Вестник Томского государственного университета. Филология. 2022. № 79. С. 131–154
Tomsk State University Journal of Philology. 2022. 79. pp. 131–154

Научная статья
УДК 81’322
doi: 10.17223/19986645/79/7

Местоимения в автоматической жанровой и гендерной атрибуции текстов

Андрей Александрович Степаненко[1], Зоя Ивановна Резанова[2]

[1, 2] Национальный исследовательский Томский государственный университет, Томск, Россия
[1] stepanenkone@mail.ru
[2] rezanovazi@mail.ru

Аннотация. Представлены результаты применения статистических методов и методов машинного обучения в решении задачи жанровой и гендерной автоматической атрибуции текстов с использования в качестве языковых маркеров форм личных местоимений *я*, *ты*, *мы*. Результаты анализа показали, что при решении задач автоматической классификации текстов по признаку гендерной принадлежности автора текста необходимо учитывать жанровую форму текста, так как в силу жанровых особенностей языковые средства выражения интенций могут влиять на частоту использования личных местоимений.

Ключевые слова: автоматическая атрибуция текста, личные местоимения, гендер, жанр, устная диалогическая речь, жанр интервью, социальная сеть ВКонтакте, стена ВКонтакте, диалоги ВКонтакте

Источник финансирования: исследование выполнено при поддержке Программы развития Томского государственного университета (Приоритет-2030).

Для цитирования: Степаненко А.А., Резанова З.И. Местоимения в автоматической жанровой и гендерной атрибуции текстов // Вестник Томского государственного университета. Филология. 2022. № 79. С. 131–154. doi: 10.17223/19986645/79/7

Original article
doi: 10.17223/19986645/79/7

Pronouns as machine learning markers in genre and gender text attribution

Andrei A. Stepanenko[1], Zoya I. Rezanova[2]

[1, 2] Tomsk State University, Tomsk, Russian Federation
[1] stepanenkone@mail.ru
[2] rezanovazi@mail.ru

Abstract. This article focuses on the statistical and machine learning approaches in sloving the tasks of automatic genre and gender text attribution using the forms of personal pronouns *ya* [I], *ty* [you], *my* [we] as language markers. The research materi-

al was texts of discourses with a specific feature: spontaneous informal speech. Computer-mediated communication was represented by texts of several genres of the social network VKontakte: (a) personal written dialogues between men and women from VKontakte. The full text size was 114 046 words. The number of dialogue participants was 38 people (19 men, 19 women) aged 18 to 20; (b) 287 walls of the social network VKontakte with 9 951 001 words. All the participants were students of Tomsk State University. Transcriptions of oral texts were extracted from the RuTuBiC database of the Russian speech corpus of Turkic-Russian bilinguals. The number of respondents was 138 people. The texts consist of 617 846 words. During the investigation, the authors used such methods as correlation analysis, the method of generalized linear models (GLM), criteria for testing statistical hypotheses (Wilcoxon, Kruskal-Wallis test), and machine learning. The analysis was implemented in the programming language R 4.0.5 using the quanteda library. The analysis was carried out in two stages: (1) the diagnostic power of the pronouns in the tasks of classification by genre forms; (2) gender opposition within genre forms of the texts. The authors proved the dependence of using groups of pronouns on the genre form of the text. The machine learning methods showed effectiveness of models using formal metrics and confirmed a significant degree of similarity in the use of pronouns in the texts of the VKontakte wall and VKontakte dialogues. These two types of communication are opposed to the genre forms of oral public communication. The studied groups of pronouns are better used for the classification of the genre of the text than for gender attribution. Gender differentiation is confirmed only in the texts of the VKontakte wall genre. The result of the "full dataset" is a classification within two genres (VKontakte dialogues combined with VKontakte Walls) and oral public communication. There is an actual significant increase in the accuracy of the classifier, which indicates the similarity of these two genres and their opposition to oral public communication in the binary classification. The results of the analysis show problems of automatic text classification based on the gender of the text's author. It is necessary to pay attention to the genre form of the text. Such differences can be explained by genre features. Linguistic means of expressing intentions can affect the frequency of personal pronouns in the texts.

Keywords: automatic text attribution, personal pronouns, gender, genre, oral dialogue speech, interview genre, VKontakte social network, VKontakte walls, VKontakte dialogues

Financial Support: The study was supported by the Tomsk State University Development Programme (Priority 2030).

For citation: Stepanenko, A.A. & Rezanova, Z.I. (2022) Pronouns as machine learning markers in genre and gender text attribution. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology.* 79. pp. 131–154. (In Russian). doi: 10.17223/19986645/79/7

Введение

Гендерные исследования, имея весьма длительную традицию развития, начиная с первой половины XX в., представлены практически во всех гуманитарных науках, в том числе и лингвистике. В современном языкознании исследования гендера как социокультурного феномена, соотнесенного с биологическим различием мужского и женского в человеке, представлены несколькими сложившимися направлениями. На основании соотноше-

ния частной предметной сферы направления противопоставляются: 1) исследования гендерных различий, отраженных в структурах языков: в грамматике – наличие грамматических категорий рода, в лексике – лексическая разработанность семантики «мужского» и «женского» и под.; 2) типы гендерных стереотипов, отраженных во фразеологии, лексике языков; 3) различия речевого поведения мужчин и женщин, проявляющиеся в выборе и использовании языковых единиц разных уровней.

В данной статье излагаются результаты анализа, выполненного в рамках третьего направления, в котором далее выделяются значимые для обоснования представленного подхода теоретико-методоло-гические «развилки». Охарактеризуем их.

Во-первых, при исследовании речи, речевого поведения авторы сосредоточиваются либо на общей характеристике их особенностей, рассматривая явления разных уровней языковой системы как обобщенно характеризующие мужскую и женскую речь «вообще», без привязки к определенным типам коммуникации, как, например, в [1; 2; 3. С. 216, 4; 5; 6; 7. С. 11–15], либо данная противопоставленность исследуется в пределах определенных дискурсов или жанровых форм [8–14].

В настоящее время фокусировка внимания на отдельных типах дискурсивных практик как поля выявления гендерных различий видится более продуктивной. Данный подход мотивирован теоретическими взглядами на гендер как социокультурную характеристику субъекта коммуникации, которая не может не вступать во взаимодействие с другими параметрами дискурсивных практик [15]. Сам тип дискурса может оказывать влияние на проявленность гендерного своеобразия субъекта коммуникации, так как в институциональных дискурсах выбор языковых средств и их комбинация, коммуникативные стратегии и т.д. в значительной мере находятся под влиянием социально детерминированных стандартов коммуникации. В личностной, неинституциональной коммуникации, как было неоднократно отмечено, говорящий человек более непосредственно проявляет индивидуальность (см., например, широко известные в российской гендерной лингвистике положения о противопоставленности двух типов дискурсов в работах В.В. Карасика [16]), что может служить основанием для более непосредственной актуализации женской или мужской стратегий развертывания речи. В последнее время наряду с анализом устной обыденной коммуникации интенсивно развиваются исследования непосредственной личностной коммуникации на материале текстов компьютерно опосредствованного общения, рождаемых в частной переписке, в социальных сетях (СС) [17–21]. При этом и устная, и компьютерно опосредствованная личностная коммуникация имеют широкую жанровую палитру, что также проявляется и в характере использования языковых единиц, и в вариантах интеракций с многочисленными факторами дикурсообразования, в том числе гендерными.

В данной статье мы обращаемся к анализу двух базовых вариантов коммуникации: устного непосредственного общения, представленного

комплексными жанровыми формами устных интервью и бесед, и жанров компьютерно опосредствованного общения – личной переписки и текстов стены социальных сетей.

Во-вторых, исследователи при обращении к проблеме языковых маркеров гендера в речи либо стремятся выделить особенности, проявляющиеся на всех уровнях языковой системы, либо обращаются к характеристике одного или какой-либо группы признаков. Мы отмечали ранее значительное различие в подходе к выделению маркеров гендерных различий в собственно лингвистических исследованиях и исследованиях, выполненных с использованием методов автоматической обработки текстов [22]. В собственно лингвистических исследованиях обычно выделяются отдельные языковые единицы, являющие собой результат выбора из синонимических рядов (лексических, деривационных, морфолого-синтаксических), количественное преобладание которых формирует варьирование смыслового развития текста. Применительно к русскому языку, начиная с работы Е.А. Земской и ее коллег, накоплены знания о маркерах речевого поведения, противопоставляющие мужскую и женскую речь, в числе которых, как правило, выделяются степень и типы эмоционального фона коммуникации, степень конкретности представления тождественных тем, варианты личностной, коммуникативной фокусировки общения и под. Стоит отметить, что, когда лингвисты пишут о различиях в использовании языковых средств мужчинами и женщинами, речь идет о количественном преобладании, а не абсолютном отсутствии каких-либо единиц.

Автоматический анализ текста позволяет выделять признаки морфолого-синтаксической структуры: различия в использовании грамматических классов слов, n-граммы символов, знаки препинания, длина предложений и т.д.

В работах с использованием методов автоматического анализа текстов в целеполагание авторов включается проверка степени устойчивости выявляемого языкового признака, степени статистической релевантности полученных выводов, на основе чего ставится вопрос о возможности опоры на данный выделяемый признак в решении задач автоматической гендерной классификации текстов. Актуальность этого направления гендерных исследований определяется наличием социального заказа на создание методов автоматического определения авторства текста, субъект которого намеренно скрывается, прежде всего в криминалистической практике [23, 24] (см. также обзор в [22]).

Использование личных местоимений как маркеров гендерных различий, являющееся предметом нашего исследования, отмечено в качестве дифференцирующего фактора в двух охарактеризованных выше направлениях, чему способствует, с одной стороны, глубокая отрефлексированность личных местоимений как коммуникативно актуальных единиц. С другой стороны, эффективность автоматического анализа функциональных позиций местоимений в коммуникации обеспечивается тем, что они последовательно формально маркируются. Как следствие, исследователь может использовать ресурс существующих автоматических морфологических анализа-

торов (для текстов русского языка это прежде всего морфологический анализатор Mystem), не прибегая к технике предварительной первичной ручной разметки. В лингвистических исследованиях местоимений учеными было доказано, что этот класс лексических единиц служит одним из значимых средств актуализации позиции говорящего по отношению к другим коммуникантам (см. работы Pennebaker J.W. и др., Е.М. Вольф [25. P. 563–565; 26. С. 356112]). Было отмечено, что местоимения имеют, во-первых, собственное внеконтекстное постоянное значение и, во-вторых, контекстуальное значение, определяемое дейктической функцией. Учет типов ведущей референции местоимений, используемых говорящим в том или ином дискурсе наряду с другими словообразовательными, лексическими и синтаксическими средствами, может свидетельствовать о коммуникативных установках говорящих: об эгоцентричной или партнерской направленности. Наиболее ярким маркером выражения эгоцентризма в речи является соотношение местоимения *я* и его производных форм и форм местоимений *ты, вы*.

В работах по гендерной лингвистике отмечено различие в использовании местоимений мужчинами и женщинами. Так, B.C. Verhoeven доказывает на материале мультилингвальных электронных корпусов текстов, что женщины используют местоимение *я* чаще, чем мужчины [27. P. 1632–1633]. Подобное исследование проводилось А.Н. Барановым на материале художественных текстов [28]. В проведенном нами ранее исследовании использования местоимений и маркеров экспрессивности в текстах компьютерной коммуникации сделаны выводы о взаимодействии гендерного фактора с другими социально значимыми параметрами компьютерной коммуникации – темой текста, ролевыми и социальными позициями коммуникантов в диалоге [29].

Функциональная направленность местоимений на маркирование позиций говорящего и его отношений с коммуникантами обусловливает различие актуализации разных классов местоимений в различных дискурсивных практиках. Особенно значимым видится различие в институциональных и личностных дискурсах. Так, например, в системе русскоязычной научной коммуникации обозначение я-позиции предписывается замещать мы-позицией, представлять результаты исследования в системе безличных обозначений. Личностная коммуникация, напротив, наиболее открыта и более свободна в выражении межличностных отношений, значительную роль в которых играют личные местоимения, однако значительные жанровые различия также могут обусловить вариантность в использовании местоимений в данном типе дискурса.

Представленное в статье исследование выполнено на основе применения методов автоматической обработки текстов.

Наша гипотеза заключался в том, что использование местоимений в текстах обыденной коммуникации может быть маркером гендерных различий, однако диагностирующая сила использования данного признака в задачах автоматической атрибуции находится в зависимости от жанровых форм.

Цель проведенного исследования состояла в выявлении влияния жанровых различий коммуникации на диагностирующую силу личных местоимений в задачах автоматической атрибуции текстов и наличия корреляций между жанровым и гендерным признаками в решении данной задачи.

Материал и методы исследования

Материалом исследования послужили тексты дискурсов, интегральной чертой которых является непосредственность, спонтанность речи, ее преимущественно личностная ориентированность, которая, как было отмечено нами ранее, имеет разные детерминации в условиях коммуникации в среде социальной сети, созданной в качестве площадки для личностного общения, и в устной непосредственной коммуникации, протекающей в условиях контактирования говорящего и слушающего, меняющего спонтанно свои ролевые позиции в диалогах.

I. Компьютерно-опосредованная коммуникация представлена текстовыми материалами СС ВКонтакте, анализировались тексты двух жанровых форм:

а) тексты диалогов личной коммуникации между мужчинами и женщинами в СС ВКонтакте. Материалы были собраны в рамках учебной практики студентов ТГУ, объем материала – 114 046 слов, тексты личных сообщений 38 человек (19 мужчин, 19 женщин) в возрасте 18–20 лет;

б) текстовые материалы стен СС ВКонтакте – 9 951 001 слово, тексты 287 стен СС ВКонтакте учащихся первых и вторых курсов разных факультетов ТГУ.

СС ВКонтакте исследователями в жанровом аспекте интерпретируется как гипержанр, персональная страница – как наджанровое макрообразование, включающее жанры «анкета», «статус», «записи на стене», «личные сообщения», «обсуждения», «комментарии» [17. С. 24], «типичными стилевыми чертами» которых «являются эмоциональность, субъективность и имитация разговорной спонтанности при помощи экспрессивно-окрашенной лексики, разговорного синтаксиса, звукового письма, эмотиконов и экспрессивной пунктуации» [17. С. 23]. В литературе в качестве новой жанровой формы, реализованной в СС «ВКонтакте», отмечается «статус», который наряду с комментарием «фиксирует коммуникативные установки пользователя» и «ориентирован не только на поддержание контакта, но и на активизацию диалога» [18. С. 6]. Авторы данных работ не анализируют местоимения в составе коммуникативно актуальных средств, но мы полагаем, что их значимость вытекает из выявленной коммуникативной характеристики – диалогичности.

2. Транскрипции текстов устной речи. Транскрибированные тексты устной речи извлечены из базы данных корпуса русской речи тюркско-русских билингвов RuTuBiC, созданного в рамках проекта «Языковое и культурное своеобразие Южной Сибири: взаимодействие языков и культур» (описание корпуса см. [30]). Текстовые материалы корпуса записи

диалогической устной речи в жанровых формах интервью, разговора и беседы, различающихся степенью институциональной стандартизации: тематика диалога в интервью определена заранее, однако в диалогическом разведывании речи коммуниканты могут в большей или меньшей степени спонтанно изменять тематическую направленность общения и интервью может «перетекать» в разговор на не определенные заранее темы. Речевые жанры интервью, разговора и беседы объединяет наряду с политематичностью их диалогическая природа. В собственно лингвистических (работы по русской разговорной речи [31, 32] и др.) и психолингвистических исследованиях [33] отмечаются коммуникативные, психолингвистические особенности данного типа коммуникации и, как следствие, поверхностно-текстуальные, среди которых отмечается и высокая степень частотности использования местоимений [33. С. 261–262; 34. С. 71].

В корпусе представлена русская речь билингвов, важнейшим признаком его является абсолютное функциональное доминирование русского языка, при котором родной язык находится в состоянии утраты и вытесняется в сферу домашнего и семейного общения. Все респонденты получили образование на русском языке, что дает основание использовать данный материал недифференцированно в этом аспекте по отношению к материалу текстов компьютерно опосредстовованной коммуникации. В анализ включено 138 текстов респондентов. Объем текстового массива данных составляет 617 846 слов.

Анализируемые записи стен СС ВКонтакте относятся к открытым данным, записи личной коммуникации получены при условии личного согласия информантов (перед сбором диалогов в СС и записью устных интервью и бесед респонденты заполняли «Форму информационного согласия» и были проинформированы о ФЗ-152 РФ «О персональных данных».

Методика анализа

Для дальнейшей статистической обработки все текстовые материалы были разделены на файлы, содержащие мужские и женские реплики по 49-50 Кб каждая.

Массивы текстовых данных подверглись предобработке, которая включала в себя: 1) токенизацию – выделение лексических единиц в массиве символов; 2) лемматизацию – приведение всех словоформ к единой (начальной) форме при помощи программы «Mystem 3.0»; 3) перевод слов в единый (нижний) регистр; 4) удаление знаков пунктуации; 5) разделение на реплики на основании гендерной принадлежности респондента. В текстах транскриптов устной речи корпуса RuTuBiC была проведена дополнительная корректировка: удалены реплики интервьюера и техническая информация при помощи регулярных выражений; 6) формирование частотной матрицы относительных величин. В качестве нормализации абсолютных величин была использована формула условной вероятности, которая позволяет минимизировать влияние объема текстов на результаты статистического анализа и машинного обучения.

Все действия были реализованы в языке программирования R 4.0.5 и библиотеки quanteda. Были применены следующие частные методы – корреляционный анализ, метод обобщенных линейных моделей (GLM) критерии проверки статистических гипотез (критерий Уилкоксона, Краскела–Уоллиса), методы машинного обучения.

Результаты анализа

Так как мы исследовали влияние жанровых и гендерных различий коммуникации на диагностирующую силу личных местоимений в автоматической атрибуции текстов и наличие корреляций между жанровым и гендерным признаками в решении данной задачи, анализ проводился в два этапа. С применением методов статистического анализа и машинного обучения на первом этапе была проанализирована диагностическая сила местоимений в задачах классификации по жанровым формам, на втором этапе в пределах жанровых форм – по гендерной оппозиции.

I. Автоматическая жанровая атрибуция текстов.

На первом этапе выявлялось соотношение жанровых форм и частоты использования личных местоимений, маркирующих личностные позиции.

Мы сфокусировались на противопоставлении эгоцентрической позиции (маркирование я) vs кооперативной (в двух вариантах – обращение к собеседнику vs мы-позиция (кооперации субъекта с группой). Был проведен статистический анализ использования трех местоимений в совокупности их словоизменительных форм (далее: «Группа-я»; «Группа-ты», «Группа-мы») в трех жанровых вариантах личностной коммуникации (далее: Диалоги ВК, Стены ВК, устная публичная коммуникация) без дифференциации по гендерному признаку.

Сначала было проведено сравнение относительных частот использования местоимений методом векторизации BagOfWords.

Относительная частота местоимений (событий) A определяется как отношение NA / N, где N – число повторений местоимений, а NA – число тех повторений, в которых осуществилось событие A (повторение местоимений в группе). В итоге значение относительных местоимений в группах приобретает значения от 0 до 1. Данный подход нормализации позволяет минимизировать влияние несбалансированности корпуса.

Как можно видеть на рис. 1, наибольшее различие относительных частот местоимений наблюдается между исследуемыми жанрами компьютерно опосредствованной коммуникации и материалами записей бесед и интервью, что, на наш взгляд, определяется в значительной мере дискурсивным и жанровым своеобразием записей текстов устной коммуникации. Текстовые данные корпуса RuTuBic были получены в ходе направленных формализованных и полуформализованных интервью и бесед. Беседы и интервью имели личностную направленность, однако интервьюеры выступали в институциональной позиции. Интервьюерами были как члены лаборатории с широким возрастным диапазоном, так и студенты. Возраст-

ной, гендерный статус, уровень образования также варьировались в значительной степени. Вследствие этого местоимения группы *ты* замещались вежливой формой *вы* в большинстве случаев. (Примененные в работе способы выборки единиц на данном этапе не позволяют разграничить дискурсивные варианты значения местоимения *вы* – форму вежливого обращения к одному лицу – этикетный эквивалент *ты* и обозначение группы лиц, поэтому частотность данного местоимения в работе не подсчитывалась). Во всех жанровых формах количественно преобладают я-формы, однако в Стенах ВК и в Диалогах ВК второе место по частотности употребления занимает местоимение *ты*, маркирующее включение собеседника в смысловые поля диалога.

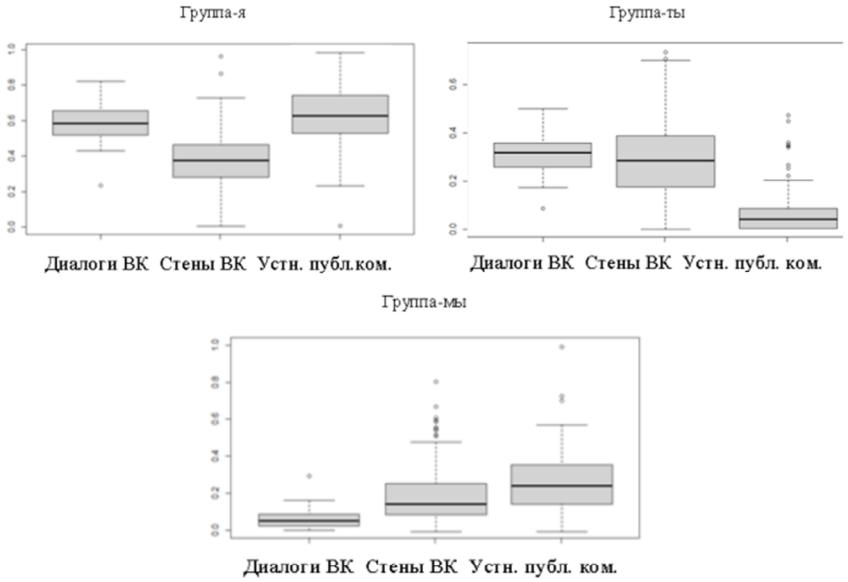

Рис. 1. Диаграмма размаха относительных частот групп местоимений

Диаграммы размаха показывают медиану, нижний и верхний квартили, минимальное и максимальное значение частот распределения групп местоимений и выбросы. Как видно из диаграммы, существует ряд отличий в частоте использования местоимений в большинстве групп. Местоимения «Группы-я» чаще используются в Диалогах ВК и устной коммуникации, меньше всего – текстах Стен ВК, жанровая форма которых более ориентирует на представление позиций «другого», с которым солидаризуется автор, что маркируется самим фактом расположения информации на странице. В устной публичной коммуникации преобладает «Группа-я», остальные группы местоимений используются значительно меньше, чем в других вариантах жанровых форм. В персонализированной личностной коммуникации говорящий склонен выступать от своего имени, но не от имени группы, что наиболее ярко проявлено в текстах собственно спонтанных неограниченных институциональными рамками диалогов. Мы можем объ-

яснить большее количество форм *мы* в текстах интервью и бесед устной коммуникации тематикой и условиями сбора материала, так как актуализация приобщения к групповому опыту стимулировалась вопросами интервьюеров.

Относительно Диалогов ВК и Стен ВК следует отметить, что в них преобладает использование местоимений «Группы-я» и «Группы-ты».

Однако было бы некорректно принимать нашу гипотезу о выявленных различиях, основываясь только на визуализации данных, так как различия могут быть случайными и не иметь статистически значимых показателей. Поэтому считаем необходимым применить статистические методы, выявляющие различия использования групп местоимений в разных жанровых формах текстов. Далее представим результаты: а) корреляционного анализа, выявляющего взаимосвязи частот группы местоимений, а также сопутствующую ему диаграмму рассеяния, которая используется для демонстрации наличия или отсутствия корреляции между переменными; б) применения статистических критериев проверки гипотез, выявляющего достоверность различий в генеральных совокупностях (в группах местоимений, жанрах, гендерной принадлежности автора); в) применения метода обобщенных линейных моделей, позволяющего учитывать взаимодействие между факторами, вид распределения зависимой переменной и предположения о характере регрессионной зависимости.

Так как распределение относительных частот не соответствуют гауссовскому критерию нормальности распределения в качестве критерия оценивания корреляций были выбраны непараметрические (ранговые) статистические критерии. Для выявления корреляций был использован критерий Спирмена (количественная оценка статистического изучения связи между явлениями, используемая в непараметрических методах). В данном случае корреляционный анализ был проведен в два этапа: на первом – без дифференциации по жанрам текста (1), на втором – с учетом различий жанровой формы (2).

(1) В результате проведенного корреляционного анализа установлены сильные и средние отрицательные корреляции (чем чаще используется одно местоимение, тем реже другое) во всех группах. Сильная отрицательная корреляция выявлена для частот использования местоимений «Групп-я» – «Групп-ты» ($r = –0,52$), «Групп-Ты» – «Групп-Мы» ($r = –0,53$), умеренная корреляция – для частот использования местоимений «Групп-Я» – «Групп-Мы» ($r = –0,30$). Корреляцию считали достоверной при $p < 0,05$.

(2) Диалоги ВК: «Группа-я» – «Группа-ты», $r = –0,85$ ($p < 0,05$); «Группа-ты» – «Группа-мы», $r = –0,26$ ($p > 0,05$). «Группа-я» – «Группа-мы», $r = –0.15$.

Стены ВК: «Группа-я» – «Группа-ты», $r = –0,25$; «Группа-я» – «Группа-мы», $r = –0,42$; «Группа-ты» – «Группа-мы», $r = –0,62$. Для всех уровень значимости $p < 0,05$.

Устная публичная коммуникация: «Группа-я» – «Группа-ты», $r = –0,25$; «Группа-я» – «Группа-мы», $r = –0,77$; «Группа-ты» – «Группа-мы», $r = –0,18$. Для всех уровень значимости $p < 0,05$.

Данные корреляции свидетельствуют о том, что в диалогическом общении, как правило, осуществляется фокусировка на одном из участников коммуникации, выводя его в коммуникативно сильную позицию, снижая частотность маркирования другого участника. Отрицательная корреляция статистически незначима только между группами местоимений «Я» и «Мы» в «Диалогах ВК».

На рис. 2 можно видеть маркирование диалогичности исследуемых жанров – количественное преобладание местоименных форм, базовых операторов диалогического общения.

Диаграмма демонстрирует попарное сравнение частоты использования групп местоимений относительно друг друга (вне зависимости от жанра текста). Чтобы отличить принадлежность текста к тому или иному жанру, точки в пространстве окрашены.

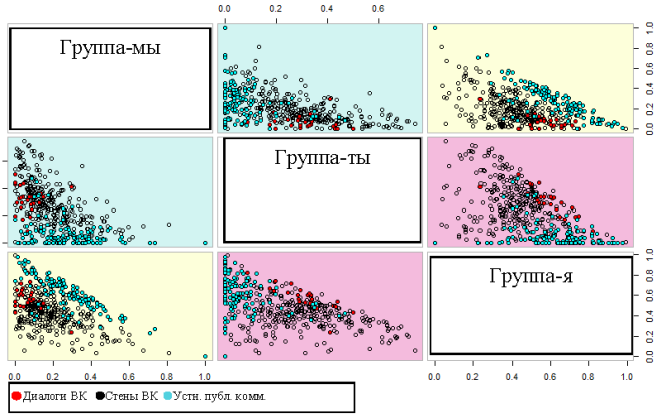

Рис. 2. Диаграмма рассеивания относительных частот групп местоимений в текстах трех типах жанров

Частоты группы всех типов местоимений «сливаются» в одно пространство по двум текстовым жанрам (Стены ВК и Диалоги ВК) и одновременно противопоставляются частотам местоимений в устной публичной коммуникации, которые имеют явно выраженную отрицательную линейную зависимость во всех группах местоимений. Результат визуализации корреляций позволяет предположить, во-первых, существование линейной отрицательной зависимости использования групп местоимений от типа жанра; во-вторых, ярко выраженное противопоставление жанра устной публичной коммуникации относительно двух других жанровых форм. Данную гипотезу мы подтверждаем в последующих статистических анализах.

Далее нами была проверена гипотеза о принадлежности сравниваемых независимых выборок к одной и той же генеральной совокупности с помощью непараметрического Краскела – Уоллиса. Формулируем гипотезы: H_0 – выбранные группы не имеют значимых различий по исследуемому признаку (нулевая гипотеза); H_1 – выбранные группы значимо различаются

по исследуемому признаку (альтернативная). Если эмпирическое значение равно или превышает теоретическое значение критерия ($p < 0,05$), то принимаем гипотезу H_0 и отклоняем гипотезу H_1.

В результате применения теста Краскела – Уоллиса получены следующие результаты: местоимения «Группа-я»: chi-squared = 151.56, df = 2, p-value < 2.2e-16; местоимения «Группы-ты»: chi-squared = 210.97, df = 2, p-value < 2.2e-16; местоимения «Группы-мы»: chi-squared = 55.703, df = 2, p-value < 8.022e-13. Следовательно, мы опровергаем статистическую гипотезу о равенстве групп местоимений (H_0) «Группы-я» и «Группы-ты» в использовании в трех исследуемых жанровых формах: разница в использовании данных групп местоимений статистически значима в отличие от местоимений «Группы-мы», значимая разница которой в использовании в текстах трех жанровых форм не подтвердилась. Результаты представлены в табл. 1.

Таблица 1

Результаты применения теста гипотезы Краскела – Уоллиса: равенство / различие дисперсий использования групп местоимений в текстов трех жанровых форм

Зависимая «Группа-я»	Kruskal – Wallis test: H (2, N = 449) = 151,5601 p = 0,000		
	Стены ВК	Диалоги ВК	Устн. публ. коммуникация
Стены ВК		0.000	0. 000
Диалоги ВК	0,000		1
Устн. публ. коммуникация	0,000	1	
Зависимая «Группа-я»	Kruskal – Wallis test: H (2, N= 449) =151,5601 p =0,000		
	Стены ВК	Диалоги ВК	Устн. публ. коммуникация
Стены ВК		0,000	0,000
Диалоги ВК	0,000		1
Устн. публ. коммуникация	0,000	1	
Зависимая «Группа-ты»	Kruskal – Wallis test: H (2, N= 449) =210,9674 p = 0,000		
	Стены ВК	Диалоги ВК	Устн. публ. коммуникация
Стены ВК		1	0,000
Диалоги ВК	1		0,000
Устн. публ. коммуникация	0,000	0,000	
«Группа-мы»	Kruskal – Wallis test: H (2, N= 449) =55,70287 p = 0,0000		
	Стены ВК	Диалоги ВК	Устн. публ. коммуникация
Стены ВК		0,000	0,000
Диалоги ВК	0,000		0,000
Устн. публ. коммуникация	0,000	0,000	

Таким образом, проведенный статистический анализ подтверждает нашу гипотезу о том, что использование групп местоимений отличается в рассмотренных жанровых вариантах личностной коммуникации. В частности, в «Группе-мы» зафиксированы статистически значимые отличия частотности

их использования во всех вариантах соотношений жанров; в «Группе-ты» выявлены статистические различия частотности их использования во всех типах соотношений жанровых форм, кроме оппозиции «Диалоги ВК – Стены ВК»; в «Группе-я» не выявлены значимые различия в частотности в соотношении жанровых форм Диалогов ВК и устной публичной коммуникации. На основе данных подтверждается гипотеза о том, что использование групп местоимений отличается во всех трех типах жанровых форм.

Линейная зависимость в относительной частотности использования местоимений в исследуемых трех типах жанровых форм также была подтверждена с помощью метода обобщенных линейных моделей (GLM). Результаты анализа представлены на рис. 3.

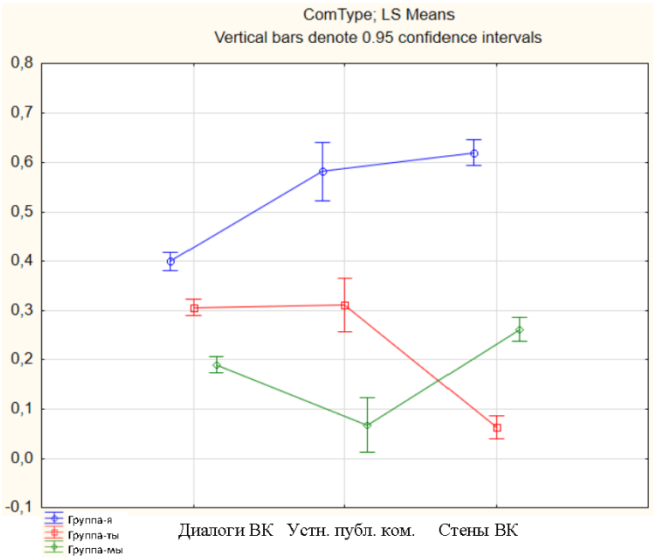

Рис. 3. Обобщенные линейные модели относительной частотности использования местоимений

Обобщенная линейная модель связывает зависимую переменную с факторами и ковариатами посредством задаваемой функции, что позволяет определить динамику использования групп местоимений в трех типах жанровых форм. В результате анализа была выявлена отрицательная линейная зависимость относительных частот всех групп местоимений ($p < 0,05$). Так, например, если в Стенах ВК используются чаще местоимения «Группы-я», то, соответственно, меньше используются местоимения «Группы-ты». Результаты этого анализа подтверждают данные, представленные в диаграмме рассеивания частот использования групп местоимений.

На графиках также визуализируется: 1) общее преобладание частот местоимений «я» во всех типах жанровых форм; 2) значительно меньший разброс частотностей в использовании групп местоимений в Диалогах ВК

(от 0,18 до 0,4) и наибольшая – в Стенах ВК (от 0,06 до 0, 62); 3) представленные в анализе диалогические жанры устной публичной коммуникации и Стены ВК, сближаясь по уровню разброса частотности местоимений, противопоставлены по соотношению частотностей групп «ты» и «мы» с другими группами местоимений; 4) проанализированные жанровые формы устной публичной коммуникации противопоставлены жанровым формам интернет-коммуникации по степени различий.

Далее была проверена эффективность использования местоимений *я, мы, ты* в решении задач автоматической атрибуции рассмотренных типов текстов по жанровому признаку с использованием методов машинного обучения.

Все тексты были разбиты на обучающую и тестовую выборки в пропорции 70 к 30%. Точность классификации основывается на формуле F1 (F-мера), которая позволяет нивелировать разброс классов (объем корпуса). Результаты анализа эффективности использования групп местоимений в задачах автоматической атрибуции с использованием семи алгоритмов представлены в табл. 2. Мы включили в таблицу среднюю точность работы классификаторов для последующей оценки влияния типа текстов на точность работы автоматического классификатора.

Таблица 2

Формальная точность классификации текстов методами машинного обучения (три жанровые формы текстов)

Модель машинного обучения	**F1**
Линейный дискриминантный анализ	0,92
Случайный лес	0,60
Метод опорных векторов	0,74
Деревья решений	0,91
Наивный байесовский классификатор	0,70
Логистическая регрессия	0,77
NN LSTM	0,98
Средняя точность	**0,98**

По данным таблицы, самым точным алгоритмом классификации текстов коммуникации на основе трех групп местоимений являются нейронные сети (NN LSTM).

Исходя из предыдущего анализа, который показал, что Стены ВК и Диалоги ВК сильно коррелируют между собой, мы объединили тексты этих типов жанровых форм и провели обучение модели с бинарной классификацией: Стена ВК и Диалоги ВКонтакте vs устная публичная коммуникация. Результаты анализа представлены в табл. 3.

Как видно из разных вариантов классификаций, представленных в табл. 2, 3, наилучший результат показывает бинарная классификация (LSTM). При этом средняя точность бинарной классификации методами машинного обучения увеличилась на 0.01, что подтверждает гипотезу о равенстве частотности использования местоимений в Диалогах ВК и в

Стенах ВК. Рекуррентные нейронные сети и в данном случае показывают большую точность.

Таблица 3

Формальная точность классификации текстов методами машинного обучения (две жанровые формы текстов)

Модель машинного обучения	F1
Линейный дискриминантный анализ	0,80
Случайный лес	0,92
Метод опорных векторов	0,74
Деревья решений	0,84
Наивный	0,88
Логистическая регрессия	0,90
NN LSTM	0,99
Средняя точность	**0,99**

Таким образом, статистическими методами была доказана зависимость использования групп местоимений от жанровой формы текста. Применение методов машинного обучения, проверка результативности моделей с использованием формальных метрик подтвердила значительную степень близости в использовании местоимений в текстах исследуемых жанровых форм (Стены ВК и Диалоги ВК), то, что интернет-коммуникация в этом аспекте противопоставлена жанровым формам устной публичной коммуникации.

II. Автоматическая гендерная атрибуция текстов.

Далее был проведен анализ маркирования гендерных различий в исследуемых типах текстов в той же последовательности, которая была применена для анализа жанровой дифференциации.

Результаты сравнения относительных частот использования местоимений мужчинами и женщинами в исследуемых текстах представлены на рис. 4 (группы субъектов коммуникации маркированы по гендерному признаку: ж. – женщины и м. – мужчины).

Как можно видеть на графиках, практически отсутствуют значительные различия в относительной частотности использования местоимений всех трех групп в исследуемых типах жанровых форм мужчинами и женщинами. Однако все же прослеживаются слабо проявленные различия в частоте их использования в Стенах ВК: женщины чаще используют местоимения групп «Я» и «Ты», однако меньше используют местоимений «Группы-Мы»; в «Диалогах ВК» женщины чаще используют местоимения «Группы-Я», а мужчины, наоборот, чаще используют местоимения «Группы-Ты», и, судя по медиане, местоимения «Группы-Мы». В устной публичной коммуникации гендерно обусловленные различия в использовании местоимений исследуемых групп не отмечены.

Зафиксировав степень отличий в использовании групп местоимений в текстах трех жанровых форм, написанных мужчинами и женщинами, проведем корреляционный анализ, вычленив фрагменты текстов по гендерному признаку их авторов и установив характер попарных корреляций групп местоимений.

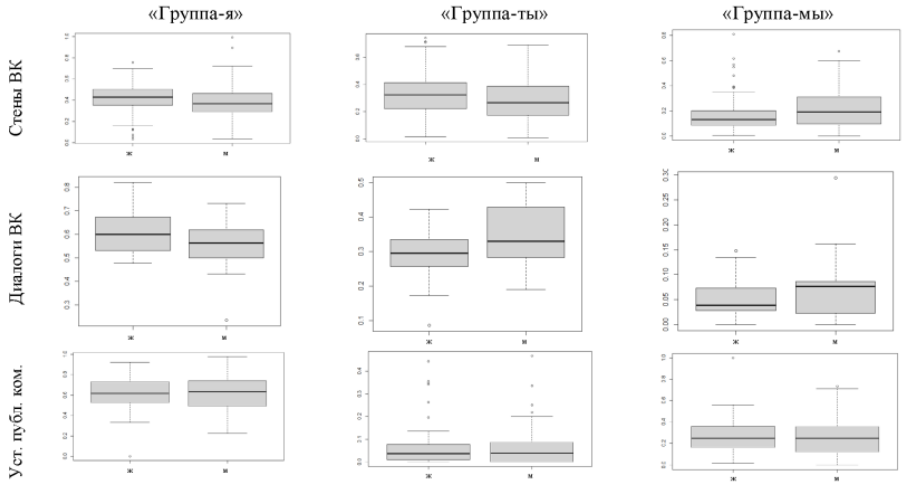

Рис. 4. Диаграмма размаха относительной частоты использования групп местоимений мужчинами и женщинами в текстах трех жанровых форм

1. В мужских фрагментах: «Группа-я» – «Группа ты», r = –0,47; «Группа-я» – «Группа-мы», r = –0,43; «Группа-ты» – «Группа-мы», r = –0,48. Для всех уровень выделенных групп установлена значимость $p < 0,05$.

2. В женских фрагментах: «Группа-я» – «Группа-ты», r = –0,61; «Группа-я» – «Группа-мы», r = –0,11 (отсутствует статистическая значимость); «Группа-ты» – «Группа-мы», r = –0,58. Для всех уровень значимости $p < 0,05$.

Исходя из полученных значений корреляций можно предположить, что, во-первых, преобладает средний уровень корреляционной зависимости; во-вторых, сохраняется отрицательная динамика использования частот местоимений в мужских и женских фрагментах текста, т.е. увеличение частоты одной группы местоимений коррелирует с уменьшением частоты другой. Если сопоставить с предыдущим анализом о выявлении корреляций групп местоимений с учетом жанра, то можно предположить, что отрицательная динамика частоты групп местоимений будет зависеть не от гендерной принадлежности участника коммуникации, а от жанровой принадлежности текста, соотносимой со своеобразием коммуникативных стратегий и интенций коммуникантов, тем коммуникации. Однако есть различия в характере корреляционных связей в мужских и женских фрагментах в целом: в мужских фрагмента установлен близкий уровень корреляций (степень различии не превышает 0,05) , в женских наблюдаются различия: между «Группой-я» и «Группой-мы» отсутствует статистически значима корреляция, в то время как между «Группой-я» и «Группой-ты» установлен наиболее высокий уровень отрицательной корреляции (–0,61).

Представим распределение частот местоимений во всех типах текстов, визуализировав двумя цветами частоты местоимений во фрагментах, написанных мужчинами (красный) и женщинами (синий) на рис. 5.

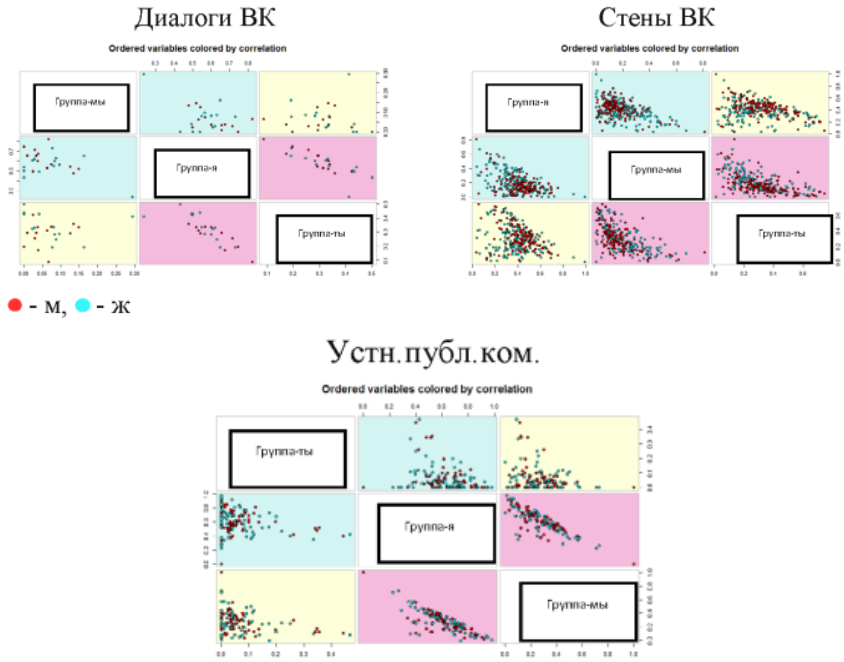

Рис. 5. Диаграмма рассеивания относительной частоты групп местоимений в мужских и женских фрагментах трех типов жанровых форм

Частоты всех типов местоимений в текстовых фрагментах, написанных мужчинами и женщинами, «сливаются» в одно пространство в текстах Стен ВК и Диалогов ВК) и одновременно противопоставляются устной публичной коммуникации.

Частоты всех типов местоимений имеют отрицательную линейную зависимость от типа гендерной и жанровой принадлежности текста. Гендерная диверсификация не прослеживается: частоты расположены одном пространстве, нет условного визуального разграничения относительно гендерной принадлежности автора текста. Все текстовые фрагменты, авторами которых являются мужчины и женщины, расположены относительно друг друга в одном гиперпространстве, без визуальных отличий. Однако отмечаются незначительные отличия дисперсий во фрагментах текстов Стен ВК, написанных мужчинами и женщинами (в Стенах ВК прослеживаются различия в плотности распределения относительных частот).

Это наблюдение подтверждается сравнением независимых выборок одной и той же генеральной совокупности с помощью непараметрического U-критерия Манна – Уитни (чем меньше значение критерия, тем вероятнее, что различия между значениями параметра в выборках достоверны). В качестве вывода выносим альтернативные гипотезы: а) H_0 – выбранные группы не имеют значимых различий по исследуемому признаку; б) H_1 – выбранные группы значимо различаются по исследуемому признаку. Если

эмпирическое значение равно или превышает теоретическое значение критерия, то отклоняем гипотезу H_0 (при $p > 0,05$) и принимаем гипотезу H_1 ($p < 0,05$). Исходя из результатов статистического критерия Манна – Уитни, делаем вывод, что существуют гендерные отличия в использовании всех группах местоимений в текстах Стен ВК, что представлено в табл. 4.

Таблица 4

Результаты применения теста Краскела – Уоллиса: частотность использования групп местоимений мужчинами и женщинами в текстах Стен ВК

Группа	Макс. отр. разн.	Макс. пол. разн.	p-val	Среднее m	Среднее F	Стат. откл. m	Стат. откл. F	N m	N f
Группа-я	–0,260	0,019	p < ,001	0,378	0,424	0,156	0,133	152	135
Группа-ты	–0,260	0,046	p < ,05	0,291	0,322	0,164	0,154	152	135
Группа-мы	–0,031	0,265	p < ,001	0,216	0,160	0,146	0,123	152	135

Вследствие того, что была подтверждена альтернативная гипотеза только в текстах жанра Стены ВК, далее представим результаты анализа гендерных различий в текстах данного жанра.

При анализе данных Стены ВК мы принимаем альтернативную гипотезу (H_1). В остальных типах коммуникаций статистически значимых различий в использовании групп местоимений относительно гендерной принадлежности автора текста не наблюдается (H_0). Другими словами, анализ подтверждает отличия частот в использовании местоимений «Группы-мы» в текстах, написанных мужчинами и женщинами, только в текстах Стен ВК. Остальные отличия носят случайный характер и не могут быть использованы в качестве подтверждающей гипотезы о различии использования местоимений мужчинами и женщинами. Эти данные соотносятся с полученными нами ранее результатами и выводами о том, что использование местоимений в речи зависит именно от темы коммуникации, а не от гендерной принадлежности коммуниканта. В статье представлены данные кластерного анализа (метод Уорда), который показал распределение текстов в кластеры текстов в зависимости от темы коммуникации, а не от гендерной принадлежности автора текста [35].

Кроме этого при помощи метода обобщенных линейных моделей была подтверждена линейная зависимость частот использования местоимений всех групп местоимений мужчинами и женщинами в текстах Стены ВК ($p < 0,05$) (рис. 6). Таким образом, можно сделать вывод, что существуют статистически значимые различия в использовании местоимений мужчинами и женщинами в текстах Стен ВК. И так как в других типах дискурса гендерных статистически значимых различий в использовании местоимений не наблюдается, были получены более низкие результаты классификации текстов на основе местоимений рассматриваемых групп в качестве маркеров гендерных различий. Для классификации текстов были применены те же методы машинного обучения для автоматической гендерной атрибуции текста, что и ранее для жанровой атрибуции. Результаты представлены в табл. 5.

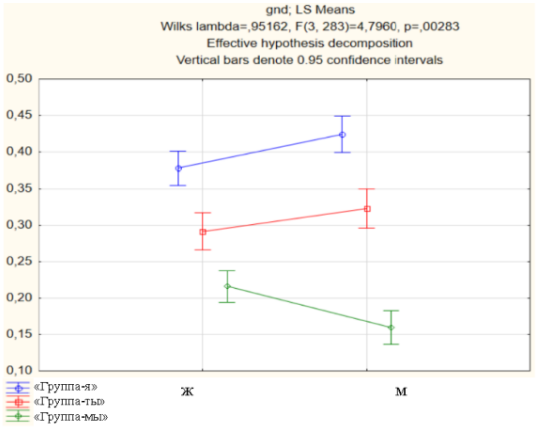

Рис. 6. Обобщенные линейные модели частотности групп местоимений в текстах Стены ВК

Исходя из проведенного анализа можно сделать вывод, что группы местоимений лучше используются в классификации жанровой принадлежности текста, чем для гендерной атрибуции. Гендерная дифференциация подтверждена только в текстах жанра Стены ВК. Результат анализа «полного датасета» представляет собой классификацию, в которой Диалоги ВК объединены со Стенами ВК и противопоставлены устной публичной коммуникации. Как видно, существует значимое увеличение точности классификатора, что свидетельствует о «схожести» этих двух жанров и их противопоставлении устной публичной коммуникации.

Таблица 5

Формальная точность гендерной классификации текстов методами машинного обучения

Модель машинного обучения	F1	Тип дискурса
Линейный дискриминантный анализ	0,64	Стены ВК
Случайный лес	0,68	
Метод опорных векторов	0,79	
Деревья решений	0,56	
Наивный	0,63	
Логистическая регрессия	0,64	
NN LSTM	0,56	
Средняя	**56,25**	
Линейный дискриминантный анализ	0,29	Диалоги ВК
Случайный лес	0,69	
Метод опорных векторов	0,86	
Деревья решений	0,86	
Наивный	0,25	
Логистическая регрессия	0,29	
NN LSTM	0,67	
Средняя	**49,69**	

Модель машинного обучения	F1	Тип дискурса
Линейный дискриминантный анализ	0,37	Устная публичная коммуникация
Случайный лес	0,55	
Метод опорных векторов	0,38	
Деревья решений	0,52	
Наивный	0,66	
Логистическая регрессия	0,37	
NN LSTM	0,58	
Средняя	**43,63**	
Линейный дискриминантный анализ	0,74	Полный датасет
Случайный лес	0,77	
Метод опорных векторов	0,77	
Деревья решений	0,58	
Наивный	0,64	
Логистическая регрессия	0,74	
NN LSTM	0,54	
Средняя	**57,96**	

Исходя из полученного результата анализа можно сделать вывод о частичном подтверждении влияния жанровых и гендерных различий коммуникации на диагностическую силу личных местоимений в автоматической атрибуции текстов. Кроме этого, установлена близость жанров Стены ВК и Диалоги ВК и их противопоставленность жанру устной публичной коммуникации. Результаты анализа свидетельствуют о том, что при решении задач автоматической классификации текстов по признаку гендерной принадлежности автора текста необходимо учитывать жанровую форму текста, так как в силу жанровых особенностей языковые средства выражения интенций могут влиять на частоту использования личных местоимений. В результате проведенного исследования гипотеза была подтверждена на основе использования статистических методов и методов машинного обучения. В качестве важного результата отметим также наличие статистически значимых отличий только в использовании местоимений «Группы-я», «Группы-мы» в Стенах ВК.

Список источников

1. *Земская Е.А., Китайгородская М.А., Розанова Н.Н.* Особенности мужской и женской речи // Русский язык и его функционирование. М., 1993. С. 90–136.

2. *Земская Е.А., Китайгородская М.А., Розанова Н.Н.* О чем и как говорят женщины и мужчины // Русская речь. 1989. № 1. С. 2–46. URL: https://russkayarech.ru/ru/archive/1989-1/42-46

3. *Колесов В.В.* Язык и ментальность. СПб., 2004. 237 с.

4. *Попова Е.А.* Об особенностях речи мужчин и женщин // Русская речь. 2007. № 3. С. 40–49. URL: https://russkayarech.ru/ru/archive/2007-3/40-49

5. *Новикова И.Н., Хамидулина Л.Ю.* К вопросу об особенностях мужской и женской речи // Наука и современность – 2013. Филологические науки. Новосибирск, 2013. С. 78–83.

6. *Беляева А.Ю.* Особенности речевого поведения мужчин и женщин : На материале русской разговорной речи : автореф. дис. ... канд. филол. наук. Саратов, 2002. 19 с.

7. *Стернин И.А.* Общение с разными типами собеседников. Воронеж : Истоки, 2012. 42 с.

8. *Mukherjee A., Liu B.* Improving Gender Classication of Blog Authors // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010. P. 32–38.

9. *Yan X., Yan L.* Gender Classification of Weblog Authors // Computational Approaches to Analyzing Weblogs. AAAI, 2006. P. 18–26.

10. *Shlomo A.* Gender, Genre, and Writing Style in Formal Written Texts // Shlomo Argamon, Moshe Koppel, Jonathan Fine, Anat Rachel Shimoni Springer, Sex Roles. 2010 Jun. № 62 (11-12). P. 705–720.

11. *Marcelo Luiz.* Brocardo Authorship Verification for Short Messages using Stylometry, 2014. URL: https://www.deepdyve.com/lp/institute-ofelectrical-and-electronics-engineers/authorship-verification-for-short-messages-using-stylometry-JM5XWbkHyN (дата обращения: 07.07.2016).

12. *Arroju M.* Age, Gender and Personality Recognition using Tweets in a Multilingual Setting // 6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction. 2015. P. 23–31.

13. *Васильева А.В.* Коммуникативно-прагматические аспекты проявления экспрессивности в мужских и женских коротких электронных сообщениях // Вестник науки Сибири. 2014. № 4 (14). С. 190–195.

14. *Горошко Е.* Особенности мужского и женского стиля письма // Преображение. Русский феминистский альманах. М., 1998. № 6. С. 48–64.

15. *Кирилина А.В.* Гендер: лингвистические аспекты. М. : Ин-т социологии РАН, 1999. 189 с.

16. *Карасик В.В., Карасик В.И.* О типах дискурса // Языковая личность: институциональный и персональный дискурс. Волгограда, 2000. С.5–20.

17. *Алтухова Т.В.* Социальная компьютерная сеть «ВКонтакте»: жанровая характеристика // Вестник Кемеровского государственного университета. 2012. № 4 (52). Т. 3: Филология. С. 21–25.

18. *Марченко Н.Г.* Социальная сеть «ВКонтакте»: лингвопрагматический аспект : автореф. ... канд. филол. наук. Ростов н/Д, 2013. 21 с.

19. *Кобрин Н.В.* Твиттинг – новый социокоммуникативный жанр интернет-коммуникации // Филологические науки. Вопросы теории и практики. 2016. № 9 (63) : в 3 ч. Ч. 3. С. 109–111.

20. *Ковальчукова М.А.* Новостной анонс в сети Интернет как речевой жанр дискурса СМИ : автореф. дис. ... канд. филол. наук. Ижевск, 2009. 24 с.

21. *Кириллов А.Г.* Трансформация жанра блога в программах обмена мгновенными сообщениями // Жанры речи. 2017. № 2 (16). С. 260–267.

22. *Резанова З.И., Романов А.С., Мещеряков Р.В.* Задачи авторской атрибуции текста в аспекте гендерной принадлежности (к проблеме междисциплинарного взаимодействия лингвистики и информатики) // Вестник Томского государственного университета. 2013. № 370. С. 24–28.

23. *Дроздова Т.Н.* Диагностические и классификационные задачи в авторoведческой экспертизе блогов // Актуальные проблемы российского права. 2010. № 2 (15). С. 394–404.

24. *Романов А.С.* Методика и программный комплекс для идентификации автора неизвестного текста : автореф. дис. … канд. техн. наук. Томск, 2010. 27 с.

25. *Pennebaker J.W., MR Mehl, Niederhoffer K.G.* Psychological aspects of natural language use: Our words, our selves // Annual review of psychology. 2003. P. 548–571.

26. *Вольф Е.М.* Грамматика и семантика местоимений. М. : Наука, 1974. 223 с.

27. *Verhoeven B.C.* TWISTY: A Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling // Ben Verhoeven, Walter Daelemans and Barbara Plank CLiPS Research Center, University of Antwerp, Belgium University of Groningen, The Netherlands, 2015. P. 1632–1637.

28. *Баранов А.Н.* Введение в прикладную лингвистику. М. : Эдиториал УРСС, 2001. 347 с.

29. *Степаненко А.А.* Гендерная атрибуция текстов компьютерной коммуникации: статистический анализ использования местоимений // Вестник Томского государственного университета. 2017. № 415. С. 17–25. doi: 10.17223/15617793/415/3

30. *Резанова З.И.* Подкорпус устной речи русско-тюркских билингвов Южной Сибири: типологически релевантные признаки // Вопросы лексикографии. 2017. № 11. С. 105–118. doi: 10.17223/22274200/11/7

31. *Земская Е.А., Китайгородская М.В., Ширяев Е.Н.* Русская разговорная речь: Общие вопросы. Словообразование. Синтаксис. М. : Наука, 1981.

32. *Русская* разговорная речь: Фонетика. Морфология. Лексика. Жест / отв. ред. Е.А. Земская. М. : Наука, 1983.

33. *Лурия А.Р.* Язык и сознание. Ростов н/Д, 1998. 416 с.

34. *Резанова З.И., Мишанкина Н.А.* Семиотический синтез в коммуникативном пространстве интернет-текстов (на материале чат-коммуникации) // Сибирский филологический журнал. 2006. № 1–2. С. 70–74.

35. *Степаненко А.А., Резанова З.И.* Экспрессивность как маркер гендерных различий компьютерной коммуникации (к проблеме автоматической гендерной атрибуции текста) // Вестник Томского государственного университета. 2018. № 433. С. 38–46. doi: 10.17223/15617793/433/5

References

1. Zemskaya, E.A., Kitaygorodskaya, M.A. & Rozanova, N.N. Osobennosti muzhskoy i zhenskoy rechi [Features of male and female speech]. In: Zemskaya, E.A. & Shmelev, D.N. (eds) (1993) *Russkiy yazyk i ego funktsionirovanie* [Russian Language and Its Functioning]. Moscow: Nauka. pp. 90–136.

2. Zemskaya, E.A., Kitaygorodskaya, M.A. & Rozanova, N.N. (1989) O chem i kak govoryat zhenshchiny i muzhchiny [What and how women and men talk about]. *Russkaya rech' – Russian Speech*. 1. pp. 2–46. [Online] Available from: https://russkayarech.ru/ru/archive/1989-1/42-46.

3. Kolesov, V.V. (2004) *Yazyk i mental'nost'* [Language and Mentality]. Saint Petersburg: Saint Petersburg State University.

4. Popova, E.A. (2007) Ob osobennostyakh rechi muzhchin i zhenshchin [About the peculiarities of speech of men and women]. *Russkaya rech' – Russian Speech*. 3. pp. 40–49. [Online] Available from: https://russkayarech.ru/ru/archive/2007-3/40-49.

5. Novikova, I.N. & Khamidulina, L.Yu. (2013) [On the question of the peculiarities of male and female speech]. *Nauka i sovremennost' – 2013. Filologicheskie nauki* [Science and Modernity – 2013. Philological sciences]. Proceedings of the 23rd International Conference. Novosibirsk. 8 July 2013. Novosibirsk: Sibprint: TsRNS. pp. 78–83. (In Russian).

6. Belyaeva, A.Yu. (2002) *Osobennosti rechevogo povedeniya muzhchin i zhenshchin: Na materiale russkoy razgovornoy rechi* [Features of speech behavior of men and women: On the material of Russian colloquial speech]. Abstract of Philology Cand. Diss. Saratov.

7. Sternin, I.A. (2012) *Obshchenie s raznymi tipami sobesednikov* [Communication with Different Types of Interlocutors]. Voronezh: Istoki.

8. Mukherjee, A. & Liu, B. (2010) [Improving Gender Classication of Blog Authors]. *2010 Conference on Empirical Methods in Natural Language Processing.* Proceedings of the International Conference. Cambridge, MA. 9–11 October 2010. Stroudsburg, PA: Association for Computational Linguistics. pp. 32–38.

9. Yan, X. & Yan, L. (2006) [Gender Classification of Weblog Authors]. *Computational Approaches to Analyzing Weblogs*. Proceedings of the 2006 AAAI Spring Symposium. Stanford, CA. 27–29 March 2006. AAAI. pp. 18–26.

10. Shlomo, A. et al. (2003) Gender, Genre, and Writing Style in Formal Written Texts. *Text*. 23 (3). pp. 321–346. DOI: 10.1515/text.2003.014

11. Marcelo, L. (2014) Brocardo Authorship Verification for Short Messages using Stylometry, 2014. *Deepdyve*. [Online] Available from: https://www.deepdyve.com/lp/institute-ofelectrical-and-electronics-engineers/authorship-verification-for-short-messages-using-stylometry-JM5XWbkHyN. (Accessed: 07.07.2016).

12. Arroju, M. (2015) [Age, Gender and Personality Recognition using Tweets in a Multilingual Setting]. *6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction*. Proceedings of the 6th Conference. Toulouse, France. 8–11 September 2015. Switzerland: Springer. pp. 23–31.

13. Vasil'eva, A.V. (2014) Kommunikativno-pragmaticheskie aspekty proyavleniya ekspressivnosti v muzhskikh i zhenskikh korotkikh elektronnykh soobshcheniyakh [Communicative and pragmatic aspects of expressiveness in male and female short electronic messages]. *Vestnik nauki Sibiri – Siberian Journal of Science*. 4 (14). pp. 190–195.

14. Goroshko, E. (1998) Osobennosti muzhskogo i zhenskogo stilya pis'ma [Features of male and female writing style]. *Preobrazhenie. Russkiy feministskiy al'manakh*. 6. pp. 48–64.

15. Kirilina, A.V. (1999) *Gender: lingvisticheskie aspekty* [Gender: Linguistic aspects]. Moscow: Sociological Institute of the RAS.

16. Karasik, V.V. & Karasik, V.I. (2000) O tipakh diskursa [About the types of discourse]. In: Karasik, V.I. & Slyshkin, G.G. (eds) *Yazykovaya lichnost': institutsional'nyy i personal'nyy diskurs* [Linguistic Personality: Institutional and personal discourse]. Volgograd: Peremena. pp. 5–20.

17. Altukhova, T.V. (2012) Social computer network Vkontakte: genre characterization. *Vestnik Kemerovskogo gosudarstvennogo universiteta – Bulletin of Kemerovo State University*. 4-3 (52). pp. 21–25. (In Russian).

18. Marchenko, N.G. (2013) *Sotsial'naya set' "VKontakte": lingvopragmaticheskiy aspekt* [Social network "VKontakte": linguopragmatic aspect]. Abstract of Philology Cand. Diss. Rostov-on-Don.

19. Kobrin, N.V. (2016) Twitting – new socio-communicative genre of Internet communication. *Filologicheskie nauki. Voprosy teorii i praktiki – Philology. Theory & Practice*. 9-3 (63). pp. 109–111. (In Russian).

20. Koval'chukova, M.A. (2009) *Novostnoy anons v seti Internet kak rechevoy zhanr diskursa SMI* [News announcement on the Internet as a speech genre of media discourse]. Abstract of Philology Cand. Diss. Izhevsk.

21. Kirillov, A.G. (2017) Transformation of blogs as a genre in instant messaging applications. *Zhanry rechi – Speech Genres*. 2 (16). pp. 260–267. (In Russian). DOI: 10.18500/2311-0740-2017-2-16-260-267

22. Rezanova, Z.I., Romanov, A.S. & Meshcheryakov, R.V. (2013) Tasks of author attribution of text in the aspect of gender (on interdisciplinary interaction of linguistics and computer science). *Vestnik Tomskogo gosudarstvennogo universiteta –Tomsk State University Journal*. 370. pp. 24–28. (In Russian).

23. Drozdova, T.N. (2010) Diagnosticheskie i klassifikatsionnye zadachi v avtorovedcheskoy ekspertize blogov [Diagnostic and classification tasks in the author's expertise of blogs]. *Aktual'nye problemy rossiyskogo prava – Actual Problems of Russian Law*. 2 (15). pp. 394–404.

24. Romanov, A.S. (2010) *Metodika i programmnyy kompleks dlya identifikatsii avtora neizvestnogo teksta* [Methodology and software package for identifying the author of an unknown text]. Abstract of Technics Cand. Diss. Tomsk.

25. Pennebaker, J.W., Mehl, M.R. & Niederhoffer, K.G. (2003) Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*. pp. 548–571. DOI: 10.1146/annurev.psych.54.101601.145041

26. Vol'f, E.M. (1974) *Grammatika i semantika mestoimeniy* [Grammar and Semantics of Pronouns]. Moscow: Nauka.

27. Verhoeven, B.S. (2016) TWISTY: A Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling. *LREC*. pp. 1632–1637.

28. Baranov, A.N. (2001) *Vvedenie v prikladnuyu lingvistiku* [Introduction to Applied Linguistics]. Moscow: Editorial URSS.

29. Stepanenko, A.A. (2017) Gender attribution in social network communication: the statistical analysis of pronouns frequency. *Vestnik Tomskogo gosudarstvennogo universiteta – Tomsk State University Journal*. 415. pp. 17–25. (In Russian). DOI: 10.17223/15617793/ 415/3

30. Rezanova, Z.I. (2017) Subcorpus of oral speech of Russian-Turkic bilinguals of Southern Siberia: typologically relevant signs. *Voprosy leksikografii – Russian Journal of Lexicography*. 11. pp. 105–118. (In Russian). DOI: 10.17223/22274200/11/7

31. Zemskaya, E.A., Kitaygorodskaya, M.V. & Shiryaev, E.N. (1981) *Russkaya razgovornaya rech'. Obshchie voprosy. Slovoobrazovanie. Sintaksis* [Russian Colloquial Speech. General questions. Word formation. Syntax]. Moscow: Nauka.

32. Zemskaya, E.A. (ed.) (1983) *Russkaya razgovornaya rech'. Fonetika. Morfologiya. Leksika. Zhest* [Russian Colloquial Speech. Phonetics. Morphology. Vocabulary. Gesture]. Moscow: Nauka.

33. Luriya, A.R. (1998) *Yazyk i soznanie* [Language and Consciousness]. Rostov-on-Don: Feniks.

34. Rezanova, Z.I. & Mishankina, N.A. (2006) Semioticheskiy sintez v kommunikativnom prostranstve internet-tekstov (na materiale chat-kommunikatsii) [Semiotic synthesis in the communicative space of Internet texts (based on chat communication)]. *Sibirskiy filologicheskiy zhurnal – Siberian Journal of Philology*. 1–2. pp. 70–74.

35. Stepanenko, A.A. & Rezanova, Z.I. (2018) Expressiveness as a marker of gender differences in computer communication (the problem of automatic gender attribution of the text). *Vestnik Tomskogo gosudarstvennogo universiteta – Tomsk State University Journal*. 433. pp. 38–46. (In Russian). DOI: 10.17223/15617793/433/5

Информация об авторах:

Степаненко А.А. – научный сотрудник лаборатории лингвистической антропологии Национального исследовательского Томского государственного университета (Томск, Россия). E-mail: stepanenkone@mail.ru

Резанова З.И. – д-р филол. наук, зав. кафедрой общей, компьютерной и когнитивной лингвистики, зам. заведующего лабораторией лингвистической антропологии Национального исследовательского Томского государственного университета (Томск, Россия). E-mail: rezanovazi@mail.ru

Авторы заявляют об отсутствии конфликта интересов.

Information about the authors:

A.A. Stepanenko, researcher, Tomsk State University (Tomsk, Russian Federation). E-mail: stepanenkone@mail.ru

Z.I. Rezanova, Dr. Sci. (Philology), head of the Department of General, Computational and Cognitive Linguistics, deputy head of the Linguistic Anthropology Laboratory, Tomsk State University (Tomsk, Russian Federation). E-mail: rezanovazi@mail.ru

The authors declare no conflicts of interests.

Статья поступила в редакцию 14.04.2022;
одобрена после рецензирования 11.09.2022; принята к публикации 22.09.2022.

The article was submitted 14.04.2021;
approved after reviewing 11.09.2022; accepted for publication 22.09.2022.