Санкт-Петербургский Государственный Университет
Математико-механический факультет

Кафедра системного программирования

АЛГОРИТМ КЛАССИФИКАЦИИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ ПО ВОЗРАСТУ И ГЕНДЕРНОМУ ПРИЗНАКУ АВТОРА

Дипломная работа студентки 545 группы

Тумановой Кристины Сергеевны

Научный руководитель	…… / подпись /	д.ф.-м.н., проф. Б.А. Новиков
Рецензент	…… / подпись /	к.ф-м.н., доцент К.В. Вяткина
“Допустить к защите” заведующий кафедрой,	…… / подпись /	д.ф.-м.н., проф. А.Н. Терехов

Санкт-Петербург
2011

Saint-Petersburg State University
Mathematics and Mechanics Faculty

Software Engineering Department

ALGORITHM OF RUSSIAN TEXT CLASSIFICATION BY AUTHOR'S GENDER AND AGE

Graduate paper by

Kristina Tumanova

545 group

Scientific advisor	……	Dr. Sci., Prof. B. A. Novikov
Reviewer	……	Ph. D., Assoc. Prof. K.V. Vyatkina
"Approved by" Head of Department	……	Dr. Sci., Prof. A. N. Terekhov

Saint Petersburg
2011

Оглавление

1. Введение 4
2. Гендерная лингвистика 7
3. Постановка задачи 11
4. Обзор литературы 12
5. Предлагаемое решение 23
   5.1. Плоская («Flat») классификация 23
   5.2. Иерархическая классификация 25
6. Экспериментальная среда 26
   6.1. Корпус текстов 26
   6.2. Представление текстов 29
   6.3. Алгоритмы классификации 35
   6.4. Обработка текстов 37
7. Эксперименты 38
   7.1. Результаты и их анализ 39
8. Заключение 46
9. Список использованной литературы 47
10. Приложения 55
   Приложение 1 55
   Приложение 2 56
   Приложение 3 57
   Приложение 4 58
   Приложение 5 59
   Приложение 6 59
   Приложение 7 60
   Приложение 8 60

1. Введение

В последнее десятилетие область обработки естественных языков (англ. Natural Language Processing) и, в частности, её подраздел «классификация текстов» развивается очень интенсивно [38]. Это во многом связано с тем, что с каждым годом объем информации, хранимой на электронных носителях, значительно возрастает, и требуются эффективные алгоритмы для обработки и анализа документов, написанных на естественных языках. Усовершенствование алгоритмов, в свою очередь, возможно благодаря увеличению мощности и производительности современных компьютеров.

Существующие алгоритмы классификации можно использовать не только непосредственно для классификации текстов, но и, например, для извлечения из них дополнительной информации. В данной работе будет рассматриваться одно из таких направлений, а именно, автоматическое профилирование автора анонимного текста. Профилирование автора – это установление некоторых значимых характеристик человека на основе написанного им текста. В рамках автоматического профилирования автора подобные задачи чаще всего решаются с помощью построения на основе методов машинного обучения моделей, которые для любого входного текста будут возвращать необходимую информацию об авторе этого текста, например, пол, возраст, принадлежность к той или иной социальной группе, психологический тип или нечто другое.

Автоматическое профилирование имеет множество применений. Одно из них – это судебная автороведческая экспертиза. Получение каких-либо данных о личности преступника позволяет значительно сузить область поиска и сэкономить время. Актуальной в настоящий момент является и проблема поиска интернет-злоумышленников, когда у следствия нет никаких других улик, кроме нескольких сообщений с угрозами или обвинениями.

Согласно [60], «…интенсивное распространение электронных изданий в сети Интернет в последнее десятилетие привело к резкому увеличению нарушений прав правообладателей и авторских прав. Литературные и научные произведения неправомерно копируются, заимствуются, иногда слегка редактируются и переиздаются под другим именем». В этих условиях важной задачей становится выявление незаконного присвоения авторства. Также широко известны несколько проблем, касающихся авторства литературных произведений. Одна из наиболее активно исследовавшихся – проблема авторства романа «Тихий Дон», опубликованного Михаилом Александровичем Шолоховым в 1928 году. Немало обсуждений возникло и относительно критико-литературного творчества А.С. Пушкина, Ф.М. Достоевского, Н.Г. Чернышевского, В.Г. Белинского. Большое количество проблем возникает из-за того, что значительная часть произведений остается неопубликованной при жизни автора, либо бывает опубликована с неточностями и искажениями.

В литературных и судебных автороведческих экспертизах чаще выполняется атрибуция текстов, то есть, определение авторства текста. Несмотря на то, что таким образом можно установить непосредственно личность автора текста, алгоритмы профилирования могут оказаться полезными в случаях, когда у экспертов нет никакой дополнительной информации, и сделать предположения о возможных авторах трудно. В подобных случаях важны любые данные об авторе.

Профилирование автора также находит свое применение в сочетании с сентимент-анализом (англ. Sentiment Analysis) и анализом высказываний (англ. Opinion Mining). В этих областях решаются проблемы выявления в тексте эмоционально окрашенной лексики и эмоциональной оценки разного рода объектов. Например, получив данные о мнениях пользователей разных возрастов относительного какого-либо продукта или устройства, производящая компания может точнее определить целевую аудиторию и

скорректировать маркетинговую политику. Также это может применяться для анализа политических настроений людей разных социальных групп или показа на сайтах рекламных баннеров, ориентированных на конкретного пользователя.

В зависимости от области применения (судебная экспертиза, медицина, маркетинг) набор извлекаемых характеристик в автоматическом профилировании может меняться. Но одними из наиболее значимых для идентификации личности автора характеристик являются его гендерный признак и возраст. При этом очевидно, что люди одного пола, но разных возрастов имеют разный словарный запас, используют разное количество речевых оборотов и строят предложения различной сложности. Подобные связи можно обнаружить и между другими признаками, например, между возрастом и уровнем образования. Поэтому результаты выделения какой-то одной характеристики без учета другой могут оказаться недостаточно точными.

Задача данной дипломной работы заключается в исследовании возможности профилирования авторов анонимных текстов по нескольким признакам одновременно. Для её решения будут предложены два подхода. Основанные на них алгоритмы будут реализованы. Их сравнение будет выполнено на корпусе русских текстов, поэтому в рамках данной работы будет также проведено исследование возможностей профилирования автора в применении к текстам на русском языке. По результатам экспериментов будет проведен анализ предложенных подходов и классификации русских текстов в целом.

2. Гендерная лингвистика

Речь является для человека одним из главных средств для самовыражения, которое направлено на создание и поддержание социальной и личностной идентичности [82]. Содержание речи отдельного человека и её стилевые особенности создают у собеседников психологический портрет высказывающегося.

На речь человека оказывает влияние множество факторов, среди которых присутствуют как социально-демографические, так и психофизиологические особенности человека: пол, возраст, уровень образования, социальная ориентация, психологический тип личности, эмоциональное состояние и др. [78].

Очевидно, что одним из самых значимым среди них является пол.

Следует сделать замечание относительно понятий «пол» и «гендер». В современной социальной науке эти два понятия имеют разное значение. Слово «пол» отождествляется с понятием «биологический пол», которое используется для обозначения тех анатомо-физиологических особенностей людей, на основе которых человеческие существа определяются как мужчины или женщины. Под словом «гендер» понимается совокупность социальных и культурных норм, которые общество предписывает выполнять людям в зависимости от их биологического пола [65]. Иначе, гендер – это социальный пол, определяющий поведение человека в обществе и то, как это поведение воспринимается [58]. Но в данной работе эти два понятия будут использоваться как равнозначные.

В области гендерной лингвистики было опубликовано множество работ, в которых подтверждается наличие различий в речи мужчин и

женщин, причем как в устной, так и в письменной. Наиболее часто упоминаемыми различиями являются следующие:

- мужчины чаще, чем женщины, обращаются к стилистически сниженной и бранной лексике [57, 66, 67];

- в речи женщин чаще, чем в речи мужчин, встречаются вопросительные и восклицательные высказывания [57, 67];

- женщины чаще мужчин выражают неуверенность, и делают это с помощью модальных слов «наверное», «по-моему», таких выражений, как «мне кажется», «я не знаю …» и с помощью уточняющих конструкций «ну разве я не права?», «так ведь?» [57, 62, 68];

- женская речь более эмоциональна и содержит больше эмоциональной оценочной лексики, чем мужская [57, 68, 72, 75];

- словарь женщин имеет некое большое ядро, широко используемое всеми женщинами, у мужчин слой общеупотребительной лексики меньше, но они лучше владеют периферийными разделами словаря, иначе говоря, в речи мужчин обнаруживается больше слов, которые встречаются один или два раза [56, 69, 75];

- женщины чаще мужчин используют в своей речи речевые клише и штампы [62,75];

Некоторые исследователи также отмечают следующие различия:

- мужчины чаще, чем женщины, употребляют существительные, местоимения и глаголы; женщины в своей разговорной речи чаще, чем мужчины, используют прилагательные, наречия и союзы [57];

- в речи мужчин чаще, чем в речи женщин, встречаются неполные предложения [57];

Т.А. Гомон в работе [62] также указала следующие характеристики, отражающие психолингвистические навыки мужской и женской письменной речи, которые не могут быть относительно легко сфальсифицированы, а значит, представляют особые возможности для идентификации личности автора анонимного текста:

- в женской письменной речи больше вводных слов, определений, обстоятельств, местоименных подлежащих, дополнений;

- высокочастотным для женщин является использование конструкций наречие+наречие («слишком безжалостно», «очень хорошо») и наречие+прилагательное («ужасно красивый»), простых и сложносочиненных предложений, синтаксических оборотов с двойным отрицанием;

- женскую речь отличает частое использование знаков пунктуации и высокая эмоциональная окраска речи в целом.

Е.С. Ощепкова указала еще одну особенность женской речи, которая при намеренном искажении будет изменяться незначительно, это более частое использование женщинами отрицательных частиц [75].

Отдельной темой для исследований в последнее время становится изучение влияния гендерного фактора на электронную коммуникацию (электронные письма, форумы, чаты, блоги).

Несколько исследований в этой области проводила Е.И. Горошко. На основе 100 сообщений электронной почты ею были выявлены следующие статистически значимые характеристики:

- мужские предложения длиннее женских;

- женщинам свойственны более вежливые формы обращений;

- женщины чаще употребляют личные местоимения;

- словарь женщин беднее;

- в речи мужчин встречается больше слов с текстовой частотой один и два.

По остальным показателям достоверных статистических различий получено не было [64].

В другой её работе материалом для исследования послужил банк данных почтовых сообщений слушателей некоторого дистанционного курса. [63]. Статистически значимыми получились следующие характеристики:

- женские тексты длиннее;

- женщины используют гораздо больше вежливой лексики;

- смайлики используются исключительно женской виртуальной аудиторией;

- словарь женщин беднее;

- в речи мужчин встречается больше слов с частотой один и два.

По остальным показателям достоверных статистических различий найдено не было.

Помимо указанных особенностей мужской и женской речи, выявленных разными исследователями, существуют и такие, которые устанавливались одними учеными и не подтверждались другими, например, Т.Б. Крючковой были получены другие данные о частоте использования мужчинами и женщинами разных частей речи [69].

На основе этих данных и дополнительных исследований лингвисты делают вывод о том, что на речь во многом оказывают влияние и другие факторы: возраст, профессия, социальное положение, вид коммуникации и пр. [61, 63].

3. Постановка задачи

Исходя из результатов лингвистических исследований, описанных в предыдущей главе, можно сделать вывод о том, что в речи мужчин и женщин действительно существуют заметные отличия. При этом они могут быть представлены в форме, приемлемой для компьютерной обработки, а значит, их можно использовать и в методах машинного обучения в применении к классификации текстов.

Другие характеристики личности автора, по данным лингвистов, также оказывают влияние на речь человека. Поэтому, используя эту информацию при классификации, можно надеяться на получение более точных данных об авторе анонимного текста.

Задача данной дипломной работы – исследование возможности проведения классификации русских текстов по гендерному признаку и возрасту автора, учитывая одновременное влияние этих факторов на речь человека. В рамках этой задачи требуется разрешить вопрос выбора достоверных характеристик для классификации, которые бы отражали различия в текстах людей разных полов и возрастов.

Стоит сказать о наиболее очевидном способе автоматического определения пола автора русского текста с помощью глагольных окончаний. Данный метод рассматриваться не будет в силу следующих причин. Во-первых, задача состоит в определении характеристик автора текста, который может принадлежать произвольному жанру. Текст может являться частью художественного произведения, научного исследования или относиться к разговорно-бытовому стилю письма. Использование же глагольных окончаний возможно лишь для текстов последнего типа. В текстах другого жанра и стиля очень часто речь автора состоит из безличных конструкций («я думаю», «мне кажется», «меня удивило»), поэтому данный поход

оказывается неприменимым. Во-вторых, глагольные окончания являются своеобразными маркерами гендера лишь ввиду морфологических особенностей русского языка и не отражают глубинных свойств письма мужчин и женщин. Данный признак является поверхностным, он легко может быть сфальсифицирован, а значит, не может рассматриваться как достоверный, особенно в судебных и литературоведческих экспертизах.

Таким образом, задача стоит в разработке алгоритмов, которые бы не имели упомянутых ограничений и при этом основывались на достоверной информации.

4. Обзор литературы

Итак, задача профилирования автора заключается в построении на основе алгоритмов классификации модели, которая для любого входного текста будет выдавать информацию об определенных характеристиках его автора.

В большинстве случаев для классификации текстов с помощью методов машинного обучения нужно осуществить следующие шаги:

- Выбрать характеристики, по которым будет производиться классификация, и построить для каждого текста соответствующее векторное представление;
- Выбрать алгоритмы классификации;
- Выбрать методы оценки работы алгоритма.

Далее в соответствии с выбранным алгоритмом строится модель и оценивается точность или другие параметры классификации.

Важным вопросом при этом является выбор характеристик. Характеристики должны хорошо отражать особенности текстов, а именно

различия между текстами, относящимся к разным классам, и близость текстов одного класса. От этого зависит соответствие модели реальности и точность определения класса для новых входных данных.

Когда говорят о классификации текстов, обычно подразумевают классификацию текстов по содержанию, по теме, о которой идет речь в тексте, например, политика, медицина, культура. В подобных случаях необходимо использовать характеристики, зависящие от контекста (англ. «Content-based»). Профилирование автора отличается от данной задачи тем, что требуется выбрать характеристики, которые, наоборот, не зависят от содержания текстов. При этом они должны отражать авторский стиль письма. Для такой классификации и таких характеристик в англоязычных работах используется термин «Style-based».

Помимо профилирования автора, «style-based» классификация используется для автоматической атрибуции текстов, когда по анонимному тексту следует установить автора из заранее заданного набора имен. Обычно в этом случае к характеристикам дополнительно предъявляют требование того, чтобы человек во время написания текста не мог сознательно их контролировать, они должны быть такими, о которых автор даже не задумывается. В научной литературе такие характеристики называются «авторскими инвариантами» [79].

Таким образом, задачи, решаемые в рамках и атрибуции текстов, и профилирования автора, сводятся к выявлению значимых признаков письменной речи авторов тех или иных групп и использованию этих признаков для классификации текстов. Эти проблемы частично относятся к стилометрии – научной дисциплине, в основном, занимающейся вопросами исследования стилистики текстов и их статистического анализа.

Поэтому сначала рассмотрим работы, которые относятся к атрибуции текстов и касаются вопросов классификации текстов по стилю в общем.

Довольно много научных работ посвящено атрибуции литературных текстов на русском языке. Согласно обзору, сделанному Е.С. Родионовой [76], исследования в этой области начались еще в начале XX века. Так Н.А. Морозов в 1915 году для выявления индивидуального стиля письма предложил использовать информацию о частоте употребления в тексте служебных слов, т.е. предлогов, союзов и частиц [73]. Исследовались также возможности атрибуции текстов на основе количества повторений той или иной части речи [83] и на основе синтаксических особенностей [71, 77]. В частности, в [77] исследовались графы синтаксических связей в рамках типических фраз и предложений. Но Е.С. Родионова указывает недостаток предложенного подхода в том, что с помощью него можно получить информацию только о предложениях, но не о тексте в целом.

В рамках решения проблемы авторства «Тихого Дона» группа шведских и норвежских ученых одна из первых предложила методику, основанную на автоматическом анализе частотных словарей и статистических данных текстов [81]. Были изучены такие характеристики как частота использования различных сочетаний грамматических классов, длина предложений, длина слов, насыщенность словаря, и была доказана возможность их применения к задаче атрибуции текстов.

Интересна работа двух ученых В.П. Фоменко и Т.Г. Фоменко [79]. В ней описывается статический эксперимент, проведенный для более 20 классических писателей, и устанавливается, что доля служебных слов является авторским инвариантом и может быть использована для атрибуции текстов. Другие исследованные признаки (длина предложений, длина слов, частота употребления существительных, глаголов, прилагательных, предлога «в», частицы «не»), по мнению ученых, оказывают существенно меньшее влияние на индивидуальный авторский стиль.

В работе [80] предлагается применимый для большого числа авторов подход, который основан на исследовании последовательностей букв текста как реализации цепи Маркова. Ученым были проведены эксперименты для большого числа авторов (82), и получена достаточно высокая точность алгоритма – 84% для двухбуквенных сочетаний. Под точностью понимается отношение количества текстов, автор которых был определен правильно, к общему числу текстов.

Продолжением данного исследования стала работа [70], в которой для построения цепей Маркова использовались пары грамматических классов. В экспериментах по атрибуции текстов для 82 авторов точность алгоритма составила 73%.

Огромное количество исследований в области стилометрии проводилось и зарубежными учеными. Подробные обзоры методов и подходов, предлагавшихся для выявления индивидуального стиля письма, сделали D. Holmes в 1998 году [15] и E. Stamatatos в 2009 году [40]. Большое количество из упоминавшихся в этих обзорах работ посвящено анализу одной или нескольких характеристик текста и исследованию возможности проведения классификации текстов на их основе. Перечислим основные характеристики, которые в разное время применялись теми или иными учеными:

Лексика

- Частота использования служебных слов (артикли, предлоги, союзы) [5, 6, 34, 52]
- Информация о позиции слов и пар слов в отдельных предложениях и в тексте целиком [13, 29, 30, 31, 32]
- Длина предложений и слов [28, 49]
- Богатство словарного запаса [16, 43, 50]

Символы

- Частота использования символов (букв, цифр, знаков пунктуации) [43, 54]
- N-граммы символов [17, 18, 26, 36, 39]

Синтаксис

- Части речи [3, 23, 53]
- Синтаксические связи [7, 11, 14]

Семантика [6,11, 27]

Интересную работу в 2003 году представили M. Koppel и его коллеги [21]. Для «style-based» классификации текстов ученые предложили использовать «нестабильные» характеристики, т.е. такие, которые могут быть заменены в тексте так, что смысл текста при этом не изменится. Самими простыми характеристиками, обладающими таким свойством, является большинство служебных слов. Примечательно то, что введенная авторами мера стабильности может быть применена к любому типу характеристик: лексическим, синтаксическим и др. Эксперименты по атрибуции текстов и определению пола автора показали достаточно хорошую точность, около 85% и 79% соответственно.

Как уже было сказано, автоматическое профилирование автора относится к классификации текстов по стилевым особенностям. В связи с этим в работах по профилированию автора чаще всего используются стандартные характеристики, относящиеся к тем или иным описанным группам. Однако некоторые ученые обращались и к контентно-зависимым характеристикам, таким, как частота появления слов или групп слов [4, 20, 24, 33, 37, 41, 48, 51]. Обычно подобные характеристики не дают высоких результатов в силу того, что тексты корпуса относятся к разным жанрам и темам. Однако в работе [20] использование таких характеристик для

определения принадлежности автора текста к религиозной группе и идеологической организации стало оправданным, так как все тексты были одной тематики.

Важным вопросом в автоматическом профилировании также является выбор алгоритма классификации. Помимо широко применяемых алгоритмов типа SVM (Support Vector Machines) и Naïve Bayes (Наивный Байесовский классификатор), некоторые исследователи использовали алгоритмы, позволяющие одновременно с классификацией проводить и отбор информативных признаков [21, 22, 37]. Подобные алгоритмы дают хорошие результаты при использовании большого набора характеристик.

В таблице 1 приводится краткая информация о наиболее значительных исследованиях в области автоматического профилирования. SB и CB – обозначения для стилистических характеристик («Style-based») и характеристик, зависящих от содержания текстов («Content-based»). Точность – параметр, равный отношению количества текстов, для которых извлекаемый признак был определен правильно, к общему числу текстов.

<table>
<tr><th>Признаки</th><th>Тип текстов (Язык)</th><th>Характеристики</th><th>Алгоритмы классификации</th><th>Точность</th></tr>
<tr><td colspan="5">O. de Vel et al. (2002) [44]</td></tr>
<tr><td>Пол, родной язык</td><td>E-mail (Англ.)</td><td>SB (длина предложений, длина слов, богатство словарного запаса, различные метрики, статистика относительно появления тех или иных символов, служебные слова, специфические характеристики, напр., количество вложений в письмо или количество</td><td>SVM</td><td>Пол: 71,1% (максимум)*
Язык: 80,8% (максимум)*

*Вместо точности для оценки использовалась мера, равная 2RP/(R+P), где R – Recall, P – Precision. Recall = TP / (TP + FN). Precision = TP / (TP + FP)
<table><tr><td>Actual \ Predicted</td><td>YES</td><td>NO</td></tr><tr><td>YES</td><td>TP</td><td>FN</td></tr><tr><td>NO</td><td>FP</td><td>TN</td></tr></table></td></tr>
</table>

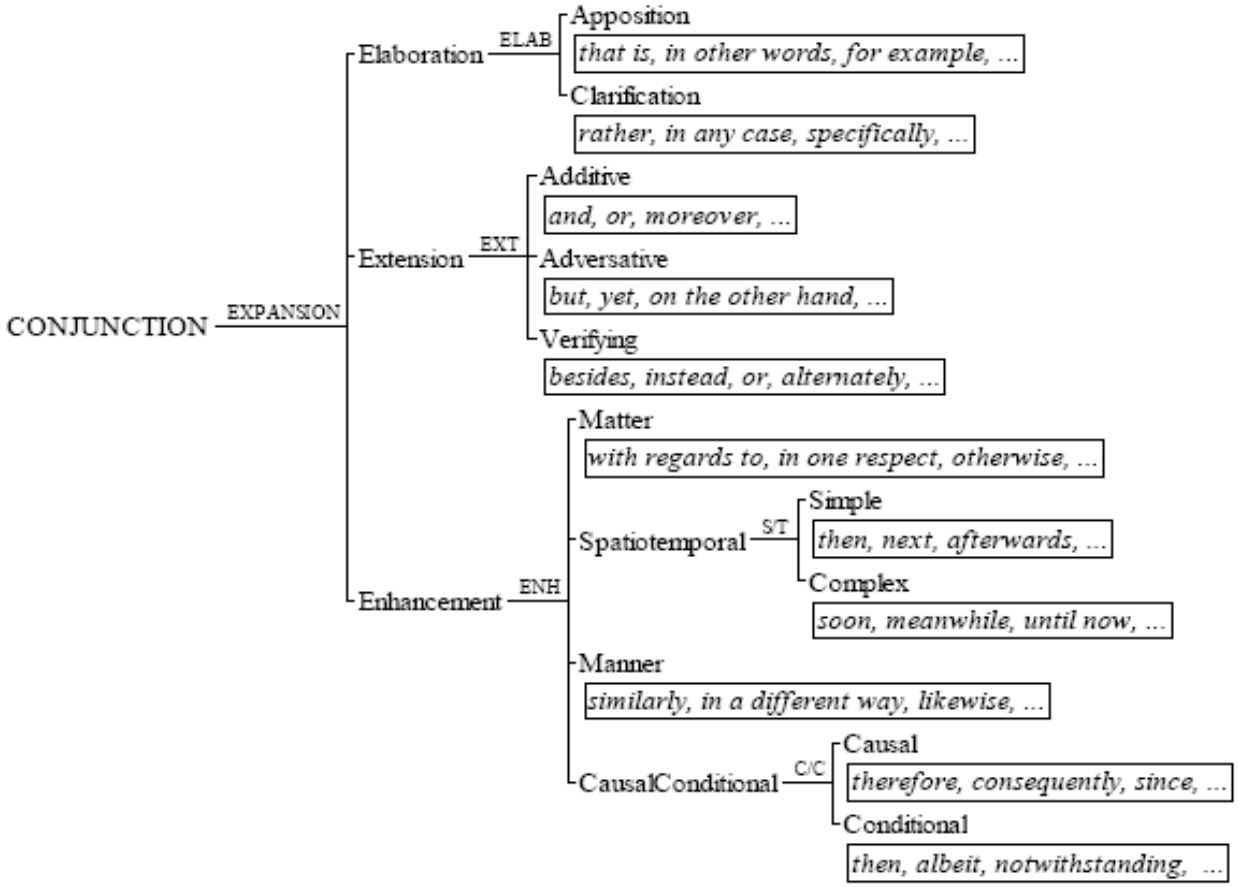
		приветсвий, особенности мужской и женской речи)		
M.Koppel et al. (2002) [22]				
Пол	Худож-ая и не худож-ая литература (Англ.)	SB (служебные слова, N-граммы частей речи для N = 1, 2, 3)	Exponential Gradient (частный случай Balanced Winnow)	Пол (худож-ая): 79,5%; Пол (не худож-ая): 82,6%; Пол (худож-ая и не худож-ая): 77,3%
M. Koppel et al. (2003) [21]				
Пол	Худож-ая и не худож-ая (Англ.)	SB (служебные слова, N-граммы частей речи)	Balanced Winnow	Пол: 72%
S. Argamon et al. (2005) [2]				
Нервозность, экстравертность	Эссе (Англ.)	SB (служебные слова, семантическая информация)	SVM/SMO	Нервозность: 58%; Экстравертность: 58%
J. Schler et al. (2006) [37]				
Пол, возраст	Блоги (Англ.)	SB (части речи, служебные слова, специальные слова) CB (1000 наиболее часто встретившихся в корпусе слов)	Multi-Class Real Winnow	Пол: 80,1%; Возраст: 76%
X.Yan, L.Yan (2006) [48]				
Пол	Блоги (Англ.)	SB (цвет фона и шрифты, использованные на странице блога; знаки пунктуации; смайлики) CB (слова)	Naïve Bayes	Пол: 65%
D. Kobayashi et al. (2007) [19]				
Пол	Блоги (Япон.)	SB (N-граммы частей речи для N = 1…10)	SVM	Пол: 69%
D. Estival et al. (2007a) [9]				
Пол, возраст, родной язык, образование, страна, покладистость, честность, экстравертность,	Email (Англ.)	SB (длина слов, знаки пунктуации, прописные и заглавные буквы; служебные слова, части речи, именованные	SVM, Random Forest, Bagging, Instance-Based learner with	Пол: 69%; Возраст: 56%; Язык: 84%; Образование: 80%; Страна: 81%; Поклад-ть: 53%; Честность: 54%;

нервозность, открытость		сущности; количество параграфов; др.)	fixed neighborhood	Экстравертность: 57%; Нервозность: 54%; Открытость: 55%
D. Estival et al. (2007b) [10]				
Пол, возраст, родной язык, образование, экстравертность, склонность к лжи, нервозность	Email (Араб.)	SB (длина слов, знаки пунктуации; именованные сущности; части речи и др. морфологические признаки; лексические признаки)	SVM/SMO, Bagging	Пол: 81%; Возраст: 72%; Образование: 94%; Экстравертность: 54%; Склонность к лжи: 52%; Нервозность: 55%
J. Lin (2007) [25]				
Пол, возраст	Чаты (Англ.)	SB (длина предложений, богатство словарного запаса, знаки пунктуации, смайлики)	Naive Bayes	Пол: 53% - 55%; Возраст: от 33% до 88% для разных случаев
T. Kucukyilmaz et al. (2008) [24]				
Пол, возраст, школа	Чаты (Турец.)	SB (длина сообщений, слов, знаки пунктуации, частота символов, смайлки, богатство словарного запаса) CB	k-Nearest neighbor, Naive Bayes, Patient rule induction method, SVM	Пол: 72% - 82%; Возраст: 27% - 75%; Школа: 28% - 69%
R.Strous, M. Koppel et al. (2008) [41]				
Шизофрения	- Эссе (Англ.)	SB (буквенные 3-граммы, повторения слов) CB (25 наиболее часто встретившихся в корпусе слов)	SVM, Bayesian regression	Шизофрения: 83,3%
S. Argamon et al. (2009) [4]				
Пол, возраст, родной язык, нервозность.	Для пола и возраста: блоги (Англ.); Для языка: International Corpus of Learner English (Англ.);	SB (части речи, служебные слова CB (1000 наиболее часто встретившихся в корпусе слов)	Bayesian Multinomial Regression	Пол: 76,1%; Возраст: 77,7%; Язык: 82.3%; Нервозность: 65,7%

	Для личных качеств: эссе (Англ.)			
M. Koppel et al. (2009) [20]				
Религиозная группа, идеологическая организация	- (Араб.)	CB (1000 наиболее часто встретившихся в корпусе слов)	Bayesian Multinomial Regression	Религиозная группа: 100%; Идеологическая организация: 99%
A. Mukherjee, B. Liu (2010) [33]				
Пол	Блоги (Англ.)	SB (части речи, специфические слова, особенности мужской и женской речи) CB (группы слов)	Naïve Bayes, SVM classification SVM Regression	Пол: 88,7%
C.Zhang, P. Zhang (2010) [51]				
Пол	Блоги (Англ.)	SB (длина слов, длина предложений, части речи) CB (группы слов)	Naïve Bayes, SVM, Linear discriminant analysis	Пол: 72,1%

Таблица 1. Обзор литературы в области профилирования автора.

Практически единственной работой, в которой в качестве характеристик использовались семантические особенности текста, является работа [2]. Характеристики, используемые в этой работе, основываются на теории Системной Функциональной Грамматики языка, разработанной в 60-е года XX века (Systemic Functional Grammar, SFG). В соответствии с этой моделью, язык представляется как набор некоторых схем (направленных ациклических графов), отражающих возможные значения слов определенной группы. Например, союзные слова английского языка могут быть организованы в следующую иерархию:



Согласно этой схеме текст, следующий за союзным словом, может либо углублять смысл сказанного (Elaboration), либо расширять (Extension), либо уточнять или ограничивать (Enhancement). Каждый из этих вариантов разделяется еще на несколько случаев. В [2] использовались характеристики, которые отражают количество слов, принадлежащих каждой из этих подгрупп, и подобные величины для схем других типов. Несмотря на то, что точность алгоритмов, основанных на данных характеристиках, оказалась не очень высока (58%), в последующей своей работе та же исследовательская группа добилась улучшения результатов за счет добавления характеристик, зависящих от контекста [4]. Результаты новых экспериментов показали точность алгоритма 65,7%.

Важное исследование было проведено в [33]. Во-первых, в работе был описан новый метод для выбора информативных характеристик, в котором для оценки характеристик одновременно используется несколько

стандартных критериев (Information Gain, Mutual Information, χ^2-Statistic). Эксперименты показали, что использование этого метода по сравнению с выбором характеристик только по одному критерию дает увеличение точности классификации от 6,8% до 14,4% для разных алгоритмов и критериев. Во-вторых, вместо стандартных характеристик, основанных на количестве появлений в тексте N-грамм частей речи для N = 1, 2, 3, в этой работе использовались наиболее часто встречающиеся последовательности частей речи неограниченной длины, извлеченные с помощью предложенного алгоритма. В итоге, точность классификации английских текстов по гендерному признаку составила 88,7%. Это выше, чем во всех других работах, посвященных той же проблеме.

Задача данной работы заключается в исследовании классификации текстов по возрастному и гендерному признакам автора, поэтому стоит отдельно выделить исследования, выполненные в этих направлениях. В таблицах 2 и 3 приводится информация о работах, имеющих отношение к классификации текстов по полу и возрасту автора соответственно.

Авторы (год)	**Тип текстов (Язык)**	**Точность**
M.Koppel et al. (2002)	Худож-ая и не худож-ая (Англ.)	77,3%; 79,5%
J. Schler, M. Koppel et al. (2006)	Блоги (Англ.)	80,1%
X.Yan, L.Yan (2006)	Блоги (Англ.)	65%
D. Kobayashi, N. Matsumura, M. Ishizuka (2007)	Блоги (Япон.)	69%
D. Estival et al. (2007a)	Email (Англ.)	69%
D. Estival et al. (2007b)	Email (Араб.)	81%
J. Lin (2007)	Чаты (Англ.)	53% - 55%
T. Kucukyilmaz et al. (2008)	Чаты (Турец.)	72% - 82%
S. Argamon, M. Koppel et al. (2009)	Блоги (Англ.)	76,1%
A. Mukherjee, B. Liu (2010)	Блоги (Англ.)	88,7%
C.Zhang, P. Zhang (2010)	Блоги (Англ.)	72,1%

Таблица 2. Работы по классификации текстов по полу.

Авторы (год)	Тип текстов (Язык)	Точность
J. Schler, M. Koppel et al. (2006)	Блоги (Англ.)	76%
D. Estival et al. (2007a)	Email (Англ.)	56%
D. Estival et al. (2007b)	Email (Араб.)	72%
J. Lin (2007)	Чаты (Англ.)	33% - 88%
T. Kucukyilmaz et al. (2008)	Чаты (Турец.)	27% - 75%
S. Argamon, M. Koppel et al. (2009)	Блоги (Англ.)	77,7%

Таблица 3. Работы по классификации текстов по возрасту.

Как видно из таблиц, ни в одном из исследований не проводились эксперименты на корпусе из русских текстов, а также не учитывались взаимосвязи признаков личности автора и их влияние на речь, т.е. все признаки извлекались независимо.

5. Предлагаемое решение

В данной работе предлагаются подходы для автоматического профилирования автора, которые будут учитывать возможные взаимосвязи характеристик личности автора. Для простоты будем рассматривать случай двух характеристик, но модели могут быть естественным образом расширены. Каждый признак автора, который необходимо установить, назовем измерением. Тогда для классификации текстов по двум измерениям одновременно могут быть использованы следующие подходы:

5.1. Плоская («Flat») классификация

Имея в первом измерении N групп, а во втором M групп, определим (N * M) новых классов, которые представляют собой всевозможные сочетания этих групп. Далее в соответствии с выбранным алгоритмом классификации и

протоколом тестирования выполним классификацию текстов. Иллюстрация данного алгоритма для N = 2 и M = 3 представлена на рисунке 1.

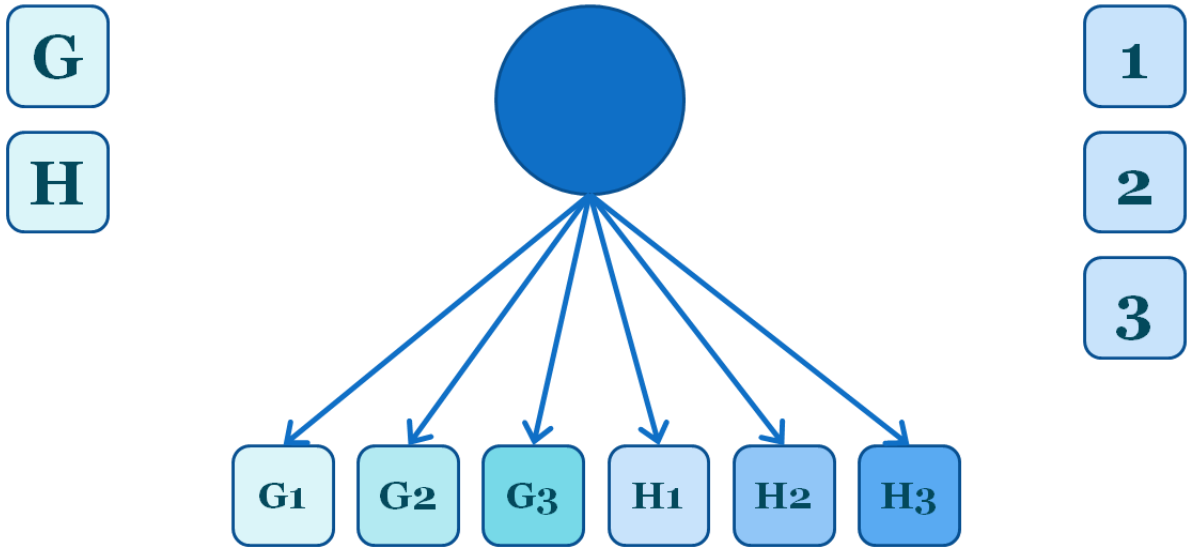


Рисунок 1. Плоская классификация.

Этот метод обладает следующими особенностями. Во-первых, при построении модели на каждый класс приходится относительно небольшой объем входных данных. Во-вторых, при больших числах N и M точность классификации может оказаться низкой, так как между некоторыми классами будут существовать лишь незначительные различия. Это затруднит определение класса, к которому относится рассматриваемый объект.

Данный метод, на самом деле, является лишь модификацией классической классификации, когда каждому объекту сопоставляется вектор из нескольких значений, обозначающих принадлежность объекта соответствующему классу.

5.2. Иерархическая классификация

Данный подход заключатся в том, что сначала классификация производится по одному признаку, а затем для получившихся классов независимо друг от друга выполняется классификация по второму признаку. Каждому объекту присваивается несколько классов, в соответствии с количеством извлекаемых признаков.

При этом в случае двух измерений возможны два варианта алгоритма, в зависимости от того, по какому признаку классификация производится в первую очередь. Рисунок 2 иллюстрирует данный подход.

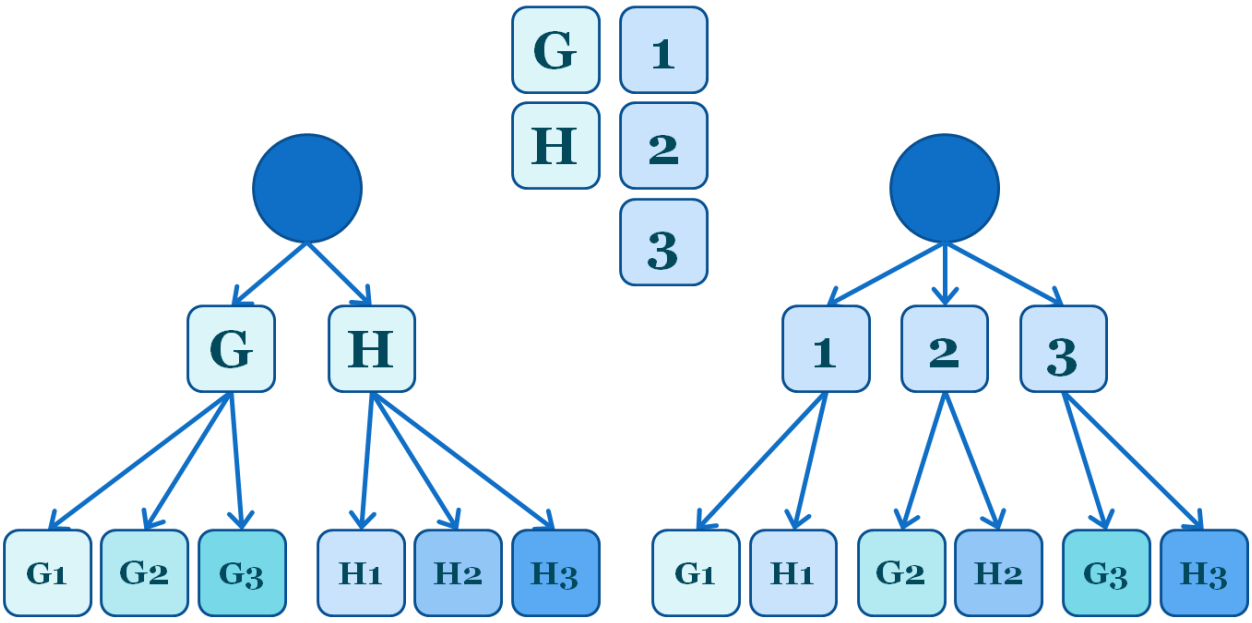


Рисунок 2. Иерархическая классификация.

Особенность этого метода заключается в том, что на первом уровне на каждый класс приходится больший объем данных, чем на втором, что может сказаться на точности алгоритма.

Разные модификации иерархической классификации применяются для автоматической индексации и фильтрации документов, для организации

документов, например, веб-страниц, в заранее определенную структуру, для разрешения проблем лексической неоднозначности и др. [8, 12, 42, 45].

При автоматическом извлечении из текста информации об его авторе иерархическая классификация служит для улучшения определения признака, по которому классификация проводится на нижнем уровне. Если, например, известно, что возрастной признак оказывает на речь большее влияние, чем гендерный, то для более точного определения пола автора, необходимо сначала провести классификацию текстов по возрасту, и далее для текстов каждой возрастной группы – классификацию по полу.

Иерархическая модель также характеризуются тем, что во время выполнения алгоритмов на её основе запускается несколько процедур классификации для каждого уровня модели. Поэтому для повышения точности можно применять, во-первых, различные характеристики, а во-вторых, разные алгоритмы классификации. Разные наборы характеристик могут потребоваться в случае, когда для классификации по одному признаку более информативными являются одни характеристики, а для классификации по другому признаку – другие. Комбинируя различные характеристики и алгоритмы на разных уровнях иерархической модели, можно получить более значительные результаты.

Если же для классификации используется один набор характеристик, то важно, чтобы характеристики хорошо отражали различия текстов, как для одного признака, так и для другого.

6. Экспериментальная среда

6.1. Корпус текстов

Стоит сказать, что в свободном доступе корпусов русских текстов, на которых возможно проведение необходимых экспериментов, нет, поэтому отдельной задачей стало составление корпуса и разметка текстов.

В данной работе в качестве текстов для проведения исследований используются блоги или интернет-дневники. Данное решение основано на следующих фактах: во-первых, речь в блогах достаточно неформальна и близка к разговорной, во-вторых, люди, ведущие дневники, обладают возможностями выражать свои эмоции как непосредственно в словах, так и через специальные сочетания символов, смайлики. Поэтому для письменных текстов такого формата становятся значимыми различия между мужской и женской речью, указанные в разделе «Гендерная лингвистика». Эти различия можно использовать для разделения текстов на классы.

Также для блогов характерны записи относительно небольшого объема, что делает их похожими на записи с отзывами о качестве продукта или устройства, размещенные на специализированных сайтах, например, Яндекс Маркет. Таким образом, в сочетании с алгоритмами из области сентимент-анализа, возможно использование рассматриваемых моделей для анализа мнений пользователей.

Тексты извлекались с русскоязычного сайта для размещения интернет-дневников www.livejournal.com с помощью программы «Alpha Parser» [1]. Значениями возрастного и гендерного признака автора считались те, что были указаны на странице с профилем автора. В связи с тем, что случаи подмены личной информации все-таки существуют, каждый блог изучался на предмет соответствия указанной в профиле информации и реальных данных, но полностью ошибки не могут быть исключены. Из всех извлеченных текстов были удалены цитаты, заключенные в кавычки. Использование авторами цитат без кавычек – еще один источник возможных неточностей результатов экспериментов.

Блоги выбирались в основном из тех, которые располагаются в рейтинге популярности на первых пятистах местах. Благодаря этому для каждого автора было собрано достаточно большое количество записей: в

среднем на каждого автора приходится 450 записей. Минимальное собранное количество записей одного автора - 12, максимальное - 1008 записей. Записи одного автора объединялись в один текстовый файл. Подобные текстовые файлы и составляют корпус текстов.

Для выполнения классификации авторов блогов было решено разделить на 4 возрастные группы: не старше 18, от 20 до 27, от 30 до 37, от 40 и старше. Выбрано именно такое разделение групп, исходя из соображений относительно очевидных различий между людьми разных возрастов, при этом учитывались некоторые переломные моменты в жизни большинства людей, например, поступление в высшее учебное заведение в возрасте 18 лет и т.п. Промежуточные группы были исключены из рассмотрения для обеспечения большего различия между представителями соседних групп, а также по причине того, что к моменту извлечения многие блоги велись авторами уже несколько лет.

В таблице 4 приводится распределение количества блогов по возрастным и гендерным группам. Для каждой группы приводится информация о количестве блогов и процентном соотношении:

	До 18	20 - 27	30 – 37	От 40	Всего
Мужчины	14 / 4,0%	48 / 13,7%	64 / 18,2%	69 / 19,6%	195
Женщины	55 / 15,7%	50 / 14,2%	30 / 8,5%	21 / 6,0%	156
Всего	69	98	94	90	351

Таблица 4. Распределение количества блогов по группам.

В таблице 5 указано количество записей, опубликованных авторами определенной группы, и процентное соотношение:

	До 18	20 - 27	30 – 37	От 40	Всего
Мужчины	2174 / 1,4%	21248 / 13,7%	38641 / 24,3%	38040 / 23,9%	100103
Женщины	6723 / 4,2%	22737 / 14,3%	15321 / 9,7%	14075 / 8,8%	58856
Всего	8897	43985	53962	52115	158959

Таблица 5. Распределение количества записей по группам.

В таблице 6 для каждой группы приводится количество предложений совместно для авторов определенной группы и процентное соотношение:

	До 18	20 - 27	30 – 37	От 40	Всего
Мужчины	28173 / 1,7%	191827 / 11,8%	376878 / 23,2%	419892 / 25,8%	1016770
Женщины	76503 / 4,7%	210741/ 12,9%	145442/ 8,9%	178463 / 11,0%	611149
Всего	104676	402568	522320	598355	1627919

Таблица 6. Распределение количества предложений по группам.

6.2. Представление текстов

Выбор характеристик для классификации текстов основывался, прежде всего, на указанных в главе «Гендерная лингвистика» различиях в речи мужчин и женщин. Эти характеристики имеют то преимущество, что они не зависят от контекста и имеют лингвистическую интерпретацию, так как, например, на основе данных о частоте различных буквенных сочетаний нельзя каким-либо образом охарактеризовать речь людей тех или иных групп. Все используемые различия были переведены в форму, приемлемую

для программной обработки. Получившиеся характеристики можно разделить на следующие группы:

1) *Данные о частоте использования знаков пунктуации и специальных символов.*

Исследовались 33 символа, которые используются в основной русскоязычной раскладке клавиатуры компьютера:

`	~	!	@	#
$	%	^	&	*
(	)	-	_	+
=	[	{	]	}
\	|	;	:	‘
“	,	<	.	>
/	?	№		

Таблица 7. Пунктуационные знаки и специальные символы, использованные в классификации.

Для каждого знака высчитывались следующие величины:

- Количество появлений знака в тексте, деленное на общее количество предложений;

- Количество предложений, в которых встретился знак, деленное на общее количество предложений.

Также вычислялись следующие четыре характеристики:

- Количество появлений какого-либо знака, деленное на общее количество предложений;

- Количество предложений, в которых встретился хотя бы один знак, деленное на общее количество предложений;

- Среднее количество различных знаков в предложении;

- Широта использования автором знаков пунктуации, равная максимуму от количества разных знаков в предложениях, деленного на количество всевозможных знаков.

2) *Данные о частоте использования различных частей речи и их сочетаний.*

Исследовались основные части речи русского языка и их формы: Имя существительное, Глагол, Местоимение личное, Местоимение (все другие виды), Имя прилагательное, краткая форма Имени прилагательного, Наречие, Предикатив («жаль», «хорошо», «пора»), Вводное слово, Служебная часть речи (Предлог, Союз, Частица), а также два сочетания: «Наречие + Имя прилагательное» и «Наречие + Наречие».

Для каждой из перечисленных групп высчитывались следующие величины:

- Количество появления части речи или сочетания в тексте, деленное на общее количество предложений;

- Количество предложений, в которых есть определенная часть речи или сочетание, деленное на общее количество предложений.

Для определения части речи использовался морфологический анализатор, разработанный группой АОТ [55]. По сведениям, представленным на сайте группы, данный анализатор базируется на грамматическом словаре А.А.Зализняка и включает 161 тысячу слов.

При обработке слова, не найденные в словаре, учитывались в отдельной характеристике, которая отнесена к шестой группе «Данные о словарном запасе».

Также в отдельную характеристику было выделено «Количество появлений в тексте частицы ‘не’, деленное на количество предложений».

3) Данные о частоте использования речевых оборотов и фразеологизмов.

Речевые обороты и фразеологизмы (устойчивые словосочетания) способствуют большему разнообразию речи и её оживлению. Использование речевых оборотов может говорить о возрасте, образованности, настроении говорящего или пишущего человека.

С сайта Национального Корпуса Русского языка [74] было взято пять списков с наиболее употребляемыми лексическими оборотами, выполняющими функции предлога («без оглядки на», «на волне»), наречия или предикатива («абы где», «без задних ног»), союза или союзного слова («постольку поскольку», «что касается»), частицы («всего-навсего», «нет-нет да и»), а также с наиболее употребляемыми вводными оборотами («как ни крути», «по правде сказать»).

В таблице 8 представлена информация о количестве оборотов в каждой группе:

Тип оборота	Количество оборотов
Наречия и предикативы	2307
Вводные слова	256
Союзы и союзные слова	165
Предлоги	310
Частицы	35

Таблица 8. Речевые обороты, использованные в экспериментах.

Для каждой группы речевых оборотов высчитывались следующие характеристики:

- Количество оборотов, встретившихся в тексте, деленное на общее количество предложений;

- Количество предложений, в которых встретился хотя бы один оборот из списка, деленное на общее количество предложений.

Также были использована характеристика «Количество всех оборотов, встретившихся в тексте, деленное на количество предложений».

Из списка фразеологизмов в Викисловаре [59] был отобран 1121 фразеологизм. Для каждого текста корпуса вычислялись характеристики, аналогичные предыдущим:

- Количество фразеологизмов, встретившихся в тексте, деленное на общее количество предложений;

- Количество предложений, в которых встретился хотя бы один фразеологизм из списка, деленное на общее количество предложений.

4) Данные о частоте использования смайликов

По результатам гендерных исследований, смайлики являются объектами, используемыми преимущественно женской аудиторией, и поэтому частота и объем их употребления в письменной речи, могут служить хорошими характеристиками для классификации. Для исследований было отобрано 13 чаще всего встречающихся типов-эмоций. Каждый тип содержал от 1 до 9 смайликов, при этом учитывались возможные написания смайликов английскими и русскими буквами. У смайликов из первых шести типов (за исключением «^_^» и «^_-»), последний символ может повторяться сколь угодно раз, например, «:)))))», эта особенность также была учтена, и подобные смайлики рассматривались как один, а не как несколько. В общую статистику добавлялась информация и о смайликах «)» и «(», если количество появлений каждой скобки в рамках одного предложения оказывалось не равным.

:)	:-)	^_^	))	)				
:(	:-(	((	(					
;)	;-)	^_-						
8-)	%)	%-)	8)					
:')	:'-)	:,)	:,-)					
:,(	:,-(	:'(	:'-(					
:*	:-*	=*						
o_O	o_O	=O	=O	o_o	o_o	O_O	O_O	0_0
:-b	:-P	":-p	:-P	:-p				
:-D	:D	;-D	;D					
:-[								
:-/	:-\							
>_<								

Таблица 9. Смайлики, использованные в экспериментах.

Для каждой группы смайликов высчитывались следующие величины:

- Количество смайликов определенного типа, встретившихся в тексте, деленное на общее количество предложений;

- Количество предложений, в которых встретился хотя бы один смайлик определенного типа, деленное на общее количество предложений.

Следующие характеристики выражают общую частоту и объем использования смайлов:

- Количество предложений, в которых встретился хотя бы один смайлик, деленное на общее количество предложений;

- Количество разных типов смайлов, использованных в тексте, деленное на всевозможное количество типов.

5) Данные о длине предложений и слов

- Средняя длина предложений в тексте, выраженная в словах;

- Средняя длина слов в тексте, выраженная в символах.

6) Данные о словарном запасе

- Богатство словарного запаса (Количество разных слов, использованных в тексте, деленное на общее количество предложений);

- Количество слов с частотой появления в тексте 1 и 2, деленное на общее количество предложений;

- Количество слов, не найденных в словаре, деленное на общее количество предложений.

Последняя характеристика выражает количество сленговых выражений и слов, написанных с ошибками.

Как видно, большая часть использованных для классификации характеристик не зависит от языка, на котором написан текст.

6.3. Алгоритмы классификации

Непосредственно классификация текстов проводилась с помощью программы «Weka» [47], в которой реализовано большое количество алгоритмов машинного обучения, включающих алгоритмы классификации, кластеризации, выбора значимых характеристик и др.

Пробные эксперименты показали, что метод опорных векторов (Support Vector Machines) в сочетании с алгоритмом «Sequential Minimization Optimization» для его обучения и «Байесовские сети» показывают лучшие результаты по сравнению с другими опробованными алгоритмами, поэтому именно эти два метода было решено использовать для основных экспериментов.

Метод опорных векторов для линейного случая заключается в построении в пространстве объектов оптимальной разделяющей гиперплоскости. Под оптимальностью понимается то, что элементы классов должны быть удалены от гиперплоскости настолько далеко, насколько это возможно. Обучение SVM сводится к задаче квадратичного программирования, имеющей единственное решение. При этом положение оптимальной гиперплоскости зависит лишь от небольшой доли обучающих объектов, которые называются опорными векторами. Метод обобщается на случай нелинейных разделяющих поверхностей с помощью введения функции ядра. Для быстрого и эффективного решения задачи квадратичного программирования в «Weka» реализован SMO – алгоритм, предложенный в 1998 году. Подробную информацию об этом алгоритме можно найти в [35].

В экспериментах метод опорных векторов использовался со стандартными настройками и полиномиальной функцией ядра. Задача классификации текстов по четырем возрастным группам автоматически преобразовывалась в несколько задач бинарной классификации.

Байесовская сеть – это вероятностная модель, представляющая собой множество переменных и их вероятностных зависимостей. Байесовская сеть представляется в виде направленного ациклического графа, вершины которого обозначают переменные, а ребра кодируют условные зависимости между переменными. Задача классификации в терминах байесовских сетей состоит в классификации переменой Y , заданной набором переменных $X =$

$X_1 \ldots X_n$. Баейсовская сеть строится на основе некоторого тренировочного множества пар (X, Y) и далее используется для вычисления величины $\text{argmax}_Y P(Y|X)$. Обучение алгоритма происходит в две фазы: настройки сети и настройки вероятностных таблиц. Для обоих фаз в Weka предусмотрено несколько алгоритмов. В данной работе для построения сети использовался алгоритм «Hill Climbing», а для построения вероятностных таблиц – «SimpleEstimator». Детали этих алгоритмов можно найти в [46].

6.4. Обработка текстов

Как уже было сказано, для извлечения записей из блогов была использована программа «Alpha Parser» [1]. Эта программа автоматически извлекает из блога записи и сохраняет их в текстовом формате. При этом из текстов удаляются все гиперссылки и HTML-тэги.

Для последующей обработки текстов на языке программирования Java был написан модуль, основная функция которого – извлечение и подсчет характеристик, описанных в подразделе «Представление текстов».

Основными этапами работы программы являются следующие:

1) Создание текстового файла со всеми извлеченными записями для каждого автора. Одновременно с этим происходит удаление всех английских символов, а также буквы «ё» и «Ё» заменяются на «е» и «Е» соответственно.
2) Поиск в тексте смайликов и подсчет количества их появлений.
3) Разделение текста на предложения. Маркерами конца предложений являются смайлики, точка ‘.’, знак восклицания ‘!’, вопросительный знак ‘?’, знак перехода на новую строку ‘\n’, знак возврата каретки ‘\r’ и знак табуляции ‘\t’.

Далее для каждого предложения производится:

4) Подсчет количества знаков препинания.
5) Разделение предложения на слова.
6) Определения части речи и начальной формы для каждого слова.
7) Подсчет количества различных частей речи.
8) Подсчет количества употреблений речевых оборотов и фразеологизмов.

Далее на основе полученной информации для каждого текста вычисляются характеристики, используемые непосредственно для классификации. В зависимости от типа классификации (плоская, иерархическая, многомерная) результаты по всем текстам собираются в один или несколько файлов формата ".arff", которые передаются одному из используемых классификаторов, реализованных с помощью «Weka».

7. Эксперименты

Базой для сравнения в экспериментах стал алгоритм классификации текстов по одному признаку.

Классификация по гендерному признаку и классификация по возрастному признаку производились в соответствии с базовым алгоритмом и двумя предложенными подходами: плоская классификация и иерархическая классификация. Для извлечения пола использовался вариант иерархической классификации «сначала по возрасту, потом по полу», а для извлечения информации о возрасте автора – «сначала по полу, потом по возрасту». Для каждого случая проводилось 10 экспериментов: 5 для метода опорных векторов и 5 для алгоритма классификации на основе Байесовских сетей. Результатом считалось усредненное значение пяти экспериментов.

Для построения моделей в качестве тренировочного набора использовалось 75% текстов корпуса, и для проверки в качестве тестового набора – 25% текстов.

Помимо основных экспериментов, в ходе работы были проведены эксперименты по выявлению оптимального набора характеристик, дающих наилучшую точность. Для данных экспериментов использовался метод кросс-валидации с параметром 5 (5-fold cross validation). Он заключается в том, что сначала весь корпус текстов делится на пять равных частей. Затем на каждой итерации четыре части объединяются в тренировочный набор, а пятая часть становится тестовым набором, на котором проверяется точность классификации. Результатами классификации являются совмещенные результаты всех пяти итераций.

Качество классификации во всех экспериментах оценивалось с помощью параметра «Точность», равному отношению правильно классифицированных текстов к количеству всех текстов в тестовом наборе.

7.1. Результаты и их анализ

В ходе экспериментов для выбора и настройки алгоритмов, а также для анализа использованных характеристик, было проведено более 300 опытов. В итоге, получены следующие результаты:

Анализ характеристик

В таблице 10 указана точность классификации по гендерному и возрастному признакам автора, проведенная на всем корпусе текстов с помощью метода кросс-валидации.

	Пол	Возраст
SVM	71,2%	50,1%
Bayesian Network	68,1%	48.2%

Таблица 10. Результаты классификации текстов по одному признаку.

Как видно из результатов эксперимента, точность классификации текстов по полу автора оказалась значительно выше, чем точность классификации по возрасту. Это связано с тем, что характеристики, использованные в классификации, основывались на различиях, выявленных между мужской и женской речью, различия между людьми разных возрастов использовались гораздо в меньшей степени. Данные эксперименты показывают, что выбранные характеристики не обеспечивают достаточное разделение текстов на группы при классификации по возрасту, и для подобных экспериментов требуются более информативные признаки, отражающие различия в письменной речи людей разных возрастов.

SVM-алгоритм в обоих случаях показал более высокую точность, чем Байесовские сети.

Интересно рассмотреть результаты классификации более подробно.

В таблице 11 для классификации текстов по полу представлены данные о количестве текстов, попавших в те или иные классы, а также точность определения каждого класса, выраженная в процентах:

Реальный класс	Определен в класс «Мужчины»	Определен в класс «Женщины»	Точность
Мужчины	152	43	77,9%
Женщины	58	98	62,8%

Таблица 11. Матрица исходов классификации текстов по гендерному признаку по SVM-алгоритму.

Из таблицы видно, что точность определения «мужских» текстов на 15% выше, чем точность определения «женских» текстов.

В таблице 12 аналогичные результаты представлены для классификации текстов по возрасту в соответствии с SVM-алгоритмом:

Реальный класс	Определен в класс «До 18»	Определен в класс «20 - 27»	Определен в класс «30 - 37»	Определен в класс «От 40»	Точность
До 18	47	20	1	1	68,1%
20 - 27	14	49	22	13	50,0%
30 - 37	4	28	38	24	40,4%
Старше 40	3	9	36	42	46,7%

Таблица 12. Матрица исходов классификации текстов по возрастному признаку по SVM-алгоритму.

Наиболее точнее класс был определен для текстов авторов, относящихся к группе «До 18». Класс «20 - 27» определялся менее точно, при этом неправильно классифицированные тексты этой группы были отнесены к другим классам примерно в одинаковом соотношении. Тексты, относящиеся к группе «30 - 37» чаще определялись к классам «20 - 27» и «От 40». А основные ошибки при определении класса для текстов последней возрастной группы были связаны с незначительными их различиями по сравнению с соседним классом «30 - 37».

Анализ характеристик, на основе которого были отобраны наиболее значимые для классификации, проводился с помощью критерия «хи-квадрат» χ^2. С помощью этого критерия можно измерить зависимости между характеристиками и классами. В общем случае таблица сопряженности для характеристики F и двух классов P1 и P2 выглядит следующим образом:

Значения признака P	P1	P2
Наличие характеристики F	*A*	*B*
Отсутствие характеристики F	*C*	*D*

Таблица 13. Таблица сопряженности 2x2 для характеристики F.

A – количество текстов, относящихся к классу P1, в которых встретилась характеристика F. *B* – количество текстов, относящихся к классу P2, в которых встретилась характеристика F. *C* – количество текстов, относящихся к классу P1, в которых не встретилась характеристика F. *D* – количество текстов, относящихся к классу P2, в которых не встретилась характеристика F.

Для данной таблицы статистика «хи-квадрат» вычисляется по следующему выражению:

$$\chi^2(F,P) = \frac{N \times (A \times D - B \times C)^2}{(A+B) \times (C+D) \times (A+C) \times (B+D)}$$

где $N=A+B+C+D$.

На основе реализации этого критерия в «Weka» было выявлено 25 характеристик для SVM-алгоритма и 24 характеристик для Байесовских сетей, классификация по которым дает более высокую точность определения гендерного признака автора. В таблице 14 представлены данные результатов эксперимента.

	Точность	Точность (улучш.)
SVM	71,2%	76,1%
Bayesian Network	68,1%	70,7%

Таблица 14. Результаты классификации текстов по гендерному признаку с улучшенным набором характеристик.

Для метода опорных векторов отбор характерстик дал улучшение примерно 5%, для Байесовских сетей чуть меньше – примерно 2,5%.

Список характеристик, дающих улучшенную точность, приведен в приложениях 1 – 2.

В итоге, классификация текстов по гендерному признаку показывает достаточно хорошую точность, и по этому параметру превосходит половину аналогичных исследований, рассмотренных в разделе «Обзор литературы».

В результате анализа характеристик для классификации текстов по возрастному признаку были определены наборы из 27 и 30 характеристик для метода опорных векторов и Байесовских сетей соответственно. Точность классификации на основе этих характеристик указана в таблице 15. Список отобранных характеристик можно найти в приложениях 3 – 4.

	Точность	Точность (улучш.)
SVM	50,1%	50,7%
Bayesian Network	48,2%	49,3%

Таблица 15. Результаты классификации текстов по возрастному признаку с улучшенным набором характеристик.

Отбор информативных для классификации по возрасту признаков дал улучшение в полпроцента для SVM – алгоритма и в 1% для Байесовских сетей.

Результаты классификации по возрастному признаку также превосходят некоторые из предшествующих исследований.

С другой стороны, данная работа не может быть сравнима с аналогичными, сделанными ранее, так как корпус русскоязычных текстов рассматривался в ней впервые.

Классификация текстов

В таблице 16 представлена точность определения пола и возраста в соответствии с тремя алгоритмами: базовым и двумя предложенными в этой

работе. Для каждой строки выделен подход, с помощью которого удалось получить самую высокую точность. Подробные результаты экспериментов можно найти в приложениях 5 – 8.

Признак (Алгоритм)	Baseline	Плоская	Иерархическая
Пол (SVM)	73,18%	70,82%	57,66%
Пол (Байесовские сети)	72,94%	69,64%	61,38%
Возраст (SVM)	55,28%	52,46%	45,18%
Возраст (Байесовские сети)	48,94%	37,86%	37,64%

Таблица 16. Результаты классификации текстов.

Как видно из таблицы SVM-алгоритм показал более высокую точность, чем Байесовские сети во всех случаях.

Базовый алгоритм оказался точнее по сравнению с двумя подходами, предложенными в данной работе. Однако результаты, полученные для плоской классификации, сопоставимы с результатами базового алгоритма и в лучшем случае уступают ему всего лишь на 2,8% процента. Иерархическая классификация показала самую низкую точность.

Такие результаты можно связать со следующей причиной. Разделение текстов на возрастные группы перед классификацией по гендерному признаку не дает увеличения точности, так как различия в текстах исходного корпуса, написанных людьми одного пола, но разных возрастов, незначительные. Аналогично, не существует серьезных различий и в текстах, написанных людьми одного возраста, но разного пола. Разделение текстов на группы приводит к тому, что значения характеристик для текстов одной группы оказываются «разбросанными», различия между классами при этом «размываются». Поэтому процесс классификации затрудняется. Для плоской классификации, благодаря тому, что алгоритмы обучаются на всех текстах тренировочного набора одновременно, удается добиться более высокой точности, чем для иерархической.

В целом, можно сделать вывод, что в текстах использованного корпуса возрастной и гендерный признаки не коррелируют. Подобная особенность может быть следствием специфических свойств данного вида электронной коммуникации или стиля интернет-дневников в общем.

Возможно и то, что выбор блогов для составления корпуса оказался не достаточно объективным, так как авторы большинства блогов являются известными в этой среде личностями, чьи записи читают тысячи людей. Ведение блогов для широкой аудитории скорей всего оказывает существенное влияние на содержание и построение речи.

Предложенные в работе подходы могут дать более высокие результаты для извлечения других признаков автора, связь между которыми в тексте значительней.

Однако проверка всех этих гипотез требует дополнительных исследований, которые выходят за рамки данной работы.

Для улучшения точности классификации текстов по данным признакам в целом можно предложить несколько вариантов. Один из них – более тщательный выбор информативных характеристик. Характеристики должны четче отражать особенности текстов разных классов. Другой вариант, применимый только для иерархической модели – это использование разных алгоритмов и наборов характеристик для извлечения разных признаков. Возможность подобного упоминалась в подразделе «Иерархическая классификация».

8. Заключение

В данной работе получены следующие результаты.

Во-первых, для автоматического профилирования автора по нескольким признакам одновременно предложено два подхода – плоская классификация и иерархическая классификация. Реализованы основанные на этих подходах алгоритмы. Для оценки их работы создана экспериментальная среда и подготовлены данные – корпус русских текстов, извлеченных из блогов. В ходе анализа результатов экспериментов установлено, что предложенные подходы не дают увеличения точности определения гендерного и возрастного признаков для текстов на русском языке. Однако результаты плоской классификации, в целом, сопоставимы с результатами базового алгоритма.

Во-вторых, для классификации русских текстов по полу и возрасту исследован ряд характеристик, которые отражают глубинные особенности письменной речи людей разного пола и возраста. Доказана возможность использования данных характеристик в задаче профилирования автора анонимного текста. Также установлены характеристики, дающие более высокую точность классификации, чем исходный набор. Полученные результаты сопоставимы с результатами предшествующих исследований и превосходят более половины из них. В то же время результаты данного исследования не могут быть сравнимы с другими, так как в данной работе корпус русских текстов рассматривался впервые. Таким образом, в работе создана основа для последующих исследований в области автоматического профилирования авторов русскоязычных текстов.

9. Список использованной литературы

1. Alpha-parser, Electronic source, http://makeprosoft.ru/alpha-parser/ (14.05.11)
2. Argamon S., Dawhle S., Koppel M., Pennebaker J. Lexical Predictors of Personality Type. In Proceedings of Classification Society of North America, St. Louis MI, June 2005.
3. Argamon S., Koppel M., Avneri G. Style-based text categorization: What newspaper am I reading? In Proceedings of AAAI Workshop on Learning for Text Categorization, 1–4, 1998
4. Argamon S., Koppel M., Pennebaker J., Schler J. Automatically Profiling the Author of an Anonymous Text. Communications of the ACM , 52 (2): 119-123, 2009
5. Argamon S., Levitan S. Measuring the usefulness of function words for authorship attribution. In Proceedings of ACH/ALLC 2005, Association for Computing and the Humanities, Victoria, BC, 2005.
6. Argamon S., Whitelaw C., Chase P., Hota S.R., Garg N., Levitan S. Stylistic text classification using functional lexical features. Journal of the American Society for Information Science and Technology, 58(6): 802–822, 2007
7. Baayen R., Halteren H., Tweedie F. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing, 11(3): 121–131, 1996
8. Dumais, S., Chen, H. Hierarchical Classification of Web Content. In Proceeding of the 23rd ACM International Conference on Reach and Development in Information Retrieval, Athens, Greece, 256–263, 2000
9. Estival D., Gaustad T., Pham S., Radford W., Hutchinson B. Author Profiling for English Emails. In Proceeding of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007), 262-272, 2007

10. Estival D., Gaustad T., Pham S., Radford W., Hutchinson B. TAT: an author profiling tool with application to Arabic emails. In Proceedings of the 5th Australasian Language Technology Workshop (ALTA 2007), Melbourne, 21–30., 2007
11. Gamon, M. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In Proceedings of the 20th International Conference on Computational Linguistics, Morristown, NJ: Association for Computational Linguistics, 611–617, 2004
12. Granitzer M. Hierarchical Text Classification using Methods from Machine Learning. Master's thesis, Graz University of Technology, 2003.
13. Grieve J. Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing, 22(3): 251–270, 2007
14. Hirst G., Feiguina O. Bigrams of syntactic labels for authorship discrimination of short texts. Literary and Linguistic Computing, 22(4): 405–417, 2007
15. Holmes D. The Evolution of Stylometry in Humanities Scholarship. Literary and Linguistic Computing 13(3):111-117, 1998
16. Honore A. Some simple measures of richness of vocabulary. Association for Literary and Linguistic Computing Bulletin, 7(2): 172–177, 1979
17. Juola P. Authorship attribution for electronic documents. In M. Olivier & S. Shenoi (Eds.), Advances in digital forensics II, Boston: Springer, 119–130, 2006
18. Keselj V., Peng F., Cercone N., Thomas C. N-gram-based author profiles for authorship attribution. In Proceedings of the Pacific Association for Computational Linguistics, 255–264, 2003
19. Kobayashi D., Matsumura N., Ishizuka M. Automatic Estimation of Bloggers' Gender. In Procceedings of the Int'l Conf. on Weblogs and Social Media (ICWSM 2007), Boulder, Colorado, USA, 279-280, 2007

20. Koppel M., Akiva N., Alshech E., Bar K. Automatically Classifying Documents by Ideological and Organizational Affiliation. In Proccedings of IEEE Intelligence and Security Informatics, Dallas TX, June 2009
21. Koppel M., Akiva N., Dagan I. A Corpus-Independent Feature Set for Style-Based Text Categorization. In Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico, 2003
22. Koppel M., Argamon S., Shimoni A, Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4): 401-412, November 2002
23. Koppel M., Schler J. Exploiting stylistic idiosyncrasies for authorship attribution. In Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, 69–72, 2003
24. Kucukyilmaz, T., Cambazoglu, B.B., Can, F., Aykanat, C. Chat mining: predicting user and message attributes in computer-mediated communication, Information Processing & Management, Elsevier, Volume 44, Issue 4, 1448-1466, July 2008
25. Lin J. Automatic Author Profiling of Online Chat Logs, Master's thesis, Naval Postgraduate School, Monterey, 2007.
26. Matsuura T., Kanada Y. Extraction of authors' characteristics from Japanese modern sentences via n-gram distribution. In Proceedings of the 3rd International Conference on Discovery Science, Berlin, Germany: Springer, 315–319, 2000
27. McCarthy P., Lewis G., Dufty D., McNamara D. Analyzing writing styles with coh-metrix. In Proceedings of the Florida Artificial Intelligence Research Society International Conference, 764–769, 2006
28. Mendenhall T.C. The characteristic curves of composition. Science, IX, 237–249, 1887
29. Merriam, T. What Shakespeare Wrote in Henry VIII (Part 1). The Bard, 2: 81-94, 1979

30. Merriam, T. What Shakespeare Wrote in Henry VIII (Part 2). The Bard, 2:111-8, 1980
31. Merriam, T. The Authorship of Sir Thomas More. Association for Literary and Linguistic Computing Bulletin, 10: 1-7, 1982
32. Morton, A. Q. Literary Detection. Scribners, New York, 1978
33. Mukherjee A., Liu B. Improving Gender Classication of Blog Authors. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010.
34. Mosteller F., Wallace D. L. Applied Bayesian and Classical Inference: The Case of the Federalist Papers. Addison-Wesley, Reading, MA, 1964
35. Platt J., Fast Training of Support Vector Machines using Sequential Minimal Optimization. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf, C. Burges, and A. Smola, eds., MIT Press, 185-208, 1999
36. Peng F., Shuurmans D., Keselj V., Wang S. Language independent authorship attribution using character level language models. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, Morristown, NJ: Association for Computational Linguistics, 267–274, 2003
37. Schler J., Koppel M., Argamon S., Pennebaker J. Effects of Age and Gender on Blogging. In Procceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, March 2006
38. Sebastiani F., Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1–47, 2002
39. Stamatatos E. Ensemble-based author identification using character n-grams. In Proceedings of the 3rd International Workshop on Text-Based Information Retrieval, 41–46, 2006
40. Stamatatos E., Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology, Wiley, 60(3): 538-556, 2009

41. Strous, R., Koppel, M., Fine, J., Nahaliel, S., Shaked, G. and Zivotofsky, A. Automated Characterization and Identification of Schizophrenia in Writing, J. of Nervous and Mental Disorders, 197(8): 585-588, 2008
42. Sun, A., Lim, E. Hierarchical text classification and evaluation. In Proceedings of the 2001 International Conference on Data Mining, IEEE Press, 521–528, 2001
43. Vel O., Anderson A., Corney M., Mohay G. Mining e-mail content for author identification forensics. SIGMODRecord, 30(4): 55–64, 2001
44. 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
45. Wang Y., Gong Z. Hierarchical Classification of Web Pages Using Support Vector Machine. In Proceedings of the ICADL 08 Proceedings of the 11th International Conference on Asian Digital Libraries: Universal and Ubiquitous Access to Information, 2008
46. WEKA Manual for Version 3-6-4, Electronic source, http://prdownloads.sourceforge.net/weka/WekaManual-3-6-4.pdf?download (23.05.11)
47. Witten I., Frank E. Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco, 2000
48. Yan, X., Yan, L. Gender Classification of Weblog Authors. Computational Approaches to Analyzing Weblogs, AAAI, 2006
49. Yule G.U. On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. Biometrika, 30: 363–390, 1938
50. Yule G.U. The statistical study of literary vocabulary. Cambridge University Press, 1944
51. Zhang C., Zhang P. Predicting gender from blog posts, 2010, Electronic source, http://www.cs.umass.edu/~pyzhang/course/genderClassify.pdf (14.05.11)

52. Zhao Y., & Zobel J. Effective and scalable authorship attribution using function words. In Proceedings of the 2nd Asia Information Retrieval Symposium, Berlin, Germany: Springer, 2005
53. Zhao Y., Zobel J. Searching with style: Authorship attribution in classic literature. In Proceedings of the 30th Australasian Computer Science Conference, NewYork: ACM Press, 59–68, 2007
54. Zheng R., Li J., Chen H., Huang Z. A framework for authorship identification of online messages: Writing style features and classification techniques. Journal of the American Society of Information Science and Technology, 57(3): 378–393, 2006
55. Автоматическая Обработка Текста. Электронный ресурс, http://www.aot.ru/index.html (14.05.11)
56. Беляев Н.Н., Хренова Н.Ф. Сохранение в коммуникации гендерных различий. Культура общения и её формирование, Вып. 9, Воронеж, с. 82, 2002.
57. Беляева А.Ю. Особенности речевого поведения мужчин и женщин. Дисс. канд. филол. наук, Саратов, 2002
58. Википедия – свободная энциклопедия. Гендер. Электронный ресурс, http://ru.wikipedia.org/wiki/Гендер (14.05.11)
59. Викисловарь. Список фразеологизмов русского языка. Электронный ресурс, http://ru.wiktionary.org/wiki/Приложение:Список_фразеологизмов_русского_языка (14.05.11)
60. Галяшина Е.И. Лингвистическая безопасность речевой коммуникации. Электронный ресурс, http://www.rusexpert.ru/magazine/034.htm (14.05.11)
61. Гетте Е.Ю. Речевое поведение в гендерном аспекте (Проблемы теории и методики описания). Дисс. канд. филол. Наук, Воронеж, 2004
62. Гомон Т. В. Исследование документов с деформированной внутренней структурой. Дисс. канд. юридич. наук, М., 1990

63. Горошко Е. И. Гендерные особенности русскоязычного Интернета. Электронный ресурс, http://www.textology.ru/article.aspx?aId=22 (14.05.11)
64. Горошко Е.И., Интернет-коммуникации в гендерном измерении. Вестник пермского университета, Выпуск 3 «Язык – культура – цивилизация», Пермь, с. 219-229, 2006
65. Денисова. А. А. Словарь гендерных терминов. Под. ред. А. А. Денисовой, М., Информация 21 век, 2002
66. Жельвис В.И. Стратегия и тактика брани: гендерный аспект проблемы. Доклады Первой Международной конференции «Гендер: язык, культура, коммуникация», М., с. 180-187, 2001
67. Земская Е.А., Китайгородская М.В., Розанова Н.Н. Особенности мужской и женской речи. Русский язык в его функционировании. Коммуникативно-прагматический аспект, М., Наука, с. 90-135, 1993
68. Кирилина А.В. Гендер: лингвистические аспекты. М. – Изд-во «Институт социологии РАН» - 1999.
69. Крючкова Т.Б. Некоторые экспериментальные исследования особенностей использования русского языка мужчинами и женщинами. Проблемы психолингвистики, М., с. 186-199, 1975
70. Кукушкина О.В., Поликарпов А.А., Хмелёв Д.В.. Определение авторства текста с использованием буквенной и грамматической информации. Проблемы передачи информации, 37(2):96-108, 2001
71. Мартыненко Г.Я., Многомерный синтаксический анализ художественной прозы. Структурная и прикладная лингвистика, Л, Изд-во ЛГУ, Вып.2, с.58-72, 1983
72. Мартынюк А.П. Прагматические особенности текста в зависимости от пола автора, Вестник Харьковского университета, №322, с. 56-60, 1988
73. Морозов Н.А. Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного известного

автора. Стилеметрический этюд, Известия отд. русского языка и словесности Имп.акад.наук, Т.20, Кн.4, 1915

74. Национальный Корпус Русского языка. Электронный ресурс, http://ruscorpora.ru (14.05.11)

75. Ощепкова Е. С. Идентификация пола автора по письменному тексту (Лексико-грамматический аспект). Дисс. канд. филол. наук, М., 2003

76. Родионова Е.С. Методы атрибуции художественных текстов. Структурная и прикладная лингвистика, Вып. 7, Межвуз. сб., Под ред. А.С. Герда, СПб, Изд-во С.-Петерб. Ун-та, с. 118-127, 2008

77. Севбо И.П. Графическое представление синтаксических структур и стилистическая диагнстика, Киев, 1981

78. Сиротинина О.Б., Гольдин В.Е. Речевая коммуникация и ее изучение. Проблемы речевой коммуникации, Межвуз. сб. науч. тр., Саратов, с.3-5, 2000

79. Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов. Предисловие А.Т. Фоменко. Фоменко А.Т. Новая хронология Греции: Античность в средневековье, Т. 2., М, Изд-во МГУ, с.768-820, 1996

80. Хмелёв Д.В. Распознавание автора текста с использованием цепей А.А. Маркова. Вестн. МГУ, Сер. 9, Филология, N02, с.115-126, 2000

81. Хьетсо Г., Густавссон С., Бекман Б., Гил С. «Кто написал «Тихий Дон», М., “Книга”, 1989

82. Шкуратова И.П. Речь как средство самовыражения личности. Северо-Кавказский психологический вестник, 2009

83. Якубайтис Т.А., Скляревич А.Н. Вероятностная атрибуция типа по нескольким морфологическим признакам. Рига, 1982

10. Приложения

Приложение 1

Список характеристик для классификации текстов по гендерному признаку, дающих более высокую точность SVM- алгоритма.

№	Характеристика
1	Средняя длина слов
2	Кол-во употреблений личных местоимений, деленное на кол-во предложений
3	Кол-во предложений, в которых встретилось хотя бы одно личное местоимение, деленное на кол-во предложений
4	Кол-во предложений, в которых встретился знак ‘$’, деленное на кол-во предложений
5	Кол-во употреблений знака ‘$’, деленное на кол-во предложений
6	Кол-во предложений, в которых встретилось хотя бы одно существительное, деленное на кол-во предложений
7	Кол-во предложений, в которых встретился знак ‘/’, деленное на кол-во предложений
8	Кол-во употреблений знака ‘/’, деленное на кол-во предложений
9	Кол-во употреблений знака ‘@’, деленное на кол-во предложений
10	Кол-во предложений, в которых встретился знак ‘:’, деленное на кол-во предложений
11	Кол-во предложений, в которых встретился знак ‘@’, деленное на кол-во предложений
12	Кол-во употреблений знака ‘#’, деленное на кол-во предложений
13	Кол-во предложений, в которых встретился знак ‘#’, деленное на кол-во предложений
14	Кол-во употреблений знака ‘:’, деленное на кол-во предложений
15	Кол-во употреблений сочетания «Наречие + Наречие», деленное на кол-во предложений
16	Кол-во предложений, в которых встретился знак ‘%’, деленное на кол-во предложений
17	Кол-во употреблений вводных оборотов в функции предлога, деленное на кол-во предложений
18	Кол-во предложений, в которых встретилось сочетание «Наречие + Наречие», деленное на кол-во предложений
19	Кол-во употреблений частицы «не», деленное на кол-во предложений
20	Кол-во употреблений знака ‘%’, деленное на кол-во предложений
21	Кол-во употреблений наречий, деленное на кол-во предложений
22	Кол-во предложений, в которых встретилось прилагательное, деленное на кол-во предложений
23	Кол-во употреблений речевых оборотов в функции союза и союзного слова, деленное на кол-во предложений
24	Кол-во употреблений глаголов, деленное на кол-во предложений
25	Кол-во предложений, в которых встретился речевой оборот в функции предлога, деленное на кол-во предложений

Приложение 2

Список характеристик для классификации текстов по гендерному признаку, дающих более высокую точность алгоритма «Байесовские сети».

№	Характеристика
1	Средняя длина слов
2	Кол-во употреблений личных местоимений, деленное на кол-во предложений
3	Кол-во предложений, в которых встретилось хотя бы одно личное местоимение, деленное на кол-во предложений
4	Кол-во предложений, в которых встретился знак ‘$’, деленное на кол-во предложений
5	Кол-во употреблений знака ‘$’, деленное на кол-во предложений
6	Кол-во предложений, в которых встретилось хотя бы одно существительное, деленное на кол-во предложений
7	Кол-во предложений, в которых встретился знак ‘/’, деленное на кол-во предложений
8	Кол-во употреблений знака ‘/’, деленное на кол-во предложений
9	Кол-во употреблений знака ‘@’, деленное на кол-во предложений
10	Кол-во предложений, в которых встретился знак ‘:’, деленное на кол-во предложений
11	Кол-во предложений, в которых встретился знак ‘@’, деленное на кол-во предложений
12	Кол-во употреблений знака ‘#’, деленное на кол-во предложений
13	Кол-во предложений, в которых встретился знак ‘#’, деленное на кол-во предложений
14	Кол-во употреблений знака ‘:’, деленное на кол-во предложений
15	Кол-во употреблений сочетания «Наречие+Наречие», деленное на кол-во предложений
16	Кол-во предложений, в которых встретился знак ‘%’, деленное на кол-во предложений
17	Кол-во употреблений речевых оборотов в функции предлога, деленное на кол-во предложений
18	Кол-во предложений, в которых встретилось сочетание «Наречие+Наречие», деленное на кол-во предложений
19	Кол-во употреблений частицы «не», деленное на кол-во предложений
20	Кол-во употреблений знака ‘%’, деленное на кол-во предложений
21	Кол-во употреблений наречий, деленное на кол-во предложений
22	Кол-во предложений, в которых встретилось прилагательное, деленное на кол-во предложений
23	Кол-во употреблений речевых оборотов в функции союза и союзного слова, деленное на кол-во предложений
24	Кол-во употреблений глаголов, деленное на кол-во предложений

Приложение 3

Список характеристик для классификации текстов по возрастному признаку, дающих более высокую точность точность SVM- алгоритма.

№	Характеристика
1	Средняя длина слов
2	Кол-во употреблений личных местоимений, деленное на кол-во предложений
3	Кол-во предложений, в которых встретилось хотя бы одно личное местоимение, деленное на кол-во предложений
4	Кол-во употреблений знака ‘#’, деленное на кол-во предложений
5	Кол-во предложений, в которых встретился знак ‘#’, деленное на кол-во предложений
6	Кол-во предложений, в которых встретился знак ‘%’, деленное на кол-во предложений
7	Кол-во употреблений знака ‘%’, деленное на кол-во предложений
8	Кол-во употреблений смайлика из категории {";)", ";-)", "^_-"}, деленное на кол-во предложений
9	Кол-во употреблений знака ‘@’, деленное на кол-во предложений
10	Кол-во предложений, в которых встретился знак ‘@’, деленное на кол-во предложений
11	Кол-во употреблений знака ‘$’, деленное на кол-во предложений
12	Кол-во предложений, в которых встретился знак ‘$’, деленное на кол-во предложений
13	Кол-во предложений, в которых встретился знак ‘^’, деленное на кол-во предложений
14	Кол-во предложений, в которых встретился знак ‘/’, деленное на кол-во предложений
15	Кол-во употреблений смайлика из категории {":-D", ":D", ";-D", ";D"}, деленное на кол-во предложений
16	Кол-во употреблений знака ‘_’, деленное на кол-во предложений
17	Кол-во предложений, в которых встретился знак ‘_’, деленное на кол-во предложений
18	Кол-во употреблений знака ‘=’, деленное на кол-во предложений
19	Кол-во употреблений знака ‘^’, деленное на кол-во предложений
20	Кол-во употреблений сочетания «Наречие + Наречие», деленное на кол-во предложений
21	Кол-во употреблений знака ‘№’, деленное на кол-во предложений
22	Кол-во предложений, в которых встретился знак ‘№’, деленное на кол-во предложений
23	Кол-во предложений, в которых встретился знак ‘:’, деленное на кол-во предложений
24	Кол-во предложений, в которых встретилось существительное, деленное на кол-во предложений
25	Кол-во употреблений смайликов, деленное на кол-во предложений
26	Кол-во употреблений знака ‘””, деленное на кол-во предложений
27	Кол-во предложений, в которых встретился знак ‘*’, деленное на кол-во предложений

Приложение 4

Список характеристик для классификации текстов по возрастному признаку, дающих более высокую точность алгоритма «Байесовские сети».

№	Характеристика
1	Средняя длина слов
2	Кол-во употреблений личных местоимений, деленное на кол-во предложений
3	Кол-во предложений, в которых встретилось хотя бы одно личное местоимение, деленное на кол-во предложений
4	Кол-во употреблений знака ‘#’, деленное на кол-во предложений
5	Кол-во предложений, в которых встретился знак ‘#’, деленное на кол-во предложений
6	Кол-во предложений, в которых встретился знак ‘%’, деленное на кол-во предложений
7	Кол-во употреблений знака ‘%’, деленное на кол-во предложений
8	Кол-во употреблений смайлика из категории {";)", ";-)", "^_-"}, деленное на кол-во предложений
9	Кол-во употреблений знака ‘@’, деленное на кол-во предложений
10	Кол-во предложений, в которых встретился знак ‘@’, деленное на кол-во предложений
11	Кол-во употреблений знака ‘$’, деленное на кол-во предложений
12	Кол-во предложений, в которых встретился знак ‘$’, деленное на кол-во предложений
13	Кол-во предложений, в которых встретился знак ‘^’, деленное на кол-во предложений
14	Кол-во предложений, в которых встретился знак ‘/’, деленное на кол-во предложений
15	Кол-во употреблений смайлика из категории {":-D", ":D", ";-D", ";D"}, деленное на кол-во предложений
16	Кол-во употреблений знака ‘_’, деленное на кол-во предложений
17	Кол-во предложений, в которых встретился знак ‘_’, деленное на кол-во предложений
18	Кол-во употреблений знака ‘=’, деленное на кол-во предложений
19	Кол-во употреблений знака ‘^’, деленное на кол-во предложений
20	Кол-во употреблений сочетания «Наречие + Наречие», деленное на кол-во предложений
21	Кол-во употреблений знака ‘№’, деленное на кол-во предложений
22	Кол-во предложений, в которых встретился знак ‘№’, деленное на кол-во предложений
23	Кол-во предложений, в которых встретился знак ‘:’, деленное на кол-во предложений
24	Кол-во предложений, в которых встретилось существительное, деленное на кол-во предложений
25	Кол-во употреблений смайликов, деленное на кол-во предложений
26	Кол-во употреблений знака ‘”’, деленное на кол-во предложений
27	Кол-во предложений, в которых встретился знак ‘*’, деленное на кол-во предложений
28	Кол-во предложений, в которых встретился знак ‘”’, деленное на кол-во предложений
29	Кол-во употреблений смайлика из категории {":)", ":-)", "^_^", "))", “)”}, деленное на кол-во предложений
30	Кол-во употреблений знака ‘:’, деленное на кол-во предложений

Приложение 5

Результаты классификации текстов по гендерному признаку в соответствии с SVM-алгоритмом. Числа в первом столбце обозначают номер эксперимента. В последней строке приведены усредненные результаты пяти экспериментов.

	Baseline	Плоская	Иерархическая
1	71,8%	70,6%	65,9%
2	71,8%	68,2%	51,8%
3	72,9%	67,1%	54,1%
4	72,9%	75,3%	54,1%
5	76,5%	72,9%	62,4%
	73,18%	70,82%	57,66%

Приложение 6

Результаты классификации текстов по гендерному признаку на основе «Байесовских сетей». Числа в первом столбце обозначают номер эксперимента. В последней строке приведены усредненные результаты пяти экспериментов.

	Обычная	Плоская	Иерархическая
1	71,8%	65,9%	63,5%
2	68,2%	65,9%	61,1%
3	75,3%	65,9%	69,4%
4	72,9%	72,9%	61,1%
5	76,5%	77,6%	51,8%
	72,94%	69,64%	61,38%

Приложение 7

Результаты классификации текстов по возрастному признаку в соответствии с SVM-алгоритмом. Числа в первом столбце обозначают номер эксперимента. В последней строке приведены усредненные результаты пяти экспериментов.

	Обычная	Плоская	Иерархическая
1	54,1%	48,2%	36,5%
2	51,8%	48,2%	45,9%
3	58,8%	60,0%	48,2%
4	52,9%	50,6%	47,1%
5	58,8%	55,3%	48,2%
	55,28%	52,46%	45,18%

Приложение 8

Результаты классификации текстов по возрастному признаку на основе «Байесовских сетей». Числа в первом столбце обозначают номер эксперимента. В последней строке приведены усредненные результаты пяти экспериментов.

	Обычная	Плоская	Иерархическая
1	43,5%	37,6%	41,2%
2	50,6%	38,8%	35,3%
3	57,6%	38,8%	34,1%
4	50,6%	35,3%	35,3%
5	42,4%	38,8%	42,3%
	48,94%	37,86%	37,64%