Population and Economics

Population and Economics 4(3): 98–103
DOI 10.3897/popecon.4.e60519

DATA PAPER

Database "Pro-family (pronatalist) communities in the social network VKontakte"

Irina E. Kalabikhina[1], Evgeny P. Banin[2,3]

1 *Lomonosov Moscow State University, Moscow, 119991, Russia*
2 *Bauman Moscow State Technical University, Moscow, 105005, Russia*
3 *Research Center "Kurchatov Institute", Moscow, 123182, Russia*

Received 5 November 2020 ♦ **Accepted** 20 November 2020 ♦ **Published** 18 December 2020

Citation: Kalabikhina IE, Banin EP (2020) Data base "Familiist (pro-natalist) communities in the social network VKontakte". Population and Economics 4(3): 98-103. https://doi.org/10.3897/popecon.4.e60519

Abstract

The database contains uploading text comments from the social network VKontakte **in .csv format (UTF-8 encoding)**. The comments are collected from communities discussing pregnancy, childhood, motherhood, etc. Uploading contains comments to posts with which the interaction took place. The absolute number of likes was used as a criterion (comments were collected where the number of likes is greater than or equal to 5). Text data was pre-processed (stemmization and lemmatization).

The data is suitable for thematic analysis (e.g. LDA – Latent Dirichlet Allocation), for modelling the graph structure of communities (the link_comment variable contains a unique post identifier, link_author contains a unique user identifier), for analysis of tonalities of statements and formation of a dictionary of demographic connotation in Russian. Analysis of the tonalities of statements enables measuring the dynamics of "demographic temperature" in pro-family (pronatalist) communities.

Keywords

Database, big data, pronatalism, VKontakte, social networks, communities, family values

JEL codes: J1, C31, C32, D71, E71

Data access and data format

Database name "Pro-family (pronatalist) communities in the social network VKontakte" Copyright I.E. Kalabikhina, E.P.Banin The database is in the public domain and under the Creative Commons Attribution license (CC-BY 4.0) can be used, distributed and reproduced without limitation on any medium subject to indication of the authors and the source. Irina Kalabikhina, Evgeny Banin: Pro-family (pronatalist) communities in the social net-

work VKontakte. Access mode: https://doi.org/10.5281/zenodo.4244361. Data format .csv (UTF-8 encoding).Description: Data can be downloaded from an open source (Zenodo online depository), where the database of Pro-family (pronatalist) communities in the social network VKontakte is located. Data file vk_posts_stem_lemm.scv 117.0 MB

Brief overview of literature and motivation of research. Two major projects with developed mood dictionaries are known in the Russian-language segment: RusentiLeX (Loukachevitch and Levchik 2016) and LINIS Crowd (Koltsova, Alexeeva and Kolcov 2016). Both projects are developed dictionaries with an assessment of tonality (from positive to negative) for each word or combination of words without characterizing emotional colouring. More complex models of tonality are offered in the (Baccianella, Esuli and Sebastiani 2010) SentiWordNet and SenticNet projects (Cambria et al. 2016, 2018) based on the analysis of the tonality of the English language. At present, deep neural network learning techniques are most actively developed, which demonstrate (Tang, Qin and Liu 2015; Tang and Zhang 2018) the best current results of tonality assessment compared to the rest of approaches. The methods of tonality analysis based on machine learning are characterized by the need for pre-learning on large sets of marked texts. Attempts to combine the two approaches presented (based on rules and methods of machine learning), for example, work, are known (Meškele and Frasincar 2019; Kumar et al. 2020).

In the review work on applied analysis of tonalities, it is (Smetanin 2020) noted that for the Russian language this direction is not sufficiently studied (the author notes 27 most relevant research papers on tonalities analysis in Russian). Much of the research focuses on analyzing the tonalities of tweets (short posts) on the social network Twitter. In Russia in 2019 this social network was used by about 650,000 active users. The social network VKontakte has the greatest coverage of the Russian-speaking population. According to a report by the consultancy Deloitte (Zemlyanskaya et al. 2018), VKontakte covers up to 70% of the Russian population.

High coverage of the population predetermined the choice of the social network VKontakte as a source of textual data for analysis. The analysis of various sources made it possible to conclude that there is insufficient elaboration of models of tonalities evaluation in the Russian-speaking segment as a whole, and even existing works set limited tasks and do not move from the level of individual small communities to the regional or country level. Most difficult for tonality analysis are demographic topics from the field of reproductive behaviour (compared to self-preservation or migratory behaviour). These circumstances led to the creation of a database in the Russian-language segment using machine learning on text data from the social network VKontakte on topics in the field of reproductive behaviour.

Data collection methodology. This study attempts to test machine learning tools on text data obtained from the social network VKontakte. Collection of unstructured text data from communities was carried out, preliminary data processing (cleaning, lemmatization, stemmization and removal of punctuation) was carried out, a structured array (body) of texts was formed. Thematic clusters have been identified based on Latent Dirichlet Allocation, LDA. After thematic analysis, the tonalities of texts were evaluated for each cluster and the dynamics of change of tonality in time was constructed for comments (publication on this model testing in print).

The thematic model is a text document collection model that determines which topics the document refers to. In addition to highlighting the structure of text collection, thematic modelling allows for semantic information retrieval (as opposed to keyword search, where meaning is not explicitly represented).

TensorFlow and tflearn libraries are used for tonality analysis. Neural network training is carried out on a marked database of short messages from twitter (Rubtsova 2015). Neural network training is performed in the Google Colab environment using a graphical accelerator (GPU, *graphics processing unit*). About 24 GB of RAM is used to teach the neural network with the training dictionary amounting to 5,000 words. Before training, the data was stemmized (brought to the basic form of the word), all non-Cyrillic characters were eliminated from the sample. The test sample size is 30% of the entire sample. The number of eras for training is 30. The resulting accuracy on the training sample is 93.4%, on the test sample — 69%. The probability threshold for assigning a comment as positive or negative is 0.5.

Data sources. The source of text data is thematic communities in the social network VKontakte (vk.com). At the first stage of processing using the built-in API (application programming interface) unique address numbers of thematic communities in the form *vk.com/* were collected by keywords ("mom", "mommies", "kids", "child", "baby", "health", "birth", "pregnancy", "parents"). In the first phase, about 1,000 unique group addresses were collected with data on the number of participants. In the second stage, ad-related communities as well as communities with low member activity were excluded from the sample (the overall dynamics of changes in the number of posts, likes and reposts was assessed) together with those with a number of subscribers under 10000.

Information about the sample

- Only comments with the number of likes >= 5 were gathered
- Comments are gathered only by communities (list of communities below), which discuss issues related to childhood, motherhood, pregnancy, etc.
- The sample of communities on average contains 309 thousand subscribers (maximum value – 1,482,303, minimum value – 72,570, total number of subscribers excluding intersections – 11,743,295)
- The comment sample contains a total of 112,900 user comments

Following the formation of the final list of communities, textual information from the communities was gathered. In this paper, the collected texts are limited only to posts and comments. Based on the information gathered, a language body was formed: all words were brought to lower case, stop words were removed using functions from the nltk or gensim library, punctuation was removed, numerical data were excluded. To reduce the volume of text data, stemmization (deletion of suffixes) or lemmatization (bringing the word to the initial form using the mySteM lemmatizer) was additionally carried out. We have determined that the comment body is most appropriate to assess tonality.

For the **sample structure and list of major groups**, see the database description in the International Depository https://doi.org/10.5281/zenodo.4244361.

Database Application Areas

The data is suitable for thematic analysis (e.g. LDA - Latent Dirichlet Allocation), for modelling the graph structure of communities (the link_comment variable contains a unique post identifier, link_author contains a unique user identifier), for analysis of tonalities of statements and formation of a dictionary of demographic connotation in Russian.

Analysis of the tonalities of statements enables measuring the dynamics of "demographic temperature" in pro-family (pronatalist) communities. By demographic temperature we mean the emotional background or the predominance of positive or negative tonality of statements on topics related to family values, childbirth and other topics in the field of reproductive behaviour. Demographic temperature is measured as the difference between the number of positive and the number of negative statements over a certain period of time.

Let us emphasize once again that the demographic temperature in this case is measured in communities of people with pronatalist views, that is, reproductive attitudes towards creating a family and having children. In our view, the measurement of population temperature in such communities best shows the state of the demographic climate, shaped by demographic and family policies, economic trends (but purified from the influence of the population structure according to the criterion of pronatalist or anti-natalist reproductive attitudes).

The presented database enables comparing the demographic temperature in individual clusters of communities in social networks, study the dynamics of positive and negative comments of women and men on demographic topics in the areas of childbirth, parenthood and family values.

In the next versions, we plan to add anti-natalist groups and more tonalities (besides positive and negative ones).

Example of demographic temperature measurement using the database described

Figure 1 shows the distribution of comments by month since the beginning of 2014, taking into account the author's gender. It should be noted that women show the greatest activity in the represented pronatalist groups. A surge and increasing activity has been observed since June 2017.

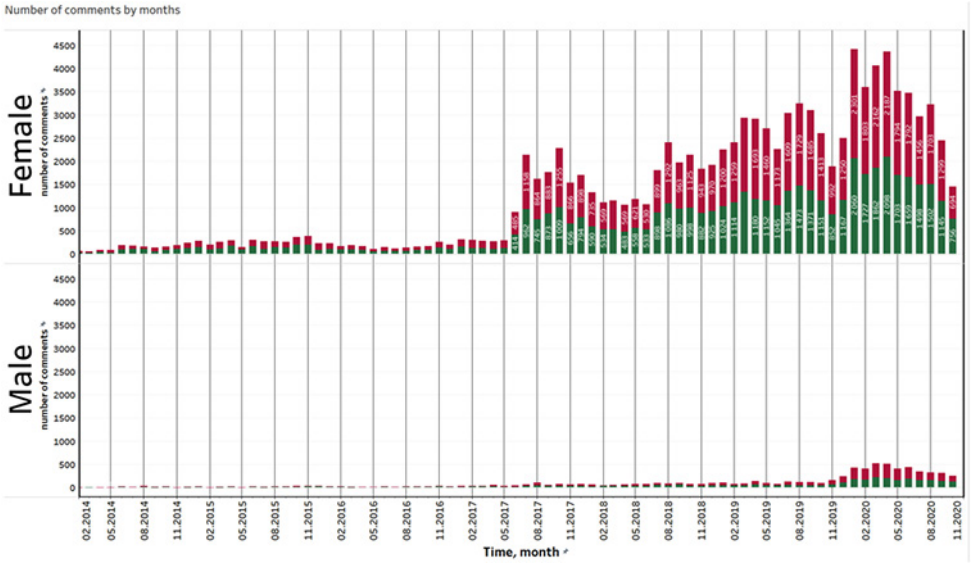

Figure 1. Distribution of the number of comments in the body by month. "Red" is negative comments, "green" is positive. *Source*: compiled by the authors in Tableau 19.3 based on gathered data from the social network "VKontakte"

The demographic temperature in the investigated pronatalist communities (the difference between positive and negative comments) is shown in Figure 2. It can be noticed that since the end of 2016 the number of negative comments has increased significantly, the demographic temperature has become negative. In the recent history of Russia, the historical maximum of the total fertility rate was observed in 2015 — 1.78 children per 1 woman of the conditional generation, then the decline of the birth rate to 1.50 began in 2019 (and we estimate to 1.4 in 2020).

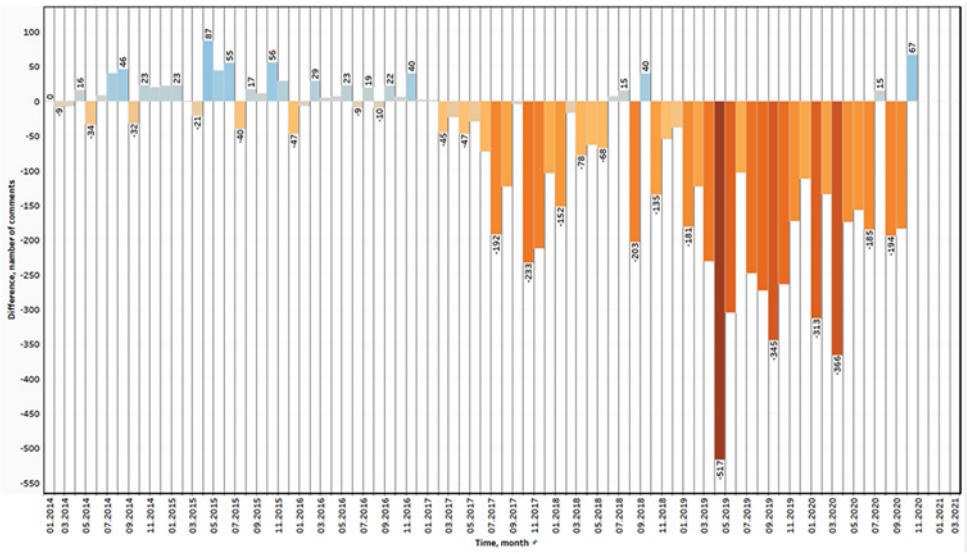

Figure 2. Difference between negative and positive comments in the body of comments by month. "-" is the predominance of negative comments. *Source*: compiled by the authors in Tableau 19.3 based on gathered data from the social network "VKontakte"

Contribution to the creation and development of the database: The idea and concept of creating a database based on the developed range of applications of the database in the demographic analysis of fertility, reproductive behaviour, population response to population policy and other factors of reproductive behavior — Doctor of Economics. I.E. Kalabikhina. Methods of base creation, creation of the first variant of the base — Banin E.P. The base was created within the framework of the implementation of the internal grant of the Faculty of Economics of Lomonosov Moscow State University. The authors thank project colleagues for their assistance in formulating thematic words and phrases, searching for samples of pronatalist groups in the social network: Abduselimova I.A., Arkhangelskiy V.N., Klimenko G.V., Kolotusha A.V., Nikolaeva U.G., Shamsutdinova V.Sh.

Reference list

Baccianella S, Esuli A and Sebastiani F (2010) SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010, 2200–2204. Available at: http://wordnetcode.princeton. (Accessed: 28 September 2020).

Cambria E et al. (2016) SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In: COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers, 2666–2677. Available at: https://www.aclweb.org/anthology/C16-1251.pdf (Accessed: 28 September 2020).

Cambria E et al. (2018) SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In: 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 1795–1802. Available at: www.aaai.org (Accessed: 28 September 2020).

Koltsova OY, Alexeeva SV and Kolcov SN (2016) An opinion word lexicon and a training dataset for Russian sentiment analysis of social media. In: Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, 277–287. Available at: http://linis-crowd.org. (Accessed: 28 September 2020).

Kumar A et al. (2020) Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data, Information Processing and Management, 57(1). doi: 10.1016/j.ipm.2019.102141.

Loukachevitch N, Levchik A (2016) Creating a general Russian sentiment lexicon. In: Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, 1171–1176. Available at: http://www.labinform.ru/pub/ruthes/index.htm (Accessed: 28 September 2020).

Meškele D, Frasincar F (2019) ALDONA: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalised domain ontology and a neural attention model. In: Proceedings of the ACM Symposium on Applied Computing. Association for Computing Machinery, 2489–2496. doi: 10.1145/3297280.3297525.

Rubtsova YV (2015) Constructing a corpus for sentiment classification training // International journal "Software & Systems" 27: 72–78. doi: 10.15827/0236-235X.109.072-078.

Zemlyanskaya L, Gordeev M, Afanasyeva Y (2018) Recovery of tolerance for Internet advertising: Media consumption in Russia 2018, Deloitte CIS Research Center, link: https://www2.deloitte.com/content/dam/Deloitte/ru/Documents/research-center/media-consumption-in-russia-2018-en.pdf (access: 08.10.2020)

Smetanin S (2020) The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives. IEEE Access, 8: 110693–110719. doi: 10.1109/ACCESS.2020.3002215.

Tang D, Qin B, Liu T (2015) Deep learning for sentiment analysis: Successful approaches and future challenges // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. Wiley-Blackwell 5(6): 292–303. doi: 10.1002/widm.1171.

Tang D, Zhang M (2018) Deep learning in sentiment analysis. In: Deep Learning in Natural Language Processing. Springer International Publishing, 219–253. doi: 10.1007/978-981-10-5209-5_8.

Information about the authors

- Irina Evgenievna Kalabikhina, Doctor of Sciences (Economics), Professor, Head of the Population Department, Faculty of Economics, Lomonosov Moscow State University. E-mail: kalabikhina@econ.msu.ru
- Evgeny Petrovich Banin, Research Engineer, Bauman Moscow State Technical University, Research Center "Kurchatov Institute"; PhD student, Bauman Moscow State Technical University. E-mail: evg.banin@gmail.com