УДК 81'33
DOI 10.25205/1818-7935-2020-18-1-22-32

Автоматическое определение половой принадлежности автора текста: феномен русской женской прозы

А. Б. Хазова

*Национальный исследовательский университет
«Высшая школа экономики»
Москва, Россия*

Аннотация

Статья посвящена экспериментальной апробации методов автоматического определения половой принадлежности авторов на материале русской художественной прозы 1960–2000 гг. Цель данной работы – выявление оптимальных методов автоматического определения половой принадлежности авторов, подкрепленное экспериментом по определению наиболее результативного алгоритма классификации художественных текстов при помощи машинного обучения. Задачи данного исследования включают в себя выделение грамматических и лексико-стилистических особенностей женской художественной прозы 1960–2000 гг. на фоне мужской прозы того же периода, а также и текстов XVIII–XIX вв.; отслеживание изменений в распределении употребления частей речи и знаков пунктуации на протяжении обозначенного периода. В женских и мужских текстах анализировались наиболее частотные части речи. Произведен анализ употребления особо часто используемых знаков препинания из заданного списка: вопросительный знак, восклицательный знак, запятая, двоеточие, точка с запятой, точка, запятая. В статье показан анализ распределения частей речи и знаков пунктуации в художественных текстах мужчин и женщин XVIII–XIX вв. Также проведен эксперимент по выявлению наиболее действенного алгоритма определения половой принадлежности автора. В результате исследования установлено, что частотность употребления различных частей речи в женских и мужских текстах заданного периода выглядит следующим образом: существительные, глаголы, предлоги, местоименные существительные, союзы и прилагательные, что отражает специфику художественного стиля. Можно заключить, что в современной литературе женщины более активно используют возможности пунктуации как выразительного средства: доля употребления восклицательных, вопросительных знаков и запятых у писательниц значительно превышает значение, полученное посредством анализа мужских текстов. Выяснено, что наиболее эффективными классификаторами художественной литературы являются такие реализации алгоритмов, как BayesNet и SMO.

Ключевые слова

женская проза, пол, гендер, классификация, машинное обучение, кластеризация

Для цитирования

Хазова А. Б. Автоматическое определение половой принадлежности автора текста: феномен русской женской прозы // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2020. Т. 18, № 1. С. 22–32. DOI 10.25205/1818-7935-2020-18-1-22-32

Automatic Detection of Gender Identity: The Phenomenon of Russian Women's Prose

Anastasiya B. Khazova

*National Research University Higher School of Economics
Moscow, Russian Federation*

Abstract

The article deals with the method of automatic detection of authors' gender identity on the material of fiction prose of 1980–2000. During this period, there is a special construct, called “women's prose”, which is characterized by a special genre and stylistic originality. We set ourselves the task to find out whether the concept of “women's prose” refers only to the non-text reality or is clearly reflected at the level of language. We have collected corpus of texts 1980–

© А. Б. Хазова, 2020

ISSN 1818-7935

Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2020. Т. 18, № 1
Vestnik NSU. Series: Linguistics and Intercultural Communication, 2020, vol. 18, no. 1

2000 and conducted that identified the most effective machine learning algorithms for the classification of male and female prose. This research focuses on methods for automatically determining the gender identity of authors on the material of prose from 1960 to 2000. The purpose of this work is to identify optimal methods for automatically determining the gender identity of the authors. The objectives of this study include highlighting the grammatical and stylistic features of prose from 1960 to 2000 and, in particular, women's prose and texts of 18th – 19th centuries; tracing the changes in the distribution of usage different parts of speech and punctuation for a specified period and conducting an experiment to identify the most effective algorithm for the classification of literary texts by using machine learning. The analysis revealed that women and men often use in their texts the following parts of speech: nouns, verbs, prepositions, pronominal nouns, conjunctions, and adjectives that reflects the specific artistic style. In addition, analysis was made of the use of the most commonly used punctuation marks from the given list: question mark, exclamation point, comma, colon, semicolon, period, comma. It has been observed that women are more actively using the means of punctuation as a means of expression in modern literature: the share of the use of exclamation, question marks and commas the writers is much higher than the value obtained through the analysis of men's texts. The work also contains an analysis of the distribution of parts of speech and punctuation of literary texts of men and women of 18th – 19th centuries. We performed experiment to identify the most effective algorithm for determining the gender identity of the author. It was found that the most effective classifiers of literature are the implementation of algorithms as BayesNet and SMO.

Keywords

women's prose, classification, sex, gender, machine learning, clusterization

For citation

Khazova, Anastasiya B. Automatic Detection of Gender Identity: The Phenomenon of Russian Women's Prose. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2020, vol. 18, no. 1, p. 22–32. DOI 10.25205/1818-7935-2020-18-1-22-32

Введение

Значительную долю современного литературного пространства занимают произведения, написанные авторами-женщинами. Исследователи считают женскую эмансипацию конца XX в. важной предпосылкой активного включения женщин в литературный процесс. Но в действительности точнее (в историко-культурном смысле) говорить о второй волне или пике женской эмансипации, поскольку первые «женские» (но не феминистские) тексты появились в России еще в конце XVIII в. (например, А. И. Вельяшева-Волынцева, Екатерина II, Е. Е. Голицына, Е. Р. Дашкова и др.), а во Франции даже раньше (Мари Мадлен де Лафайет, Мадам де Севинье и др.).

В Европе понятие «женская литература» начало утверждаться в 30–40-х гг. XIX в. [Воробьева, 2006. С. 6]. В России начала XX в. происходит расцвет женской поэзии – появляется плеяда поэтесс Серебряного века (Ахматова, Гиппиус, Цветаева, Арсенева, Волошина и др.). Этот процесс оказался созвучен процессу борьбы за равноправие женщин во всех сферах жизни. Как следствие, к концу века в России появляется ряд сборников, которые становятся своеобразным манифестом женщин-авторов: «Новые амазонки», «Брызги шампанского», «Не помнящая зла» [Серго, 2006].

Отметим, вместе с тем, что термины «женская литература» и «женская проза» не синонимичны. Если до 1980-х гг. исследователи называли женской прозой или литературой все художественные прозаические тексты, написанные писателями-женщинами, то в начале 1980-х возникает специфическое понятие женской прозы как отдельного конструкта, применяемого к текстам, созданным в СССР и современной России в 1980–2000 гг. Она обладает своими тематическими и стилистическими особенностями, а также очерчивает вполне конкретный круг авторов, среди которых фигурируют, в первую очередь, Т. Толстая, Л. Улицкая, В. Токарева, Л. Петрушевская, Г. Щербакова.

Первые отечественные работы о феномене этой женской прозы возникли под влиянием социальных перемен и с приходом в Россию гендерных идей в 1980–1990-х гг. Первопроходцами в этой области стали литературоведы Н. М. Габриэлян, Е. И. Трофимова, И. А. Жеребкина, которые описали место этого европейского и американского явления в рамках современной российской культуры и общества [Титарева, 2015. С. 4].

В дальнейшем исследователи начали рассматривать различные аспекты этого явления. Некоторые обращали внимание на особенности творчества отдельных авторов (ср. Ю. Л. Высочина «Интертекстуальность в прозе Татьяны Толстой: на материале романа “Кысь”», Е. А. Меркотун «Поэтика одноактной драматургии Л. Петрушевской», Е. А. Стрельцова «Языковые особенности художественных текстов В. Токаревой, Л. Петрушевской, Л. Улицкой»). Другие же пытались установить общие литературные особенности функционирования данного социокультурного феномена: Н. В. Воробьева «Женская проза 1980–2000 годов: динамика, проблематика, поэтика», Г. А. Пушкарь «О категории “гендер” и гендерных исследованиях. Художественная оппозиция феминность / маскулинность в современной женской прозе. Художественная специфика конфликта и хронотопа в женской прозе. Уровни гендерных художественных конфликтов».

Кроме того, многие авторы уделяют внимание факторам возникновения и становления женской прозы, а также пытаются определить признаки, с помощью которых ее можно идентифицировать как специфический род художественной литературы. Часть исследователей ориентируется при этом на гендерную принадлежность автора, которая, впрочем, заявляется им самим. Другие считают основанием для выделения женской прозы наличие специфических жанрово-стилевых характеристик текста, общих для всех писателей-женщин и не свойственных писателям-мужчинам.

Действительно, в исследовательских работах уже достаточно давно встречаются утверждения о существовании мужских и женских вариантов языка. Так, А. Пиперски считает, что, хотя и существует различие между словами и конструкциями, которые свойственно употреблять мужчинами и женщинам, но это «не абсолютные различия полов, а различия стилей: женский язык обычно нейтральный, а мужской – более грубый» [Пиперски, 2006]. Впрочем, на это различие было указано гораздо раньше в работе московских исследователей: мужчины чаще используют отрицательную оценку и сниженную лексику [Земская и др., 1993]. Что касается нейтральности женской речи, то она может и отрицаться. Так, Е. А. Стрельцова [2014] утверждает, что женской речи, напротив, свойственна экспрессивность, причем гиперболизированная.

Гендер, нередко игнорировавшийся в более ранних языковедческих исследованиях, приобретает особое значение с развитием гендерной лингвистики. Категория пола является продуктом культуры и фиксирует представления о различиях между мужским и женским. Она находит свое отражение как в народном творчестве и традициях, так и в самом языке.

В современной гендерной лингвистике выделяют две основные проблемы изучения этого фактора в языке.

1. Речевое и коммуникативное поведение людей подразумевает наличие типичных стратегий выбора определенных лексических единиц и синтаксических конструкций, «специфика мужского и женского говорения».

2. Отражение пола в языке. Этот подход состоит в описании стратегий выбора различных единиц в «номинативной системе, синтаксисе, категории рода» [Holmes, Meyerhoff, 2003].

Понятие гендерной идентичности, т. е. категоризации автором самого себя, включает в себя ряд категорий, таких как агендерность, тернарные гендерные модели и мн. др. Мы же используем в данном исследовании бинарную модель гендера, основанную на биологическом поле автора. Она диктуется спецификой корпуса, поскольку авторы-женщины, произведения которых попали в корпус, намеренно выделяют себя в литературной среде и не стараются скрыть свою половую принадлежность.

Задача половой идентификации методами машинного обучения становится проще благодаря множеству текстов, написанных авторами-женщинами в 1980–2000-х гг. Исследователи выделяют их в отдельный конструкт, поэтому получение этих данных позволяет проследить, относится ли понятие «женская проза» только к внетекстовой реальности или находит явное отражение на уровне языка. Кроме того, они помогают определить, способствуют ли различия в женском и мужском письме этого периода точнее идентифицировать пол автора.

Анализ художественной прозы всегда осложнен индивидуальными стилистическими чертами автора. Но вместе с тем в письменной речи появляются различные особенности на уровне грамматики и синтаксиса художественного текста, позволяющие отличить мужское письмо от женского.

Литературный процесс XX в. нашел отражение на всех уровнях языковой системы. Постоянное изменение лексического состава языка, использование разговорной и обценной лексики, отсутствие стилистической строгости и многое другое сформировали новый облик художественных произведений и значительно повлияли на женское письмо.

Теоретические предпосылки автоматической идентификации

Задача автоматической атрибуции текста осложняется серьезными ограничениями, связанными с возможностями программных средств. Машине трудно выявить качественные (такие, как тематика, тональность, просодия) и многие количественные признаки предположительно женского или мужского художественного текста, вполне очевидные литературоведу, критику и даже лингвисту. Е. А. Стрельцова, например, относит к характерным особенностям женской прозы «частое использование междометий, использование положительных оценок, разговорной речи, диминутивов» [2014]. Но попытка автоматически выявить в ней текстовую плотность положительных оценок и разговорной речи закончится, вероятнее всего, провалом из-за сложности формализации в лингвистике текста понятий «положительность» и «разговорность». Трудность может вызвать даже выявление диминутивов, несмотря на наличие у них достаточно характерной морфологии. А вот утверждения А. Пиперски, убежденного, что «для мужских текстов характерны существительные, прилагательные и предлоги, а для женских – местоимения, глаголы, наречия и междометия» [2006], достаточно легко поддаются автоматической верификации, поскольку речь идет всего лишь о классификации единиц текста по частям речи.

Применение подхода с использованием машинного обучения является актуальным методом для решения задач данного типа в основном потому, что решающими в процессе идентификации авторства текста признаются именно количественные показатели. Степень их репрезентативности и валидности выше, что дает экспертам возможность выносить более объективные заключения.

Таким образом, мы приходим к заключению, что методами автоматической идентификации можно квантифицировать только те признаки условно женского и мужского текстов, которые машина имеет возможность «посчитать», т. е. решать самые простые задачи. Выявить особенности текста, скажем, биологического мужчины, но демонстрирующего иные гендерные модели поведения и пишущего «по-женски», не представляется возможным на данный момент. Поэтому одной из ключевых исходных посылок для отнесения текста к женской прозе являются особенности выбора автором лексики, грамматики и синтаксических конструкций. Однако нужно иметь в виду, что этот выбор подчинен в еще большей степени законам художественного (литературного) жанра, общим для писателей-мужчин и писателей-женщин.

В настоящее время проведено уже множество экспериментов по автоматической классификации различных текстов [Argamon, 2009], в том числе относящихся к художественной литературе [Diederich et al., 2003], методами машинного обучения. Многие исследователи используют в качестве классов пол и возраст, акцентируя внимание на текстах из блогов и социальных сетей. Ведущими специалистами в этой области являются М. Koppel, S. Argamon, E. Stammatos, М. Liwicki, предлагающие различные способы определения половых классов при помощи ряда алгоритмов машинного обучения (balanced winnow algorithm, naïve bayes, decision trees и др.). Так, в работе [Koppel et al., 2002] апробирован вариант алгоритма экспоненциального градиента, который, в свою очередь, является обобщением алгоритма Balanced Winnow.

Также в задаче автоматической классификации текстов популярными являются алгоритмы SVM и k-nearest neighbor, описание которых даны в работах [Vel et al., 2002; Kucukyilmaz et al., 2008].

Несмотря на то что семантические особенности текста лучше выявляют половые различия и потенциально могут являться маркером авторского языка, они трудно измеримы, поэтому исследований в этой области не так много. Знаковой является статья [Argamon et al., 2005], в которой авторы, апеллируя к семантическим особенностям текста, представляют каждый набор частей речи в виде графов, способных расширять, углублять или ограничивать смысл текста.

В большинстве случаев исследователи предпочитают использовать стандартные лингвистические и потому стабильные характеристики слова и текста: части речи, служебные слова, длина предложений и слов, синтаксические связи, знаки пунктуации, поскольку они дают высокие результаты в области атрибуции текстов и, в частности, гендерной идентификации их автора.

Несмотря на актуальность проблемы авторского профилирования, задача гендерной классификации текстов на материале русской женской прозы еще не решалась. Ключевой конференцией в этом направлении является PAN. Также существует много ежегодных соревнований по определению пола автора по набору текстовых данных. Таким образом, российские исследователи активно вовлекаются в разработку методов авторского профилирования. Важными являются работы Т. А. Литвиновой и О. В. Загоровской, авторов корпуса RusGendAttr, созданного на базе интернет-текстов.

Совместная лаборатория корпусной социолингвистики и автороведческих исследований (Воронеж) работает в области автоматического профилирования интернет-текстов, а также занимается созданием корпусов текстов с нейропсихологической разметкой, т. е. указывает соответствующие характеристики их авторов.

К сожалению, современные алгоритмы не позволяют точно определить пол пишущего, но для идентификации достаточно и уровня 90 %, которые обеспечиваются составлением статистической модели гендерно маркированного текста на основе ряда его формальных признаков.

Задачи и метод

Задача данного исследования заключается в определении наиболее результативных алгоритмов классификации из числа принятых для решения задач классификации текстов с самыми разными целями, чтобы выявить их способность корректно отнести документ к категории мужских или женских произведений из числа отобранных для эксперимента.

Подобные эксперименты активно проводятся на материале английского, итальянского, датского и других языков, но на русском материале этот вопрос не исследовался.

Существует достаточное количество алгоритмов машинного обучения для решения задачи классификации текстов. Рассмотрим некоторые из них.

1. Машина опорных векторов, или SVM (support vector machines).

Данный метод реализует бинарную классификацию [Platt, 1998], а именно разделяет входные векторы на две заданные группы. Например, разделяет отзывы на положительные и отрицательные. Лучшим типом данных для его применения является набор текстов с максимальной разностью между двумя категориями (в частности, мужские и женские тексты).

При его реализации точки в пространстве, выраженные векторами, разделяются гиперплоскостями. Задачей алгоритма является нахождение такой плоскости, которая разделяла бы векторы на две или более группы. В случае, когда категорий две, классификация называется бинарной, во всех остальных случаях используется термин «многоклассовая».

Алгоритм может работать как с уже обозначенными классами, тогда это будет «обучение с учителем», так и без них, в таком случае перед нами встает задача кластеризации.

Для обучения классификатора используется метод SMO (sequential minimal optimization), изменяющий параметры SVM для соответствия обучающим данным.

В нашем случае мы столкнулись с задачей бинарной классификации с учителем, поскольку у нас имеется обучающая выборка с двумя предварительно размеченными классами: мужские и женские тексты.

2. Наивный байесовский классификатор. По своей сути это вероятностный классификатор, который на основе ряда переменных и их зависимостей вычисляет зависимости по теореме Байеса, выражающей степень уверенности в истинности суждения.

Особенность алгоритма в том, что все классы рассматриваются независимо друг от друга, и это нужно рассматривать как допущение, поскольку в большинстве случаев данные каким-либо образом все же связаны друг с другом.

Классификатор на основе Наивного Байеса требует обучения на размеченных вручную данных. При всей своей простоте использование алгоритма позволяет достичь высокой точности в задачах, связанных с анализом текста.

3. Метод k-ближайших соседей. Он представляет собой алгоритм классификации, который оценивает сходство ближайших соседей и на основании этого сходства относит объект к одной из групп [Jain, 2010]. Все объекты находятся в n -мерном пространстве, где n является числом признаков, на основании которых производится классификация. Для обучения этого алгоритма используется выборка, в которой нужно предварительно обозначить отношение объектов к одному из классов.

4. «Случайный лес» (Random Forest). Этот алгоритм представляет собой множество решающих «деревьев», которые при реализации задачи классификации основываются на «голосовании» по выделенным в исходной текстовой выборке признакам. Дерево принятия решений – это метод со следующей структурой. Существует корневой узел, от которого отходят узлы. В них располагаются атрибуты, на основе которых распределяются входные данные. От узлов отходят листья, в которых находятся классифицированные объекты. Для принятия решения о классификации объекта необходимо пройти от корневого узла через все дочерние узлы, на основе которых принимается решение, к листу.

Алгоритм успешно показывает себя в решении задач кластеризации, классификации и регрессии, поскольку способен включать в себя большое количество данных и анализируемых классов [Liaw et al., 2002].

Наиболее популярными для решения задачи классификации текстов являются Weka – программное обеспечение для анализа данных, пакет для языка Python Gensim, использующийся для тематического моделирования, и библиотека Scikit-Learn, содержащая большое количество алгоритмов машинного обучения.

Для проведения эксперимента по выявлению лучшего алгоритма для классификации текстов художественной прозы 1980–2000 гг. по гендерному признаку нами была выбрана программа Weka, поскольку в ней наряду с удобным механизмом предобработки материала, состоящем в очистке от знаков препинания и сохранении в формате arff, представлен гибко настраиваемый набор фильтров (StringToWordVector, StringToNominal) других параметров (выбор стеммера, параметры токенизации, использования стоп-слов и др.).

Сбор корпуса и предобработка

Для решения задачи классификации нами был собран корпус, состоящий из произведений авторов второй половины XX – начала XXI в. В состав корпуса вошли следующие произведения:

- Быков Д. «Эвакуатор» (2005), «Ночные электрички» (1995), «Борис Пастернак» (2005), «Оправдание» (2001), «Орфография» (2003), «ЖД» (2006) – URL: <http://www.litres.ru>;
- Ерофеев В. «Пять рек жизни» (1958), «Русская красавица» (1990), «Энциклопедия русской души» (1999), «Мужчины: тираны и подкаблучники» (2008), «Дневник. Записки сума-

сшедшего» (1957), «Москва – Петушки» (1989), «Карманный апокалипсис», «Вальпургиева ночь, или Шаги Командора» (1989) – URL: <http://thelibrary.ru/>;

- Лимонов Э. «История его слуги» (1981), «Убийство часового» (1992), «Анатомия героя» (1998), «В плену у мертвецов» (2002), «Священные монстры» (2001), «Как мы строили будущее России» (2001), «Другая Россия» (2001), «Книга мертвых» (2000) – URL: <http://thelibrary.ru/>;

- Пелевин В. «ГКЧП как тетраграмматон» (2013), «Зеленая коробочка» (1996), «День бульдозериста» (1991), «Жизнь насекомых» (1993), «Затворник и Шестипалый» (1990) «Зигмунд в кафе» (2013), «Иван Кублаханов» (2013), «Мардонги» (2013), «Миттельшпиль» (2013), «Музыка со столба» (2013) – URL: <http://www.litres.ru/>;

- Петрушевская Л. «Бал последнего человека» (1995), «Время ночь» (1992), «Гирлянда птичек» (1996), «Как цветок на заре» (2002), «Перезимуем» (1995), «Два царства» (2009) – URL: <http://thelibrary.ru/>;

- Токарева В. «На каникулах» (2010), «Вместо меня» (1995), «Из жизни миллионеров» (2003), «Коррида» (1993), «Птица счастья» (2004) – URL: <http://thelibrary.ru/>;

- Толстая Т. «Кысь» (2000), «Лимпопо» (1990), «Ночь» (1987), «Не кысь» (2004), «Сюжет» (1991), «Надежда и опора» (1990) – URL: <http://thelibrary.ru/>;

- Улицкая Л. «Сонечка» (1992), «Медя и ее дети» (1996), «Сквозная линия» (2002), «Даниэль Штайн, переводчик» (2006), «Бедные родственники» (1996) – URL: <http://www.litres.ru/>;

- Щербакова Г. «Лизонька и все остальные» (2009), «Прошло и это» (2004), «Скелет в шкафу» (1996), «Вам и не снилось» (1979), «Трое в доме, не считая собаки» (2005) – URL: <http://thelibrary.ru/>.

Данные авторы были выбраны на основе различий стиля и тематики текстов. Это позволило учесть возможные различия, накладываемые этими характеристиками.

Корпус текстов собирался на основе открытых источников сети Интернет, и потому часть текстов представлена (в связи с авторскими правами) в неполном, но достаточном для проведения эксперимента объеме. Для анализа отбирались произведения разных жанров и длины: романы, повести, рассказы, сказки. Благодаря этому удалось добиться разнообразия текстового материала для обеих групп, о чем можно судить по различиям в употреблении частей речи и пунктуационных знаков (табл. 1).

Таблица 1

Различия в употреблении частей речи и пунктуационных знаков
в мужских и женских текстах корпуса (1960–2000)

Table 1

Differences in the use of parts of speech and punctuation marks
in the corpus of male and female texts (1960–2000)

Часть речи		Количество слов, %		Знак препинания	Количество знаков, %	
		Men	Women		Men	Women
ADV	наречие	5,46	5,54	!	6,92	9,84
ADVPRO	мест., нар.	1,9	2,1	?	10,9	10,73
CONJ	союз	8,8	10	,	0	0
INTJ	междомет.	0,095	0,15	.	86,63	81,33
PART	частица	4,58	5,16	:	1,06	1,73
SPRO	мест.-сущ.	9,06	10,55	;	2,07	0,64
V	глагол	16,35	18,87			
					Men	Women
Среднее количество слов в предложении					11,7	10,47
Количество предложений					16 902	18 881

Собранные корпуса женской прозы и мужских текстов впоследствии были объединены и разделены на обучающую и тестовую выборки. Размер обучающей выборки составил 89 000 словоупотреблений в 12 мужских и 12 женских текстах. Тексты обучающей выборки были размечены с точки зрения половой принадлежности авторов.

Тестовая выборка состояла из 20 мужских и 22 женских текстов и составила 317 000 словоупотреблений.

Все тексты в формате txt и кодировке utf-8 прошли процедуру предобработки – были очищены от знаков препинания, способных вызвать ошибку при прочтении данных в Weka. Затем они были использованы при создании arff файлов для последующей обработки.

Процедура классификации текстов

При использовании коллекции из 24 (training) и 42 (test) текстов была произведена процедура классификации текстов с учетом выбранных алгоритмов. При проведении эксперимента использовался метод 5-fold cross validation, разделяющий корпус на 5 равных частей, в которых одна часть становится объектом проверки обучения, а остальные четыре – тренировочными данными. После прохождения пяти итераций результаты объединялись, что позволило сделать оценку качества формирования обучающейся модели.

Для эксперимента по выявлению наиболее действенного алгоритма определения половой принадлежности автора нами были выделены пять наиболее популярных алгоритмов для решения этой задачи: «машина опорных векторов», «случайный лес», «наивный Байес», «BayesNet», «k-ближайших соседей».

Для построения моделей в Weka были использованы следующие классификаторы: Naive Bayes, BayesNet, SMO (вариация SVM), Lbk (k-ближайших соседей), LWL RandomForest.

В качестве фильтра использован StringToWordVector, преобразующий строки в набор числовых атрибутов. В качестве стеммера нами был выбран Lovins Stemmer, а для токенизации – WordTokenizer. Эта комбинация способствовала получению лучших результатов для работы моделей с использованием стандартных алгоритмов без дополнительных атрибутов.

Использование стоп-слов значительно снижало количество верно классифицированных текстов, что подтверждает суждение о том, что служебные слова в текстах русскоязычных авторов также являются одним из признаков стилового своеобразие [Морозов, 1915].

TFIDF и токенизация на n -граммы не давали улучшения качества работы моделей, поэтому было принято решение отказаться от их использования.

В ходе применения моделей, обученных на тестовой выборке с использованием обозначенных выше алгоритмов, были получены результаты, представленные в табл. 2.

Таблица 2

Результаты классификации мужских и женских текстов

Table 2

Results of classification of male and female texts

Тексты	Количество, ед.	SMO		Naive Bayes		BayesNet		KNN		LWL Random Forest	
		ед.	%	ед.	%	ед.	%	ед.	%	ед.	%
Мужские	20	15	75	15	75	13	65	15	75	14	70
Женские	22	15	68,18	15	68,18	12	54,54	12	54,54	14	63,63
Всего	42	30	71,43	30	71,43	25	59,52	27	64,29	28	66,67

Интерпретация результатов

Наиболее точным в процессе классификации художественных текстов на женские и мужские показал себя алгоритм SVM, ему удалось верно классифицировать 68,18 % женских и 75 % мужских текстов. Такие же результаты показал NaiveBayes.

Худший результат в задаче классификации женских текстов показали алгоритмы Bayes-Net и KNN. Эти инструменты плохо соотносят женскую прозу с нужной группой (54,54 %). Парадоксально, однако, то, что «мужские» тексты (и признаки маскулинности в языке и стиле) они распознали лучше – 65 и 75 % соответственно. Такой парадокс говорит либо о том, что язык и стиль женских текстов ориентируется на единые с мужскими жанрово-стилистические нормы художественной литературы, либо об инерции авторов, считающих себя феминистками, придерживаться на деле «патриархальных» канонов литературности.

Средние результаты показал RandomForest. Им было верно классифицировано 63,63 % женских текстов и 70 % мужских.

Таблица 2 демонстрирует, что наиболее эффективными классификаторами для художественной литературы по признакам маскулинности / феминности являются такие реализации алгоритмов, как NaiveBayes и SMO. Лучший результат в определении половой принадлежности авторов художественных произведений показали классификаторы на основе Наивного Байесовского классификатора и метода опорных векторов (71,43 %). Остальные алгоритмы демонстрировали не столь высокий результат.

Было выявлено 7 женских (6 из них – тексты Л. Петрушевской) и 2 мужских текста, половая принадлежность которых не классифицировались ни одной моделью. Образцы текстов этих авторов присутствуют в тренировочной выборке, но этого оказалось недостаточно для их корректной обработки (следует, пожалуй, отметить, что все они характеризуются разным объемом).

Безошибочно определяются стандартными алгоритмами классификации 11 женских и 8 мужских текстов, а в остальных случаях (4 и 8 текстов) классификаторы периодически допускают ошибку.

Как выяснилось, не все женские тексты обладают достаточно устойчивыми феминными сигналами, не вполне понятно также, на каком уровне (тексто- или лексико-семантическом) они располагаются. Одним из наиболее доступных и возможных путей улучшения моделей является учет пунктуационных особенностей текстов, что даст возможность в определенной степени выявить их синтаксическую специфику. Что касается различных аспектов тональности текста, реализуемой как через лексику, так и через синтаксис, то ее машинный анализ выполнить крайне трудно.

Как бы то ни было, мы можем сделать вывод, что универсального формального кода конструкта женской прозы на уровне текста, реализуемого в художественном творчестве всех писателей-женщин, не существует. Этот конструкт есть не более чем теоретическая гипотеза, выдвигаемая и защищаемая представителями нового литературоведения. Однако во многих женских произведениях мы можем выявить феминные сигналы разной интенсивности, которые и регистрируются в ходе процедуры автоматической идентификации.

Заключение

Самым важным результатом эксперимента является, пожалуй, то, что мужские тексты распознаются **всеми** алгоритмами гораздо лучше, чем женские. Это, во-первых, не позволяет списывать результат на погрешность в функционировании алгоритмов и, во-вторых, дает серьезную основу для формулировки новых гипотез относительно причин лучшей «видимости» мужского художественного языка / стиля.

Эксперимент, тем не менее, показал, что стандартные алгоритмы машинного обучения, применяемые для профилирования нехудожественных текстов, вполне удовлетворительно решают задачу идентификации и авторов художественной прозы в бинарной модели, основанной на половой принадлежности автора.

Список литературы

- Воробьева Н. В.** Женская проза 1980–2000 г.: динамика, проблематика, поэтика: Автореф. дис. ... канд. филол. наук. Пермь, 2006.
- Земская Е. А., Китайгородская М. А., Розанова Н. Н.** Особенности мужской и женской речи // Русский язык в его функционировании / Под ред. Е. А. Земской, Д. Н. Шмелева. М., 1993. С. 90–136.
- Морозов Н. А.** Лингвистические спектры, как средство для отличения плагиатов от истинных произведений того или другого известного автора и для определения их эпохи. 1915. URL: <http://www.textology.ru/library/book.aspx?bookId=1&textId=3>
- Пиперски А.** Гендер и язык: есть ли разница между мужской и женской речью? // Варламова Д. Проект Theory and Practice. 2006. 26 февр. URL: <https://theoryandpractice.ru/posts/9451-gender-language>
- Серго Ю. Н.** Женская проза России: особенности художественной философии // Филологический класс. 2006. № 16. С. 52–56.
- Стрельцова Е. А.** Языковые особенности художественных текстов В. Токаревой, Л. Петрушевской, Л. Улицкой // Вестник Череповец. гос. ун-та. 2014. № 3 (56). С. 137–140.
- Титарева Л. Д.** Женская проза как феномен современной российской культуры (на примере забайкальского края): Автореф. дис. ... канд. филол. наук. Чита, 2015.
- Argamon, S.** Automatically profiling the author of an anonymous text. *Commun. ACM*, 2009, vol. 52, no. 2, p. 119–123.
- Argamon S., Dawhle S., Koppel M., Pennebaker J.** Lexical Predictors of Personality Type. In: Proceedings of Classification Society of North America. St. Louis, MI, 2005.
- Diederich, J. J., Kindermann, J., Leopold, E., Paass, G.** Authorship Attribution with Support Vector Machines. *Applied Intelligence*, 2003, vol. 9, iss. 1. URL: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.7558&rep=rep1&type=pdf>
- Holmes, J., Meyerhoff, M.** The handbook of language and gender. Oxford, Blackwell Publ., 2003, vol. 25, p. 4–17.
- Jain, A. K.** Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 2010, vol. 31, no. 8, p. 651–666.
- Koppel, M., Argamon, S., Shimon, A. R.** Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 2002, vol. 17 (4). URL: <http://u.cs.biu.ac.il/~koppel/papers/male-female-llc-final.pdf>
- Kucukyilmaz, T., Cambazoglu, B. B., Can, F., Aykanat, C.** Chat mining: predicting user and message attributes in computer-mediated communication. *Information Processing & Management*, 2008, vol. 44 (4), p. 1448–1466. DOI 10.1016-j.ipm.2007.12.009
- Liaw, A. et al.** Classification and regression by randomForest. *R news*, 2002, vol. 2, no. 3, p. 18–22.
- Platt, J. C.** Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. In: Microsoft Research, Technical Report MSR-TR-98-14, April 21, 1998.
- Vel, O., Corney, M., Anderson, A., Mohay, G.** Language and Gender Author Cohort Analysis of E-mail for Computer Forensics. In: Second Digital Forensics Research Workshop, 2002. URL: https://www.dfrws.org/sites/default/files/session-files/paper-language_and_gender_author_cohort_analysis_of_e-mail_for_computer_forensics.pdf

References

- Argamon, S.** Automatically profiling the author of an anonymous text. *Commun. ACM*, 2009, vol. 52, no. 2, p. 119–123.
- Argamon S., Dawhle S., Koppel M., Pennebaker J.** Lexical Predictors of Personality Type. In: Proceedings of Classification Society of North America. St. Louis, MI, 2005.

- Diederich, J. J., Kindermann, J., Leopold, E., Paass, G.** Authorship Attribution with Support Vector Machines. *Applied Intelligence*, 2003, vol. 9, iss. 1. URL: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.7558&rep=rep1&type=pdf>
- Holmes, J., Meyerhoff, M.** The handbook of language and gender. Oxford, Blackwell Publ., 2003, vol. 25, p. 4–17.
- Jain, A. K.** Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 2010, vol. 31, no. 8, p. 651–666.
- Koppel, M., Argamon, S., Shimoni, A. R.** Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 2002, vol. 17 (4). URL: <http://u.cs.biu.ac.il/~koppel/papers/male-female-llc-final.pdf>
- Kucukyilmaz, T., Cambazoglu, B. B., Can, F., Aykanat, C.** Chat mining: predicting user and message attributes in computer-mediated communication. *Information Processing & Management*, 2008, vol. 44 (4), p. 1448–1466. DOI 10.1016-j.ipm.2007.12.009
- Liaw, A. et al.** Classification and regression by randomForest. *R news*, 2002, vol. 2, no. 3, p. 18–22.
- Morozov, N. A.** Linguistic spectra as a means to distinguish plagiarism from the true works of one or another famous author and to determine their era. 1915. (in Russ.) URL: <http://www.textology.ru/library/book.aspx?bookId=1&textId=3>
- Piperski, A.** Gender and language: is there a difference between male and female speech? In: Varlamova D. Project Theory and Practice. 2006. February 26. (in Russ.) URL: <https://theoryandpractice.ru/posts/9451-gender-language>
- Platt, J. C.** Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. In: Microsoft Research, Technical Report MSR-TR-98-14, April 21, 1998.
- Sergo, Yu. N.** Women's prose in Russia: features of artistic philosophy. *Philological class*, 2006, no. 16, p. 52–56. (in Russ.)
- Streltsova, E. A.** Language features of literary texts By V. Tokareva, L. Petrushevskaya, L. Ulitskaya. *Bulletin of Cherepovets State University*, 2014, no. 3 (56), p. 137–140. (in Russ.)
- Titareva, L. D.** Women's prose as a phenomenon of modern Russian culture (on the example of the transbaikal territory). Doctoral dissertation. Chita, 2015. (in Russ.)
- Vel, O., Corney, M., Anderson, A., Mohay, G.** Language and Gender Author Cohort Analysis of E-mail for Computer Forensics. In: Second Digital Forensics Research Workshop, 2002. URL: https://www.dfrws.org/sites/default/files/session-files/paper-language_and_gender_author_cohort_analysis_of_e-mail_for_computer_forensics.pdf
- Vorobyova, N. V.** Women's prose 1980–2000: dynamics, problems, poetics. Doctoral dissertation. Perm, 2006. (in Russ.)
- Zemskaya, E. A., Kitaygorodskaya, M. A., Rozanova, N. N.** Features of male and female speech. In: Zemskaya, E. A., Shmeleva, D. N. (eds.). Russian language in its functioning. Moscow, 1993, p. 90–136. (in Russ.)

Материал поступил в редколлегию
Date of submission
16.09.2019

Сведения об авторах / Information about the Authors

Хазова Анастасия Борисовна, ассистент Школы лингвистики Национального исследовательского университета «Высшая школа экономики» (ул. Мясницкая, 20, Москва, 101000, Россия)

Anastasiya B. Khazova, Instructor, National Research University Higher School of Economics (20 Myasnitskaya Str., Moscow, 101000, Russian Federation)

ahazova@hse.ru

ORCID 0000-0002-1612-2016