УДК 004.021

ПОДХОД К КЛАССИФИКАЦИИ ТЕКСТОВОЙ ИНФОРМАЦИИ

Ермаков Г.А. (ВКА), **Теткин С.В.** (ВКА), **Андрушкевич Д.В.** (ВКА)
Научный руководитель – кандидат технических наук Андрушкевич Д.В.
(Военно-космическая академия имени А.Ф.Можайского)

Введение. В условиях стремительного роста объема текстовой информации, доступной в цифровом формате, организации и предприятия сталкиваются с необходимостью автоматизации обработки данных. Текстовые данные, поступающие из множества источников, таких как социальные сети, деловая переписка и аналитические отчеты, требуют структурирования и анализа для принятия обоснованных решений. Однако текущие методы обработки часто не обеспечивают должной оперативности, точности и гибкости, особенно при работе с большими объемами разнородной информации [1].

Существующие подходы к обработке текстовой информации включают методы на основе правил, статистические методы, методы машинного обучения (ML) и методы глубокого обучения (DL) [2]. Методы на основе правил требуют ручного определения ключевых слов и шаблонов, обладают высокой интерпретируемостью, но плохо масштабируются. Статистические методы эффективны для анализа частотности слов, но ограничены в обработке сложных текстов. Методы машинного обучения демонстрируют высокую точность и гибкость, но требуют значительных временных и трудозатрат на подготовку данных [3]. Методы глубокого обучения, в частности трансформеры, обеспечивают высокую точность и возможность анализа контекстуальных связей, но требуют значительных вычислительных ресурсов [4]. Эти подходы не всегда способны эффективно справляться с задачами, где необходимо сочетание оперативности, высокой точности и структурного анализа [5].

При этом для эффективной обработки и систематизации информации важную роль играет классификация данных [6]. В России классификация информации и понятий осуществляется на основе как отечественных, так и международных методологий. Одним из основополагающих нормативных документов является Федеральный закон «Об информации, информационных технологиях и о защите информации», который определяет понятие информации как «сведения (сообщения, данные) независимо от формы их представления» [7]. В мировой практике применяются различные модели и стандарты классификации информации. Наиболее значимыми являются:

1. **Dewey Decimal Classification (DDC) –** система, разработанная в США и широко используемая в библиотечной практике [8].
2. **Library of Congress Classification (LCC) –** принятая в США система классификации, используемая в крупнейших библиотеках мира [8].
3. **International Standard Industrial Classification (ISIC) –** международная система классификации экономической деятельности, разработанная ООН [9].
4. **North American Industry Classification System (NAICS) –** система, используемая в США, Канаде и Мексике для классификации видов экономической деятельности [9].
5. **Medical Subject Headings (MeSH) –** специализированная система классификации медицинской информации, разработанная Национальной медицинской библиотекой США [9].

Основная часть. С целью повышения оперативности обработки текстовых данных предлагается разработать автоматизированный унифицированный классификатор на основе современных алгоритмов обработки естественного языка (Natural Language Processing, NLP) [10]. Среди существующих классификаторов выделяются УДК (Универсальная десятичная классификация), ББК (Библиотечно-библиографическая классификация) и МПК (Международная патентная классификация) [8]. УДК представляет собой наиболее

универсальную систему, используемую в различных сферах деятельности, включая науку, образование и аналитическую работу [7]. Именно ее универсальность, гибкость, масштабируемость, комбинируемость и широкая адаптация к различным типам документов делают УДК предпочтительным выбором для задач автоматической классификации текстов [8]. Классификация текстовой информации на основе УДК позволит более точно идентифицировать ключевые понятия и сущности (названия объектов инфраструктуры, организаций, персоналий и геолокации) [6].

Выводы. Таким образом, предлагаемый подход к классификации текстовой информации основан на методах глубокого обучения, включая архитектуры трансформеров, по системе УДК, что обеспечит высокую точность обработки даже сложных текстов, гибкость в применении к различным отраслям и возможность масштабирования и интеграции с существующими аналитическими системами [5]. Данный подход реализован в программном комплексе с применением наиболее популярной на сегодняшний день модели NLP BERT (Bidirectional Encoder Representations from Transformers) [4].

В дальнейшем планируется адаптация решения задачи классификации под конкретные отрасли, разработка пользовательского интерфейса для удобной интеграции в существующие системы и проведение масштабных испытаний на реальных данных для оценки эффективности. Внедрение программного комплекса позволит повысить качество и скорость обработки текстовой информации, минимизировать влияние человеческого фактора и создать основу для принятия обоснованных решений в различных сферах деятельности. Это особенно важно в условиях растущего объема информации и необходимости быстрого доступа к релевантным данным.

Список использованных источников:

1. Федеральный закон от 27 июля 2006 г. № 149-ФЗ «Об информации, информационных технологиях и о защите информации» (с изм. и доп.). – [Электронный ресурс]. – URL: http://www.consultant.ru/document/cons_doc_LAW_61798/ (дата обращения: 23.12.2024).
2. Журавлёв А. В., Калинин С. А. Методы машинного обучения в задачах классификации текстов // Вопросы искусственного интеллекта. – 2021. – Т. 12, № 1. – С. 45-59.
3. Смирнов В. А., Петров И. Н. Автоматическая классификация документов по УДК с применением алгоритмов NLP // Информационные технологии и вычислительные системы. – 2022. – № 1. – С. 75-89.
4. Vaswani A. et al. Attention is All You Need // arXiv preprint arXiv:1706.03762, 2017. – [Электронный ресурс]. – URL: https://arxiv.org/abs/1706.03762 (дата обращения: 04.01.2025).
5. Байков А. Ю. Методы классификации текстов на основе глубокого обучения // Вестник компьютерных наук. – 2023. – Т. 16, № 2. – С. 33-48.
6. Гиляревский Р. С. Информационные ресурсы и технологии: классификация и перспективы развития // Научные и технические библиотеки. – 2020. – № 3. – С. 12-19. – DOI: 10.33186/1027-3689-2020-3-12-19.
7. Универсальная десятичная классификация. Основные таблицы. – [Электронный ресурс]. – URL: https://www.udcc.org/ (дата обращения: 29.12.2024).
8. Российская государственная библиотека. Библиотечно-библиографическая классификация (ББК) и ее применение. – [Электронный ресурс]. – URL: https://www.rsl.ru/ru/prof/bbk (дата обращения: 01.02.2025).
9. Национальная электронная библиотека (НЭБ). Универсальная десятичная классификация и ее использование в систематизации знаний. – [Электронный ресурс]. – URL: https://нэб.рф/ (дата обращения: 21.01.2025).
10. Google AI Blog. BERT: Pre-training of Deep Bidirectional Transformers for Language

Understanding. – [Электронный ресурс]. – URL: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html (дата обращения: 08.01.2025).