Application of Text Mining to Analysis of Social Groups in Blogosphere

Bogdan Gliwa, Anna Zygmunt, Jarosław Koźlak, and Krzysztof Cetnarowicz

Abstract. The paper concerns analysis of social groups in blogosphere using text mining methods to discover additional knowledge about groups and users. Two methods to distinguish messages (the first one - between messages from main and secondary thread, the second one - between facts and opinions) in blogosphere were proposed and their quality was assessed on manually annotated dataset. Both tasks are very important and proposed methods deal with them in a fully automatic way. The results were obtained on real-world data from Polish blogosphere.

Keywords: social network analysis, group topics, subjectivity detection, blogosphere, text mining.

1 Introduction

Nowadays, more and more elements of our everyday life are transferred to the virtual reality, especially communication with other people: we participate in discussions on forums, comment on blogs, chat and express our opinions using social media. Many companies are interested in automatic way of extraction information from users messages left in forums, blogs etc.

For analysis of user activity in social media, the application of methods of social network analysis is very popular. Discussions between people in blogs or forums can be modeled as social network and in such a network there are formed some groups of users that are more strongly connected between themselves than with the rest of network. This approach lets us analyse groups of people at different angles. Analysing groups in the context of written messages is the main goal of the paper.

Bogdan Gliwa · Anna Zygmunt · Jarosław Koźlak · Krzysztof Cetnarowicz
AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Kraków, Poland
e-mail: {bgliwa,azygmunt,kozlak,cetnar}@agh.edu.pl

P. Contucci et al. (eds.), *Complex Networks V*,
Studies in Computational Intelligence 549,
DOI: 10.1007/978-3-319-05401-8_27,

2 Related Work

Many methods for groups detection were proposed [3], [8]. They can find overlapping or non-overlapping groups, changing in time or not, etc. One of the most popular representative of algorithms finding overlapping groups is CPM method (Clique Percolation Method) [10].

In different methods regarding dynamics of groups, many events in groups lifecycle (also called groups evolution) were proposed. Palla et al. [11] described some events that can be identified during groups evolution: growth, contraction, merging, splitting, birth and death.

Topic Modeling [9] is a statistical technique that detects abstract "topics" existing in a collection of documents. "Topic" can be defined as a set of words that tend to co-occur in multiple documents, and, therefore, they are expected to have similar semantics. One of the biggest advantage of this method is that similar texts can be discovered even if they use different vocabulary, which is hard to achieve using other methods. Latent Dirichlet Allocation (LDA) [1] is one of the most popular methods in topic modeling and aims to reduce dimensionality by grouping words with similar semantics together.

In literature most applications of Text Mining in the field of Social Network Analysis regard some specific cases [2]. In [7] the authors showed usefulness of topic modeling to analysis of groups dynamics in social networks in blogosphere.

3 Analysis of Text Messages and Groups in Blogs

This section provides the concept of methods used to further analysis. In 3.1 and 3.2 we describe methods used to find out whether a message is a fact or an opinion and whether given message relates to the main topic of discussion thread (called *in the main thread*) or not (called *in the secondary thread*). Next, we depict methods used to analyse groups in dynamic social network.

3.1 Finding Messages in the Main and the Secondary Thread for Comments

Distinction between messages in the main and the secondary thread is based on topics uncovered by LDA method (and manually labelled) for given comment and post in analysed conversation thread. Additionally, one from LDA topics was labelled as *various* (it was hard to annotate as one particular topic), so in this method during comparison of topics in the case when post has topic *various* and comments has topic *various*, we assumed that they are different ones.

Let us define c as analysed comment, $post_c$ - post in thread where comment c was written and $topics(m)$ as topics for message m. Method is quite simple

and can be described in the following way (MS is a function assigning label for a comment whether such comment is in the main thread or in the secondary one):

$$MS(c, post_c) = \begin{cases} \text{main}, & \text{if } topics(c) = \emptyset \vee |topics(c) \cap topics(post_c)| > 0, \\ \text{secondary}, & \text{otherwise.} \end{cases} \tag{1}$$

3.2 Finding Facts and Opinions in Comments

To distinguish messages containing only facts from messages containing opinions we also employed detection of topics (by means of LDA method) for a comment and a post in the same discussion thread.

Method consists of 2 steps:

Step 1. Analysing content of message to find out some symptoms of opinions, which we defined as occurence one of opinion words (manually defined, about 20 words such as (translated to English) *think*, *convince*, *respect*...), containing exclamation sign or have one topic from annotated as *opinions* and *critics* (LDA method uncovered such clusters). If any of above mentioned conditions are fulfilled, then message is annotated as *opinion* and second step is omitted.

Step 2. Analysing similarity of topics for given comment and post in thread. If they are similar i.e. $|topics(c) \cap topics(post_c)| > 0$, then we assumed that message c is a fact. When there are no topics for given post and comment then such comment is treated as *opinion*, but when there are no topics for post and comment has some topics – the comment is labelled as *fact*. Otherwise, we marked comment as *opinion*.

Above conditions can be expressed also as an assumption that when topic of comment and post matches then people discuss facts (except the case when we found some symptoms of opinions) and when they introduce new topic, then they express their opinions (or attacks personally between themselves).

3.3 Groups in Dynamic Social Network

Data from whole time range is divided into series of time slots and each time slot contains static snapshot of network from defined period of time.

In each slot we used the comments model for building graph, introduced by us in [4] - the users are nodes and relations between them are built in the following way: from user who wrote the comment to the user who was commented on (if the user whose comment was commented on is not explicitly

referenced by using @ and name of author of comment in title of comment by commenting author, the target of the relation is the author of post).

In every static snapshot of social network groups were detected. Groups from adjacent time slots can be matched to find continuation of groups from different periods of time. For this goal, the SGCI (Stable Group Changes Identification) [5] method was applied. SGCI algorithm consists of the following steps: identification of short-lived groups in each time slot; identification of groups continuation, separation of the stable groups (lasting for a certain time interval) and the identification of types of group changes (transitions between stable groups).

Identification of group continuation is conducted using modified Jaccard measure with minimal threshold equals 0.5 (A and B are examined groups from the consecutive time slots):

$$MJ(A,B) = \begin{cases} 0, & \text{if } A = \emptyset \vee B = \emptyset, \\ max(\frac{|A \cap B|}{|A|}, \frac{|A \cap B|}{|B|}), & \text{otherwise.} \end{cases} \quad (2)$$

and ratio of group size with maximal threshold equals 50:

$$ds(A,B) = max(\frac{|A|}{|B|}, \frac{|B|}{|A|}). \quad (3)$$

4 Results

4.1 Description of Experiments

The experiments were conducted on data set containing data from the portal *salon24*[1]. The data set consists of 31 750 users (12 750 of them have their own blog), 380 700 posts and 5 703 140 comments within the period 1.01.2008 - 1.07.2013. The analysed period was divided into time slots, each lasting 7 days and neighboring slots overlap each other by 4 days. In the examined period there are 503 time slots.

For group detection we used CPM [10] method (directed version of CPM from CFinder[2]).

Topic for messages were extracted using LDA algorithm from mallet tool[3]. The method discovered 350 clusters of topics, which were manually annotated and some of them were manually joined. After this operation the number of clusters shortened to 67.

[1] Mainly focused towards politics, `www.salon24.pl`
[2] `http://www.cfinder.org/`
[3] `http://mallet.cs.umass.edu/`

4.2 Testing Quality of the Methods for Detection of Opinions, Facts, Messages in the Main and in the Second Thread

To assess quality of proposed methods we prepared set of discussion threads chosen in a random way from threads having at least 10 comments inside. Test dataset consists of 30 threads and 833 comments. Each comment was manually annotated whether contains only facts or contains opinions (possibly mixed with facts), and whether is related to the main topic in discussion thread or maybe is related to other one (e.g. personal messages between bloggers are annotated as messages not related to the main topic). The shortest thread has 11 comments and the longest one – 69 comments.

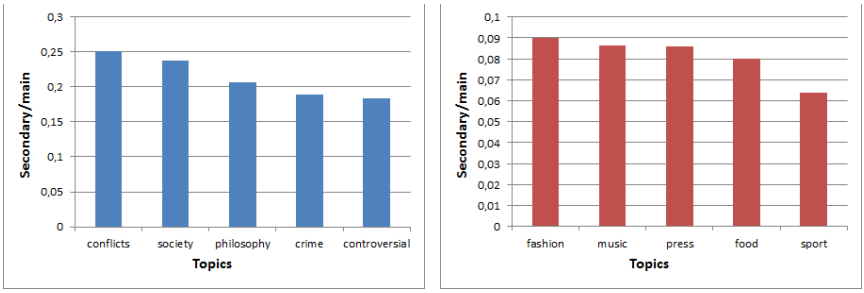
We evaluated F-measure (the harmonic mean of precision and recall) for each thread expressing quality of both methods. The results are presented in table 1. One can observe that results are quite good for both tasks. The lowest values (below or equal 0.5) in task determining whether message is fact or opinion are for 3 cases and in task assessing the fact that message belongs to the main topic of discussion thread or not are for 2 cases.

Table 1 Number of cases with given F-measure for methods detecting facts/opinions and main/secondary thread on manually annotated set of threads

range	main/secondary	opinion/fact
0-0.5	2	3
0.51-0.6	0	2
0.61-0.7	5	7
0.71-0.8	11	10
0.81-0.9	10	6
0.91-1	2	2

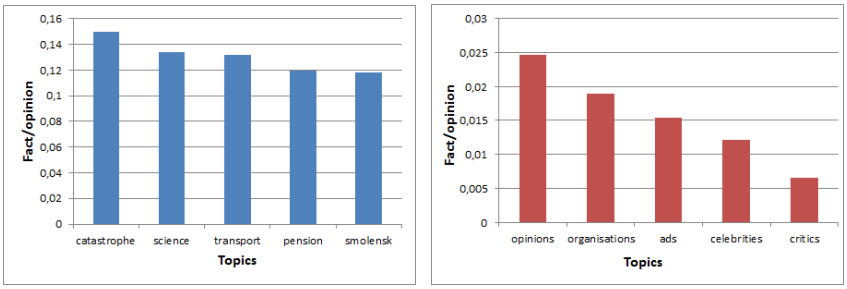
4.3 Discussion Threads with Messages Related to the Main and the Secondary Topic

We analysed the impact of the discussion thread topic on tendency to moving discussion to other topics. In figure 1a we can see topics of discussion threads, in which users most frequently discussed also other topics. Figure shows for topics the number of messages in the secondary thread divided by the number of messages in the main thread. We can observe that people often change topic of discussion in e.g. discussion threads with controversial content (like *abortion*) or in *philosophical* threads. Opposite situation is described in figure 1b – there are topics, in which users rarely change the subject of discussion. Among them we can find such topics as *sport* and *music*.



(a) Max secondary part. (b) Min secondary part.

Fig. 1 Top 5 topics of discussion threads with max and min secondary part



(a) Max fact part. (b) Min fact part.

Fig. 2 Top 5 topics of thread with max and min fact part

4.4 Discussion Threads with Facts and Opinions

We conducted similar analysis – we tried to find out the topics where users mostly express their opinions and where they discuss also about facts. Fig. 2a presents topics with the highest number of facts in comments in discussion threads. It is not surprise that we can find there *science* topic. On the other hand, fig. 2b shows topics with the lowest number of facts in comments in discussion threads. Among such topics, one can notice topics related with *critics*, *celebrities* and *opinions*.

4.5 Topics in Groups

In fig. 3 the most popular topics in groups with different size are shown. We can notice that the most popular topics in groups are *various* and *politics*. *Science* topic is dominated by groups of medium size (11-50 members). Another interesting observation is that the topic of *religion* mostly occurs in small and

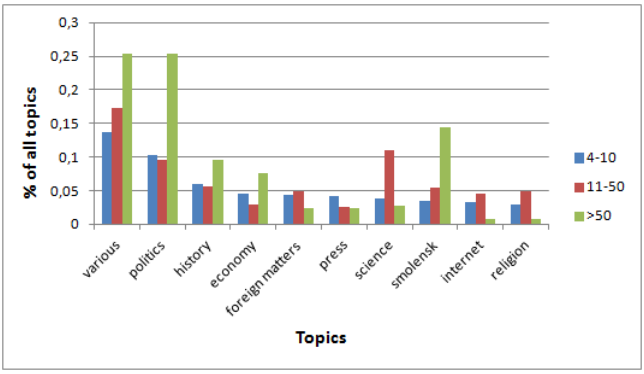


Fig. 3 Most popular topics in groups with different size

medium size groups. One can see also *smolensk* topic which is very popular. This topic concerns event of Polish President airplane crash in Smolensk and some other events related with investigation of this catastrophe.

4.6 Groups Formed Around Messages Deviating from the Main Topic

In fig. 4 we can observe what part of a group consitutes the part related with the main thread of discussion or, in other words, in how many groups the people during their discussions are stuck to the main topic. One can notice that for most groups the fraction of the main threads in discussions established inside them is very high, which means that people form groups to discuss particular topics. The highest variety can be noted for small groups (with 4-10 members) and it decreases when size of groups increases.

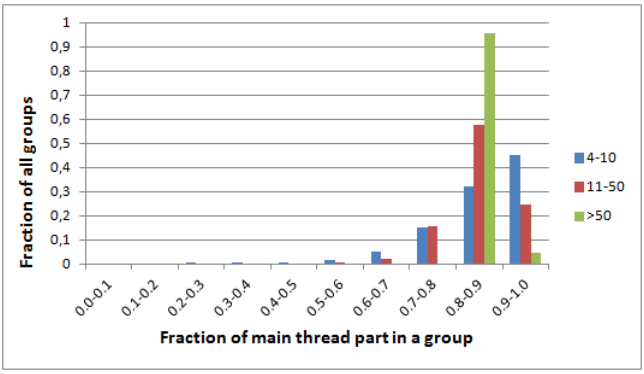


Fig. 4 Fraction of the main thread part in groups with different size

4.7 Groups Formed around Facts and Opinions

Fig. 5 presents how many groups with different size talk mostly about opinions. As we could anticipate, in blogs people in groups mostly share their opinions with others. However, we can notice that there are some small groups that talk almost completely about facts without expressing their own opinions. Moreover, in large groups (with more than 50 members) for most of them the part related with facts is quite high (about 20%) which is different from small and medium size groups.

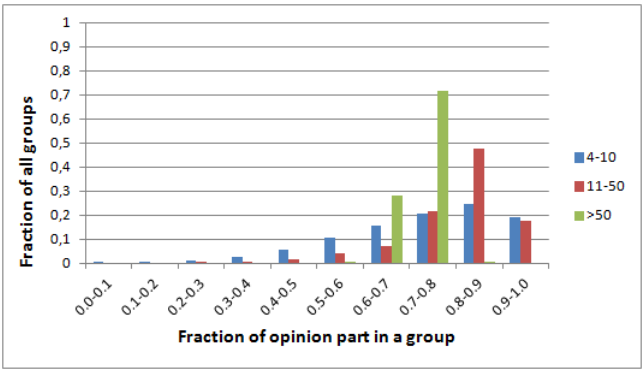


Fig. 5 Fraction of opinion part in groups with different size

5 Conclusion

In this paper we proposed 2 methods – the first for the distinction of messages in the main and in the secondary thread and the second one – detection opinions and facts, both in blogosphere. We assessed quality of these methods on manually annotated subset of whole analysed data and achieved results seem promising. Moreover, we analysed groups in social network under those angles. The obtained results allow us to better understand structure of groups.

Future work may follow in several directions. Firstly, there is a place to improve these methods (e.g. maybe some assumptions are not well suited for all types of topics). The second is to analyse roles of users in groups who e.g. change the main discussed topic or express mostly facts. For this purpose we want to employ our method of detecting roles of users [6]. Another interesting direction is to conduct experiments on other datasets including datasets in English language.

Acknowledgements. The research leading to these results has received funding from the dean grant.

References

1. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Aggarwal, C., Wang, H.: Text mining in social networks. In: Aggarwal, C. (ed.) Social Network Data Analytics, pp. 353–378. Springer (2011)
3. Fortunato, S.: Community detection in graphs. In: Phys. Rep., ch. 486 (2010)
4. Gliwa, B., Koźlak, J., Zygmunt, A., Cetnarowicz, K.: Models of social groups in blogosphere based on information about comment addressees and sentiments. In: Aberer, K., Flache, A., Jager, W., Liu, L., Tang, J., Guéret, C. (eds.) SocInfo 2012. LNCS, vol. 7710, pp. 475–488. Springer, Heidelberg (2012)
5. Gliwa, B., Saganowski, S., Zygmunt, A., Bródka, P., Kazienko, P., Kozlak, J.: Identification of group changes in blogosphere. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012, Istanbul, Turkey, August 26-29. IEEE Computer Society (2012)
6. Gliwa, B., Zygmunt, A., Kozlak, J.: Analysis of roles and groups in blogosphere. In: Burduk, R., Jackowski, K., Kurzynski, M., Wozniak, M., Zolnierek, A. (eds.) CORES 2013. AISC, vol. 226, pp. 299–308. Springer, Heidelberg (2013)
7. Gliwa, B., Zygmunt, A., Podgórski, S.: Incorporating text analysis into evolution of social groups in blogosphere. In: Federated Conf. on Computer Science and Information Systems, FedCSIS 2013, Krakow, Poland, September 8-11 (2013)
8. Greene, D., Doyle, D., Cunningham, P.: Tracking the evolution of communities in dynamic social networks. In: Proc. International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2010. IEEE (2010)
9. Huang, Y.: Support vector machines for text categorization based on latent semantic indexing. Tech. rep., Electrical and Computer Engineering Department, The Johns Hopkins University (2003)
10. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814–818 (2005)
11. Palla, G., Barabási, A.L., Vicsek, T., Hungary, B.: Quantifying social group evolution. Nature 446, 2007 (2007)