УДК 81’322
DOI: 10.17223/19986645/52/8

А.А. Степаненко, К.С. Шиляев, З.И. Резанова

АТРИБУЦИЯ ПРОФЕССИОНАЛЬНЫХ ИНТЕРЕСОВ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНОЙ СЕТИ «ВКОНТАКТЕ» НА ОСНОВЕ ТЕКСТОВ ТЕМАТИЧЕСКИХ ГРУПП И ПЕРСОНАЛЬНЫХ СТРАНИЦ[1]

Исследование выполнено в рамках междисциплинарного проекта, посвященного разработке системы прогнозирования профессиональных интересов абитуриентов на основе данных социальных сетей. Автоматические методы анализа и классификации текстов на основе специальных тематических тезаурусов применяются для проверки гипотезы, согласно которой тексты персональной страницы могут выражать научные и профессиональные интересы абитуриента в рамках противопоставления «гуманитарный, естественнонаучный или математический профиль».

Ключевые слова: *обработка естественного языка, компьютерная лингвистика, классификация текстов, профориентация, социальная сеть.*

Введение

Бурное развитие социальных сетей, являющихся площадкой разноаспектного выражения знаний, интересов, мнений, оценок представителей практически всех социальных групп населения, обусловливает вовлечение их контента в современные исследования широкого спектра гуманитарных наук: социологии, психологии, культурологии, политологии и др. Один из актуальных вопросов, разрешаемых при привлечении эмпирического материала текстов социальной сети «ВКонтакте», – моделирование психологических, социальных и других аспектов личностей, формирующих данный контент.

Современная лингвистика вовлекается в исследование текстов социальных сетей как с целью рефлексии появляющихся новых дискурсов, жанров, особенностей реализации единиц всех языковых уровней в новых условиях коммуникации, так и для решения проблем, формируемых в предметном поле смежных гуманитарных наук. В таком случае наряду с лингвистическим инструментарием зачастую используются методы и данные других наук.

Работа выполнена в логике междисциплинарных исследований, объединяющих усилия социологов, лингвистов, математиков и программистов, и представляет результаты одного из этапов реализации комплексного междисциплинарного проекта. Его конечная цель – разработка системы прогнозировании профессиональной ориентации потенциальных абитури-

[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 17-16-70004.

ентов на основе автоматического анализа текстов открытого тематически немаркированного общения выпускников школ в социальных сетях. Прогнозирование профессиональной ориентации осуществляется в соответствии со сложившейся системой противопоставления направлений подготовки студентов (гуманитарный, естественно-научный и математический профиль)[2].

Решение этой задачи позволит на ранних этапах их профессионализации выявлять абитуриентов для дальнейшей работы с ними в центрах профориентации и довузовской подготовки. Таким образом, научная проблема формируется в предметном поле социологии, а ее решение ставит перед исследователями междисциплинарной команды ряд конкретных задач, решаемых при использовании теоретико-методоло-гического аппарата отдельных наук [1–3].

В данной статье представлены результаты первого этапа выполнения проекта с использованием лингвистических методов анализа текста, результаты которого применяются в методиках автоматической обработки текстовых массивов, в первую очередь – через создание автоматических классификаторов текстов личных страниц «ВКонтакте» по трём группам интересов их авторов: гуманитарное, естественное и математическое.

Разработка данного этапа является проверкой исследовательской гипотезы о том, что пользовательские данные потенциальных абитуриентов в социальной сети «ВКонтакте» содержат сведения об их интересах к той или иной предметной области и о том, что эти сведения могут быть формализованы и в конечном итоге станут основой автоматических моделей определения направлений профессиональной ориентации исследуемых текстов.

При выдвижении данной гипотезы и при определении способов и этапов ее проверки мы опирались на ряд предшествующих социологических исследований интересов пользователей социальных сетей [4–7], на лингвистические работы, посвященные анализу дискурсивно-жанрового своеобразия анализируемых текстов и лингвистическому моделированию языковой дискурсивной личности [8]. Кроме того, был использован опыт применения междисциплинарного использования методов лингвистического анализа текста, автоматической обработки значительных текстовых массивов и статистической проверки лингвистических гипотез [9–11].

---

[2] Отметим, что в ситуации нарастающей потребности в междисциплинарности при подготовке ряда современных профессий в системе современного вузовского образования уровня бакалавриата доминирует определяемый госстандартами внутридисциплинарный подход, например направление «Филология» и под. Процессы междисциплинарной интеграции, как правило, реализуются на следующих уровнях современной системы российского образования – магистратуры и аспирантуры.

Материал и источники анализа

Основным источником данных для лингвистического портретирования абитуриента явилась *стена* персональной интернет-страницы пользователя социальной сети «ВКонтакте».

Стена персональной страницы как одно из средств конструирования и отражения интернет-идентичности пользователя исследовалась в работах Т.В. Алтуховой, Л.И. Ермоленкиной, Е.А. Костяшиной, З.И. Резановой, А.В. Щекотурова [12–15] и др. Исследователи характеризуют различные речевые жанры в текстовом пространстве стены личных страниц в социальных сетях: поздравление, реклама, информационные и развлекательные сообщения [16, 17], среди которых преобладают ссылки на другие интернет-ресурсы и репосты новостей, развлекательных видеороликов, вербальных и визуальных мемов. В целом стена отражает выделенные многочисленными исследователями особенности виртуальной коммуникации [18, 19]: гипертекстуальность и интерактивность медийной среды (наличие комментариев и репостов, причем последние составляют преимущественное наполнение стен многих пользователей); интенсивность использования мультимедиа (наличие мультимедиа-контента, нередко в составе репоста); нарастающее присутствие элементов невербального общения (высокий процент сообщений, содержащих изображение либо состоящих из одного изображения). Клиповый и преимущественно визуальный характер содержимого стены, краткость и репродуктивный характер ее текстового содержимого ограничивают применимость эффективных в других случаях традиционных методов анализа текста: грамматико-синтаксического, лексического, прагматического, фреймового и т.п. Однако при дополнении их другими методами, в том числе, как мы попытаемся показать далее на примере лексического анализа, математического моделирования с использованием автоматических классификаторов текстов, они могут продемонстрировать свою эффективность.

Все традиционно выделяемые в составе стены речевые жанры могут быть разделены на два функциональных типа, противопоствляемых по отношению к создателю страницы: **посты** содаются автором страницы, **репосты** представляют собой копии чужих постов, распространяемых автором страницы вследствие совпадения интересов и точки зрения автора и транслятора репоста. Именно эта особенность явилась основанием включения в анализ при моделирования профессиональных направленностей пользователей социальной сети «ВКонтакте» **текстовых материалов постов и репостов**. Большинство постов и репостов, составивших материал автоматизированного контент-анализа, представляют собой краткое (300 знаков) сообщение научно-популярного или развлекательного характера, как правило, снабженное фотографией или изображением иного рода (рисунок, инфографика), передающее в доступной форме актуальные события и новости научного направления, которому посвящено сообщество.

В качестве конкретного материала анализа нами были привлечены следующие:

1. На первом этапе исследованы посты и репосты стен *профессионально ориентированных* сообществ «ВКонтакте», соответствующие трем научным направлениям: математическое, гуманитарное и естественное. К анализу привлекались: тексты стен сообществ, сгруппированные по отнесенности к тематическим группам, были исследованы тексты 22 стен групп математического направления – «Роботы и робототехника», «МЕХатроника и BIOS», «Информационные технологии и системы» и др.; 14 стен групп естественного направления – «Клуб National Geographic Россия», «Добрая Экология» и др.; 48 стен групп гуманитарного – «Латынь – это интересно!», «Лучшие стихи великих поэтов» и др.

Привлечение к лингвистическому анализу данного типа текстов создало основание для определения состава лексических маркеров их тематической дифференциации. Статистический анализ выделенных групп маркеров позволил обосновать возможность атрибуции текстов.

2. На втором этапе мы изменили принципы отбора материала: к анализу был привлечен весь текстовый контент открытых стен пользователей социальной сети «ВКонтакте», являющихся студентами первого курса Томского государственного университета. Были изменены и принципы организации первичных текстовых источников – они также были разделены в соответствии с тремя направлениями образования. Принципиально важным был отбор текстов всего контента стены пользователя: на данном материале проверялась гипотеза о том, что выделенный на первом этапе состав лексических маркеров может стать основой классификации текстов тематически неориентированного общения пользователей социальной сети «ВКонтакте».

Анализ и обсуждение результатов

Как уже было отмечено ранее, на первом этапе с целью выявления маркеров профессиональных интересов абитуриентов ализировались тексты тематических групп, относящихся к одной из предметных областей знаний. Поскольку стена является отражением виртуального самопозиционирования абитуриента, мы склонны считать, что встречаемость релевантных лингвистических маркеров демонстрирует интерес к предметной области.

Для выявления ключевых слов мы намеренно не использовали терминологические словари, ориентированные на профессиональных пользователей, но предпочли выбирать лексику путем контент-анализа сообществ. Основой создания тематических тезаурусов стал частотный анализ лексических единиц, использованных в текстах стен сообществ: общий объем всех стен различных сообществ составил 1254 Мб (114 592 246 слов), в том числе стен математического направления – 217 Мб (16752913 слов), естественного – 428 Мб (28 533 920 слов) и гуманитарного – 609 Мб (69 305 413 слов).

Анализ проводился на лемматизированном программой «MyStem 3.0» [20] текстовом материале. Предварительный этап анализа текста включал также приведение всех лексических единиц к единому формату (индексация), а также исключал из анализа служебные части речи (предлоги, союзы, частицы), знаки препинания, не несущие смысловой нагрузки. Все эти действия позволили представить исследуемые тексты (стены пользователей) в виде вектора. Другими словами, каждый текст был представлен как набор лексических единиц (атрибутов), приведенных к единому формату (нормализация текста), что позволило осуществить более точный формально-количественный анализ. Нормализация включала лингвистические компоненты: лемматизацию (приведение всех словоформ к начальной форме) и редуцирование так называемых «стоп-слов» (служебных частей речи), а также технические: удаление знаков препинания и фиксацию единого регистра (нижнего) для всех слов в тексте. Тем самым исключается влияние на относительную частоту маркеров неинформативных признаков и повторяющихся лексем.

Далее слова в нормализованных текстах были выбраны вручную по принципу «один список – несколько сообществ по одной теме». Например, список слов по биологический тематике был составлен на основе сообществ «Клуб National Geographic Россия», «Экология | Пермакультура | ЭКО-Поселения Природа», «Углубленный биолог» и др. Затем списки были укрупнены: объединенные списки по философии, социологии, филологии и лингвистике, журналистике, юридическим наукам, истории составили гуманитарный тезаурус; списки по физике, химии, биологии, химии – тезаурус естественных наук; по математическим дисциплинам – математический тезаурус.

Отбор лексических единиц по критерию частотности осуществлялся следующим образом: при включении слова в тезаурус относительная частотность слов должна превышать 0,001% и быть меньше 0,01%. В случае сдвига диапазона в сторону увеличения в выборку попадает большое количество общеупотребительных слов, не связанных с конкретной предметной областью. При сдвиге диапазона в сторону более низких значений из выборки начинают исчезать наиболее популярные слова, относящиеся к тематическим группам, важным для идентификации научной ориентации абитуриента[3]. В тезаурус для последующего автоматического анализа текста были отобраны слова, отвечающие следующим критериям: а) терминологический характер или соотнесенность с определенной предметной областью (*аминокислота, биоразнообразие, гоминид, ихтиология, геоинформационный, картографический, космосъемка, герундий, безударный, зарисовка, аллегория, великокняжеский, воцарение, априорный, идеалистический, вменять, дознание* и т.п.); б) принадлежность к классу имен соб-

[3] Указанный диапазон был установлен эмпирически для конкретной выборки текстов и не может быть использован в качестве безусловного ориентира для решения подобных задач.

ственных, связанных с данной областью знания (*Альтюссер, Вундт, Гуссерль, Бэкон, Геродот* и т.п.), более характерные для гуманитарных наук. В случае затруднений при классификации лексики для справки использовались терминологические и энциклопедические словари различных областей знания, а также тезаурус WordNet (привлекался иностранный эквивалент). В данном исследовании использовались укрупненные тематические группы (деление на гуманитарные, естественные и математические науки), что позволило снизить остроту проблемы вхождения некоторых лексем-терминов в несколько тематических групп одновременно (например, *атом, давление, площадь* и т.п.).

Неспецифическая и общенаучная лексика (такие слова, как *анализ, исследование, разработка* и т.п.), интернет-сленг и молодежный сленг, попавшие в выборку, отсеивались вручную. Получившиеся в результате словари содержали лексику, демонстрирующую низкий уровень полисемии и омонимии, что способствовало снижению количества ложных срабатываний алгоритма и уточнению вектора слов. В то же время эта лексика обладала достаточной частотой встречаемости в сообществах, подвергшихся контент-анализу, чтобы присутствовать на странице в форме репостов.

В результате был составлен словарь ключевых слов, включающий 432 лексических единицы по гуманитарным направлениям, 120 единиц – по естественным и 126 – математическим. Примеры выявленных лексических маркеров (по 5 слов, относящихся к числу наиболее частотных, занимающих срединное положение и наименее частотных), приведены в табл. 1.

Таблица 1

Примеры лексических маркеров, используемых в атрибуции текстов

Естественный	**Гуманитарный**	**Математический**
Состав – 0,00165	Древний – 0,0744	Высота – 0,00168
Реактив – 0,00163	Империя – 0,0721	Множество– 0,00150
Химик – 0,00143	Дворянство – 0,0048	График – 0,00146
Химический – 0,00130	Анализ – 0,0048	Луч – 0,00016
Бензол – 0,00029	Граница – 0,0214	Квантовый – 0,00016

С целью проверки корректности выбора маркеров (признаков), которые в дальнейшем служат для классификации стен пользователей, был предложен тест для определения уровня значимости по критерию Стьюдента, выявляющий различия использования лексических единиц тематического тезауруса в естественных и математических стенах сообществ, и тест Краскела – Уоллиса – для гуманитарных. В результате статистического анализа было установлено, что использование лексических единиц отличается в естественных и математических тезаурусах со значением p-value: 0,004247 естественного тезауруса и со значением p-value: 0,004323 для математического. Что касается критерия значимости гуманитарного тезауруса, то уровень значимости p-value составил 0,4232. Так как для принятия гипотезы о значимости был выбран стандартный порог 0,05 (означающий,

что с вероятностью 95% данный результат не является случайным), мы принимаем альтернативную гипотезу для математического и естественного тезаурусов. Другими словами, разработанные тезаурусы действительно влияют на классификацию текстов.

Визуально различия дисперсий частот маркеров тематических тезаурусов относительно текстов стен сообществ можно представить в виде диаграмы размаха (рис. 1).

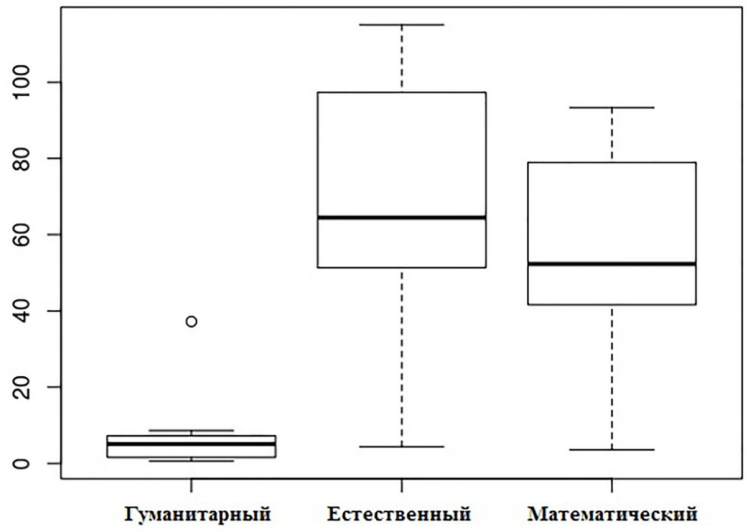

Рис. 1. Диаграмма размаха частот лексических единиц тематических тезаурусов относительно стен групп «ВКонтакте»

На оси ординат представлены данные о средней частоте использования лексических единиц по тематическим тезаурусам, на оси абсцисс – тезаурусы.

Значение p-value по критерию Стьюдента в постах естественных и математических сообществ позволило сделать вывод, что тематические тезаурусы могут быть использованы при классификации текстов стен пользователей по исследуемым аспектам. Однако было замечено, что уровень значимости маркеров тезауруса гуманитарного направления превышает допустимый порог p = 0,05, т.е. вероятность того, что маркеры гуманитарного тезауруса являются уникальными относительно маркеров других научных направлений, меньше стандартных 95%, что может повлиять на точность работы классификатора.

2. **На втором этапе** к анализу были привлечены тексты персональных страниц пользователей социальной сети «ВКонтакте», обучающихся в Томском государственном университете на факультетах трех основных направлений – гуманитарного, естественно-научного и математического

профилей. Основная цель данного анализа – проверить возможность использования маркеров, выделенных на первом этапе, в процедурах автоматической классификации текстов по установленным параметрам.

Как было отмечено ранее, в анализируемый материал были включены все публикации от имени автора и репосты стен пользователей. Минимальный объем текста одной стены составил 20 Кб (около 1 600 слов). Тексты стен персональных страниц были разделены на три группы по признаку обучения студента на факультете одного из трех направлений. Основная процедура анализа на данном этапе – расчет частотности единиц выделенных тезаурусов в текстах каждого из направлений.

Частотный анализ лексических единиц рассчитывался по классической формуле определения вероятности $P = \frac{m}{n}$, где P – вероятность использования лексики, принадлежащей определенной тематической группе тезауруса; m – абсолютная частота термина; n – сумма частот всех лексических единиц в словаре.

Таким образом, редуцируя абсолютные частоты лексических единиц, мы исключили возможность влияния объема текста на значимость атрибутов (единиц тематических тезаурусов).

Каждый атрибут a (лексическая единица тематического тезауруса), представленный в тексте $a_i \in T$ (T – совокупность всех текстов), имеет вес w_{ij} по отношению к тематически направленному документу $d_j \in D$ (D – совокупность всех документов). Таким образом, каждый текст T_i представлен в виде вектора весов его атрибутов $d_j = < p_{1j} \ldots p_{|T|j} > \in D$.

Пример представления вычислений объединенных векторов слов, взятых случайным образом из матрицы, но относящихся к разным предметным сферам, представлен в табл. 2. Строки представляют тексты персональных страниц студентов с указанием направления подготовки, в столбцах представлены значения относительных частот атрибутов (лексем).

Нами была проведена валидизация тезаурусов методами статистики: использован непараметрический критерий Краскела – Уоллиса ($p = 0{,}0456$), что говорит о значимом различии текстов по исследуемому параметру, т.е. тезаурусы различают тексты персональных страниц пользователей социальной сети «ВКонтакте», противопоставленные по направлениям обучения.

Различия использования тематических маркеров наглядно отображает диаграмма размаха частот (рис. 2): на оси ординат представлена средняя частота использования лексических единиц по тематическим тезаурусам на страницах пользователя социальной сети «ВКонтакте», на оси абсцисс – тезаурусы.

Таким образом, статистический анализ свидетельствует о возможности использования маркеров в процессе атрибуции текстов персональных страниц в соответствии с профессиональными интересами их авторов.

Таблица 2

Относительные частоты использования маркеров в текстах персональных страниц студентов трех направлений подготовки

Направление подготовки пользователей	Атрибуты текстов (лексемы)			
	Мысль	Ферма	Процессор	Трек
Текст 1(d_1), гуманитарное	0.0416	0.042	0	0
Текст 2 (d_2), гуманитарное	0	0	0	0
Текст 3(d_3), гуманитарное	0.222	0	0	0
Текст 4(d_4), гуманитарное	0.667	0	0	0
Текст 5(d_5), естественное	0	0	0	0
Текст 6(d_6), естественное	0	0	0	0
Текст 7(d_7), естественное	0.143	0.429	0	0
Текст 8(d_8), естественное	0	0	0	0
Текст 9(d_9), естественное	0.008	0	0.008	0.038
Текст 10(d_{10}), естественное	0.045	0	0.045	0.136
Текст 11(d_{11}), математическое	0.143	0	0	0
Текст 12(d_{12}), математическое	0.032	0.032	0	0
Текст 13(d_{13}), математическое	0.143	0.429	0	0
Текст 14(d_{14}), математическое	0.008	0	0.008	0.038
Текст 15(d_{15}), математическое	0.045	0	0.045	0.136
Текст 16(d_{16}), математическое	0	0	0	0

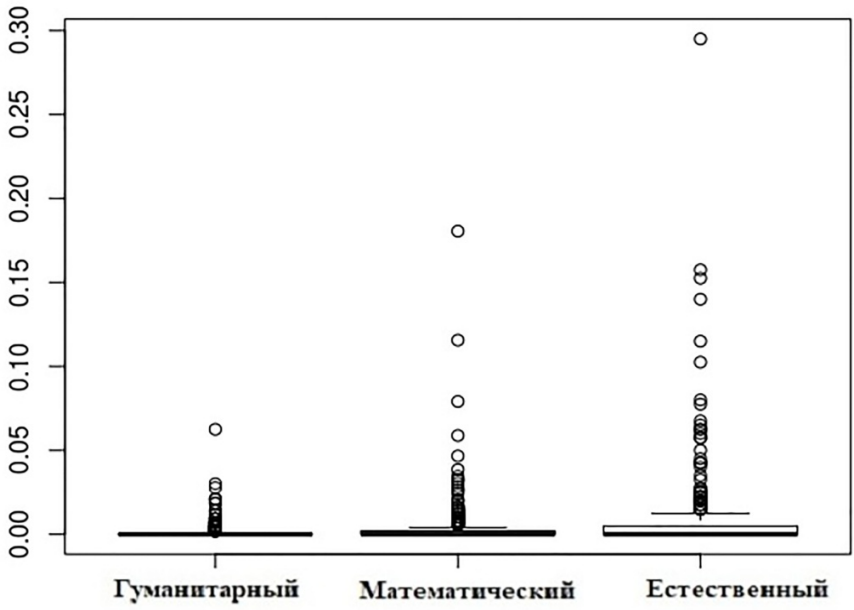

Рис. 2. Диаграмма размаха частот лексических единиц по тематическим тезаурусам относительно стен пользователей «ВКонтакте»

На следующем этапе исследования был осуществлен поиск наиболее точного классификатора текстов по заданным маркерам (в терминологии статистического анализа – атрибутам)[4].

При построении алгоритма машинного обучения на вышеуказанной формальной модели текстов пользователей «ВКонтакте» с заранее известным классом (факультет обучения: гуманитарный, естественный, математический) весь массив текстов был разбит на две непересекающиеся части 70:30%: «Обучающую» (Tr) и «Тестовую» (Te), где «Обучающая» – $Tr = \{d1 \dots d|Tr|\}$, по которой создается классификатор Ф', а также «Тестовая» ($Te = \{d|Tr|+1 \dots d|Q|\}$) – массив стен пользователей, на которых производится качество работы классификатора. Обе группы текстов (X) содержат вектор атрибутов с относительной частотой лексических единицм(x_i), сохраненных в тезаурусе $X \rightarrow \{x_1, x_2, \dots x_i\}$.

Для классификации текстов нами использовались следующие виды классификаторов: линейный дискриминантный анализ (LDA), метод опорных векторов (SVM), логистическая регрессия (LR), деревья решений (Trees), случайный лес (RF). Данные виды классификаторов были выбраны, так как эффективность их использования была установлена при решении задачи классификации текстов по лексическим маркерам в ряде предшествующих работ (см., например: [21, 22]).

В табл. 3 представлены результаты классификации текстов (Tr_{acc} обозначает точность работы классификатора обучающей выборки, Te_{acc} – точность работы классификатора тестовой выборки).

Таблица 3

Оценка точности классификации текстов на основе группы классификаторов

Классификатор	Tr_{acc}	Te_{acc}
LDA	62,67	60,71
LR	65,33	64,29
SVM	64	60,71
Trees	49,33	50
RF	65,33	57,14

Как видно из таблицы, наиболее успешный классификатор – LR (Логистическая регрессия), показывающий стабильный результат распределения стен пользователей. На рис. 2 представлены результаты использования данного классификатора.

[4] Алгоритм классификатора текстов: задано конечное множество сущностей (атрибутов) $C = \{c1 \dots c|C|\}$, конечное множество стен пользователей $D = \{d1 \dots d|D|\}$ и неизвестная функция Ф, которая определяет, соответствуют ли они друг другу: Ф: D×C $\rightarrow$ {0,1}. Задача состояла в том, чтобы найти максимально близкую к функции Ф функцию Ф'.

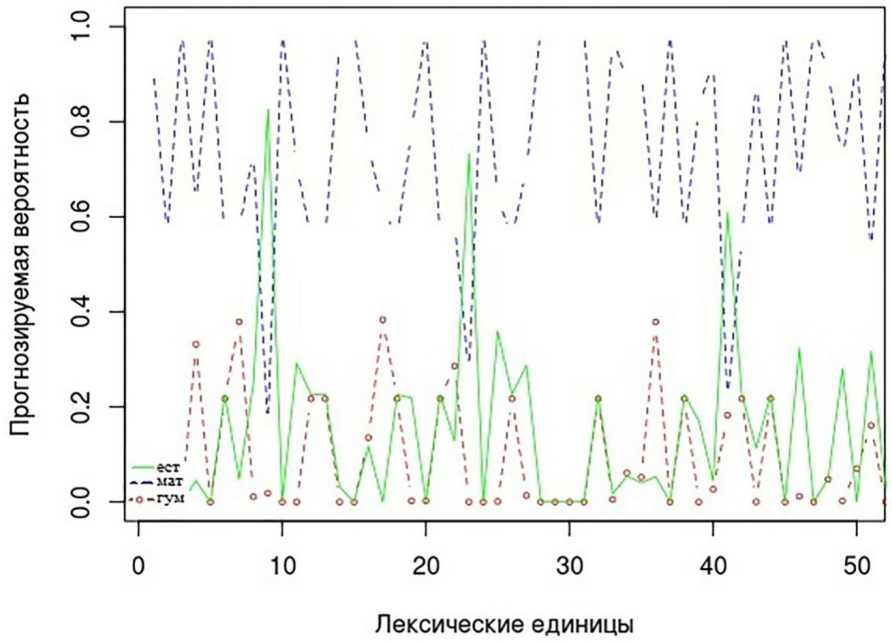

Рис. 3. Прогнозируемая вероятность лексических единиц по научным направлениям на основе логистической регрессии

Как показывает график классификации, все тексты разделяются на три класса: гуманитарный, естественный и математический, что свидетельствует о корректности работы составленных нами тезаурусов. Однако погрешности анализа вызывают тексты, относящиеся к гуманитарному научному направлению. Полагаем, что это обусловлено наличием большого количества лексических единиц в гуманитарном тезаурусе, которые также встречаются в постах (или стенах) пользователей социальных сетей двух других научных направлений, об этом свидетельствует и вышеописанный критерий проверки уровня значимости гуманитарного тезауруса относительно стен сообществ ($p = 0{,}4232$). Подобный результат требует более детального изучения атрибутов тезауруса (в частности, гуманитарного направления), поиска оптимальных моделей классификации и применения методов мультиколлинеарности.

Данную проблему можно объяснить тем что количество наблюдений (текстов) меньше числа атрибутов. Главная задача мультиколлинеарности – выявить наиболее информативные атрибуты, влияющие на классификацию текстов, что позволит найти более подходящие признаки для гуманитарного тезауруса и улучшить точность работы классификатора.

Таким образом, исходная гипотеза о возможности формализации сведений об интересах пользователей социальной сети «ВКонтакте» к той или иной предметной области и об их использовании в системах автоматиче-

ской классификации текстов была подтверждена. Проведенный анализ свидетельствует, во-первых, о результативности автоматической классификации текстов тематически свободного общения в социальной сети «ВКонтакте» с использованием в качестве атрибутов выделенных на первом этапе ключевых слов текстов групп профессионально ориентированного общения; во-вторых, о меньшей степени релевантности в рассматриваемом аспекте данных гуманитарного блока профессионализации. В последнем случае необходимо работать в направлении как уточнения состава маркеров, так и поиска более точных систем автоматической обработки текста. Однако уже на этом уровне можно говорить о возможности применять данную модель классификации к текстам других групп пользователей социальной сети «ВКонтакте», что предполагается осуществить на следующем этапе работы.

Литература

1. *Можаева Г.В., Слободская А.В., Фещенко А.В.* Информационный потенциал социальных сетей для выявления образовательных потребностей школьников // Открытое и дистанционное образование. 2017. № 3 (67). С. 25–30. DOI: 10.17223/16095944/67/4

2. *Feshchenko A., Goiko V., Stepanenko A.* Recruiting university entrants via social networks // EDULEARN17 Proceedings 9th International Conference on Education and New Learning Technologies. P. 6077–6082. DOI: 10.21125/edulearn.2017.2375

3. *Feshchenko A., Goiko V., Mozhaeva G., Shilyaev K., Stepanenko A.* Analysis of user profiles in social networks to search for promising entrants // INTED2017 Proceedings, 11th International Technology, Education and Development Conference. March 6th–8th, 2017. P. 5188–5194. DOI: 10.21125/inted.2017.1203

4. *Коршунов А.В.* Задачи и методы определения атрибутов пользователей социальных сетей // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Ярославль, 2013. С. 380–390.

5. *Kim J. et al.* Extracting User Interests on Facebook //International Journal of Distributed Sensor Networks. 2014. Vol. 10, №. 6. P. 1–5.

6. *Ahmed A. et al.* Scalable distributed inference of dynamic user interests for behavioral targeting // Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011. P. 114–122.

7. *Michelson M., Macskassy S.A.* Discovering users' topics of interest on twitter: a first look // Proceedings of the fourth workshop on Analytics for noisy unstructured text data. ACM, 2010. P. 73–80.

8. *Резанова З.И., Скрипко Ю.К.* Личность в среде дискурса: языковая репрезентация социально-психологических типов (на материале дискурса виртуальных фансообществ музыкальной направленности) // Вестник Томского государственного университета. Филология. 2016. № 3 (41). DOI: 10.17223/19986645/41/4

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

10. *Резанова З.И., Романов А.С., Мещеряков Р.В.* О выборе признаков текста, релевантных в автороведческой экспертной деятельности // Вестник Томского государственного университета. Филология. 2013. № 6 (26). С. 38–52.

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

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

13. *Ермоленкина Л.И., Костяшина Е.А.* Коммуникативно-языковые механизмы формирования этнокультурной идентичности в дискурсивном пространстве интернета // Вестник Томского государственного университета. Культурология и искусствоведение. 2013. № 3 (11). С. 5–15.

14. *Щекотуров А.В.* Конструирование виртуальной гендерной идентичности подростков на страницах социальной сети «ВКонтакте» // Женщина в российском обществе. 2012. № 4 (65). С. 31–43.

15. *Резанова З.И.* Институциональная и личностная презентация национально-культурной идентичности в интернет-коммуникации: жанровые формы и дискурсивные стратегии // Вестник Томского государственного университета. 2013. № 375. С. 33–41.

16. *Алтухова Т.В.* Электронные и рукописные жанры естественной письменной речи: сопоставительный аспект (на примере граффити и записей на электронной стене) // Вестник Кемеровского государственного университета.. 2012. № 2 (50). С. 110–116.

17. *Марковская А.С.* Особенности поздравления с днем рождения в социальных сетях // Вестник Московского государственного университета. Сер. 19. Лингвистика и межкультурная коммуникация. 2013. № 4. С. 153–159.

18. *Горошко Е.И., Полякова Т.Л.* К построению типологии жанров социальных медий // Жанры речи. 2015. № 2 (12). С. 119–127.

19. *Горошко Е.И.* Современные интернет-коммуникации: структура и основные параметры // Интернет-коммуникации как новая речевая формация. М., 2012. С. 9–52.

20. *MyStem* // Яндекс. 2014–2017. URL: https://tech.yandex.ru/mystem/ (дата обращения: 1.11.2017).

21. *Sheshasaayee A., Thailambal G.* Comparison of Classification Algorithms in Text // International Journal of Pure and Applied Mathematics. 2017. Vol. 116, № 22. P. 425–433.

22. *Singhal A., Gopalakrishnan K., Khaitan S.K.* Predicting Budget from Transportation Research Grant Description: An Exploratory Analysis of Text Mining and Machine Learning Techniques // Journal of Soft Computing in Civil Engineering. 2017. № 1–2. P. 89–102.

ATTRIBUTION OF PROFESSIONAL INTERESTS OF SOCIAL NETWORK USERS BASED ON SUBJECT-ORIENTED GROUPS AND PERSONAL PAGES

Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology. 2018. 52. 130–144. DOI: 10.17223/19986645/52/8

Andrey A. Stepanenko, Konstantin S. Shilyaev, Zoya I. Rezanova, Tomsk State University (Tomsk, Russian Federation). E-mail: stepanenkone@mail.ru / shilyaevc@gmail.com / rezanovazi@mail.ru

Keywords: natural language processing, computational linguistics, text classification, vocational training, social network.

The present study is part of an interdisciplinary project that unites the efforts of sociologists, linguists, mathematicians and IT specialists. The authors' ultimate aim is to create a system for predicting the career choice of prospective students that would use automatic text analysis of freely available and topically unconstrained data produced by school leavers in social networks.

The current stage presents the testing of a hypothesis that user data of prospective students in the social network VKontakte contain information about their interests for a certain subject, and this information can be formalized and become the basis of automatic models for determining the career choice of text producers.

The main source of data for the prospective student's linguistic portrayal is the wall of the personal profile which contains both original posts and reposts that are shared by the page owner.

The first stage of analysis consists in studying posts and reposts on the walls of subject-oriented communities in VKontakte which were matched to three subject areas: mathematics, humanities and natural science. This content was used to determine which lexical markers could differentiate between texts of different subject areas and serve as their markers.

At the second stage of analysis, the authors used the textual content of freely available user walls that belonged to TSU first-year students. The principles of organizing primary textual sources consisted in dividing them into the three aforementioned subject areas. Downloading all the texts from a user's wall was highly important, since the material was used to test the hypothesis that the set of lexical markers discovered at the first stage could be used to automatically classify the texts of thematically unconstrained communication of VKontakte users.

The statistical analysis showed that it was possible to apply the markers in text attribution tasks according to the three subject areas the prospective student might be most interested in. Among the classifiers tested using the markers derived in the first stage, Logistic Regression proved to be the most successful in dividing the texts into three classes: humanitites, natural science and mathematics; this also proves that the subject-area thesauri used functioned correctly.

Overall, the study shows the effectiveness of using automatic text classification of thematically unconstrained communication in VKontakte with keywords derived from texts belonging to particular subject areas. The problem that awaits its resolution is the relatively low discriminatory power of humanities keywords, probably due to their widespread usage in social network communication on the whole.

References

1. Mozhaeva, G.V., Slobodskaya, A.V. & Feshchenko, A.V. (2017) Informational potential of social networks for revealing pupils educational needs. *Otkrytoe i distantsionnoe obrazovanie – Open and distance education*. 3(67). pp. 25–30. DOI: 10.17223/16095944/67/4

2. Feshchenko, A., Goiko, V. & Stepanenko, A. (2017) Recruiting university entrants via social networks. *EDULEARN17*. Proceedings 9th International Conference on Education and New Learning Technologies. pp. 6077–6082. DOI: 10.21125/edulearn.2017.2375

3. Feshchenko, A. et al. (2017) Analysis of user profiles in social networks to search for promising entrants. *INTED2017*. Proceedings, 11th International Technology, Education and Development Conference. 6–8 March 2017. pp. 5188–5194. DOI: 10.21125/inted.2017.1203

4. Korshunov, A.V. (2013) Problems and methods for attribute detection of social network users. *Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii* [Electronic libraries: promising methods and technologies, electronic collections]. Proceedings of the conference. Yaroslavl: Yaroslavl State Technical University. pp. 380–390.

5. Kim, J. et al. (2014) Extracting User Interests on Facebook. *International Journal of Distributed Sensor Networks*. 2. pp. 1–5. DOI: 10.1155/2014/146967

6. Ahmed, A. et al. (2011) Scalable distributed inference of dynamic user interests for behavioral targeting. *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM. pp. 114–122.

7. Michelson, M. & Macskassy, S.A. (2010) Discovering users' topics of interest on twitter: a first look. *Proceedings of the fourth workshop on analytics for noisy unstructured text data*. ACM. pp. 73–80.

8. Rezanova, Z.I. & Skripko, Yu.K. (2016) Personality in the medium of discourse: linguistic representation of psychological types (based on the discourse of virtual music fan communities). *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology*. 3 (41). pp. 37–56. (In Russian). DOI: 10.17223/19986645/41/4

9. Rezanova, Z.I., Romanov, A.S. & Meshcheryakov, R.V. (2013) Tasks of author attribution of text in the aspect of gender (on interdisciplinary interaction of linguistics and

computer science). *Vestnik Tomskogo gosudarstvennogo universiteta – Tomsk State Universi-University Journal*. 370. pp. 24–28. (In Russian).

10. Rezanova, Z.I., Romanov, A.S. & Meshcheryakov, R.V. (2013) Selecting text features relevant for authorship attribution. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology*. 6 (26). C. 38–52. (In Russian). DOI: 10.17223/19986645/26/4

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

12. Altukhova, T.V. (2012) Sotsial'naya komp'yuternaya set' "VKontakte": zhanrovaya kharakteristika [Social computer network "VKontakte": genre characteristics]. *Vestnik KemGU – Bulletin of Kemerovo State University*. 4 (52):3. pp. 21–25.

13. Ermolenkina, L.I. & Kostyashina, E.A. (2013) Communicative and linguistic mechanisms of ethnic and cultural identity in the discourse space of the Internet. *Vestnik Tomskogo gosudarstvennogo universiteta. Kul'turologiya i iskusstvovedenie – Tomsk State University Journal of Cultural Studies and Art History*. 3 (11). pp. 5–15. (In Russian).

14. Shchekoturov, A.V. (2012) Konstruirovanie virtual'noy gendernoy identichnosti podrostkov na stranitsakh sotsial'noy seti "VKontakte" [Construction of virtual gender identity of adolescents on the pages of social network "VKontakte"]. *Zhenshchina v rossiyskom obshchestve*. 4 (65). pp. 31–43.

15. Rezanova, Z.I. (2013) Institutional and personal presentation of national and cultural identity in Internet communication: genre forms and discursive strategies. *Vestnik Tomskogo gosudarstvennogo universiteta – Tomsk State University Journal*. 375. pp. 33–41. (In Russian). DOI: 10.17223/15617793/375/6

16. Altukhova, T.V. (2012) Elektronnye i rukopisnye zhanry estestvennoy pis'mennoy rechi: sopostavitel'nyy aspekt (na primere graffiti i zapisey na elektronnoy stene) [Electronic and hand-written genres of natural written speech: a comparative aspect (on the example of graffiti and posts on an electronic wall)]. *Vestnik KemGU – Bulletin of Kemerovo State University*. 2 (50). pp. 110–116.

17. Markovskaya, A.S. (2013) Osobennosti pozdravleniya s dnem rozhdeniya v sotsial'nykh setyakh [Features of birthday greetings in social networks]. *Vestn. Mosk. un-ta. Ser. 19. Lingvistika i mezhkul'turnaya kommunikatsiya – Bulletin of Moscow University. Series 19. Linguistics and Cross-Cultural Communication*. 4. pp. 153–159.

18. Goroshko, E.I. & Polyakova, T.L. (2015) The construction of genre typology of the social media. *Zhanry rechi – Speech Genres*. 2 (12). pp. 119–127. (In Russian).

19. Goroshko, E.I. (2012) *Sovremennaya internet-kommunikatsiya: struktura i osnovnye parametry* [Modern Internet communication: structure and basic parameters]. Moscow: Flinta. pp. 9–52.

20. Yandex. (2014–2017) *MyStem* [Online] Available from: https://tech.yandex.ru/mystem/. (Accessed: 1st November 2017)

21. Sheshasaayee, A. & Thailambal, G. (2017) Comparison of Classification Algorithms in Text. *International Journal of Pure and Applied Mathematics*. 116(22). pp. 425–433.

22. Singhal, A., Gopalakrishnan, K. & Khaitan, S.K. (2017) Predicting Budget from Transportation Research Grant Description: An Exploratory Analysis of Text Mining and Machine Learning Techniques. *Journal of Soft Computing in Civil Engineering*. 1–2. pp. 89–102. DOI: 10.22115/scce.2017.49604