

Analyzing Forum Members by Their Comments*

Bianka Nagy, Attila Kiss

ELTE Eötvös Loránd University, Faculty of Informatics, Budapest, Hungary

nmjpc1@inf.elte.hu

kiss@inf.elte.hu

Abstract

In the paper we present a forum analysis from the messages of the users in Hungarian forum topics. We represent the life of a forum topic, and the peoples inside the forum. We make categories and analyze the users from their word usage. For a user, we want to represent them with their common words, their most important sentences, plus we can compare them to another user or to the average of the topic behaviors. The paper analyzes the differences from the average word usage and the average behaviors in the topic and represent it in given time, and in a timeline. We can represent a forum topic's life.

Keywords: social network analysis, centrality, social influence, natural language processing, outlier, hidden conversation

MSC: 91D30, 68U15, 62H30

1. Introduction

Forums and social networking sites contain a huge amount of text information about members. Analyzing these texts is also important from a practical point of view, as you can explore the impact that events and products can have on different user groups [1]. Analyzes can also help influence individuals, as seen in politics and corporate campaigns that try to get customers to buy new products and services. Influence is now being done by bots, programs that, for some purpose, comment on messages that try to influence a community's beliefs, thereby hacking natural communication [7]. It is therefore an important task to find users who act and communicate in an unnatural way [8] [9].

*The project was supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002).

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Of course, there are natural language processing research, open systems [6] that perform tokenization / lemmatization / POS tagging, and other semantic analyzes, but these can be used primarily for English language analyzes, not for Hungarian. Comprehensive free tools are not currently available in Hungarian, so the English language solutions had to be redesigned and implemented so that they could be applied to Hungarian forums as well.

We note that there would be another possibility to automatically translate Hungarian texts into English with google translate or another free translator and perform the linguistic analysis, emotional analysis, classification on the resulting raw English translation in English, but we found this method based on our previous measurements. that the error would be too large, so we discarded it.

In this publication we choose the website <https://prohardver.hu> for our research, because we want a Hungarian webpage which is big enough, so we can get enough data to our calculations. On this site there are more than 100 forum topics (it is the largest Hungarian website on Informatics). In this research, our calculations and methods can be used in any of the topics and in the topic's every pages. This is a real time calculation, that's why in long term we can get more precise data. We will publish our code also, so if anybody wants to reuse or get data to their research, they only needs to change the url in the code. Our developed system can be easily applied to other websites. If somebody wants to use it in a different website then the data collecting methods need to be changed, because every website's html code is different, that is why the forum messages will be in different context in that website's html, then our method still can be usable to the new data.

For the data collecting methods, we use Python language with “BeautifulSoup version 4” which is a web scraping function to collect data from any webpages. The bigger webpages use protection against the bots and web scraping methods. In this case our or any other methods are not usable on those pages. However, before these protections, there are lot of older data collections available online from that pages, which are downloadable.

2. Method of data collecting

To our research and to get the statistics we use Python language, and we concentrate on Hungarian language in websites, but the methods are based on purely mathematics, that is why they can be used in any other language. In Hungarian language we use a lot of conjunction words, definite and indefinite articles, and pronouns. We call the base set ‘alap’ which contains these words. In the research we will not count the words which are the members of the ‘alap’ set so we use corrected statistics calculations, because these words do not contain any help or information to analyse the users.

Conjunction words: “és, vagy, de, …” which mean “and, or, but, …”

(in)definite articles: “a, az” which mean “a, an/the”

pronouns: “én, te, ő, mi, ti, ők” which mean “I, you, she/he/it, we, you, they”

We also do not count the special characters like: “., ;!?” because if we use them, then for example, the last words in sentences may be different words.

apple ≠ apple. ≠ apple! ≠ ...

After the data collection, first we transform the data to our format. We delete the words from our starting dataset, which are in the ‘alap’ set, search for the different words and count them. After that remove the special characters and modify the characters to lower case.

Since the Hungarian language is an agglutinative language, we consider the words as different even if they have the same root.

At the data collecting methods we do not only access the data from the messages, but we also store the information such as who writes the comment (user name, status in the web page), when she writes it, if it is a reaction to someone then we collect the user names also.

We need this information because we want to create time-line statistics with the dates of the messages. With the user status, we search for correlations between the status and the word usage, word frequency. Based on the data of reactions to each other, we want to represent the users in graphs and evaluate the centrality measures of the users to find how much a given user affects the community.

3. Word usage statistics

First, we make statistics ‘between two users’ and ‘between a user and the whole data set’ for this purpose we create two data:

One is the set of the different words that the user uses, and the second is the frequencies to these words. Creating these data to the analysed peoples and to the ‘whole data set’ and counting the number of words we get the estimated probability to the words. For the calculations we also create files from the 100 most used words, and the frequencies to each of them.

Using these files we can make some comparisons (in these statistics we always write X. and Y. comment writers, not with specific user names, that’s why our calculations and graphs are dynamically usable.):

We determine which are the words only used by the first person:

Similarly, we can see the intersection of the two sets, i.e. which are the words used by both users. After that we calculate the difference of the frequencies. In the graphics, if the y parameter is high, then the first user used them more, if it is low, then the second user used them more.

3.1. Sock-puppet account

We defined the user type ‘sock-puppet account’ by comparing the statistics of a given user to an another user. A user can create multiple instances of these cheaters,

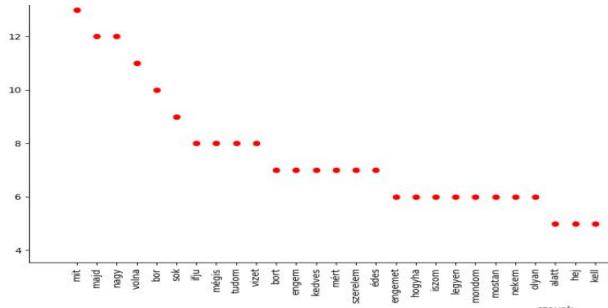


Figure 1: user1 - user2

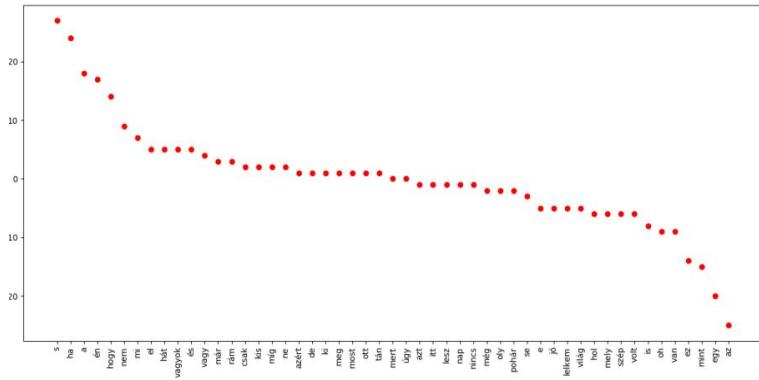


Figure 2: user1 ∩ user2

which we can detect by analyzing the statistics, since they have the same purpose and thus communicate in a very similar way with other users. We guess if they got highly similar statistics and their intersection diagram is close to flat, then it is possibly a second account of the user. This account is called a ‘sock-puppet’ if it mostly reacted to the main account. Such users are hiding, using their primary account only to send messages to the community under the alias they have created so that they influence the opinions of others anonymously [10] [2]. If we find such replicated users, then we may consider them as ‘sock-puppet’ account of the same user. Of course, this suspicion can only be considered to be certain through further investigations.

3.2. Power of the sentences

By examining the statistics, we can determine which of the most common words, phrases and sentences a user has. For sentences, the frequency of words is added up and divided by the number of words in the sentence. This formula maximizing sentence is considered to be the most representative sentence for the user.

4. Users interaction

In this section we create statistics from the reactions between the users:

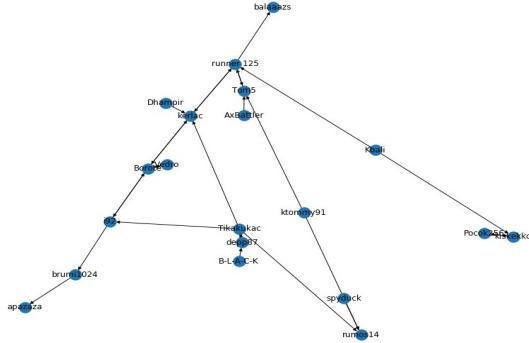


Figure 3: who responds to whom

In this graph the arrows represent who responds to whom (or they respond to each other).

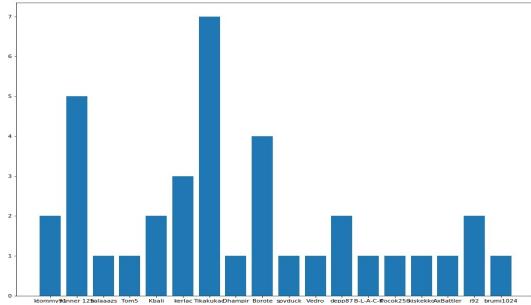


Figure 4: who got the most reactions

The community, the forum, the topic is characterized by the users who elicit the most reactions from the community members, i.e. the ones who comment the most and / or who receive the most messages and comments. Such users spin, keep the forums alive. Such users are, in a sense, central elements of the network, the graph. There are several degrees of centrality, such as closeness, which are used to find the most important members.

We calculate the closeness centrality to each user, and with those values we create a closeness graph (which does not include the peoples with 0 centrality).

In this graph the arrows represent who reacted to whom (or they reacted to each other). The size of the circles is proportional to the value of centrality. We colour the graph as follows: the users with the biggest centrality values got red color, the average got blue color, and the other accounts got mint color.

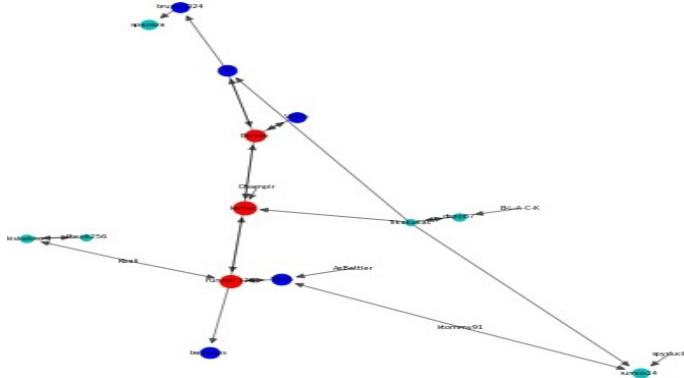


Figure 5: Closeness centrality

In our analysed forum pages, we did not find strong correlation between the most reacted users and the biggest closeness centralities. For example, the user with the highest centrality value is just the fourth most reacted person and the most reacted user got one of the smallest values. The largest centrality was 0,157 in the whole data and the user with the most reaction got 0,07. With the reactions to each other we can find the people whose activity keep the topic rolling, and the users who create their own “talk” in the topic. Peoples who ‘keep the topic rolling’ have the most centrality in the graph. In another way, if we delete those nodes from the graph, the graph falls into two disjoint parts.

4.1. Suspicious users

Today, there is a lot to read about how terrorist groups use social networks to hide their own messages. If our goal is to find out that there is such hidden communication in the forum, then we need to look for a smaller, suspicious group whose members communicate with each other with abnormal language characteristics and, moreover, only respond to each other’s messages [4]. The most suspicious users are who had many comments, but who only react to 1-4 different persons and these people also react to these small group of people. Like they have a disjoint part of the graph. If we find a group whose members talk to each other differently from the forum topic, using the forum only as a means of communication, then according to our purpose, we can either warn them to discuss elsewhere, or keep an eye on their conversations to uncover secret terrorist messages.

The first graph is a full graph, and the second is a subgraph of an another graph, which represents a suspicious group, whose members have their own talk which does not belong to the topic. The first is suspicious because one user has a mint color, but the two people only communicate with each other, and they are not the real part of the topic conversation. The second is suspicious, because they had higher colors like blue, but they only reacted in the small group, but got many

reactions.

If we find someone like this, we will analyse them as the previous part of the research comparing their word usage to the whole data set.

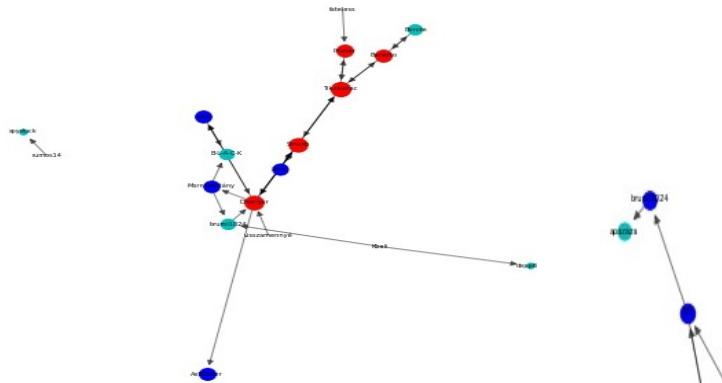


Figure 6

5. Centrality and Entropy

With the help of the reactions we also examined whether there is any relation between the centrality and the word distribution of the individuals and the entropy of the distribution.

5.1. Centrality

Four types of centrality were used to all the users [5]:

- Betweenness centrality
 - Closeness centrality
 - Eigenvector centrality
 - Degree centrality

For each type of centrality we search for the users with the maximum and minimum centrality.

In the graph we see the one of the centralities. Lighter (red) coloration indicates greater centrality and darker (blue) coloration indicates lower centrality.

5.2. Entropy

For analysing users, we compute the entropy to all the messages, to get a better representation for the user behaviours. In every forum topic there are many frequently used words, which are the topic's speciality, and we will use this feature of the topics. We want to compare the user to the all data set's entropy. Therefore we

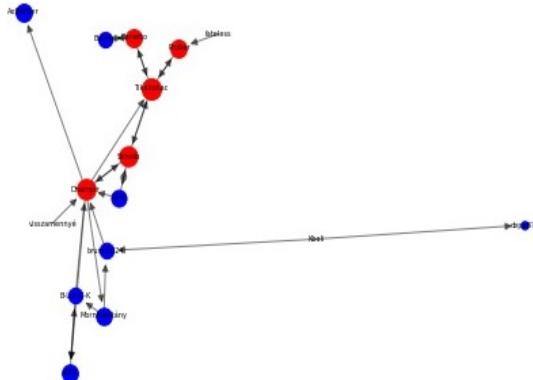


Figure 7: Subgraph

calculate two different entropy analyzes. Firstly, the entropy to every ten messages of the user, where the messages follow each other in time and secondly determine it in time. Which means, we recalculate the entropy in every new ten comments, but we calculate with the all messages until that time.

For the comparison we also need to get the entropy of the all data set. For that we applied the two different type of entropy calculations. With these pre-made calculations we can compare the user to the all dataset.

For illustrating better their relation, we create two more types of analysis, which are similar to the previous ones. In the calculations of the user entropy, we do the same as previously. But for the all data set, we recompute the previously defined two entropy types, to one of the subset of the all data set. The subset is where we count with the ‘all messages - user messages’. We take this subset, because if the given users are counted in the whole group, then we will compare them to themselves. With this improvement we get more accurate representation of the real world.

Our goal is to show the entropy difference alteration in time: If the difference gets less in time, then the user starts to adopt to the word usage of the topic, and will be a long term user in the forum. If the difference starts to get bigger in time, then the user starts their ‘secession life’ in the forum. Which means the user got high possibility to exit the forum.

Predicting customer churning is a very important practical, economic goal for service providers. If the individual's vocabulary moves away from the common language or does not follow the linguistic changes of others, then he or she may want to leave the group. In this case, the service providers may try to keep it by offering promotional offers. Just like in direct marketing, here we can save some of our expenses by not giving everyone a discount, just a smaller group.

Our last entropy analysis we made, where we compare the all date set to the ‘all messages - user messages’ subset.

The goal with that is to show, if the user changed the forum word usage by

bringing new words to the topic. We can see which words after that are frequently used in the topic by other users. If the two type's entropy difference start to get less, that means, the forum also starts to adopt to the new user and the new words start to spread in the topic. If the difference is constant or starts to get bigger, then the user did not adopt to the topic usage.

5.3. Relation between centrality and entropy

In the last part we compare the 4 types of centralities to the entropies: In the previous parts we got 4 people, whose got the highest centrality in each type, and we also found the peoples with the least centralities in each type.

To the analysis, we repeat the entropy calculations to these peoples, to analyse if there is any correlation between the centralities and entropy.

In our dataset we found that users with highest centralities are not related to the entropies. In the forum, if somebody talks about other things than the topic theme, or use not related words, they likely get more attention to their comments. For example, where the other users ask them to talk in private. We observed that most of the users with the lower centralities have very similar entropies to the all data set, because they mostly just repeat the other user messages or did not add new information to the communication.

5.4. Conclusion and future works

This paper describes the tests that can be performed on individuals and average behaviors in any forum using purely linguistic statistics. We have also tried different methods for Hungarian language forums. English-language software is available for this, but our methods also work for the Hungarian language. These can be used to determine the members of the presumption, the change of the group, the attitude of the individuals towards the group. In addition, it is predicted that individuals will drop out or that terrorists will hide messages within the forum. In the future, we would like to define more user types, explore more analyzes and integrate them into a framework that can be used to analyze any Hungarian forum. As this is just the beginning of a larger, comprehensive research, we have designed and implemented the tools for the time being, and this will be followed by running them on several large data sets and drawing conclusions and useful information from the obtained measurements and analyzes.

References

- [1] AGGARWAL, C. C., ZHAI, C. A Survey of Text Classification Algorithms. *Mining Text Data* 163–222. 2012.
- [2] ATANASOV, A., MORALES, G. D. F., & NAKOV, P. Predicting the Role of Political Trolls in Social Media. arXiv preprint arXiv:1910.02001. 2019.

- [3] ZHILI CHEN Effective-Linguistic-Steganography-Detection In 2008 International Conference on MultiMedia and Information Technology 217-220. IEEE, 2018.
- [4] ZHILI CHEN, LIUSHENG HUANG, HAIBO MIAO, WEI YANG, PENG MENG Steganalysis against substitution-based linguistic steganography based on context clusters. Computers & Electrical Engineering, 37(6), 1071-1081. 2011.
- [5] KLEIN, A., AHLF, H., & SHARMA, V. Social activity and structural centrality in online social networks. Telematics and Informatics, 32(2), 321-332. 2015.
- [6] MANNING, C. D., SURDEANU, M., BAUER, J., FINKEL, J. R., BETHARD, S., & McCLOSKY, D. The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations 55-60. 2014.
- [7] TORRES-SORIANO, M. R. The dynamics of the creation, evolution, and disappearance of terrorist Internet forums. International Journal of Conflict and Violence (IJCV), 7(1), 164-178. 2013.
- [8] PETER VOROS, PETER HUBA, ATTILA KISS Steganography and Cryptography for User Data in Calendars In Asian Conference on Intelligent Information and Database Systems 241-252. Springer, Cham, 2019.
- [9] LINGYUN XIANG, XINGMING SUN, GANG LUO, BIN XIA Linguistic steganalysis using the features derived from synonym frequency. Multimedia tools and applications, 71(3), 1893-1911. 2014.
- [10] ZHENG, X., LAI, Y. M., CHOW, K. P., HUI, L. C., & YIU, S. M. Sockpuppet detection in online discussion forums. In 2011 Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing 374-377. 2011