PE Population and Economics

Население и экономика 4(3): 98–103
DOI 10.3897/popecon.4.e60915

СТАТЬЯ О БАЗЕ ДАННЫХ

База данных «Просемейные (пронаталистские) 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 5 November 2020 ♦ **Принято в печать** 20 November 2020 ♦ **Опубликовано** 18 December 2020

Цитирование: 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.e60915

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 просемейных (пронаталистских) группах.

Ключевые слова

База данных, большие данные, пронатализм, ВКонтакте, социальные сети, сообщества, семейные ценности

Коды JEL: J1, C31, C32, D71, E71

Доступ к данным и формат данных

Название базы данных: Просемейные (пронаталистские) 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 использо-

Copyright *И.Е. Калабихина, Е.П. Банин*. Статья 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.

ваться, распространяться и воспроизводиться на любых носителях при условии указания авторов и источника. Ирина Калабихина, Евгений Банин: Просемейные (пронаталистские) сообщества в социальной сети «ВКонтакте». Режим доступа: https://doi.org/10.5281/zenodo.4244361. Формат данных .csv (кодировка UTF-8). Описание: Данные могут быть скачаны с открытого источника (онлайн-депозитарий Zenodo), где размещена база данных Просемейные (пронаталистские) сообщества в социальной сети «ВКонтакте». Файл данных vk_posts_stem_lemm.scv 117.0 MB

Краткий обзор литературы и мотивация исследования. В русскоязычном сегменте известны два крупных проекта с разработанными словарями настроений: RuSentiLex [Loukachevitch and Levchik, 2016] и LINIS Crowd [Koltsova, Alexeeva and Kolcov, 2016]. Оба проекта представляют собой разработанные словари с оценкой тональности (от позитивной до негативной) для каждого слова или сочетания слов без характеристики эмоциональной окраски. Более сложные модели тональности предлагаются в проектах SentiWordNet (Baccianella, Esuli and Sebastiani, 2010) и SenticNet [Cambria et al., 2016, 2018], построенных на анализе тональности английского языка. В настоящее время наиболее активно развиваются методы глубокого обучения нейронных сетей, которые демонстрируют наилучшие [Tang, Qin and Liu, 2015; Tang and Zhang, 2018] на данный момент результаты оценки тональности по сравнению с остальными подходами. Для методов анализа тональности на основе машинного обучения характерна потребность в предобучении на больших наборах размеченных текстов. Известны попытки объединения двух представленных подходов (на основе правил и методов машинного обучения), например работы [Meškele and Frasincar, 2019; Kumar et al., 2020].

В обзорной работе по прикладному анализу тональностей [Smetanin, 2020] отмечено, что для русского языка данное направление исследовано мало (автор отмечает 27 наиболее релевантных исследований по анализу тональностей на русском языке). Большая часть исследований сосредоточена на анализе тональностей твиттов (коротких сообщений) в социальной сети Twitter. В России в 2019 г. этой социальной сетью пользовалось около 650 тысяч активных пользователей. Социальная сеть «ВКонтакте» обладает наибольшим охватом русскоязычного населения. Согласно отчету консалтинговой компании Deloitte [Zemlyanskaya et al., 2018], «ВКонтакте» охватывает до 70 % населения России.

Высокий охват населения предопределил выбор социальной сети «ВКонтакте» в качестве источника текстовых данных для анализа. Анализ различных источников позволил сделать вывод о недостаточной проработанности моделей оценки тональностей в русскоязычном сегменте в целом, причем даже существующие работы ставят ограниченные задачи и не переходят с уровня отдельных малых сообществ на региональный или страновой уровень. Наиболее сложными для анализа тональности являются демографические темы из области репродуктивного поведения (по сравнению с самосохранительным или миграционным поведением). Указанные обстоятельства привели к созданию нами базы данных в русскоязычном сегменте с использованием машинного обучения на текстовых данных из социальной сети «ВКонтакте» по темам в области репродуктивного поведения.

Методология сбора данных. В настоящем исследовании предпринята попытка апробации инструментария машинного обучения на текстовых данных, полученных из социальной сети «ВКонтакте». Проведён 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 (публикация по данной апробации модели в печати).

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 («мама», «мамочки», «дети», «ребенок», «малыш», «здоровье», «рождение», «беременность», «родители») были собраны уникальные номера адресов тематических групп в виде *vk.com/<уникальный идентификатор группы>*. На первом этапе было собрано около 1000 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 10 000.

Информация о выборке

- Собраны только комментарии с количеством лайков ≥ 5
- Комментарии собираются только по группам (список групп ниже), в которых обсуждаются вопросы, связанные с детством, материнством, беременностью и т.п.
- Выборка групп в среднем содержит 309 тысяч подписчиков (максимальное значение — 1 482 303, минимальное значение — 72 570, общее количество подписчиков без учета пересечений — 11 743 295)
- Выборка комментариев содержит 112 900 пользовательских комментариев

После формирования окончательного списка групп осуществлялся сбор текстовой информации из групп. В настоящей работе собираемые тексты ограничены только постами и комментариями. На основании собираемой информации формировался корпус языка: все слова приводились к нижнему регистру, стоп-слова удалялись при помощи функций из библиотеки nltk или gensim, удалялась пунктуация, исключались числовые данные. Для уменьшения объема текстовых данных дополнительно проводилась стеммизация (удаление окончаний слова) или лемматизация (приведение слова к начальной форме при помощи лемматизатора MyStem). Мы определили, что для оценки тональности в наибольшей степени подходит корпус комментариев.

Структура выборки и список основных групп: см. в описании базы данных в международном депозитарии https://doi.org/10.5281/zenodo.4244361.

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 просемейных (пронаталистских) 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 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 пронаталистскими взглядами, то есть с репродуктивными установками на создание семьи и рождение детей. На наш взгляд, измерение демографической температуры в таких сообществах в лучшей степени показывает состояние демографического климата, сформированного демографической и семейной политикой, экономическими тенденциями (но очищенного от влияния структуры населения по критерию пронаталистских или антинаталистских репродуктивных установок).

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.

В следующих версиях мы планируем добавить антинаталистские группы и большее число тональностей (помимо позитивных и негативных).

Пример измерения демографической температуры с использование описываемой базы данных

На рисунке 1 показано распределение комментариев по месяцам с начала 2014 года с учетом пола автора. Отметим, что наибольшую активность в представленных пронаталистских группах проявляют женщины. Всплеск и возрастающая активность наблюдаются с июня 2017 года.

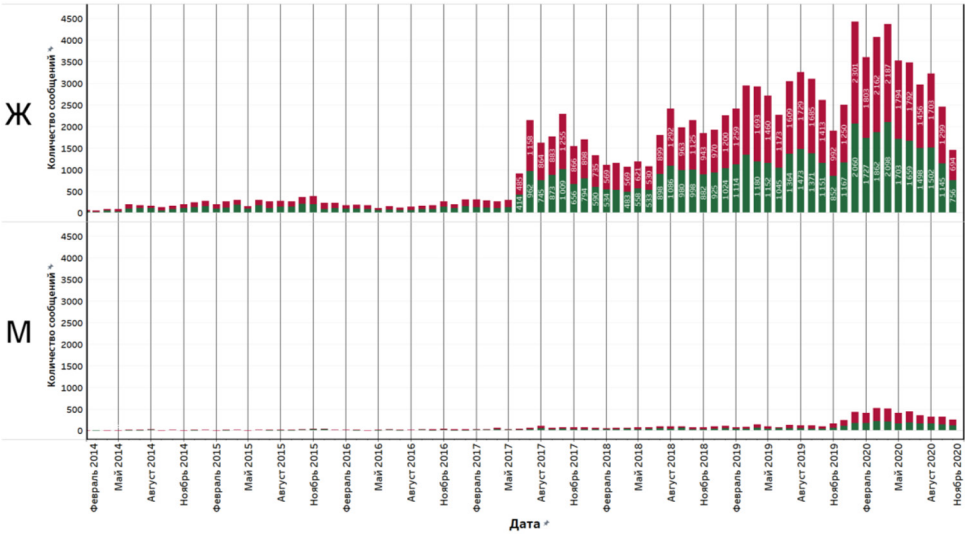

Рисунок 1. Распределение количества комментариев в корпусе по месяцам. «Красный» — негативные комментарии, «зеленый» — позитивные. *Источник*: Построено авторами в Tableau 19.3 на основе собранных данных из социальной сети «ВКонтакте»

Демографическая температура в исследуемых пронаталистских сообществах (разница между позитивными и негативными комментариями) представлена на рисунке 2. Можно заметить, что с конца 2016 года количество негативных комментариев существенно выросло, демографическая температура стала отрицательной. В новейшей истории России исторический максимум суммарного коэффициента рождаемости наблюдался в 2015 году — 1,78 детей на 1 женщину условного поколения, затем начался спад уровня рождаемости до 1,50 в 2019 году (и, по нашим оценкам, до 1,4 в 2020 году).

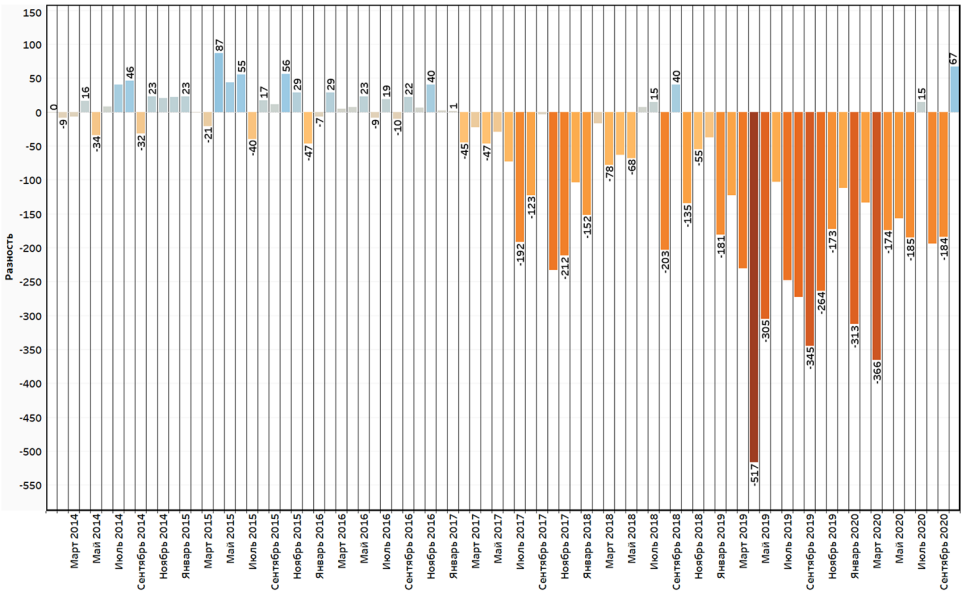

Рисунок 2. Разность негативных и позитивных комментариев в корпусе комментариев по месяцам. «-» — преобладание негативных комментариев. *Источник*: Построено авторами в Tableau 19.3 на основе собранных данных из социальной сети «ВКонтакте»

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. Калабихина И.Е. Методика создания базы, создание первого варианта базы — Банин Е.П. 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: Абдуселимову И.А., Архангельского В.Н., Клименко Г.В., Колотушу А.В., Николаеву У.Г., Шамсутдинову В.Ш.

Список литературы

Рубцова Ю.В. 2015. Построение корпуса текстов для настройки тонового классификатора // Международный журнал «Программные продукты и системы», 27: 72–78. doi: 10.15827/0236-235X.109.072-078.

Baccianella S., Esuli A., 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 O. Y., Alexeeva S. V., Kolcov S. N. 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, pp. 2489–2496. doi: 10.1145/3297280.3297525.

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.

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, инженер-исследователь, Научно-исследовательский центр «Курчатовский институт»; аспирант, факультет «Робототехника и комплексная автоматизация» МГТУ им. Н.Э. Баумана, evg.banin@gmail.com