A Slovenian Retweet Network 2018-2020

Bojan Evkoski
Jožef Stefan International
Postgraduate School,
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
Bojan.Evkoski@ijs.si

Igor Mozetič &
Nikola Ljubešić &
Petra Kralj Novak
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

As the popularity of social media has been growing steadily since the beginning of their era, the use of data from these platforms to analyze social phenomena is becoming more and more reliable. In this paper, we use tweets posted over a period of two years (2018-2020) to analyze the socio-political environment in Slovenia. We use network analysis by applying community detection and influence identification on the retweet network, as well as content analysis of tweets by using hashtags and URLs. Our study shows that Slovenian Twitter users are mainly grouped in three major socio-political communities: Left, Center and Right. Although the Left community is the most numerous, the most influential users belong to the Right and Center communities. Finally, we show that different communities prefer different online media to inform themselves, and that they also prioritize topics differently.

Keywords

Complex networks, Twitter, community detection, influencers

1. INTRODUCTION

Since the rise of the social networks, their data has been extensively used in social analysis. As the popularity of these platforms continues to grow daily, using them as a proxy to analyze specific phenomena is becoming more and more reliable. Their popularity, accessibility and availability made them the go-to way to share one's opinion, support another and even get in conflict with an opposing one. Recently, with the targeted advertising advancements, social media became the most important cultural and political battlefront.

In this paper, the country of interest is Slovenia and the proxy is Twitter data. By following the methodology developed in [3, 2, 4, 8], we address the following questions:

- Are there groups of densely connected Twitter users in the Slovenian retweet network 2018-2020?
- Who are the leading influencers in these groups?
- What is the content of the tweets in these groups and how much does it overlap?

This paper is organised as follows. In Section 2, the data acquisition process and the collected Twitter data are presented. Section 3 discusses the communities in the retweet network and their properties. Section 4 covers the notion of influencers and identifies the main influencers in the Slovenian retweet network. Section 5 investigates the content of the tweets in terms of hashtags and URLs. We draw conclusions in Section 6.

2. DATA

We acquired 5,147,970 tweets in the period from January 2018 to January 2020 with the TweetCat tool [6], built specifically for collecting Twitter data written in "smaller" languages. The tool identifies users tweeting in the focus language by searching for most common words in that language through the Twitter Search API, and collects these users' tweets through the whole data collection period. On average, the dataset containis around 8,000 tweets per day, with the three highest volume peaks on March 13, 2018 (11,556 tweets, the resignation of Slovenia's PM, Miro Cerar), June 1, 2018 (13,506 tweets, the last day of the 2018 Slovenian parliamentary elections campaign), and May 9, 2019 (12,381 tweets, Eurovision semi-final in which Slovenia had a successful run). The variation of the daily volume of tweets is affected by many phenomena, but the more evident are: a weekly seasonality with high volumes on working days and low volumes on weekends, extraordinary periods for the country (e.g. the 2018 Slovenian parliamentary elections campaign, boosting average daily tweets by around 2,000), and holidays (e.g. 2018 and 2019 Easters as local minima with 5,174 and 4,887 tweets, respectively).

3. COMMUNITY DETECTION

We used the collected tweets to construct a retweet network for the purpose of community detection. A retweet network is a directed weighted graph, where nodes represent Twitter users and edges represent the retweet relations. An edge from node (user) A to node B exists if B retweeted A at least once, indicating the information spread from A to B, or A influenced B. Note that retweeting a retweet is actually retweeting the original tweet (source), thus ignoring all intermediate retweets. The weight of an edge is the number of times user B retweeted user A. We removed all self-retweets, since they did not provide us additional information for community and influence detection. Consequently, we formed a network with 10,876 users (94% of all users) and 1,576,792 retweets (92% of all retweets).

This network can be simplified if the direction of the edges is ignored, meaning that two users are linked if one retweets the other while the source and destination are irrelevant. It turns out that such undirected retweet graphs between Twitter users are useful to detect communities of like-minded users who typically share common views on specific topics.

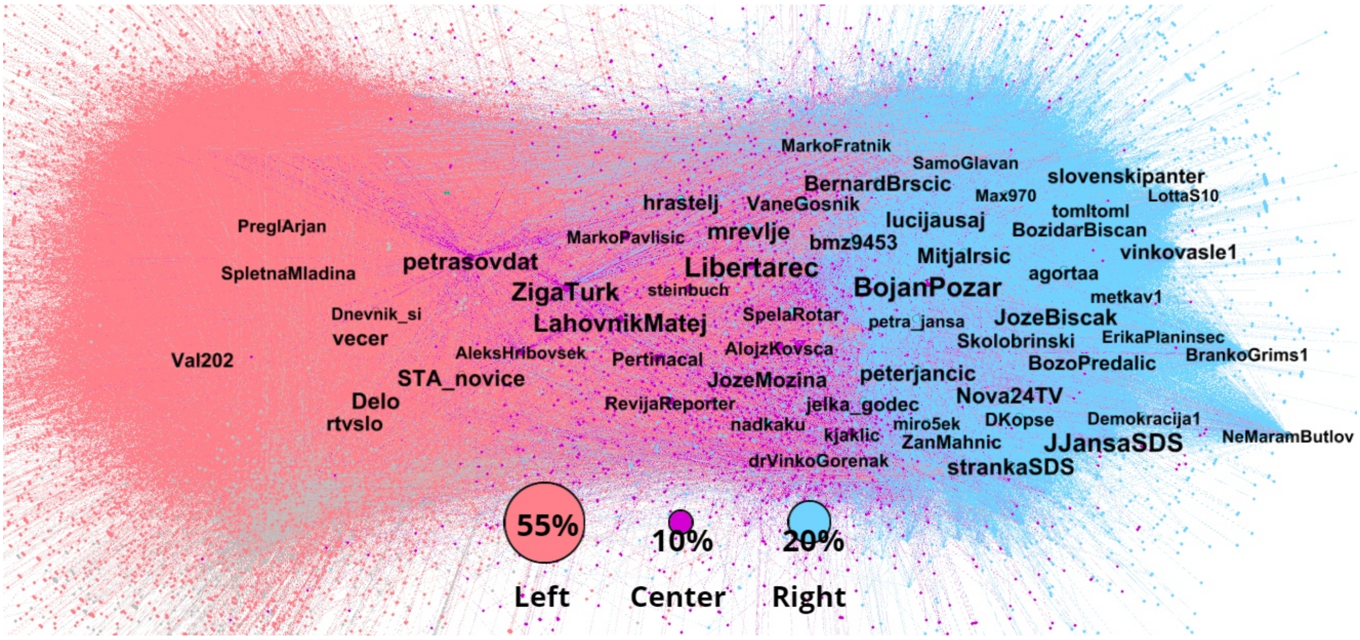

Figure 1: The Slovenian retweet network (2018-2020) colored according to the detected communities, with shares of the total number of users. The label size of a node corresponds to the number of unique users that retweeted it. Only nodes with at least 700 unique retweeters are included.

In complex networks, a community is defined as a subset of nodes that are more closely connected to each other than to other nodes. For the purpose of this paper, we apply a standard algorithm for community detection, the Louvain method [1]. The method partitions the nodes into communities by maximizing modularity (which measures the difference between the actual fraction of edges within the community and such fraction expected in a randomized graph with the same degree sequence) [7]. Modularity values range from -0.5 to 1.0, where a value of 0.0 indicates that the edges are randomly distributed, and larger values indicate a higher community density.

We ran the Louvain method (resolution = 1.05) on our undirected retweet network resulting in 183 communities with a modularity value of 0.382, which indicates a strong connectedness within communities. Only the three largest communities each have more than 5% of all users, while combined they contain 85% of all users. The three main detected communities are presented in Fig. 1. We observe the following:

- The three largest communities are labeled as Left, Center and Right with 55%, 20% and 10% as their respective shares of all users. The labeling of the communities does not necessarily represent their political orientation.
- The Left community, even though the largest, contains the smallest number of users with more than 700 unique retweeters.
- The Left community is well separated from the Center and the Right communities, which are more tightly interlinked.

We performed an exploratory data analysis and calculated the community properties presented in Table 1, to compare the communities. Most of the properties are normalized by the user to ease the comparison between communities.

- Nodes – unique users count
- Central user – user with most retweets
- Central user retweets – times the central user is retweeted
- Central user retweeters – unique users retweeting the central user
- HHI ($n = 50$) – Herfindahl–Hirschman index [9] measures the distribution of influence of the top n influential users. Higher value reflects the community influence concentrated only in few influential users, while lower value indicates more dispersed and balanced influence distribution.
- Edges in/node – edges remaining in the community per user (source and destination in the same community)
- Edges out/node – edges going out of the community per user (destination in a different community)
- Weighted edges in/node – weighted edges remaining in the community per user
- Weighted edges out/node – weighted edges going out of the community per user
- Out/In ratio – "Edges out" divided by "Edges in"
- Weighted out/in ratio – "Weighted edges out" divided by "Weighted edges in"

4. INFLUENCERS

We use two simple, but powerful metrics to detect influencers in the retweet network: the weighted out-degree and the Hirsch index (h-index) [5]. Both metrics are calculated from the number of retweets, thus known as retweet influence metrics, indicating the ability of a user to post content of interest to others.

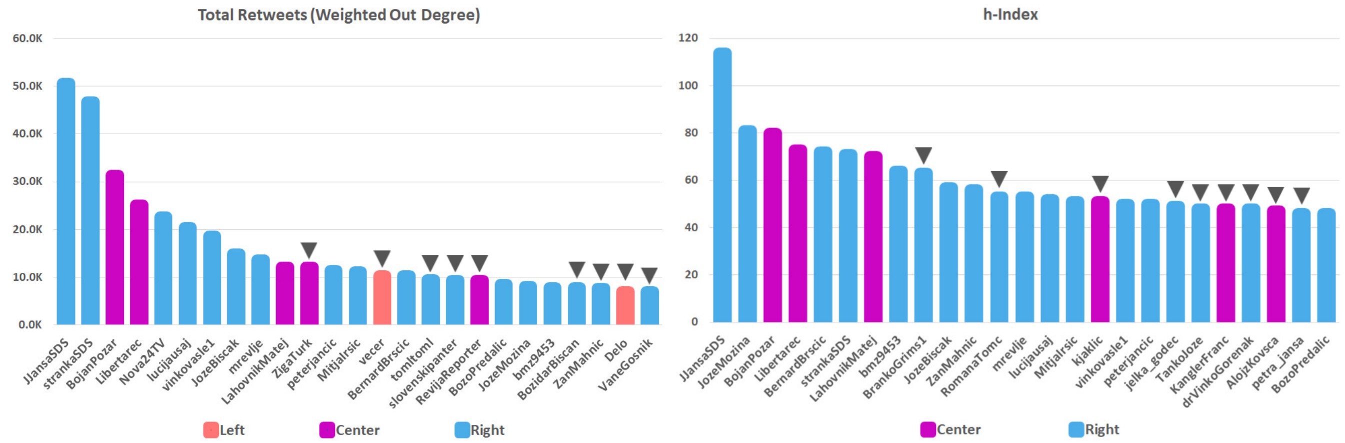

Figure 2: Weighted out-degree (total retweets) and h-index comparison. Both charts include the top 25 most influential Slovenian Twitter users according to their respective metric. Bar colors represent the community of a user. Triangles point to users exclusive to one of the charts.

Table 1: Community properties

	Left	Center	Right
Nodes	7,030	1,223	2,519
Central user	vecer	BojanPozar	JJansaSDS
Central user retweets	10,398	31,432	50,688
Central user retweeters	973	1,325	1,242
HHI ($n = 50$)	0.031	0.066	0.042
Edges in/node	19.32	14.53	69.30
Edges out/node	4.47	37.11	13.19
Weighted edges in/node	52.91	83.68	308.33
Weighted edges out/node	6.95	119.42	36.14
Out/In ratio	0.23	2.55	0.19
Weighted Out/In ratio	0.13	1.43	0.12

Weighted out-degree is simply the total number of retweets of a particular user, while the h-index is an author-level bibliometric indicator that measures the scientific output of a scholar by quantifying both the number of publications (i.e., productivity) and the number of citations per publication (i.e., citation impact). Adapted to a Twitter network, it would be described as: a user with an index of h has posted h tweets and each of them was retweeted at least h times.

Let RT be the function indicating the number of retweets for each original tweet. The values of RT are ordered in decreasing order, from the largest to the lowest, while i indicates the ranking position in the ordered list. The *h-index* is then defined as follows:

$$h\text{-}index(RT) = \max_i \min(RT(i), i)$$

The top 25 most influential users by weighted out-degree and h-index are shown in Fig. 2. The two metrics provide fairly similar results (they differ only in 9 users). Both results confirm the already visible phenomena from the previous observations: The Right community has the most influential users, while the Left community, even though the biggest, does not have nearly as popular users as the ones from the other two communities.

5. CONTENT ANALYSIS

We refer to content analysis in terms of getting knowledge from the text of the tweets. In this paper, we perform two kinds of content analysis: domain URLs and hashtags.

For domain URLs, we filtered the 2,297,008 tweets which contain a URL. Then, we extracted the domain part of the URLs and removed the domains with no specific meaning for Slovenia's content analysis (e.g. social networks: twitter.com, facebook.com, instagram.com, etc., and URL shorteners: ift.tt, bit.ly, ow.ly, etc.). This results in 512,308 tweets (approximately 22% of all the tweets with links). The most frequently occurring domains are owned by Slovenian media with nova24tv.si, rtvslo.si and delo.si as the top three URL domains with 23,879, 20,210 and 17,360 occurrences respectively. If instead of the total number of occurrences we count only the unique number of users which posted a domain URL, the top three domains are rtvslo.si, siol.net and delo.si with 2,802, 2,193 and 2,186 unique users respectively.

For the hashtag analysis, we filtered only tweets which contain a hashtag, ending up with 701,266 tweets. The top three hashtags are the following: #volitve2018 (the 2018 Slovenian parliamentary elections), #plts (the Slovenian First Football League) and #sdszate (Slovenian Democratic Party hashtag, meaning: SDS for you) with 9,845, 9,318 and 7,308 occurrences respectively. If we count only the unique number of users using a particular hashtag, the results for the top three Slovenian hashtags are as follows: #volitve2018 with 2,473, #slovenija with 1,611 and #fakenews with 1,343 users.

To see these results in the context of communities, we look at the tweets authored by members of the three largest communities, resulting in 84% of the tweets with relevant domain URLs and 83% of the tweets with relevant hashtags. We summed the domain URL counts, while grouping them by the community in which their user belongs. We applied the same procedure to the hashtags. Finally, we filtered the top eight domain URLs and hashtags for each community and put them on a single Sankey diagram in Fig. 3. Even though overlaps exist, the most popular hashtags and media very much differ from community to community, meaning that all three main communities prioritize topics differently and they inform themselves via different media.

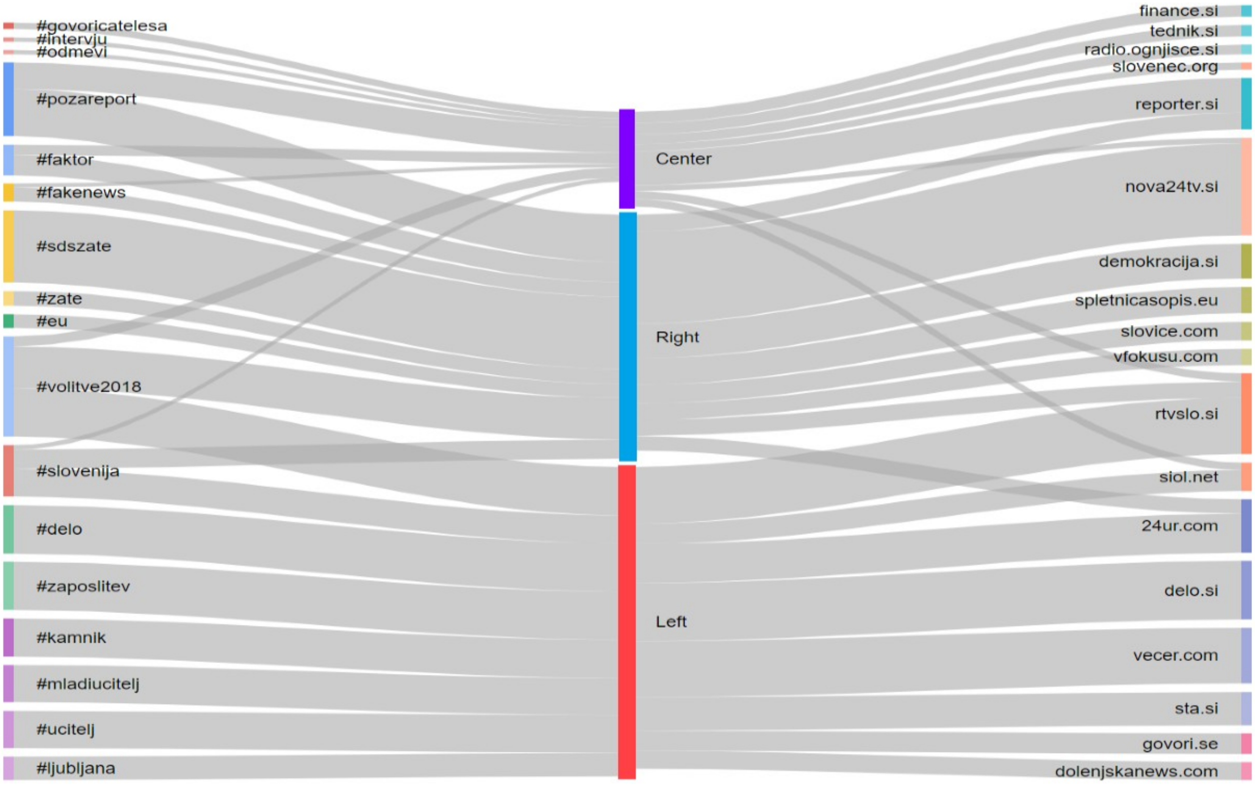

Figure 3: A Sankey diagram depicts the use of the eight most common hashtags (left-hand side) and URLs (right-hand side) by the three largest detected communities.

6. CONCLUSIONS

In this paper we explored the Slovenian twitter network from January 2018 until January 2020. We applied community detection, identifying three main communities: Left, Center and Right. We identified the most influential and the central users of each community by calculating the weighted out-degree and the h-index of the nodes. We used the Herfindahl–Hirschman index to estimate the distribution of influence within the top communities in the network. Finally, by analysis of hashtags and URL domains in tweets, we discovered the most popular topics for Slovenians as well as the most referred Slovenian media on Twitter. We showed that users from different communities prioritize different topics and use different media to inform themselves.

7. ACKNOWLEDGMENTS

The authors acknowledge financial support from the Slovenian Research Agency (research core funding no. P2-103 and P6-0411), and the European Union's Rights, Equality and Citizenship Programme (2014-2020) project IMSyPP (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech, grant no. 875263).

8. REFERENCES

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[2] D. Cherepnalkoski, A. Karpf, I. Mozetič, and M. Grčar. Cohesion and coalition formation in the European Parliament: Roll-call votes and Twitter activities. *PLoS ONE*, 11(11):e0166586, 2016.

[3] D. Cherepnalkoski and I. Mozetič. Retweet networks of the European Parliament: Evaluation of the community structure. *Applied Network Science*, 1(1):2, 2016.

[4] M. Grčar, D. Cherepnalkoski, I. Mozetič, and P. Kralj Novak. Stance and influence of Twitter users regarding the Brexit referendum. *Computational Social Networks*, 4(1):6, 2017.

[5] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, pages 16569–16572, 2005.

[6] N. Ljubešić, D. Fišer, and T. Erjavec. TweetCaT: a tool for building Twitter corpora of smaller languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2279–2283, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[7] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[8] P. K. Novak, L. D. Amicis, and I. Mozetič. Impact investing market on twitter: influential users and communities. *Applied Network Science*, 3(1):40, 2018.

[9] G. J. Werden. Using the Herfindahl–Hirschman index. In L. Phlips, editor, *Applied Industrial Economics*, number 2, pages 368–374. Cambridge University Press, 1998.