Vienna University of Technology

Department of Computer Science

Interdisciplinary Project Report @Complexity Science Hub Vienna - Newspaper Similarity

Author: Juraj Simkovic

Supervisor: Vito D. P. Servedio & Pietro Gravino

Interdisciplinary Lecture: 105.762 VU 2025S AKFVM Introduction to Financial Networks

GitHub of Project:
https://github.com/JurajkoDS/Intedisciplinary-Project–Newspapers-Similarity—SS25–CSH-Vienna

Abstract

This project will explore the relationship between newspapers and their similarity based on retweet patterns. The project is based on Twitter data that ranges from 2018 to 2022. To analyse the patterns of newspapers and their audience, a bipartite network will be created. Due to the high dimensionality of the data, dimensionality reduction techniques such as PCA will be implemented to reduce complexity. Finally, based on monthly snapshots of time data and the dynamics of change, statistical evaluations will be produced to draw insights about newspapers' change over time. Towards the end of the project, community clustering using the Louvain algorithm was calculated. Based on these results, the modularity of the graph was also calculated. The project has found that the statistical features calculated throughout the project are mostly statistically significant.

Keywords: newspaper similarity, cosine similarity, bipartite graph, principal component analysis, dimensionality reduction, community detection, Louvain clustering, modularity

Contents

Chapter 1

Introduction

1.1 Background

This project focuses on newspaper similarity in Italy and is based on Twitter data collected between the years of 2018 and 2022. The newspapers will be also referred to as "leaders" in this project. For example, throughout these 5 years of data, YouTube is the leader with by far the highest amount of retweets. As a further example, YouTube is then followed by Giorgia Meloni. So, examples of leaders are media outlets, politicians, etc. The leaders are retweeted by other Twitter users, which are being referred to as "followers".
The motivation behind this project is to find out how are individual leaders similar to each other on the basis of retweets of their followers. We assume that if a follower retweets one leader and also retweets another that this follower shares the opinion of both of the retweeted leaders.

1.2 Aims and objectives

Based on the similarity, statistical measures such as mean, standard deviation (STD) and velocities will be created to draw insights on individual time frames, such as months, consecutive months (i.e. January-February, February-March, etc.) or specific years, between the years 2018 and 2022. Velocities refer to the consecutive monthly changes of individual graph nodes, which represent the leaders. Futhermore, based on the velocities, more statistical measures will be calculated. Eventually, modularity algorithms will be also implemented to analyse and detect communities within the graph. Based on these statistical measures, conclusions, or at least partial conclusions, about the polarity and dynamics of the Italian society can be drawn.

1.3 Dataset

The dataset consists of 3 sub datasets, namely the tweets, the retweets and information about users (or the Twitter accounts). Each of these 3 dataframes were ultimately merged on, first author_ID and ID between the tweets and users dataframes. After that this newly merged dataframe was merged with the retweets dataframe on original post_ID (original, because it was originally posted by a leader) and post_ID, because that it was retweeted by another user (a follower).
While the dataset offered a lot of features such as like count, reply count, retweet count, quote count, description (i.e. the text of the post), which offer more room for analysis, our project focused only on a subset of the features. The final features that were selected to be worked

with are tweet_created_at, original_post_id, author_id, author_name, author_username and retweeter_id.
The first five rows of the final cleaned dataframe can be seen in Figure 1.1.

	tweet_created_at	original_post_id	author_id	author_name	author_username	retweeter_id
0	2018-12-31 22:46:18	1079886497279561728	622354597	Salvo Di Grazia	MedBunker	951848540
1	2018-12-31 22:46:18	1079886497279561728	622354597	Salvo Di Grazia	MedBunker	135554444
2	2018-12-31 22:46:18	1079886497279561728	622354597	Salvo Di Grazia	MedBunker	433418060
3	2018-12-31 22:46:18	1079886497279561728	622354597	Salvo Di Grazia	MedBunker	1668533642
4	2018-12-31 22:46:18	1079886497279561728	622354597	Salvo Di Grazia	MedBunker	1623208790

Figure 1.1: Final Dataframe

The final dataframe spans roughly 29 million data points and has 6 features.

Chapter 2

Literature Review

The following section will explore scientific articles relevant to this project in the Italian context and explain why this project could be a new scientific contribution.

Federico et al. (2024) has shown the community structure in Italy is similar to the actual affiliated political parties in 2022. Caldarelli et al. (2014) has investigated how do volumes of Tweets from individual political leaders correlate to the actual election results and found that it can be a good indication of the election outcome. The data used in this analysis is, however, from 2013 and the results may not be applicable today, as the platform's user base, communication strategies and algorithm dynamics may have changed since then, making the the findings potentially outdated. Using an entropy model Becatti et al. (2019) looks at inferring the polarisation of a user based on his/her relational activity on Twitter, instead of looking at the post content itself. Afterwards Becatti et al. (2019) also looks at the structure of the network.
As Zollo (2019) has shown, strong polarisation offers fertile ground for echo chambers fueled by confirmation bias. Because of that, Pierri et al. (2020) also looks at the disinformation spreading on online social networks preceding the 2019 European Parliament elections. Tognini et al. (2020) monitored on a daily basis how has the structural evolution of the semantic networks induced by the communities they identified change. The communities were developed based on a similar retweeting behavior. While the method of community detection based on retweeting patters is similar to what this project offers, ultimately the authors investigate semantic networks.
Inevitable, as the data which this project works with gather posts and information throughout the Covid-19 pandemic, the analysis of semantic network in that period and the methods used for that are also relevant. In this context, Mattei et al. (2021) have reviewed how communities are exposed to misinformation using a entropy based null model with a similar method as used in this project - relating two users based on retweets of the same content.

While all of these articles describe the societal structures in Italy based on the Twitter data, none of them describe the dynamics of change of the individual political leaders or media outlets. That is the scientific gap that this project is trying to address by calculating statistical measures such as mean, standard deviation and change of cosine distance between consecutive months (i.e. January-February, February-March, etc.) and years.

2.1 Summary

The articles referred to in this short literature review looked community structure in Italy, as well as the correlation of Tweet volumes and election results. They also looked at polarisation and the spread of misinformation on social media in Italy.

Chapter 3

Methodology

3.1 Calculating Similarity

The similarity between individual leaders is the foundation of this project. Based on this measure, further statistical features will be calculated.
To calculate the similarity, a biadjacency matrix consisting of followers and authors needs to be created. The values of this matrix are represented by the count of how many times each follower has retweeted a specific leader. For example, follower 1 has retweeted the leaders 5 and 8 in the month of March. Number 5 was retweeted 10 times and leader number 7 was retweeted another 23 times. We do this count for each of the followers for each of the leaders. For illustration purposes, the figure 3.1. depicts this counting.

Authors	0	1	2	3	4	5	6	7	8	9
Retweeters										
0	5	0	0	0	0	0	0	0	0	0
1	26	0	0	0	0	0	0	0	0	0
2	2	0	0	0	0	0	0	0	0	0
3	90	0	0	0	0	0	0	0	0	0
4	92	87	2	57	2	1	0	0	5	0
...	...	...	...	...	...	...	...	...	...	...
107388	0	0	0	0	0	0	0	0	0	0
107389	0	0	0	0	0	0	0	0	0	0
107390	0	0	0	0	0	0	0	0	0	0
107391	0	0	0	0	0	0	0	0	0	0
107392	0	0	0	0	0	0	0	0	0	0

Figure 3.1: Retweet Count

After that, we count the pairwise similarities of leaders, which can be seen in Figure 3.2.

3.1.1 Sanity Check

To make sure that the results of the similarity are not off and actually do make sense, a sanity check has been created. In order to do that, top 10 and bottom 10 similarities were calculated and displayed. The results were then checked with the supervisors how are knowledgeable on the topic of Italian politics and were able to verify, whether the similarities make sense. Further verification can be done by researching which leaders, so politicians or media sources, are leaning towards which part of the political spectrum, for example liberal vs. conservative.

	Il Fatto Quotidiano_January	Repubblica_January	RTL 102.5_January	Corriere della Sera_January	Matteo Salvini_January	Maurizio Blondet_January
Il Fatto Quotidiano	1.000000	0.236390	0.004611	0.256925	0.055982	0.064071
Repubblica	0.236390	1.000000	0.025082	0.598046	0.031441	0.017273
RTL 102.5	0.004611	0.025082	1.000000	0.038037	0.023648	0.008027
Corriere della Sera	0.256925	0.598046	0.038037	1.000000	0.051669	0.027889
Matteo Salvini	0.055982	0.031441	0.023648	0.051669	1.000000	0.128856
...	...	...	...	...	...	...
Il Giorno	0.024398	0.058705	0.002201	0.113085	0.026930	0.012760
Agenzia Agenpress.it	0.022378	0.023254	0.002050	0.023230	0.002821	0.001114
Luigi de Magistris	0.039963	0.038401	0.001640	0.044199	0.003011	0.003566
Jeda News	0.034494	0.013104	0.001688	0.028425	0.057832	0.078290
Quotidiano di Puglia	0.000000	0.000336	0.309461	0.010933	0.000000	0.000000

151 rows × 151 columns

Figure 3.2: Leaders Similarity - January 2018

Figure 3.3. depicts this check.

Author 1	Author 2	Similarity
L'Unione Sarda	Leggo	0.8713639806260215
Corriere del Veneto	VeronaSera	0.8052286598892232
Il Centro	L'Unione Sarda	0.7890297413806676
Corriere del Veneto	L'Unione Sarda	0.7462845585234967
L'Unione Sarda	VeronaSera	0.74312603891876
Il Centro	VeronaSera	0.732309674082556
Il Centro	Leggo	0.7191032890974549
Corriere del Veneto	Il Centro	0.7059516143286899
VeronaSera	Leggo	0.6807349670571833
Corriere del Veneto	Leggo	0.678945406362099

Figure 3.3: Sanity Check - Top 10 Leader Similarities in March 2018

3.1.2 Measure Selection

Before ultimately deciding for a specific measure selection, several measures were tried. Firstly, Euclidian distances were tried as a similarity measure. However, after a bit of researching, it was clear that Euclidian distances are not suitable for highly dimensional data (curse of dimensionality).
After that Jaccard similarity was also computed. This method was computationally taxing and took a lot longer to calculate than anticipated. Last but not least, the cosine similarity was calculated and this is also the measure that was kept.
Further justification of this measure and possibly other suitable measure could be looked at.

3.2 Calculating Velocity

The velocity of two nodes (or leaders), based on the similarity measure, has been calculated as well. It was calculated between two consecutive months. This means that for each year, we get 11 velocity scores. For example, we have the velocity of January-February, February-March, ..., November-December of 2018, ..., 2022. There is no velocity between December-January, which could be improved in future work. Based on the velocity, further statistical measures such as mean and standard deviation were calculated.
There are two cases of calculating the velocity between two months. Firstly, if a leader has a similarity score in the in one month and a month that follows (so for example January and February), then there is a velocity score. This means that the leader has tweeted and been retweeted as well in both of January and February. Secondly, if the leader has not tweeted and/or has been not retweeted in either of the two months, then there is no score, depicted by a NaN, because there is no velocity between a NaN and an X value.
Figure 3.4. depicts the results of this procedure. Figure 3.5. depicts the mean and standard deviation of the velocities.

	24Emilia	@qtsicilia	@tp24	Adnkronos	Affaritaliani.it	Agenzia ANSA	Agenzia Agenpress.it
Il Fatto Quotidiano_January	0.0	0.0	0.0	0.196985	0.038067	0.221752	0.022378
Repubblica_January	0.0	0.0	0.0	0.279038	0.005115	0.415242	0.023254
RTL 102.5_January	0.0	0.0	0.0	0.186057	0.000779	0.043834	0.002050
Corriere della Sera_January	0.0	0.0	0.0	0.299739	0.009079	0.437389	0.023230
Matteo Salvini_January	0.0	0.0	0.0	0.045127	0.095048	0.050655	0.002821
...	...	...	...	...	...	...	...
DinamoPress_December	0.0	0.0	0.0	0.008679	0.000000	0.023254	0.000000
Reuters Italia_December	0.0	0.0	0.0	0.000000	0.000000	0.022000	0.000000
L'Osservatore Romano_December	0.0	0.0	0.0	0.014625	0.000000	0.001292	0.000000
Fondazione Umberto Veronesi_December	0.0	0.0	0.0	0.010667	0.015915	0.026697	0.000000
Today_December	0.0	0.0	0.0	0.000000	0.000000	0.006886	0.000000

Figure 3.4: Velocity Scores - 2018

3.3 Calculating Modularity

Last but not least based on the similarities described in section 3.1., the modularity of the matrix was calculated as well.
Before calculating the actual modularity, the matrix was converted into a network using the networkX Python package. Using the Louvain algorithm for community detection, the clustering of the algorithm was created. After that, the modularity was calculated using the same networkX package.
The average Louvain modularity, based again on two consecutive months (so 11 values for one year) was then also plotted, which can be seen in figure 3.6.

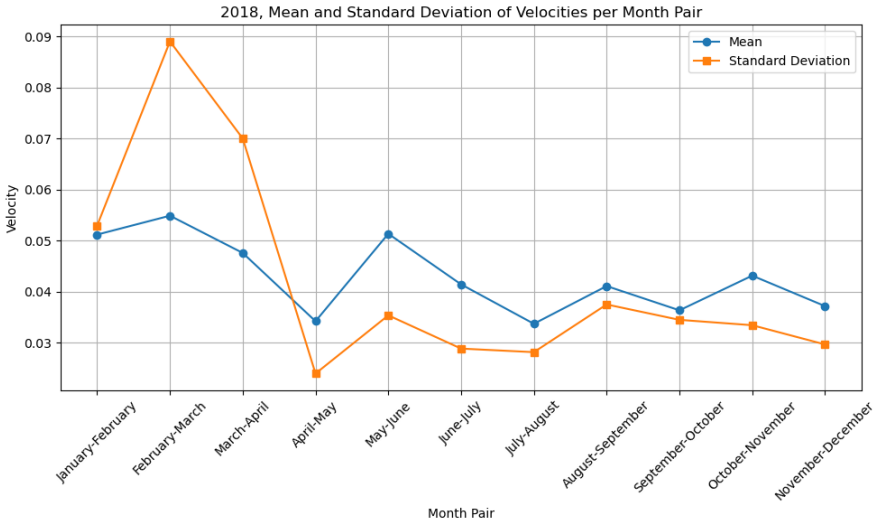

Figure 3.5: Mean and Standard Deviation of the Velocities

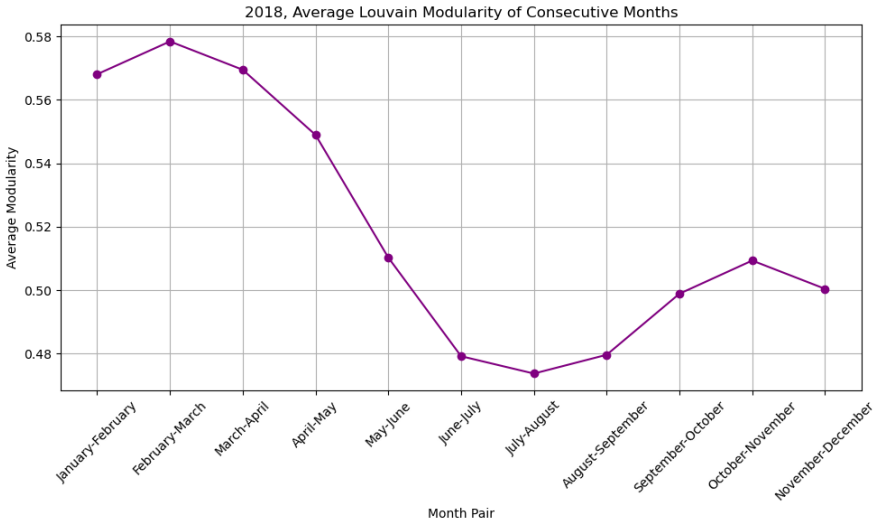

Figure 3.6: Average modularity for year 2018

Chapter 4

Results & Discussion

In this part of the report, the results will be presented along with the interpretation of them.

4.1 Calculating Statistical Measures

As a result of this project, statistical measures such as average mean and average standard deviation based on the similarity have been calculated. Furthermore, modularity averages of two consecutive months (for example January-February) were calculated as well. Last but not least the mean and standard deviation of the velocities has been calculated as well. These results will be demonstrated in the following sections. The statistics table can be also found in the appendix.

4.2 Plotting Reduced Dimensionality Graphs

To understand better how is the similarity of individual nodes (or leaders) changing on a monthly basis, a snapshot of these similarities has been visualised. Because the data is high-dimensional, for example, in 2018 there are 255 unique leaders who have tweeted and have been retweeted, a principal component analysis (PCA) has been applied. The top 2 components have been chosen to visualise the results. The figures 4.1. and 4.2. depict this reduced dimensionality graphs.
In 2018, elections were held in Italy. Based on the graphs, it can be interesting to see how are the leaders spreading out as time passes. This could be possibly interpreted as polarisation of the Italian society, though more in depth research would be needed to validate this statement. All 12 graphs can be found in the appendix. All graphs from all years can be found on the GitHub link.

4.3 Plotting Time Series Modularity Analysis

After all of the 5 years have been analysed, a joint modularity time series graph has been created. The mean velocity and mean standard deviation stays roughly in the same margins, indicating that the spread of the data around the mean does not change much, except for one outlier right at the beginning of year 2018.
The modularity, however, tends to change somewhat dynamically. As can be seen in the 2018, after the election in the beginning of March, the modularity starts gradually dropping, reaching an all time low in July-August 2018. This means the nodes, as could be seen in

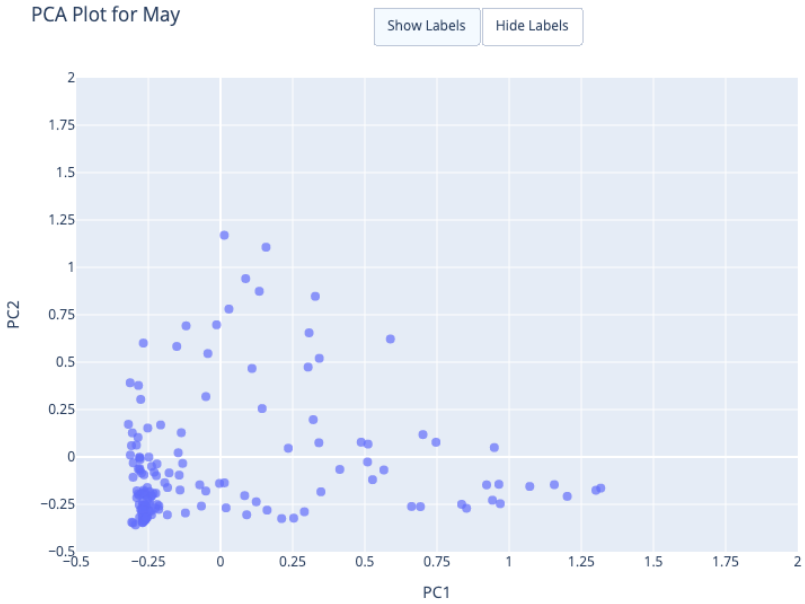

Figure 4.1: PCA Plot - May 2018

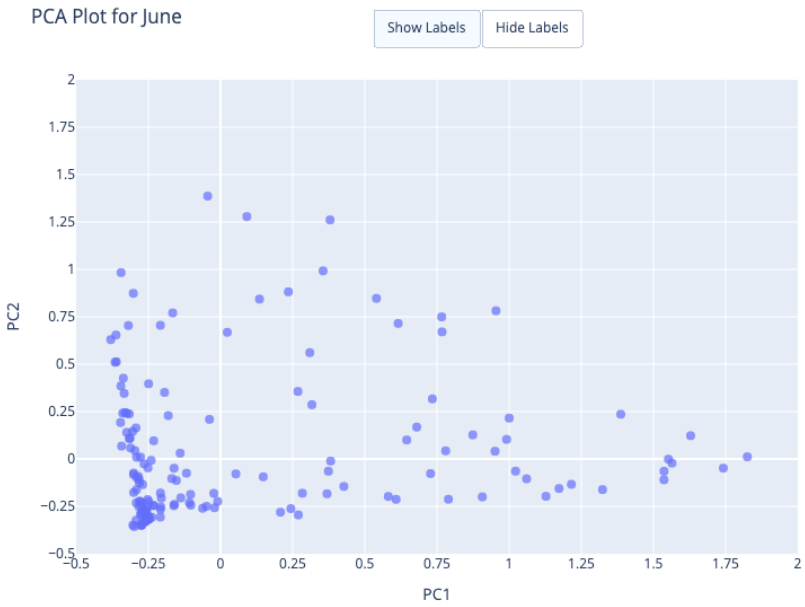

Figure 4.2: PCA Plot - June 2018

the PCA graphs, start to spread out more - there is less clustering in the graph. This could support the statement mentioned before about the polarisation of the society.

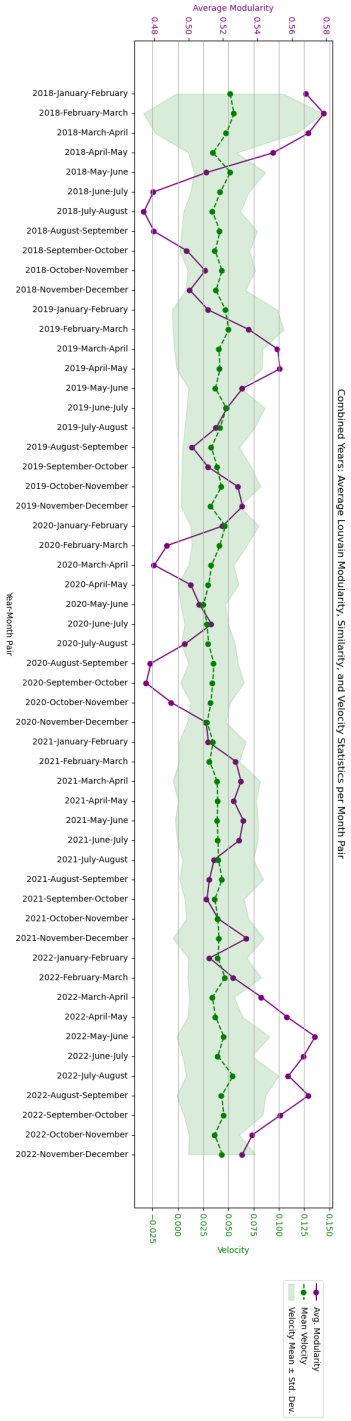

Figure 4.3: Time Series Modularity Analysis - 5 Years

4.4 Plotting Correlation Heatmap

Last but not least, a Spearman correlation heatmap has been developed to see which features are correlating. The Spearman correlation has been chosen, because, for example, the velocities are not normally distributed, which is why Spearman is suitable. The heatmap displays the p-values of the individual correlations between features.
Apart from the average standard deviation and modularity correlation, all of the features are correlated and statistically significant.

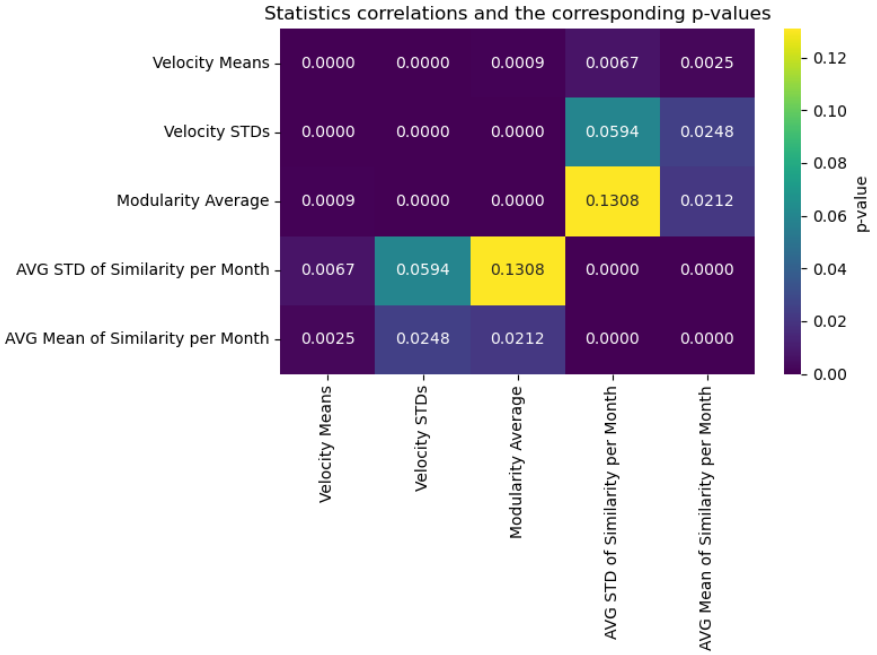

Figure 4.4: Correlation Heatmap of Statistics Results

References

Becatti, C., Caldarelli, G., Lambiotte, R. and Saracco, F. (2019), 'Extracting significant signal of news consumption from social networks: the case of twitter in italian political elections', *Palgrave Communications* **5**(91).
URL: *https://doi.org/10.1057/s41599-019-0300-3*

Caldarelli, G., Chessa, A., Pammolli, F., Pompa, G., Puliga, M. et al. (2014), 'A multi-level geographical study of italian political elections from twitter data', *PLOS ONE* **9**(5), e95809.
URL: *https://doi.org/10.1371/journal.pone.0095809*

Federico, L., Mounim, E. Y., Caldarelli, G. et al. (2024), 'Multi-scale analysis of the community structure of the twitter discourse around the italian general elections of september 2022', *Scientific Reports* **14**, 15980.
URL: *https://doi.org/10.1038/s41598-024-65564-6*

Mattei, M., Caldarelli, G., Squartini, T. and Saracco, F. (2021), 'Italian twitter semantic network during the covid-19 epidemic', *EPJ Data Science* **10**(1), 47. Epub published September 9, 2021. PMCID: PMC8427161.
URL: *https://doi.org/10.1140/epjds/s13688-021-00301-x*

Pierri, F., Artoni, A. and Ceri, S. (2020), 'Investigating italian disinformation spreading on twitter in the context of 2019 european elections', *PLOS ONE* **15**(1), e0227821.
URL: *https://doi.org/10.1371/journal.pone.0227821*

Tognini, F., Aiello, L. M., Chiericatti, A., Ruffo, G., Persello, A., Gallotti, R. and Quattrociocchi, W. (2020), 'Analyzing the diffusion of disinformation on social media during the sars-cov-2 pandemic', arXiv preprint arXiv:2009.02960.
URL: *https://arxiv.org/abs/2009.02960*

Zollo, F. (2019), 'Dealing with digital misinformation: a polarised context of narratives and tribes', *EFSA Journal* **17**(S1), e170720.
URL: *https://doi.org/10.2903/j.efsa.2019.e170720*

Appendix A

	year	veloc_means_arr	veloc_stds_arr	consecutive_modularity_averages	sim_avg_std_per_month	sim_avg_mean_per_month
0	2018	[0.051214008463286664, 0.0549131733276637 6, 0....	[0.05299022692465565, 0.0890761021085431, 0.07...	[0.5680653341085709, 0.5783983258293317, 0.569...	{'January': 0.0543861732402 6037, 'February': 0...	{'January': 0.0159100819555 7077, 'February': 0...
1	2019	[0.04648863740507882, 0.0498119932926 5121, 0.0...	[0.05207777475178581, 0.0548885655185 3223, 0.0...	[0.5111162670097047, 0.5347026476729275, 0.551...	{'January': 0.0672210938472 1667, 'February': 0...	{'January': 0.0242961808514 97488, 'February': ...
2	2020	[0.04632057040232594, 0.0410118904438 0686, 0.0...	[0.03383944252611081, 0.0273367669425 1247, 0.0...	[0.51963245106091, 0.4870908112851826, 0.47975...	{'January': 0.0609053897634 0603, 'February': 0...	{'January': 0.0209976843248 14293, 'February': ...
3	2021	[0.03419731597636642, 0.0309870885831 0372, 0.0...	[0.0330341457502581, 0.0299191556220 6083, 0.04...	[0.5112193727677552, 0.5272539271862956, 0.530...	{'January': 0.0603327618496 2287, 'February': 0...	{'January': 0.0207721620649 79664, 'February': ...
4	2022	[0.039197716300539656, 0.0460604823365 7261, 0....	[0.02893567128447 4378, 0.0364886002716 8035, 0....	[0.5118907310648402, 0.5256815905362684, 0.542...	{'January': 0.0574289172224 4035, 'February': 0...	{'January': 0.0187371447263 81, 'February': 0.0...

Figure A.1: Statistics Results Table

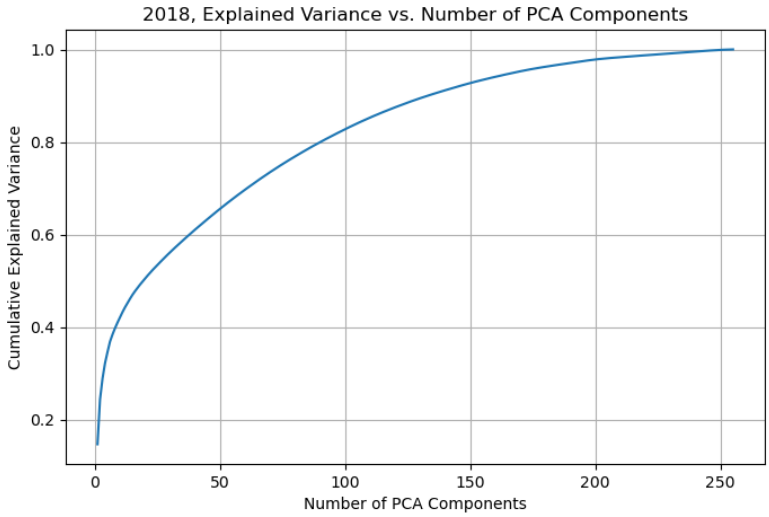

Figure A.2: PCA - Explained Variance

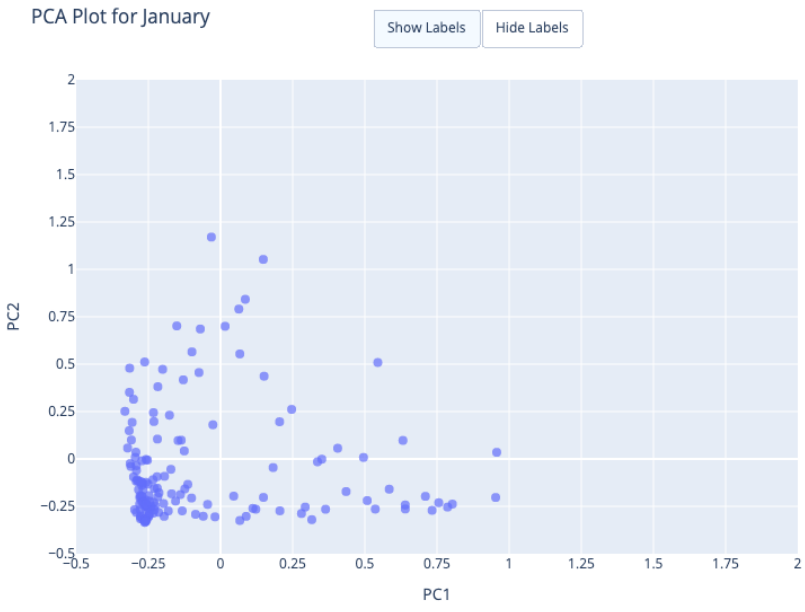

Figure A.3: PCA Plot - January 2018

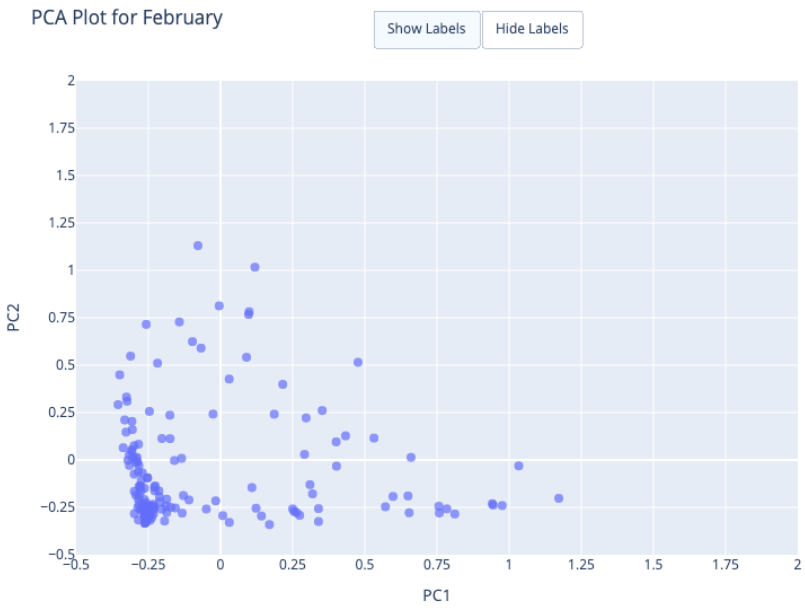

Figure A.4: PCA Plot - February 2018

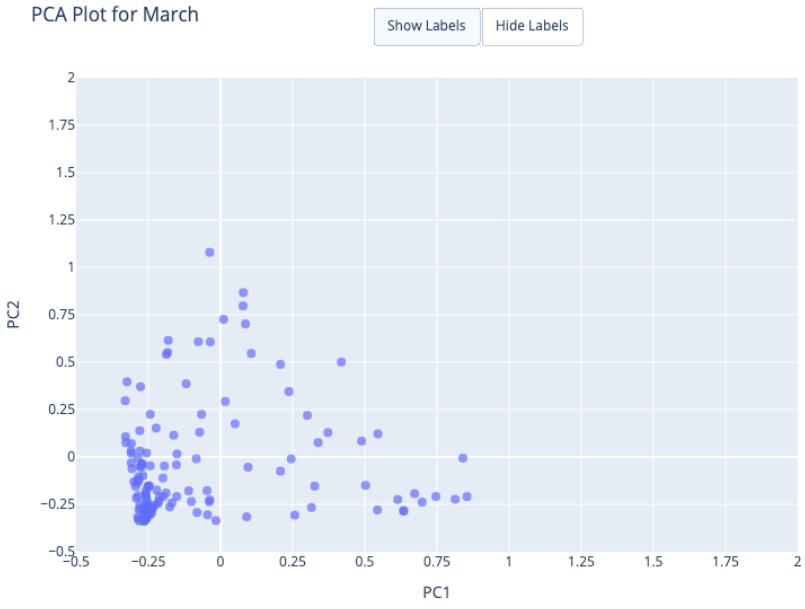

Figure A.5: PCA Plot - March 2018

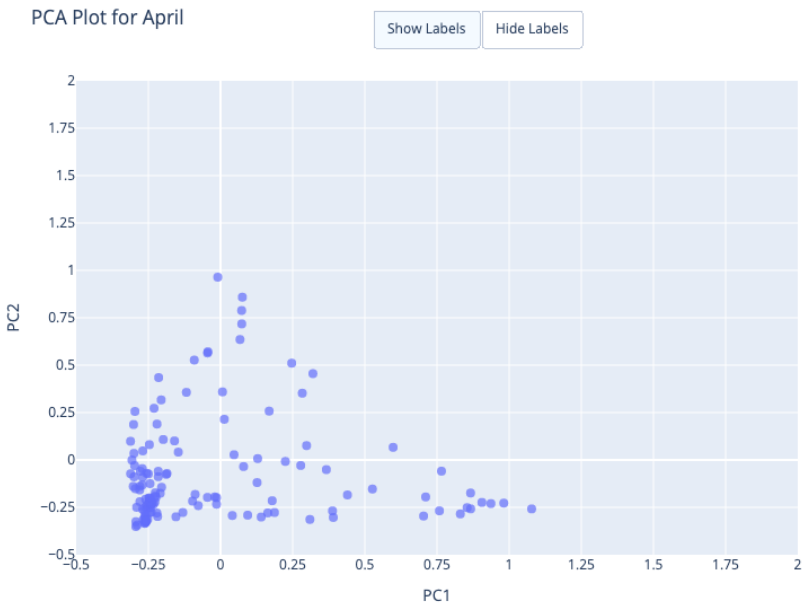

Figure A.6: PCA Plot - April 2018

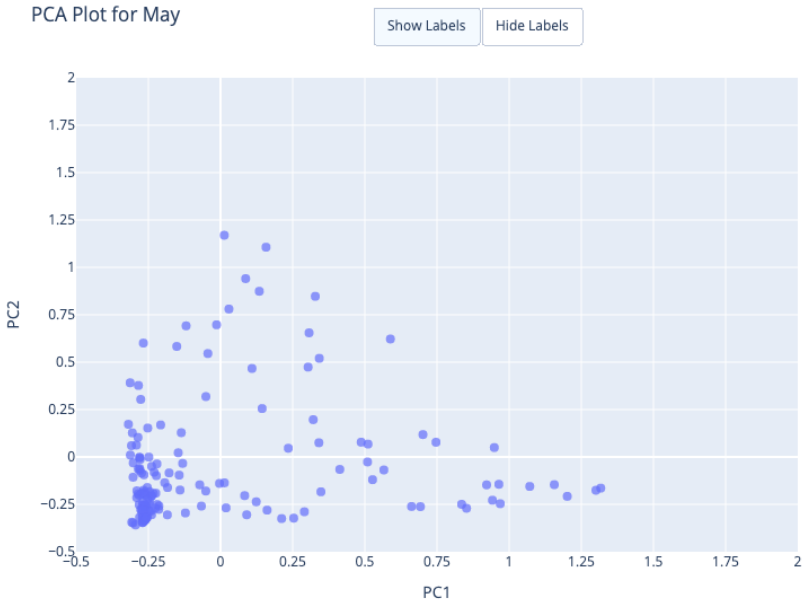

Figure A.7: PCA Plot - May 2018

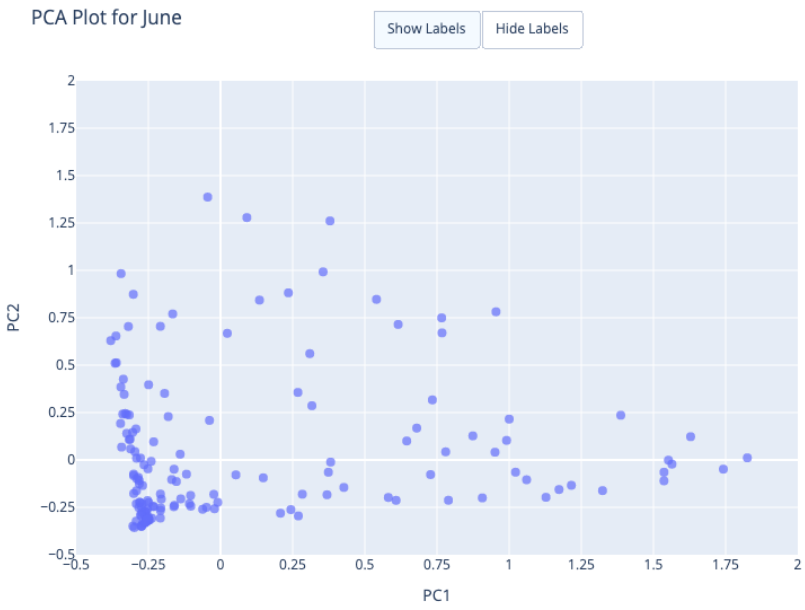

Figure A.8: PCA Plot - June 2018

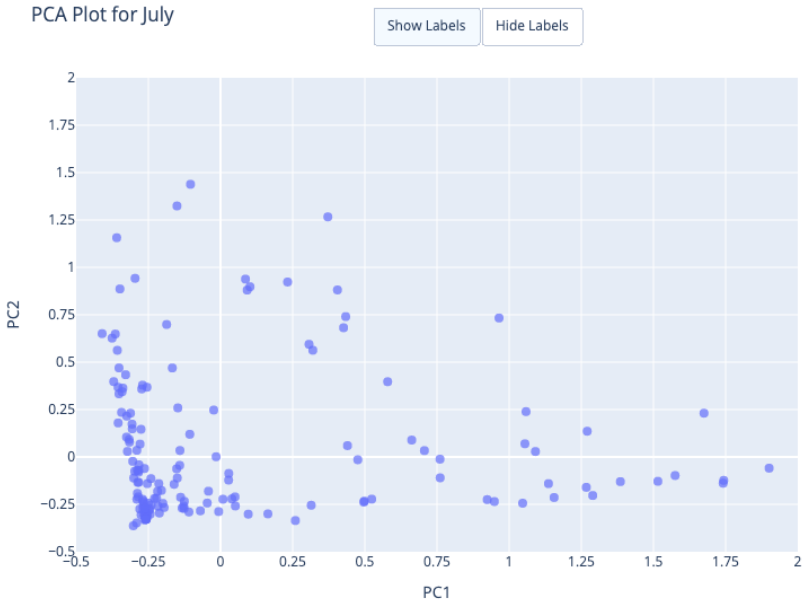

Figure A.9: PCA Plot - July 2018

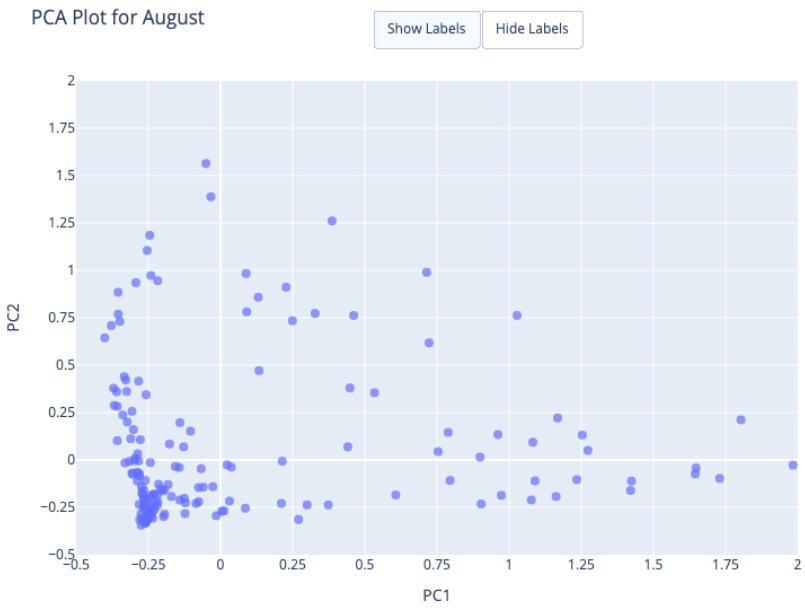

Figure A.10: PCA Plot - August 2018

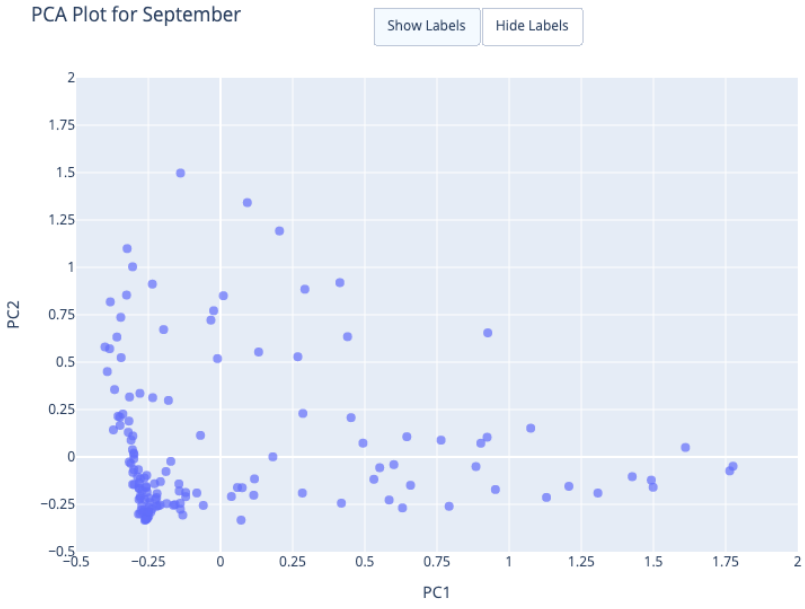

Figure A.11: PCA Plot - September 2018

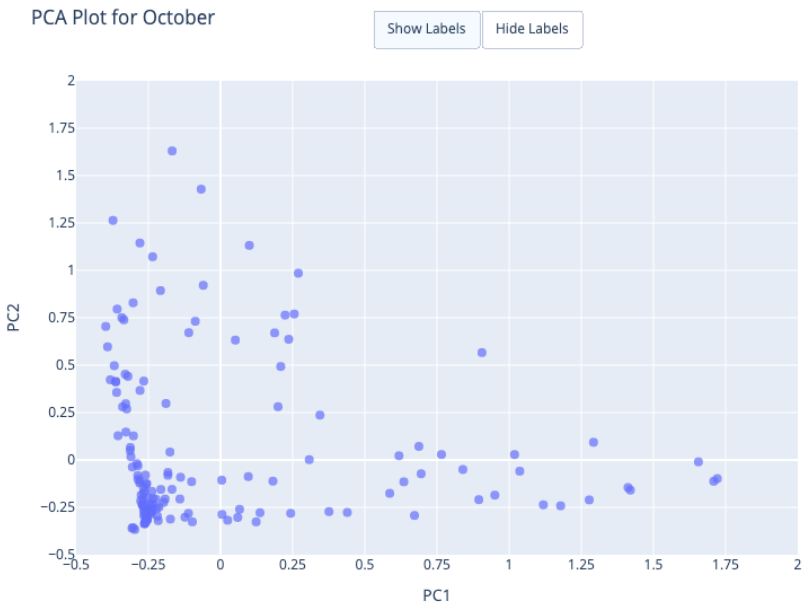

Figure A.12: PCA Plot - October 2018

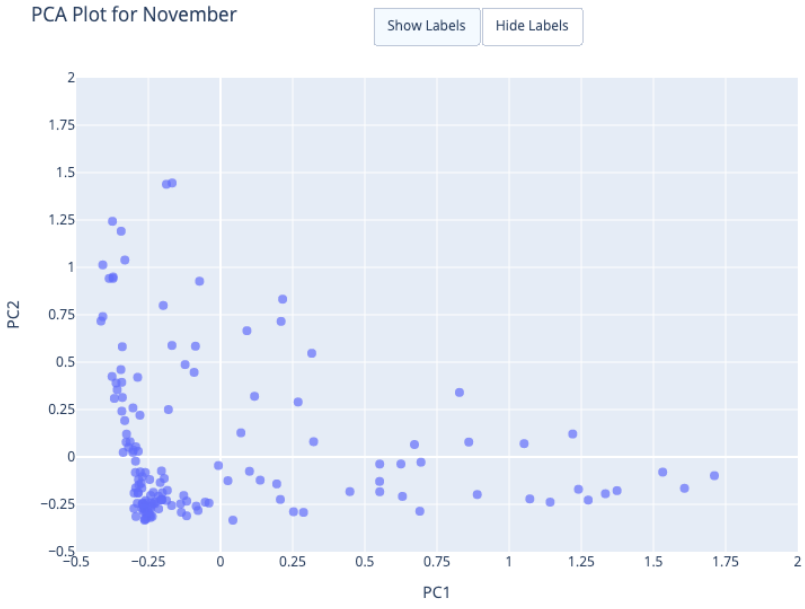

Figure A.13: PCA Plot - November 2018

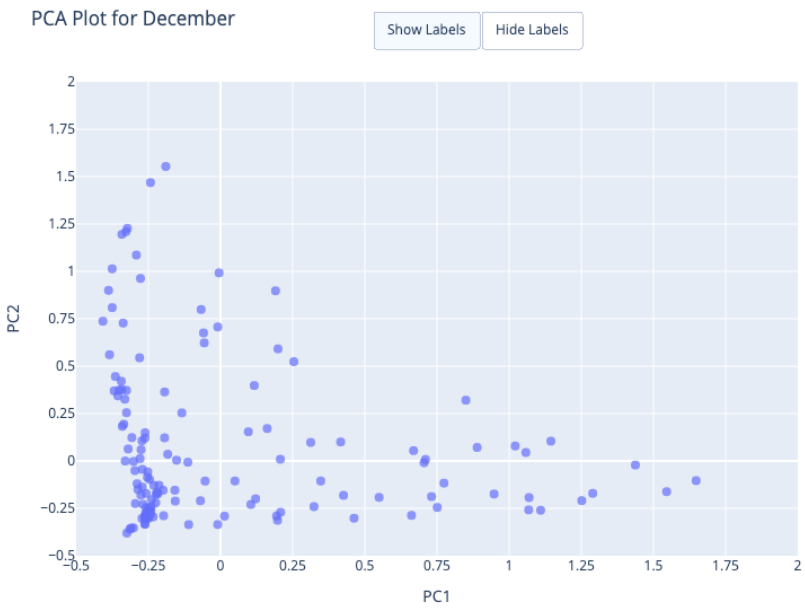

Figure A.14: PCA Plot - December 2018