УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2018 Том 18 № 3 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2018 Vol. 18 No 3 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.89

МЕЖСАЙТОВАЯ ЛИНГВИСТИЧЕСКАЯ ИДЕНТИФИКАЦИЯ ИНТЕРНЕТ-ПОЛЬЗОВАТЕЛЕЙ

А.А. Воробьева[a], В.А. Позволенко[a], А.С. Коробицына[a], А.А. Шарафиев[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: alice_w@mail.ru

Информация о статье
Поступила в редакцию 01.03.18, принята к печати 05.04.18
doi: 10.17586/2226-1494-2018-18-3-447-456
Язык статьи – русский

Ссылка для цитирования: Воробьева А.А., Позволенко В.А., Коробицына А.С., Шарафиев А.А. Межсайтовая лингвистическая идентификация интернет-пользователей // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 3. С. 447–456. doi: 10.17586/2226-1494-2018-18-3-447-456

Аннотация
Исследованы вопросы межсайтовой лингвистической идентификации пользователей интернет-ресурсов по коротким электронным сообщениям, полученным из нескольких источников (сайтов, средств онлайн-коммуникации). Рассмотрена возможность идентификации пользователя одного интернет-ресурса по его сообщениям на другом интернет-ресурсе. Определена возможность формирования единого признакового пространства для сообщений, полученных из различных источников, обеспечивающая достаточную точность лингвистической идентификации. Показано, что существует стилистическая разница между текстами сообщений, созданными одним пользователем, но с использованием различных средств коммуникации. Рассмотрены две задачи межсайтовой идентификации: 1) идентификация по смешанным данным – обучающая и тестовая выборки сформированы из сообщений, полученных из нескольких источников (сайтов); 2) разделенные источники – обучающая выборка сформирована из сообщений одного источника, тестовая из сообщений другого источника. Результаты экспериментов показали, что при обучении на смешанных данных достоверность идентификации составляет 0,82, при обучении на данных различных источников достоверность идентификации – 0,74. Сделаны выводы, что существует стилистическая разница между текстами сообщений, созданными одним пользователем, но с использованием различных средств коммуникации. Но в то же время существует возможность сформировать единое признаковое пространство для сообщений, полученных из различных источников, обеспечивающее достаточную точность лингвистической идентификации.

Ключевые слова
идентификация интернет-пользователей, лингвистическая идентификация, межсайтовая идентификация.

CROSS-DOMAIN WEB AUTHOR IDENTIFICATION

A.A. Vorobeva[a], V.A. Pozvolenko[a], A.S. Korobitsyna[a], A.A. Sharafiev[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: alice_w@mail.ru

Article info
Received 01.03.18, accepted 05.04.18
doi: 10.17586/2226-1494-2018-18-3-447-456
Article in Russian

For citation: Vorobeva A.A., Pozvolenko V.A., Korobitsyna A.S., Sharafiev A.A. Cross-domain web author identification. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 3, pp. 447–456 (in Russian). doi: 10.17586/2226-1494-2018-18-3-447-456

Abstract
The paper is devoted to the cross-domain web author attribution (identification), where user's messages are obtained from several sources (web-sites). We focused on the problem of one web-site user identification by his messages from another web-site. We found that there is a stylistic difference between the texts of messages created by one user on different web-sites. The possibility of a single feature space forming for texts received from various sources was determined providing sufficient accuracy of linguistic identification. Two subtasks were studied: 1) mixed sources – training and test datasets include messages from mixed sources (web-sites); 2) separated sources – the text messages sources of the training and test datasets do not intersect; training dataset includes texts from one source, test dataset includes texts from another. The experiment results showed that identification accuracy in mixed sources task is 0.82. The accuracy in separated sources task is 0.74. It is concluded that there is a stylistic difference between texts created by one user, but on the various web-sites. But at the same time, it is possible to form a single feature space for text messages received from various web-sites, ensuring sufficient identification accuracy.

Keywords
web users' identification, forensic linguistics, linguistic identification, cross-domain identification, author attribution

Введение

Быстрое развитие информационных технологий привело к повсеместному использованию средств онлайн-коммуникации, таких как электронная почта, мессенджеры, социальные сети, блог-платформы и форумы. При этом современные исследователи отмечают рост числа преступлений, совершаемых с помощью компьютеров и сети Интернет. В качестве примеров подобного злонамеренного использования можно привести распространение спама, домогательства и преследования (так называемый «кибербуллинг»), мошенничество и вымогательство, корпоративный шпионаж. Террористские и экстремистские организации могут использовать современные информационные технологии для внутренней коммуникации, а также для распространения различных публичных сообщений, связанных с деятельностью подобных организаций. Интернет-технологии могут использоваться для манипуляции общественным мнением, например, путем размещения многочисленных заказных публикаций, оформленных как независимые мнения частных лиц. Данное явление получило название «астротурфинг». Подобные методы используются как в политической, так и в коммерческой сфере [1, 2].

При этом, как правило, онлайн-сообщения являются анонимными. Для большинства средств онлайн-коммуникации не является необходимым предоставление своих персональных данных – имени, возраста, пола, адреса, и пр. [3]. В случае злонамеренного использования пользователь также будет дополнительно пытаться скрыть свою личность, чтобы избежать наказания – например, использовать чужие или поддельные данные, маскировать IP-адрес отправителя с помощью специальных proxy-серверов.

Зачастую единственными данными, позволяющими идентифицировать автора-злоумышленника, являются непосредственно отправленные сообщения [4].

Основной сложностью при исследовании онлайн-сообщений является малый объем данных. Интернет-сообщения часто ограничены по длине, так, например, длина сообщений Twitter ограничена 140 символами. Помимо этого, зачастую сами пользователи предпочитают использовать в онлайн-коммуникации тексты меньшей длины. Также стилистика сообщений может существенно меняться в зависимости от используемого метода коммуникации. Данное явление более заметно на коротких сообщениях [5].

Особенности задачи межсайтовой лингвистической идентификации

История лингвистической идентификации и предшествующих ей методов атрибуции автора насчитывает уже несколько веков. Современный этап развития начался в конце XX века, когда была поставлена задача идентификации интернет-пользователя, являющегося автором электронного сообщения. Данная задача сегодня классически решается с помощью различных методов машинного обучения, исследования показывают достаточно высокие показатели точности идентификации.

Сегодня среднестатистический российский пользователь сети Интернет имеет несколько (2–3 и более) аккаунтов в различных социальных медиа, чаще всего в социальных сетях vk.com, Facebook, Twitter, сайтах источниках видеоконтента (Youtube). Однако практически все предыдущие исследования использовали только один источник текстов для идентификации (например, только одну социальную сеть или блог-платформу). Научная новизна данного исследования заключается в использовании нескольких источников текстовой информации для идентификации автора.

Задача межсайтовой лингвистической идентификации, поставленная в данном исследовании, является достаточно новой. Всего несколько работ посвящено вопросам межсайтовой или междоменной (cross-domain) идентификации [6, 7]. Максимальная достоверность (accuracy), полученная авторами [7] в экспериментах для 10 пользователей составила около 0,4.

В данной работе рассматривается проблема лингвистической идентификации пользователя сети Интернет по текстовым сообщениям, полученным из нескольких источников (сайтов, средств онлайн-коммуникации).

Сбор текстовой информации из нескольких источников имеет следующие преимущества:

1. может существенно повлиять на ход расследования преступления в сети Интернет. Так, например, у пользователя может существовать аккаунт в социальной сети, содержащий идентифицирующую его информацию, но не причастный ни к какой криминальной активности, и анонимный аккаунт в другой социальной сети, при помощи которого было совершено преступление. В этом случает при помощи лингвистической идентификации можно связать данные аккаунты и доказать, что оба этих аккаунта принадлежат одному лицу, что позволяет предпринимать соответствующие меры;
2. позволяет существенно увеличить объем текстовой информации для одного пользователя, тем самым решая проблему малой длины сообщений и улучшая точность идентификации.

Однако при сборе текстовой информации из различных источников возникают затруднения:

1. стилистика интернет-сообщений может меняться в зависимости от используемого средства коммуникации, что существенно затрудняет построение единого лингвистического профиля пользователя;
2. может возникнуть проблема несбалансированности данных, когда имеется значительно большее количество сообщений одного пользователя из какого-либо источника.

Выдвинем две гипотезы данного исследования.

Гипотеза 1. Существует стилистическая разница между текстами сообщений, созданными одним пользователем, но с использованием различных средств коммуникации.

Гипотеза 2. Существует возможность сформировать единое признаковое пространство для сообщений, полученных из различных источников, обеспечивающее достаточную точность лингвистической идентификации.

Задача межсайтовой лингвистической идентификации

Примером задачи межсайтовой идентификации может служить ситуация, когда имеется некоторое нелегальное сообщение, размещенное в сети Twitter, и необходимо установить, кто из авторов – пользователей сети vk.com является автором данного сообщения.

Математическая постановка задачи межсайтовой лингвистической идентификации интернет-пользователей формулируется следующим образом.

Имеется множество источников электронных сообщений $D=\{D_1,...,D_q\}$, некоторое конечное множество пользователей (авторов-кандидатов) $U=\{u_1,...,u_m\}$, множество сообщений $T=\{T_{d1},...,T_{dq}\}$, полученных из источников D, где q – количество источников электронных текстовых сообщений, m – количество пользователей. Имеется некоторое сообщение неизвестного пользователя t_{ux}, полученное из источника D_1. Необходимо установить, кто из множества U является автором сообщения t_{ux}.

Решение данной задачи сводится к решению задачи классификации и состоит в том, что необходимо построить алгоритм $a: t_{ux} \to U$, способный определить, кому из пользователей U принадлежит t_{ux}.

Каждый пользователь представляется как набор его сообщений: $u_a = T_a = \{t_{a1},...,t_{al}\}$, где T_a – множество сообщений u_a, l – количество сообщений u_a, а каждое сообщение – как $t_{aj}=(f_{aj1},...,f_{ajn})$, где $t_{aj} \in T_a$, f – некоторый признак, n – количество признаков. Таким образом, каждый пользователь есть множество векторных представлений его сообщений $u_a = T_a = <(f_{a11},\ldots,f_{a1n}),\ldots,(f_{al1},\ldots,f_{aln})>$.

$T_{d1}=\{t_1,...,t_y\}, T_{d1} \in T$ – множество сообщений, полученных из источника D_1, $T_{dn}=\{t_1,...,t_z\}, T_{dn} \in T$ – множество сообщений, полученных из источника D_n, где y – количество сообщений из источника D_1, z – количество сообщений из источника D_n.

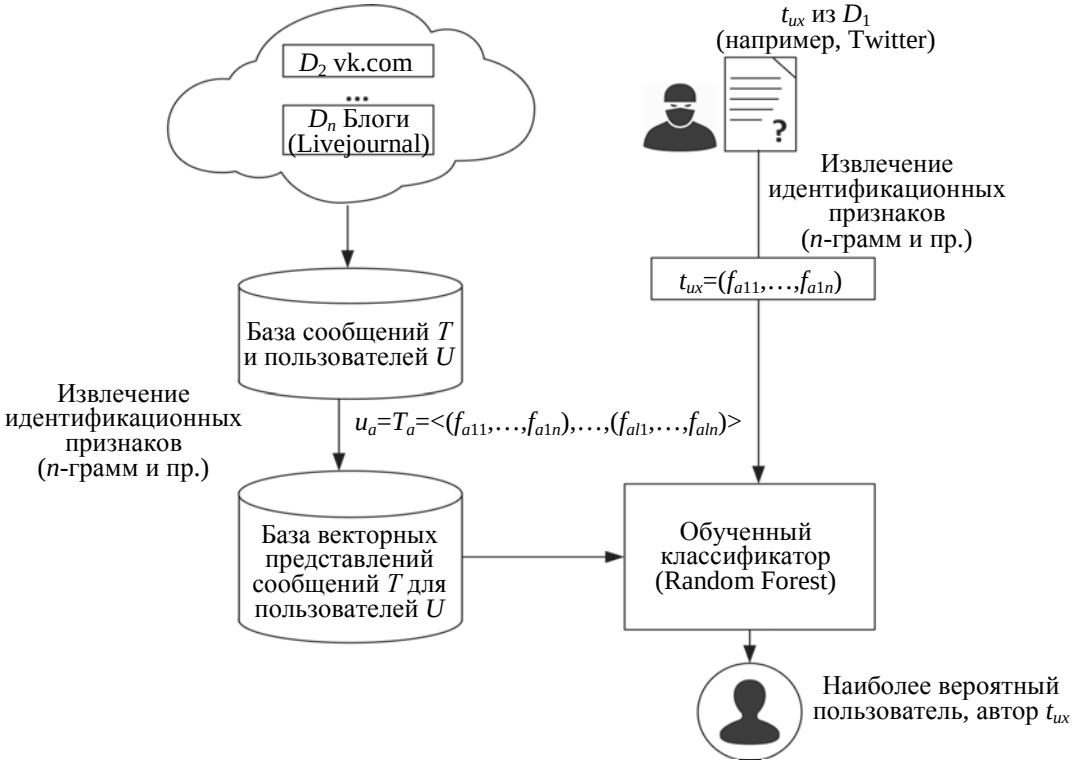

Рис. 1. Общая схема межсайтовой лингвистической идентификации для задачи 1

Исходя из состава исходных данных, можно выделить две самостоятельные задачи межсайтовой идентификации.

Задача 1. Имеются сообщения пользователей из нескольких источников D_1 и D_2. Обучение производится на основе набора данных, сформированном из текстов $T = T_{d1} \bigcup T_{d2}$.

Задача 2. Отсутствуют сообщения источника D_1, из которого получено сообщение t_{ux}. Обучение производится полностью по текстам T_{d2}. Схема данной задачи межсайтовой лингвистической идентификации приведена на рис. 1.

Отличие этих двух задач от классической задачи идентификации пользователя-автора состоит в том, что сообщения пользователей, созданные с использованием различных средств электронной коммуникации (сайтов), могут существенно отличаться по стилю.

Признаковое пространство

В качестве признаков могут быть использованы различные характерные особенности языка сообщений пользователя. Важной их особенностью является то, что они должны быть внесены автором подсознательно [8]. Сюда может относиться используемая лексика, пунктуация, грамматические ошибки, форматирование и компоновка текста.

В работе [9] показано, что совместное использование характеристик различных типов позволяет повысить точность идентификации. Также существуют более сложные методы, базирующиеся на извлечении синтаксической и семантической информации.

В случае электронных сообщений сюда добавляются некоторые особенности электронной коммуникации (время активности пользователя, используемые изображения, особенности формирования подписи в email-сообщениях) [10].

Важно отметить также различные специфические наборы признаков в зависимости от источника электронных сообщений. Так, в работах, посвященных сверхкоротким сообщениям, упоминаются различные средства выражения эмоций (например, «эмотиконы»), а в работах, использующих Twitter в качестве источника сообщений, – количество и позиция хэштегов. В случаях, когда рассматривается электронная почта, большое внимание уделяется приветствиям и особенностям форматирования (например, использованию курсива).

В целом для идентификации пользователя по электронным сообщениям сегодня наиболее часто используемыми признаками являются: частоты *n*-граммы слов, частоты *n*-граммы символов, частоты служебных слов, наиболее часто используемые слова, частоты знаков пунктуации, частоты прописных и строчных символов, частоты цифр и т.д. Подробный обзор различных зарубежных исследований в области лингвистической идентификации можно найти в работах [4, 11].

Существует два основных подхода к представлению текста в задачах обработки естественного языка и извлечения информации. Первый – это так называемая модель «bag-of-words», где текст представляется как набор слов без учета их порядка, связей и грамматики, близость двух текстов оценивается по количеству совпадающих слов [12]. Второй – это представление текста в виде набора *n*-грамм слов или символов [13, 14].

В данной работе рассматривается второй подход, где в качестве признаков используются *n*-граммы слов и *n*-граммы символов различной длины, так как в предыдущих исследованиях доказано, что именно *n*-граммы дают стабильные результаты как на коротких, так и на длинных текстах.

Установлено, что на коротких сообщениях эти *n*-граммы дают более высокую точность, так как длина электронных сообщений обычно недостаточно велика для подсчета *n*-грамм высших порядков.

Корпус текстов

Для проведения исследования был составлен корпус текстов, состоящий из сообщений, полученных из ВКонтакте, Twitter и Livejournal. В общей сложности в корпус вошли 15359 сообщений трех русскоязычных авторов, причем по каждому автору тексты собирались из двух и более источников. Корпус был размечен, в него была добавлена информация для обозначения источника. Из текстов удалены гиперссылки, специфические для отдельных социальных медиа токены (хэштеги) и не текстовый контент (изображения, видеозаписи). Тексты сообщений корпуса были проанализированы, и из них были извлечены униграммы, биграммы и триграммы слов, а также триграммы, 4-граммы, 5-граммы символов.

Распределение сообщений по длинам в символах приведено на рис. 2. Корпус сообщений является сильно несбалансированным: доля текстов первого автора сильно превышает долю текстов второго и третьего. Сбалансированный корпус был сформирован путем случайного удаления некоторого количества примеров мажоритарного класса (undersampling). Распределение количества сообщений и количества символов в сообщениях по авторам приведены на рис. 3. Из приведенных гистограмм (рис. 3) видно, что продуктивность авторов сильно отличается. Это позволяет исследовать вопросы межсайтовой идентификации как на сбалансированных, так и на несбалансированных данных.

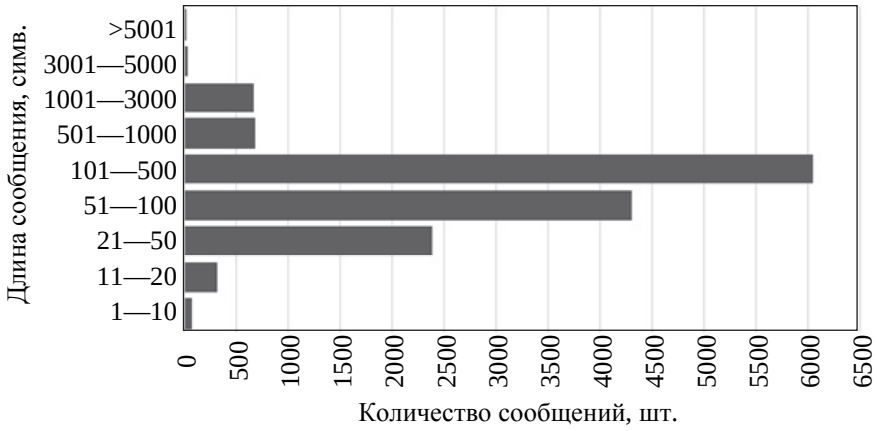

Рис. 2. Распределение длин сообщений

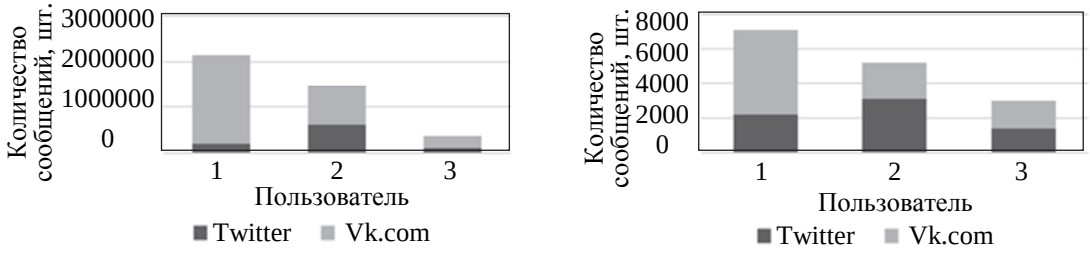

Рис. 3. Распределение количества символов и количества сообщений по пользователям

Эксперименты и оценка результатов

В качестве классификатора для решения задачи межсайтовой идентификации был выбран алгоритм Random Forest (RF), так как во многих предыдущих исследованиях он показывал наивысшую точность классификации [10, 15–17]. Для повышения качества экспериментов применялся метод кросс-валидации.

Сравнительный анализ точности межсайтовой идентификации пользователей проводился на корпусе сообщений, описанном выше, конкатенация или обрезка текстов не производились. Все эксперименты выполнялись на сбалансированных данных, за исключением эксперимента с прицельным исследованием несбалансированных данных.

Все множество сообщений T разделяется на обучающую (T_{train}) и тестовую (T_{test}) выборку. На основе обучающей выборки производится обучение классификатора, далее по тестовой выборке оценивается качество построенной модели. Размер тестовой выборки составил 20%.

Для каждого эксперимента фиксировались показатели достоверности (A, accuracy), точности (P, precision) и полноты (R, recall).

Под достоверностью идентификации (A) понимается отношение количества правильно идентифицированных пользователей для сообщений ($TP+TN$) к общему числу сообщений тестовой выборки – $TP+TN+FP+FN$, где TP – количество истинно положительных результатов; TN – количество истинно отрицательных результатов; FP – количество ложно положительных результатов; FN – количество ложно отрицательных результатов:

$$A = \frac{TP+TN}{TP+TN+FP+FN}.$$

Достоверность не всегда является лучшим показателем качества классификации. Существует такое явление, как «парадокс достоверности» (accuracy paradox) [18], когда модель с меньшей достоверностью дает лучшие результаты на тестовых данных. Так, в случае сильной несбалансированности исходных данных классификатор может верно предсказывать значения для мажоритарного класса (с наибольшим количеством экземпляров), вследствие чего иметь высокую достоверность, но плохо работать для миноритарных классов (с малым количеством экземпляров), и измерение достоверности не сможет показать этот недостаток.

Для более точной оценки качества работы классификатора используются такие метрики, как точность и полнота, а также метрики на их основе.

Точность идентификации (P) – это доля сообщений, действительно принадлежащих пользователю, относительно всех сообщений, которые были отнесены к данному пользователю:

$$P = \frac{TP}{TP + FP}.$$

Полнота (R) – доля отнесенных к пользователю сообщений относительно всех сообщений данного пользователя в тестовой выборке:

$$R = \frac{TP}{TP + FN}.$$

Зависимость качества идентификации от используемого признакового пространства

Проводились эксперименты по оценке качества идентификации в зависимости от используемого признакового пространства и различных алгоритмов классификации: только униграммы, комбинация униграмм и биграмм, комбинация биграмм и триграмм, так как в ходе теоретического исследования было установлено, что комбинация униграмм, биграмм, триграмм позволяет повысить качество классификации.

Также были рассчитаны частоты n-грамм символов для n = 3, 4, 5. В экспериментах оценивалось качество идентификации по триграммам, комбинации триграмм и 4-грамм, комбинации триграмм, 4-грамм и 5-грамм. Помимо этого, проводилась оценка совместного использования n-грамм по символам и словам (для слов n = 1 и 2; 1, 2 и 3, для символов n = 3 и 4; 3, 4 и 5).

Номер группы признаков	Признаки
1	Униграммы слов
2	Униграммы и биграммы слов
3	Униграммы, биграммы и триграммы слов
4	Триграммы символов
5	Триграммы и 4-граммы символов
6	Триграммы, 4-граммы, 5-граммы символов
7	Униграммы и биграммы слов и триграммы и 4-граммы символов
8	Униграммы, биграммы и триграммы слов и триграммы, 4-граммы, 5-граммы символов

Таблица 1. Группы признаков

Полученные результаты экспериментов по группам идентификационных признаков приведены на рис. 4. Нумерация групп и перечень входящих в них признаков приведены в табл. 1.

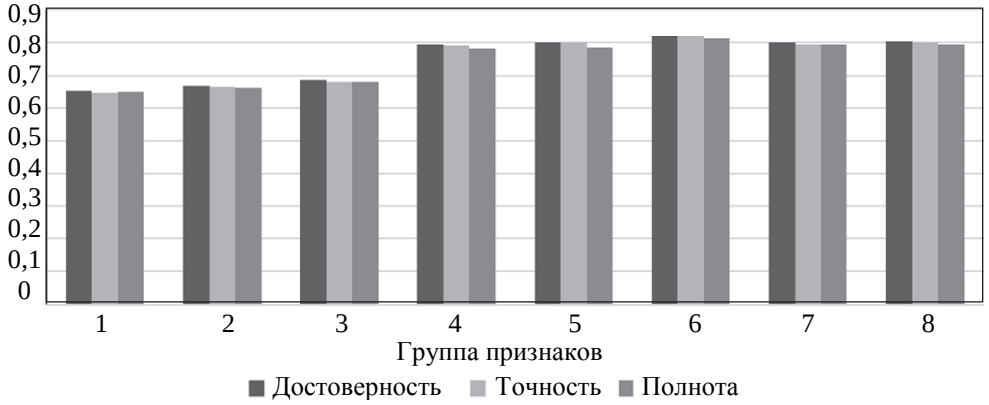

Рис. 4. Зависимость качества идентификации от используемого признакового пространства

В эксперименте результаты использования n-грамм по символам близки к результатам совместного использования n-грамм по символам и словам. При этом n-граммы исключительно по словам дают существенно более низкую точность, что объясняется малой длиной текстов в корпусе. Тем не менее, по результатам эксперимента можно сделать выводы, что совместное использование n-грамм по символам и словам дает приемлемые результаты для задачи межсайтовой лингвистической идентификации.

Стоит отметить, что совместное использование 1,2,3-грамм по словам и 3,4,5-грамм по символам дало чуть более высокую точность по сравнению с 1,2-граммами по словам и 3,4-граммами по символам. Однако во втором варианте получается признаковое пространство меньшей размерности, что ведет к

снижению времени вычисления. Для дальнейших экспериментов был отобран именно этот вариант признакового пространства.

Оценка качества классификации в эксперименте на несбалансированных данных

Эксперименты проводились с использованием алгоритма RF. Число деревьев в ансамбле составляет 100, данное значение было подобрано в ходе предварительных экспериментов. Результаты эксперимента представлены на рис. 5.

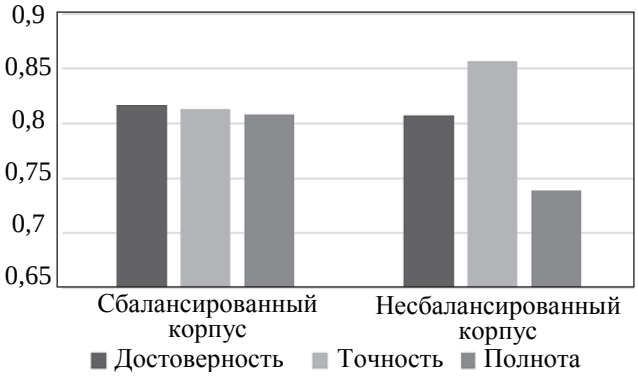

Рис. 5. Сравнение качества идентификации для сбалансированных и несбалансированных данных

Из результатов эксперимента можно увидеть упоминавшийся ранее «парадокс достоверности»: несмотря на то, что достоверность для обоих случаев является сопоставимой, однако точность идентификации во втором случае существенно ниже, так как мы принимаем слишком много положительных решений об отнесении текста к первому автора.

Эксперимент с тестовой и обучающей выборкой, сформированной из текстов сообщений различных источников

Были проведены исследования экстремального случая, когда обучающая и выборка тестовая сформированы из полностью непересекающихся источников информации (т.е. обучение проводится на данных одного источника, а тестирование – на данных, полученных из иного источника). Схема формирования обучающей и тестовой выборки приведена на рис. 6.

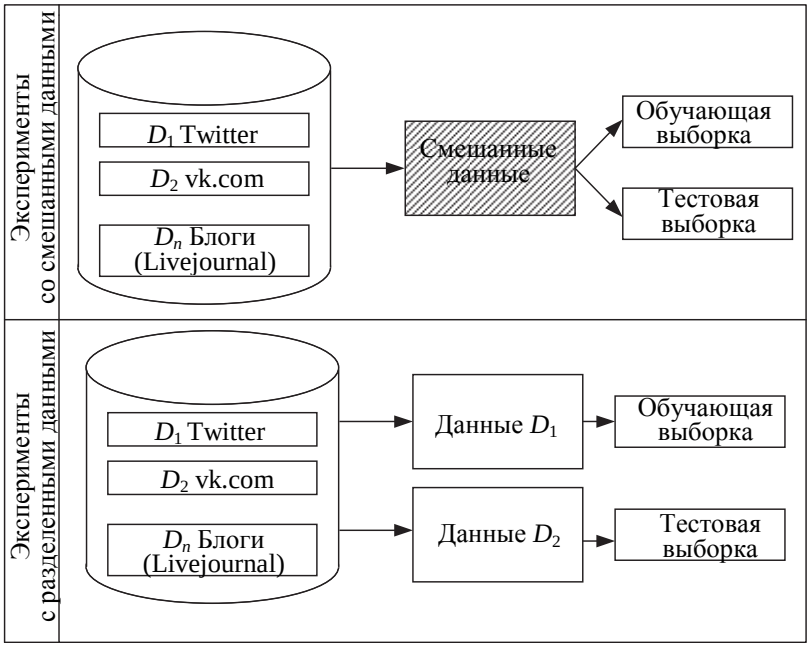

Рис. 6. Схема формирования обучающей и тестовой выборки

Результаты эксперимента приведены на рис. 7. В первую очередь стоит отметить, что качество идентификации при разделении источников информации для тестовой и обучающей выборки заметно снизилось. Это подтверждает гипотезу, породившую данное исследование: существует стилистическая разница между текстами сообщений, написанными для различных социальных медиа.

Однако при этом точность остается достаточно высокой (около 0,7) и не снижается до критических значений, что подтверждает вторую гипотезу исследования: для различных источников информации можно составить единое признаковое пространство, обеспечивающее достаточную точность лингвистической идентификации.

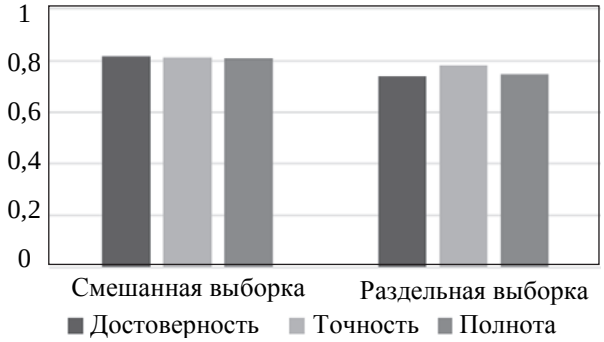

Рис. 7. Сравнение качества идентификации для смешанных и разделенных источников информации

Дальнейшие исследования

Дальнейшие исследования по проблеме межсайтовой лингвистической идентификации можно вести в нескольких направлениях. В первую очередь остается открытым вопрос о признаковом пространстве, обладающем максимальной разрешающей способностью для данной задачи, так как был исследован только один способ построения признакового пространства и его вариации (*n*-граммы слов и символов).

Интересными для исследования являются вопросы минимального размера обучающей выборки, минимальной длины текста, максимального количества пользователей, несбалансированности данных.

Также необходимо исследовать перспективы использования различных подходов к классификации и формированию модели пользователя – так называемый «профильный подход» (все тексты объединяются в один), обучение на примерах (каждый текст является обучающим примером) и их комбинации.

Часто возникает вопрос об изменении стилистики автора с течением времени, со сменой собеседника и круга общения в целом, сменой настроения и привычек. Множество исследований доказывают, что подсознательно внесенные в текст особенности остаются относительно неизменными в указанных ситуациях. Также в работе [19] было установлено, что в литературных произведениях, относящихся к разным периодам творчества писателя, доля служебных слов остается постоянной. Интересным является более глубокое исследование вопроса изменчивости стиля автора в различных средствах коммуникации и его изменений с течением времени.

Заключение

В работе рассмотрена важная для информационной безопасности проблема – идентификация интернет-пользователей по сообщениям, собранным из различных источников (межсайтовая лингвистическая идентификация). Данная проблема является достаточно новой и сложной для решения.

Для экспериментов был создан уникальный корпус русскоязычных текстов, собранных из открытых источников, включающий электронные текстовые сообщения пользователей из различных социальных медиа. Были проведены исследования и эксперименты с использованием данного корпуса. В качестве классификатора использовался алгоритм Random Forest. Идентификация производилась по признаковому пространству, которое включало в себя комбинации *n*-грамм слов и символов. Было установлено, что существует возможность построения общего признакового пространства для межсайтовой лингвистической идентификации.

В работе рассмотрены две самостоятельные задачи межсайтовой идентификации и получены следующие результаты экспериментов.

Задача 1. Имеются сообщения пользователей из нескольких источников D_1 и D_2. Обучение производится на основе набора данных, сформированном из текстов $T = T_{d1} \bigcup T_{d2}$. Достоверность идентификации – 0,82.

Задача 2. Отсутствуют сообщения источника D_1, из которого получено сообщение t_{ux}. Обучение производится полностью по текстам T_{d2}. Достоверность идентификации – 0,74.

Тем самым удалось подтвердить экспериментально две выдвинутые гипотезы исследования:

1. существует стилистическая разница между текстами сообщений, созданными одним пользователем, но с использованием различных средств коммуникации;
2. существует возможность сформировать единое признаковое пространство для сообщений, полученных из различных источников, обеспечивающее достаточную точность лингвистической идентификации.

Интересным представляется более глубокое исследование вопросов поиска единого признакового пространства, выявления информативных признаков, общих для разных средств коммуникации, минимального размера обучающей выборки, минимальной длины текста, максимального количества пользователей, несбалансированности данных, степени изменчивости стиля автора в различных средствах коммуникации и с течением времени.

Литература

1. Chen C., Wu K., Srinivasan V., Zhang X. Battling the internet water army: detection of hidden paid posters // Proc. IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM). Niagara Falls, Canada, 2013. P. 116–120. doi: 10.1145/2492517.2492637
2. Лебедев И.С., Борисов Ю.Б. Анализ текстовых сообщений в системах мониторинга информационной безопасности // Информационно-управляющие системы. 2011. № 2. С. 37–43.
3. Катаева В.А., Пантюхин И.С., Юрин И.В. Метод оценки степени связанности профилей пользователей социальной сети на основе открытых данных // Открытое образование. 2017. Т. 21. № 6. С. 14–22. doi: 10.21686/1818-4243-2017-6-14-22
4. Воробьева А.А. Отбор информативных признаков для идентификации интернет-пользователей по коротким электронным сообщениям // Научно-технический вестник информационных технологий, механики и оптики. 2017. Т. 17. № 1. С. 117–128. doi: 10.17586/2226-1494-2017-17-1-117-128
5. Сидорова М.Ю. Интернет-лингвистика: русский язык. Межличностное общение. М.: 1989.ру, 2006. 193 с.
6. Schwartz M.B. An Examination of Cross-Domain Authorship Attribution Techniques. CUNY Academic Works. 2016. 32 p.
7. Overdorf R., Greenstadt R. Blogs, twitter feeds, and reddit comments: cross-domain authorship attribution // Proceedings on Privacy Enhancing Technologies. 2016. N 3. P. 155–171.
8. Воробьева А.А. Анализ возможности применения различных лингвистических характеристик для идентификации автора анонимных коротких сообщений в глобальной сети Интернет // Информация и космос. 2014. № 1. С. 42–46.
9. 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 for Information Science and Technology. 2006. V. 57. N 3. P. 378–393. doi: 10.1002/asi.20316
10. Воробьева А.А. Методика идентификации интернет-пользователя на основе стилистических и лингвистических характеристик коротких электронных сообщений // Информация и космос. 2017. № 1. С. 127–130.
11. Stamatatos E.A survey of modern authorship attribution methods // Journal of the American Society for information Science and Technology. 2009. V. 60. N 3. P. 538–556. doi: 10.1002/asi.21001
12. Нугуманова А.Б., Бессмертный И.А., Пецина П., Байбурин Е.М. Обогащение модели Bag-of-Words семантическими связями для повышения качества классификации текстов предметной области // Программные продукты и системы. 2016. № 2. С. 89–99. doi: 10.15827/0236-235X.114.089-099
13. Houvardas J., Stamatatos E. N-gram feature selection for authorship identification // Lecture Notes in Computer Science. 2006. V. 4183. P. 77–86.
14. Gomez-Adorno H. et al. Document embeddings learned on various types of *n*-grams for cross-topic authorship attribution // Computing. 2018. P. 1–16. doi: 10.1007/s00607-018-0587-8
15. Maitra P., Ghosh S., Das D. Authorship verification: an approach based on random forest // Proc. 6th Conference and Labs of the Evaluation Forum (CLEF 2015). Toulouse, France, 2015.
16. Pacheco M.L., Fernandes K., Porco A. Random forest with increased generalization: a universal background approach for authorship verification // Proc. Conference and Labs of the Evaluation Forum (CLEF 2015). Toulouse, France, 2015.
17. Vorobeva A.A. Influence of features discretization on accuracy of random forest classifier for web user identification // Proc.

References

1. Chen C., Wu K., Srinivasan V., Zhang X. Battling the internet water army: detection of hidden paid posters. *Proc. IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining, ASONAM*. Niagara Falls, Canada, 2013, pp. 116–120. doi: 10.1145/2492517.2492637
2. Lebedev I.S., Borisov Y.B. Formalization models of natural-language messages in information security monitoring systems of open computer networks. *Information and Control Systems*, 2011, no. 2, pp. 37–43. (in Russian)
3. Kataeva V.A., Pantyuhin I.S., Yurin I.V. Estimation method of the cohesion degree for the users' profiles of social network based on open data. *Open Education*, 2017, vol. 21, no. 6, pp. 14–22. (in Russian) doi: 10.21686/1818-4243-2017-6-14-22
4. Vorobeva A.A. Dynamic feature selection for web user identification on linguistic and stylistic features of online texts. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2017, vol. 17, no. 1, pp. 117–128. (in Russian) doi: 10.17586/2226-1494-2017-17-1-117-128
5. Sidorova M.Yu. *Internet Linguistics: Russian Language. Interpersonal Communication*. Moscow, 1989.ru Publ., 2006, 193 p. (in Russian)
6. Schwartz M.B. *An Examination of Cross-Domain Authorship Attribution Techniques*. CUNY Academic Works, 2016, 32 p.
7. Overdorf R., Greenstadt R. Blogs, twitter feeds, and reddit comments: cross-domain authorship attribution. *Proceedings on Privacy Enhancing Technologies*, 2016, no. 3, pp. 155–171.
8. Vorobeva A.A. Analiz vozmozhnosti primeneniya razlichnykh lingvisticheskikh kharakteristik dlya identifikatsii avtora anonimnykh korotkikh soobshchenii v global'noi seti Internet. *Informatsia i Kosmos*, 2014, no. 1, pp. 42–46. (in Russian)
9. 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 for Information Science and Technology*, 2006, vol. 57, no. 3, pp. 378–393. doi: 10.1002/asi.20316
10. Vorobeva A.A. Technique of web-user identification based on stylistic and linguistic features of online texts. *Informatsia i Kosmos*, 2017, no. 1, pp. 127–130. (in Russian)
11. Stamatatos E.A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 2009, vol. 60, no. 3, pp. 538–556. doi: 10.1002/asi.21001
12. Nugumanova A.B., Bessmertnyi I.A., Petsina P., Baiburin E.M. Semantic relations in text classification based on Bag-of-words model. *Software & Systems*, 2016, no. 2, pp. 89–99. (in Russian)
13. Houvardas J., Stamatatos E. N-gram feature selection for authorship identification. *Lecture Notes in Computer Science*, 2006, vol. 4183, pp. 77–86.
14. Gomez-Adorno H. et al. Document embeddings learned on various types of *n*-grams for cross-topic authorship attribution. *Computing*, 2018, pp. 1–16. doi: 10.1007/s00607-018-0587-8
15. Maitra P., Ghosh S., Das D. Authorship verification: an approach based on random forest. *Proc. 6th Conference and Labs of the Evaluation Forum, CLEF 2015*. Toulouse, France, 2015.
16. Pacheco M.L., Fernandes K., Porco A. Random forest with increased generalization: a universal background approach for authorship verification. *Proc. Conference and Labs of the Evaluation Forum, CLEF 2015*. Toulouse, France, 2015.
17. Vorobeva A.A. Influence of features discretization on accuracy of random forest classifier for web user identification. *Proc. 20th Conf. on Open Innovations*

20th Conf. on Open Innovations Association (FRUCT). St. Petersburg, Russia, 2017. P. 498–504. doi: 10.23919/FRUCT.2017.8071354
18. Brownlee J. Classification Accuracy is Not Enough: More Performance Measures You Can Use [Электронный ресурс]. 2014. URL: http://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/ (дата обращения 20.03.2018).
19. Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов / В кн. Фоменко А.Т. Новая хронология Греции. Т. 2. М.: МГУ, 1995.

Авторы

Воробьева Алиса Андреевна – кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 57191359167, ORCID ID: 0000-0001-6691-6167, Alice_w@mail.ru

Позволенко Виталий Александрович – инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, ORCID ID: 0000-0002-9370-399X, frozensculpture@gmail.com

Коробицына Анастасия Сергеевна – студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, ORCID ID: 0000-0002-6238-2919, korobitsyna.as@korobochka.org

Шарафиев Азамат Азатович – студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, ORCID ID: 0000-0002-6502-6733, whoisazamat@gmail.com

Association, FRUCT. St. Petersburg, Russia, 2017, pp. 498–504. doi: 10.23919/FRUCT.2017.8071354
18. Brownlee J. *Classification Accuracy is Not Enough: More Performance Measures You Can Use*. 2014. Available at: http://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/ (accesssed 20.03.2018).
19. Fomenko V.P., Fomenko T.G. Avtorskii invariant russkikh literaturnykh tekstov. In Fomenko A.T. *Novaya Khronologiya Gretsii*. Moscow, MSU Publ., 1995, vol. 2. (in Russian)

Authors

Alice A. Vorobeva – PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 57191359167, ORCID ID: 0000-0001-6691-6167, Alice_w@mail.ru

Vitaliy A. Pozvolenko – engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, ORCID ID: 0000-0002-9370-399X, frozensculpture@gmail.com

Anastasiya S. Korobitsyna – student, ITMO University, Saint Petersburg, 197101, Russian Federation, ORCID ID: 0000-0002-6238-2919, korobitsyna.as@korobochka.org

Azamat A. Sharafiev – student, ITMO University, Saint Petersburg, 197101, Russian Federation, ORCID ID: 0000-0002-6502-6733, whoisazamat@gmail.com