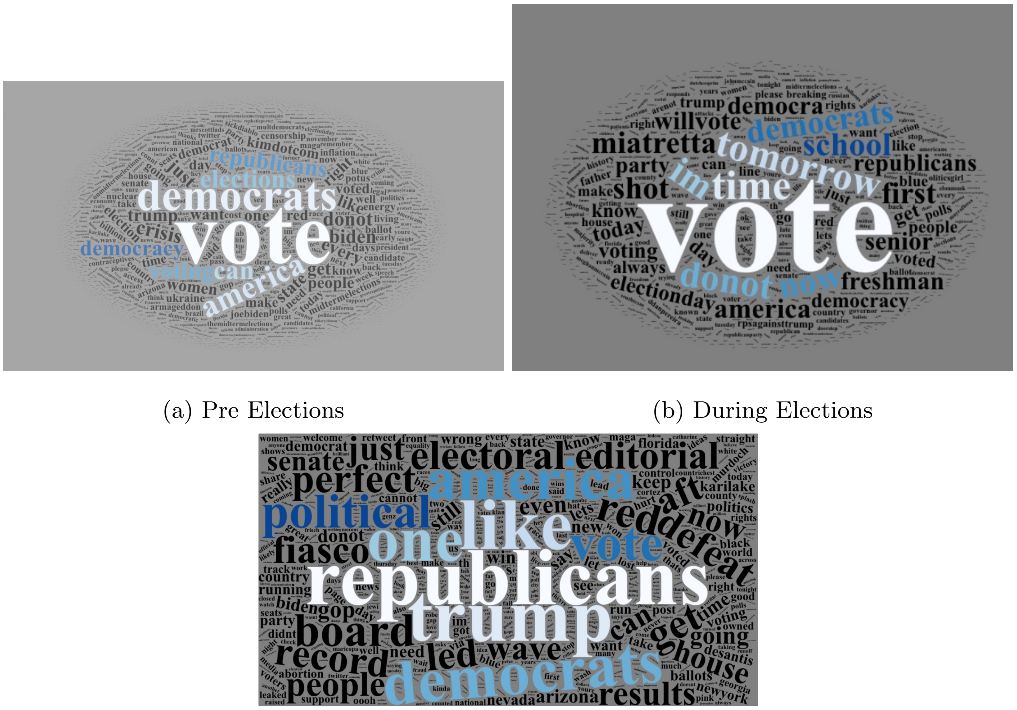
Figure 3: Top Hashtags Barplot

While some similarities are of course present (*#vote*, *#mitermelections2022* are clearly widely diffused across different periods), it is possible to notice some interesting happenings: just to cite some, it is surpising how *#trump* was sharply the most used hashtag after elections took place, appearing almost four times more than *#biden*.

Furthermore, after elections took place hashtags referring to some specific political situations started appearing, like all hashtags of USA states names, but also *#laurenboebert* and *#desantis*, indicating that (as predictable), after a very general discourse regarding midterm elections and USA as a whole, different topics started focusing on specific situations once elections took place.

3. **Most frequent Words**: starting from the clean text, for each of the three dataset a wordcloud containing the most frequent words is presented, as shown in figure 4. It is interesting to see how of course some words are constant within the different periods, but there are many significant differences, in particular between the pre-during datasets and the post elections one, where words like "trump", "fiasco", "defeat" appear.

   Of course this analysis allows us to have just a partial idea of the main topics. In order to have a more in depth view, a network analysis is performed.



(a) Pre Elections

(b) During Elections

(c) Post Elections

Figure 4: Top Words Wordcloud

5 Network Analysis

Network analysis was performed using skipgrams to identify the words that occurred most frequently together in the tweets. A skipgram is a sequence of words that occur in a certain order within a text. For example, the skipgram "vote for candidate" would indicate that the words "vote", "for", and "candidate" occurred in that order within a tweet. By analyzing the skipgrams in the tweets, we can identify the words that are most commonly associated with each other and subsenquently have an idea of the most frequent topics occuring. The biggest connected component of the network was identified and extracted. Furthermore, for each network the degree, betweenness, and closeness centrality measures were analyzed. These measures indicate how important a word is within the network, based on the number of connections it has and its position in the network. Then, the Louvain method was used for community detection, which identifies groups of words that are more densely connected than others. This allowed us to see how the topics in tweets were related to each other, and to identify any subtopics within the data.

For each dataset, four figures are presented with a first view of the network (Figures 5a, 6a and 7a), the network biggest connected component, barplots for centrality measures and a screenshot of a dynamic view of the network itself[1], where nodes are coloured according to the community they belong to.

Furthermore, in order to have comparable networks, not all tokens have been considered but only a top-n percentage, varying from dataset to dataset, that we remind are different in size. Indeed, it is shown in the pictures that the weight threshold for the pre-elections dataset is 300, while for the second one is 1300 and for the third 250. What follows is a general analysis of the main results that can be derived from these networks.

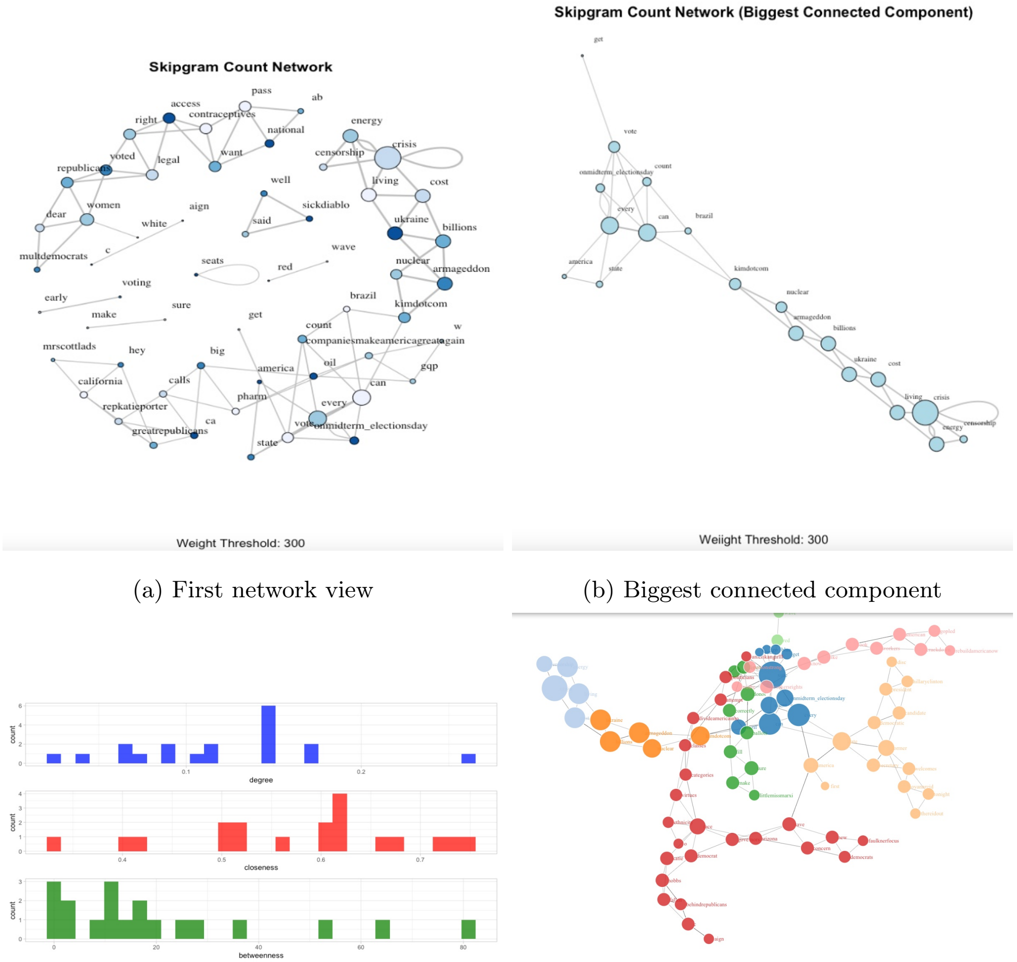
5.1 Pre Elections Period

From a first view of picture 5a, we can see two main different connected components. The top-left one refers to a set of tweets on right wing politicians' view regarding contraceptives and abortion, an argument often mentioned by democrats. The tweet leading this topic can be seen here.

For the other connected component, also highlighted in picture 5b, the situation is a little more difficult. In particular, the top-right part containing words like *Ukraine*, *Nuclear Armageddon* etc. comes from this tweet which is very critical to many different choices of the current government. This is a classical case in which the only way to completely obtain the meaning of the topic is to go back to the original tweet: picture 5d indeed fails in order to detect the correct component, since it separates in light blue and orange (top-left) words actually regarding the same topic.

It has been decided to include this example since it is quite explanatory of the methodology adopted for this project: starting from raw tweets, Data Science instruments can be and indeed are very useful in order to detect and highlight information, but then the only way to completely grasp knowledge is indeed to go back to the source with new information gained.

[1] The actual dynamic version of the network is available here for pre-elections network, here for during-elections network and here for post-elections network (In order to be properly visualized, the file needs to be dowloaded and loaded on a browser).



(a) First network view

(b) Biggest connected component

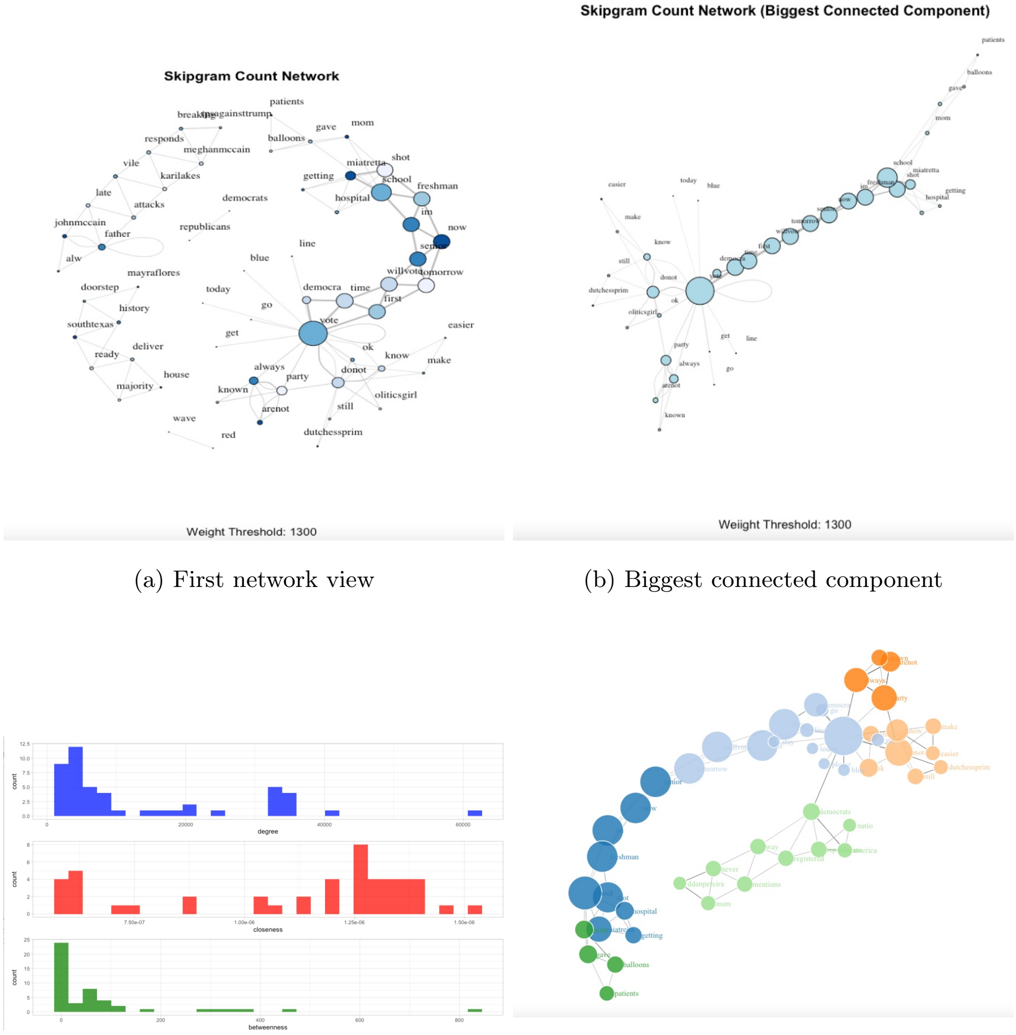
(c) Centrality Measures

(d) Network with communities

Figure 5: Network Analysis - Pre Elections

5.2 During Elections Period

From a first look at picture 6a, we can see that different topics are way better separated than before. Furthermore, we have only one community with a high number of occurrences, while all other topics and communities are less represented. The main topic, that corresponds also to a great part of the biggest connected component (picture 6b), is the one containing the words *vote*, *democrat*, *first time*, *freshman* etc. This comes from this very engaged tweet published on election day by Mia Tretta, a young woman endorsing Democrats which was victim of a school mass shooting when she was a student. This tweet had lots of engagement and retweets, resulting in quite a case during election day. Furthermore, there are two other quite explainable topics: the one regarding *meghanmccain* and *karilake* refers to a strong argue between the political analyst and son of former republican representative John McCain and the candidate governor in Arizona for the Republican party Kari Lake. It would take a whole paper to explain all socio-political



(a) First network view

(b) Biggest connected component

(c) Centrality Measures

(d) Network with communities

Figure 6: Network Analysis - During Elections

conflicts between USA representatives, and this is not the place. So, without going any further, we can just say that this was indeed a very big topic during election days, correctly detected by our methods.

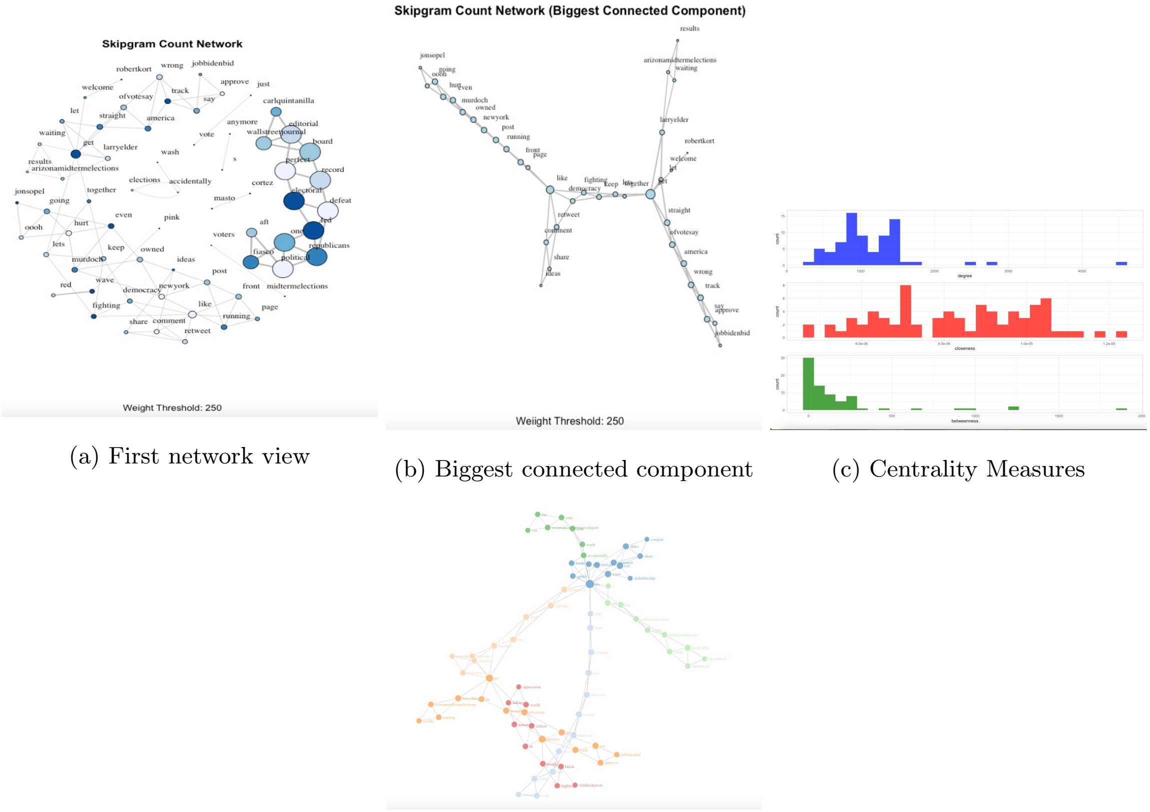
Finally the other well defined component is the one regarding *South Texas* and *Mayra Flores*, the republican candidate in that county. The topic itself refers to a tweet of the same Mayra Flores and do not need further explainations, but what it is worthy mentioning (and can not be extracted directly from our dat) is that Mayra Flores started being under the spotlight after this direct endorsement by Elon Musk in June.

Fig. 6d, finally, is very useful in order to detect properly which words belong to which topic. For example, while looking only at biggest connected component one is not sure if words "school", "balloons" "patients" are part of the Mia Tretta topic, community detections explain quite well that that is a completely different community and topic.

5.3 After Elections Period

We can see that topics and words present in the network appear quite different than in the previous ones. In particular, the most frequent topic refers, naturally, to the elections results, consisting in, if not a loose, surely not in a win for the republican party. This tweet received a very strong engagement, becoming the reference point for lots of voters to stress the bad result of Republicans and, in particular, Donald Trumps' loss as former representative of the party (remember picture 3c where trump was indeed one of the most frequent words in the post elections period).

Finally, it is worth stressing that specific terms referring to specific situations and people appear more often than before (Murdoch, Arizonamidtermelections, Jon Sopel etc.), implying the presence of more spotlights for Twitter users.



(a) First network view

(b) Biggest connected component

(c) Centrality Measures

(d) Network with communities

Figure 7: Network Analysis - Post Elections

6 Conclusion

As mentioned before, the main goal of this project is to deal with Twitter for sociological purposes. To do so, different tools have been used to make as clear as possible all necessary steps to extract information from raw data and subsequently gain demanded knowledge.

In particular, it has been firstly shown how to properly handle raw text data in order to get rid of noise and useless information and actually obtain a good starting point to perform various analysis. Furthermore, different kind of textual analysis have been proposed, from a more impactful and visual representation of most frequent words and hashtags to a network representation of words to stress and highlight frequent topics through people. Finally, in order to fully understand all spotlights enlighted thanks to the instruments used, it is necessary to take a step back and observe again with social lens what emerged.

To summarize, what this work aimed to show is the possibility to be able, starting from an unkwnown social phenomenon, to obtain actual knowledge regarding many parts of this phenomenon, properly exploiting suited tools.

Side note: Some of the plots presented in this paper are screenshots of a dynamic version of the same. Upon request, all the original files can be shared, together with the R scripts containing all codes used for the processes presented in this work together with the original dataset.