УДК 004.89:004.912

АНСАМБЛЕВЫЕ МЕТОДЫ МНОГОАСПЕКТНОГО АНАЛИЗА ТЕКСТОВ В ЗАДАЧАХ КАТЕГОРИЗАЦИИ ДОКУМЕНТОВ

А. И. Парамонов, И. А. Труханович

Белорусский государственный университет информатики и радиоэлектроники, Минск, Беларусь, a.paramonov@bsuir.by, ilya.trukhanovich@gmail.com

В статье представлен комплексный подход к задаче категоризации документов на основе многоаспектного анализа естественно-языковых текстов с учётом их различных особенностей. Описана общая схема ансамблевого метода, где каждая модель представления текста исследует его со своей перспективы: статистический анализ, семантическое представление, а также квантовые признаки текстовой информации. Представлены основные результаты сравнительного эксперимента, который показал, что стратифицированный подход обеспечивает более высокую точность и надёжность идентификации авторов, раскрывая глубинные уровни индивидуального стиля.

Ключевые слова: анализ текста; ансамблевые методы; идентификация авторства; категоризация документов.

ENSEMBLE METHODS OF MULTI-ASPECT TEXTS ANALYSIS IN DOCUMENT CATEGORIZATION TASKS

A. I. Paramonov, I. A. Trukhanovich

Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus, a.paramonov@bsuir.by, ilya.trukhanovich@gmail.com

The article presents a comprehensive approach to the task of document categorization based on a multi-aspect analysis of natural language texts taking into account their statistical, semantic and quantum features. The scheme of the ensemble method is described, where each model of text representation studies it from its own perspective: statistical analysis, semantic representation, and quantum features of text information. The main results of a comparative experiment are presented, which showed that the stratified approach provides higher accuracy and reliability of author identification, revealing deep levels of individual style.

Keywords: text analysis; ensemble methods; authorship identification; document categorization.

1. Введение

В последние годы проблема категоризации документов приобретает все большую значимость. Этому способствуют колоссальные темпы роста цифровых документов и развитие технологий автоматического анализа больших объемов неструктурированного текста. В современных условиях

активного использования генеративного ИИ, доступности электронных баз документов, глобального отсутствия контроля за распространением и искажением авторских текстов, в том числе появление фейковых новостей, особую актуальность приобрели задачи идентификации авторства текстов. Современные подходы к авторской атрибуции стремятся учитывать комплексные особенности текста, что позволяет повысить точность и надёжность решений. Однако большинство методов анализируют стиль письма преимущественно с одной точки зрения, не раскрывая глубинные уровни текстовой структуры. Поэтому предлагается уделять больше внимания исследованию ансамблевых методов, в которых текст изучается одновременно на разных уровнях с разных точек зрения. Таким образом, можно рассмотреть создание и экспериментальное тестирование стратифицированной модели, объединяющей разные типы признаков, и выявить преимущества подобной интеграции для задач авторской идентификации.

2. Признаки и модели

Статистические признаки определяют индивидуальный стиль автора на основании частотного анализа элементов текста, таких как буквы, слоги, слова, n-граммы и их сочетания. К основным характеристикам относятся распределения частот буквосочетаний, наиболее часто встречающиеся слова, длины предложений, а также показатели энтропии и сложности текста. В рамках исследования используются профили авторов, формируемые по частотам ключевых статистических элементов, что позволяет вычислять «расстояние» между профильным эталоном известного автора и анализируемым текстом [1]. Статистический подход эффективен для распознавания писательской манеры, особенно при работе с объемными текстовыми корпусами [2]. Применение ансамбля таких признаков повышает устойчивость модели к лексической вариативности и способствует более точной идентификации авторства.

Для обработки статистических признаков и решения задачи идентификации авторства широко применяются методы машинного обучения. Выбор конкретной модели зависит от объёма данных, сложности признаков и требований к точности. К популярным алгоритмам относятся машины опорных векторов (SVM), способные эффективно работать с высокоразмерными векторами признаков, а также деревья решений и ансамблевые методы, такие как случайный лес (Random Forest), обеспечивающие интерпретируемость и устойчивость модели [3]. Оптимальный выбор модели достигается путём сравнения качества классификации с помощью кросс-валидации и анализа ошибок, что позволяет адаптировать систему под специфику исследуемого корпуса [4].

Семантические признаки отражают смысловое содержание текста и учитывают лексические и контекстуальные связи между словами и фразами. К ним относятся векторные представления слов и текстов (word2vec, GloVe, BERT), семантические сети и онтологии, а также тематическое моделирование (LDA). Эти признаки позволяют учитывать стилистические особенности автора на уровне смысловых предпочтений и тематических паттернов, которые не всегда уловимы при чисто статистическом анализе [5, 6].

Для работы с семантическими признаками обычно применяются модели машинного обучения на основе глубоких нейросетей, включая трансформеры и рекуррентные сети, которые способны захватывать контекстные зависимости и генерировать плотные векторные представления текста [7, 8]. Кроме того, для классификации применяются методы, адаптированные к работе с высокоразмерными семантическими векторами, такие как градиентный бустинг.

Квантовые признаки в идентификации авторства базируются на использовании принципов квантовой механики для анализа текстовой информации, что позволяет описывать сложные зависимости и контексты в многомерных пространствах состояний. Такие признаки могут включать квантовые суперпозиции, запутанность и параллелизм, обеспечивающие более глубокое представление о структуре текста и авторском стиле, выходящее за рамки классического статистического и семантического анализа [9, 10].

В качестве модели для квантовых признаков в задачах идентификации авторства можно использовать вариационные квантовые схемы (Variational Quantum Circuits, VQC). Это гибридные алгоритмы, сочетающие квантовые вычисления с классической оптимизацией параметров. VQC состоит из параметризованного квантового оператора, в котором параметры обучаются с помощью классического оптимизатора путём минимизации некоторой функции стоимости. Такая архитектура позволяет использовать квантовые суперпозиции и запутанность, что расширяет пространство признаков и обеспечивает более глубокий анализ сложных зависимостей в тексте [11].

Квантовая модель $g(\vec{x})_{\vec{\theta}}$ является математическим ожиданием некоторой наблюдаемой величины *M*, оценённой на основе состояния, полученного с помощью параметризованной схемы $U(\vec{x},\vec{\theta})$:

$$g(\vec{x})_{\vec{\theta}} = \left\langle 0 \left| U^{\dagger}(\vec{x},\vec{\theta}) \right| M \left| \psi U(\vec{x},\vec{\theta}) \right| 0 \right\rangle. \quad (1)$$

В этой схеме *U* представляет собой унитарный оператор, реализуемый квантовой схемой. $\langle 0|$ является начальным базовым состоянием, а $U^{\dagger}$

представляет собой эрмитово-сопряжённый оператор к U. $\vec{x}$ является переменной сигнала (входные данные), а $\vec{\theta}$ представляет собой параметры квантовой схемы.

3. Ансамблевый метод

Для повышения точности и устойчивости авторской идентификации можно применить ансамблевый метод, который объединяет три группы признаков: статистические, семантические и квантовые. Такой подход позволяет использовать достоинства каждого типа признаков, компенсируя недостатки отдельных моделей [12].

Ансамбль строится на основе нескольких базовых моделей, каждая из которых обучена на конкретном подмножестве признаков. Итоговое решение формируется путём объединения прогнозов базовых моделей. Это может быть реализовано через механизм большинства голосов, взвешенное голосование или обучение мета-классификатора.

В частности, для каждой текстовой единицы извлекаются статистические, семантические и квантовые признаки, которые подаются на вход соответствующим классификаторам. Результаты классификации агрегируются для получения финального решения.

Такой ансамбль повышает качество распознавания авторства за счёт использования комплексной информации, включая традиционные лингвистические параметры и инновационные квантовые характеристики.

Кроме того, ансамбли способствуют снижению переобучения и повышают обобщающую способность системы, что критично для задач авторской идентификации с разнообразными и сложными текстовыми данными. За счёт адаптивного комбинирования результаты отдельных моделей интегрируются в единый прогноз с улучшенными характеристиками качества и надёжности.

Общая схема предложенного ансамбля приведена на рис. 1.

4. Эксперименты

В качестве корпуса документов для обучения и тестирования использовались тексты из электронной библиотеки М. Мошкова, которая является крупным собранием произведений классических и современных писателей [13]. В выборке содержатся тексты 50 авторов, с общим количеством около 250 документов. Средний объём каждого текста превышает 100 000 символов, что позволяет проводить детальный анализ авторских признаков.

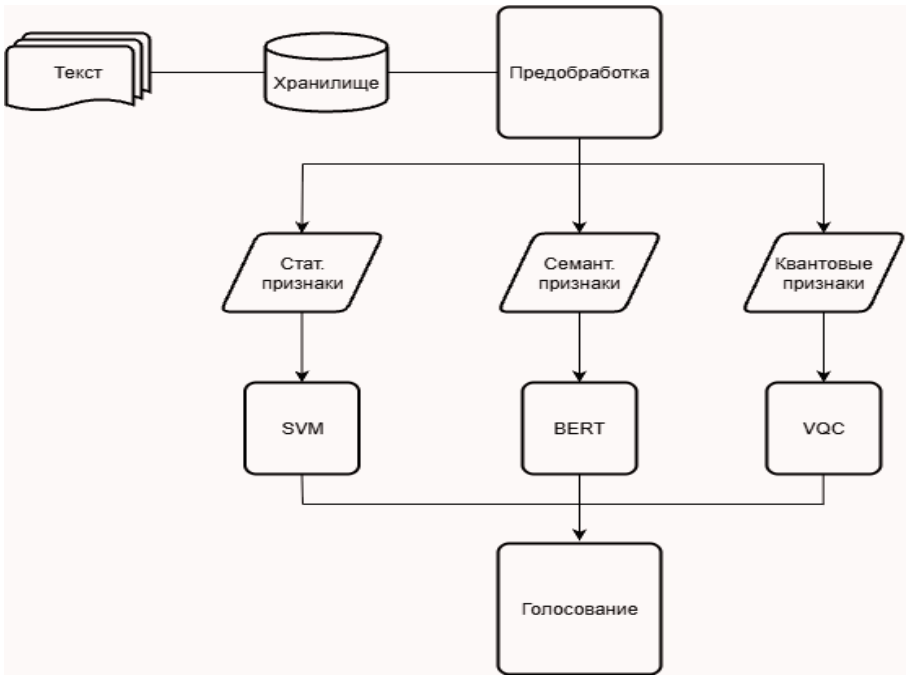

Рис. 1. Схема ансамбля с учетом разнородных признаков

Обучение моделей проводилось на 80% выборки, тестирование – на оставшихся 20%.

Все тексты корпуса прошли предобработку: токенизация, лемматизация и удаление стоп-слов.

Всего в эксперименте было представлено 5 моделей классификации.

В трех подходах использованы одиночные модели: SVM, логистическая регрессия и случайный лес.

SVM выбрана с линейным ядром, так как текстовые признаки обычно высокоразмерны и линейно разделимы. Параметр регуляризации C выставлен около 1, чтобы сбалансировать обобщающую способность.

Логистическая регрессия использовала L2-регуляризацию с параметром C около 1, оптимизирована с помощью solver lbfgs, что обеспечивает стабилизацию обучения.

Случайный лес сформирован из примерно 100–200 деревьев с максимальной глубиной примерно 15–30 для уравновешивания между переобучением и сложностью, а также с параметром max_features = sqrt, чтобы повысить разнообразие деревьев.

В качестве четвертой применён ансамбль классических моделей, включающий метод опорных векторов, логистическую регрессию и случайный лес. Параметры у моделей те же, что указаны выше. Для компенсации возможного дисбаланса классов в каждой модели применено взвешивание классов (class_weight='balanced'). Объединение результатов ан-

самбля проводилось через soft voting, при котором вероятности предсказаний каждой модели усредняются, а итоговый класс выбирается с максимальной усреднённой вероятностью.

В качестве пятой применена целевая гибридная модель с объединением различных признаков (классических, семантических, квантовых). Для классических использовалась SVM с теми же параметрами. Для семантических была применена модель BERT в режиме fine-tuning, количеством эпох 3–4 и batch size 16–32, чтобы эффективно захватывать семантические особенности текста, а также применять методы устранения переобучения, включая dropout и раннюю остановку. Вариационная квантовая схема использовала 4–6 слоев параметризованных квантовых вентилей (RX, RY, CZ), обучалась с помощью классического оптимизатора Adam с batch size 32. Объединение результатов ансамбля проводилось через soft voting.

Результаты экспериментов представлены ниже (таблица).

Результаты эксперимента

Условия	Точность
TF-IDF + SVM	0,77
TF-IDF + логистическая регрессия	0,75
TF-IDF + случайный лес	0,73
Ансамбль классических моделей (SVM + логистическая регрессия + случайный лес)	0,79
Гибридная модель = ансамбль группы признаков (классический + семантический + квантовый)	0,83

По результат экспериментов можно сделать выводы об эффективности, достигнутой при разных подходах.

Классические модели в отдельности показали достойные результаты, что соответствует ожиданиям для многоклассовой задачи с большим количеством авторов. Их ансамблирование улучшило показатели, что подтверждает пользу объединения различных классических подходов для повышения устойчивости и точности.

Вместе с тем интеграция семантических признаков с помощью языковых моделей и применение квантовых методов позволили дополнительно заметно повысить точность. Это демонстрирует, что комбинирование статистического, семантического и квантового анализа может давать более глубокое и комплексное представление об авторском стиле, особенно при работе с объемными текстовыми корпусами.

5. Заключение

В настоящей работе рассмотрен комплексный подход авторской идентификации текстов на основе объединения различных слоев признаков. В ходе проведённого эксперимента, выполненного на крупном корпусе с русскоязычными текстами, подтверждается гипотеза о высокой эффективности предложенного ансамблевого метода.

Дальнейшие исследования будут направлены на оптимизацию набора признаков, усовершенствование квантовых моделей и применение более продвинутых методов объединения. Также рассматривается возможность применение предложенных методов к более широким корпусам, включая многоязычные тексты, и обобщение подхода для решения других задач анализа неструктурированных текстов для эффективной категоризации документов.

Библиографические ссылки

1. *Труханович И. А*. Стилометрические признаки в задаче идентификации автора текста // Компьютерные системы и сети : сборник статей 58-й научной конференции аспирантов, магистрантов и студентов, Минск, 18–22 апреля 2022 г. / Белорус. гос. ун-т информатики и радиоэлектроники. Минск : 2022. С. 147–148.

2. Определение авторства текста по частотным характеристикам // Молодой ученый : сайт. URL: https://moluch.ru/conf/tech/archive/286/13237 (дата обращения: 05.07.2025).

3. *Rabinkin H.* Asymmetric Semantic Search Using Multi-Dimensional Vector Text Data Representation // Open Semantic Technologies for Intelligent Systems (OSTIS) : Research Papers Collection / Belarusian State University of Informatics and Radioelectronics ; eds.: V. V. Golenkov [et al.]. Minsk, 2025. Vol. 9. P. 339–348.

4. *Парамонов А. И.* Модификации методов машинного обучения для решения задачи идентификации автора текста // Информационно-коммуникационные технологии: достижения, проблемы, инновации (ИКТ-2022) : электронный сб. статей II Междунар. науч.-практ. конф., Полоцк, 30–31 марта 2022 г. / Полоцкий гос. ун-т им. Евфросинии Полоцкой ; редкол.: О. А. Романов [и др.]. Новополоцк : 2022. С. 78–81.

5. *Козинец А. Н.* Применение методов глубокого обучения для анализа удовлетворенности сотрудников на основе текстовых данных = Application of deep learning methods for employee satisfaction analysis based on text data // Цифровая трансформация. 2025. Т. 31, № 2. С. 13–20.

6. *Савоневская М. О.* Сравнительный анализ методов векторного представления слов в векторных базах данных // Информационные технологии и управление : материалы 61-ой научной конференции аспирантов, магистрантов и студентов, Минск, 21–25 апреля 2025 года / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: Л. Ю. Шилин [и др.]. Минск : 2025. С. 22.

7. *Шабля В. О., Коноваленко С. А., Орлов Е. О.* Методы семантического анализа на основе моделей машинного обучения с использованием искусственных нейронных сетей // Наука и реальность = Science & Reality. 2025. №1 (21). С. 113–122.

8. *Andrenko K. V.* The use of distilled large language models to determine the sentiment of a text // Open Semantic Technologies for Intelligent Systems (OSTIS) : Research Papers Collection / Belarusian State University of Informatics and Radioelectronics ; eds.: V. V. Golenkov [et al.]. Minsk, 2025. Vol. 9. P. 229–234.

9. Квантовое машинное обучение // Ultralytics. URL: https://www.ultralytics.com/ru/glossary/quantum-machine-learning (дата обращения: 05.07.2025).

10. Johns Hopkins APL Demonstrates Quantum Speedup for Text Analysis // Quantum Zeitgeist. URL: https://quantumzeitgeist.com/johns-hopkins-apl-demonstrates-quantum-speedup-for-text-analysis (date of access: 07.07.2025).

11. Variational circuits // Pennylane : URL: https://pennylane.ai/qml/glossary/variational_circuit (date of access: 10.07.2025).

12. *Труханович И. А.* Ансамблевый метод в задаче идентификации автора текста // Информационные технологии и системы 2022 (ИТС 2022) = Information Technologies and Systems 2022 (ITS 2022) : материалы Междун. науч. конф., Минск, 23 ноября 2022 / Белорус. гос. ун-т информатики и радиоэлектроники ; редкол.: Л. Ю. Шилин [и др.]. Минск, 2022. С. 171–172.

13. Lib.Ru: Библиотека Максима Мошкова : сайт. URL: https://lib.ru/ (дата обращения: 19.07.2025).