Министерство образования и науки Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)
Филологический факультет
Кафедра общего, славяно-русского языкознания и классической филологии

ДОПУСТИТЬ К ПРЕДСТАВЛЕНИЮ В ГЭК
Руководитель ООП
д-р филол. наук, профессор
_____Т.А. Демешкина
«01» июня 2018 г.

НАУЧНЫЙ ДОКЛАД

об основных результатах подготовленной научно – квалификационной работы (диссертации)

ГЕНДЕРНАЯ АТРИБУЦИЯ ТЕКСТОВ МЕТОДАМИ КОПЬЮТЕРНОЙ ЛИНГВИСТИКИ

к государственному экзамену по основной образовательной программе подготовки научно-педагогических кадров в аспирантуре
направление подготовки 45.06.01 – Языкознание и литературоведение

Степаненко Андрей

Научный руководитель
доктор филол наук, профессор
_____З.И. Резанова

Автор работы
аспирант
_____А. Степаненко

Томск-2018

В современной компьютерной лингвистике особое место занимает гендерная атрибуция текстов компьютерной коммуникации. Проблемы, связанные с определением пола участников коммуникации, решают два научных направления: лингвистика (Горошко Е.И. Кирилина А.В., 1999; Кирилина, 1998; Jespersen O., 1998; Crystal D., 2001; Spender D., 1980; Кирилина А.В., Томская М.В., 2005) и информатика (Argamon Sh., Koppel M., Fine J., Shimoni A., 2003; Arroju, 2015; Гомон Т.В., 1990). Однако до последнего времени методы данных наук не применялись в одном исследовании одного объекта. В настоящее время стали появляться работы, интегрирующие лингвистические и статистические методы (Argamon Sh., Koppel M., Fine J., Shimoni A., 2003; Sboev A., 2016), что привело к улучшению классификации текстов по исследуемым параметрам. Тем не менее, большинство результатов исследований автоматической классификации текстов по гендерному принципу неустойчивы при анализе *новых текстов*, что свидетельствует о необходимости как дальнейших поисков маркеров речи, характерных для представителей того или иного пола, так и совершенствования методов их исследования.

Диссертационное исследование посвящено поиску гендерных маркеров в текстах компьютерной коммуникации с применением лингвистических и статистических методов, а также выявлению их влияния на качество автоматической классификации текстов. Недостаточная разработанность проблемы интеграции методов математики, информатики и лингвистики в отечественных исследованиях обусловливает **актуальность** диссертационной работы, направленной на разработку способов автоматической классификации текстов по признаку гендерной принадлежности их авторов, предполагающей выявление языковых маркеров гендерных оппозиций и применение алгоритмов машинного обучения.

Несмотря на значительное количество работ, посвященных изучению различий в коммуникации мужчин и женщин, считаем необходимым отметить, что характерной чертой большинства предыдущих исследований является отсутствие или недостаточное использование статистических методов исследования, что сказывается на точности результатов противопоставления мужской и женской письменной речи.

Не менее важным остается вопрос выбора типов языковых маркеров в процессах авторской атрибуции текста. лингвистические аспекты в классификации текстов. Большинство работ, основывающихся на машинном обучении, не учитывает лингвистических маркеров в алгоритмах классификации и основывается на общепринятых методах стилеметрии (Мартыненко, 1988), (Аверьянов, 2007); (Марусенко, 1990), (Гилилов, 2007). Все эти работы ограничены использованием формальных или асемантических маркеров членения текста (n-граммы, количество слов, предложений, служебных слов и т.д.). В них не затрагивается вопрос влияния лингвистических маркеров на качество классификации текстов. Это обусловлено тем, во-первых, что такие маркеры легче найти и провести квантативный анализ, во-вторых, подобные признаки сложнее имитировать обрабатывать автору текста. Однако мы считаем, что лингвистические маркеры могут соответствовать и этим двум требованиям (З.И. Резанова, А.С. Романов, Р.В. Мещеряков, 2013).

В работе в качестве маркеров гендерных различий исследуются две группы лексических единиц: экспрессивные лексические единицы и местоимения.

Выделение первой группы лексем в качестве маркеров гендерных различий основывается на выводах отечественных исследователей в области гендерной лингвистики о преобладании использования эмоционально-оценочных и уменьшительно-ласкательных лексических единиц в женской речи (Горошко Е.И. Кирилина А.В., 1999; Земская, Китайгородская)

При выделении второй группы лексических маркеров – личных местоимений – мы, во-первых, опираемся на так называемую психолингвистическую «Я-концепцию» в контексте общего развития личности. (Рубинштейн С.Л., 1989). Во-вторых, выделение

личных местоимений в качестве гендерных текстовых маркеров также опирается на данные, полученные в гендерных исследованиях (Verhoeven B.C., 2015).

Исходя из всего перечисленного выше, считаем, что две группы лексических маркеров могут быть взяты за основу для дальнейшей автоматической классификации текстов и должны повлиять на эффективность работы алгоритмов машинного обучения, в чем и заключается следующий аспект актуальности.

Интеграция методов лингвистики и математической статистики позволит улучшить качество классификации текстов по гендерному принципу, а выделенные нами маркеры сложно имитировать, так как они являются неосознанным выбором автора текста.

Данная работа находится на стыке таких научных направлений, как лингвистика, статистика и программирование. Необходимость сочетания данных методов в лингвистических исследованиях обусловлено увеличением объема информации и необходимостью быстрого и точного анализа текстов естественного языка. Сегодня существует ряд работ, авторы которых разделяют данную точку зрения (Shlomo Argamon, Moshe Koppel, Jonathan Fine, Anat Rachel Shimoni, 2010;). В отечественной лингвистике следует выделить работу А. Сбоева, который, используя материал корпуса текстов, осуществил методы машинного обучения, основываясь на маркерах положительной/отрицательной оценки (Sboev A., 2016). Точность работы классификатора составляет 0,64 (F-score), однако сам автор считает, что расширение лингвистических маркеров для их дальнейшего применения в машинном обучении позволит увеличить точность классификации до 0,93.

Объектом исследования является лексический уровень текстов компьютерно опосредствованной коммуникации.

Предметом исследования данной работы являются лексические маркеры гендерных различий текстов компьютерно опосредствованной коммуникации в поставленной задаче автоматической обработки текстов.

Мы выдвигаем следующую **гипотезу**: существуют классы лексических единиц, которые могут использоваться как гендерные маркеры в автоматической классификации текстов.

Цель диссертационной работы – выявление диагностирующих возможностей лексических маркеров двух типов – личных местоимений и экспрессивных единиц – в методиках автоматической гендерной атрибуции текстов компьютерно опосредствованной коммуникации межличностной коммуникации.

Для достижения поставленной цели необходимо решить следующие **задачи**.

1. На основе анализа работ по гендерной лингвистике и гендерной математической атрибуции текстов сформировать гипотезу о признаках текста, способных быть маркерами гендерных различий.

2. Провести системный анализ лингвистических и математических методов, используемых при решении задач гендерной атрибуции текстов, сформировать гипотезу об их применении для решения поставленной в работе цели.

3. Сформировать репрезентативный корпус текстов, отвечающих заданным параметрам – межличностное компьютерно опосредствованное общение, охарактеризовать дискурсивно жанровое своеобразие текстов сформированного корпуса.

4. С применением лингвистических методов выделить классы лексических единиц, которые, согласно выдвинутой гипотезе, могут выступать в качестве маркеров гендерных различий в текстах компьютерно опосредствованной коммуникации.

5. Разработать программное обеспечение, осуществляющее полный цикл работы с текстом по модели: извлечение текстовой информации, токенизация, лемматизация, классификация.

6. На основе статистического анализа выявить наиболее устойчивые маркеры, обеспечивающие разбиение объектов на классы по гендерному принципу.

7. Оценить точность классификатора и сравнить его с существующими решениями.

Материалом исследования стали 24 диалога межличностной коммуникации студентов филологического факультета Томского государственного университета социальной сети «ВКонтакте». Диалоги были собраны в рамках учебной практики студентов направления Фундаментальной и прикладной лингвистики Томского государственного университета. Обработка и анализ диалогов личной коммуникации были согласованы с участниками коммуникации в рамках норм регламента Этического комитета междисциплинарных исследований ТГУ (http://lab.tsu.ru/cognitivestudies/node/14) и в соответствии с Федеральным законом №152 РФ «О персональных данных». Возраст участников коммуникации составил 18-20 лет.

Объем текстов – 85954 лексических единиц. Все полученные диалоги разбиты на реплики относительно гендерной принадлежности участника коммуникации. Средний объем одного текстового файла, содержащий мужские или женские реплики, составил 4000 слов.

Теоретико-методологическую основу диссертации составили основополагающие работы по гендерной лингвистике, дискурс-анализу и математическим методам авторской атрибуции текстов.

Работы по гендерной лингвистике послужили основой формирования гипотезы о гендерно специфичных элементах и характеристиках текстов (Горошко Е. И., 2004; Crystal D., 2001; Spender D., 1980; Herring, 1992; Koppel M., Schler J., 2003; Tannen D., 1996; Горошко Е.И. Кирилина А.В., 1999; Земская Е.А., Китайгородская М.А., Розанова Н.Н., 1993; Кирилина А.В., Томская М.В., 2005; Крючкова Т.Б., 1975), представления о возможной диагностирующей силе которых в текстах межличностной компьютерно опосредствованной коммуникации были уточнены с опорой на работы по теории дискурс-анализа, жанрологии, дискурсивно-жанровой специфике компьютерно опосредствованной межличностной коммуникации.

При разработке принципов квантитативного анализа при решении задач авторской атрибуции текстов, использования математические методов (корреляционный анализ, сравнительная и описательная статистика, линейный дискриминантный анализ, машинное обучение: логистическая регрессия, случайный лес, деревья решений) мы опирались на исследования следующих авторов (Argamon Sh., Koppel M., Fine J., Shimoni A., 2003; Arroju, 2015; Sboev A., 2016 ; Shlomo Argamon, Moshe Koppel, Jonathan Fine, Anat Rachel Shimoni, 2010; Yan X., Yan L., 2006; Поликарпов А.А., Поддубный В.В., Кукушкина О.В., Кубарев А.И., Варламов А.А., 2013; Фоменко В.П., Фоменко Т.Г.,, 1996; Марусенко М.А., 1990).

Для проведения статистических методов и нормализации текста были применены следующие **инструменты**: Язык программирования R 3.0.1; Лемматизатор компании «Яндекс» – «Mystem» 3.0; Программа «STATISTICA» 12

Новизна исследования заключается в том, что в ней впервые разработано программное обеспечение, осуществляющее полный цикл работы с текстом по модели: извлечение текстовой информации, токенизация, лемматизация, классификация для решения задач авторской атрибуции текстов; осуществлен статистический анализ местоимений, экспрессивных и оценочных лексических единиц, используемых в качестве маркеров при классификации текстов по гендерному принципу; применено машинное обучение и классификация текстов, где выявлены и охарактеризованы диагностирующие возможности лексических маркеров двух типов – личных местоимений и экспрессивных единиц – в методиках автоматической гендерной атрибуции текстов компьютерно опосредствованной коммуникации межличностной коммуникации.

Теоретическая значимость диссертационной работы определяется тем, что исследование вносит вклад в современную гендерную лингвистику, расширяя представление о типах маркеров гендерных различий и их диагностирующей силе при атрибуции текстов по гендерному принципу, а также способствуя расширению методологической базы гендерных исследований. Новые данные о взаимном влиянии

социальных факторов в процессах порождения текста углубляют представления об особенностях компьютерно опосредствованной коммуникации, что значимо для развития современной дискурсологии. Разработанная методика и соответствующее ей программное обеспечение демонстрируют эффективность интеграции лингвистических и математических методов при решении задач авторской атрибуции текста, способствуя развитию междисциплинарных методов современной лингвистики.

Практическая значимость. Результаты исследования способствуют развитию математических и прикладных методов изучения языка и вносят вклад в развитие практики атрибуции текстов.

Разработанная программа полного цикла классификации текстов по гендерному принципу может использоваться при решении данной задачи применительно к другим типам текстов, при решении смежных задач, а также в практике преподавания курсов компьютерной лингвистики, спецкурсов по гендерной лингвистике и дискурс-анализу.

Результаты исследования могут быть применены в криминалистической практике, направленной на борьбу с фактами мошенничества, целенаправленным скрытием пола участника коммуникации.

Основные положения, выносимые на защиту.

1. Существуют значимые экспрессивные лексические единицы, различающие мужские и женские тексты компьютерной коммуникации.

2. Разработанная методика определения значений параметров на материале текстов компьютерной коммуникации позволяет решать задачу определения гендера автора анонимного текста.

3. Выявленные статистически значимые экспрессивные маркеры могут точнее влиять на классификацию текстов по гендерному принципу в сравнении с формальными и асемантическими признаками.

Апробация работы. Основные положения диссертационной работы были представлены в докладах на трех международных научных конференциях: III (XVII) Международной научно-практической конференции молодых ученых «Актуальные проблемы лингвистики и литературоведения» (Томск, 2016 г.), IV (XVIII) Международной научно-практической конференции молодых ученых «Актуальные проблемы лингвистики и литературоведения» (Томск, 2017 г.), V (XIV) Международной научно-практической конференции молодых ученых «Актуальные проблемы лингвистики и литературоведения» (Томск, 2018 г.), XXVIII Международной научно-практической конференции молодых ученых «Язык и культура» (Томск, 2018 г.),. XXVII Международной научно-практической конференции молодых ученых «Язык и культура» (Томск, 2017 г.).

Структура работы состоит из введения, двух глав, заключения, списка использованных источников и литературы. Во введении обосновывается выбор темы исследования, ее цель, задачи, методы, материалы и источники, теоретическая и практическая значимость, структура работы и положения, выносимые на защиту.

В первой главе «Проблемы и методы гендерной атрибуции текста» рассматриваются проблемы становления гендерной лингвистики в отечественных и зарубежных исследованиях, изучаются языковые маркеры различных уровней, противопоставляющие речевые произведения по гендерному принципу. Обсуждаются результаты исследований, в которых выявлены и описаны признаки коммуникативного поведения мужчин и женщин, отражающиеся в своеобразии текстовых структур на разных языковых уровнях (Беляев Н.Н., Хренова Н.Ф.,, 2002; Горошко Е. И., 2004; Кирилина А. В., 2002; Гриценко, 2005; Васильева А.В., 2015; Кирилина А.В., Томская М.В., 2005; Вернер Ф., 1984). Отмечается, что в гендерной лингвистике ученые пришли к выводу о существовании эмоциональных различий в речевом поведении мужчин и женщин, выражающихся в своеобразии использования эмоциональных, оценочных лексических единиц и местоимений, дается их структуризация, классификация, роль и их место в процессе коммуникации. (Argamon Sh., Koppel M., Fine J., Shimoni A., 2003). Характеризуется специфика и типы этих

лингвистических маркеров, сравниваются математические и лингвистические методы поиска гендерных различий, выделяемых при гендерной атрибуции текстов. Вместе с тем отмечается, что гендерное речевое позиционирование не может осуществляться изолированно, вне других социальных ролей, репрезентируемых в тех или иных речевых практиках (Кузниченко, 2005).

В главе также обсуждаются теоретические и прикладные аспекты интеграции лингвистики и математики при решении проблем гендерной атрибуции текстов. Характеризуются методы стилеметрии, их роль в применении автоматической классификации текстов с учетом специфики лингвистических маркеров как признаков атрибуции класса текста по гендерному принципу. Изучаются результаты лингвистических (Spender D., 1980; Crystal D., 2001; Biber D., Burges J., 2000; Danet B., 1996; Flannery R., 1946; Crowston K., Kammerer E., 1998; Hall K., 1995; Herring, Susan C., 1996; Miller P., 2004; Sussman N. M., Tyson D. H., 2000) и математических (Argamon Sh., Koppel M., Fine J., Shimoni A., 2003; Arroju, 2015; Горошко Е. И., 2004; Виноградов В. В., 1961; Мартыненко, 1988; Марусенко М.А., 1990) исследований. Более пристальное внимание уделяется работам, находящимся на стыке лингвистики, статистики и машинного обучения (Arroju, 2015; Sboev A., 2016 ; Yan X., Yan L., 2006).

Во второй главе «Гендерная атрибуция текстов компьютерной коммуникации» характеризуются дискурсивные особенности компьютерно опосредствованной коммуникации, дискурсивное своеобразие социальных сетей, вслед за рядом авторов выделяется и описывается макрожанр «стена» социальной сети «ВКонтакте», определяется жанровое своеобразие частных жанров в структуре макрожанра. Особое внимание уделяется характеристике ролевых позиций коммуникаций, языковых особенностей текстов анализируемых жанров.

Далее в качестве лингвистических маркеров гендерных различий в текстах компьютерной коммуникации характеризуются два класса единиц – личные местоимения и экспрессивные лексемы. Представлены результаты статистического исследования местоимений, экспрессивных и оценочных лексических единиц и на их основе составлены алгоритмы машинного обучения.

Основные выводы второй главы:

На рис.1 схематично представлен алгоритм работы программы, показывающий поэтапную реализацию обработки текстовой информации и классификации текста относительно гендерной принадлежности участника коммуникации.

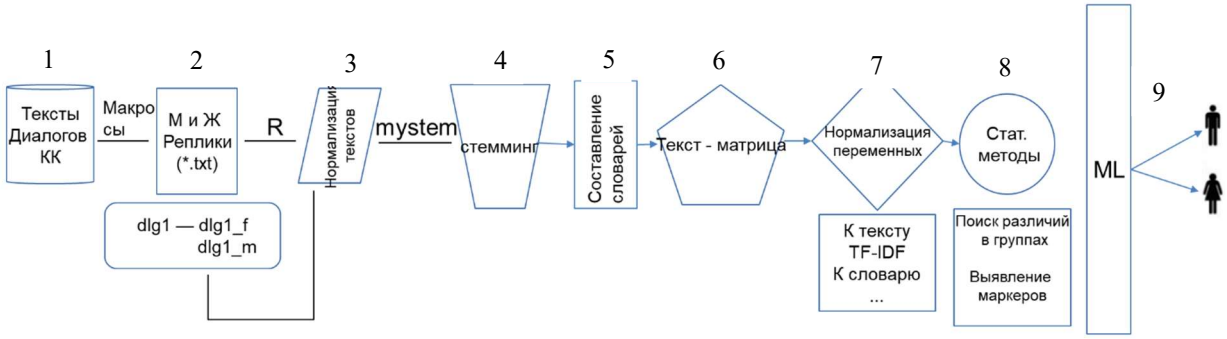


Рисунок 1. Этапы автоматической обработки текста

Указанный выше рисунок демонстрирует реализацию блоков алгоритма программы:

1. Извлечение текстовой информации из диалогов компьютерной коммуникации (мужчина-женщина).
2. Разбиение диалога на файлы, содержащие отдельно мужские и женские реплики.
3. Приведение в тексте всех слов к нижнему регистру, удаление знаков пунктуации, неинформативных элементов (ссылки, картинки, знаки пунктуации и т.д.).
4. Лемматизация текстовых массивов (поиск начальной формы слова).
5. Извлечение из текста составленных ранее лингвистических маркеров.

6-7. Частотное представление лексических единиц и нормализация абсолютных величин.
8. Выявление наиболее устойчивых лингвистических маркеров.
9. Применение алгоритмов машинного обучения.

Предварительно различия лингвистических маркеров были проверены на статистическую значимость. Исходя из типа распределения, был выбран ранговый критерий проверки значимости Вальда-Вольфовица для местоимений, экспрессивных и эмоциональных лексических единиц. Как показал результат исследования, статистические различия были подтверждены в группе местоимений «Я» (рис. 2) и его производных супплетивных форм ($p<0.05$ (уровень значимости), где $F=4,36$; 1; $p=0,0075$). По остальным группам лексических единиц статистически значимых различий найдено не было.

Рисунок 2. Различия в использовании местоимений относительно общего числа слов в тексте.

Что касается экспрессивных и оценочных лексических единиц, то отсутствие статистически значимых различий обусловлено высоким уровнем дисперсий (критерий Фридмана) внутри групп, что можно увидеть на рис. 3.

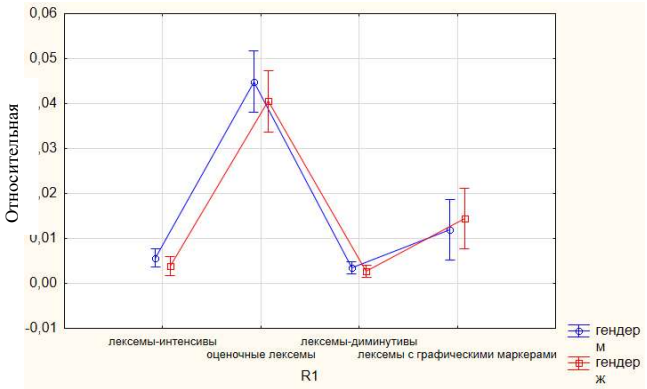


Рисунок 3. Дисперсии относительных частот использования групп экспрессивных единиц.

Исходя из результатов дисперсий по относительным величинам групп словарей, показанных на рис. 3, мы делаем вывод о том, что дисперсия разграничивает группы лексем, но не противопоставляет их относительно гендерной принадлежности участника коммуникации (уровень значимости составил 0,88).

Вследствие этого мы предположили, что подобный результат вызван высокими показателями внутригрупповых дисперсий, другими словами, частота использования маркеров может варьироваться в каждом отдельном диалоге вне зависимости от пола. Размах дисперсий представлен на рис. 4.

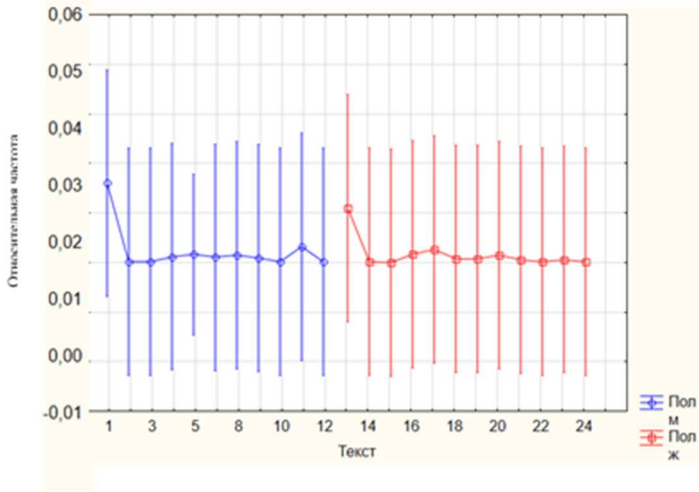


Рисунок 4. Дисперсии относительных частот использования лексем с графическими маркерами экспрессивности.

В качестве примера приведен дисперсионный анализ лексем с графическими маркерами. Разброс дисперсий, показанный на рисунке 4, свидетельствует в пользу ранее выдвинутого предположения и говорит о том, что использование экспрессивных и эмоциональных лексических единиц зависит от цели коммуникации. Последнее утверждение было проверено кластерным анализом «Уорда» (рис. 5), а также изучением коммуникативных интенций участников коммуникации.

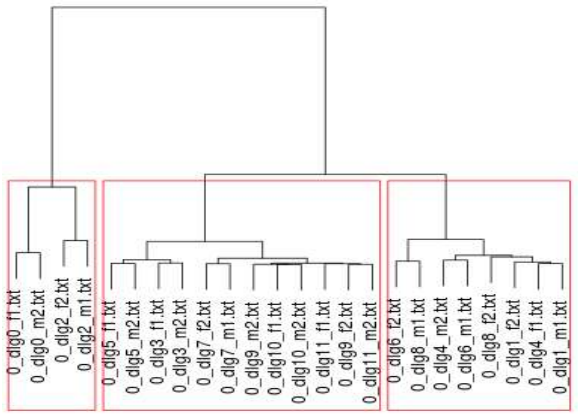


Рисунок 5. Выявление сходств в использовании групп экспрессивов на основе кластерного анализа «Уорда».

Как видно на рис. 5, все реплики разбиваются на два больших класса. Анализ интенций диалогов показал, что использование экспрессивных лексических единиц зависит, в первую очередь, от темы коммуникации и интенции участников. Подобный анализ подверждает вывод о том, что использование экспрессивных и оценочных лексических единиц в большей степени зависит от интенции коммуникантов, чем от гендера участника коммуникации

В результате мы пришли к выводу, что использование словарей неприемлемо для алгоритмов машинного обучения и предположили, что определенные лексические едиицы могут выступать в качестве маркеров. Следующим шагом исследования стало выявление подобных маркеров в тексте и применение их в машинном обучении. Для машинного обучения на «вход» подавались статистически значимые местоимения, экспрессивные и

оценочные лексические единицы. В таблице 1 представлена точность (f-мера Ван Ризенберга) машинного обучения.

Модель машинного обучения	Точность	Тип маркеров
Линейный дискриминантный анализ	75.3	Экспрессивные и оценочные лексические единицы
Логистическая регрессия	75.2	
Линейный дискриминантный анализ	85,7	Местоимения
Случайный лес	83,7	
Метод опорных векторов	64,3	

Таблица 1. Точность работы классификаторов для разных типов маркеров.

Как видно из таблицы, самым результативным классификатором оказался линейный дискриминантный анализ, основанный на маркерах местоимений. Таким образом, изучив результаты эксперимента, мы смогли подвести итог. Выявленные нами лингвистические маркеры можно использовать в машинном обучении. Дальнейшая задача подобных исследований, как мы считаем, состоит в увеличении точности работы классификатора и осуществлении поиска других лингвистических маркеров в текстах коммуникантов.

В заключении исследования подведены основные итоги работы и намечены возможные перспективы в работе.

Работы, опубликованные по теме диссертации

Статьи в журналах, включенных в Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций:

Степаненко А. А. Гендерная атрибуция текстов компьютерной коммуникации: статистический анализ использования местоимений // Вестн. Том. гос. ун-та. 2017. № 415. С. 17–25.

Степаненко А.А., Шиляев К.С., Резанова З.И. Атрибуция профессиональных интересов пользователей социальной сети «ВКонтакте» на основе текстов тематических групп и персональных страниц // Вестн. Том. гос. ун-та. Филология. 2018. № 52. С. 130-144

Степаненко А.А., Резанова З.И. Экспрессивность как маркер гендерных различий компьютерной коммуникации (к проблеме автоматической гендерной атрибуции текста) // Вестн. Том. гос. ун-та. Филология. 2018. В печати

Feshchenko, V. Goiko, G. Mozhaeva, K. Shilyaev, A. Stepanenko. Analysis of user profiles in social networks to search for promising entrants // INTED2017 Proceedings. 11th International Technology, Education and Development Conference. 2017. p. 5188-5194

Публикации в других научных изданиях:

Степаненко А.А. Местоимения как средство маркирования гендерных различий электронно-опосредствованной коммуникации // Актуальные проблемы лингвистики и литературоведения: сборник материалов V (XVII) Международной конференции молодых ученых. Томск, 20—22 апреля 2016 г. — Томск, 2016. С. 17–25

АНТИПЛАГИАТ
ТВОРИТЕ СОБСТВЕННЫМ УМОМ

Отчет о проверке на заимствования №1

Автор: stepanenkone@mail.ru / ID: 2669076
Проверяющий: (stepanenkone@mail.ru / ID: 2669076)

Отчет предоставлен сервисом «Антиплагиат»- http://www.antiplagiat.ru

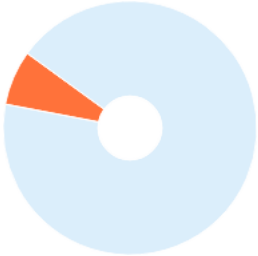
ИНФОРМАЦИЯ О ДОКУМЕНТЕ

№ документа: 7
Начало загрузки: 31.05.2018 19:42:33
Длительность загрузки: 00:00:01
Имя исходного файла:
Степаненко_доклад_23_05_2018
Размер текста: 25 кБ
Символов в тексте: 24290
Слов в тексте: 2779
Число предложений: 286

ИНФОРМАЦИЯ ОБ ОТЧЕТЕ

Последний готовый отчет (ред.)
Начало проверки: 31.05.2018 19:42:35
Длительность проверки: 00:00:01
Комментарии: не указано
Модули поиска:

ЗАИМСТВОВАНИЯ	ЦИТИРОВАНИЯ	ОРИГИНАЛЬНОСТЬ
6,74%	0%	93,26%



Заимствования — доля всех найденных текстовых пересечений, за исключением тех, которые система отнесла к цитированиям, по отношению к общему объему документа.
Цитирования — доля текстовых пересечений, которые не являются авторскими, но система посчитала их использование корректным, по отношению к общему объему документа. Сюда относятся оформленные по ГОСТу цитаты; общеупотребительные выражения; фрагменты текста, найденные в источниках из коллекций нормативно-правовой документации.
Текстовое пересечение — фрагмент текста проверяемого документа, совпадающий или почти совпадающий с фрагментом текста источника.
Источник — документ, проиндексированный в системе и содержащийся в модуле поиска, по которому проводится проверка.
Оригинальность — доля фрагментов текста проверяемого документа, не обнаруженных ни в одном источнике, по которым шла проверка, по отношению к общему объему документа.
Заимствования, цитирования и оригинальность являются отдельными показателями и в сумме дают 100%, что соответствует всему тексту проверяемого документа.
Обращаем Ваше внимание, что система находит текстовые пересечения проверяемого документа с проиндексированными в системе текстовыми источниками. При этом система является вспомогательным инструментом, определение корректности и правомерности заимствований или цитирований, а также авторства текстовых фрагментов проверяемого документа остается в компетенции проверяющего.

№	Доля в отчете	Доля в тексте	Источник	Ссылка	Актуален на	Модуль поиска	Блоков в отчете	Блоков в тексте
[01]	2,71%	2,71%	Скачать полнотекстовую ве...	http://journals.tsu.ru	30 Июл 2017	Модуль поиска Интернет	7	7
[02]	1,98%	2,07%	Doc2All.ru - Ключевая тексто...	http://doc2all.ru	11 Мая 2016	Модуль поиска Интернет	5	6
[03]	0,86%	0,86%	Информационный потенци...	http://journals.tsu.ru	20 Дек 2017	Модуль поиска Интернет	2	2

Еще источников: 5
Еще заимствований: 1,2%