Revealing the structure of domain specific tweets via complex networks analysis

Edvin Močibob, Sanda Martinčić-Ipšić, Ana Meštrović
Department of Informatics,
University of Rijeka,
Radmile Matejčić 2, 51000 Rijeka, Croatia
Email: {emocibob, smarti, amestrovic}@uniri.hr

Abstract—**In this paper we explore the relation between different groups of tweets using complex network analysis and link prediction. The tweets were collected via the Twitter API depending on their textual content. That is, we searched for the tweets in English language containing specific predefined keywords from different domains. From the gathered tweets a complex network of words was formed as a weighted network. Nodes represent words and a link between two nodes exists if these two words co-occur in the same tweet, while weight denotes the co-occurrence frequency. The Twitter search was repeated for four different search criteria (API queries based on different tweet keywords), thus resulting in four networks with different nodes and links. The resulting networks were subjects to further network analysis, as comparison of numerical properties for different networks and link prediction for individual networks. This paper shows the tweet scraping process, our approach to building the networks, the measures we calculated for them, the differences and similarities between different networks we built and our success in predicting future links.**

I. Introduction

Twitter is a popular online social network created in 2006 that enables user to send publicly visible messages called ”tweets”. One of the main characteristics that distinguishes Twitter from other online social networks is the limit on tweet length. Twitter user are allowed to send tweets that have a maximum of 140 characters. Hence, Twitter is often categorized as a micro-blogging platform. It is estimated that in 2015 Twitter had over half a billion users. [?]

Because of its popularity, user-base size and vast amounts of tweets, Twitter has been studied in the context of person-to-person relations [?], user influence [?], economic predictions [?], predictions of political elections [?], conversational practices [?] and trends discovery [?].

Another important research domain related to Twitter is sentiment analysis. In [?] Pak et al. automatically collect from Twitter a corpus and perform linguistic analysis on it. Then they build a sentiment classifier able to determine positive, negative and neutral sentiments for a document. There has been reported research in automatic classification of tweets regarding their sentiment [?]. [?] gives a detailed revision of the field of sentiment analysis with Twitter in focus. Research by Agarwal et al. [?] examines sentiment analysis on Twitter data. In it the authors introduce POS-specific prior polarity features and explore the use of a tree kernel to eliminate the need for laborious feature engineering. In [?] Kouloumpis et al. investigate the utility of linguistic features for detecting tweets sentiment using a supervised approach, while also leveraging existing hashtags in building training data. Wang et al. [?] present hashtag-level sentiment classification which aims to automatically generate the overall sentiment polarity for a given hashtag in a certain time period.

The following papers use the complex network analysis approach to Twitter data. Villazon et al. in [?] look at Twitter as a complex network, calculating the cluster coefficient, power law and average path length for it. [?] presents a model for describing the growth of scale-free networks. The model is applied only after checking that Twitter is indeed a scale-free network, and for that purpose the mentioned paper proposes a new heuristic method of finding the upper bounds of the path lengths instead of computing the exact length.

In our approach we use complex networks analysis to reveal the structure of domain specific tweets. The motivation of our research is to detect weather networks constructed from different tweets domains have different structural properties. More precisely, the goal of this research is to determine whether (and which) complex network measures can distinguish between networks of tweets with ”positive” and ”negative” aspects. Possible applications of proposed approach can be in the domain of sentiment analysis. Furthermore, link prediction enables anticipation of positive or negative attitude propagation on Twitter.

We collect positive tweets in English language using keywords with positive polarity (e.g. joy, happiness, ...) and negative tweets using keywords with negative polarity (e.g. anger, fear, ...). Then we perform the global and local complex network analysis where we compare results for four obtained networks. On the global level we use a standard set of network measures (e.g. diameter, average path length, clustering coefficient). However, for the local level analysis we apply a node selectivity measure encouraged by our previous findings [?], [?], [?] for which we show that it is an important measure for language networks analysis and differentiation.

In the second Section we present the network measures used in our research. In the third Section we describe how we construct the tweet networks. The results and discussion are given in the fourth Section. Finally, the fifth Section contains conclusions and directions for the further research.

II. Networks measures

Complex network is a graph with non-trivial topological features (e.g. high clustering coefficient, low distances, heavy-tailed degree distribution, etc.). It can be represented with a graph G, defined as a pair of two sets $G = (V, E)$; the first set V consisting of vertices and the second set E consisting of edges. N as the number of vertices in V and K as the number of edges in E. In the domain of network analysis, the vertices are referred as nodes and the edges are called links.

Network analysis can be classified by the following three levels: macro-scale or global level, meso-scale level and micro-scale or local level. In weighted complex networks every link connecting two nodes u and v has an associated weight w_{uv}. A node degree is the number of links directly connected (or incident) to that node. The set of nodes incident to a node v is denoted as $\Gamma(v)$. The number of network components is represented by ω. Next, we present network measures that will be used in the following sections.

The average network degree is the ratio of the number of links to the number of nodes. For undirected networks we multiply this ratio by 2 since undirected links always have two incident nodes:

$$\langle k \rangle = 2\frac{K}{N}. \tag{1}$$

Network strength is simply the sum of all link weights in a network:

$$S = \sum_{u,v \in V} w_{uv}. \tag{2}$$

For the average network strength we divide a networks strength with its number of nodes:

$$\langle s \rangle = \frac{S}{N}. \tag{3}$$

Node selectivity for a node v corresponds to the sum of weights of all incident links divided by that nodes degree (denoted as $\deg(v)$):

$$e(v) = \frac{\sum_{u \in \Gamma(v)} w_{uv}}{\deg(v)}. \tag{4}$$

Average network selectivity is the sum of all individual node selectivities divided by the number of nodes:

$$\langle e \rangle = \frac{\sum_{v \in V} e(v)}{N}. \tag{5}$$

Network density is represented as the ratio between the number of existing links and the number of all possible links:

$$d = \frac{K}{N(N-1)}. \tag{6}$$

Average path length for a network, where d_{uv} denotes the number of links lying on the shortest path between $u, v \in V$, is computed as following:

$$L = \sum_{u,v} \frac{d_{uv}}{N(N-1)}. \tag{7}$$

The network diameter represents the longest shortest path in a network ($u, v \in V$):

$$D = \max(d_{uv}). \tag{8}$$

The network radius denotes the shortest $\epsilon(v)$, where $\epsilon(v)$ is defined as the maximum distance between $v \in V$ and any other node:

$$R = \min(\epsilon(v)). \tag{9}$$

Network transitivity where possible triangles are identified by the number of triads (two links with a shared node):

$$T = 3\frac{\#triangles}{\#triads}. \tag{10}$$

Average clustering coefficient, where $c(v)$ is the clustering coefficient for a node v, sums all the individual clustering coefficients and divides them by the number of nodes:

$$C = \frac{1}{N} \sum_{v \in V} c(v). \tag{11}$$

The global network efficiency is the reciprocal value of a networks average path length:

$$E = \frac{1}{L}. \tag{12}$$

In the context of link prediction we use the following measures.

Weighted Common Neighbors, adapted from [?], where weights of links connecting u and v to their common neighbors are summed:

$$CN(u,v) = \sum_{z \in \Gamma(u) \cap z \in \Gamma(v)} w_{uz} + w_{vz}. \tag{13}$$

Weighted Jaccard's Coefficient, adapted from [?], which divides the weighted Common Neighbors value for u and v by the summed weights of all links incident to u and/or v:

$$JC(u,v) = \frac{\sum_{z \in \Gamma(u) \cap z \in \Gamma(v)} w_{uz} + w_{vz}}{\sum_{a \in \Gamma(u)} w_{au} + \sum_{b \in \Gamma(v)} w_{bv}}. \tag{14}$$

Lastly, we present the link prediction precision as the ratio between the number of correctly predicted links and the total number of predicted links. That is, we divide the number of true positives ($|TP|$) by the number of true and false positives ($|TP| + |FP|$). [?]

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|} \tag{15}$$

III. Networks construction

The first step in constructing networks is the collection of data. Initially, we searched for four sets of tweets according to the following criteria: a) tweets associated to recent immigrant and war related events; b) tweets containing negatively polarized words; c) tweets associated to house pets and d) tweets containing positively polarized words. The subset of positive and negative polarized words is extracted from the sentiment lexicon in [**?**]. From now on we will refer to the networks built from their respective sets as: a) emo-net^a, b) emo-net^b, c) emo-net^c and d) emo-net^d.

For the data collection process we use Python in combination with the Python Twitter Tools package, which provides an easy-to-use interface for the official Twitter API. In the API request arguments we specified we are searching for a mix of recent and popular tweets in the English language. We scraped about 10000 tweets for each of four different queries, resulting in a dataset of 39882 tweets. It is worth to mention that the official Twitter API documentation states that the language detection is based on the "best-effort" principle [**?**].

In the text (tweets) preparation step first we eliminate stopwords[1], and from the remaining text we compute the 100 most frequent words for each of the four subsets. We selected top 100 words as the reasonable list which provides the best trade-off between computation time and link prediction results. Note that the former computation was case-insensitive and we used the list of English stopwords presented at http://www.ranks.nl/stopwords.

From the words of preprocessed tweets extended with the set of explicit keywords (e.g. joy, puppy) used for retrieving each of the tweets we form the nodes of the networks. Link between two nodes (words) is established if these two word appear together in the same tweet. Weight on the link represents words co-occurrence frequencies, that is, the number of tweets in which two high-frequency words from the top 100 list co-occurred. That makes the generated networks weighted and undirected. Hence, based on the high-frequency words, we generate four different networks for each of the four data sets.

We build 16 distinct networks from four datasets: the first network is built from 25% of the data, the second from 50%, the third from 75% and the fourth from 100% of the data in one dataset. We will denote those networks, respectively, as emo-net$_1^x$, emo-net$_2^x$, emo-net$_3^x$ and emo-net$_4^x$, where $x \in \{a, b, c, d\}$. That means we, as previously mentioned, generate a total of 16 different networks, four per each dataset.

Some other used Python packages not previously mentioned are NetworkX [**?**] and LaNCoA [**?**]. The first one is a popular Python tool for creating and manipulating complex networks. It also provides a rich collection of functions for studying complex networks on various levels. The LaNCoA toolkit provides procedures for construction and analysis of complex language networks.

IV. Results

1) Global and local network measures: Here we present the computed global and local network measures for emo-net$_4^a$, emo-net$_4^b$, emo-net$_4^c$ and emo-net$_4^d$. Table **??** shows the calculated measures that were previously described in Section **??**.

TABLE I
Global and local network measures

Measure	emo-net$_4^a$	emo-net$_4^b$	emo-net$_4^c$	emo-net$_4^d$
N	101	101	103	104
K	3454	3958	2854	3848
$\langle k \rangle$	68.396	78.3762	55.4175	74
$\langle s \rangle$	1025.9406	830.505	747.0291	1310.25
$\langle e \rangle$	29.4867	24.0104	42.9054	44.7693
d	0.684	0.7838	0.5433	0.7184
ω	1	1	1	1
L	1.316	1.2162	1.4582	1.2816
D	2	2	3	2
R	1	1	2	1
T	0.7965	0.875	0.7774	0.8595
C	0.0088	0.0208	0.0532	0.0077
A	-0.1257	-0.0933	-0.0587	0.0442
E	0.7599	0.8222	0.6858	0.7803

The first visualization we present (Figure **??**) is for the node degrees across all emo-net$_4$ networks. We see no major differences for node degrees across those networks.

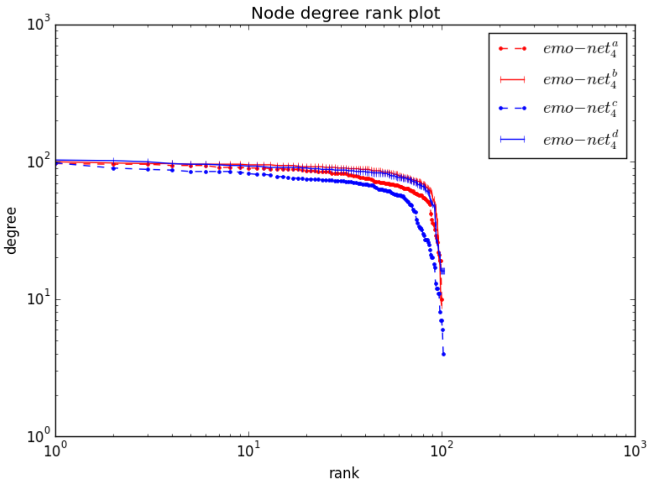

Fig. 1. Node degrees for all emo-net$_4$ networks on a log-log scale

Lets recall that emo-net$_4^a$ and emo-net$_4^b$ were based on data from queries with negative connotations. In contrast, emo-net$_4^c$ and emo-net$_4^d$ were based on queries with positive connotations. The most obvious difference between the first two "positive" and the last two "negative" networks in Table **??** is $\langle e \rangle$, which represent the value of average network selectivity. $\langle e \rangle$ is notably lower for emo-net$_4^a$ and emo-net$_4^b$ than for emo-net$_4^c$ and emo-net$_4^d$. Average network selectivity can be interpreted as how "heavy" the links across a network are. We see how our positive networks have on average stronger ties between nodes.

[1] Stopwords are a list of the most common, short function words which do not carry strong semantic properties, but are needed for the syntax of a language (pronouns, prepositions, conjunctions, abbreviations, ...).

In Figure **??** we visualize the node selectivities for the networks mentioned above. Note that the plot in Figure **??** uses a log-log scale.

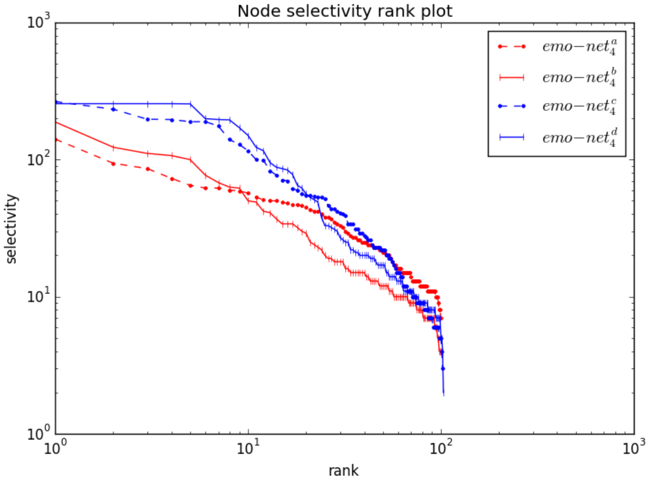

Fig. 2. Node selectivities for all emo-net$_4$ networks on a log-log scale

2) Link prediction: Next we present the results for the link predictions. Here we computed the most likely future links for emo-net$_1^x$, emo-net$_2^x$ and emo-net$_3^x$ where $x \in \{a, b, c, d\}$. The prediction were made using two measures: weighted Common Neighbors (Table **??**) and weighted Jaccard's Coefficient (Table **??**). The definitions of both measures can be found in Section **??**.

We will briefly describe the link prediction process which is the same for both measures. First compute the ranks for all non-existing links in emo-net$_i^x$, $x \in \{a, b, c, d\}$, $i \in \{1, 2, 3\}$. Generate the first set that contains the top n ranked non-existing links in emo-net$_i^x$ (n is the number of new links in emo-net$_{i+1}^x$). Next, generate the second set that holds links which appear in emo-net$_{i+1}^x$ but not in emo-net$_i^x$. Calculate the prediction precision by looking at the intersection of the first and second set.

TABLE II
PREDICTION PRECISION BASED ON THE WEIGHTED COMMON NEIGHBORS MEASURE

Network	emo-net^a	emo-net^b	emo-net^c	emo-net^d
(25%) emo-net$_1$	29.96%	45.92%	30.13%	26.84%
(50%) emo-net$_2$	24.57%	34.88%	21.43%	17.73%
(75%) emo-net$_3$	28.49%	27.49%	18.4%	13.45%

TABLE III
PREDICTION PRECISION BASED ON THE WEIGHTED JACCARD COEFFICIENT MEASURE

Network	emo-net^a	emo-net^b	emo-net^c	emo-net^d
(25%) emo-net$_1$	35.88%	47.96%	37.95%	50.26%
(50%) emo-net$_2$	29.69%	37.72%	28.97%	41.14%
(75%) emo-net$_3$	12.85%	32.16%	30.06%	32.75%

We see from Tables **??** and **??** that all predictions had a precision rate above 10%, with some going as high as 50%. The predictions are, by a large margin, most precise for emo-net$_1$ networks. Generally, those networks will not have all of the probable links already in them. With more data all the probable links are added. In most cases the prediction precision for networks with more links tends to fall. That is the only obvious trend for precision rates across query domains, network sizes and prediction measures.

V. CONCLUSION

In this paper we present how we construct multiple complex networks based on four different data sets. Each data set featured a collection of tweets gathered by predefined Twitter API queries. Two of those queries retrieved "negative" oriented tweets, while the other two gathered "positive" oriented tweets. We investigate global and local network measures across four query categories and compare them between "negative" and "positive" networks. In this paper we also predict future links for networks across all query domains. For that purpose we use networks built form a lower percentage of data and compare them with networks built from a higher percentage of the same data.

Regarding network measures, we found that the average network selectivity is the only measure that discriminates between "negative" and "positive" networks, favoring the positive ones. This preliminary results indicate that selectivity based network measures could be used in the Twitter sentiment analysis tasks.

The link prediction process gave no obvious patterns, except the higher prediction precision for networks built from the smallest amount of data. Also, for all of our networks the link prediction precision was above 10%. It should be noted that all our results are preliminary and a more complex analysis would be in order. Such analysis should primarily consider larger and more diverse data sets. Expanding the list of computed network measures would be also worth considering, along with community detection algorithms.

REFERENCES

[1] "Twitter - Wikipedia, the free encyclopedia," 20-Feb-2016. [Online]. Available: https://en.wikipedia.org/wiki/Twitter. [Accessed: 21-Feb-2016].

[2] B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," arXiv:0812.1045 [physics], Dec. 2008.

[3] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," in ICWSM, 2010, vol. 10, pp. 1017.

[4] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," Journal of Computational Science, vol. 2, no. 1, pp. 18, Mar. 2011.

[5] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," in Fourth International AAAI Conference on Weblogs and Social Media, 2010.

[6] D. Boyd, S. Golder, and G. Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," in 2010 43rd Hawaii International Conference on System Sciences (HICSS), 2010, pp. 110.

[7] M. Mathioudakis and N. Koudas, "TwitterMonitor: Trend Detection over the Twitter Stream," in Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 2010, pp. 11551158.

[8] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, vol. 1, p. 12, 2009.

[9] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining.,” in LREc, 2010, vol. 10, pp. 13201326.

[10] E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña-López, and A. R. Montejo-Ráez, “Sentiment analysis in Twitter,” Natural Language Engineering, vol. 20, no. 01, pp. 128, Jan. 2014.

[11] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, “Sentiment Analysis of Twitter Data,” in Proceedings of the Workshop on Languages in Social Media, Stroudsburg, PA, USA, 2011, pp. 3038.

[12] E. Kouloumpis, T. Wilson, and J. Moore, “Twitter Sentiment Analysis: The Good the Bad and the OMG!,” in Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[13] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, “Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach,” in Proceedings of the 20th ACM International Conference on Information and Knowledge Management, New York, NY, USA, 2011, pp. 10311040.

[14] J. Villazon-Terrazas, S. Aparicio, and G. Alvarez, “Study on Twitter as a Complex Network,” in The Third International Conference on Digital Enterprise and Information Systems (DEIS2015), 2015, p. 54.

[15] S. Aparicio, J. Villazón-Terrazas, and G. Álvarez, “A Model for Scale-Free Networks: Application to Twitter,” Entropy, vol. 17, no. 8, pp. 58485867, Aug. 2015.

[16] S. Beliga, A. Meštrović and S. Martinčić-Ipšić, “Toward selectivity based keyword extraction for Croatian news,” In Proceedings of the Workshop on Surfacing the Deep and the Social Web, CEUR, Vol. 1301, Italy, pp. 1-14, 2014.

[17] D. Margan, A. Meštrović, and S. Martinčić-Ipšić, “Complex networks measures for differentiation between normal and shuffled Croatian texts,” In Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, Croatia, IEEE, pp. 1819-1823, 2014.

[18] S. Šišović, S. Martinčić-Ipšić, and A. Meštrović, “Comparison of the language networks from literature and blogs,” In Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, IEEE, Croatia, pp. 1824-1829, 2014.

[19] L. Lu and T. Zhou, “Role of Weak Ties in Link Prediction of Complex Networks,” arXiv:0907.1728 [cs], Jul. 2009.

[20] H. R. de Sa and R. B. C. Prudencio, “Supervised link prediction in weighted networks,” in The 2011 International Joint Conference on Neural Networks (IJCNN), 2011, pp. 22812288.

[21] Y. Yang, R. N. Lichtenwalter, and N. V. Chawla, “Evaluating Link Prediction Methods,” Knowledge and Information Systems, vol. 45, no. 3, pp. 751782, Dec. 2015.

[22] S. Baccianella, A. Esuli, and F. Sebastiani, “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.”, In: LREC, 2010, pp. 2200-2204.

[23] “GET search/tweets — Twitter Developers.” [Online]. Available: https://dev.twitter.com/rest/reference/get/search/tweets. [Accessed: 21-Feb-2016].

[24] D. A. Schult and P. Swart, “Exploring network structure, dynamics, and function using NetworkX,” in Proceedings of the 7th Python in Science Conferences (SciPy 2008), 2008, vol. 2008, pp. 1116.

[25] D. Margan and A. Meštrović, “LaNCoA: A Python Toolkit for Language Networks Construction and Analysis,” in Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on, 2015, pp. 16281633.