Идентификация авторства коротких текстов методами машинного обучения

Identification of authorship of short texts with machine learning techniques

Романов А. С. (alexx.romanov@gmail.com),
Мещеряков Р. В. (mrv@keva.tusur.ru)

ГОУ ВПО «Томский государственный университет систем управления и радиоэлектроники», Томск

В статье рассматривается проблема идентификации авторства коротких текстов. Описан процесс формирования модели автора и алгоритм разбора текста. Приведено описание и результаты экспериментов по идентификации авторства коротких электронных сообщений в случае двух возможных альтернатив с помощью разновидностей искусственных нейронных сетей и аппарата опорных векторов.

1. Введение

Задача идентификации авторства коротких текстов возникает чаще, чем задача определения авторства текстов больших объемов и является в настоящее время актуальной проблемой. Это связано, прежде всего, с широким распространением программ для обмена сообщениями в сети Интернет (интернет-мессенджеров), возросшей роли электронной почты в деловой переписке, высокой популярности интернет-форумов и блогов. Пользователи имеют возможность отправлять сообщения без регистрации и указания какой-либо информации о себе, а регистрация сама по себе зачастую носит чисто символический характер. То же самое касается интернет-мессенджеров и электронной почты — регистрационные данные не позволяют однозначно идентифицировать личность собеседника, адрес отправителя можно легко изменить.

Идентификации авторства коротких текстов посвящено сравнительно небольшое количество работ. Стоит отметить, что авторам не известны подобные работы отечественных исследователей для русского языка. Судить о точности тех или иных методов по результатам исследования для английского и др. языков не корректно в силу особенностей строя каждого языка. В частности главной особенностью русского языка в сравнении с английским, для которого представлено большинство результатов, является его флективность, а, следовательно, и более сложное словообразование.

Эксперименты на корпусе электронных писем (всего 253 письма 4 авторов) в работе [1] дали итоговую максимальную точность 82,4 % при идентификации на основе 184 характеристик уровней символов и слов и признаков электронного письма (позиции цитат, доли слов приветствия, прощания, подписи к общей длине письма, количество вложений). При этом исследователи утверждают, что минимально необходимый объем письма для определения авторства составляет 200 слов, а для обучения модели достаточно 20 таких писем.

В работе [2] исследовался метод опорных векторов (SVM) на примере корпуса немецких газет «Berliner Zeitung» (2652 статьи, средний размер каждой из которых 200–300 слов). 2121 статья использовалась для обучения, тестирование проводилось на оставшихся 531 статьях, с последующей заменой тестовой и обучающей частей в соотношении 4 к 1. Средняя точность классификации по 7 авторам на основе словоформ (всего около 120 000 признаков) составила 99,7 %, а на основе сочетания грамматических классов, их биграмм и распределения длин слов в тексте — 99,2 %. Эксперименты показали, что выбор того или иного ядра для классификатора не играет существенной роли. Также в этой работе авторы сравнивают SVM с нейронными сетями и деревьями решений — машина опорных векторов и перцептрон показали сравнимые результаты (100 % и 93,3 % соответственно), тогда как деревья решений с задачей успешно справиться не смогли (точность 22,7 %).

В работе [3] для идентификации автора электронных писем применялся метод k ближайших соседей, точность при этом в среднем составляла 80 %.

Эксперт в области криминалистической лингвистики Кэрол Часки в работе [4] для идентификации авторства коротких текстов из области крими-

налистики применяла линейный дискриминантный анализ и текстовые аномалии на всех лингвистических уровнях. Точность идентификации в зависимости от используемого типа ошибок, допущенных автором, в её работе колебался от 65 % до 92 %.

Аббаси, исследовавший авторство сообщений на английском и арабском языке с интернет-форума экстремисткой группы, утверждает, что метод опорных векторов справляется с этой задачей лучше, чем деревья принятия решений, искусственные нейронные сети и линейный дискриминантный анализ, которые в свою очередь превосходят по точности методы неконтролируемого обучения такие как метод главных компонент. Точность классификации с помощью SVM в его работе [5] по 5 авторам для английского языка составляет 97 %, для арабского языка — 94,83 %. Аналогичные исследования для онлайн сообщений на китайском языке проводились в работе [6]. Точность классификации с помощью SVM составила 88,33 %, с помощью нейронных сетей — 83,05 %.

Целью данной работы является проверка способности современных машинных методов обучения, таких как искусственные нейронные сети и машина опорных векторов, идентифицировать автора короткого электронного сообщения на русском языке.

2. Модель автора и текста

Проблему идентификации автора текста при ограниченном наборе альтернатив сформулируем следующем образом. Имеется множество текстов $T = \{t_1,...,t_k\}$ и множество авторов $A = \{a_1,...,a_l\}$. Для некоторого подмножества текстов $T' = \{t_1,...,t_m\} \subseteq T$ авторы известны, т. е. существует множество пар «текст-автор» $D = \{(t_i, a_j)\}_{i=1}^{m}$. Необходимо установить, кто из множества A является истинным автором остальных текстов (анонимных или спорных) $T'' = \{t_{m+1},...,t_k\} \subseteq T$.

В данной постановке задачу идентификации автора можно рассматривать как задачу классификации с несколькими классами [7]. В этом случае множество *A* составляет множество предопределенных классов и их меток, *D* — обучающие примеры, а множество T'' — классифицируемые объекты. Целью является построение классификатора, решающего данную задачу, т. е. нахождение некоторой целевой функции $F : T \times A \to [-1,1]$, относящей произвольный текст множества *T* к его истинному автору. Значения функции интерпретируется как степень принадлежности объекта классу: 1 соответствует полностью положительному решению, -1 — отрицательному.

Для решения задачи можно использовать любой из разработанных на данный момент алгоритмов классификации. IDEF0 модель процесса формирования модели автора показана на рис. 1.

Для определения отличий стилей авторов предлагается следующая последовательность действий:

1. Разбиение имеющегося множества текстов на две группы. Первая используется для обучения модели классификатора. Вторая — для проверки точности идентификации автора с помощью обученной модели.
2. Формирование модели текста путем выбора модели представления текстовой информации и выделения определенных информативных групп характеристик текста. Отличия в стилях авторов характеризуются главным образом употреблением и частотой встречаемости определенных признаков в тексте — вектором $(x_1, x_2,..., x_n)$.
3. Приведение значений признаков в единый диапазон с помощью операций нормирования и шкалирования.
4. Корректировка параметров классификатора, позволяющих обеспечить высокую разделяющую способность исследуемых авторов путем обучения классификатора на нормированных векторах признаков группы обучающих текстов и проверке точности обученного классификатора на векторах признаков тестовой группы текстов. Первоначальное обучение классификатора происходит с параметрами по умолчанию или при заданных параметрах.
5. Изменение перечня групп характеристик и/или признаков, составляющих группу, в случае если изменением параметров классификатора достичь требуемой точности не удается.

Итогом является обученный классификатор, веса связей которого настроены таким образом, чтобы он был способен разделить стили авторов, на текстах которых проводилось обучение, при подаче на его входы подобранного набора признаков.

Таким образом, конечная модель помимо информативности признаков текста, учитывающихся в статистических методах идентификации авторства, учитывает влияние общей способности классификатора к разделению данных и его точность.

В данной работе были выбраны два инструмента — искусственные нейронные сети (многослойные перцептрон и сети каскадной корреляции) и аппарат опорных векторов. Эксперименты отечественных и зарубежных исследователей показывают, что на сегодняшний день эти два инструмента, при должной настройке и выборе входных параме-

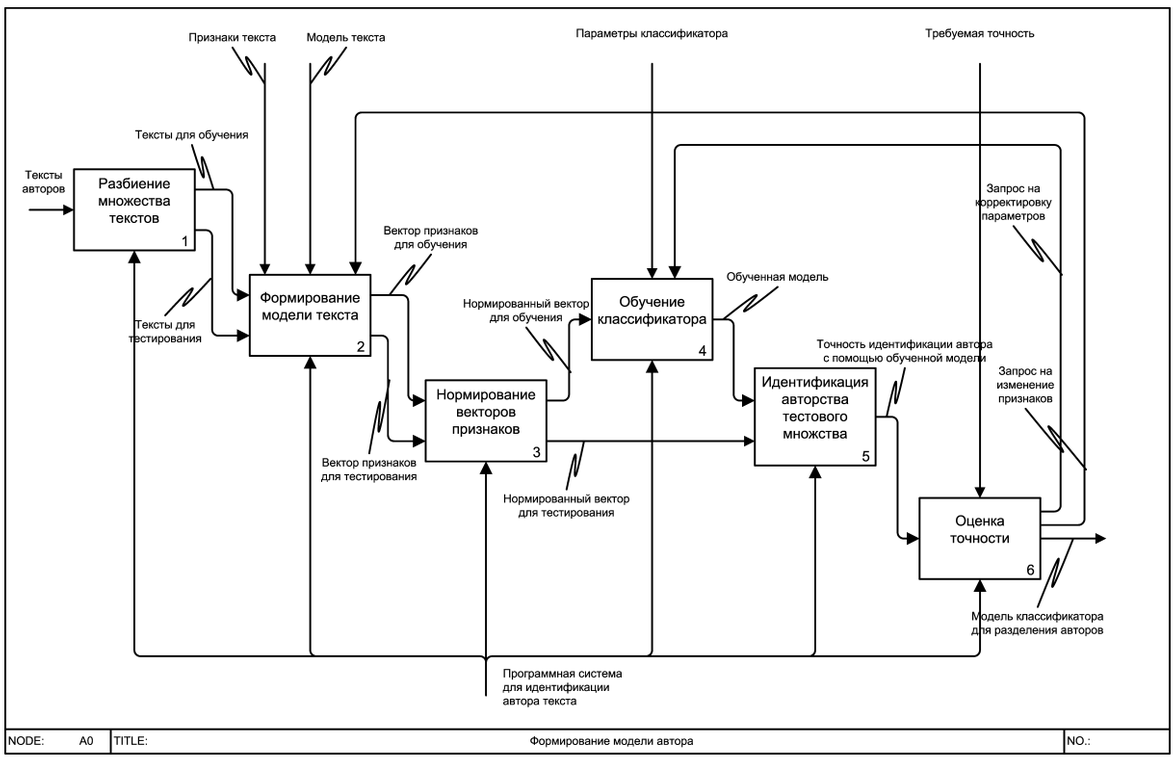

Рис. 1. IDEF0 модель процесса формирования модели автора

тров, являются лучшими в своем классе. Но каждому из них присущи свои достоинства и недостатки. Так, например, многослойный перцептрон (MLP) [8] долго обучается и не может работать с большим признаковым пространством. Временные издержки на подбор топологи сети и обучение можно сократить, применяя алгоритмы построения сетей с оптимальной топологией, к которым относятся сети каскадной корреляции (CCN) [9,10], однако точность классификации при этом может существенно снизиться. Метод опорных векторов [11, 12] лишен всех этих недостатков, однако слишком чувствителен к шумам во входных данных. Поэтому необходимо детально изучить возможность применения этих инструментов к задаче идентификации авторства, сравнить их производительность и эффективность.

В модели используется векторное представление текста, когда каждый текст представлен точкой в N-мерном пространстве. Элементами вектора могут быть характеристики уровней символов, слов, предложений, структурные признаки текста.

3. Алгоритм разбора текста

Особенностью коротких электронных сообщений является использование авторами эмотиконов («смайликов»), отсутствие пунктуации, намеренное искажение слов и т. д. Учет этой информации при определении авторства требует использования модифицированных алгоритмов разбора текста.

Специфика алгоритмов разбора текста состоит в том, что в процессе их работы происходит посимвольный или пословный анализ всего текста, в результате которого проверяется, удовлетворяет ли данная последовательность символов или слов определенной группе эвристических правил. Таким образом, для вынесения итогового решения, проверяется несколько промежуточных условий. Текущее состояние решения и очередной символ, поданный на вход алгоритма, определяют следующее состояние. Очевидно, что наилучшим техническим решением при реализации данной группы алгоритмов является использование конечных автоматов, несомненным преимуществом которых является возможность расширения функциональности алгоритма за счет добавления новых состояний.

Диаграмма состояний графа для определения границ предложений в коротких электронных сообщениях на рис. 2.

Началом предложения считается первый печатный символ текста. Концом предложения помимо последнего символа сообщения считается точка, вопросительный или восклицательный знак или их группа, а также любой эмотикон. Эмотиконы в большинстве случаев выражают законченность мысли и служат для придания написанным словам дополнительной эмоциональной окраски, тогда как в середине предложения употребляются редко. Также они используются в начале сообщения,

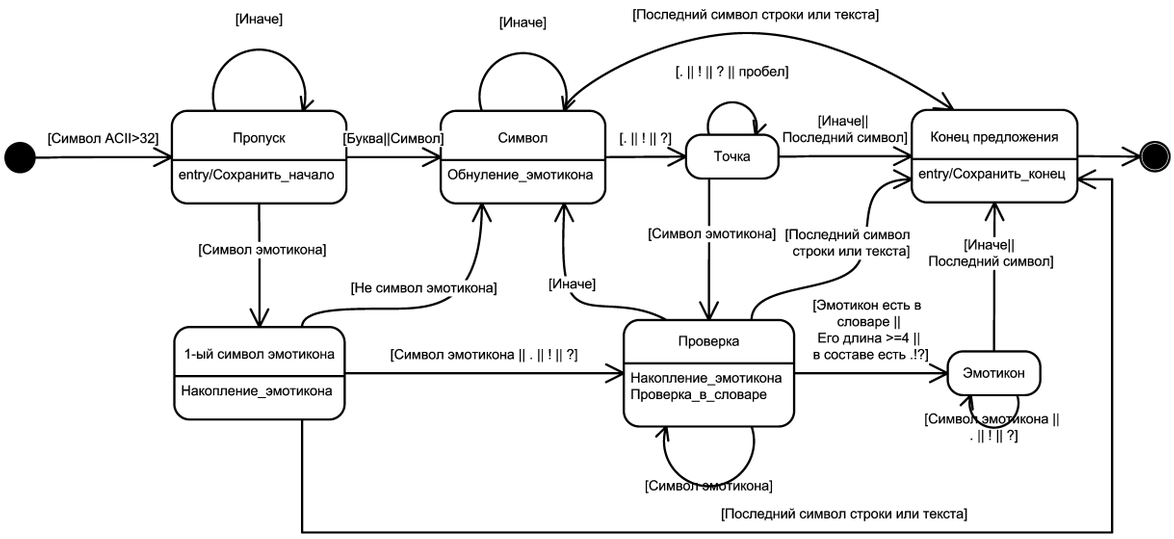

Рис. 2. Алгоритм определения границ предложений в коротких сообщениях

чтобы выразить эмоции по отношению, например, к предыдущей фразе собеседника — в этом случае алгоритм не выделяет эмотикон как отдельное предложение, а включает его в состав первого предложения сообщения.

Разделителями слов выступают все символы, не относящиеся к русскому и латинскому алфавиту, а также цифры. Один или несколько последовательно идущих символов «.», «!», «?» рассматриваются как отдельная группа символов конца предложения. В алгоритме также предусмотрен ряд проверок символа «-», позволяющих разделять случаи, когда он является переносом, частью составных слов, написанных через дефис и т. д.

4. Экспериментальная часть

В качества корпуса для исследования были взяты сообщения 10 авторов, собранные на Интернет-форуме http://forum.tomsk.ru. Информация о корпусе представлена в табл. 1 (длина текста в символах).

Все сообщения были разбиты на тексты длиной 100 символов (сообщения меньшего объема использовались целиком). Для обучения классификаторов использовалось по 50 таких текстов, тестирование проводилось на 25 сообщениях каждого автора. После этого обучающая и тестовая части менялись и опыты повторялись, пока каждый из текстов не был использован в тестовой части. Всего было проведено около 3000 таких экспериментов с различными комбинациями авторов и сообщений, что обеспечило необходимое покрытие сочетаний букв и слов русского языка.

В работе был рассмотрен самый простой случай, когда для спорного текста существует два предполагаемых автора. Данный частный случай был рассмотрен специально, чтобы раскрыть все потенциальные возможности метода опорных векторов, который по умолчанию является бинарным классификатором и поставить его в равные условия с двумя другими методами применительно к русскому языку.

Таблица 1. Корпус текстов для исследований

Автор	Количество текстов	Средняя длина текста	Минимальная длина текста	Максимальная длина текста
1	126	310,8	80	1368
2	102	438,2	83	2100
3	180	243,6	69	869
4	145	308,4	81	1665
5	141	238	72	928
6	146	261,1	83	1264
7	186	425,2	89	1649
8	116	292,4	68	1130
9	140	329	94	1957
10	157	539,8	95	3407
Все авторы	143,9	339,1	68	3407

Исследования проводились с использованием программной системы для идентификации автора письменной речи «Авторовед» [13].

Параметры обучения нейронных сетей архитектуры многослойный перцептрон были выбраны следующие:

- алгоритм обучения — обратного распространения ошибки;
- функция активации скрытых слоев — сигмоид;
- функция активации выходного слоя — сигмоид;
- скорость обучения — 0,7;
- момент — 0,0;
- количество скрытых слоев — 1;
- количество нейронов в скрытом слое — 10;
- максимальное количество эпох обучения — 20 000;
- допустимый уровень ошибки — 0,00 001.

Параметры обучения нейронных сетей каскадных корреляций были выбраны следующие:

- алгоритм обучения — быстрого распространения ошибки;
- функция активации скрытых слоев — сигмоид;
- функция активации выходного слоя — сигмоид;
- максимальное количество нейронов, которое можно добавить — 20;
- допустимый уровень ошибки — 0,00 001.

Параметры обучения метода опорных векторов были выбраны следующие:

- алгоритм обучения — метод последовательной оптимизации;
- ядро — линейное;
- параметр регуляризации $C = 1$;
- допустимый уровень ошибки — 0,00 001.

В качестве характеристик использовались признаки текста, показавшие наилучшие результаты на литературных текстах большего объема в ранней работе авторв [7]:

- частоты наиболее частых слов русского языка (согласно словарю Шарова [14]);
- частоты наиболее частых триграмм русского языка;
- частоты униграмм символов;

К признакам также добавлены частоты отдельных знаков препинания, составные знаки препинания (многоточие, «!?!» и т. п.) и эмотиконы.

Итоговый вектор признаков содержит 1024 компоненты. Данный вектор характеризует строй русского языка как лексико-зависимый, имеющий четко выраженную структуру, использующий базу наиболее употребимых слов и трехбуквенных сочетаний символов.

Результаты исследований представлены в табл. 2.

Таблица 2. Результаты экспериментов

	MLP	**CCN**	**SVM**
Точность классификации	0,62±0,01	0,70±0,17	0,69±0,12
Время обучения, с.	503,1	206,15	0,6

Из таблицы видно, что наименее точным в данном случае является многослойный перцептрон. Также этот метод классификации оказался наиболее затратным по времени. Применение сетей каскадных корреляций позволяет существенно уменьшить временные затраты на обучение модели и повысить точность идентификации. Дополнительные исследования по подбору архитектуры многослойного перцептрона, возможно, позволили бы повысить его точность до уровня сетей каскадных корреляций, однако это потребовало бы существенных временных затрат. Сравнимую с сетями каскадных корреляций точность показывает классификатор на основе машины опорных векторов. Небольшие потери в точности компенсируются высокой скоростью обучения моделей.

Точность идентификации выше 0,5 позволяет позволяет сделать вывод о принципиальной возможности идентификации авторства коротких текстов именно для русского языка. Несомненным положительным результатом можно признать учет лексической и синтаксической специфики русского языка, на основе которой удалось определить характеристики короткого сообщения, имеющие преимущественное значение для использования в методиках определения автора короткого сообщения и отличающиеся от других языков.

Дальнейшие исследования авторов по данной тематике будут связаны с применением ансамблей классификаторов для идентификации авторства и поиском статистически устойчивых характеристик на малых текстовых фрагментах.

Работа поддержана грантом ФСРМПНТ.

Литература

1. *Corney M., Anderson A., Mohay G., De Vel O.* Identifying the Authors of Suspect E-mail [Электронный ресурс] // Computers and Security, 2001. — Режим доступа: http://eprints.qut.edu.au/8021/1/CompSecurityPaper.pdf, свободный.
2. *Diederich J., Kindermann J., Leopold E., Paass G.* 2003. Authorship attribution with support vector machines. Appl. Intell. 19, pp. 109–123.
3. *Calix K., Connors M., Levy D.* Stylometry for E-mail Author Identification and Authentication // Proceedings of CSIS Research Day, Pace University, May 2008. — Режим доступа: http://csis.pace.edu/~ctappert/srd2008/c2.pdf, свободный.
4. *Chaski C. E.* Who's at the keyboard: Authorship attribution in digital evidence investigations // International Journal of Digital Evidence, vol. 4, no. 1, p. n/a, Electronic-only journal: http://www.ijde.org, accessed May 31, 2007, 2005.
5. *Abbasi, Chen H.* Applying authorship analysis to extremist-group web forum messages // IEEE Intelligent Systems, vol. 20, no. 6, pp. 67–75, 2005.
6. *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, vol. 57, no. 3, pp. 378–393, 2006.
7. *Романов А. С., Мещеряков Р. В.* Идентификация автора текста с помощью аппарата опорных векторов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). — М.: РГГУ, 2009. — С. 432–437.
8. *Хайкен С.* Нейронные сети: полный курс / Саймон Хайкен. — 2-е изд. — М. : Вильямс, 2006. — 1104 с.
9. *Fahlman S. E., Lebiere C.* The cascade-correlation learning architecture. Tech. Rep. CMU-CS-90–100, School of Computer Science, Carnegie Mellon University, August 1991.
10. *Hoefeld M., Fahlman S. E.* Learning with limited numerical precision using the cascade-correlation algorithm. Tech. Rep. CMU-CS-91-130, School of Computer Science, Carnegie Mellon University, 1991.
11. *Vapnik V. N.* Statistical Learning Theory // Wiley, New York, 1998. — 732 pages.
12. *Vapnik V. N.* The nature of statistical learning theory // Springer-Verlag, New York, 2000. — 332 pages.
13. *Романов А. С.* Структура программного комплекса для исследования подходов к идентификации авторства текстов // Доклады Томского государственного университета систем управления и радиоэлектроники. Томск: Изд-во ТУСУР, 2008. Ч.1. №2(18). С. 106–109.
14. *Шаров С. А.* Частотный словарь [Электронный ресурс]. — Режим доступа: http://www.artint.ru/projects/frqlist.asp, свободный. — Загл. с экрана.