Научная статья
УДК 004.85
DOI: 10.14529/ctcr250304

ГИБРИДНЫЙ МЕТОД КЛАССИФИКАЦИИ ТЕКСТОВЫХ ДАННЫХ С УЗКОСПЕЦИАЛИЗИРОВАННОЙ ТЕРМИНОЛОГИЕЙ

***В.С. Серова**, vladislava.serova.98@mail.ru, https://orcid.org/0000-0001-9045-1048*
***А.В. Голлай**, gollaiav@susu.ru, https://orcid.org/0000-0002-5070-6779*
***Е.В. Бунова**, bunovaev@susu.ru, https://orcid.org/0000-0002-0997-8000*
Южно-Уральский государственный университет, Челябинск, Россия

Аннотация. В условиях экспоненциального роста объемов текстовой информации, особенно в предметно-ориентированных областях (технических, медицинских, юридических), задача автоматической классификации текстов, насыщенных узкоспециализированной терминологией, приобретает критическую важность. Существующие подходы, включая трансформерные модели (BERT), часто демонстрируют снижение точности при работе с редкой или доменно-специфической лексикой из-за обучения на общеупотребительных корпусах. **Целью исследования** является разработка гибридного метода Combined Neural BERT (CNB), обеспечивающего максимальную точность классификации (100 %) для текстов со специализированной терминологией за счет синергетического объединения преимуществ контекстуальных языковых моделей, лексико-статистических методов и инструментов визуализации. **Материалы и методы.** Предложенный метод CNB интегрирует три ключевых компонента: 1) BERT (или его производные) для генерации глубоких контекстуальных эмбеддингов, учитывающих семантику и порядок слов; 2) полносвязные нейронные сети (FCNN), выступающие как классификатор на основе признаков от BERT и/или обрабатывающие лексико-статистические признаки; 3) метод «Облако слов» и TF-IDF для выделения и визуализации ключевых терминов домена, формирования словаря признаков и повышения интерпретируемости. Архитектура метода включает этапы: предобработка текста (нормализация, очистка), параллельное извлечение признаков (контекстуальные эмбеддинги BERT + TF-IDF векторы), объединение признаковых пространств, классификация с помощью FCNN, интерактивная настройка на основе анализа «Облака слов». **Результаты**. Гибридный подход CNB протестирован на реальном корпусе из 10 000 обращений жителей Челябинской области (7 тематических категорий) с использованием 70 ключевых терминов и 150 стоп-слов. Метод продемонстрировал 100%-ную точность классификации после трех итераций обучения (общее время 90 мин). Ключевые преимущества: высшая точность за счет компенсации слабых мест BERT в специализированных доменах лексико-статистическими признаками; улучшенная интерпретируемость благодаря визуализации ключевых терминов «Облаком слов»; эффективность обработки больших объемов специализированных текстов. **Заключение.** Разработанный гибридный метод CNB доказал свою исключительную эффективность для классификации текстов с узкоспециализированной терминологией. Он представляет собой мощный инструмент для аналитики предметно-ориентированных текстовых массивов (юридические документы, техническая документация, медицинские заключения и т. п.) в условиях постоянно растущих объемов данных. Перспективы включают адаптацию метода для других доменов и оптимизацию вычислительной эффективности.

Ключевые слова: классификация текстов, BERT, FCNN, гибридные модели, специализированная терминология, облако слов, семантический анализ

Для цитирования: Серова В.С., Голлай А.В., Бунова Е.В. Гибридный метод классификации текстовых данных с узкоспециализированной терминологией // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». 2025. Т. 25, № 3. С. 42–52. DOI: 10.14529/ctcr250304

Original article
DOI: 10.14529/ctcr250304

HYBRID METHOD OF CLASSIFICATION OF TEXT DATA WITH SPECIALIZED TERMINOLOGY

***V.S. Serova**, vladislava.serova.98@mail.ru, https://orcid.org/0000-0001-9045-1048*
***A.V. Hollay**, gollaiav@susu.ru, https://orcid.org/0000-0002-5070-6779*
***E.V. Bunova**, bunovaev@susu.ru, https://orcid.org/0000-0002-0997-8000*
South Ural State University, Chelyabinsk, Russia

Abstract. In the context of exponential growth of text information, especially in domain-specific areas (technical, medical, legal), the task of automatic classification of texts saturated with highly specialized terminology is of critical importance. Existing approaches, including transformer models (BERT), often demonstrate a decrease in accuracy when working with rare or domain-specific vocabulary due to training on common corpora. **The aim of the study** is to develop a hybrid method Combined Neural BERT (CNB), which provides maximum classification accuracy (100 %) for texts with specialized terminology due to the synergistic combination of the advantages of contextual language models, lexical-statistical methods, and visualization tools. **Materials and methods.** The proposed CNB method integrates three key components: 1) BERT (or its derivatives) for generating deep contextual embeddings that take into account semantics and word order; 2) fully connected neural networks (FCNN) acting as a classifier based on BERT features and/or processing lexical-statistical features; 3) the Word Cloud method and TF-IDF for extracting and visualizing key domain terms, forming a feature dictionary and improving interpretability. The architecture of the method includes the following stages: text preprocessing (normalization, cleaning), parallel feature extraction (BERT contextual embeddings + TF-IDF vectors), merging feature spaces, classification using FCNN, interactive tuning based on the Word Cloud analysis. **Results.** The hybrid CNB approach was tested on a real corpus of 10,000 requests from residents of the Chelyabinsk region (7 thematic categories) using 70 key terms and 150 stop words. The method demonstrated 100 % classification accuracy after three training iterations (total time is 90 minutes). Key benefits: Higher accuracy due to compensation of BERT's weaknesses in specialized domains with lexical-statistical features; Improved interpretability due to visualization of key terms with the "Word Cloud"; Efficiency of processing large volumes of specialized texts. **Conclusion**. The developed hybrid CNB method has proven its exceptional efficiency for classifying texts with highly specialized terminology. It is a powerful tool for analyzing domain-specific text arrays (legal documents, technical documentation, medical reports, etc.) in the context of constantly growing data volumes. Prospects include adapting the method to other domains and optimizing computational efficiency.

Keywords: text classification, BERT, FCNN, hybrid models, specialized terminology, word cloud, semantic analysis

For citation: Serova V.S., Hollay A.V., Bunova E.V. Hybrid method of classification of text data with specialized terminology. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*. 2025;25(3):42–52. (In Russ.) DOI: 10.14529/ctcr250304

Введение

В современном мире объемы текстовой информации растут экспоненциально. Эффективная обработка и анализ этих данных становится критически важным для многих организаций, позволяя им принимать обоснованные решения и оперативность обработки информации. Именно поэтому задача классификации текстов, позволяющая автоматически распределять текстовые данные по категориям, является одной из самых востребованных в области интеллектуального анализа данных. Методы классификации прошли долгий путь развития: от простых алгоритмов машинного обучения до современных мощных языковых моделей. Внедрение нейронных сетей, особенно трансформеров, совершило революцию в этой области, открыв возможности для более глубокого понимания смысла текста [1, 2]. Однако, несмотря на достигнутые успехи, задача классификации текстов остается актуальной и требует дальнейших исследований в нескольких направлениях:

1) повышение точности и надежности классификации в условиях наличия специализированных терминов. Кроме того, реальные текстовые данные часто содержат ошибки, опечатки, сленг, сокращения и другие особенности, которые могут затруднить классификацию [3–5];

2) оптимизация алгоритмов классификации для работы в режиме реального времени с большими объемами текстовых данных. Необходимо разрабатывать эффективные алгоритмы, которые могут быстро обрабатывать большие объемы текстовых данных [6].

Однако, несмотря на существенный прогресс, особую сложность продолжают представлять задачи, связанные с анализом текстов, содержащих узкоспециализированную терминологию, поскольку большинство моделей обучаются на корпусах, отражающих общеупотребительную лексику [7, 8].

Таким образом, развитие и совершенствование методов классификации текстов остается важной и актуальной задачей, требующей комплексного подхода и применения передовых технологий.

Целью данной работы является разработка метода классификации текстовых данных, позволяющего оперативно и с высокой точностью (100 %) осуществлять классификацию текстовых данных, содержащих узкоспециализированную лексику.

Традиционные методы представления текста, такие как Bag-of-Words (BoW), TF-IDF и n-граммы, в значительной степени игнорируют порядок слов и не учитывают контекст, что является особенно критичным при работе с предметно-ориентированными текстами [9]. Методы векторизации следующего поколения Word2Vec, GloVe, FastText позволили моделировать семантические взаимосвязи между словами за счёт использования плотных векторов и контекстов. Однако они попрежнему страдают от ограничения статической природы векторных представлений: каждое слово кодируется одним вектором вне зависимости от контекста, в котором оно используется [10].

Существенный прорыв в области NLP связан с внедрением трансформерных архитектур, в первую очередь модели BERT (Bidirectional Encoder Representations from Transformers), которая впервые предложила двунаправленное контекстуальное представление слов. BERT и его производные (RoBERTa, ALBERT, DistilBERT) показали выдающиеся результаты в ряде стандартных NLP-задач, включая классификацию, вопросы-ответы. Эти модели обучаются на обширных корпусах, что позволяет им обрабатывать текстовые данные [11, 12]. Тем не менее даже при высоких общих показателях точности производительность BERT-систем может снижаться при применении к текстам, содержащим редкую или предметно-специфическую лексику, например, в технической, медицинской или юридической областях. Это объясняется тем, что BERT предобучается на универсальных корпусах, таких как Wikipedia, книги, журналы, и слабо адаптирован к узким доменам без дополнительной донастройки [13, 14].

Методы классификации текстовых данных могут существенно различаться по точности, сложности и требуемым ресурсам. Полносвязные нейронные сети (FCNN) традиционно применяются с различными техниками векторизации текста: BoW, TF-IDF, а также статическими эмбеддингами Word2Vec, GloVe, FastText. Более современные подходы включают sentence embeddings, такие как Doc2Vec или Sentence-BERT. Эти методы формируют числовые представления текста, которые подаются в FCNN, – архитектуру с полносвязными слоями, иногда дополненную автоэнкодерами или картами Кохонена (SOM) для визуализации и извлечения признаков. В качестве классификатора могут выступать FCNN либо внешние алгоритмы, такие как K-means [15]. Дополнительные улучшения включают регуляризацию (Dropout, L2), использование функции активации ReLU в скрытых слоях и Softmax на выходе. Оптимизация модели производится с помощью Adam. Несмотря на простоту, такие сети ограничены в учёте контекста слов и часто служат промежуточным этапом для выделения признаков [16].

В отличие от FCNN, модель BERT опирается на архитектуру трансформеров с механизмом самовнимания (self-attention), позволяющим учитывать контекст слова в обоих направлениях. Обучение BERT включает два этапа: предобучение (Masked Language Modeling и Next Sentence Prediction) и последующее дообучение под конкретную задачу. Модель использует токенизацию WordPiece, специальные токены ([CLS], [SEP]) и генерирует динамические эмбеддинги на основе окружения слов. Для задач классификации над BERT добавляется полносвязный слой, который использует вектор [CLS] как обобщенное представление текста [16].

В данной статье предложен новый гибридный метод CNB (Combined Neural BERT), который обеспечивает баланс между точностью, интерпретируемостью и вычислительной эффективностью.

Одной из перспективных стратегий является комбинирование BERT с более простыми, но адаптивными архитектурами, такими как полносвязные нейронные сети (FCNN), которые позволяют встраивать пользовательские или доменно-ориентированные признаки. Такие признаки могут быть получены, например, с использованием TF-IDF или специализированных словарей,

а затем интегрированы в архитектуру вместе с эмбеддингами от BERT. Гибридные модели позволяют достичь баланса между мощностью предобученной языковой модели и адаптивностью пользовательской логики [17]. Кроме того, использование FCNN как классификатора на основе BERT-эмбеддингов позволяет снизить вычислительную сложность и ускорить обучение, что особенно важно в условиях ограниченных ресурсов [18].

Отдельного внимания заслуживает этап интерпретации и уточнения ключевых лексических признаков, особенно в случае текстов с высокой терминологической насыщенностью. В данной задаче успешно применяются методы визуализации частотных распределений, такие как «Облако слов», позволяющее выделить наиболее релевантные термины в корпусе и понять тематику текстов до этапа обучения модели. Использование «облаков слов» как на этапе предварительного анализа, так и для уточнения гипотез об информационной насыщенности классов может способствовать лучшему отбору признаков и формированию эффективных признаковых пространств [19]. Это особенно важно, если необходимо ориентироваться на лексические особенности отраслевого языка, а не на общеупотребительные слова, типичные для обучающих корпусов большинства трансформерных моделей [20].

Рассмотрим этапы разработки гибридного метода CNB (Combined Neural BERT) классификации текстовых данных.

1-й этап. Анализ, отражающий общие используемые методы и различия в подходах к реализации данных методов для полносвязных нейронных сетей (FCNN) и нейросети BERT при классификации текстовых данных.

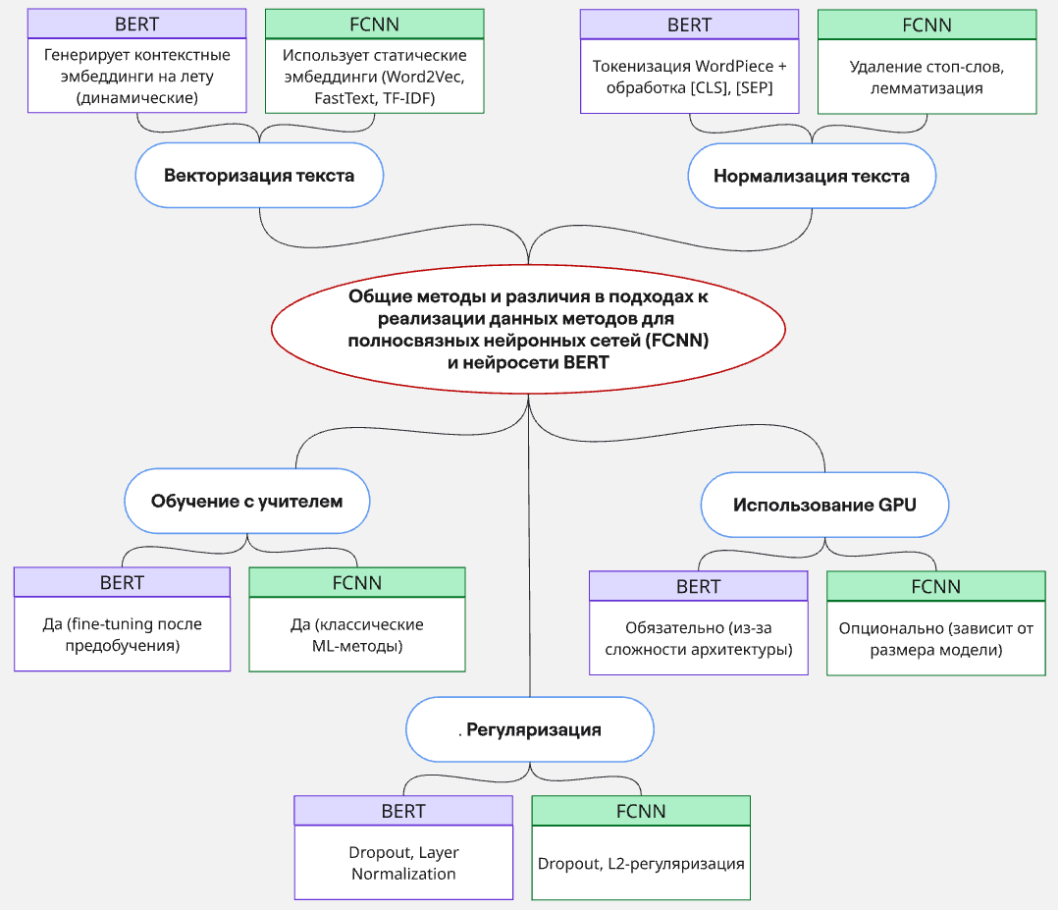

Рис. 1. Общие методы и различия в подходах к реализации данных методов для полносвязных нейронных сетей (FCNN) и модели BERT
Fig. 1. Common methods and differences in approaches to the implementation of these methods for fully connected neural networks (FCNN) and the BERT model

Обе модели (FCNN и BERT) используют общие методы: нормализацию/векторизацию текста, обучение с учителем, GPU и регуляризацию (рис. 1). Однако реализации различаются: FCNN

полагается на классическое ML и простую предобработку (удаление стоп-слов, лемматизацию), используя статические эмбеддинги (Word2Vec, TF-IDF). BERT требует предварительного обучения на больших данных и последующего дообучения (fine-tuning), применяя контекстно-зависимые эмбеддинги и токенизацию WordPiece. В вычислительном аспекте FCNN менее требователен (GPU опционален), а BERT критически зависит от GPU. Для регуляризации FCNN часто использует Dropout + L2, а BERT – Dropout + Layer Normalization для стабилизации глубокой архитектуры.

2-й этап. Анализ архитектурных и эксплуатационных характеристик для полносвязных нейронных сетей (FCNN) и нейросети BERT при классификации текстовых данных.

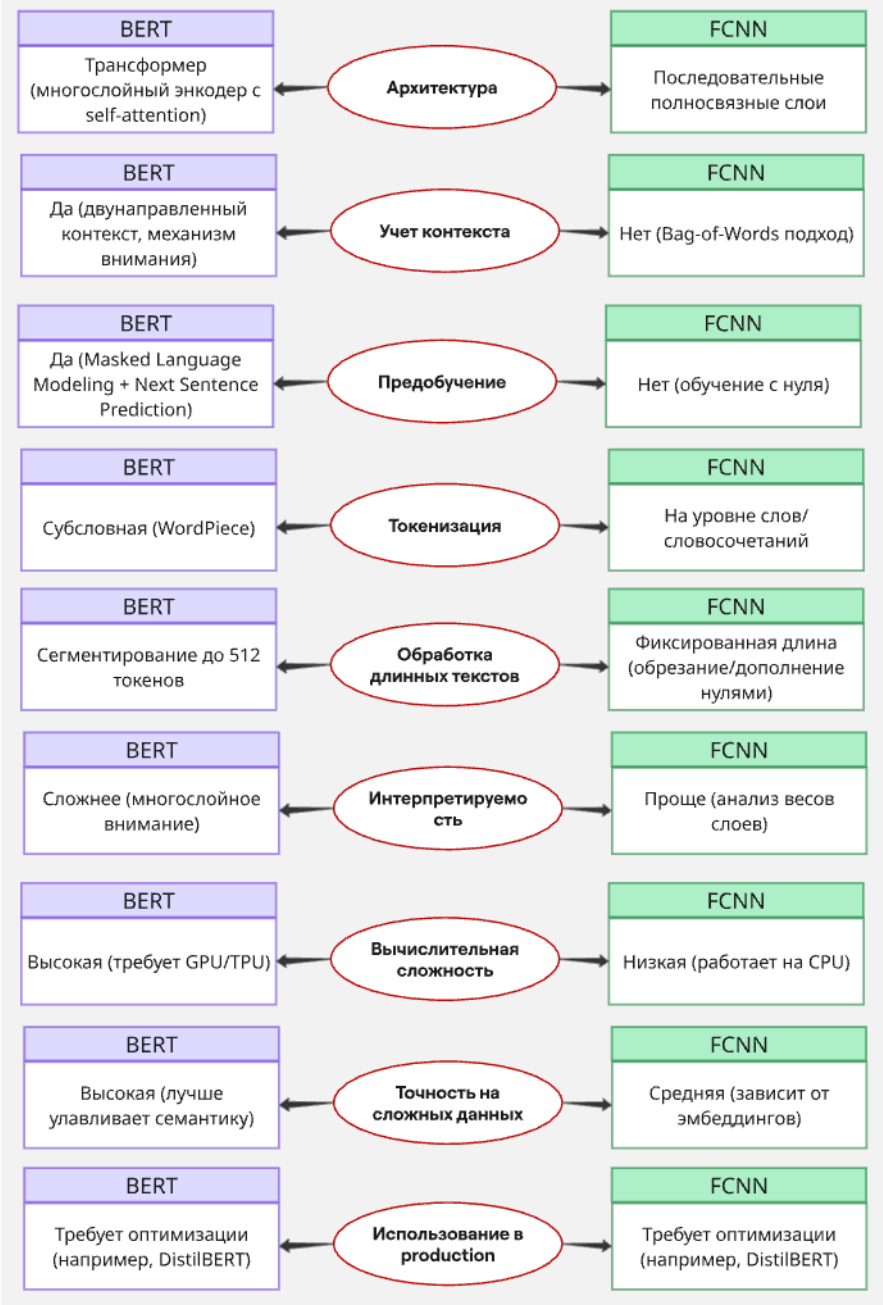

Рис. 2. Сравнительный анализ архитектурных и эксплуатационных характеристик
Fig. 2. Comparative analysis of architectural and operational characteristics

Результаты сравнительного анализа архитектурных и эксплуатационных характеристик FCNN и BERT представлены на рис. 2 и позволяют сделать следующие выводы: **FCNN** – простая

и быстрая модель, идеальная для базовых задач без сложного контекста, работает даже на обычных процессорах. BERT предпочтительнее для сложных лингвистических задач, где важен глубокий контекст, при наличии соответствующих вычислительных ресурсов или при использовании её упрощённых версий (например, DistilBERT).

3-й этап. Технологии учета специализированных терминов при классификации текстовых данных.

Для интерпретации и уточнения ключевых лексических признаков в текстах с высокой терминологической насыщенностью используется метод «Облако слов» (Word Cloud), который позволяет визуально оценить наиболее частотные лексемы в корпусе. Особенно полезно для предварительного анализа:

– выявление доминирующих терминов;

– оценка «информационной насыщенности» каждого класса;

– поддержка гипотез при ручной разметке или отборе признаков.

Преимущества метода «Облако слов»:

– метод позволяет наглядно и мгновенно определить, какие термины чаще всего встречаются в корпусе. Это важно на этапе предварительной разведки данных (exploratory data analysis);

– не требует сложной настройки или вычислительных ресурсов;

– может быть использован даже специалистами без знаний в NLP;

– удобно для анализа терминов при работе с текстами в специализированных областях:

○ помогает сразу увидеть лексические «якоря» тематики, часто недоступные при стандартной токенизации;

○ подсвечивает термины, отсутствующие в общеупотребительных моделях, как, например, у BERT;

– гибкость использования – облако может быть построено как для всего корпуса, так и по отдельным классам.

Преимущество метода «Облако слов» перед ручными отраслевыми словарями и онтологиями заключается в его способности оперативно выявлять новые или редкие термины, неологизмы, аббревиатуры и жаргонизмы (например, в ИТ или медицине) по их частотности. Словари же часто обновляются медленно и пропускают такую лексику. Это делает «Облако слов» более эффективным инструментом на ранних этапах обработки текста, особенно при адаптации к конкретному корпусу.

Особая ценность метода проявляется при работе с узкоспециализированной терминологией. Универсальные эмбеддинги (Word2Vec, BERT) обучены на общих данных и могут плохо отражать отраслевую специфику. «Облако слов» позволяет вручную или полуавтоматически выделить ключевые отраслевые термины (напр., «трансмиссия», «гепатит В») до построения признаков и обучения модели, формируя релевантный словарь ключевых признаков.

4-й этап. Разработка метода классификации текстовых данных Combined Neural BERT (CNB).

Гибридный метод Neuro Clustering System (NCS) объединяет полносвязные нейронные сети (FCNN), BERT и «Облако слов» для эффективной классификации текстовых данных. Он использует сильные стороны каждого компонента: простоту и эффективность FCNN при обработке числовых признаков и мощь BERT в работе с контекстом и семантикой текста. Из метода FCNN гибридный подход заимствует технологии очистки и нормализации текста, включая удаление стоп-слов, приведение к нижнему регистру и токенизацию, что обеспечивает качественное входное представление данных для последующего анализа (рис. 3).

Технологии, использованные в гибридном методе из нейросети BERT, представлены на рис. 4.

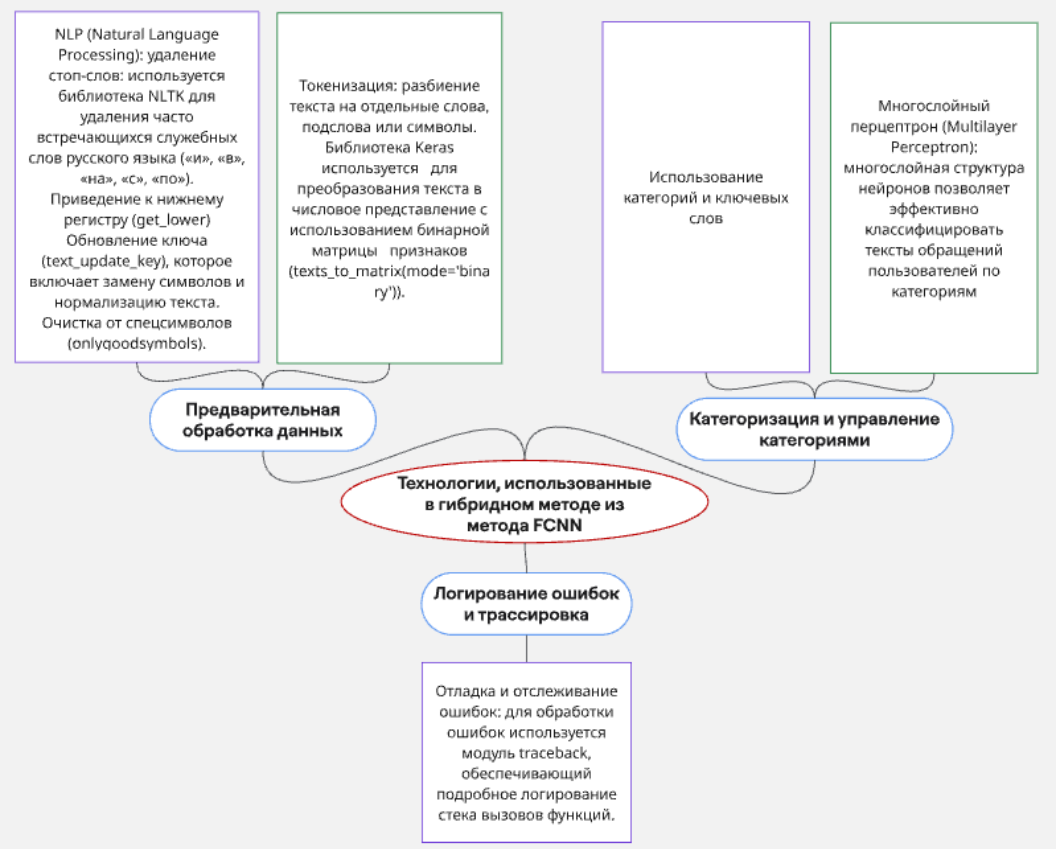

Рис. 3. Технологии, использованные в гибридном методе из нейросети FCNN
Fig. 3. Technologies used in the hybrid method from the PNN neural network

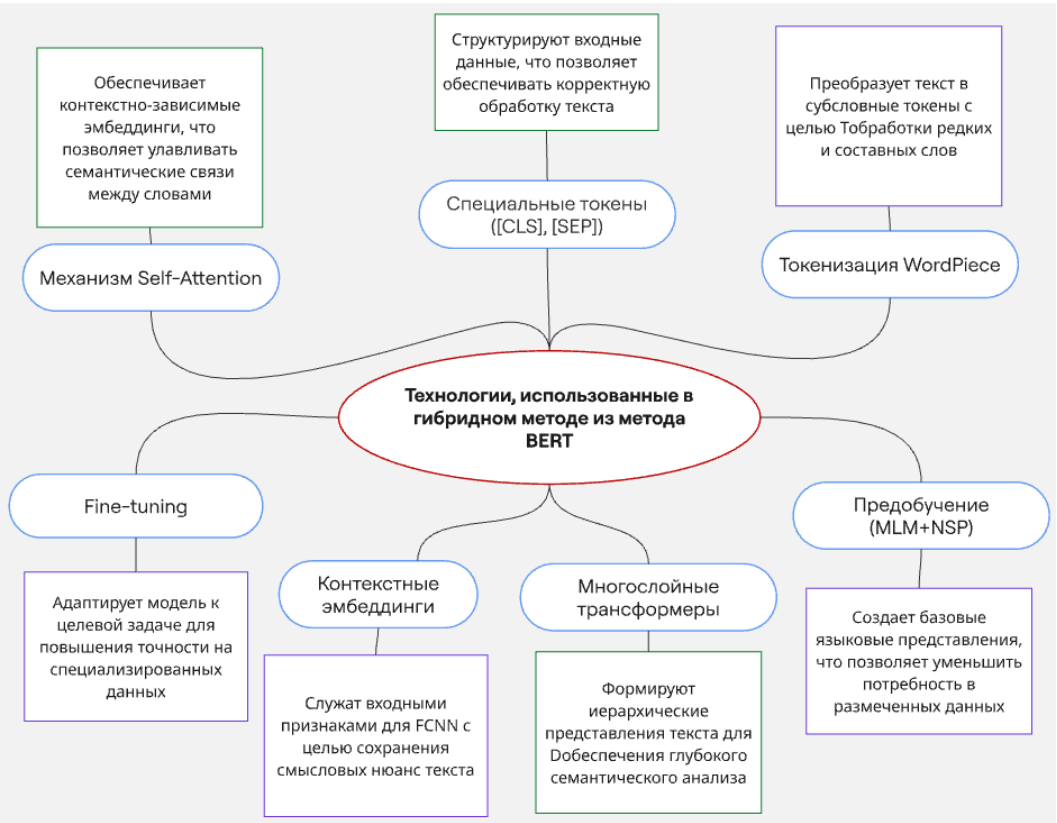

Рис. 4. Технологии, использованные в гибридном методе из нейросети BERT
Fig. 4. Technologies used in the hybrid method from the BERT neural network

Гибридный метод на основе BERT применяет передовые технологии глубокого обучения для качественного анализа текста.

1. Self-Attention и многослойные трансформеры обеспечивают глубокий контекстуальный и семантический анализ текста, выявляя связи между словами вне зависимости от их расположения.

2. Токенизация WordPiece и использование специальных токенов ([CLS], [SEP]) позволяют эффективно обрабатывать сложные и редкие слова, а также структурировать входные данные.

3. Предобучение с использованием методов Masked Language Modeling (MLM) и Next Sentence Prediction (NSP) формирует универсальные языковые представления, уменьшая потребность в большом количестве размеченных данных.

4. Контекстные эмбеддинги, полученные на выходе BERT, служат информативным входом для последующей классификации с использованием FCNN.

5. Fine-tuning позволяет адаптировать предобученную модель к конкретной задаче, что повышает точность и релевантность классификации.

В целом метод BERT в составе гибридной системы обеспечивает высокую точность за счет глубокой языковой модели и гибкости при адаптации к конкретным прикладным задачам.

Таким образом, разработан алгоритм работы гибридного метода при классификации текстовых данных с узкоспециализированной терминологией, который представлен на рис. 5.

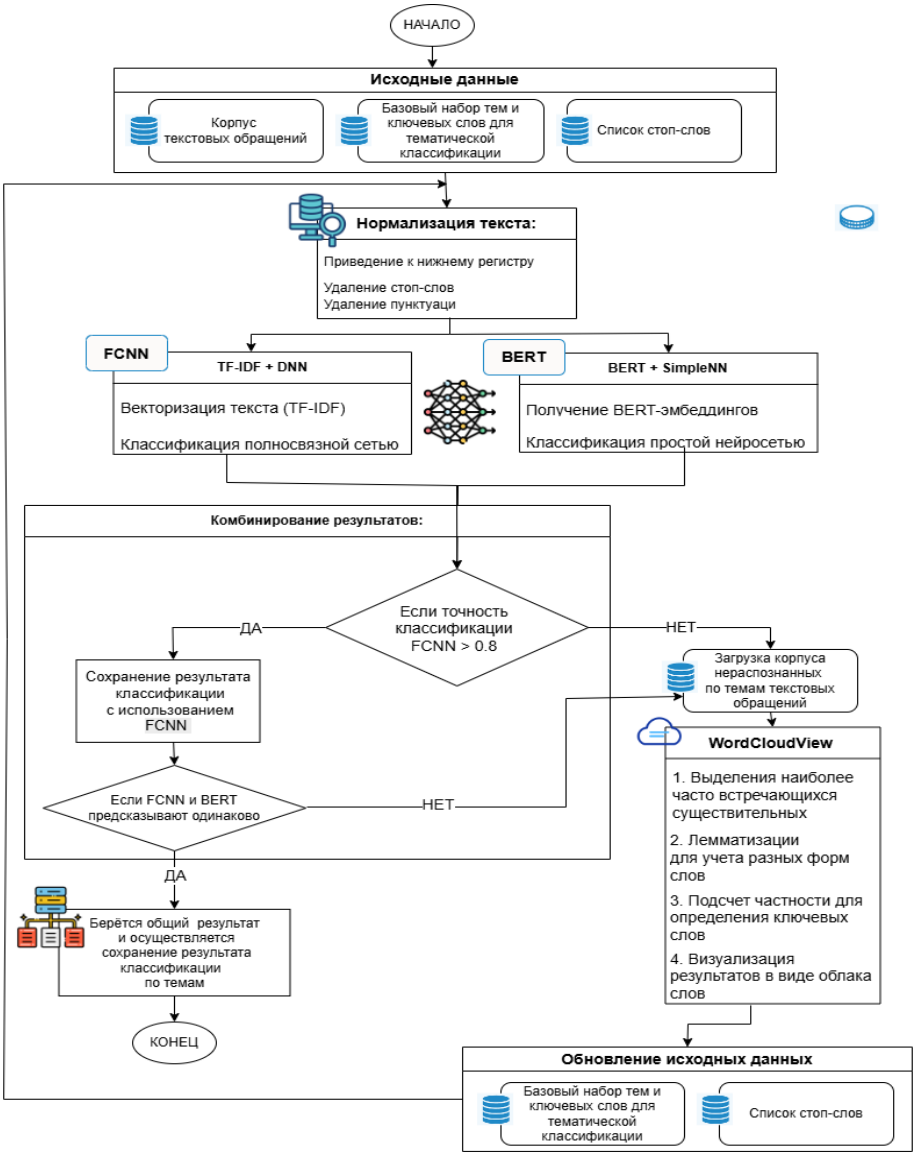

Рис. 5. Алгоритм гибридного метода при классификации текстовых данных с узкоспециализированной терминологией
Fig. 5. The algorithm of the hybrid method for classifying text data with highly specialized terminology

Гибридный метод CNB, сочетающий в себе сильные стороны полносвязных нейронных сетей и нейросети BERT для классификации текстовых данных, отличается от существующих методов двумя ключевыми особенностями: во-первых, он использует встроенную систему перекрестной проверки, где результаты полносвязной нейронной сети верифицируются с помощью BERT, повышая надежность классификации. Во-вторых, CNB позволяет интерактивно настраивать ключевые слова с помощью «Облака слов», адаптируя метод к терминологии предметной области и улучшая обработку специализированной лексики.

Данный гибридный метод протестирован на текстовых данных, включающих 10 тысяч обращений и запросов жителей Челябинской области, направленных на горячую линию Президента и губернатора региона по семи тематическим направлениям. Базовый набор ключевых слов по темам включает 70 слов. Базовый список стоп-слов включает 150 шт.

Гибридный метод CNB показал максимальную точность (100 %) после трёх итераций, которые были проведены за 90 мин.

Таким образом, современное развитие методов классификации текстов демонстрирует движение в сторону гибридных решений, способных объединять преимущества предобученных трансформеров и адаптивных пользовательских компонентов. В частности, сочетание FCNN и BERT в единой архитектуре позволяет учитывать как контекстуальные особенности текста, так и специфические признаки узкоспециализированной лексики. Кроме того, визуальные методы, такие как «Облака слов», могут служить вспомогательным инструментом для выявления терминологической структуры предметной области, обеспечивая дополнительную интерпретируемость модели и повышение точности классификации.

Список литературы

1. Attention is All You Need / A. Vaswani, N. Shazeer, N. Parmar et al. // Advances in Neural Information Processing Systems (NeurIPS). 2017. Vol. 30. DOI: 10.48550/arXiv.1706.03762

2. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of NAACL-HLT. 2019. DOI: 10.18653/v1/N19-1423

3. RoBERTa: A Robustly Optimized BERT Pretraining Approach / Y. Liu, M. Ott, N. Goyal et al. // arXiv preprint. 2019. DOI: 10.48550/arXiv.1907.11692

4. Pennington J., Socher R., Manning C.D. GloVe: Global Vectors for Word Representation // EMNLP 2014. 2014. P. 1532–1543. DOI: 10.3115/v1/D14-1162

5. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // EMNLP 2019. 2019. DOI: 10.18653/v1/D19-1410

6. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations / Z. Lan, M. Chen, S. Goodman et al. // ICLR 2020. 2020. DOI: 10.48550/arXiv.1909.11942

7. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter / V. Sanh, L. Debut, J. Chaumond, T. Wolf // arXiv preprint. – 2019. DOI: 10.48550/arXiv.1910.01108

8. Efficient Estimation of Word Representations in Vector Space / T. Mikolov, K. Chen, G. Corrado, J. Dean // arXiv preprint. 2013. DOI: 10.48550/arXiv.1301.3781

9. Enriching Word Vectors with Subword Information / P. Bojanowski, E. Grave, A. Joulin, T. Mikolov // Transactions of the Association for Computational Linguistics. 2017. Vol. 5. P. 135–146. DOI: 10.1162/tacl_a_00051

10. Платонов А.В., Мартынова И.С. Семантический анализ отзывов об организациях методами машинного обучения // Моделирование и анализ данных. 2024. Т. 14, № 1. С. 7–26. DOI: 10.17759/mda.2024140101

11. Платонов Е.Н., Руденко Е.А. Выявление и классификация токсичных высказываний методами машинного обучения // Моделирование и анализ данных. 2022. Т. 12, № 1. С. 27–48. DOI: 10.17759/mda.2022120103

12. Zhang Y., Jin R. Understanding Bag-of-Words Model: A Statistical Framework // International Journal of Machine Learning and Cybernetics. 2010. Vol. 1, no. 1–4. P. 43–52. DOI 10.1007/s13042-010-0001-0

13. Орловский К.А. Сравнительный анализ алгоритмов классификации текстов: от классических моделей к трансформерам // Вестник науки. 2023. № 5 (86). С. 1430–1443.

14. Сологуб С.Ю., Пухов В.А. Проблемы классификации текстов естественного языка методами классического машинного обучения // Моделирование и анализ данных. 2023. Т. 13, № 2. С. 64–76. DOI: 10.17759/mda.2023130203

15. Мансур А.М. Алгоритм на основе трансформеров для классификации длинных текстов // Известия ЮФУ. Технические науки. 2024. № 3 (239). С. 187–196. DOI: 10.18522/2311-3103-2024-3-187-196

16. Оськина К.А. Оптимизация метода классификации текстов, основанного на TF-IDF, за счёт введения дополнительных коэффициентов // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2016. № 15 (754). С. 175–187.

17. Игнатова Т.В., Ивичев В.А. Технологии потоковой обработки новостей для персонализированной поисковой выдачи новостного контента // Всероссийский экономический журнал «ЭКО». 2013. Vol. 43, no. 3. С. 183–189. DOI: 10.30680/ECO0131-7652-2013-3-183-189

18. Эффективная классификация текстов на естественном языке и определение тональности речи с использованием выбранных методов машинного обучения / Е.С. Плешакова, С.Т. Гатауллин, А.В. Осипов и др. // Вопросы безопасности. 2022. № 4. С. 1–14. DOI: 10.25136/2409-7543.2022.4.38658

19. Епрев А.С. Автоматическая классификация текстовых документов // Математические структуры и моделирование. 2010. Вып. 21. С. 65–81.

20. Сергиенко С.В. Отбор признаков для классификации текстов на основе ограничений для весов терминов // Сибирский аэрокосмический журнал. 2023. № 1. С. 55–60.

References

1. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*. 2017;30. DOI: 10.48550/arXiv.1706.03762

2. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*. 2019. DOI: 10.18653/v1/N19-1423

3. Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. 2019. DOI: 10.48550/arXiv.1907.11692

4. Pennington J., Socher R., Manning C.D. GloVe: Global Vectors for Word Representation. In: *EMNLP 2014*. 2014. P. 1532–1543. DOI: 10.3115/v1/D14-1162

5. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: *EMNLP 2019*. 2019. DOI: 10.18653/v1/D19-1410

6. Lan Z., Chen M., Goodman S., Gimpel K., Sharma P., Soricut R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In: *ICLR 2020*. 2020. DOI: 10.48550/arXiv.1909.11942

7. Sanh V., Debut L., Chaumond J., Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint*. 2019. DOI: 10.48550/arXiv.1910.01108

8. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint*. 2013. DOI: 10.48550/arXiv.1301.3781

9. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*. 2017;5:135–146. DOI: 10.1162/tacl_a_00051

10. Platonov A.V., Martynova I.S. Semantic Analysis of Reviews About Organizations Using Machine Learning Methods. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*. 2024;14(1):7–26. (In Russ.) DOI: 10.17759/mda.2024140101

11. Platonov E.N., Rudenko V.Y. Identification and Classification of Toxic Statements by Machine Learning Methods. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*. 2022;12(1):27–48. (In Russ.) DOI: 10.17759/mda.2022120103

12. Zhang Y., Jin R. Understanding Bag-of-Words Model: A Statistical Framework. *International Journal of Machine Learning and Cybernetics*. 2010;1(1–4):43–52. DOI: 10.1007/s13042-010-0001-0

13. Orlovsky K.A. Comparative analysis of text classification algorithms: from classical models to transformers. *Vestnik nauki*. 2023;5(86):1430–1443. (In Russ.)

14. Sologub G.B., Pukhov V.A. Problems of Natural Language Classification Using Methods of Classical Machine Learning. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*. 2023;13(2):64–76. (In Russ.) DOI: 10.17759/mda.2023130203

15. Mansour A.M. A transformer-based algorithm for classifying long texts. *Izvestiya SFedU. Engineering sciences*. 2024;3(239):187–196. (In Russ.) DOI: 10.18522/2311-3103-2024-3-187-196

16. Oskina K. A. Optimisation of TF-IDf text classification method by introducing additional weighting coefficients. *Vestnik of Moscow State Linguistic University. Humanities*. 2016;15(754):175–187. (In Russ.)

17. Ignatova T.V., Ivichev V.A. Technology of News Streaming Processing for News Personalized Recommendations. *ECO journal*. 2013;43(3):183–189. (In Russ.) DOI: 10.30680/ECO0131-7652-2013-3-183-189

18. Pleshakova E.S., Gataullin S.T., Osipov A.V., Romanova E.V., Samburov N.S. [Efficient Natural Language Text Classification and Sentiment Analysis Using Selected Machine Learning Methods]. *Security Issues*. 2022;(4):1–14. (In Russ.) DOI: 10.25136/2409-7543.2022.4.38658

19. Eprev A.S. [Automatic Classification of Text Documents]. *Mathematical Structures and Modeling*. 2010;21:65–81. (In Russ.)

20. Sergienko S.V. [Feature Selection for Text Classification Based on Term Weight Constraints]. *Siberian Aerospace Journal*. 2023;(1):55–60. (In Russ.)

Информация об авторах

Серова Влада Сергеевна, аспирант кафедры информационно-аналитического обеспечения управления в социальных и экономических системах, Южно-Уральский государственный университет, Челябинск, Россия; vladislava.serova.98@mail.ru.

Голлай Александр Владимирович, д-р техн. наук, доц., проф. кафедры информационно-аналитического обеспечения управления в социальных и экономических системах, директор Высшей школы электроники и компьютерных наук, Южно-Уральский государственный университет, Челябинск, Россия; gollaiav@susu.ru.

Бунова Елена Вячеславовна, канд. техн. наук, доц., доц. кафедры прикладной математики и информатики, Южно-Уральский государственный университет, Челябинск, Россия; bunovaev@susu.ru.

Information about the authors

Vlada S. Serova, Postgraduate student of the Department of Informational and Analytical Support of Control in Social and Economic Systems, South Ural State University, Chelyabinsk, Russia; vladislava.serova.98@mail.ru.

Alexander V. Hollay, Dr. Sci. (Eng.), Ass. Prof., Prof. of the Department of Information and Analytical Support of Management in Social and Economic Systems, Director of the Higher School of Electronics and Computer Science, South Ural State University, Chelyabinsk, Russia; gollaiav@susu.ru.

Elena V. Bunova, Cand. Sci. (Eng.), Ass. Prof., Ass. Prof. of the Department of Applied Mathematics and Programming, South Ural State University, Chelyabinsk, Russia; bunovaev@susu.ru.

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации.

Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors contributed equally to this article.

The authors declare no conflicts of interests.

Статья поступила в редакцию 03.02.2025

The article was submitted 03.02.2025