PE Population and Economics

Население и экономика 5(2): 92–96
DOI 10.3897/popecon.5.e70786

СТАТЬЯ О БАЗЕ ДАННЫХ

База данных «Чайлдфри (антинаталистские) сообщества в социальной сети “ВКонтакте”»

Ирина Е. Калабихина[1], Евгений П. Банин[2,3]

1 *МГУ имени М. В. Ломоносова, Москва, 119991, Россия*

2 *Московский государственный технический университет имени Н.Э. Баумана, Москва, 105005, Россия*

3 *Научно-исследовательский центр «Курчатовский институт», Москва, 123182, Россия*

Получено 28 June 2021 ♦ **Принято в печать** 30 June 2021 ♦ **Опубликовано** 27 July 2021

Цитирование: 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

Аннотация

База данных содержит выгрузку текстовых комментариев из социальной сети «ВКонтакте» в формате **.csv (кодировка UTF-8)**. Комментарии собраны из групп, в которых обсуждаются вопросы беременности, детства, материнства и т.п. Выгрузка содержит комментарии под постами, с которыми происходило взаимодействие. В качестве критерия использовалось абсолютное количество лайков (собирались комментарии, где количество лайков больше или равно 5). Текстовые данные проходили предобработку (стеммизация и лемматизация).

Данные подходят для тематического анализа (например, LDA — Latent Dirichlet Allocation), для моделирования графовой структуры групп (переменная link_comment содержит уникальный идентификатор поста, link_author содержит уникальный идентификатор пользователя), для анализа тональностей высказываний и формирования словаря демографической коннотации на русском языке. Анализ тональностей высказываний позволяет измерять динамику «демографической температуры» в антинаталистских группах.

База данных является комплиментарной к опубликованной ранее базе данных 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.

Ключевые слова

база данных, большие данные, пронатализм, «ВКонтакте», социальные сети, сообщества, семейные ценности, чайлдфри

Коды JEL: J1, C31, C32, D71, E71

Доступ к данным и формат данных

Название базы данных: Чайлдфри (антинаталистские) сообщества в социальной сети «ВКонтакте». Copyright I.E. Kalabikhina, E.P. Banin. База данных находится в открытом доступе и в соответствии с лицензией Creative Commons Attribution (CC-BY 4.0) может без ограничений использоваться, распространяться и воспроизводиться на любых носителях при условии указания авторов и источника. Ирина Калабихина, Евгений Банин: Чайлдфри (антинаталистские) сообщества в социальной сети «ВКонтакте». Режим доступа: https://doi.org/10.5281/zenodo.4612131. Формат данных: .csv (кодировка UTF-8). Описание: Данные могут быть скачаны с открытого источника (онлайн-депозитарий Zenodo), где размещена база данных «Чайлдфри (антинаталистские) сообщества в социальной сети "ВКонтакте"». Файл данных: Antinata_vk_sentiments_preparing.csv?download=1/. 1.2 GB.

База данных является комплиментарной к опубликованной ранее базе 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. Краткий обзор литературы см. (Kalabikhina, Banin, 2020).

Методология сбора данных. В настоящем исследовании предпринята попытка апробации инструментария машинного обучения на текстовых данных, полученных из социальной сети «ВКонтакте». Проведен сбор неструктурированных текстовых данных из групп, осуществлена предобработка данных (очистка, лемматизация, стеммизация и удаление пунктуации), сформирован структурированный массив (корпус) текстов. На основе латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) были выявлены тематические кластеры. После проведения тематического анализа для каждого кластера проведена оценка тональностей текстов и построена динамика изменения тональности во времени для комментариев.

Тематической моделью является модель коллекции текстовых документов, которая определяет, к каким темам относится данный документ. Помимо выделения структуры текстовой коллекции, тематическое моделирование позволяет осуществлять смысловой поиск информации (в отличие от поиска по ключевым словам, где смысл явно не представлен).

Для анализа тональности использованы библиотеки TensorFlow и tflearn. Обучение нейросети осуществлено на размеченной базе данных коротких сообщений из твиттера [Рубцова, 2015]. Обучение нейросети произведено в среде Google Colab с использованием графического ускорителя (GPU, *graphics processing unit*). Для обучения нейросети использовано около 24 Гб оперативной памяти при размере обучающего словаря 5000 слов. Перед обучением данные проходили стеммизацию (приведение к базовой форме слова), все некириллические символы из выборки устранялись. Объем тестовой выборки составляет 30 % от всей выборки. Количество эпох для обучения — 30. Результирующая точность на обучающей выборке — 93,4 %, на тестовой — 69 %. Пороговое значение вероятности для отнесения комментария к позитивному или негативному равно 0,5.

Источники данных. Источник текстовых данных — тематические группы в социальной сети «ВКонтакте» (vk.com). На первом этапе обработки с помощью встроенного API (application programming interface) по ключевым словам («чайлдфри», «ребенок», «здоровье», «рождение», «родители» и др.) были собраны уникальные номера адресов тематических групп в виде *vk.com/<уникальный идентификатор группы>*. На первом этапе было собрано около 100 уникальных адресов групп с данными о количестве участников. На втором этапе из выборки были исключены группы, связанные с рекламой, а также группы с малой активностью участников (оценивалась общая динамика изменения количества постов, лайков и репостов) и количеством подписчиков меньше 500.

Информация о выборке

- В выборке содержится 8 групп (количество пользователей без учета самопересечений около 100 тысяч)
- Содержательный тип групп: группы, в которых пользователи делают преимущественно негативные замечания на темы, связанные с детством, материнством, беременностью и т.п. Несмотря на описанную процедуру отбора, в базе могут встречаться отдельные пользователи с пронаталистскими взглядами
- Исключены группы с числом подписчиков менее 500
- Собраны только комментарии с количеством лайков ≥ 5
- Комментарии собираются только по группам, в которых обсуждаются вопросы, связанные с детством, материнством, беременностью и т.п.
- Выборка содержит 670 тысяч пользовательских комментариев.

Список вошедших в выборку групп:

- 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

Структура выборки и описание переменных в базе:

- **link_author** — ссылка на автора комментария в виде https://vk.com/*author identificator*
- **gender of author —** пол автора комментария (F — женский, M — мужской, NaN — данные отсутствуют)
- **link_comment** — ссылка на комментарий в виде https://vk.com/* post identificatior on a *community wall*?reply=*comment id *
- **date_time** — дата и время публикации (формат “YYYY-MM-DD HH:MM:SS”)
- **text** — необработанный (исходный) текст комментария
- **likes** — количество лайков, поставленных под комментарием
- **text_prep** — текст комментария после предварительной обработки (удалены пунктуационные знаки, все слова переведены в строчное написание)
- **text_stem** — обработанный текст комментария (проведена стеммизация информации из колонки text_prep с использованием SnowBallstemmer (“Russian”) из библиотеки nltk)
- **text_sw** — обработанный текст комментария (проведено удаление стоп-слов из информации из колонки text_prep с ипользованием word_tokenize (text) из библиотеки nltk)
- **text_lemm** — обработанный текст комментария (проведена лемматизация информации из колонки text_prep с использованием mystem.lemmatize (text) из библиотеки pymystem3)

Области применения базы данных

Данные подходят для тематического анализа (например, LDA — Latent Dirichlet Allocation), для моделирования графовой структуры групп (переменная link_comment содержит уникальный идентификатор поста, link_author содержит уникальный идентификатор пользователя), для анализа тональностей высказываний и формирования словаря демографической коннотации на русском языке.

Анализ тональностей высказываний позволяет измерять динамику «демографической температуры» в антинаталистских группах. Под демографической температурой мы понимаем эмоциональный фон или преобладание позитивной или негативной тональности высказываний на темы, связанные с семейными ценностями, рождением детей и прочими темами в области репродуктивного поведения. Измеряется демографическая температура как разница (или отношение) между числом позитивных и числом негативных высказываний за определенный период времени.

Еще раз подчеркнем, что демографическая температура в данном случае измеряется в сообществах людей с антинаталистскими взглядами, то есть с репродуктивными установками на отрицание создания семьи и рождения детей.

Представленная база данных позволяет сравнивать демографическую температуру в отдельных кластерах групп в социальных сетях, изучать динамику позитивных и негативных комментариев женщин и мужчин на демографические темы в области рождения детей, родительства и семейных ценностей.

Первая публикация по измерению демографической температуры с использованием методологии измерения тональности высказываний в социальной сети «ВКонтакте» [Kalabikhina et al., 2021] основывается на двух базах данных: описываемой в этой статье и базе, описанной в [Kalabikhina, Banin, 2020]. Это первая попытка анализа тональности русскоязычных комментариев в социальной сети «ВКонтакте» для определения демографической температуры в различных социально-демографических группах в сети. В частности, используя данные с 2014 г., мы обнаружили асинхронные структурные сдвиги по периодам в корпусах пронаталистских и антинаталистских групп [Kalabikhina et al., 2021].

Вклад в создание и развитие базы данных: Идея и концепция создания базы данных на основе разработанного спектра направлений применения базы данных в демографическом анализе рождаемости, репродуктивного поведения, реакции населения на демографическую политику и другие факторы репродуктивного поведения — д.э.н. И.Е. Калабихина. Методика создания базы, создание первого варианта базы — Е.П. Банин. База создана в рамках реализации внутреннего гранта Экономического факультета МГУ имени М.В. Ломоносова. Авторы благодарят коллег по проекту за помощь в формулировке тематических слов и словосочетаний, в поиске образцов пронаталистских групп в социальной сети: И.А. Абдуселимову, В.Н. Архангельского, Г.В. Клименко, А.В. Колотушу, У.Г. Николаеву, В.Ш. Шамсутдинову.

Список литературы

Рубцова Ю.В. (2015) Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы 27: 72–8. https://doi.org/10.15827/0236-235X.109.072-078

Kalabikhina I.E., Banin E.P. (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 I.E., Banin E.P., Abduselimova I.A., Klimenko G.A., Kolotusha A.V. (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

Сведения об авторах

- Калабихина Ирина Евгеньевна, доктор экономических наук, профессор, заведующая кафедрой народонаселения экономического факультета МГУ имени М.В. Ломоносова, kalabikhina@econ.msu.ru
- Банин Евгений Петрович, кандидат технических наук, инженер-исследователь, Научно-исследовательский центр «Курчатовский институт», Московский государственный технический университет имени Н.Э. Баумана, evg.banin@gmail.com