УДК 81.33

DOI 10.25205/1818-7935-2019-17-3-71-82

Использование текстов жанра «интернет-откровение» в контексте решения задач сентимент-анализа

А. В. Колмогорова

*Сибирский федеральный университет
Красноярск, Россия*

Аннотация

Статья посвящена анализу валидности текстов жанра «интернет-откровение» в качестве источника данных для обучающей выборки в целях тренировки модели компьютерного классификатора интернет-текстов на русском языке по критерию их эмоциональной тональности. Целью осуществляемого проекта является создание программы, способной автоматически оценивать тексты как вербализующие одну из восьми эмоций (модель Г. Левхейма) или как эмоционально нейтральные. Для достижения цели возникла необходимость в формировании обучающей выборки – коллекции интернет-текстов, где каждому из них уже приписана определенная эмоция или оценка «нейтрально». В качестве источника подобных данных выбраны тексты из публичной группы «Подслушано» в социальной сети «ВКонтакте». Всем текстам публика присущи специфические черты, позволяющие рассматривать их совокупность как отдельный нарративный жанр интернет-откровения, свойства которого описаны с применением метода нарративной семиотики А. Ж. Греймаса. Выявлены качественные и формальные преимущества текстов данного жанра в качестве источника данных для сентимент-анализа.

В качестве ограничения, накладываемого текстами жанра «интернет-откровение» на выборку, выступает их нарративный характер, что исключает из коллекции данных иные типы дискурсивных единиц, например, аргументативы или дескриптивы.

Ключевые слова

сентимент-анализ, обучающая выборка, интернет-тексты, жанр «интернет-откровение», социальные сети, нарратив

Благодарности

Исследовательская группа благодарит Российский Фонд Фундаментальных Исследований за финансовую поддержку проекта (проект № 19-012-00205 «Разработка классификатора русскоязычных интернет-текстов по критерию их тональности на основе модели эмоций “Куб Левхейма”»)

Для цитирования

Колмогорова А. В. Использование текстов жанра «интернет-откровение» в контексте решения задач сентимент-анализа // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2019. Т. 17, № 3. С. 71–82. DOI 10.25205/1818-7935-2019-17-3-71-82

Texts of “Internet Confessions” as a Source for Training Data Set for the Research on the Sentiment-Analysis Field

Anastasia V. Kolmogorova

*Siberian Federal University
Krasnoyarsk, Russian Federation*

Abstract

The article aims to analyze the validity of Internet confession texts used as a source of training data set for designing computer classifier of Internet texts in Russian according to their emotional tonality. Thus, the classifier, backed by Lövheim’s emotional cube model, is expected to detect eight classes of emotions represented in the text or to assign the text to the emotionally neutral class. The first and one of the most important stages of the classifier creation is the training data set selection. The training data set in Machine Learning is the actual dataset used to train the model for

© А. В. Колмогорова, 2019

ISSN 1818-7935

Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2019. Том 17, № 3
Vestnik NSU. Series: Linguistics and Intercultural Communication, 2019, vol. 17, no. 3

performing various actions. The internet text genres that are traditionally used in sentiment analysis to train two or three tonalities classifiers are tweets, films and market reviews, blogs and financial reports. The novelty of our project consists in designing multiclass classifier that requires a new non-trivial training data. As such, we have chosen the texts from public group Overheard in Russian social network VKontakte. As all texts show similarities, we united them under the genre name "Internet confession". To feature the genre, we applied the method of narrative semiotics describing six positions forming the deep narrative structure of "Internet confession": Addresser – a person aware of her/his separateness from the society; Addressee – society / public opinion; Subject – a narrator describing his / her emotional state; Object – the person's self-image; Helper – the person's frankness; Adversary – the person's shame. The above mentioned genre features determine its primary advantage – a qualitative one – to be especially focused on the emotionality while more traditional sources of textual data are based on such categories as expressivity (tweets) or axiological estimations (all sorts of reviews).

The structural analysis of texts under discussion has also demonstrated several advantages due to the technological basis of the Overheard project: the text hashtagging prevents the researcher from submitting the whole collection to the crowdsourcing assessment; its size is optimal for assessment by experts; despite their hyperbolized emotionality, the texts of Internet confession genre share the stylistic features typical of different types of personal internet discourse.

However, the narrative character of all Internet confession texts implies some restrictions in their use within sentiment analysis project.

Keywords

sentiment analysis, training data set, Internet texts, Internet confession genre, social networks, narratives

Acknowledgements

The research is supported by the Russian Foundation for Basic Research, project № 19-012-00205 "Design of sentiment classifier for Internet-texts in Russian backed by Lovheim's Cube emotional model"

For citation

Kolmogorova, Anastasia V. Texts of "Internet Confessions" as a Source for Training Data Set for the Research on the Sentiment-Analysis Field. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2019, vol. 17, no. 3, p. 71–82. (in Russ.) DOI 10.25205/1818-7935-2019-17-3-71-82

Введение

В современном глобальном мире общение в социальных сетях стало не только нормой, но и доминирующей формой коммуникации. В сети обсуждаются политические новости, события, хобби, подробности личной жизни звезд, но, кроме того, в виртуальное пространство выплескивается и субстанция психической и эмоциональной жизни реальной личности. Социальные медиа гибко приспосабливаются к актуальным потребностям – появляются новые интернет-жанры.

В центре внимания данной публикации один из таких жанров – нарративный жанр интернет-откровений, появившийся несколько лет назад благодаря новому медийному формату, предложенному создателем публичной группы «Подслушано» в русскоязычной социальной сети «ВКонтакте».

Предметом анализа являются характеристики текстов данного жанра, обеспечивающие эффективность их использования в качестве источника данных для тренировки алгоритма машинного обучения по прецедентам в контексте задач сентимент-анализа русскоязычных интернет-текстов.

Основной исследовательский вопрос статьи может быть сформулирован следующим образом: каковы преимущества и ограничения данных, полученных из коллекции текстов подобных интернет-откровений, по сравнению с другими источниками, для решения задачи определения дискриминантных черт текстов, вербализующих разные эмоции?

Для ответа на данный вопрос мы последовательно охарактеризуем специфику осуществляемого исследовательской группой проекта в области сентимент-анализа, жанр интернет-откровений в целом, выделим его свойства, позволяющие эффективно использовать тексты данного жанра для целей сентимент-анализа, а также некоторые ограничения материала, сравнив его с другими типами данных, традиционно используемых для решения обсуждаемых задач.

Обсуждение проблемы

Краткая характеристика проекта: цель, теоретическая основа и методология

Цель исследовательского проекта – разработка компьютерной программы, на вход которой подается интернет-текст на русском языке, а на выходе пользователь получает оценку эмоциональной тональности, доминирующей в тексте. Например, этот текст – грустный, а тот – вербализует эмоцию злости.

Актуальность такой технологии в том, что она позволяет автоматически обрабатывать большие массивы текстов для мониторинга общественного мнения, проведения маркетинговых кампаний и оценки успешности, а также для выявления случаев эмоционального насилия, как, например, на шумевшая история с группами суицидальной направленности, подписчиками которых являются дети.

Ведущим методом проекта является сентимент-анализ текстов. Это выявление эмоциональной тональности текста при помощи методов обработки естественного языка, статистики, машинного обучения. Впервые этот термин был использован в статьях S. R. Das и M. Y. Chen [2001], B. Pang, L. Lee и Sh. Vaithyanathan [2002].

Теоретической основой для выделения репертуара эмоциональных тональностей послужила классификация эмоций Г. Левхейма, визуализированная ученым в виде куба – так называемый «куб Левхейма». Шведский нейрофизиолог разработал модель корреляции базовых эмоций с комбинацией уровней трех моноаминов – серотонина, дофамина и норадреналина [Lövhim, 2012. С. 342], визуализировав ее в виде куба на координатной плоскости с осями 5-HT (серотонин), NE (норадреналин), DA (дофамин). В зависимости от сочетания уровня данных гормонов в крови субъекта эмоции исследователь выделяет восемь эмоций, где первая номинация класса отражает наименее выраженную степень интенсивности эмоции-аффекта, а вторая – ее высшую точку: Интерес / Возбуждение; Удовольствие / Радость; Удивление; Страдание / Тоска; Гнев / Ярость; Страх / Ужас; Презрение / Отвращение; Стыд / Унижение.

В качестве метода для разработки программы-классификатора выбрана технология машинного обучения по прецедентам, основной принцип которой – по частным данным, представляющим набор пар «объект, ответ», выявить закономерности, присущие не только конкретной обучающей выборке, но и генеральной совокупности данных.

Для подобной технологии важнейшим этапом является формирование обучающей выборки – коллекции текстов, где каждому из них уже приписан эмоциональный класс: этот текст вербализует эмоцию радости, а другой, например, отмечен страданием. Данная выборка служит основой для «обучения» алгоритма, который, обрабатывая размеченные данные, строит статистически релевантные модели, устанавливает зависимости. Но при создании выборки обычно возникают два ключевых вопроса: 1) из какого источника брать тексты? и 2) кто и как их будет аннотировать? Ответы на них были найдены благодаря постам из публичной группы «Подслушано» «ВКонтакте», написанным в жанре интернет-откровений.

Общая характеристика жанра интернет-откровений

Поскольку жанр интернет-откровений является в определенном смысле порождением публика «Подслушано», охарактеризуем контент публичной группы согласно модели, предложенной в [Гуськова, Левина, 2016].

1. *Количество участников группы, социальный портрет участников группы.* По словам создателя группы, дающего интервью под псевдонимом Владимир Огурцов, у группы 9 млн пользователей в месяц и 150 тысяч активных пользователей в сутки¹; по его словам, аудитория у «Подслушано» разная: «нам пишут как мужчины, так и женщины всех возрастов.

¹ <https://vc.ru/story/23626-podslushano>

Можно встретить секреты студентов, рабочих, пенсионеров, богатых и бедных людей – словом, читателей самых разных социальных прослоек» (МК, 02.02.2015, № 19).

2. *Краткое описание группы, созданное ее администрацией.* На приветственном экране мобильного приложения, с помощью которого подписчики могут отправлять модераторам группы свои истории, значится: «Мы – социальный развлекательный проект, в котором люди каждый день анонимно делятся своими секретами, откровениями и жизненными ситуациями перед огромной аудиторией»².

3. *Контент стены группы* составляют эмоционально окрашенные нарративы от первого лица, в которых авторы анонимно («для того, чтобы открыть душу, не боясь социального прессинга», – отмечает В. Огурцов (МК, 02.02.2015, №19)) делятся самыми интимными личностными переживаниями, вопреки традициям сохранения лица.

4. *Активность участников группы* при наполнении контента группы и комментировании информации на стене группы. Архитектура группы предполагает сложное взаимодействие и взаимопереплетение активностей участников: одни участники загружают свои истории в приложение, другие, зарекомендовавшие себя в качестве «продвинутых пользователей», отбирают те из них, которые они хотели бы увидеть на стене группы, и уже эти истории попадают «в руки» редакторов; они выбирают самые интересные, не затрагивающие табуированные темы истории, проверяют их подлинность и публикуют под определенным тематическим хештегом; наконец, опубликованные после редакторской селекции анонимные нарративы подписчики могут комментировать.

5. *Тематика контента*, создаваемого участниками группы разнообразна и отражена в хештегах группы: смешное; пошлое; страшное; бабушки; ненависть; «паштет»; странное; похоть; мистика; чернуха; стыдно; зависть; жестокость; изнанка; лень; одиночество; коты; бесит; детство; мечты; семья; наблюдения; сны; алчность; работа; лайфхак; фууу; добро; хомьяки; счастье; дружба; предательство; хобби; провал; любовь; общение; пьянь и т. д.

6. *Языковые особенности контента*, созданного участниками группы, и комментариев. Тексты написаны, как правило, в рамках разговорного письменно-речевого стиля, сформировавшегося в интернет-среде, однако нередки и включения из обшечной лексики.

7. *Ценностные ориентиры участников группы*, создающих контент на стене группы и комментирующих информацию. Хотя на заре развития сетей нередко высказывались гипотезы о грядущем расщепления идентичности пользователей на «сетевую» и «реальную», современные исследователи фиксируют совершенно иную тенденцию – стремление личности к достоверности и самовыражению как в объективной реальности, так и в интернет-пространстве [Евсюкова, Шилова, 2015]; именно такое стремление, глубоко фундированное в ценности приятия противоречивости реальной человеческой личности, включая те ее грани, которые традиция сохранения социального лица запрещала демонстрировать, объединяет подписчиков группы.

Впрочем, нельзя не отметить, что стремление к предельной откровенности признаний зачастую приближается к другой крайности, которая, на наш взгляд, удачно была определена в [Пожидаева, Карамалак, 2018] как «киберэкзгибиционизм».

Для того чтобы сделать краткое описание жанра интернет-откровений мы воспользовались методологией нарративной семиотики [Greimas, 1987], поскольку все истории основаны на принципах нарративного дискурса.

А. Ж. Греймас предлагает использовать для описания глубинной структуры нарратива – фундаментальной системы ценностей, встроенной в текст, – систему из шести актантов, которые не обязательно являются акторами: а) адресант – определенная сила, которая устанавливает правила и ценности в действии и представляет идеологию текста; б) адресат – несет ценности (а), поэтому имеет отношение к объекту, на который адресант распространяет цен-

² https://app.ideer.ru/?referrer=appmetrica_tracking_id%3D530071196532780788%26ym_tracking_id%3D18326465027784365313

ности; в) субъект – выполняет главную роль в повествовании; г) объект – цель, на которую направлен интерес субъекта; д) помощник – поддерживающая сила, которая помогает субъекту в его стараниях, направленных на объект; е) противник – препятствующая сила, которая олицетворяет все, что удерживает субъекта от достижения цели [Титчер, Мейер и др., 2017. С. 177].

Если описать специфику каждого актанта в нарративном жанре интернет-откровения, то получится следующая система ролей.

Адресантом в интернет-откровениях является, по большому счету человеческая личность, охваченная чувством отчужденности, по Э. Фромму [Fromm, 1956. С. 23], которое «является источником всех тревог. <...> Быть отдельным от всех означает быть беспомощным, неспособным активно контактировать с миром <...>; это значит, что мир может подмять меня, лишенного способности сопротивляться» [Там же]. Это человек, который хочет рассказать о самом сокровенном – иногда отвратительном, пугающем – переживании, чтобы, разделив его с другими, как бы влиться в коллективное человеческое «Я», преодолев губительное чувство отчужденности. Как пишут редакторы паблика, суть «Подслушано» – «это не любопытство, а потребность почувствовать себя не одиноким, ощутить то самое единение с незнакомцем, который читает в метро твою любимую книгу» [Люди слишком верят...]. Это самое «коллективное Я» и заполняет актантную позицию адресата, обобщенного носителя противоречивой ценностной установки – с одной стороны, это стремление принять и понять, что все мы разные, уникальные, но, с другой – это и одновременная приверженность идеологии конформизма, спрятанная в тайной надежде, что и у других такое было: «если я такой же, как все, если у меня нет мыслей и чувств, которые бы отличали меня от других <...>, то я спасен от пугающего одиночества» [Fromm, 1956. Р. 29].

Субъектом в анализируемых нарративах выступает «я» рассказчика – он переживает определенное эмоциональное состояние, спровоцированное собственным ощущением жизни (1) либо поведением других в этой жизни (2):

(1) Работаю в очень известном заведении, бармейд, отлично выгляжу, общительная, начитанная, начальство, подчиненные, гости восхищаются мной, поклонников много (с работы, естественно). Но стоит мне выйти с работы, превращаюсь в абсолютного мизантропа. Отвергаю все свидания. Сажу дома, пью и смотрю фильмы, сериалы. Любое общение вне работы отвергаю. Мне уже 27. Ни разу не было серьезных отношений. Мне одиноко и страшно. Очень одиноко.

(2) Устроилась в фирму, где директором работала моя мать. Решила не афишировать, что являюсь её дочерью, хотя постепенно все узнали. Директор она строгий и требовательный, что многим работникам не нравится. Работу свою выполняю хорошо, на уровне. И как же меня бесит, что люди, выше меня должностью, срываются на мне, если у них с моей мамой произошёл конфликт. Торкают, принижают, пытаются как-то задеть. Я понимаю, пытаются отыграться, но бесит неимоверно! Думаю переводиться в другой филиал. Достали.

Объектом же является образ самого себя, получившийся в результате отражения «в глазах» других людей: я честно рассказал вам о том, что я чувствую; кто я после этого по-вашему? М. М. Бахтин писал: «Я сижу всегда на двух стульях. Я строю свой образ (осознаю себя) одновременно и из себя, и с точки зрения другого» [Бахтин, 2000. С. 236].

Заполнителем актантной позиции помощника – некоторой силы, помогающей субъекту в его усилиях – является откровенность. Именно она оправдывает порой неприглядность демонстрируемой изнанки личностного «я» рассказчика, как бы говорящего: да, я, может быть, смешон или безобразен в своих чувствах и поступках, но я ничего от вас не скрываю, я показываю все как есть. Противоборствующей силой, противником, как правило, выступает чувство стыда. Кто-то находит в себе силы этот стыд артикулировать (*мучает совесть, очень*

стыдно (3)), а кто-то пытается камуфлировать его оправданиями (*нечем кормить ребенка, проблемы на работе, счета за квартиру* (4)):

(3) *Мой папа сидел в тюрьме за убийство почти 15 лет, я живу с мамой. Она всегда боялась, что я вырасту такой же агрессивной, как он. Я считала это бредом, до того момента, когда я избила её за то, что она, придя с работы в плохом настроении, накричала на меня. Теперь мучает совесть, не знаю как это произошло... очень стыдно;*

(4) *Продала поддельную брошь Шанель по цене настоящей...Подделка качественная, не дешёвая, но подделка все-таки. Проблемы на работе, ребёнка кормить нечем, а тут ещё за квартиру счета пришли. Знаю, что поступила неправильно, совесть мучает уже несколько дней, но понимаю, что собственный ребёнок дороже, чем та фифа в розовом мини, которая купила ту самую брошь. Ненавижу себя.*

Таким образом, нарративный жанр интернет-откровений – это сложное взаимодействие личностного «я» индивида и коллективного человеческого «я», разворачивающееся в поле притяжения двух полюсов – откровения и стыда, результатом которого становится воссоздаваемый индивидом образ себя, достойный самопритятия. Данный жанр принадлежит жанрам персонального интернет-дискурса [Пожидаева, Карамалак, 2018].

*Преимущества и ограничения текстов жанра интернет-откровений
как источника данных для обучающей выборки
в контексте задач по сентимент-анализу*

Прежде всего, стоит отметить, что большинство кейсов, выполненных с использованием технологий сентимент-анализа, решают задачу двух- или трехчастной классификации текстов на позитивно-оценочные, негативно-оценочные и нейтральные [Liu et al., 2007; Hogenboom et al., 2015], часто объединяемые в группу *opinion mining technologies* [Pang, Lee, 2008]. Подчеркнем, что наш проект преследует цель создать алгоритм для классификации текстов согласно схеме «8 эмоциональных классов + 1 нейтральный», поэтому дискуссионным остается вопрос о том, насколько адекватны для данной цели традиционные для *opinion mining* источники данных: твиты [Bollen et al., 2011], рецензии интернет-пользователей на фильмы, товары, посты в блогах [Boiy, Marie-Francine, 2009], финансовые отчеты [Ghose, Ipeirotis, 2007].

Все перечисленные тексты так или иначе находятся в поле сопряжения трех лингвистических категорий: экспрессивности, оценочности и эмотивности. Первая определяется как «такое свойство текста или части текста, которое передает смысл с увеличенной интенсивностью, выражая внутреннее состояние говорящего» [Арнольд, 1975. С. 5]; оценочность трактуется как семантическая категория, реализующаяся в рамках аксиологических полюсов «хорошо – плохо» [Вольф, 1985], а эмотивность – как языковое выражение эмоциональности при помощи различных средств языка [Ленько, 2015. С. 86]. Если мы расположим различные жанровые разновидности текстов-источников данных для сентимент-анализа между тремя этими категориями, то получим следующую визуализацию (см. рисунок).

Если в новостных текстах и финансовых отчетах доминирует оценочность, а экспрессивность и эмотивность практически не выражены ввиду ограничений, накладываемых самими жанрами, то тексты маркет- и кинорецензий с необходимостью включают в себя языковые средства оценки, комбинируемые, как правило, со средствами выражения экспрессивности и эмоционально окрашенными языковыми единицами. Формат микроблоггинга в Твиттере предполагает создание достаточно короткого экспрессивного текста-реакции на какое-либо событие политической, социальной или личной жизни пользователя. И хотя твиты нередко бывают эмоционально окрашенными, эта эмоциональность всегда подчинена экспрессии. Для выражения собственно эмоций необходимо более обширное текстовое пространство, которое как раз предполагается форматами жанров интернет-откровений и постов в блогах.

Распределение жанров интернет-текстов между категориями экспрессивности,
оценочности и эмотивности

Different Internet text genres bias towards categories of expressivity,
appraisal and emotivity

С размером текста в жанре интернет-откровения связано и одно из его формальных преимуществ для использования в качестве источника данных в целях сентимент-анализа: чтобы точно оценить эмоцию, вербализованную в том или ином фрагменте текста, эксперту, занимающемуся аннотированием обучающей выборки, необходимо, чтобы этот фрагмент не был ни слишком большим (тогда велика вероятность размывания эмоции, ее смешения с другой эмоцией), ни слишком маленьким – в последнем случае контекст будет недостаточным для идентификации эмоции. В пилотном исследовании подобного рода на материале художественной прозы [Колмогорова, 2018] нами опытным путем был определен оптимальный объем текстового фрагмента для аннотирования – 60–80 слов. Истории в жанре интернет-откровений, публикуемые в «Подслушано», именно такого размера.

Другим преимуществом является то, что истории в жанре «интернет-откровение» публикуются уже после их разметки при помощи хештега редакторами паблика. Поскольку ведущей интенцией текстов интернет-откровений является желание искренне рассказать о пережитом эмоциональном состоянии, получив тем самым ощущение единения с другими, а хештег отражает коммуникативную цель интернет-публикации [Патрушева, 2018], редакторы, приписывая тот или иной хештег тексту, по сути дела, фиксируют эмоцию, о которой хочет рассказать данный автор, выступая, не подозревая об этом, в качестве ассессоров. Обычно это пользователи краудсорсинговых платформ, за небольшое вознаграждение оценивающие тексты в соответствии с теми критериями, которые задают разработчики классификатора.

Путем проб и ошибок нам удалось соотнести ряд хештегов с нужными нам 8 эмоциональными классами текстов (см. таблицу). Отметим также, что жанр интернет-откровений предполагает не-нейтральную оценку автора к высказываемым фактам, поэтому в настоящий момент в нашей обучающей выборке собраны только тексты, размеченные в соответствии с 8 классами эмоций по Левхейму. В дальнейшем в работе в качестве нейтрального корпуса планируется использовать внешний корпус нейтральных текстов – например, корпус новостей Яндекса.

Объем подкорпусов и их соотношение с хештегами
Emotional subcorpora size and its hashtagging

Эмоциональный класс текстов (подкорпус)	Объем подкорпуса в токенах	Хештег в «Подслушано»
Страдание / Тоска	56 470	#Подслушано_одиночество
Интерес / Возбуждение	184 074	#Подслушано_успех
Удовольствие / Радость	85 117	#Подслушано_счастье
Страх / Ужас	230 730	#Подслушано_страшное
Брезгливость/Отвращение	45 868	#Подслушано_фууу
Злость / Гнев	131 564	#Подслушано_БЕСИТ
Стыд / Унижение	70 232	#Подслушано_стыдно
Удивление	288 272	#Подслушано_наблюдения #Подслушано_странное

Из таблицы следует, что ряд хештегов прямо указывают на эмоциональное состояние: #Подслушано_счастье (я хочу рассказать вам о той ситуации, когда я испытывал чувство счастья (5)), #Подслушано_страшное (...когда мне было страшно (6)), Подслушано_стыдно (...когда мне было стыдно (7)):

(5) *Переехала в Питер, случайно в баре встретила с очень известным исполнителем, и всё — искра! Я понимаю, что недолговечно, что таких как я, вероятно, миллион. Но когда он в Питере, то пропадает у меня в квартире. Недавно сказал: «Люблю просыпаться здесь не меньше, чем засыпать...» Счастлива;*

(6) *Как-то ночью проснулась от детского смеха в стороне балкона. Отчётливый и то приближается, то отдаляется. Причем злоеющий такой. Светит луна, занавеска колыхается, ребёнок злорадно смеётся и напевает. Кот лежит на моём плече и смотрит огромными глазами на балкон... Собрал кишки в кулак, иду к балкону. Оказалось, это форточка открылась от ветра, колыхая занавески. Какая-то пьянь наматывала круги под моими окнами, ожидая такси, и у неё бесконечно звонил телефон с этим самым смехом и песенкой;*

(7) *Когда училась в школе, жили вдвоём с мамой. Она всегда готовила, придерживалась мнения, что есть надо дома, что это полезно. Я же после школы наедалась всякого говна с друзьями и, приходя домой, выливая суп. Мамы не стало полгода назад, но перед глазами она, уставшая, стоящая у плиты после работы, варящая мне суп, чтобы я ела здоровую пищу. И я, сливающая ее труд в туалет, выбрасывающая котлеты в мусоропровод. Сердце разрывается, не понимаю, как я могла быть такой ублюдкой, стыдно до трясушки.*

Другие хештеги указывают на эмоцию опосредованно – через типичные для ситуации переживания этой эмоции слово, реплику, тематическую референцию. Например, #Подслушано_фууу – междометие *фуу* является вербальным проявлением эмоции отвращения; #Подслушано_БЕСИТ – реплика *бесит* стала в последнее время общепринятым «зачином» для выражения злости, возмущения. Есть такие сюжеты, темы, которые устойчиво ассоциируются в социальной реальности с той или иной эмоцией, например, успех – это радость, удовольствие (#Подслушано_успех), а одиночество неминуемо связано с тоской, депрессией, грустью (#Подслушано_одиночество). Следует признать, что наибольшие трудности вызвало выявление хештега, коррелирующего с эмоцией удивления по Г. Левхейму. В итоге были отобраны два хештега: #Подслушано_наблюдения и #Подслушано_странное. Однако в процессе рандомизированного предъявления группе ассессоров тексты именно из этих двух тематических групп вызвали наибольший разброс оценок, что привело к необходимости по-

вторного аннотирования ассессорами всей коллекции историй под данными двумя хештегами.

Следующее преимущество текстов интернет-откровений состоит в том, что, обладая наибольшей выраженностью категории эмотивности по сравнению с традиционными источниками данных для тренировки алгоритмов сентимент-анализа, они сохраняют стилистические черты письменно-разговорного интернет-дискурса: предпочтительность вводных (*мол, ...*), неполных (*Сижу, плачу*), параллельных (*Ты покупаешь себе модную одежду...; Ты идешь в пустую квартиру...; Ты клянешься себе никогда больше не любить...*) и эллиптических конструкций (*Я простила, он – нет*), обилие повторов (*Мой муж ничего не хочет делать со мной вместе. На какие-то курсы – нет. В спортзал – нет. На выставку – нет*), вопросительных (*Я понимаю, что фиксированных остановок нет, я ещё могу понять остановки через улицу, но блин, просить остановить «за поворотом», когда только что отъехали от остановки «перед поворотом»?*) и восклицательных предложений (*Надоели суицидники в метро! Почему бы не сделать это, например, дома? Работники метро и сотни пассажиров не должны страдать из-за чьих-то неудач!)* [Горошко, 2008. С. 391], частотность сокращений, разговорных, просторечных ЛЕ (*Тётка 50+, которой каждую смену звонят какие-то хмыри с сайтов знакомств, и она им по полчаса (!) орёт, что на работе и разговаривать не может... Твою ж мать, если не можешь разговаривать, сделай отбой и работай*).

Наконец, нельзя не упомянуть и об одном недостатке текстов обсуждаемого жанра, который, как это часто бывает, является продолжением их достоинств – все они представляют собой примеры исключительно нарративного дискурса. В сформированной таким образом обучающей выборке нет текстов-дескриптивов или текстов-аргументативов. Окажет ли данный фактор негативное влияние на точность модели, обученной на такой выборке, покажет практика.

Заключение

Проведя анализ специфики жанра интернет-откровений на материале публикаций в публичке «Подслушано» в социальной сети «ВКонтакте», мы можем ответить на поставленный в начале статьи вопрос о преимуществах и ограничениях подобных текстов в качестве источника данных для обучающей выборки в целях сентимент-анализа, следующим образом.

Данный жанр имеет преимущества: а) качественные: тот факт, что жанровая роль адресанта в данных текстах выполняет человеческая личность, пытающаяся преодолеть чувство отчужденности, а ее адресатом становится общечеловеческое «я» коллектива, обуславливает тяготение таких текстов к реализации именно категории эмотивности, а не экспрессивности или оценочности, которые вторичны; б) формальные: 1) тексты уже подверглись первичной эмоциональной разметке при помощи хештегов, что облегчает их извлечение и позволяет избежать затратной в финансовом и временном плане процедуры разметки ассессорами для значительной части данных; 2) размер текстов оптимален для восприятия экспертами или ассессорами; 3) при выраженной категории эмотивности тексты сохраняют лексическую, синтаксическую и графическую специфику, свойственную в целом персональному интернет-дискурсу социальных сетей.

Ограничением является нарративный характер всех текстов жанра «интернет-откровение», исключающий из обучающей выборки иные типы дискурсивных единиц по критерию их дискурсивной организации.

Сейчас ведется расширение обучающей выборки для тренировки многоклассового классификатора русскоязычных интернет-текстов по их эмоциональной тональности за счет текстов данного жанра. В настоящий момент максимальный показатель *weighted average f1-score*, представляющий собой взвешенное по доле каждого класса гармоническое среднее значений точности и полноты классификации, приближается к 50 % (0.50). Ведется поиск

оптимальных дискриминантных черт текстов различных эмоциональных классов, подаваемых «на вход» классификатору.

Список литературы

- Арнольд И. В.** Интерпретация художественного текста: типы высказываний и проблема экспрессивности // Экспрессивные средства английского языка. Л., 1975. С. 5–9.
- Бахтин М. М.** Автор и герой: К философским основам гуманитарных наук. СПб.: Азбука, 2000. 336 с.
- Вольф Е. М.** Функциональная семантика оценки. М.: Наука, 1985. 256 с.
- Горошко Е. И.** Гендерные аспекты коммуникаций на примере образовательных практик Интернета // *Educational Technology & Society*. 2008. Т. 11 (2). С. 388–411.
- Гуськова С. В., Левина В. Н.** Особенности общения в открытых группах в социальной сети «ВКонтакте» как отражение ценностных ориентров молодежной аудитории // Вестник Волгogr. гос. ун-та. Сер. 2, Языкозн. 2016. Т. 15, № 3. С. 66–77.
- Евсюкова Т. В., Шилова Е. С.** Проблема самоидентификации языковой личности в социальных сетях // Вестник МГОУ. Серия: Русская филология. 2015. № 6. С. 8–15.
- Колмогорова А. В.** Вербальные маркеры эмоций в контексте решения задач сентимент-анализа // Вопросы когнитивной лингвистики. 2018. № 1. С. 83–93.
- Ленько Г. Н.** Анализ категории эмотивности и смежных с ней понятий // Вестник ЛГУ им. А. С. Пушкина. 2015. № 1. С. 84–91.
- Патрушева Л. С.** Хештегирование как новый коммуникативный процесс // Вестник Удмуртского университета. Серия История и Филология. 2018. Т. 28, вып. 3. С. 471–475.
- Пожидаева Е. В., Карамалак О. А.** Хэштеги в социальных сетях: интенции и аффордансы (на примере группы сообщений на английском языке по теме «Food» (Пища / еда)) // Вестник Томского государственного университета. Филология. 2018. № 55. С. 106–118. DOI 10.17223/19986645/55/8
- Титчер С., Мейер М., Водак Р., Веттер Е.** Методы анализа текста и дискурса / Пер. с нем. Харьков: Гуманитарный Центр, 2017. 356 с.
- Boiy E., Marie-Francine M.** A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts. *Inf. Retrieval*, 2009, no. 12 (5), p. 526–558.
- Bollen J., Mao H., Zeng X.** Twitter Mood Predicts the Stock Market. *Journal of Computational Science*, 2011, no. 2 (1), p. 1–8.
- Das S., Chen M.** Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards. In: Proc. of the Asia Pacific Finance Association Annual Conference (APFA), 2001, p. 1–16.
- Fromm E.** Art of Loving. New York, 1956, 133 p.
- Ghose A., Ipeirotis P. G.** Designing Novel Review Ranking Systems: Predicting the Usefulness and Impact of Reviews. In: Proc. of the 9th International Conference on Electronic Commerce, 2007, p. 303–310.
- Greimas A. J.** De l'Imperfection. Périgueux, P. Fanlac, 1987, 102 p. (in Fr.)
- Hogenboom A., Frasincar F., Jong F., Kaymak U.** Polarity Classification Using Structure-Based Vector Representations of Text. *Decis. Support Syst.*, 2015, no. 74, p. 46–56.
- Liu Y., Huang X., An A., Yu X.** ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs. Proc. of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007, p. 607–614.
- Lövheim H.** A New Three-Dimensional Model for Emotions and Monoamine Neurotransmitters. *Medical Hypotheses*, 2012, no. 78, p. 341–348.
- Pang B., Lee L.** Opinion Mining and Sentiment Analysis. *Foundations and Trends in Inf. Retrieval*, 2008, no. 2 (1–2), p. 1–135.

Pang B., Lee L., Vaithyanathan Sh. Thumbs up? Sentiment Classification Using Machine Learning Techniques. Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002, p. 79–86.

Список источников примеров

«Люди слишком верят в свою уникальность»: легко ли быть редактором «Подслушано». URL: <https://daily.afisha.ru/relationship/7644-lyudi-slishkom-veryat-v-svoyu-unikalnost-legko-li-byt-redaktorom-podslushano/> (дата обращения 22.02.2019).
Подслушано ВК. URL: <https://vk.com/overhear> (дата обращения 22.02.2019).

References

- Arnold I. V.** Fiction Text Interpretation: Types of Utterances and the Problem of Expressivity. In: Expressive Means in English. Leningrad, 1975, p. 5–9. (in Russ.)
- Bakhtin M. M.** Author and Hero: Towards the Philosophical Basis of Humanities. St. Petersburg, Azbuka, 2000, 336 p. (in Russ.)
- Boiy E., Marie-Francine M.** A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts. *Inf. Retrieval*, 2009, no. 12 (5), p. 526–558.
- Bollen J., Mao H., Zeng X.** Twitter Mood Predicts the Stock Market. *Journal of Computational Science*, 2011, no. 2 (1), p. 1–8.
- Das S., Chen M.** Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards. In: Proc. of the Asia Pacific Finance Association Annual Conference (APFA), 2001, p. 1–16.
- Evsyukova T. V., Shilova E. S.** The Problem of Self-identification of the Linguistic Personality in Social Networks. *Bulletin MRSU. Issue: Russian Philology*, 2015, no. 6, p. 8–15. (in Russ.)
- Fromm E.** Art of Loving. New York, 1956, 133 p.
- Ghose A., Ipeirotis P. G.** Designing Novel Review Ranking Systems: Predicting the Usefulness and Impact of Reviews. In: Proc. of the 9th International Conference on Electronic Commerce, 2007, p. 303–310.
- Goroshko E. I.** Gender Aspects of Communication (as Based on the Educational Practices on the Internet). *Educational Technology & Society*, 2008, vol. 11 (2), p. 388–411. (in Russ.)
- Greimas A. J.** De l'imperfection. Périgueux, P. Fanlac, 1987, 102 p. (in Fr.)
- Guskova S. V., Levina V. N.** Features of Communication in the Open Groups of the VKontakte Social Network as a Reflection of Value Orientations of Young Audience. *Bulletin of Volgograd State University. Issue 2, Linguistics*, 2016, vol. 15, no. 3, p. 66–77. (in Russ.)
- Hogenboom A., Frasincar F., Jong F., Kaymak U.** Polarity Classification Using Structure-Based Vector Representations of Text. *Decis. Support Syst.*, 2015, no. 74, p. 46–56.
- Kolmogorova A. V.** Verbal Markers of Lövheim's Eight Emotional Classes in Sentiment Analysis Researches. *Issues of Cognitive Linguistics*, 2018, no. 1, p. 83–93. (in Russ.)
- Lenko G. N.** Analysis of Emotivity and Related Concepts. *Bulletin of LGU Named after A. S. Pushkin*, 2015, no. 1, p. 84–91. (in Russ.)
- Liu Y., Huang X., An A., Yu X.** ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs. Proc. of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007, p. 607–614.
- Lövheim H.** A New Three-Dimensional Model for Emotions and Monoamine Neurotransmitters. *Medical Hypotheses*, 2012, no. 78, p. 341–348.
- Pang B., Lee L.** Opinion Mining and Sentiment Analysis. *Foundations and Trends in Inf. Retrieval*, 2008, no. 2 (1-2), p. 1–135.
- Pang B., Lee L., Vaithyanathan Sh.** Thumbs up? Sentiment Classification Using Machine Learning Techniques. Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002, p. 79–86.

- Patrusheva L. S.** Hashtagging as a New Process of Communication. *Bulletin of Udmurt University. Issue: History and Philology*, 2018, vol. 28, no. 3, p. 471–475. (in Russ.)
- Pojidaeva E. V., Karamalak O. A.** Hashtags in Social Networks: Intentions and Affordances (Exemplified in the English Language by Message Groups on the Topic “Food”). *Bulletin of TSU. Philology*, 2018, no. 55, p. 106–118. DOI 10.17223/19986645/55/8 (in Russ.)
- Titscher S., Meyer M., Wodak R., Vetter E.** *Method of Text and Discourse Analysis*. Westdeutscher Verlag GmbH, Wiesbaden, 1998, 356 p. (in Russ.)
- Wolf E. M.** *Functional Semantics of Appraisal*. Moscow, Nauka, 1985. 256 p. (in Russ.)

Sources of Examples

«People believe too much in their uniqueness»: is it easy to be editor in Overheard group? URL: <https://daily.afisha.ru/relationship/7644-lyudi-slishkom-veryat-v-svoyu-unikalnost-legko-li-byt-redaktorom-podslushano/> (date of access 22.02.2019).
Overheard VK. URL: <https://vk.com/overhear> (date of access: 22.02.2019).

Материал поступил в редколлегию
Date of submission
15.03.2019

Сведения об авторе / Information about the Author

Колмогорова Анастасия Владимировна, доктор филологических наук, профессор, зав. кафедрой романских языков и прикладной лингвистики, Сибирский федеральный университет (пр. Свободный, 82а, Красноярск, 660041, Россия)

Anastasia V. Kolmogorova, Doctor of Philology, Full Professor, Head of Romance Languages and Applied Linguistics Department, Siberian Federal University (82A Svobodny Ave., Krasnoyarsk, 660041, Russian Federation)
nastiakol@mail.ru
ORCID 0000-0002-6425-2050