How much Americans like Joe Biden? A network analysis to estimate the outcome of midterm election

Simona Mazzarino
s.mazzarino@studenti.unipi.it
Student ID: 620022

ABSTRACT

On the 8th November 2022, the midterm elections will take place in America in order to assess, to some extent, the action of the U.S. President, Joe Biden and his party. In this paper, I study, using Twitter data, the public debate around Joe Biden, before and after his appoint as President, in order to understand how the appreciation for him has changed over time and in order to estimate the outcome of these elections. To do so, network science and natural language processing (NLP) tools are used.

KEYWORDS

Joe Biden, Midterm Elections, Twitter, Social Network Analysis, Natural Language Processing, Topic Modeling, Sentiment Analysis

1 INTRODUCTION

The following November, the U.S. population is called to vote in the midterm elections. But what do we refer to when we talk about midterm elections? When we talk about midterm elections, we refer to a kind of political elections where people can elect their representatives and other sub-national officeholders in the middle of the term of an executive. In addition, in the U.S., the midterm elections are sometimes regarded as a referendum on the sitting President's performance. So, in order to estimate the outcome of the upcoming midterm elections, I studied, starting from Twitter data, the public debate around the current President, Joe Biden, combining two different kind of analysis. On the one hand, I examined using network science techniques the network of Twitter users, which I used as a sample of the real American population; on the other hand, I analyzed, with some NLP tools, the tweets written by the users in the network (and in the communities found in it) in order to understand which were and which are the main topics covered by the American public debate and how much Americans appreciated and appreciate the action of their President. In section 2, I described how data were collected and how the network was created. Then, section 3 reports a basic network analysis and the comparison with other synthetic networks. Sections 4, 5, 6 and 7, instead, are about advanced analysis, such as, respectively, *community discovery*, *dynamic community discovery*, *opinion dynamics* and *unsupervised link prediction*. In sections 8, I performed the NLP analysis (i.e. the topic modeling and the sentiment analysis) on the Twitter posts. In section 9, I discussed the results.

2 DATA COLLECTION

2.1 Scraping, Data Cleaning and Data Understanding

First of all, in order to reach my goal, I decided to scrape data from Twitter using *Twint*, a Python library designed to easily collect data from Twitter, which allows to download data using a keyword[1]. I defined the data collection strategy as follows: I chose, as keywords, the hashtags *#JoeBiden* and *#Biden* and, then, I set the scraper in such a way that it gathered a maximum of 1000 tweets per day from January 2020 to June 2022. After filtering the tweets, in order to have just the ones written in English, I obtained 674494 tweets about Joe Biden, before and after his appoint as President. Then, for following analysis, I divided the collected data into five semesters, according to the date of posting. In this way, I obtained, for each semester, the following amount of tweets:

Figure 1: Distribution of the number of Tweets from January 2020 to June 2022

- **First Semester** (January 2020 - June 2020): 144711;
- **Second Semester** (July 2020 - December 2020): 147576;
- **Third Semester** (January 2021 - June 2021): 125386;
- **Fourth Semester** (July 2021 - December 2021): 132830;
- **Fifth Semester** (January 2022 - June 2022): 123991.

A temporal distribution of the tweets is shown in Figure 1: as it can be seen, people tweeted more about Joe Biden in the 2020, the year of his appoint, rather than in 2021. Then, the social activity concerning Joe Biden increased around the beginning of the current year. After these previous analysis, I created the network. I considered as nodes all the users who have written a tweet in the dataset. The edges, instead, were created as follows: every time that a user replied to another one, then, between those two users, a directed link was set. Moreover, each edge was weighted according to the number of times that a user replied to another one. An example is reported in Table 1.

Project Repository: https://github.com/simonamazzarino/Social-Network-Analysis
[1]Twint repository: https://github.com/twintproject/twint

Source	Target	Weights
scienceinvestme	JoeBiden	60
elnurrik3	POTUS	40
dmaga101	JoeBiden	40
slothsforme	POTUS	30
natashaejs	ProjectLincoln	28

Table 1: Example of the Network Edgelist

Graph	# Nodes	# Edges	Average Degree
RW	22737	32438	2.8533
ER	22737	32802	2.8853
WS	22737	45474	4.0
BA	22737	45470	3.9996
CM	22737	32438	2.8533

Table 2: Basic Measures calculated on the RW and on the synthetic graphs

3 NETWORK ANALYSIS

3.1 Basic Measures, Degree Distribution, Component Analysis, Path Analysis, Density and Clustering Coefficient Analysis

Once obtained the edgelist, the network (hereafter called RW, which stands for Real World) was created. Originally, the graph was a directed multigraph, meaning that there was more than one relationship between a pair of nodes and that those relationships were not mutual, but singular (i.e. if an edge between the node A and B existed, there wasn't necessarily an edge between B and A). In order to facilitate the analysis, I decided to convert the directed multigraph in an undirected multigraph. So, I obtained a network composed by 22737 nodes and 32438 edges. Then, I calculated some **basic measures** on the network and I compared the results with the ones obtained on four different synthetic networks, defined as follows:

- **Erdős-Rényi** (ER): with 22737 nodes and the probability p for edge creation equal to 0.00012535;
- **Watts-Strogatz** (WS): with 22737 nodes, the number of nearest neighbors k equals to 4 and the probability p of rewiring each edge equal to 0.2;
- **Barabási-Albert** (BA): with 22737 nodes and the number m of edges to attach from a new node to existing nodes equal to 2;
- **Configuration Model** (CM): using the same degree distribution obtained in the RW.

The previous parameters were set in order to match some features of the RW. Table 2 reports the number of nodes, the number of edges and the average degree for each network.
Next, I observed the **degree distribution** for each network (Fig. 2): the RW, the BA and the CM follow a power law distribution, which means that there are some hubs in the network (i.e. nodes with a very high degree), while the ER and the WS follow a Poisson distribution.

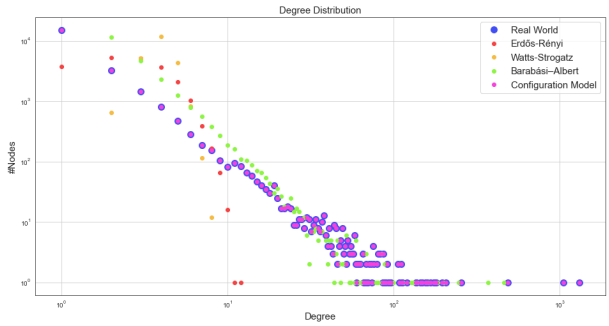

Figure 2: Degree Distribution of each graph

	Diameter	Average Shortest Path	Density	Average Clustering Coefficient
RW	15	5.2297	0.000125	0.001601
ER	23	9.3621	0.000127	0.000125
WS	18	10.528	0.000176	0.263806
BA	9	5.3325	0.000176	0.002674
CM	13	4.6659	0.000125	0.008187

Table 3: Diameter, Average Shortest Path, Density and Clustering Coefficient calculated on the RW and on the synthetic graphs

At this point, I focused on the **component analysis**. A component is a connected subgraph of the network which is not part of any larger connected subgraph. The results were the following:

- Number of connected components in the RW : 1182, where the giant component (i.e. the biggest connected component) contained 19675 nodes (which means that the 86.5% of nodes of the network were inside the giant component);
- Number of connected components in the ER: 1388, where the giant component contained 21180 nodes;
- Number of connected components in the WS: 1, where the giant component was the entire network itself;
- Number of connected components in the BA: 1, where the giant component was the entire network itself, as above;
- Number of connected components in the CM: 1927, where the giant component contained 18629.

Figure 3 shows a representation, obtained using *Force Atlas 2* layout on *Gephi*, of the giant component of the RW [2].

In addition, I calculated the **diameter** and the **average shortest path** in each graph. The diameter is the longest shortest path in a graph. I decided to compute such measures only on the giant component of each network. The diameter of the unconnected networks (RW, ER, CM), otherwise, would have been equal to infinity. Later, I computed the **density** and the **average clustering coefficient**. The results are shown in Tab. 3.

[2]In order to obtain such a color partition, the *modularity* of the graph was calculated by *Gephi* using *Louvain* algorithm.

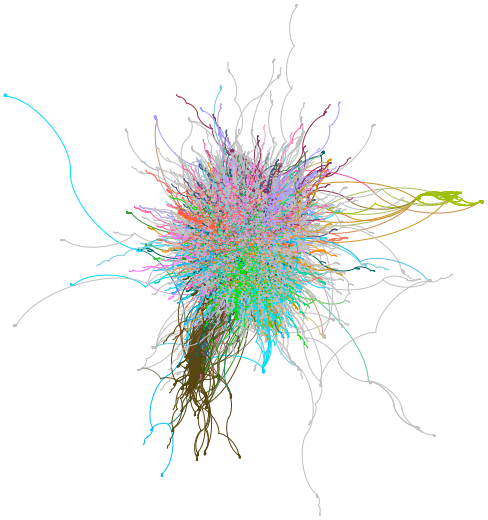

Figure 3: Visualization of the Giant Component in the RW

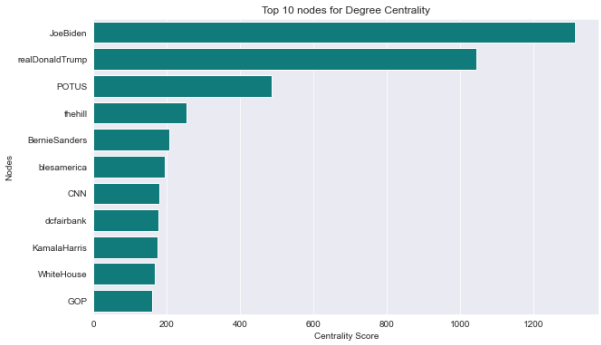

Figure 4: Top 10 Nodes for Degree Centrality in RW

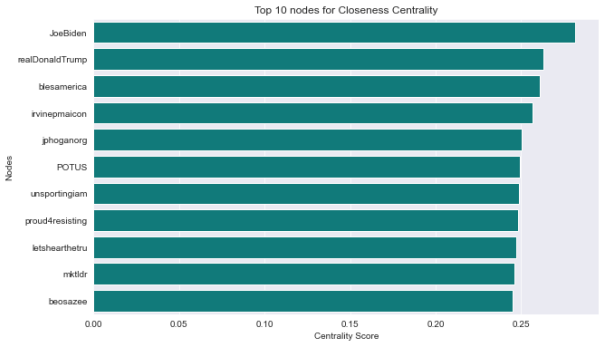

Figure 5: Top 10 Nodes for Closeness Centrality in RW

Focusing on the density results, it can be seen that all networks are sparse: in fact, the density value is very close to zero. Consequently, even the average clustering coefficients are really low, except for the value observed in the WS, which is slightly higher (in fact, the WS model tries to hold together the presence of large clustering coefficients and short distances, which are typical features of a real network).

3.2 Centrality and Assortativity Analysis

In this section, I focused on the computation of the **centrality** degree of each node in the network and on the **assortativity analysis**. The centrality of a node is the importance that such a node has in the network according to some measure. An easy way to compute the centrality of a node is to count its neighbors. An example of centrality is the degree centrality, which is computed simply by counting the neighbours of a node. Figure 4 reports the first ten nodes in the RW, according to their value of ***degree centrality***. Although very effective to understand the local structure of networks, the degree centrality does not capture some interesting aspects of complex networks. So, there are several ways to calculate the centrality, divided into two macro-families: connectivity-based methods, where the influence of a node is based on the number of links a node has to other nodes in the network, and geometric-based methods, where the importance of a node depends on some function of its distances w.r.t. other nodes.

3.2.1 ***Connectivity-based methods: PageRank and Eigenvector centrality***. The definition of centrality for the connectivity-based methods is a recursive one: in fact, important nodes are those that are connected to other important nodes. I implemented on the RW two of these methods: the ***PageRank*** method, which ranks the nodes according to the probability with whom a random walker visits that node (the higher the probability, the higher the centrality score), and the ***Eigenvector*** method, which computes the eigenvector associated with the largest absolute eigenvalue and derives the centrality score for each node from the scores of its incoming neighbors. The first five nodes (i.e. users) for *PageRank* centrality were: *JoeBiden* (0.0149), *realDonaldTrump* (0.0121), *POTUS* (0.0057), *thehill* (0.0029) and *mark3ds* (0.0022). Instead, the first five nodes for *Eigenvector* centrality were: *JoeBiden* (0.5708), *realDonaldTrump* (0.3302), *POTUS* (0.0987), *KamalaHarris* (0.0689) and *BernieSanders* (0.0657).

3.2.2 ***Geometric-based methods: Closeness Centrality, Harmonic Centrality, Betweenness Centrality***. The first geometric-based method that I performed was a method based on the concept of closeness. ***Closeness*** is the inverse of the farness, where the farness is the average of length of shortest paths to all other nodes. This means that the lower the farness, the higher the centrality score. If the farness is high, then the node is a peripheral node. The ***Harmonic centrality*** is the same, but it considers the harmonic average of length of shortest paths. The results obtained with the *Closeness* method are reported in Figure 5. Instead, the first five nodes for *Harmonic Centrality* are: *JoeBiden*, *realDonaldTrump*, *blesamerica*, *irvinepmaicon* and *POTUS*.

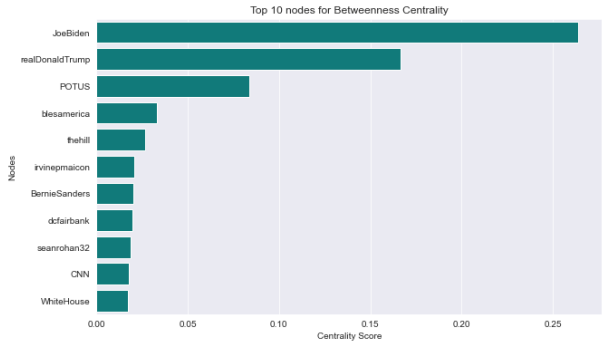

Figure 6: Top 10 Nodes for Betweeness Centrality in RW

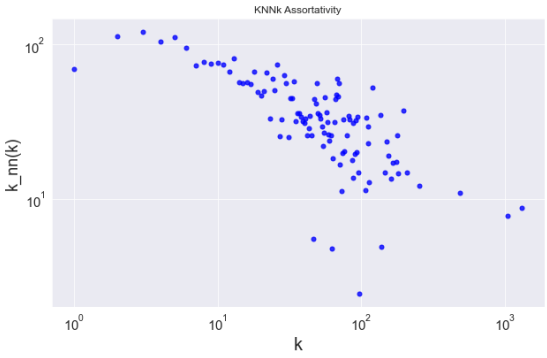

Figure 7: KNNk Assortativity of the RW

Finally, I computed the ***Betweeness centrality***. The *betweeness centrality* takes into account the number of shortest paths that go through a node. The assumption is that important vertices are bridges over which information flows, which means that if information spreads via shortest paths, important nodes are found on many shortest paths. Figure 6 reports the found results. The top 5 nodes were: *JoeBiden* (0.2638), *realDonaldTrump* (0.1664), *POTUS* (0.0839), *blesamerica* (0.0333) and *thehill* (0.02667). I tried to compute the betweeness centrality even on the synthetic graphs in order to observe if the synthetic graphs and the RW had the same centrality distribution. Just the BA and the CM presented a very similar distribution with the RW. This result seems to suggest the existence of a link between the betweeness centrality distribution and the degree distribution. In fact, these three models also share a very similar degree distribution, as mentioned in section §3.1. It is likely that the betweeness centrality is correlated with the degree distribution that follows a power law (such as the degree distribution of the RW, BA and CM) because of the presence of hubs, i.e. those nodes with large degree, through which a lot of information is very likely to pass.

In order to conclude this section about centrality, it's interesting to note that all methods returned very similar outputs which show that the most influential nodes belong to politicians, political institutions or media. In fact, among the at most top 10 nodes reported by each method, there were: *JoeBiden*, *realDonaldTrump*, which are the official accounts of Joe Biden and Donal Trump, *POTUS*, which is the official account of the President of the United States, *thehill* and *CNN* which are news accounts.

3.2.3 ***Assortativity***. The last measure that I calculated on the RW was the ***Newman's Assortativity***. Assortativity can be consider as a quantitative measure of homophily, which is the property whereby nodes with similar attitude tend to connect with each other rather than connect with nodes with different attitude. Assortativity takes values in the [-1, 1] interval: if it is included between 0 and 1 we have an assortative mixing, which means that nodes tends to connect homogeneously w.r.t. their degree (e.g., hubs with hubs), while, if it is included between 0 and -1 we have a disassortative mixing which means that nodes tends to connect in a star-like topology. If value is 0, then there's no correlation. The *Newman's assortativity* value for the RW was -0.0685, which means that the network presents a slightly disassortative behavior. The assortativity can also be observed with the **KNNk correlation plot**. The correlation is calculated as follows: all nodes of degree k are taken; then, for each of these nodes, the average degree of the nodes immediate neighbors is calculated; finally, the average of these averages is calculated. This will define a point on the graph with coordinates (k; knn). If the set of points shows an increasing trend, there will be assortative mixing; if decreasing, disassortative mixing; if (roughly) parallel to the x-axis, neutral. Observing Fig. 7, thus, it can be seen that as the k of nodes (x-axis) increases, the average of averages degrees of their direct neighbors decreases. This means that hubs are not connected with hubs, highlighting a disassortative behavior.

4 COMMUNITY DISCOVERY

An important analysis that can be performed on a network is **community discovery**, which allows to find out whether groups exist in the network whose nodes share similar features. First of all, I implemented and evaluated several algorithms on the entire RW; later, for the sake of the following analysis, I discovered static communities in each semester.

Speaking about the analysis on the entire RW, I decided to use five different algorithms: ***Louvain***, ***Label Propagation***, ***Angel***, ***Demon*** and ***K-clique***. *Louvain* adopts a bottom-up strategy to create communities: firstly, it assigns to each node a community, and then, iteration by iteration, each node is moved into the adjacent community that yields the greatest modularity increase. *Label Propagation* (LP), instead, works as follows: each node has an unique label which changes, with probability α, to one of the labels of its neighbors. Then, iteration by iteration, labels of nodes change according to the majority labels of their neighbors. The iteration process lasts until consensus is reached. *Angel* and *Demon* work like label propagation, but in a local strategy: they are based on the concept of *ego*, which is the focal node, around whom a network (called, *ego-network*) is built. For each node n, they extract the *ego-network* of n, remove n from the ego-network, perform a *Label Propagation*, insert n in each community found and update the raw community set C. Then, each community c in set C is merged with similar ones. Last but not least, *K-clique* algorithm identifies k-cliques, which are fully connected networks with k nodes, and then it creates communities, defined as set of adjacent k-cliques, that is, k-cliques that share

Louvain	weight: weight resolution: 0.7
Label Propagation	/
Angel	threshold: 0.5 min_community_size: 3
Demon	epsilon: 0.4, min_com_size: 3
K-cliques	k: 3

Table 4: Parameters for CD algorithms

	Louvain	LP	Angel	Demon	K-cliques
Community	4603	4674	18	16	35
Node coverage	1.0	1.0	0.01	0.01	0.01
Average internal degree	1.30	1.28	3.89	3.89	3.89
Internal edge density	0.71	0.72	0.83	0.83	0.82
Conductance	0.33	0.32	0.99	0.99	0.94
Newman-Girvan modularity	0.56	0.57	0.00	0.00	0.00

Table 5: CD Internal Evaluation results

exactly k-1 nodes.
The parameters chosen for each method are reported in Table 4.

The algorithms were evaluated with internal measures, in order to assess the model itself, and the external ones, in order to compare the results with the ones obtained with different algorithms. The internal evaluation results are reported in Table 5. As it can be seen, the best algorithms were *Louvain* and *Label Propagation*, which covered the entire network and obtained high enough modularity values. Even the conductance values for these two methods were good enough.

Next, I evaluated the performance of the algorithms by comparing them with each other. To do so, I used the ***Normalized Mutual Information***, a measure that compares the partitions obtained by two different algorithms. The higher the NMI, the more similar the compared partitions are. However, to use this evaluation methods, both algorithms should cover the same node set. For this reason, I compared, with NMI, just the *Louvain* and the *Label Propagation* algorithms. The matching score obtained was 0.945, a very high score, which confirms that the two algorithms found out almost the same communities. In order to compare the other methods, I used the ***Normalized F-1 Score***. The results were really low, except for the comparison between *Angel* and *Demon*, which returned a NF-1 score of 0.56, and the comparison between *Angel* and *K-cliques*, which returned a NF-1 score of 0.25. Figure 8 reports the similarity matrix among each algorithms.

At this point of the analysis, as said before, I performed two CD algorithms, the *Louvain* and the *Label Propagation*, on the network of each semester. Even in this case, the NMI matching score

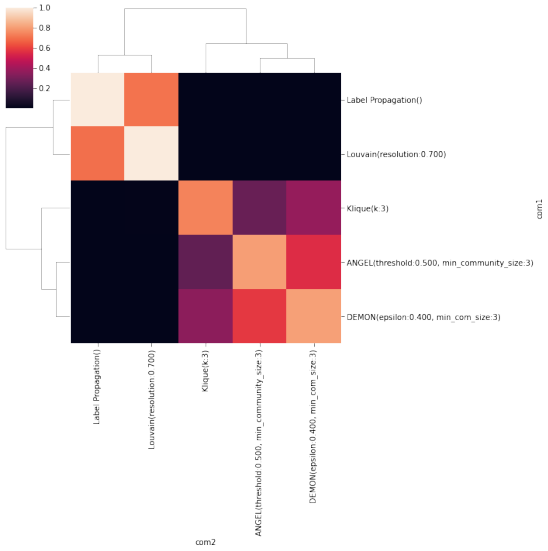

Figure 8: NF-1 similar matrix

obtained by comparing the communities found by these two algorithms in each semester, were really high (between 0.90 and 0.97). To subsequently perform the topic modeling on the communities obtained in each semester I considered only the communities found by the *Louvain* method.

5 DYNAMIC COMMUNITY DISCOVERY

After the discovery of static communities on the whole network, I decided to observe how the communities changed in time (i.e. in each semester). To do so, I considered the five networks created for each semester, as described in section §2.1. In order to reduce the computational time, I ran the DCD algorithms only on the giant component of each network. I chose to follow two different approaches: an *Instant Optimal* approach, using the ***Two-Step*** algorithm, and a *Temporal Trade-Off* approach, using the ***Tiles*** method. ***Two-Step*** works as follows: communities are detected at each time step using a static algorithm (in this case, I chose to use the *Louvain* and the *Label Propagation*); then similarities are computed between communities in consecutive steps, using a similarity measure (I chose the *Jaccard* similarity). At the end, the most similar communities are matched between t and $t+1$. So, after performing the algorithm, I observed the clustering stability trend (obtained using the ***NF-1 score*** as evaluation metric) between each pair of semesters. The results are reported in Table 6. As it can be seen, the diachronic stability was really low. One possible reason for these poor results could be the time elapsed between snapshots: it is clear that communities changed a lot between each time step, but the snapshots didn't capture these changes, considering the communities in t+1 as brand new. Perhaps, by decreasing the time between snapshots (e.g. by considering quarters, instead of semesters), the algorithms would be able to capture the changes and improve the results.

	1st Sem - 2nd Sem	2nd Sem - 3rd Sem	3rd Sem - 4th Sem	4th Sem - 5th Sem
Louvain	0.0096	0.0112	0.0142	0.0248
LP	0.0372	0.0294	0.0512	0.0492

Table 6: Two-Step Clustering Stability Trend

1st Sem - 2nd Sem	2nd Sem - 3rd Sem	3rd Sem - 4th Sem	4th Sem - 5th Sem
0.50	0.61	0.62	0.66

Table 7: Tiles Clustering Stability Trend

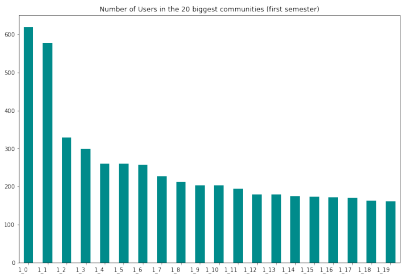

Figure 9: Numbers of users in each communities in the first semester (Two-Step algorithm)

Next, in order to improve the previous results, I implemented ***Tiles*** algorithm on the networks. *Tiles* is an algorithm whose communities found at each time step are a trade-off between the graph at t and its previous states. Using *Tiles*, the clustering stability trend increased a lot (Table 7), but the number of communities found in each semester decreased. However, this means that, even if few communities were found, these were really stable in time.

In order to conclude this section on dynamic community discovery, looking at the communities found by each method, it can be said that they follow a Zipfian law, as can be seen in Fig. 9 and 10, meaning that there are a few large communities, with many users, and many small communities, with few users. Moreover, I observed, using the *polytree* (Fig. 11), the life-cycle of two communities (obtained with the *Two-Step* method). The nodes in the graphs represent communities, identified by the *t_id* and the *community_id*. In the network on the left, a **merge** event can be seen, in which communities *1_8*, *1_9*, *1_32* converged into the *2_5* community. In the second semester, there was the **birth** of one community, the *2_14*, which, later, converged into the *3_16* community, along with the community *2_5*. Then, the community in the third semester, *3_16*, converged into the *4_31* community, and, at the end, the community of the fourth semester converged into the *5_37* community. The network on the right shows the same behavior, with only a merge event of the *1_0* and *1_47* communities, which merged into the *2_0* community.

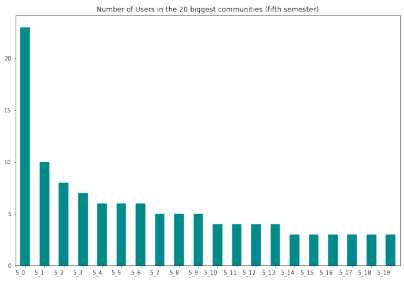

Figure 10: Numbers of users in each communities in the fifth semester (Tiles algorithm)

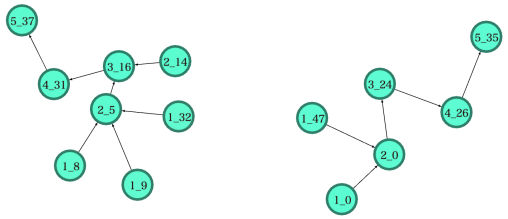

Figure 11: Polytree of two communities found by *Two-Step* algorithm

6 OPINION DYNAMICS

In this section, I tried to model the evolution of opinions about Joe Biden in the network, using some **opinion dynamics** models. The algorithms that I ran were ***Voter Model***, ***Majority Rule*** and ***Sznajd***, that are algorithms in which opinions are discrete (opinions can be 1 or -1). The results were compared with those obtained on some synthetic networks (BA and WS), in order to understand which synthetic networks better model the behavior of the real one.

The ***Voter Model*** is a method originally introduced to analyze competition of species and then applied to electoral competitions. At each iteration, a random agent *i* is selected with one of its neighbors *j*. If *i* and *j* share the same opinion, nothing happens; instead, if *i* and *j* have different opinions, then *i* changes its opinion according to *j*'s one. The algorithm was run for 1000 iterations both on the real network and on the synthetic networks. Fig. 12, 13 and 14 report the results. The *Voter Model* in RW presents, from the beginning, a clear distinction between the two opinions: iteration by iteration, the number of nodes with positive opinions increases (till to reach the percentage of 50.3% of positive opinions), and, consequently, the number of nodes with negative opinions decreased. BA shows a fuzzy behavior in the first iterations, becoming more stable only after more or less 400 iterations. Instead, WS exhibits an opposite behavior: it is negative opinion that increases, while positive opinion decreases.

The ***Majority Rule*** model was originally introduced to describe public debates (e.g., global warming, H1N1 pandemic). At each iteration, a random group of *r* agents is selected. Then the agents take the majority opinion within the group. If the selected *r* is odd, majority always exists, while if selected *r* is even, there is the possibility of tied configurations. To select a prevailing opinion in this

Figure 12: Voter Model (RW)

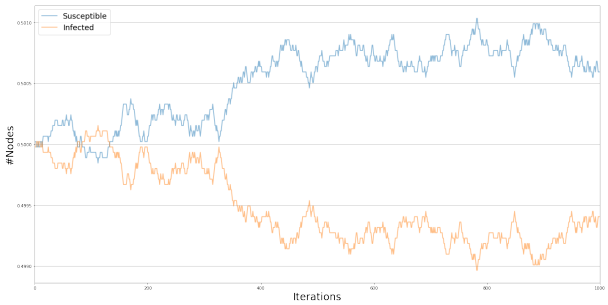

Figure 13: Voter Model (BA)

case, a bias in favor of one opinion is introduced. In the RW (Fig. 15), there is an increase in negative opinions, till to reach the percentage of 50.4%, while the positive opinion have decreased. Also with this algorithm, the BA (Fig. 16) shows a fuzzy trend in the first iterations, becoming steadier after 400 iterations. Again, negative opinions increased and, as a result, positive opinions decreased. In the WS (Fig. 17), the *Majority Rule* model presents an opposite trend, in which positive opinions increase and negative opinions decrease.

Lastly, I used the ***Sznajd*** model. The idea behind this algorithm is that a group of individuals having the same opinion can influence their neighbors more than a single individual. The model works as follows: at each iteration, a random agent i is selected with one of its neighbors j. If i and j opinions coincide, all their neighbors take that opinion, otherwise the neighbors take contrasting opinions. The model converges to one of the two contrasting stationary states. In each networks, as reported in Fig. 18, 19 and 20, the model show a clear polarization between the two opinions, with an increase of negative opinion and, consequently, a decrease of positive ones.

7 LINK PREDICTION

Another analysis that can be implement on a network is the *link prediction*, which allows to understand how a graph evolves. In other words, the link prediction is used to predict new possible links in the network. There are two different approaches: unsupervised link prediction and supervised link prediction. In this project, I only focused on the **unsupervised link prediction**. In order to reduce the computational time, I decided to sample data, extracting the sample from the giant component. In this way, I obtained an

Figure 14: Voter Model (WS)

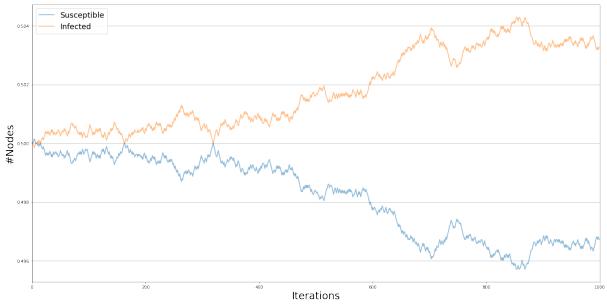

Figure 15: Majority Rule Model (RW)

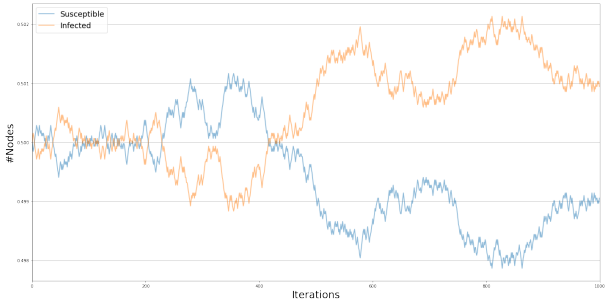

Figure 16: Majority Rule Model (BA)

Figure 17: Majority Rule Model (WS)

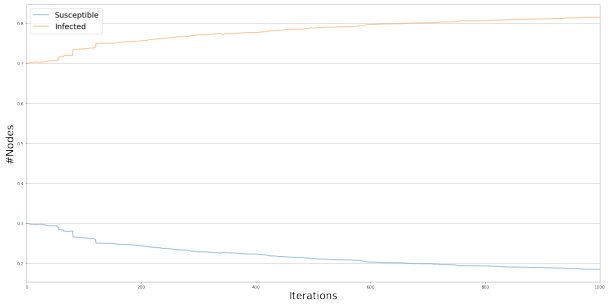

Figure 18: Sznajd Model (RW)

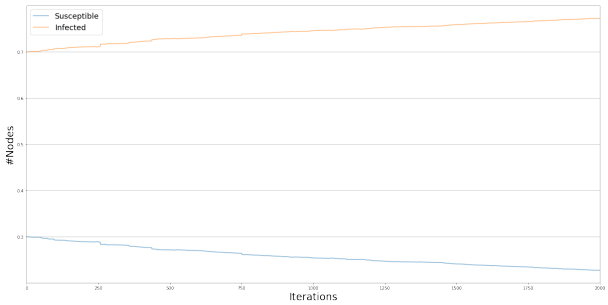

Figure 19: Sznajd Model (BA)

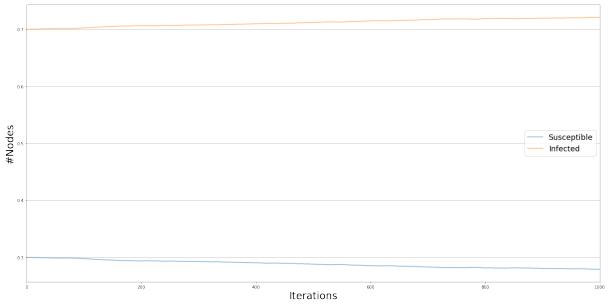

Figure 20: Sznajd Model (WS)

undirected multigraph with 963 nodes, 4810 edges and average degree of 9.98. Link predictions algorithms are divided into three families based on how they predict a new link between two nodes:

- **Neighborhood measures**, based on the idea that the more neighbors two nodes share, the more likely they are to become neighbors;
- **Path-based measures**, based on the idea that the closer two nodes are, the more likely it is that a link exists between them;
- **Ranking measures**, based on the idea that the more similar two nodes are, the more likely it is that a link exists between them.

For the first family, I used three algorithms: ***Common Neighbors***, ***Jaccard***, and ***Adamic Adar***.

The ***Common neighbors*** model defines the score by which a

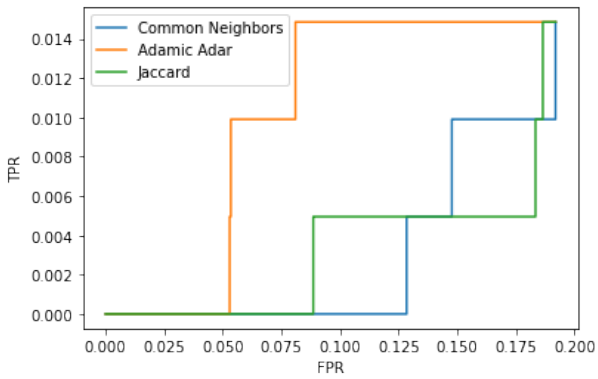

Figure 21: ROC curves (Common Neighbors, Jaccard, Adamic Adar)

link exists between two nodes u and v as described in 1, that is, taking into account the intersection between the neighborhoods $\Gamma(u)$ and $\Gamma(v)$ of the two nodes. The larger the intersection is, the more likely it is that a link exists between the two nodes.

$$score(u, v) = |\Gamma(u) \cap \Gamma(v)| \tag{1}$$

Instead, ***Jaccard*** is based on the idea that the more the neighborhoods of the two nodes overlap, the more likely it is that there is a connection between the two nodes. *Jaccard* is defined as follows:

$$score(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \tag{2}$$

Finally, the ***Adamic Adar*** model is based on the idea that the more selective our mutual friends are, the more likely we will become friends. It computes the score as follows:

$$score(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)|} \frac{1}{\log |\Gamma(z)|} \tag{3}$$

If z (which is a common neighbor of the nodes u and v) is a node with a high degree (like a hub), then it is less likely that a link exists between u and v since it is possible that z does not accurately select its neighbors. These three models were evaluated by splitting the network into a training graph (with 963 nodes and 4608 links) and a test graph (with 200 nodes and 202 links). Fig. 21 reports the ROC curves of these three models. As it can be seen, the results returned by these three algorithms were really poor.

Later, I used a path-based method, ***Katz***, and a ranking method, ***SimRank***. ***Katz*** computes the weighted sum over all the paths between two nodes. The more paths exist between two nodes, the more likely these two nodes are to connect directly. Instead, ***SimRank*** is based on the idea that two nodes are similar to the extent that their neighborhoods are similar. *Katz* and *SimRank* were also evaluated as before. As shown in Fig. 22, *Katz* performed slightly better than a random predictor, while the results obtained with *SimRank*, even if were significantly better than those of *Common Neighbors*, *Jaccard* and *Adamic Adar*, were worse than those of a random predictor.

In conclusion, even if the results were not good enough, it is interesting to note that *Common Neighbors*, *Adamic Adar* and *Katz*

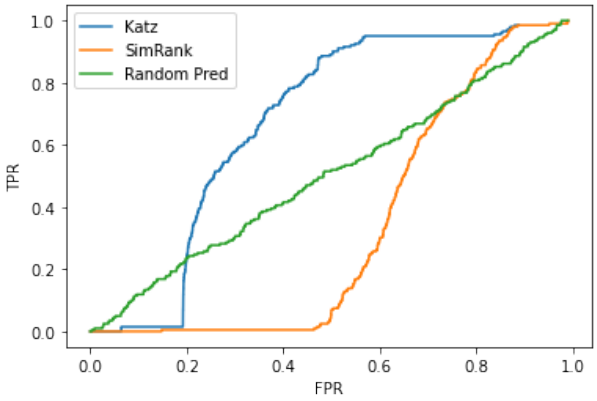

Figure 22: ROC curves (Katz, SimRank, Random Predictor)

predicted almost the same links. In fact, among the most likely new links, there are links between:

- *realDonalTrump - JoeBiden*;
- *JoeBiden - BernieSanders*;
- *KamalaHarris - JoeBiden*;
- *JoeBiden - BarackObama*;
- *JoeBiden - HillaryClinton*.

Even *Jaccard* and *SimRank* returned similar results to each other. Anyway, a supervised link prediction approach could be adopted in order to improve the results.

8 OPEN QUESTION

The goal of this work is to predict which will be the outcome of U.S. midterm election. To do so, I enriched the previous network analysis with some Natural Language Processing (NLP) techniques. First of all, for each semester, I performed a **topic modeling** algorithm in order to understand which were the topics, around Joe Biden, covered by the American public debate. Then, I did a **sentiment analysis** in order to understand how much Americans like Joe Biden. Moreover, I wanted to see whether there was a connection between the topics and the sentiments.

8.1 Data Preprocessing

For each semester, I created a new dataset that would contain not only all the information about the users and tweets, but also the information of the community to which they belonged, according to the communities found with *Louvain* algorithm, as reported in section §4. The following analysis were performed on the entire set of tweets presented in each semester's dataset, but also on a subset of each that I created by merging the top 5 communities by number of tweets from each semester. I chose to use the communities with the highest number of tweets instead of the communities with the highest number of users because I noticed that the communities with the most users are not always the most active on Twitter, that is, they are not always the ones that produce the most tweets. This could mean that the users in the first communities by number of users are not very active and focused on the issues in the public debate, while the users in the communities with more tweets turn

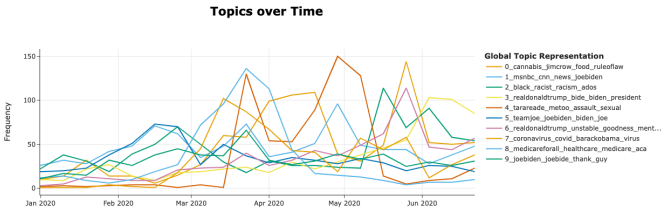

Figure 23: Topics extracted from First Semester

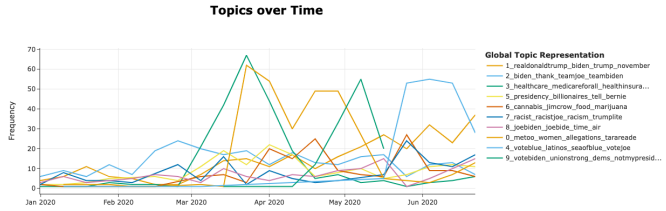

Figure 24: Topics extracted from First Semester's top 5 communities

out to be more interested in those issues. In communities with more users, moreover, having a limited number of tweets, it is more difficult to observe a significant trend of topics. In communities with more tweets, on the other hand, it is easier, as well as more interesting, to observe the trend of topics and compare it with the general trend of the semester. In this way, it is possible to understand which topics about Joe Biden, within those covered in each semester, have been most striking and touching in the most active communities. Sentiment analysis was then also conducted on these communities since I believe that the most active users are also those with a more established opinion about a certain topic and, consequently, about the President's performance.

8.2 Topic Modeling

The topic modeling was performed using a BERT model created for this kind of analysis called BERTopic[3]. First of all, tweets within the datasets were preprocessed using a function that:

- removes every links;
- removes audios and videos;
- lowers case;
- removes punctuation;
- removes double spacing;
- removes special characters;
- applies tokenization and lemmatization.

I chose to not remove mentions and hashtags because they could be useful for the topic detection. For each semester, the top 10 topics were extracted.

8.2.1 ***First Semester***. Firstly, looking at Fig. 23, it can be seen that, in the first months of 2020, Americans started talking about Joe Biden and the elections (topics 3, 5, 6 and 9). Even the topic 4

[3]BERTopic repository: https://github.com/MaartenGr/BERTopic

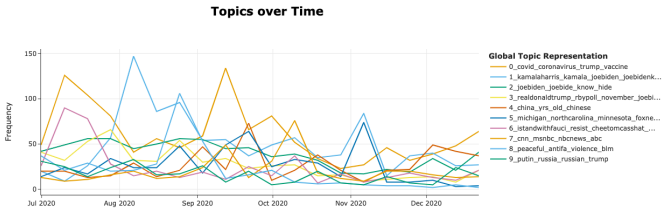

Figure 25: Topics extracted from Second Semester

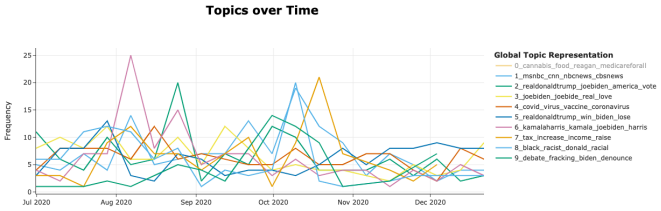

Figure 26: Topics extracted from Second Semester's top 5 communities

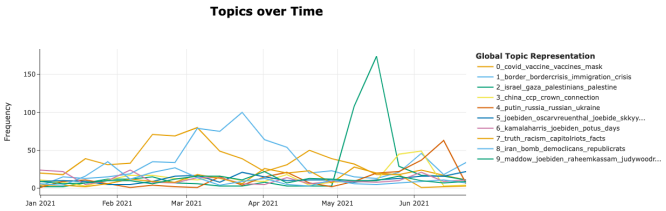

Figure 27: Topics extracted from Third Semester

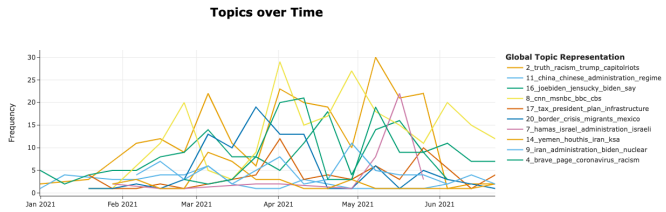

Figure 28: Topics extracted from Third Semester's top 5 communities

(*tarareade_metoo_assault_sexual*) is connected to the Joe Biden's election campaign. In fact, Tara Reade is a woman who accused Joe Biden of assaulting her in 1993. This fact has been of particular interest to Twitter users especially around May 2020. Another topic covered especially in the month of June 2020 was the legalization and decriminalization of cannabis in the U.S.. In addition, a topic which covered the public debate for all the semester was topic 8 (*medicareforall_healthcare_medicare*) related to the emergence of Covid-19 pandemic. Instead, considering only the top 5 communities by number of tweets (Fig. 24), it can be said that, generally, users in these communities talked more about the election campaigns rather than users in the global dataset (as described by topics 1, 2, 4, 5, 8, 9). Moreover, even in these 5 communities, users were really interested in Tara Reade's accuse (perhaps precisely because they are more focused on election campaigns). Finally, Covid-19 related topics were covered. In conclusion, it could be argued that the top 5 communities by number of users are populated by users who are very focused on political issues, more than the global users. In any case, there is a partial overlap between topics covered by the global users and those covered by users in the top 5 communities.

8.2.2 ***Second Semester.*** Since the second semester is marked by the election of Joe Biden as president, some of the arguments are related to this fact (Fig. 25). One of the most tweeted topics, especially in August 2020, was related to Kamala Harris (topic 1), the current U.S. vice-president, who, at the time, was running for presidency with Joe Biden. Then, coronavirus remained among the main topics. During this time, Americans tweeted about the economic problems related to the pandemic and started talking about vaccines. The coronavirus, therefore, was one of those topics that most interested the public debate, and which, arguably, influenced, if not directed, the outcome of the election. Looking at Fig. 26[4], it can be said that, even in this semester, the communities were focused on the political debate. It's interesting to note that in the top 5 communities there were both Trump's supporters and Biden's supporters, as shown, respectively, in topic 5 (*realdonaldtrump_win_joebiden_lose)* and 3 (*joebiden_joebiden_real_love*). Finally, other covered topics were racism and the tax plan release by Joe Biden, before the election, that would have raised taxes on individuals with income above $400,000.

8.2.3 ***Third Semester.*** Analyzing Fig. 27, three main topics stand out: vaccines (topic 0), Mexico's border crisis (topic 1), Israel–Palestine crisis (topic 2). Vaccines were one of those topics which characterized the public debate around the world, especially in the early months of 2021. The public debate was quite polarized between those who were pro-vaccine and those who were not. The topic about Mexico's border crisis refers to the arrest of 210.000 migrants attempting to cross the border with Mexico in March. In the first months of his term, as it is reported by Reuters [5], in fact, Biden tried to reverse many of the restrictive immigration policies imposed by Trump, encouraging, according to Republicans, illegal immigration. Israel–Palestine crisis, instead, refers to a major outbreak of violence commenced on 10 May 2021 and continued until a ceasefire came into effect on 21 May. The crisis was triggered on 6 May, when Palestinians in East Jerusalem began protesting over an anticipated decision of the Supreme Court of Israel on the eviction of six Palestinian families in the neighborhood of Sheikh Jarrah [6]. Biden's first months in office, therefore, were marked by

[4] I deleted Topic 0, because it biased the analysis, flattening all other results and affecting the readability of the plot.
[5] https://www.reuters.com/world/us/us-arrests-210000-migrants-mexico-border-march-rivaling-record-highs-2022-04-16/, consulted on July 12, 2022
[6] https://www.nytimes.com/2021/05/22/world/middleeast/israel-gaza-conflict.html, consulted on July 12, 2022

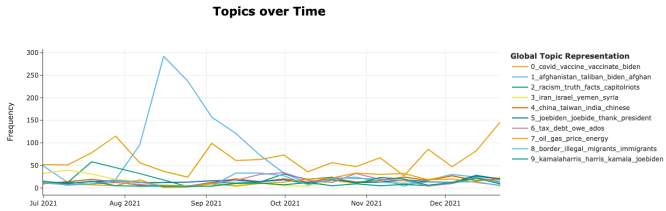

Figure 29: Topics extracted from Fourth Semester

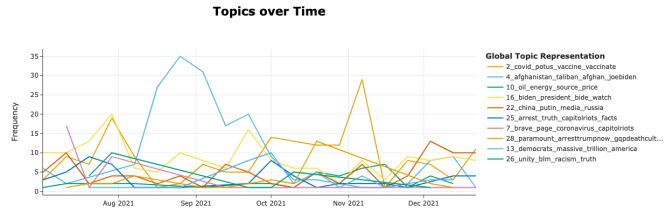

Figure 30: Topics extracted from Fourth Semester's top 5 communities

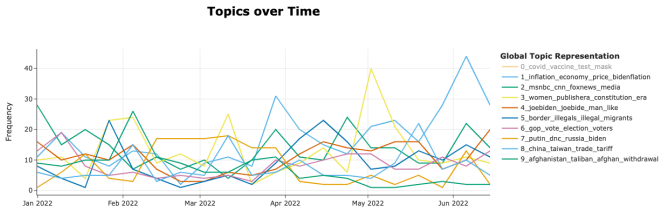

Figure 31: Topics extracted from Fifth Semester

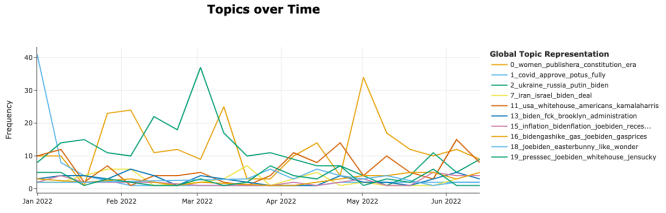

Figure 32: Topics extracted from Fifth Semester's top 5 communities

several foreign policy problems. In fact, the other topics, in addition to those already mentioned, also show that Americans have been tweeting a lot about different foreign policy issues. For example, topic 3 focuses on China, topic 8 on the airstrike, ordered by Biden, against some buildings belonging, according to the Pentagon, to Iran-backed militias[7], while topic 4 shows the emergence of the Russia-Ukraine issue. Also among the top 5 topics of the communities, Biden's foreign policy is of great interest, although there are even some domestic political issues, especially concerning the facts surrounding the January 6, 2021 attack, perpetrated by Trump supporters, on the Capitol Building in Washington D.C.. Finally, it is interesting that media started referring to Joe Biden as *sleepyJoe*.

8.2.4 ***Fourth Semester***. Fourth semester was characterized especially by a main topic, as shown in Fig. 29 (topic 1): the withdrawal of U.S. troops from Afghanistan, taken place in August 2021. In February 2020, the Trump administration and the Taliban, without the participation of the then Afghan government, signed the U.S.–Taliban deal in Doha, Qatar which stipulated fighting restrictions for both the U.S. and the Taliban, and provided for the withdrawal of all NATO forces from Afghanistan in return for the Taliban's counter-terrorism commitments [8]. The Trump administration's U.S.–Taliban deal, and then the Biden administration's decision in April 2021 to pull out all U.S. troops by September 2021 without leaving a residual force, were the two critical events that caused the collapse of the Afghan National Security Forces (ANSF)[9]. These facts leaded to the Taliban takeover of Kabul on 15 August 2021[10]. Even in this semester (Fig. 30), the second main topic was related to coronavirus. The other topics seem to reflect the topics of the previous semester (racism, China and Russia) except for the emerging issue of oil, gas and energy price (topic 7). These were more or less the topics also covered by the top 5 communities, with increased interest in November 2021, regarding coronavirus (topic 2).

8.2.5 ***Fifth Semester***. Finally, fifth semester was characterized basically by three main topics: the Russia-Ukraine war, inflation and the abolition of the constitution right to abortion (Fig. 31[11]). Russia-Ukraine war started in February 2022, when Russia invaded Ukraine with the goal, in Putin's words, of demilitarizing and denazifing Ukraine. This war has destabilized politics around the world. In the same period, inflation, i.e. a general increase in costs, affected both Europe and America and ignited public debate (topic 1). The topic about the abolition of the constitution right to abortion (topic 3) refers to the decision of the U.S. Supreme Court to overturn the *Roe v. Wade*, i.e. a landmark decision of the U.S. Supreme Court in which the Court ruled that the Constitution of the United States conferred the right to choose to have an abortion. Even in the top 5 communities, these were the main topics, with a deeper focus on Russia-Ukraine war. It is interesting to note the inflation is also called *Bidenflation* in topic 15 in Fig. 32, which could mean that Americans are blaming Joe Biden for inflation.

8.3 Sentiment Analysis

The next step in the analysis was to run a **sentiment analysis** algorithm on the tweets. As mentioned earlier, I ran the algorithm

[7] https://www.nytimes.com/2021/02/26/us/politics/biden-syria-airstrike-iran.html, consulted on July 12, 2022
[8] https://www.aljazeera.com/news/2020/2/29/afghanistans-taliban-us-sign-agreement-aimed-at-ending-war, consulted on July 12, 2022
[9] https://www.theguardian.com/world/2022/may/18/afghanistan-us-withdrawal-defeat-watchdog-report-sigar, consulted on July 12, 2022
[10] https://www.aljazeera.com/news/2022/5/18/us-withdrawal-prompted-collapse-of-afghan-army-report, consulted on July 12, 2022
[11] Even in this case, I deleted Topic 0, because it biased the analysis, flattening all other results and affecting the readability of the plot.

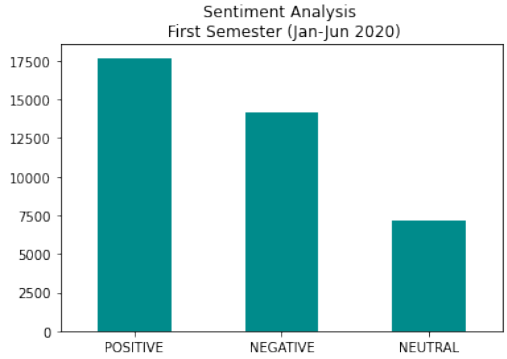

Figure 33: Sentiment Analysis First Semester

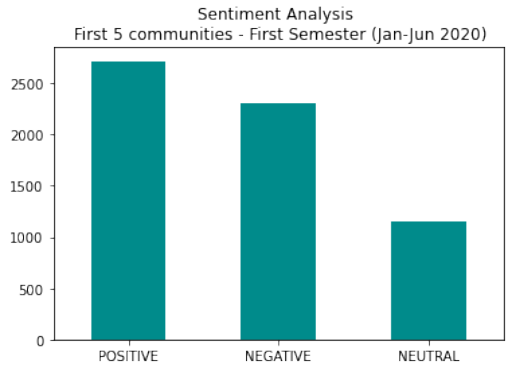

Figure 34: Sentiment Analysis First Semester's top 5 communities

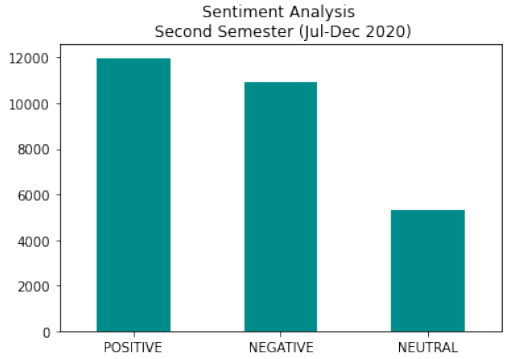

Figure 35: Sentiment Analysis Second Semester

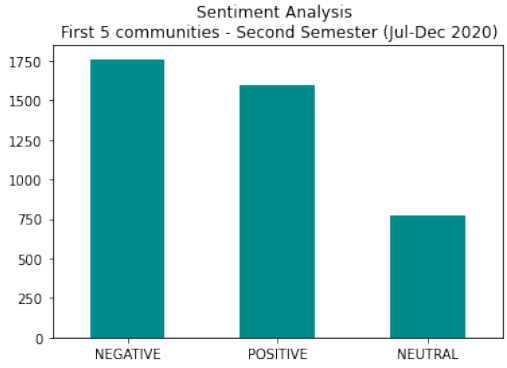

Figure 36: Sentiment Analysis Second Semester's top 5 communities

on both the entire set of tweets in each semester, and the tweets in top 5 communities by number of tweets in each semester. I chose to use as algorithm, ***VADER***[12], i.e. a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. *VADER* assigns to each tweet a score, the so called *compound* score, according to the global sentiment conveyed by the tweet. In order to perform the sentiment analysis, tweets were preprocessed as described in section §8.2.

8.3.1 ***First Semester.*** As reported in Fig. 33 and Fig. 34, in both the whole set of tweets of the semester and the top 5 communities set of tweets, the prevailing sentiment is positive. This result is not very surprising: in the first half of the 2020, in fact, Joe Biden still did not play an active and key role in the executive branch, but rather, he was beginning his run for the presidency. Therefore, it is possible that for many of the Americans, after 4 years of the Trump's administration, Biden represented a novelty (and a better option) and thus placed their hopes and trust in him.

8.3.2 ***Second Semester.*** The sentiment felt by Americans in the second half of the 2020 is really interesting: in fact, considering only the whole set of tweets of this semester (Fig. 35), it can be seen that the predominant sentiment is positive. In fact, in these months, Joe Biden has been nominated for president. The news was certainly welcomed by Biden's supporters, who tweeted positively. However, looking at Fig. 36, it can be observed that the prevailing sentiment in the top 5 communities is negative. A possible explanation for this difference could be that in the top 5 communities there were more Trump's supporters rather than in the global set. This can be confirmed by the presence of topic 5 (*realdonaldtrump_win_joebiden_lose*) in the top 5 communities' topics, as shown in Fig. 26. Another reason could be found in topic 7 which refers to the new Biden's tax plan.

8.3.3 ***Third Semester.*** In the third semester, the sentiment follows a similar trend as in the second semester: globally (Fig. 37), the sentiment conveyed by tweets is positive, although in the 5 communities (Fig. 38) negative sentiment prevails. A possible reasons for this phenomenon is that one of the most tweeted topics in the third semester's top 5 communities is about Trump and the January 6, 2021 Capitol riot. This means that the negative sentiment does not concerned Joe Biden, but rather, Donald Trump. However, it should be remembered that in these months, media started calling Joe Biden as *sleepyJoe*, which could be seen as an early symptom of discontent with Biden's administration.

8.3.4 ***Fourth Semester.*** The sentiment in fourth semester is clear: Americans started having a negative sentiment about Joe Biden's

[12]VADER repository: https://github.com/cjhutto/vaderSentiment

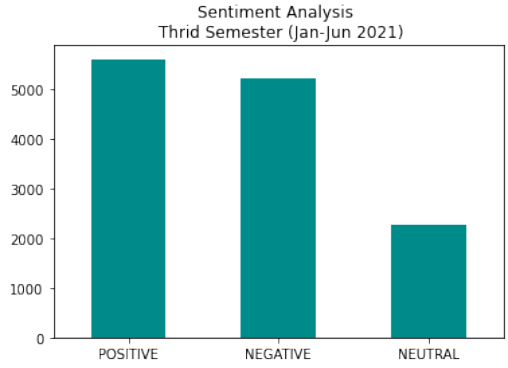

Figure 37: Sentiment Analysis Third Semester

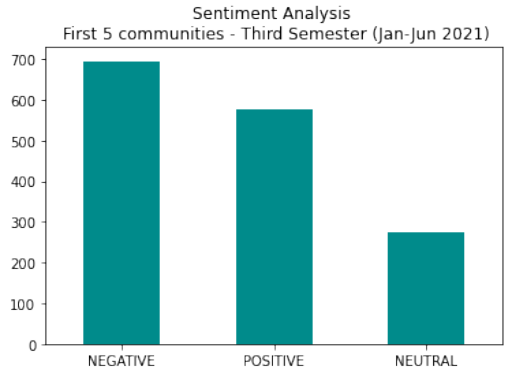

Figure 38: Sentiment Analysis Third Semester's top 5 communities

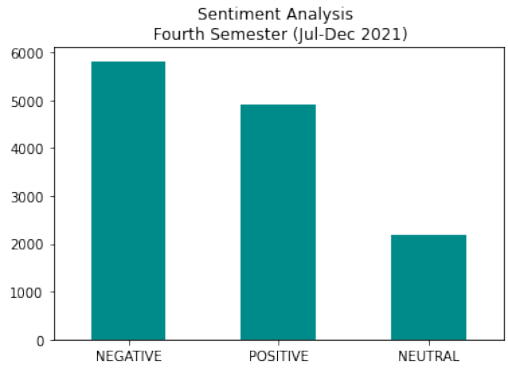

Figure 39: Sentiment Analysis Fourth Semester

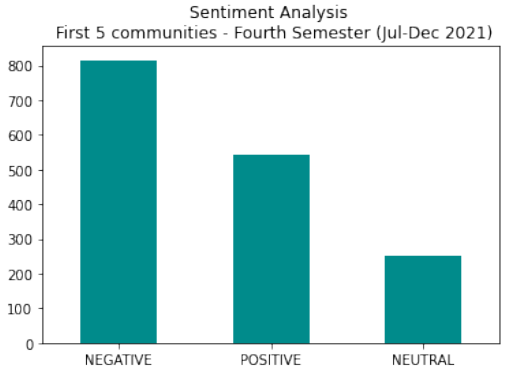

Figure 40: Sentiment Analysis Fourth Semester's top 5 communities

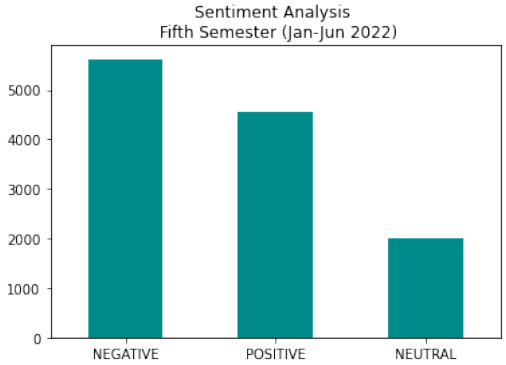

Figure 41: Sentiment Analysis Fifth Semester

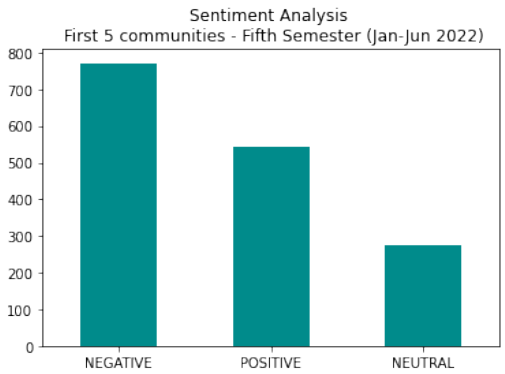

Figure 42: Sentiment Analysis Fifth Semester's top 5 communities

administration. It should be remembered that the withdrawal of U.S. troops from Afghanistan took place in this semester. Biden's decision to withdraw U.S. troops quickly from Afghanistan has divided the American (and world) public debate and American politics. Biden has received criticism from both the right and the left. Newspapers have called it a *catastrophe*, a *disaster*, and a *failure*. The criticism focused not so much on the appropriateness of the troop pullout but on its manner, for which Biden and his government were primarily responsible. This could explain the rapid change in American sentiment about Biden.

8.3.5 ***Fifth Semester.*** In the fifth semester, negative sentiment has become entrenched among Americans, as shown in both Fig. 41

and Fig. 42. Three could be the main reasons: the Russian-Ukrainian war and how the U.S. is dealing with it, the inflation that is affecting the world economy, and the Supreme Court's decision to overturn the landmark *Roe v. Wade* ruling that established the constitutional right to abortion. The trend in this last semester seems to confirmed and worsen the previous one. Biden's administration, therefore, seems to be struggling to maintain a broad consensus.

8.4 Final Discussion on Open Question

So, is it possible to predict the outcome of U.S. midterm election? Surely, after all the previous analysis, it can be said that the degree to which Americans appreciate their president has slowly decreased over the semesters, as summarized in Fig. 43. The time when Americans began to feel negative feelings toward their president was around the summer of 2021, when the events in Afghanistan occurred. However, it should be noted that already among the top 5 communities in the second and third semesters a negative sentiment was beginning to circulate (as mentioned in sections §8.3.2 and §8.3.3). The fact that the first to express negative sentiment were the most active communities, rather than the entire set of users in each semester, confirms what was stated in section §8.1: the most active communities are the most attentive and interested in public debate and, as a result, express discontent sooner. Thus, it can be said that since his nomination as president, Joe Biden has lost more and more support. A gradual loss of consensus toward midterm is physiological. In fact, as reported in [7], since World War II, in the midterm elections, the President's party loses an average of 26 seats in the House, and an average of 4 seats in the Senate. However, some U.S. newspapers, such as *Newsweek*[13], report that, according to polls, Joe Biden is bracing for the most difficult midterms in 48 years (i.e., since the Nixon administration's midterm election, after the Watergate scandal). Although with the analysis carried out, it is not possible to say whether Biden's midterm elections are more or less difficult than those faced by his predecessors, it can still be said that the Biden's administration, in these first two years in office, has had to deal with more difficult issues and make more complex decisions than other previous executives, which, as a result, have increasingly polarized public opinion. To conclude, then, it is strongly conceivable that the Biden's administration will lose several seats in both houses of Congress and, with them, much of the consensus of Americans.

9 CONCLUSION

This work shows that the combination of network science and natural language processing techniques provides a better understanding of certain aspects of reality. The objective of this study was to try to predict the outcome of the U.S. midterm elections. After downloading data from Twitter, several preliminary analyses were conducted on the network. The techniques of community discovery and dynamic community discovery allowed to observe how users were grouped within the network and, then, thanks to topic modeling techniques, to understand what were the topics Americans felt most strongly about during these two years of Biden's administration. Next, link prediction and opinion dynamics

[13]https://www.newsweek.com/joe-biden-bracing-most-difficult-midterms-48-years-1718344, consulted on July 12, 2022

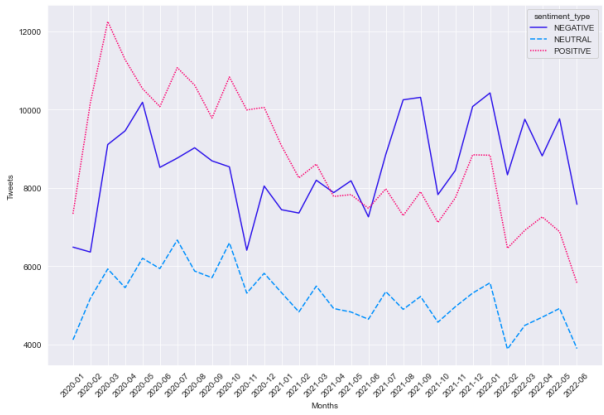

Figure 43: Sentiment Trend

were performed in order to understand, respectively, what new links could be created within the network and how opinions were spreading across the network. Finally, to understand how much Biden was liked by Americans and to define a possible midterm election outcome, sentiment analysis was performed. In conclusion, in order to improve some results, it could be interesting to rerun the dynamic community discovery algorithms reducing the elapsed time between each snapshot. Moreover, to improve the results of the link prediction, it might be performed a supervised link prediction. Finally, opinion dynamics could be run on the top 5 communities by number of tweets of each semester in order to see how opinions spread within these communities.

REFERENCES

[1] Albert-László Barabási. Network science book. *Network Science*, 625, 2014.
[2] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.
[3] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. 2009.
[4] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
[5] Gennaro Cordasco and Luisa Gargano. Community detection via semi-synchronous label propagation algorithms. In *2010 IEEE international workshop on: business applications of social network analysis (BASNA)*, pages 1–8. IEEE, 2010.
[6] Michele Coscia, Giulio Rossetti, Fosca Giannotti, and Dino Pedreschi. Demon: a local-first discovery method for overlapping communities. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 615–623, 2012.
[7] David A Crockett. *The opposition presidency: Leadership and the constraints of history*. Number 11. Texas A&M University Press, 2002.
[8] P Erdős and A Rényi. On random graphs i. publicationes mathematicae (debrecen). 1959.
[9] Paul Erdos, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.
[10] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.
[11] Serge Galam. Minority opinion spreading in random geometry. *The European Physical Journal B-Condensed Matter and Complex Systems*, 25(4):403–406, 2002.
[12] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.
[13] Richard A Holley and Thomas M Liggett. Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability*, pages 643–663, 1975.
[14] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

[15] Mathieu Jacomy, Sebastien Heymann, Tommaso Venturini, and Mathieu Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization. *Medialab center of research*, 560:4, 2011.

[16] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[17] Mark EJ Newman. Mixing patterns in networks. *Physical review E*, 67(2):026126, 2003.

[18] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[19] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, June 2005.

[20] Giulio Rossetti. Exorcising the demon: Angel, efficient node-centric community discovery. In *International Conference on Complex Networks and Their Applications*, pages 152–163. Springer, 2019.

[21] Giulio Rossetti, Letizia Milli, and Rémy Cazabet. Cdlib: a python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, 4(1):1–26, 2019.

[22] Giulio Rossetti, Letizia Milli, Salvatore Rinzivillo, Alina Sîrbu, Dino Pedreschi, and Fosca Giannotti. Ndlib: a python library to model and analyze diffusion processes over complex networks. *International Journal of Data Science and Analytics*, 5(1):61–79, 2018.

[23] Giulio Rossetti, Luca Pappalardo, Dino Pedreschi, and Fosca Giannotti. Tiles: an online algorithm for community discovery in dynamic social networks. *Machine Learning*, 106(8):1213–1241, 2017.

[24] Katarzyna Sznajd-Weron and Jozef Sznajd. Opinion evolution in closed community. *International Journal of Modern Physics C*, 11(06):1157–1165, 2000.

[25] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[26] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, 2015.