ОБРАБОТКА СИГНАЛОВ, ИЗОБРАЖЕНИЙ, РЕЧИ, ТЕКСТА И РАСПОЗНАВАНИЕ ОБРАЗОВ

SIGNAL, IMAGE, SPEECH, TEXT PROCESSING AND PATTERN RECOGNITION

УДК 004.89:004.912
https://doi.org/10.37661/1816-0301-2026-23-2-7-20

Поступила в редакцию | Received 12.03.2026
Подписана в печать | Accepted 03.04.2026
Опубликована | Published 30.06.2026

Исследование эффективности применения ансамблевых методов многоаспектного анализа текста в задачах категоризации

И. А. Труханович, А. И. Парамонов✉
✉E-mail: a.paramonov@bsuir.by

Белорусский государственный университет информатики и радиоэлектроники, ул. П. Бровки, 6, Минск, 220013, Беларусь

Аннотация

Цели. Цель представленной работы – экспериментальное исследование эффективности применения ансамблевых методов для многоаспектного анализа текстов в задачах категоризации документов на примере идентификации авторства. Особое внимание уделяется сравнению классических алгоритмов машинного обучения, их ансамблей и разработанной гибридной квантово-классической модели.

Методы. В исследовании использованы метод опорных векторов, логистическая регрессия и случайный лес, а также ансамбль этих методов и гибридная модель авторской архитектуры. Предложенный гибридный подход сочетает синтаксический анализ на основе метода опорных векторов, семантический анализ с использованием трансформерной модели BERT и квантовый вариационный модуль. Эксперименты проводились на разных корпусах текстов на английском языке с варьированием по количеству авторов. Качество оценивалось по метрикам точности, полноты и F1-меры.

Результаты. В серии экспериментов с небольшим числом авторов все модели показали высокую точность, при этом гибридная модель достигла наилучших результатов (F1-мера до 82,5 %). В экспериментах с большим числом авторов наблюдалось закономерное снижение качества, однако гибридная модель продемонстрировала лучшую устойчивость, превосходя классические ансамбли на всех корпусах. Наиболее значимый прирост точности зафиксирован на сложном корпусе коротких текстов (блогов) с большим числом авторов.

Заключение. Разработанная авторами гибридная квантово-классическая модель подтвердила свою эффективность для задач авторской атрибуции и может быть масштабирована для более широкого круга задач категоризации документов, особенно в условиях высокой размерности признаков

и большого количества классов. Применение квантового модуля позволило выявить сложные нелинейные зависимости в данных, недоступные традиционным подходам. Полученные результаты открывают перспективы для практического использования предложенного подхода в системах анализа текстов, включая обработку коротких сообщений и обширные базы авторов. Дальнейшее развитие исследования связано с расширением набора признаков, оптимизацией архитектуры квантовых схем и адаптацией модели для работы в различных прикладных областях.

Ключевые слова: ансамблевые архитектуры, категоризация документов, идентификация авторства, стилометрия, квантовые компоненты, многоаспектный анализ, обработка текста

Для цитирования. Труханович, И. А. Исследование эффективности применения ансамблевых методов многоаспектного анализа текста в задачах категоризации / И. А. Труханович, А. И. Парамонов // Информатика. – 2026. − Т. 23, № 2. – С. 7–20. – https://doi.org/10.37661/1816-0301-2026-23-2-7-20.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Research on the effectiveness of using ensemble methods of multidimensional text analysis in categorization tasks

Ilya A. Trukhanovich, Anton I. Paramonov✉
✉E-mail: a.paramonov@bsuir.by

Belarusian State University of Informatics and Radioelectronics,
st. P. Brovki, 6, Minsk, 220013, Belarus

Abstract

Objectives. The aim of the work is to experimentally investigate the effectiveness of ensemble methods for multidimensional text analysis in document categorization tasks using the example of authorship identification. Particular attention is paid to comparing classical machine learning algorithms, their ensembles, and the developed hybrid quantum-classical model.

Methods. The study uses support vector machines, logistic regression, and random forests, as well as an ensemble of these models and a hybrid model of the author's architecture. The proposed hybrid approach combines syntactic analysis based on the support vector method, semantic analysis using the BERT transformer model, and a quantum variational module. Experiments were conducted on different corpora of English texts with varying number of authors. Quality was assessed using accuracy, completeness, and F1-score metrics.

Results. In a series of experiments with a small number of authors, all models showed high accuracy, with the hybrid model achieving the best results (F1 score up to 82.5%). In experiments with a large number of authors, a regular decrease in quality was observed, but the hybrid model demonstrated better stability, outperforming classical ensembles on all corpora. The most significant increase in accuracy was recorded on a complex corpus of short texts (blogs) with a large number of authors.

Conclusion. The hybrid quantum-classical model developed by the authors has proven its effectiveness for author attribution tasks and can be scaled for a wider range of document categorization tasks, especially in conditions of high feature dimensionality and a large number of classes. The use of the quantum module made it possible to identify complex nonlinear dependencies in the data that are inaccessible to traditional approaches. The results obtained open up prospects for the practical use of the proposed approach in text analysis systems, including the processing of short messages and extensive author databases. Further development of the research is related to expanding the set of features, optimizing the architecture of quantum circuits, and adapting the model for use in various application areas.

Keywords: ensemble constructions, document categorization, authorship identification, stylometry, quantum components, multidimensional analysis, text processing

For citation. Trukhanovich I. A., Paramonov A. I. *Research on the effectiveness of using ensemble methods of multidimensional text analysis in categorization tasks*. Informatika [*Informatics*], 2026, vol. 23, no. 2, pp. 7–20 (In Russ.). https://doi.org/10.37661/1816-0301-2026-23-2-7-20.

Conflict of interests. The authors declare of no conflict of interest.

Введение

Классические модели машинного обучения, такие как метод опорных векторов (Support Vector Machine, SVM), логистическая регрессия и случайный лес, на протяжении долгого времени успешно применялись для решения прикладных задач категоризации. Несмотря на свою эффективность, эти методы часто ограничены анализом классических признаков, что может снижать точность на сложных и больших корпусах текстов. В ответ на эти ограничения в последнее время наблюдается рост интереса к ансамблевым подходам, позволяющим комбинировать преимущества различных моделей для повышения общей производительности. В работе [1] авторы ранее провели исследования возможности применения различных методов анализа текста для определения плагиата, а также предложили подходы по комбинированию методов в виде ансамбля для достижения более высокого уровня точности.

В числе перспективных исследований сегодня рассматривается направление, связанное с использованием глубокого семантического анализа текстов, который реализуется с помощью современных трансформерных моделей, например BERT. Эти модели способны извлекать более абстрактные смысловые представления, выходящие за рамки классических признаков, что особенно важно при работе с анонимными и многожанровыми текстами. В совокупности с классическими подходами на их основе можно создавать ансамблевые архитектуры, сочетающие синтаксический, семантический и статистический анализы.

С развитием квантовых вычислений формируется новый класс методов, который расширяет возможности анализа данных за счет использования квантовых вариационных схем [2]. Такие методы позволяют выявлять сложные нелинейные зависимости и скрытые паттерны в семантическом пространстве текстов, что недоступно традиционным алгоритмам. Объединение квантовых методов с классическими и семантическими моделями в составе единого ансамбля создает потенциально мощные инструменты для решения задач категоризации, в частности для авторской атрибуции.

Современные исследования демонстрируют, что комбинирование различных типов анализа может обеспечить более высокое качество категоризации документов, особенно в условиях анонимности и неструктурированности текстовых данных.

В работах [3, 4] авторы предлагают комплексный подход к задаче категоризации документов на основе многоаспектного анализа естественно-языковых текстов с учетом их различных особенностей. Предложенный подход основан на модификации ансамблевых методов машинного обучения с использованием различных слоев признаков и вычислительных моделей для их обработки.

В данной работе приведены некоторые результаты исследований эффективности предложенного многоаспектного анализа текстов в задачах категоризации документов.

Объекты исследования

В рамках исследования проведены серии экспериментов, в которых рассмотрены возможности ряда моделей машинного обучения для решения задачи идентификации авторства текстов на различных по своим характеристикам выборках корпусов документов. Были реализованы алгоритмы для таких моделей, как SVM, логистическая регрессия, случайный лес, а также для ансамбля первых трех классических моделей и ансамбля моделей с квантовым модулем (авторская архитектура).

Базовой линией при анализе текстов выступала *классическая модель на основе метода опорных векторов*, которая широко применяется и хорошо зарекомендовала себя в стилометрии [5]. Она строит гиперплоскость в пространстве признаков, оптимально разделяющую тексты разных авторов на основе группы синтаксических и лексических характеристик.

Логистическая регрессия представляет собой статистический классификатор, который моделирует вероятность принадлежности текста конкретному автору. Метод выбран по причине того, что эффективно используется при работе с большими признаковыми пространствами и хорошо интерпретируем.

Случайный лес представляет собой ансамбль решающих деревьев, который объединяет несколько моделей для повышения устойчивости и точности классификации. Он способен выявлять сложные нелинейные зависимости в данных за счет случайного выбора подмножеств признаков и образцов. Алгоритм выбран благодаря подтвержденной способности эффективно решать задачи классификации и регрессии, предлагая высокую точность и устойчивость к переобучению. В исследовании [6] показан высокий показатель эффективности данного алгоритма при решении задачи категоризации документов на сбалансированных выборках.

В качестве альтернативного алгоритма предложен *ансамбль классических моделей* в виде объединения прогнозов SVM, логистической регрессии и случайного леса. Такой ансамбль использует метод голосования или агрегирования для улучшения общего качества классификации по сравнению с отдельными моделями.

Ансамблевая модель с квантовым модулем представляет собой авторский инновационный подход, сочетающий в себе три ключевых компонента: синтаксическую модель SVM для анализа формальных признаков текста, семантическую модель BERT, способную извлекать глубокие смысловые представления, и квантовый вариационный модуль. Эта гибридная модель реализует многоаспектный анализ текстовых документов, архитектурные особенности которого описаны авторами в работах [3, 4].

Постановка экспериментов

В качестве исходных данных для формирования тестовой выборки для экспериментов используются несколько известных корпусов, которые представляют различные жанры и типы текстов на английском языке. Характеристики задействованных корпусов приведены в табл. 1.

Таблица 1

Описание тестовых корпусов

Table 1

Description of text corpora

Корпус *Corpus*	Количество текстов на автора *Number of texts per author*	Тип текстов *Type of texts*	Особенности *Peculiarities*	Источник *Source*
Blog Authorship Corpus	35	Блоговые посты	Много авторов, короткие тексты, социальные медиа	Kaggle
Project Gutenberg	20+	Литература	Классическая литература	Project Gutenberg
Reuters 50/50	50	Новостные статьи	Сбалансированный корпус новостей с авторами	UCI Repository

Первый Blog Authorship Corpus включает блоги с большим числом авторов и короткими текстами, отражающими современную цифровую культуру[1]. Второй корпус представлен Project Gutenberg, который содержит классическую литературу с разнообразными авторами и значительным текстовым массивом[2]. Корпус Reuters 50/50 включает новостные статьи по финансовой тематике с явной разметкой авторов и сбалансированным количеством текстов на каждого из них[3]. Такой набор корпусов обеспечивает широкий охват и разнородность данных, что позволяет исследовать модели и методы при работе с текстами в разных жанрах – от неформальных блогов и литературных художественных произведений до формальных новостных текстов. Все это способствует более надежной оценке моделей и определению их универсальности.

Для оценки качества категоризации предлагается использовать такие метрики, как точность, полнота и мера F1.

Точность (accuracy) отражает долю правильно классифицированных текстов по сравнению с общим числом анализируемых. Эта метрика показывает, насколько часто модель правильно определяет автора текста в целом. Высокая точность свидетельствует о том, что модель хорошо справляется с задачей авторской атрибуции, минимизируя количество ошибок. Метрика удобна для быстрой оценки общей эффективности классификатора, но при несбалансированных данных может быть недостаточно информативной.

Полнота (recall) отражает способность модели находить все тексты, принадлежащие конкретному автору. Чем выше полнота, тем меньше авторских текстов остаются нераспознанными. Эта метрика важна для оценки того, насколько полно модель охватывает все примеры автора, что критично в приложениях, где пропускать тексты

[1]Blog Authorship Corpus : [dataset]. – [San Francisco], 2026. – URL: https://www.kaggle.com/datasets/rtatman/blog-authorship-corpus (date of access: 20.01.2026).

[2]Project Gutenberg : [digital library]. – [United States], 2026. – URL: https://www.gutenberg.org (date of access: 20.01.2026).

[3]Reuters 50/50 : [dataset]. – [Irvine], 2026. – URL: https://archive.ics.uci.edu/dataset/217/reuter+50+50 (date of access: 20.01.2026).

нельзя (например, при судебной экспертизе или проверке соблюдения авторских прав). Recall помогает понять, насколько эффективно работает классификация именно по каждому из классов.

F1-мера представляет собой гармоническое среднее между точностью и полнотой. Эта метрика особенно полезна, когда важно сбалансированное качество классификации – как уменьшение ложноположительных результатов, так и минимизация пропущенных текстов конкретного автора. F1-мера учитывает обе ошибки и дает более объективную оценку модели, особенно при неоднородном распределении текстов по авторам. Метрика широко применяется в задачах авторской идентификации для комплексной характеристики работы алгоритмов.

Аппаратное и программное обеспечение эксперимента

Модели обучались на аппаратных конфигурациях следующего вида: 24-ядерный CPU (Intel Core i9), ОЗУ 64 GB DDR5, графический ускоритель NVIDIA RTX 4070.

Экспериментальная часть исследования базируется на экосистеме Python версии 3.9.x, что обеспечивает стабильную работу всех используемых компонентов и оптимальную совместимость библиотек. Основу программной среды составил фреймворк PyTorch с поддержкой CUDA, обеспечивая эффективное использование графических ускорителей при обучении нейросетевых компонентов модели.

Для реализации классических алгоритмов машинного обучения применялась библиотека scikit-learn, которая содержит реализации методов опорных векторов, логистической регрессии и случайного леса с лучшими оптимизациями. Кроме того, данная библиотека обеспечивает эффективную параллелизацию вычислений на многоядерных процессорах через механизм n_jobs.

Квантовый вариационный модуль реализован на базе фреймворка PennyLane[4], который предоставляет унифицированный интерфейс для работы с различными квантовыми симуляторами и реальными квантовыми устройствами. Для ускорения квантовых симуляций применяется оптимизированный симулятор PennyLane-Lightning, использующий векторные инструкции и параллелизацию вычислений. Этот симулятор успешно применяется на разных устройствах (как на CPU-, так и на GPU-процессорах). Поскольку эксперименты проводились на симуляции (в идеальной среде исполнения), то следует принять во внимание, что в реальных условиях на квантовых процессорах (QPU) могут быть получены иные результаты. Приведенные в работе эксперименты носят исследовательский теоретический характер с перспективой их подтверждения на реальных установках при возможности.

Результаты экспериментов

Первая серия испытаний была проведена на выборке, в которую отобрано по 10 авторов из каждого корпуса. Предполагалось, что это обеспечит равные условия для сравнения моделей на разных корпусах, снижая влияние дисбаланса по количеству классов. Таким образом был задан умеренный масштаб задачи при достаточно сложной многоавторской ситуации. Данный метод формирования выборки (отбора авторов

[4]PennyLane : [software platform]. – [Toronto], 2026. – URL: https://pennylane.ai (date of access: 20.01.2026).

и текстов) стандартизирует процесс экспериментов, упрощает интерпретацию результатов и позволяет проводить сравнения между корпусами с разной природой текстов. Результаты моделей с округлением представлены в табл. 2.

Таблица 2

Результаты эксперимента (выборка из 10 авторов)

Table 2

Results of the experiment (data sample from 10 authors)

Корпус *Corpus*	Модель *Model*	Точность *Accuracy*	F1	Полнота *Recall*	Время обучения *Training time*	Время работы *Opening hours*
Blog Authorship Corpus	SVM	73,2	74,1	71,7	12	0,03
	Логистическая регрессия	66,9	67,5	64,9	10	0,01
	Случайный лес	69,4	69,9	67,4	17	0,04
	Ансамбль классических моделей	76,7	77,9	75,3	21	0,08
	Гибридная модель	79,8	80,9	78,5	45	0,3
Project Gutenberg	SVM	74,6	75,2	72,3	14	0,05
	Логистическая регрессия	67,3	68,5	65,6	12	0,02
	Случайный лес	70,2	70,7	68,6	17	0,06
	Ансамбль классических моделей	78,7	79,7	77,2	23	0,15
	Гибридная модель	81,6	82,5	80,8	72	0,50
Reuters 50_50	SVM	71,1	71,9	69,2	19	0,05
	Логистическая регрессия	64,7	65,1	62,8	10	0,01
	Случайный лес	68,3	68,7	6,9	15	0,05
	Ансамбль классических моделей	73,5	74,2	71,6	25	0,1
	Гибридная модель	77,2	78,3	75,8	61	0,35

Время обучения указано в минутах, а время работы – в секундах на один текст. Показатели в таблице демонстрируют сравнительно одинаковую эффективность моделей, что обусловлено умеренным уровнем сложности за счет снижения влияния дисбаланса по числу классов. Ключевые выводы по первой серии испытаний можно сформулировать следующим образом:

– все модели показывают достаточно высокую точность, особенно ансамбли и гибридные модели, достигающие высоких показателей метрики F1, что указывает на хорошую способность классификаторов различать стили относительно небольшого числа авторов;

– относительно близкие результаты по разным корпусам свидетельствуют о том, что при умеренном числе классов стилистические и тематические отличия между авторами хорошо улавливаются выбранными алгоритмами, несмотря на разную природу текста;

– время работы моделей остается сравнительно небольшим, что указывает на практическую применимость алгоритмов для задач с небольшим числом авторов и достаточным объемом текстов на каждого.

Вторая серия испытаний проведена уже на комбинированной выборке с разным количеством авторов. В этом эксперименте использован набор из 100 авторов из корпуса Blog Authorship, из 30 авторов из Project Gutenberg и 50 авторов из Reuters 50/50. Выбор большого числа авторов из каждого корпуса отражает более сложную и масштабную задачу авторской идентификации с повышенной многоклассовой нагрузкой. Такая постановка эксперимента приближена к реальным условиям многоавторских систем и позволяет изучать поведение моделей в условиях, когда количество классов значительно варьируется. Методика выбора авторов и текстов обеспечивает сопоставимость результатов по разным корпусам, несмотря на различия в стиле, объеме и природе текстов. При этом структура задачи усложняется, что дает возможность оценивать устойчивость и масштабируемость моделей классификации.

В табл. 3 отражены результаты работы моделей в тяжелых многоклассовых условиях с большим количеством авторов, что моделирует более реалистичные сценарии авторской атрибуции в масштабах большого корпуса. Результаты приведены с некоторым округлением, время обучения – в минутах, время работы – в секундах на один текст.

Таблица 3

Результаты эксперимента (100, 30 и 50 авторов)

Table 3

Results of the experiment (100, 30 and 50 authors)

Корпус *Corpus*	Модель *Model*	Точность *Accuracy*	F1	Полнота *Recall*	Время обучения *Training time*	Время работы *Opening hours*
Blog Authorship Corpus	SVM	50,7	51,1	48,3	30	0,05
	Логистическая регрессия	44,4	44,9	41,9	25	0,015
	Случайный лес	46,3	46,7	44,3	32	0,05
	Ансамбль классических моделей	51,6	52,1	49,5	52	0,11
	Гибридная модель	55,3	55,8	53,1	140	0,35

Окончание табл. 3
End of table. 3

Корпус *Corpus*	Модель *Model*	Точность *Accuracy*	F1	Полнота *Recall*	Время обучения *Training time*	Время работы *Opening hours*
Project Gutenberg	SVM	58,7	59,2	56,5	34	0,03
	Логистическая регрессия	52,8	53,3	50,9	28	0,008
	Случайный лес	54,4	54,9	52,3	41	0,05
	Ансамбль классических моделей	60,2	60,6	58,2	63	0,09
	Гибридная модель	65,1	65,6	63,1	172	0,55
Reuters 50_50	SVM	62,1	62,7	59,9	30	0,05
	Логистическая регрессия	56,2	56,7	53,7	26	0,01
	Случайный лес	58,7	59,2	56,3	37	0,05
	Ансамбль классических моделей	64,2	64,8	62,1	52	0,09
	Гибридная модель	68,6	69,1	66,2	131	0,35

Ключевые выводы по данной серии испытаний можно сформулировать следующим образом:

– наблюдается заметное снижение точности и F1-меры по всем моделям, что связано с ростом числа классов и усложнением дискриминации между авторскими стилями, особенно в корпусе Blog Authorship;

– второй и третий корпусы с 30 и 50 авторами соответственно показывают более высокие метрики по сравнению с Blog Authorship при большем числе авторов, что может быть связано с особенностями корпусной структуры и большим объемом текста на каждого автора;

– ансамблевые и гибридные модели сохраняют преимущество по точности и полноте, демонстрируя лучшую устойчивость к росту числа классов, но при этом время работы существенно возрастает, особенно для гибридных решений, что влияет на производительность и требования к ресурсам.

Обсуждение результатов экспериментов

Сравнительный профиль исследуемых моделей по пяти ключевым рассмотренным характеристикам (полнота, точность, F1, превосходство над SVM, устойчивость к масштабу) изображен на рис. 1. Сплошная линия (внешний контур) отражает показатели гибридной архитектуры, контур пунктирной линией – ансамбль классических моделей, а внутренний контур построен на показателях базового метода опорных векторов.

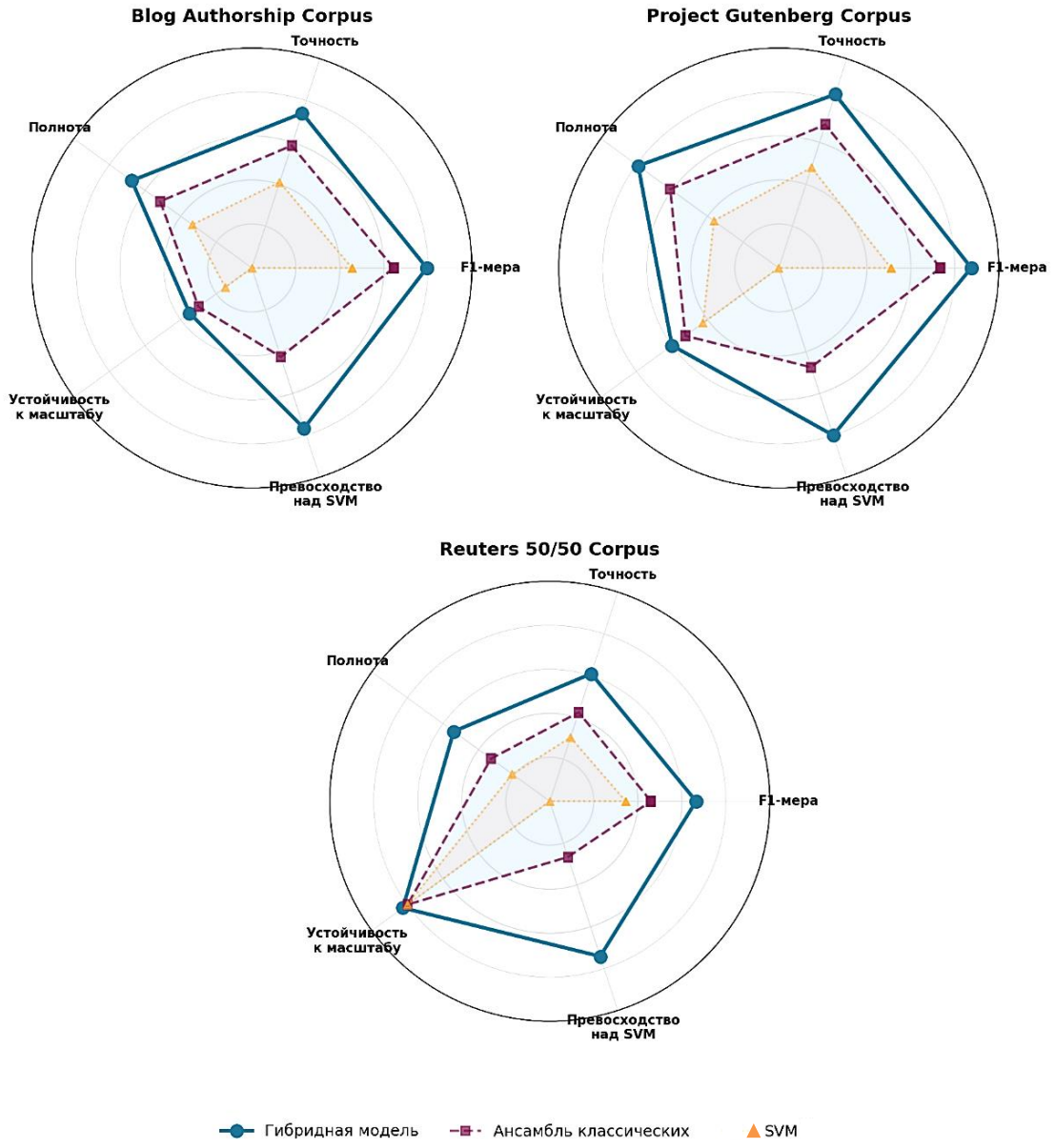

Рис. 1. Сравнительный профиль моделей
Fig. 1. Comparative profile of models

Наблюдаемое превосходство гибридной модели можно объяснить способностью квантового модуля выявлять сложные нелинейные зависимости в пространстве стилометрических признаков. Показательно, что относительный прирост качества максимален на наиболее сложном корпусе (Blog Authorship Corpus) и при большом числе классов, т. е. там, где классические методы испытывают затруднения с разделением близких авторских стилей. Это открывает перспективы применения гибридных квантово-классических архитектур в практических системах атрибуции, особенно при работе с короткими текстовыми фрагментами и обширными базами потенциальных авторов.

Помимо интегральных метрик качества важно понимать структуру ошибок классификатора: какие авторы распознаются надежно, а какие систематически смешиваются

между собой. Для этого были построены матрицы ошибок, где элемент на пересечении строки *i* и столбца *j* показывает долю текстов автора *i*, отнесенных моделью к автору *j*. Диагональные элементы соответствуют корректным предсказаниям, внедиагональные – ошибкам.

Матрица ошибок гибридной модели для Blog Authorship Corpus (10 авторов) изображена на рис. 2. Анализ матрицы показывает, что гибридная модель корректно классифицирует 76–85 % текстов каждого автора. Наибольшее смешение наблюдается между авторами C и D (семь и шесть ошибок соответственно), что объясняется схожей тематикой публикаций. Аналогичная картина и для авторов G и H, пишущих в схожем стиле.

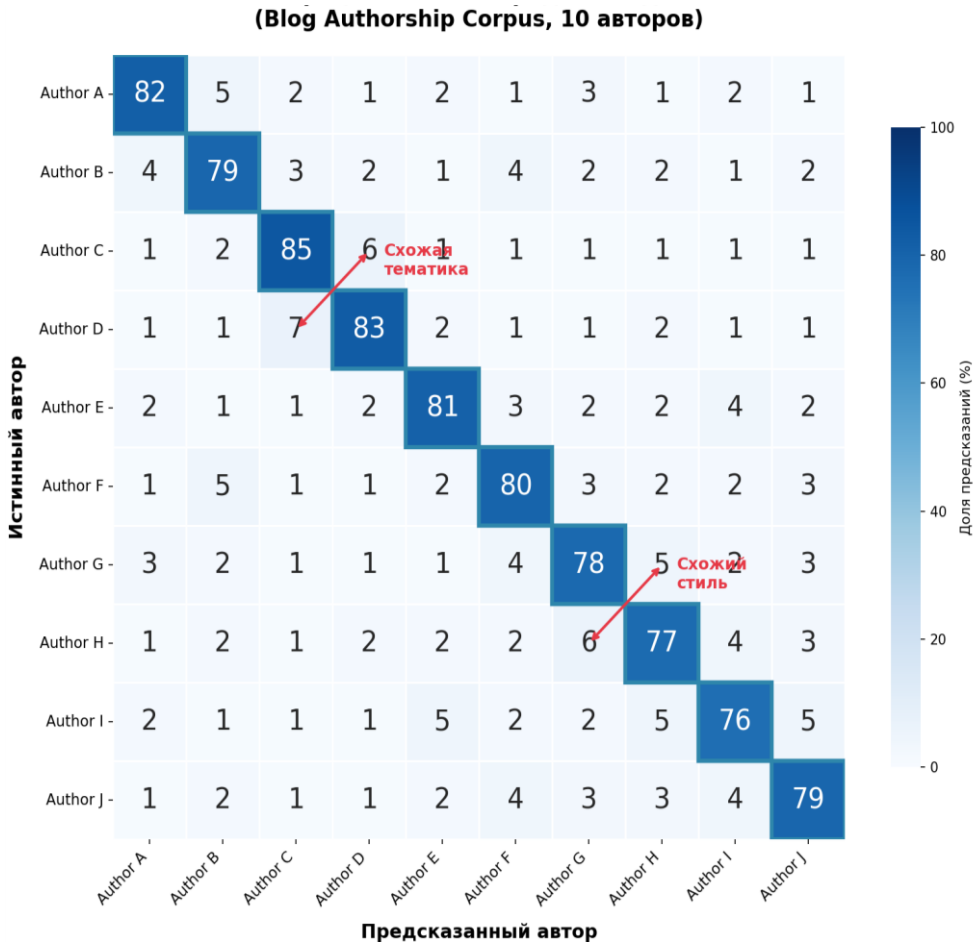

Рис. 2. Матрица ошибок гибридной модели

Fig. 2. Confusion matrix of the hybrid model

Таким образом, результаты экспериментов (табл. 2 и 3) показали, что на различных конфигурациях исследуемых корпусов и выборок гибридная модель демонстрирует устойчивые заметно лучшие показатели относительно других моделей как по точности, так и по метрике F1. Однако следует отметить, что достигнутые высокие результаты обременены большими временными затратами при обучении модели и ее работе. Это следует учитывать при построении прикладных информационных систем. В текущем ви-

де предлагаемая модель может применяться ситуативно в условиях, когда скорость работы системы некритична, а главной целью является обеспечение как можно большей точности. Кроме того, как это было в случае с некоторыми другими ML-моделями, текущая архитектура может стать гораздо эффективнее и раскрыть свой потенциал в будущем при ожидаемом появлении более совершенных технических средств эмуляции и эксплуатации [7].

Ограничения и направления дальнейшего развития

Несмотря на продемонстрированное превосходство гибридной квантово-классической модели, проведенное исследование имеет ряд ограничений, которые необходимо учитывать при интерпретации результатов и планировании практического внедрения.

Число используемых кубитов значительно ограничено, что накладывает потолок на размерность квантового пространства признаков. Масштабирование до большего количества кубитов потенциально может улучшить результаты, однако такие вычислительные ресурсы пока остаются труднодоступными. Кроме того, интерпретируемость квантового модуля остается ограниченной, поскольку в отличие от классических признаков сложно объяснить, какие именно стилистические паттерны выявляются в квантовом пространстве. Можно констатировать, что на текущий момент квантовый модуль усугубляет проблему «черного ящика». Однако проблема «черного ящика» является общей в случаях применения подобного рода ML-реализаций. Поэтому такие решения в системах поддержки принятия решения чаще могут носить консультационный характер, чтобы избежать сложностей с юридическими аспектами обоснования полученных результатов. Также заслуживает внимания исследование альтернативных архитектур квантовых схем. Варьирование глубины схемы, типов используемых вентилей и стратегий запутывания кубитов открывает пространство для оптимизации под конкретные задачи атрибуции.

Вместе с тем одним из перспективных направлений развития является расширение набора признаков. Добавление эмбеддингов, метрик читаемости и эмоциональной окраски текста может существенно повысить качество классификации.

Заключение

В рамках данной работы проведено экспериментальное исследование ансамблевых моделей и метода многоаспектного анализа текстов в задачах категоризации документов, в том числе с применением гибридного квантово-классического подхода.

Разработанная гибридная квантово-классическая модель, сочетающая классическое извлечение признаков с квантовым вариационным классификатором, была использована в серии экспериментов на различных корпусах с варьированием числа авторов. Полученные результаты позволяют утверждать, что гибридная модель превосходит классические методы и ансамбли на их основе по всем основным метрикам. Наибольшее преимущество предложенного подхода наблюдается при увеличении числа классов, что свидетельствует о лучшей способности гибридной модели работать в пространствах высокой размерности. В то же время анализ матрицы ошибок позволяет выявлять типичные паттерны смешения авторов и определять направления последующей оптимизации модели.

Дальнейшее развитие исследования связано с расширением признакового пространства, оптимизацией архитектуры квантовых схем и адаптацией модели для работы в разных условиях.

Вклад авторов. *А. И. Парамонов* предложил формулировки ключевых целей и задач исследования, определил план статьи и структуру эксперимента. Им выполнена верификация полученных результатов, а также корректировка и редактирование текста статьи; *И. А. Труханович* предложил конфигурацию гибридной модели и осуществил проведение компьютерных экспериментов. Им выполнен сбор экспериментальных данных, сформулированы основные результаты исследования и подготовлен черновик рукописи. Авторы вместе участвовали в интерпретации и анализе полученных данных компьютерных экспериментов.

Список использованных источников

1. Парамонов, А. И. Методы идентификации авторства в определении студенческого плагиата / А. И. Парамонов, И. А. Труханович // Системный анализ и прикладная информатика. – 2023. – № 3. – С. 56–59. – https://doi.org/10.21122/2309-4923-2023-3-56-59.

2. Variational quantum algorithms / M. Cerezo, A. Arrasmith, R. Babbush [et al.] // Nature Reviews Physics. – 2021. – Vol. 3, no. 9. – P. 625–644. – https://doi.org/10.1038/s42254-021-00348-9.

3. Trukhanovich, I. Intelligent analysis in text authorship identification / I. Trukhanovich, A. Paramonov // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS) : сб. науч. тр. / Бел. гос. ун-т информатики и радиоэлектроники ; редкол.: В. В. Голенков [и др.]. – Мн., 2024. – Вып. 8. – С. 327–332.

4. Парамонов, А. И. Ансамблевые методы многоаспектного анализа текстов в задачах категоризации документов / А. И. Парамонов, И. А. Труханович // Информационные системы и технологии = Information Systems and Technologies : материалы XI Междунар. науч. конгр. по информатике (CSIST-2025), Минск, 29–31 окт. 2025 г. : в 2 ч. / Бел. гос. ун-т ; редкол.: С. В. Абламейко (гл. ред.) [и др.]. – Мн., 2025. – Ч. 2. – С. 204–211.

5. Манахова, А. М. Анализ влияния стилометрических характеристик разного уровня на верификацию авторов художественных произведений / А. М. Манахова, К. В. Лагутина // Теория данных и моделирование информационных систем. – 2021. – Т. 28, № 3. – С. 260–279. – https://doi.org/10.18255/1818-1015-2021-3-260-279.

6. Веретенников, И. С. Оценка качества классификации текстовых материалов с использованием алгоритма машинного обучения «Случайный лес» / И. С. Веретенников, Е. А. Карташев, А. Л. Царегородцев // Известия Алтайского государственного университета. – 2017. – № 4(96). – URL: https://cyberleninka.ru/article/n/otsenka-kachestva-klassifikatsii-tekstovyh-materialov-s-ispolzovaniem-algoritma-mashinnogo-obucheniya-sluchaynyy-les (дата обращения: 20.01.2026).

7. Tatur, M. Open semantic technology as the foundation for new generation intelligent systems / M. Tatur, A. Paramonov // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS) : сб. науч. тр. / Бел. гос. ун-т информатики и радиоэлектроники ; редкол.: В. В. Голенков [и др.]. – Мн., 2023. – Вып. 7. – С. 61–66.

References

1. Paramonov A. I., Trukhanovich I. A. *Authorship identification methods in student plagiarism detection*. Sistemnyj analiz i prikladnaja informatika [*System Analysis and Applied Information Science*], 2023, no. 3, pp. 56–59 (In Russ.). https://doi.org/10.21122/2309-4923-2023-3-56-59.

2. Cerezo M., Arrasmith A., Babbush R., Benjamin S. C., Endo S., …, Coles P. J. Variational quantum algorithms. *Nature Reviews Physics*, 2021, vol. 3, no. 9, pp. 625–644. https://doi.org/10.1038/s42254-021-00348-9.

3. Trukhanovich I., Paramonov A. Intelligent analysis in text authorship identification. Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem: sbornik nauchnyh trudov [*Open Semantic Technologies for Intelligent Systems* (OSTIS): *Collection of Scientific Papers*]. Ed. board: V. V. Golenkov, I. S. Azarov, V. A. Golovko, A. N. Gordey, N. A. Guliakina, …, D. V. Shunkevich. Minsk, Belorusskij gosudarstvennyj universitet informatiki i radiojelektroniki, 2024, vol. 8, pp. 327–332.

4. Paramonov A. I., Trukhanovich I. A. *Ensemble methods of multi-aspect texts analysis in document categorization tasks*. Informacionnye sistemy i tehnologii: materialy XI Mezhdunarodnogo nauchnogo kongressa po informatike (CSIST-2025), Minsk, 29–31 oktjabrja 2025 goda : v 2 chastjah [*Information Systems and Technologies: Proceedings of the 2025 International Scientific Congress on Informatics (CSIST-2025), Minsk, 29–31 October 2025: in 2 parts*]. Ed. board: S. V. Ablamejko, V. V. Kazachenok, A. N. Kurbackij, V. V. Krasnoproshin. Minsk, Belorusskij gosudarstvennyj universitet, 2025, pt. 2, pp. 204–211 (In Russ.).

5. Manakhova A. M., Lagutina N. S. *Analysis of the impact of the stylometric characteristics of different levels for the verification of authors of the prose*. Teorija dannyh i modelirovanie informacionnyh sistem [*Modeling and Analysis of Information Systems*], 2021, no. 28, no. 3, pp. 260–279 (In Russ.). https://doi.org/10.18255/1818-1015-2021-3-260-279.

6. Veretennikov I. S., Kartashev E. A., Tsaregorodtsev A. L. *Assessment of the quality of text classification using the machine learning algorithm "Random forest"*. Izvestija Altajskogo gosudarstvennogo universiteta [*Izvestiya of Altai State University*], 2017, no. 4(96) (In Russ.). Available at: https://cyberleninka.ru/article/n/otsenka-kachestva-klassifikatsii-tekstovyh-materialov-s-ispolzovaniem-algoritma-mashinnogo-obucheniya-sluchaynyy-les (accessed 20.01.2026).

7. Tatur M., Paramonov A. Open semantic technology as the foundation for new generation intelligent systems: sbornik nauchnyh trudov [*Open Semantic Technologies for Intelligent Systems* (OSTIS): *Collection of Scientific Papers*]. Ed. board: V. V. Golenkov, I. S. Azarov, V. A. Golovko, A. N. Gordey, N. A. Guliakina, …, D. V. Shunkevich. Minsk, Belorusskij gosudarstvennyj universitet informatiki i radiojelektroniki, 2023, vol. 7, pp. 61–66.

Информация об авторах

Труханович Илья Александрович, соискатель, Белорусский государственный университет информатики и радиоэлектроники.
E-mail: ilya.trukhanovich@gmail.com
https://orcid.org/0000-0002-9935-1825

Парамонов Антон Иванович, кандидат технических наук, доцент, заведующий кафедрой информационных систем и технологий Института информационных технологий, Белорусский государственный университет информатики и радиоэлектроники.
E-mail: a.paramonov@bsuir.by
https://orcid.org/0000-0001-6616-2481
SPIN-код: 4280-3133

Information about the authors

Ilya A. Trukhanovich, Applicant, Belarusian State University of Informatics and Radioelectronics.
E-mail: ilya.trukhanovich@gmail.com
https://orcid.org/0000-0002-9935-1825

Anton I. Paramonov, Cand. Sci. (Eng.), Assoc. Prof., Head of the Department of Information Systems and Technologies of the Institute of Information Technologies, Belarusian State University of Informatics and Radioelectronics.
E-mail: a.paramonov@bsuir.by
https://orcid.org/0000-0001-6616-2481
SPIN code: 4280-3133