Indexing of Social Network Texts for a Psychometric Model of Social Success

Nikolai Prokopyev [1] and Pavel Ustin [2]
[1] *Institute of Computational Mathematics and Information Technologies, Kazan Federal University, Kazan, Russia*
[2] *Institute of Psychology and Education, Kazan Federal University, Kazan, Russia*
nikolai.prokopyev@gmail.com, pavust@mail.ru

Keywords: Natural Language Processing, Text Indexing, Sociological Psychology.

Abstract: Development of information technologies, mathematical methods and big data processing capabilities makes it possible to build and test formal psychometric models for their further use in creating software systems that can predict the success of a person's personal activity. This paper presents a description of methods and technologies for indexing of social network texts used in the development of a psychometric model of social success in the framework of the project "Neural network psychometric model of cognitive-behavioural predictors of life activity of a person on the basis of social networks". As data for indexing, anonymized dataset was collected based on the resumes of professionals posted on job search sites and their profiles on the Vkontakte social network. According to results of the study, such a cognitive behavioural predictor of success as texts themes, that are typical for professionals with high and low social success, was identified. The obtained results can later be used in training neural network models to predict the social success of a person.

1 INTRODUCTION

The "Neural network psychometric model of cognitive-behavioural predictors of life activity of a person on the basis of social networks" project (Prokopyev et al., 2023) is a 5-year interdisciplinary study at intersection of personality psychology and open data processing by machine learning methods. It is aimed at developing and evaluation of a neural network psychometric model that allows predicting the educational, professional and social success of a person through the structure of cognitive-behavioural predictors presented in the metrics of personal profile in social networks. Relation between life activity of a person and virtual representation, which substantiates the tasks of the project, was disclosed in papers (Ryan & Xenos, 2011) and (Kosinski et al., 2013).

Previously in the project anonymized data was collected from profiles of the Vkontakte social network from students of Kazan Federal University, as well as professionals who had their resumes on the Headhunter job search site. Predictors were identified and neural network models of academic (Gafarov et al., 2020) and professional (Gafarov et al., 2011) success of a person were built. One of the current tasks of the project is development of a neural network psychometric model of social success.

The main components of this model are predictors - numerical characteristics extracted from the profile on a social network and correlated with whether a person belongs to group of socially successful or socially unsuccessful. Sources of these predictors are two groups of data - quantitative (expressed numerically, such as, for example, number of friends, number of posts, number of subscriptions, number of photos) and qualitative (expressed not numerically, such as content of posts, photos). In this paper, we consider an approach to qualitative analysis of textual information of social network profiles to identify predictors of social success.

[1] https://orcid.org/0000-0003-0066-7465
[2] https://orcid.org/0000-0003-3950-7434

2 PRELIMINARY WORK

2.1 Data Acquisition and Analysis

Initial data for the qualitative analysis of textual information are texts of posts and reposts on profiles in the Vkontakte social network from a certain dataset. In this study, dataset of professionals was used, that is, profiles of people who posted their resumes on a job search site that were previously collected in the project.

Profile addresses were searched using a special system for automated search (Berdnikov et al., 2019). This system includes two search modules based on original search algorithms and an anonymization module. The first search module searches for personal profiles in a social network based on analysis of unique primary data (last name, first name, place of study, date of birth). The second search module includes an additional algorithm based on analysis of social graphs such as friends who belong to the same social group.

Using this system, 149 430 social profiles were found among the initial 436 031 resumes and processed for plausibility. Next, since key metrics in the social network were not found for many user profiles, the sample was reduced to 61 232 profiles. Among this sample, posts and reposts from 60 317 profiles were extracted. The total number of extracted texts was 10 113 389 posts and 11 725 410 reposts.

Next step was to divide the obtained dataset into groups of social success according to the original formula, which is the topic of a separate article. However, this formula required additional data from the social network, which was obtained using profile monitoring system. This system collects dynamic data of user activity (online presence, posting, commenting, adding friends and subscribing) as well as graphs of friends and friends of friends.

Using this system, extended data for 1980 profiles were obtained. With the original formulas, these profiles were divided into 5 groups of social success according to characteristic "Dominance" and "Friendliness". These characteristics are two different types of social success highlighted in the project.

2.2 Indexation Methodology

There are various approaches to natural language processing: rule-based, statistical (corpus linguistics and machine learning), but they are always associated with transformation of arbitrary text into a mathematically formalized and labelled array of numerical data. With this method of formalization, task of qualitative text analysis is reduced to a set of tasks of quantitative analysis.

To formalize texts, we use the text indexing method, one among the methods of latent semantic analysis (LSA) (Valdez, 2018). This method consists in dividing the text into separate words (terms) and counting the statistical characteristics of terms: number of uses, number of documents in which the term is used, resulting in a matrix called text index. Latent semantic analysis makes it possible to reveal relation between a set of texts and words found in them, thereby revealing the characteristic themes inherent in all texts in the sample.

Since text indexing is usually used to implement information retrieval, it includes the concept of a document, which means any logically coherent source of texts. In this study, a document refers to the profile page of a social network user. General algorithm of indexing process is shown in Fig. 1 and discussed below.

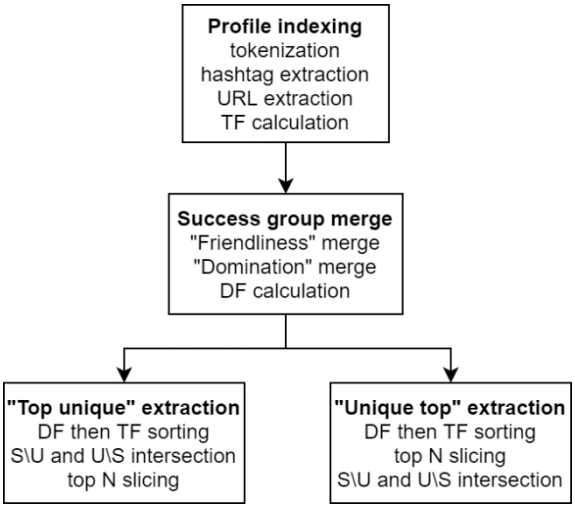

Figure 1: Indexing algorithm.

Indexing process begins with tokenization of the input text into terms, during which it removes unnecessary punctuation characters and highlights special semantic elements, such as hashtags and URLs, for which separate indexes are built. Terms are reduced to the basic grammatical form using stemming. Next, occurrence frequency of each term (TF) in the profile posts and reposts is calculated.

Thus, for each profile, 6 types of indices are built:

1. Index of words from posts;
2. Index of hashtags from posts;
3. Index of URLs from posts;
4. Index of words from reposts;
5. Index of hashtags from reposts;
6. Index of URLs from reposts.

Next, these profile indices are combined on the basis of belonging to the previously identified 5 groups of social success separately according to characteristic "Dominance" and "Friendliness", that is, the terms are collected into new index for the group of social success, and their TF are summed. Also, the document frequency (DF) is additionally calculated, that is, number of profiles in the group that use the term.

Among the groups of social success in terms of both characteristics, the most interesting for our research are the unsuccessful group (group 1) and the successful group (group 5). By studying the most frequent terms from the corresponding indices, unique for a particular group, in accordance with hypothesis of the project, predictor words which are characteristic for socially successful and socially unsuccessful persons are identified. General approach to constructing subsets of such terms is to sort the terms in index by their quantitative characteristics (primarily by DF, since prevalence of a word among group profiles is a more important predictor), identify top N terms, and find the difference between subsets of terms for a group of successful and unsuccessful in terms of set theory. However, order in which these steps are applied determines two different approaches.

The first approach is called "Top Unique". In accordance with it, for the sorted indices S (successful) and U (unsuccessful), their differences S\U and U\S are found. After that, for each difference, top N terms are extracted. Interpretation of this approach is: "Select the most frequently used words among the words uniquely used by a group."

The second approach is called "Unique Tops". In this approach, top N words are first extracted from the sorted indices, subsets of SN (successful) and UN (unsuccessful) are obtained. After that, their differences SN\UN and UN\SN are found. Interpretation of this approach is: "Select uniquely used words among the most frequently used by a group."

Both methods give different results in terms of suitability. In particular, "Top unique" method, despite the fact that it gives output terms that are truly unique for success groups, tends to produce too few words, unsuitable for extracting predictors. "Unique tops" method, in turn, does not produce terms that are really unique among all terms, only within the top N terms. Both methods are implemented in the project and their use is justified depending on the results obtained.

3 RESULTS

3.1 Indexing Results

As a result of indexing the original 1980 profiles, divided evenly into 5 groups of success, indexes of groups were obtained in various volumes. Table 1 shows number of profiles in dataset of each social success group for both characteristics and number of profiles included in the indexes.

Table 1: Indexing statistics.

Group	Domination		Friendliness	
	In dataset	In index	In dataset	In index
1	396	94	396	109
2	396	129	396	148
3	396	182	396	164
4	396	236	396	206
5	396	276	396	290

A profile is not included in the index if it has no posts or is private. From Table 1, it can be seen that for groups of lower social success, absence of posts is more characteristic, since the number of posts in itself is one of social success predictors, as revealed during the research.

“Unique top” approach was chosen as the main approach for constructing subsets of predictor words characteristic of socially successful and socially unsuccessful profiles, since it provided a sufficient number of predictor words for analysis in a small sample. Tables 2 and 3 show some selected predictor words for socially successful and unsuccessful profiles according to characteristic "Dominance", and in tables 4 and 5 - according to characteristic "Friendliness".

Table 2: Predictor words for socially successful profiles according to characteristic "dominance".

Term (Russian / English)	**DF**	**TF**
место / place	24	91
сила / power	22	105
последний / last	22	53
желание / wish	21	53
музыка / music	21	42
искать / search	19	84
добрый / kind	19	71

Term (Russian / English)	DF	TF
радость / joy	19	51
фото / photo	18	35
главный / main	17	76
группа / group	15	75

Table 3: Predictor words for socially unsuccessful profiles according to characteristic "dominance".

Term (Russian / English)	DF	TF
участие / participation	3	7
победитель / winner	3	4
оценка / grade	3	3
отечество / fatherland	2	3
защитник / defender	2	2

Table 4: Predictor words for socially successful profiles according to characteristic "friendliness".

Term (Russian / English)	DF	TF
душа / soul	20	27
музыка / music	19	38
желание / wish	18	24
ребёнок / child	16	83
песня / song	16	28
радость / joy	16	25
попробовать / try	14	64
делать / do	14	32
добрый / kind	14	19
счастливый / happy	13	30

Table 5: Predictor words for socially unsuccessful profiles according to characteristic "friendliness".

Term (Russian / English)	DF	TF
участие / participation	3	7
признание / recognition	3	5
отечество / fatherland	3	4
связь / relation	3	3
состояться / take place	2	4
аниме / anime	2	4

3.2 Discussion

Analysing the obtained subsets of predictor words for social success according to characteristics "Dominance" and "Friendliness", we can draw the following points:

1. Corresponding groups of success for characteristics "Dominance" and "Friendliness" overlap to a large extent, which indicates the similarity of these types of social success.
2. For groups of unsuccessful profiles, a serious shortcoming in the dataset size is revealed, as a result of which the subsets of words obtained for them cannot be called full-fledged predictors of these groups.
3. For the group of social success according to characteristic "Dominance", predictor words largely reflect the theme of activity and hobbies.
4. For the group of social success according to characteristic "Friendliness", predictor words largely reflect the theme of life self-realization.

4 CONCLUSIONS

The paper presents one of latent semantic analysis methods in its application to the tasks of an interdisciplinary project to build a neural network psychometric model of personal social success and the results of its application on a dataset of professional profiles in social networks. Further work:

1. Increase of the dataset and updating the subsets of predictor words to improve quality of analysis, especially in relation to the group of unsuccessful profiles, characterized by a small number of posts in the social network.
2. Application of the content analysis method using a categorical grid, since latent semantic analysis in itself does not imply an intellectual analysis by a person. The essence of this method lies in categorization of predictor words based on a psychometric model for predicting the social activity of a person. Use of categorical grid makes it possible to reduce the amount of predictor words by including only those words that are significant both statistically and within the psychometric model.
3. Verification of a categorical grid, that is, testing the hypothesis about its effectiveness. This hypothesis is based on the assumption that frequency predictor words in a certain group should monotonically decrease for all subsequent groups of success.

4. Using a categorical grid to train a neural network psychometric model of personal social success.

ACKNOWLEDGEMENTS

The study (all theoretical and empirical tasks of the research presented in this paper) was supported by a grant from the Russian Science Foundation (Project No. 19-18-00253, "Neural network psychometric model of cognitive-behavioral predictors of life activity of a person on the basis of social networks"), https://rscf.ru/project/19-18-00253/.

REFERENCES

Prokopyev, N., Ustin, P., Abdurakhmonova, N., 2023. Application of Open Data Processing in Sociological Psychology. 2022 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2023, pp. 1-4.

Ryan, T., Xenos, S., 2011. “Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage. *Computers in Human Behavior*, Vol. 27, Is. 5, 2011, pp. 1658-1664.

Kosinski, M., Stillwell, D., Graepel, T., 2013. Private traits and attributes are predictable from digital records of human behaviour. *Proceedings of the National Academy of Sciences of the United States*, Vol. 110(15).

Gafarov, F., Enikeeva, Z., Vakhitov, G., 2020. The prediction of educational success of students-humanitarians in social networks from the psychometric view. *International Journal of Pharmaceutical Research*, Vol. 12, Is. 1, 2020, pp. 804-811.

Gafarov, F.M., Nikolaev, K.S., Ustin, P.N., Berdnikov, A.A., Zakharova, V.L., Reznichenko, S.A, 2021. A Complex Neural Network Model for Predicting a Personal Success based on their Activity in Social Networks.*Eurasia Journal of Mathematics, Science and Technology Education*, Vol. 17, Is. 10, 2021, pp. 1-9.

Berdnikov, A.A., Davletshin, N.M., Gafarov, F.M., Ustin, P. N., 2019. Development of a System for Automated Search of Personal *Profiles of Students in the Social Network VKontakte. Computer Technologies and Modeling in Science, Technology, Economics, Education and Management: Trends and Developments*, pp. 302-305.

Valdez, D., Pickett, A., Goodson, P., 2018. Topic Modeling: Latent Semantic Analysis for the Social Sciences. Social Science Quarterly, Vol. 99.