Application of the principal component analysis to detect semantic differences during the content analysis of social networks

I A Rytsarev[1,2], D D Kozlov[1], N S Kravtsova[1], A V Kupriyanov[1,2], K S Liseckiy[1], S K Liseckiy[1], R A Paringer[1,2] and N Yu Samykina[1]

[1]Samara National Research University, Moskovskoe shosse 34, Samara, Russia, 443086
[2]Image Processing Systems Institute - Branch of the Federal Scientific Research Centre "Crystallography and Photonics" of Russian Academy of Sciences, Molodogvardeyskaya str. 151, Samara, Russia, 443001

Abstract. In this paper, we propose an approach to semantic differences detection in texts presented in the form of frequency dictionaries. The original text data has been obtained by collecting records on various online communities. We have implemented a specialized software module that allows us to analyze and download both posts and comments from the social network VK's open communities. To build our frequency dictionary, we have developed an algorithm that takes into account the peculiarities of the data collected from social networks. In the article, we propose an approach based on the use of methods reducing the dimension of feature spaces to identify keywords based on the analysis of their frequency of usage. The algorithm we present uses the principal component analysis technique. As a result, we have shown that by using the coefficients of the obtained linear transformation, it is possible to estimate the importance of words. With the help of these estimates, we were able to identify not only key words, but also semantic differences in social networks communities. The proposed approach can also be used to form metrics and calculate the social distance between Internet communities.

1. Introduction

Development of the newest informational and industrial technologies, dynamics of its wide implementation influence all the personal and public activities in everyday life. Nowadays we can see rapid transformation of ways of interaction between different persons, groups and organizations.

Virtual network communities can be described as social groups in terms of social psychology (there are common interests and activities in such groups, awareness of membership, sense of collective "we", and possibility of direct personal contact between members). At the same time, these communities have specific features because of its virtual, digital nature. Participation in most of virtual communities is not bounded by age, gender, social status or territory. These communities have free membership (there are no entrance fees, membership cards, formal obligations). Members of the communities are able to enter any of it or leave it at any moment as they wish. It allows people to choose a content according to their interests and motives. Besides, in most cases it is not necessary to be a member of virtual community to see its content.

Communication via internet is so wide-scaled that it requires development of new methods to study and analyze it [1] point out the following features of data from social networking services (SNS) that create certain methodological difficulties:

- subjective selectivity and social desirability of the content published by users;
- possibility to delete the content previously published by users;
- existence of fake accounts and bots producing unpersonal, unified, automatic content;
- features of SNS itself, its possibilities and interface implementations that bound and induce certain activities of users at the same time, thus creating a discrepancy between users' behavior in virtual and real life communities;
- autocompleting of forms and choices "by default", often made without awareness, probably not consistent with real choice or opinion of users;
- ethical issues concerning the use of personal information and the right of a user to delete it permanently.

Nevertheless, ability to quantify and classify digital footprints of nearly endless quantity, high speed of information processing (even online), its verifiability, transparency, and low financial costs are create new opportunities and benefits. However, we have to invent new methods to gather and analyze such data to avoid difficulties mentioned above [2].

Content users publish in SNS, and digital footprints, allow revealing and taking into account personal psychological characteristics (values, emotions, mood, self-regulation strategies and motives) essential to economic, social and political decision making [3]. Moreover, SNS make it possible to monitor these characteristics and its dynamics in real-time. There have been already developed trading algorithms that use as input variables actual social mood calculated upon recent posts in SNS (namely, Twitter, LifeJournal or Facebook) [4, 5], and these algorithms proved to be very effective in a real stock market [6]. Similar psychological characteristics influence not only economic decisions, but political and social decisions and behavior as well [7, 8].

However, prediction of psychological characteristics of SNS users is still much less precise than prediction of social or demographic characteristics. For instance, in one of the largest up-to-date studies by Kosinski et al. [9] social and demographic characteristics (such are, e.x., sex or sexual orientation) of SNS users were predicted with high accuracy, while error rate for personality traits was relatively high. Nevertheless, even such not so precise results can be useful to optimize search requests and personal ads on webpages [10]. If we want to lower the error rate in diagnostics and prognosis, we have to go deeper and analyze digital footprints not as behavior markers only. We have to follow new directions and analyze meaning and semantics of posts being published by users [11], and invent new mathematical and statistical methods to execute such analysis [12, 13].

These studies demonstrate potential benefits of analysis of big data obtained from SNS [14, 15]. However, they are the first steps in a long journey. The models proposed in these and other studies include generalized characteristics of very large samples and communities only [16]. It is necessary to develop more precise models to lower error rates in prognosis and decision making. To do this, we have to take into account subjective semantics and psychological features of particular social communities and groups in SNS. One of the tools to achieve this goal is the «Social Sonar» created at Samara University. It is based on new methods of analysis of group dynamics in SNS. In this case, group dynamics refers to all the processes of social groups' life cycle: its creation, functioning, development, stagnation, involution, and vanishing. The present study propose a new method of modelling semantic similarities and differences between SNS communities. This method allows raising the efficacy and precision of such analysis.

2. Data collection and preprocessing

We selected VK for the analysis of communities the Social network, as it is one of the most popular Russian-speaking one in the Internet. A feature of the social network is that it is available to all and does not have a strictly defined theme. In addition, this social network freely provides the application software interface for writing external applications.

Retrieving data from the social network VK is possible only when using standard tools provided by the developers of the network, it is therefore necessary to access the servers of the social network VK.

To do this, it is needed to create an application on the VK server and get the access keys by following the steps below:

1. In the developer section, follow the instructions in the create applications wizard to create an application.
2. Get the application ID and secret key to quickly connect the application to the servers.
3. In the app settings, give the app permissions to work with communities and records.

The next stage of our work was the development of a software module for data collection. The implementation was carried out in Python, using scripting library for VK. This library is developed by third-party developers and has many methods built on the official program interface's. After installing the module into the system, it is imported and an authorization object is created which stores the login information. All interaction with the social network are then carried out threw the module.

As a matter of understanding, here are some terms used in the upcoming paragraphs:. A record is any text message in the community. A post is a record that carries information about one or several events, most often prompting you to start discussing a topic. Comments are entries below the post that reflects the reaction of a specific user to the post or a comment of another participant.

The stage of collecting records from the wall consists of two procedures: the procedures for collection of posts and the collection of comments.

The procedure for collecting posts initializes a special method of the official application data collection system, which returns a list of posts of a selected user or community and processes its posts. To initialize the method, you must specify a unique user ID or a unique community address from which to download information.

The procedure of collecting comments to the post is carried out the same way: through the initialization of the program interface method's and the process of its answer. The difference is in the method used and its required set of parameters: to receive comments, you must specify a unique identifier not only for the community, but also for the post.

The use of the two procedures described above allows the collection of information from any open VK communities, but there is a software limit on the number of requests per day, which significantly complicates the collection of information. So third-party applications have a daily limit of 2500 requests and may receive no more than 100 records per request.

Also, the properties of records contain the date of publication. To reduce the number of requests, it is possible to limit the number of records received in accordance with its time interval. Comparing the time in the record with the interval of interest, you can get only the records corresponding to the time interval of interest.

3. Algorithm for frequency dictionary compilation

Records are a special information storage structure, containing both text fields (the direct content of the record) and metadata, containing information such as identifiers of the record, its author or time data.

Since the basis of the research carried out in this work is the compilation of a frequency dictionary, the developed algorithms relate primarily to the field containing the text of the records under consideration, but the rest of the information of the records can also potentially be taken into account and used for processing.

After referring to the corresponding field of the record and receiving the data, the resulting text fragment is divided into words. Word, text in this case is called sequence of alphabetic symbols, separated by spaces, numbers or punctuation marks. Since the entries in social networks are full of typos and errors, the spelling check of the data obtained was the necessary step for the correct accounting of the number of words in the preparation of the frequency dictionary. A third-party library was used to perform the spell checking procedure for this work. In parallel, the number of posts and comments is taken into account.

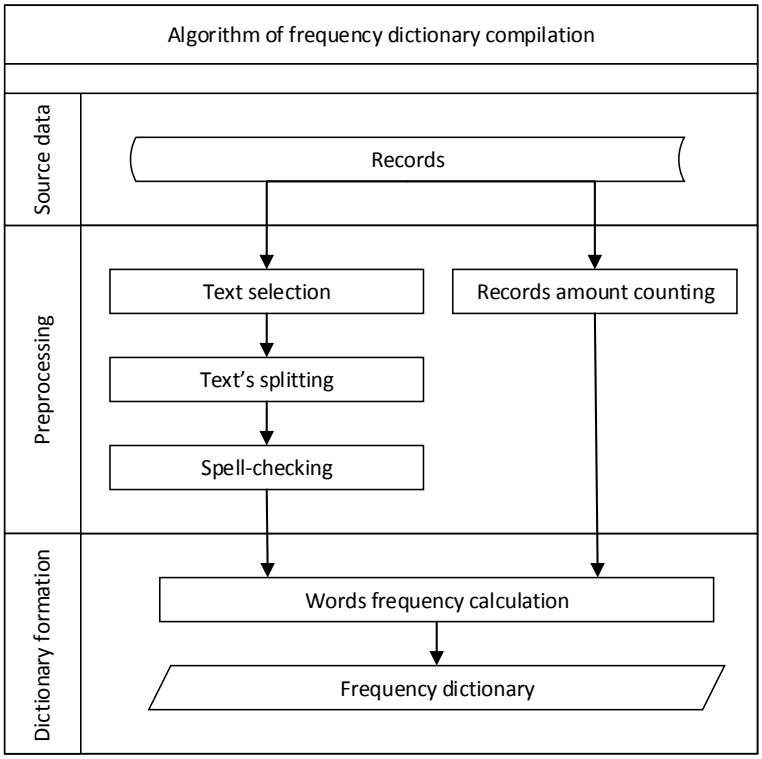

Figure 1. The sheme of algorithm for frequency dictionary compilation.

After performing the described steps, counting the number of occurrences of each unique word w in the whole set S of text data was carried out by the formula:

$$count(w) = \sum_i (w_i \in S)$$

The resulting number of words is divided by the number of entries and thus estimated frequency of use of words in the records – formed frequency dictionary. The scheme of the developed algorithm is presented in Figure 1.

4. Algorithm for semantic differences detection

We developed a matching algorithm for the identification of semantic differences on the basis dictionaries frequency analysis. The concept of this algorithm is to apply existing techniques to reduce the dimensionality of the feature space in order to rank words. The principal component analysis technique [17] was used in this article.

The initial data for the algorithm are frequency dictionaries. Each dictionary can be represented by a vector of words. With two dictionaries, it is possible to form one new attribute that will separate these dictionaries in the best way. A new feature using the principal component analysis is formed by multiplying the values of the initial space by the corresponding vector of coefficients. The values of the vector coefficients used to form a new feature can be used as an estimate of the word's contribution to the formed feature. Thus, it is possible to make a list of words that have made the greatest contribution to the feature's formation. The list of words compiled in this way describes a feature that provides the dictionaries separations and describes their semantic difference.

The first step of the algorithm is data normalization. The values of the frequency of word usage were normalized in order to obtain dictionary values in frequencies in the interval [0;1] according to the following formula:

$$y(x) = \frac{x - x_{min}}{x_{max} - x_{min}},$$

where x_{min} – minimum value among elements in the vector, x_{max} is the maximum value among elements in the vector.

The next feature of the algorithm is the application of the principal component analysis.

Principal component analysis is one of the main ways to reduce the dimension of data with the lesser loss of information. The calculation of the principal components is usually reduced to the calculation of the eigenvectors and eigenvalues of the covariance matrix of the original data. By definition, the covariance of two features X_i and X_j is calculated as follows:

$$cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = E[X_i X_j] - \mu_i \mu_j,$$

where μ_i — - mean of the i sign. The covariance matrix is a symmetrical matrix, where the diagonals represents the feature's dispersions and outside the diagonals are the covariations of the corresponding pairs of features. The Rayleigh relation [18] implies that the maximum variation of the data set will be achieved along the eigenvector of this matrix corresponding to the maximum eigenvalue. Therefore, the main components on which we project the data are simply eigenvectors of the corresponding eigenvalues of the matrix. The values of eigenvector elements are therefore the desired estimated coefficients for the formation of words describing semantic differences. To obtain a visual interpretation of the relative positions of the dictionaries in the resulting semantic space, we need to multiply the frequency of word use vector's with its corresponding eigenvector.

5. Results

To study the performance of the developed algorithms, we selected communities of similar areas of interest. All the selected communities are communities of Samara and Samara region residents and three of the five communities are open communication platforms for students and teachers of Samara's two largest Universities. Due to the existing restriction on data collection within the framework of this study, we introduced an additional criterion for selection of records based on the time of their publication (from January 1 to May 20, 2018. The following communities have been selected for data collection and processing:

I "Overheard in Samara University".
II "Overheard Samara University" (old name: "Overheard in SamSU").
III "Overheard in SamSTU".
IV "Heard Samara".
V "Overheard Samara".

Table 1. Amount of posts and comments.

Record	Community number				
	I	II	III	IV	V
Post	431	841	1486	2060	5918
Comment	2076	1711	4108	28682	257163

Table 2. A fragment of frequency dictionaries.

Words		**Community number**				
Original	**Translation**	**I**	**II**	**III**	**IV**	**V**
он	**he**	0.038	0.020	0.018	0.038	0.043
очень	**very**	0.037	0.032	0.024	0.032	0.018
вам	**you**	0.021	0.029	0.015	0.030	0.022
мы	**we**	0.029	0.019	0.015	0.025	0.023
нужно	**need**	0.023	0.027	0.021	0.017	0.015
лучше	**better**	0.014	0.014	0.014	0.018	0.015
всем	**to all**	0.013	0.025	0.019	0.017	0.015
потом	**later**	0.016	0.013	0.013	0.019	0.018
еще	**more**	0.013	0.013	0.014	0.022	0.021
время	**time**	0.020	0.026	0.013	0.014	0.014
нас	**us**	0.027	0.017	0.012	0.023	0.019
быть	**be**	0.019	0.014	0.012	0.017	0.017
много	**many**	0.014	0.013	0.014	0.013	0.011
кого	**whom**	0.053	0.018	0.028	0.011	0.011
день	**day**	0.013	0.024	0.010	0.014	0.011
всех	**of all**	0.016	0.013	0.010	0.014	0.016

After collecting records of the selected communities for the given period of time, in accordance with the frequency dictionary algorithm, we selected texts of the various records, before splitting texts into words. Spell-checking was then performed and the number of posts and comments for each of the studied communities was calculated (table 1). The frequency dictionaries obtained as a result of the algorithm are partially presented in table 2.

The graphical representation of the results of the principal component analysis application to frequency dictionaries is shown in figure 2.

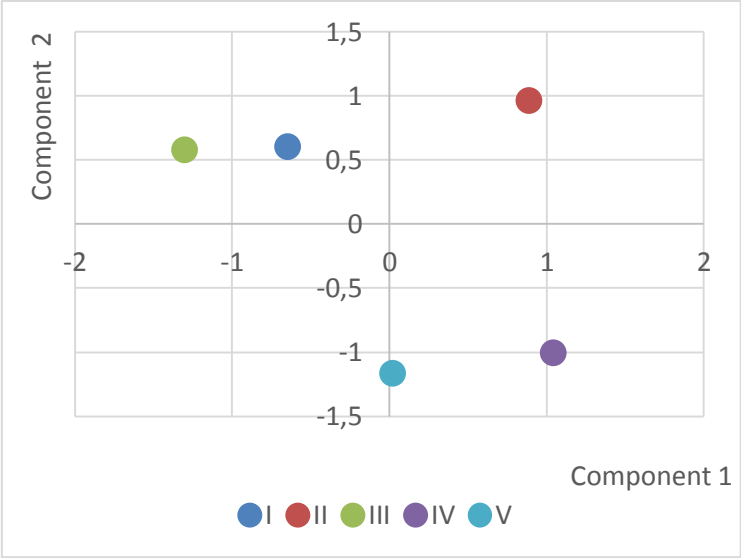

Figure 2. Graphical representation of the principal component analysis.

As can be seen from figure 2, the communities were divided into 3 groups:

- University communities "Overheard in Samara University" and "Overheard in SamSTU".
- City communities "Overheard Samara" and "Heard Samara".
- University community "Overheard Samara University" (old name: "Overheard in SamSU").

The words describing the formed components are presented in table 3.

Table 3. Formed components.

Component number			
1		2	
Original	**Translation**	**Original**	**Translation**
вам	you	имена	First names
нас	us	дома	at home
мы	we	работа	work
всем	to all	интересно	interesting
был	was	могу	can
были	were	Самара	Samara

Meaningful interpretation of the metrics is based on the semantic meaning of its unique components. There are many plural nouns (“you”, “us”, “we”, “to all”) among them. This fact points to the importance of community identity and expressing oneself by belonging to a certain community. It seems that communities with high scores on this metric represents some real or virtual community in virtual space, highlighting social identity of its members, making it more salient. Another peculiarity of the components of the first metric is prevalence of the past term verbs (“was”, “were”). It means that for the members of such communities events that have already happen are more interesting than actual or future events. This is probably because members of such communities are more conservative in their views and more thoughtful and critical about social changes.

Unique components of the second metric consist largely of first names of persons. It possibly implies that conversation in communities scored high on second metric is more personal and direct. Other components ("at home", "work", "can", "interesting") are about mundane and usual activities. It is plausible that communities high on this metric are more about daily routine, and its members are more concerned about their personal circumstances and are not much interested in discussing topics about their community or society.

Scores of the measured communities on the obtained metrics support the interpretation given above. For instance, the second metric clearly separates city communities from university communities. We can say that discussion topics of analyzed university communities are about students' life, not about university events. On the other side, topics of analyzed city communities are more concerned about city events. In other words, university publics are for students mainly (and not for the staff), but city publics are for all the citizens who are active in social networks. These results also shows that the younger members of public are, the more this public should score on the second metric.

Scores of the analyzed publics on the first metric implies that community and social identity issues are not very interesting for students, they are more concerned about everyday routine. Nevertheless, "Overheard Samara University" community gained unusually high scores on the first metric. This is most likely because members of this community still identify themselves as students and alumni of former Samara State University that has been united with Samara National Aerospace University into Samara University a few years ago. We may say that active members of "Overheard Samara University" community feel sorry and nostalgic about the former university, and such feelings are not common among the members of the new community "Overheard in Samara University" of the same university. This fact makes it plausible that there are more alumni in "Overheard Samara University" than in newer "Overheard in Samara University" community.

6. Conclusion

Based on the analysis of the results, we can conclude that the developed algorithms allows us to evaluate the semantic differences of the Internet communities of social networks. The founded components can be used as metrics for analyzing the relative positions of the considered communities. Results can be improved by eliminating filler-words from the frequency dictionaries and grouping words by parts of speech. The task of further research is the development of technology for automatic filtering of frequency dictionaries.

7. References

[1] Kosinski M, Matz S C, Gosling S D, Popov V and Stillwell D 2015 Facebook as a Research Tool for the Social Sciences: Opportunities, Challenges, Ethical Considerations, and Practical Guidelines *American Psychologist* **70(6)** 543-556

[2] Spitsyn V G, Bolotova Yu A, Phan N H and Bui T T T 2016 Using a Haar wavelet transform, principal component analysis and neural networks for OCR in the presence of impulse noise *Computer Optics* **40(2)** 249-257

[3] Raynard R 2017 *Economic Psychology* (Hoboken, NJ: John Wiley & Sons) p 512

[4] Rytsarev I A and Blagov A V 2016 Classification of Text Data from the Social Network Twitter *CEUR Workshop Proceedings* **1638** 851-856

[5] Rytsarev I and Blagov A 2017 Creating the Model of the Activity of Social Network Twitter Users *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* **9** 27-30

[6] Nofer M, Hinz O 2015 Using Twitter to Predict the Stock Market *Business & Information Systems Engineering* **57(4)** 229-242

[7] Trottier D and Fuchs C 2014 *Social Media, Politics and the State: Protests, Revolutions, Riots, Crime and Policing in the Age of Facebook, Twitter and YouTube* (New York: Routledge) p 252

[8] Housholder E E and LaMarre H L 2014 Facebook Politics: Toward a Process Model for Achieving Political Source Credibility Through Social Media *Journal of Information Technology & Politics* **11(4)** 368-382

[9] Kosinski M, Stillwell D and Graepel T 2013 Private Traits and Attributes are Predictable from Digital Records of Human Behavior *PNAS Proceedings of the National Academy of Sciences of the United States of America* **110** 5802-5805
[10] Kosinski M, Bachrach Y, Kohli P, Stillwell D J and Graepel T 2014 Manifestations of User Personality in Web Site Choice and Behaviour on Online Social Networks *Machine Learning* **95** 357-380
[11] Park G, Schwartz H A, Eichstaedt J C, Kern M L, Kosinski M, Stillwell D J, Ungar L H and Seligman M E P 2015 Automatic personality assessment through social media language *Journal of Personality and Social Psychology* **108(6)** 934-952
[12] Howlader P, Pal K K, Cuzzocrea A and Kumar S D M 2018 Predicting facebook-users' personality based on status and linguistic features via flexible regression analysis techniques *Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC '18)* 339-345
[13] Matz S C and Netzer O 2017 Using Big Data as a window into consumers' psychology *Current Opinion in Behavioral Sciences* **18** 7-12
[14] Mikhaylov D V, Kozlov A P and Emelyanov G M 2017 An approach based on analysis of *n*-grams on links of words to extract the knowledge and relevant linguistic means on subject-oriented text sets *Computer Optics* **41(3)** 461-471
[15] Bolotova Yu A, Spitsyn V G, Osina P M 2017 A review of algorithms for text detection in images and videos *Computer Optics* **41(3)** 441-452
[16] Mikhaylov D V, Kozlov A P, Emelyanov G M 2016 Extraction of knowledge and relevant linguistic means with efficiency estimation for the formation of subject-oriented text sets *Computer Optics* **40(4)** 572-582 DOI: 10.18287/2412-6179-2017-41-3-461-471
[17] Jolliffe I T 2002 *Principal Component Analysis* (Springer) p 487
[18] Wasserman L 2005 *All of Statistics: A Concise Course in Statistical Inference* (Springer) p 442

Acknowledgments
This work was partially supported by the Ministry of education and science of the Russian Federation in the framework of the implementation of the Program of increasing the competitiveness of Samara University among the world's leading scientific and educational centers for 2013-2020 years; by the Russian Foundation for Basic Research grants (# 15-29-03823, # 16-41-630761, # 17-01-00972, # 18-37-00418), in the framework of the state task #0026-2018-0102 "Optoinformation technologies for obtaining and processing hyperspectral data".