УДК 004.912+004.02

СОСТОЯНИЕ И РАЗВИТИЕ ПРОБЛЕМЫ ИДЕНТИФИКАЦИИ АВТОРСТВА ТЕКСТОВ В BIG DATA

И.А. Труханович
Магистр технических наук
ilya.trukhanovich@gmail.com

А.И. Парамонов
Заведующий кафедрой информационных систем и технологий Института информационных технологий БГУИР, кандидат технических наук, доцент
a.paramonov@bsuir.by

И.А. Труханович
Выпускник Белорусского государственного университета информатики и радиоэлектроники. Область научных интересов связана с исследованием обработки естественных языков.

А.И. Парамонов
Область научных интересов связана с исследованием проблем обработки естественно-языковых текстов и построением интеллектуальных информационных систем.

Аннотация. Выполнен анализ состояния проблемы идентификации авторства текстов в *Big Data*. Обозначены основные проблемы в классификации текстовых данных при больших объемах. Приведено описание существующих подходов к идентификации авторства. Показаны некоторые пути развития в решении проблемы. Описаны способы модификации существующих подходов с учётом специфики *Big Data*. Приведен результат эксперимента с разными методами классификации текста на больших массивах данных.

Ключевые слова: обработка естественных языков, идентификации авторства, классификация текстов, большие данные

Введение. В современном мире ежедневно генерируются огромные объемы цифровых данных, источниками которых выступают как пользователи цифровых устройств, так и сами устройства. Значительную долю в этих объемах занимают текстовые документы, причем зачастую это неструктурированные массивы данных из разнородных источников. В связи с чем все также остро стоят вопросы эффективной обработки таких данных с целью поиска и систематизации информации. Одной из актуальных и важных проблем при обработке текстовых естественно-языковых данных остается разработка эффективных способов классификации и установления авторства в обширных и разнообразных текстовых коллекциях. В ряде прикладных областей, среди которых можно выделить фильтрацию контента, выявление плагиата и литературных источников, задача установления авторства имеет огромное значение [1-3]. Колоссально возрастающий объем поступающих для обработки текстовых данных приводит к тому, что возникают потенциальные проблемы с масштабируемостью и точностью классических алгоритмов классификации текстовых документов. Преодолеть проблемы, связанные с большими данными, и получить ценные сведения из огромного количества текстовых документов, возможно сегодня только при использовании целого комплекса мер, включающий технологии распределенных вычислений, новые методы извлечения

признаков, передовые алгоритмы машинного обучения, методы уменьшения размерности, активное обучение и комбинированные методы.

Идентификация авторства текстов больших объёмов. Подход к обработке больших текстовых данных, в том числе в идентификации авторства, является комплексным и может включать в себя сочетание нескольких направлений.

Важным этапом текстовой классификации является выбор и извлечение признаков. Быстрое и эффективное извлечение необходимых характеристик из больших массивов данных имеет ключевое значение при реализации современных прикладных систем. Такие методы, как например, мешок слов и *TF-IDF*, извлекают из текста синтаксическую и семантическую информацию. Подходы к выбору признаков, в основе которых лежат методы такие как критерий хи-квадрат, помогают определить наиболее информативные и дискриминационные признаки. Особенно это актуально в задачах обработки текста в контексте больших данных. Такой подход предполагает, что найденные признаки внесут существенный вклад в решение проблемы идентификации авторства.

Идентификация авторства текста может быть выполнена с помощью различных методов машинного обучения [4]. Традиционные алгоритмы (случайный лес, деревья решений, наивный байесовский классификатор) все еще являются хорошими вариантами. Тем не менее, более сложные методы продемонстрировали впечатляющий успех в распознавании сложных паттернов и тонкостей в текстовых данных. Например, модели глубокого обучения, которые включают в себя свёрточные нейронные сети (*CNN*), рекуррентные нейронные сети (*RNN*). Используя обучающие данные с метками, эти алгоритмы способны с достаточно высокой точностью определять авторов текстов.

Решение задачи идентификации авторства путем разметки огромного объема данных может быть весьма дорогостоящей и трудоемкой процедурой. Метод, известный как «активное обучение», позволяет используемой модели активно выбирать для разметки наиболее выделяющиеся экземпляры. «Активное обучение» минимизирует усилия по разметке при сохранении эффективности классификации путем итеративного запроса дополнительных меток в зависимости от неопределенности модели или противоречивых случаев. Такой подход позволяет максимально эффективно использовать помеченные данные и повышает производительность. Это делает его полезным при работе с большими объемами данных.

Влияние на производительность модели может оказать высокая размерность векторов признаков в больших данных. Чтобы минимизировать количество признаков, сохранив наиболее важную информацию, можно использовать методы снижения размерности, такие как анализ главных компонент (*PCA*) или латентное размещение Дирихле (*LDA*). Благодаря уменьшению размерности данных эти методы повышают эффективность вычислений и устраняют недостатки размерности, что позволяет ускорить определение авторства.

Кроме того, для повышения общей эффективности классификации используются комбинированные подходы, объединяющие множество моделей. Объединение результатов нескольких моделей – популярный пример использования различных стратегий. При работе с большими массивами данных комбинированные подходы очень полезны для повышения надежности и точности моделей идентификации авторства. Благодаря использованию преимуществ многих моделей эти методы повышают надежность и эффективность классификации.

Следует отметить, что масштабируемые решения очень важны при работе с большими данными. Необходимую основу для этого обеспечивают технологии распределенных вычислений в виде различных фреймворков. С помощью этих фреймворков можно обрабатывать масштабные текстовые массивы данных параллельно, за счет распределения нагрузки между несколькими компьютерами или узлами. Таким

образом можно эффективно справиться с огромным объемом текстовых данных и ускорить их обработку.

Можно предположить, что, сосредотачиваясь на отдельных рассмотренных подходах или комбинируя их, есть шансы добиваться различных успехов в обработке больших объемов текстовых данных, в частности в определении авторов текстов.

Вместе с тем необходимо брать во внимание и недостатки вышеперечисленных подходов. Главным из них является значительная ресурсоёмкость, особенно с учётом предполагаемых размеров обрабатываемых данных.

Модификация подходов. Рассматриваемые подходы могут быть модифицированы различными способами.

Прежде всего, в области выбора и извлечения признаков могут быть добавлены уточнения, специфичные для конкретной области [5]. Идентификация авторства может быть достигнута с большей точностью за счет повышения дискриминационной способности признаков и получения более определённой картины путем интеграции лексики, семантических деталей или тематического моделирования.

Для уменьшения размерности можно использовать автокодировщики. Снижая размерность и сохраняя важную информацию, эти методы могут создавать компактные представления многомерных векторов признаков, что в конечном итоге повысит эффективность вычислений и производительность модели.

Вместе с тем для расширения комбинированных методов можно добавлять новые группы, как со старыми, так и с новыми составляющими.

Наконец, достижения в области облачных вычислений могут быть использованы для снижения затрат на инфраструктуру и повышения масштабируемости распределенных вычислений. Гибкость и возможность использования ресурсов по требованию, предоставляемых облачными системами, позволяет эффективно обрабатывать большие текстовые массивы данных.

Алгоритмы машинного обучения, в том числе нейронные сети, могут быть изменены и адаптированы под конкретную предметную область [6]. Модификация методов машинного обучения может включать в себя такие аспекты как:

– совершенствование лингвистического анализа;

– использование *n*-грамм;

– применение методов кроссдоменной идентификации.

В модификациях путем развития лингвистического анализа для извлечения более продвинутых языковых аспектов особое внимание уделяется интеграции различных методов такого анализа. Примерами таких свойств являются грамматические структуры, теги частей речи, синтаксические деревья и семантические сведения. Повышение точности идентификационных моделей достигается за счет улавливания более тонких лингвистических закономерностей.

N-граммы – текстовые последовательности, состоящие из n элементов (например, слов или символов), которые идут непрерывно. В этом случае модификация подразумевает изучение локального порядка слов или символов с помощью различных вариаций n-грамм. Например, биграммы (пары последовательных слов) или n-граммы более высокого порядка могут использоваться для захвата дополнительного контекста и улучшения идентификации авторства, вместо того чтобы полагаться только на униграммы (одиночные слова).

Кроссдоменная идентификация авторства предполагает обучение модели на исходном домене и оценивает их на целевом. Цель состоит в том, чтобы создать методы, которые при наличии небольшого количества обучающих данных, специфичных для конкретного домена, или при их отсутствии смогут точно определять стиль написания автора в тех или иных требующихся доменах. Для этого исследователи используют

различные методы извлечения признаков, которые отражают стиль автора, не опираясь на данные по конкретной схеме. Примерами таких признаков являются n-граммы, синтаксические шаблоны, частоты слов и остальные характеристики, такие как стилометрические аспекты. Для улучшения обобщения предпочтение отдается лингвистическим признакам, независимым от области.

Эксперимент с идентификацией авторства текстов для больших объёмов. В рамках эксперимента для рассмотренных методов сделана выборка данных из различных областей. В качестве тестовых наборов использованы массивы из тысяч документов объемом более двухсот мегабайт каждый по трем формальным областям: литературные произведения, учебные работы и статьи с различных ресурсов.

Для классификации текстов применялись три метода. Первый метод включает в себя кроссдоменные компоненты с целью достижения универсальности в различных областях. Второй метод более сосредоточен на статистических характеристиках и выявлении специфических признаков. Третий метод пытается учитывать характеристики более высокого порядка, нежели статистические.

Результаты эксперимента приведены в таблице 1.

Таблица 1. Результаты эксперимента

	Набор 1 (литература)	*Набор 2 (учебные работы)*	*Набор 3 (статьи)*
Метод 1	35%	34%	31%
Метод 2	43%	38%	29%
Метод 3	41%	43%	34%

Как видно из таблицы, результаты эксперимента подтверждают гипотезу, что из-за больших размеров массивов документов границы между «почерками» авторов всё ещё остаются весьма размыты. Как следствие, современные методы не позволяют решать задачу идентификации автора в ее исходной постановке и требуются новые подходы и концептуально иные решения.

Кроме того, на текущем этапе лучше ограничиться использованием специализированными методами для каждой области, а универсальный метод (Метод 1) ещё требует некоторых доработок и адаптации для разных областей. Для развития универсального метода необходимо большее количество экспериментов, в том числе с более разнообразными образцами текстов.

Заключение. Таким образом, определение авторства текстов в больших массивах данных сегодня претерпевает значительные изменения. Благодаря использованию комплексных методов стало возможным изучение огромных объемов текстового материала и нахождение ранее не обнаруженных закономерностей и шаблонов. Придерживаясь этих передовых методов, важно как можно больше использовать потенциал анализа текстовых данных и изучать дальнейшее их развитие, позволяя сохранять приемлемый уровень точности при росте объёмов данных для изучения.

Список литературы

[1] Валгина Н.С. Теория текста: учебное пособие. М.: Логос, 2003. – 280 с.

[2] Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. М.: Вильямс, 2011. – 528 с.

[3] Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. М.: URSS, 2012. – 326 с.

[4] Парамонов, А.И. Методы идентификации авторства в определении студенческого плагиата. / Парамонов А. И., Труханович И. А. // «Системный анализ и прикладная информатика». 2023; (3): С. 56–59. doi: 10.21122/2309-4923-2023-3-56-59

[5] Paramonov, A. Dynamic features selection in authorship identification problem / A. Paramonov, I. Trukhanovich, U. Kuntsevich // Open Semantic Technologies for Intelligent Systems (OSTIS-2021) : сборник научных трудов / Белорусский государственный университет информатики и радиоэлектроники ; редкол. : В. В. Голенков [и др.]. – Минск, 2021. – Вып. 5. – С. 309–312.

[6] Парамонов, А. И. Модификации методов машинного обучения для решения задачи идентификации автора текста / А. И. Парамонов, И. А. Труханович // Информационно-коммуникационные технологии: достижения, проблемы, инновации (ИКТ-2022) : электронный сборник статей II Международной научно-практической конференции, Полоцк, 30–31 марта 2022 г. / Полоцкий государственный университет имени Евфросинии Полоцкой ; ред. кол.: О. А. Романов [и др.]. – Новополоцк, 2022. – С. 78-81.

Авторский вклад

Парамонов Антон Иванович – руководство исследованием, постановка задач исследования, формирование структуры статьи.

Труханович Илья Александрович – выполнение задач исследования, проведение экспериментов, анализ полученных результатов.

STATE AND DEVELOPMENT OF THE TEXT AUTHORSHIP IDENTIFICATION PROBLEM IN BIG DATA

I.A. Trukhanovich
Master of Technical Sciences

A.I. Paramonov
Head of Information Systems and Technologies Department of Institute of Information Technologies BSUIR, PhD in Technical Sciences, Associate Professor

Abstract. The state of the text authorship identification problem in Big Data is analyzed. The main problems in classifying text data with large volumes are outlined. A description of existing approaches to identification authorship is provided. Some development paths to solving the problem are shown. The methods for modifying existing approaches considering the specifics of Big Data are described. The result of an experiment with different methods of text classification on large data sets is presented.

Keywords: natural language processing, authorship identification, text classification, big data.