Автоматический анализ репродуктивных ценностей пользователей сети ВКонтакте

И. Е. Калабихина [1], Н. В. Лукашевич [2], Е. П. Банин [3], К. В. Алибаева[4]

В работе исследуются мнения пользователей сети ВКонтакте о рождении детей. Собран датасет с разметкой мнений по трем классам позиций по отношению к шести темам, связанных с рождением детей и пронаталистской политикой. Выполнены эксперименты по автоматической классификации мнений. Лучшие результаты получены на основе применения нейросетевой модели BERT в формулировке NLI (Natural language Inference - вывод по тексту). Выявлено, что феномен сознательной бездетности активно представлен в сети, а многодетность остается слабо распространенной моделью поведения. В рамках пронаталистской политики важно формировать позитивное общественное мнение о родительстве, смягчать дефицит времени у родителей.

Ключевые слова: анализ мнений, классификация, BERT, ВКонтакте, репродуктивные ценности, пронаталистская политика.

1. Введение

Мнения пользователей социальных сетей по демографическим вопросам могут служить дополнительным источником информации в исследова-

[1] *Калабихина Ирина Евгеньевна* — заведующая кафедрой народонаселения, доктор экономических наук, экономический факультет МГУ имени М.В. Ломоносова, школа «Мозг, когнитивные системы, искусственный интеллект», e-mail: kalabikhina@econ.msu.ru.

Kalabikhina Irina Evgenievna — Head of the Department of Population, Doctor of Economic Sciences, Faculty of Economics, Lomonosov Moscow State University, School “Brain, cognitive systems, artificial intelligence".

[2] *Лукашевич Наталья Валентиновна* — ведущий научный сотрудник, доктор технических наук НИВЦ МГУ имени М.В. Ломоносова, школа «Мозг, когнитивные системы, искусственный интеллект» , e-mail: louk_nat@mail.ru.

Loukachevitch Natalia Valentinovna — leading researcher, Doctor of Technical Sciences, Research Computing Center, Lomonosov Moscow State University School “Brain, cognitive systems, artificial intelligence".

[3] *Банин Евгений Петрович* — инженер-исследователь. кандидат технических наук, Национальный исследовательский центр «Курчатовский институт», e-mail-mail: evg.banin@gmail.com.

Banin Eugene Petrovich — engineer-researcher, Phd, National Research Center “Kurchatov Institute".

[4] *Алибаева Камила Винеровна* — студентка фак-та ВМК МГУ имени М.В. Ломоносов, e-mail: camalibi@yandex.ru.

Alibaeva Kamila Vinerovna — student, Computer Science department, Lomonosov Moscow State University.

ниях и разработке социально-демографической политики. В данном исследовании рассматриваются подходы к автоматическому извлечению и анализу мнений пользователей сети ВКонтакте по вопросам отношения к (не)рождению детей и мерам государственной поддержки семей в области рождаемости. В данной работе мы представляем специализированный датасет, с разметкой мнений пользователей о репродуктивном поведении (DOI 10.5281/zenodo.5561126). Мы анализируем особенности распределение оценок «за» и «против» по конкретным аспектам репродуктивного поведения и отношения к пронаталистской политике. Созданный датасет используется для решения двух задач классификации: классификации сообщений по релевантности изучаемых тем и позиции автора по теме. Для классификации сообщений используются классические методы машинного обучения, а также нейросетевая модель BERT.

Сбор данных осуществлялся на основе групп ВКонтакте, в названиях или описаниях которых явно присутствовали слова чайлдфри и childfree и их вариации (так называемые антинаталисты), и группы, в названиях или описаниях которых присутствовали ключевые слова мама, мамочки, дети (пронаталисты). Базы данных текстов таких групп: DOI 10.3897/popecon.5.e70786; DOI 10.3897/popecon.5.e70786. Для анализа были выбраны 6 тем, отражающих отношение пользователей к (не)рож-дению детей и их оценку мер, предпринимаемых государством для повышения рождаемости. Высказывания по темам извлекались по соответствующим ключевым словам: «Аборты» (аборт), «Бездетность» (childfree, чайлдфри, бездетный, нет детей, без детей), «Многодетность» (многодетный, многодетность, много детей), «Индивидуализм» (в свое, эгоист, ответственность, для себя, личность, развиваться, эго), «Родительские отпуска» (декрет, отпуск), «Материнский капитал» (маткапитал, материнский капитал, выплаты, пособие). В дополнение к теме «Бездетность» отдельно разрабатывалась тема «Индивидуализм» (в контексте «пожить для себя»). Выбор темы «Индивидуализм» основан на гипотезе мотивации сознательной бездетности – изменении системы ценностей, увеличения набора жизненных траекторий, конкуренция ценностей самореализации и семейных ценностей, рост индивидуализации и приоритетов саморазвития.

Высказывания пользователей были разделены на предложения. Предложения из собранной выборки размечались тремя аннотаторами – профессиональными демографами и лингвистами. Поскольку в каждом предложении могли обсуждаться несколько вопросов, то аннотатор каждое предложение размечал по всем шести темам. Расхождения в разметке аннотаторами решались путем голосования. В результате разметки было размечено 5413 предложений по 6 темам с классами разметки: «нерелевантно», «за», «против», «прочее».

Собранные данные показали превалирование оценок «за» по теме «Аборты», что связано с с отношением населения к аборту как к приемлемому средству регулирования рождаемости. По теме «Бездетность» было выявлено превалирование позитивных оценок, что связано с тем, что сторонники этого паттерна достаточно эмоциональны, и тема активна в период роста представителей такой модели. Тем же можно объяснить и превалирование позитивных оценок в теме «Индивидуализм». Негативные оценки в отношении детских и семейных пособий и родительских отпусков связаны либо с неготовностью публики поддерживать родительство собственными ресурсами (налоги, рабочее время, хлопоты с сотрудниками-родителями), либо с персональными трудностями воспитания маленьких детей, дефицитом времени и запросом на большую помощь со стороны государства.

2. Сбор и разметка данных

В исследовании рассматривались две задачи классификации: классификация высказываний на релевантные/нерелевантные и классификация релевантных высказываний на три класса тональности позиций. Классификация сообщений по релевантности важна, поскольку сообщения извлекались не по хэштегам, а по ключевым словам, которые не всегда точно характеризуют тему сообщений. В качестве базовых моделей используются классические методы машинного обучения: наивный байесовский классификатор в двух вариантах мультиномиальный (MNB) и Бернулли (BNB), метод опорных векторов (SVC), Gradient Boosting (GB), Случайный лес (Random Forest).

В качестве основного метода использовалась нейросетевая модель BERT [1], в версии Conversational RuBERT, для создания которой использовалась русскоязычная модель RuBERT [2], которая была дообучена на русскоязычных диалогах и текстах социальных сетей. Использовалось три варианта обучения модели BERT: классификация целевого высказывания, а также так называемые NLI (Natural Language Inference – вывод по тексту) и QA (question-answering – вопросно-ответный) подходы (Sun et al., 2019). В NLI и QA подходах модель получала пары (текст, предположение). Для классификации релевантности NLI и классификации тональности QA этим предположением был сам аспект («Аборты», «Выплаты» и тд), для классификации тональности NLI предположение включало в себя еще и саму тональность («Негативно к абортам», «Нейтрально к выплатам» и т.д.). В работе использовались реализации классических методов машинного обучения из пакета scikit-learn. Обучение алгоритмов производилось на основе векторных представлений предложений с весами tf-idf. Для настройки параметров алгоритмов исполь-

зовалась процедура grid-search на валидационных данных. Лучшие результаты по разным темам для классических подходов были получены на основе методов: Байесовский классификатор Бернулли (BNB), метод опорных векторов (SVC) и Gradient Boosting (GB). Для оценки качества классификации используются меры: правильность классификации (Accuracy) и F-мера. Лучшие результаты классификации получены моделью BERT NLI [3], обученной на парах предложений, обучение выполнено на полной выборке для каждой темы. Следующие по качеству результаты получены моделью NLIsingle, которая для каждой темы обучалась только на обучающих данных своей темы – 66.15 Accuracy, 65.58 Macro F-меры.

3. Заключение

В данной работе представлены результаты анализа высказываний пользователей сети ВКонтакте по тематике деторождения и отношения к пронаталистской политике. Собран новый датасет, размеченный по отношению пользователей к шести темам. На собранных данных были исследованы две задачи, существенные для анализа мнений в социальных сетях в реальном времени: классификация высказываний по релевантности и классификаций релевантных мнений по тональности позиции. В обеих задачах лучшие результаты получены на основе варианта модели NLI BERT, на вход которой данные подаются в виде двух предложений, и обучение для классификации по конкретным темам производится на всем объеме обучающих данных. Выявлено, что феномен сознательной бездетности активно представлен в сети, а многодетность остается слабо распространенной моделью поведения. В контексте полученных данных можно сделать вывод о некоторых рекомендациях в отношении пронаталистской политики: формировать позитивное общественное мнение о родительстве и родителях, смягчать дефицит времени у родителей.

Благодарности

Работа выполнена в рамках НИР «Воспроизводство населения в контексте социально-экономического развития»: «Паттерны репродуктивного поведения россиян (на основе тематического анализа текстов в социальных сетях)».

Список литературы

[1] Devlin J. et al., “BERT: Pre-training of deep bidirectional transformers for language understanding”, *NAACL HLT 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **1** (2019), 4171—4186.

[2] Kuratov Y., Arkhipov M., “Adaptation of deep bidirectional multilingual transformers for Russian language”, *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, 2019, № 18, 333—339.

[3] Sun C., Huang L., Qiu X., “Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence”, *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. Association for Computational Linguistics (ACL)*, **1** (2019), 380—385.

Automated Analysis of Family Values of VKontakte Users
Kalabikhina I.E., Loukachevitch N.V., Banin E.P., Alibaeva K.V.

The paper examines opinions of VKontakte network users about having births. A dataset annotated with opinions on three classes of positions towards six childbirth-related topics. Experiments on automatic classification of opinions have been carried out. The best results were obtained using neural network model BERT in the formulation of NLI (Natural language Inference). It was revealed that the phenomenon of conscious childlessness is actively represented in the network, while having many children remains a poorly widespread model of behavior. Within the framework of a pro-natalist policy, it is important to support a positive public opinion about parenting, to develop family-life balance for parents.

Keywords: stance detection, classification, BERT, VKontakte, reproductive values, pro-natalist policy.

References

[1] Devlin J. et al., “BERT: Pre-training of deep bidirectional transformers for language understanding”, *NAACL HLT 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **1** (2019), 4171—4186.

[2] Y. Kuratov Y., Arkhipov M., “Adaptation of deep bidirectional multilingual transformers for Russian language”, *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, 2019, № 18, 333—339.

[3] Sun C., Huang L., Qiu X., "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence", *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. Association for Computational Linguistics (ACL)*, **1** (2019), 380—385.