Population and Economics

Population and Economics 5(2): 92–96
DOI 10.3897/popecon.5.e70786

DATA PAPER

Database "Childfree (antinatalist) 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 28 June 2021 ♦ **Accepted** 30 June 2021 ♦ **Published** 27 July 2021

Citation: Kalabikhina IE, Banin EP (2021) Database "Childfree (antinatalist) communities in the social network VKontakte". Population and Economics 5(2): 92-96. https://doi.org/10.3897/popecon.5.e70786

Abstract

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

The data are suitable for thematic analysis (e.g. LDA — Latent Dirichlet Allocation), sentiment analysis of statements, modelling the graph structure of communities (the link_comment variable contains a unique identifier of the post, link_author contains a unique user identifier), and forming a dictionary of demographic connotation in Russian. Sentiment analysis of statements enables measuring the dynamics of «demographic temperature» in antinatalist communities.

The database is a supplement to the publication Kalabikhina IE, Banin EP (2020) Database «Pro-family (pronatalist) communities in the social network VKontakte». Population and Economics 4(3): 98–130. https://doi.org/10.3897/popecon.4.e60915.

Keywords

database, big data, antinatalism, VKontakte, social networks, communities, family values, child free

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

Data access and data format

Database name: Childfree (antinatalist) 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) it can be used, distributed and reproduced without limitation on any medium subject to indication of the authors and the source. Irina Kalabikhina, Evgeny Banin: Childfree (antinatalist) communities in the social network VKontakte. Access mode: https://doi.org/10.5281/zenodo.4612131. Data format: .csv (UTF-8 encoding). Description: Data can be downloaded from an open source (Zenodo online depository), where the database is located. Data file Antinata_vk_sentiments_preparing.csv?download=1/. 1.2 GB.

The database is a supplement to the publication Kalabikhina IE, Banin EP (2020) Database «Pro-family (pronatalist) communities in the social network VKontakte». Population and Economics 4(3): 98-130. https://doi.org/10.3897/popecon.4.e60915. A brief overview of literature see in (Kalabikhina and Banin 2020).

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

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 sentiment 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 («childfree», «child», «health», «birth», «parents», etc.). In the first phase, about 100 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 500.

Information about the sample

- The sample of communities contains 8 groups (number of subscribers without self-intersection is about 100 000)

- Content type of communities: communities in which users make mainly negative comments about the birth of children, motherhood, parenthood and family are selected. However, some users with pro-familistic attitudes may be encountered within the database
- Only comments with the number of likes >= 5 are collected
- Groups with less than 500 subscribers are excluded
- Comments are collected only from communities (the list of communities below) discussing issues related to childfree, childhood, motherhood, pregnancy, etc.
- The sample contains 670 thousand user comments.

List of communities:

- https://vk.com/club69265846
- https://vk.com/club43946
- https://vk.com/club48085
- https://vk.com/club4687918
- https://vk.com/club38197124
- https://vk.com/club58565280
- https://vk.com/club59638638
- https://vk.com/club148257242

Sample structure and description of variables:

- **link_author** — link to the author of the comment in the form of https://vk.com/*author identificator*
- **gender of author —** (F — female, M — male, NaN — no data)
- **link_comment** — link to comment in the form of https://vk.com/* post identificatior on a *community wall*?reply=*comment id *
- **date_time** — date and time of publication (format YYYY-MM-DD HH:MM:SS)
- **text** — raw comment text
- **likes** — number of likes the comment has
- **text_prep** — processed text (punctuation marks removed, words brought down to lowercase)
- **text_stem** — processed text (based on the text_prep column stemmization using SnowBallstemmer («Russian») of the nltk library) is performed
- **text_sw** — processed text (based on the text_prep column stop words are deleted using word_tokenize (text) of the nltk library)
- **text_lemm** — processed text (lemmatization using mystem.lemmatize (text) of pymystem3 library is performed based on the text_prep column)

Database application

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 sentiment analysis of statements and formation of a dictionary of demographic connotation in Russian.

Analysis of the sentiment of statements enables measuring the dynamics of «demographic temperature» in antinatalist communities. By demographic temperature we mean the emotional background or the predominance of positive or negative sentiment 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 or ratio between the number of positive and the number of negative statements over a certain period of time.

Within this database, the demographic temperature is measured in communities of people with antinatalist views, that is, reproductive attitudes towards non-creating a family and not having children.

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.

The first publication on measuring the demographic temperature using the methodology for measuring the sentiment of statements in the social network VKontakte (Kalabikhina et al. 2021) is based on two databases: the one described in this article and the one described in (Kalabikhina and Banin 2020).

This is the first attempt to analyze the sentiment of Russian-language comments in the social network VKontakte to determine the demographic temperature in various social and demographic groups among the users of the network. In particular, using the available data in two types of groups since 2014, we find an asynchronous structural shift in comments of the corpuses of pronatalist and antinatalist thematic groups (Kalabikhina et al. 2021).

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 od Sci. (Econ.) Irina Kalabikhina. Methods of database creation, creation of the first version of the database — Evgeny Banin. The database 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

Kalabikhina IE, Banin EP (2020) Database «Pro-family (pronatalist) communities in the social network VKontakte». Population and Economics 4(3): 98–130. https://doi.org/10.3897/popecon.4.e60915

Kalabikhina IE, Banin EP, Abduselimova IA, Klimenko GA, Kolotusha AV (2021) The Measurement of Demographic Temperature Using the Sentiment Analysis of Data from the Social Network VKontakte. Mathematics 9(9): 987. https://doi.org/10.3390/math9090987

Rubtsova YV (2015) Postroenie korpusa tekstov dlya nastroiki tonovogo klassifikatora [Constructing a corpus for sentiment classification training]. Programmnye produkty i sistemy [Software & Systems] 27: 72–8. https://doi.org/10.15827/0236-235X.109.072-078 (in Russian)

Information about the authors

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