ИНТЕГРАЦИЯ МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ОБРАБОТКУ ТЕКСТОВЫХ ДАННЫХ: КОНЦЕПТУАЛЬНЫЕ РАМКИ

Худойкулов Адхамжон Суннатулло угли
Нукусский государственный технический университет
e-mail: Barlosuser00@gmail.com
Давлетов Гуванч Атажанович
Нукусский государственный технический университет
e-mail: dga.061984@gmail.com
:Сейтназаров Куанышбай Кенесбаевич
Научный руководитель,
доктор технических наук, профессор
https://doi.org/10.5281/zenodo.16950226

ARTICLE INFO

Qabul qilindi: 15-Avgust 2025 yil
Ma'qullandi: 20- Avgust 2025 yil
Nashr qilindi: 26- Avgust 2025 yil

KEY WORDS

искусственный интеллект, обработка текстов, машинное обучение, BERT, GPT, автоматический перевод, чат-боты, анализ документов..

ABSTRACT

В данной статье рассматриваются современные подходы к интеграции методов искусственного интеллекта (ИИ) в обработку текстовых данных. Подчеркивается роль алгоритмов машинного обучения и глубинного обучения (BERT, GPT, классификационные алгоритмы и др.) в решении задач автоматического перевода, анализа документов, построения чат-ботов и других приложений. Особое внимание уделяется техническим проблемам, возникающим в процессе внедрения ИИ-технологий, а также методам их решения..

Введение

Современные информационные технологии характеризуются стремительным ростом объемов текстовых данных. Для их обработки и анализа традиционные методы оказываются недостаточными, что требует применения алгоритмов искусственного интеллекта (ИИ) и машинного обучения (МО). Интеграция таких методов позволяет решать широкий спектр задач: от автоматической классификации текстов до построения интеллектуальных систем взаимодействия с пользователем.

Объем текстовой информации в современном цифровом пространстве стремительно растет. В частности, социальные сети, новостные порталы, блоги и другие цифровые источники ежедневно производят огромные массивы неструктурированных текстов. Анализ и обработка столь больших объемов данных является сложной задачей. В связи с этим исследователи совмещают методы искусственного интеллекта (ИИ) с традиционными статистическими и алгоритмическими подходами для более глубокого анализа текстовой информации. Например, в рамках подхода DAFIM («Data Analysis Framework for Information and Media») предлагается последовательность стадий: сбор текстовых данных (через API и веб-скрапинг), предварительная очистка и обогащение, а затем применение ИИ-модулей (распознавание именованных

сущностей, семантический и эмоциональный анализ, кластеризация текста) для проведения анализа. Подобные рамки позволяют выявлять тенденции, паттерны в информационных потоках, автоматизировать коммуникацию и принимать решения на основе больших данных.

Обработка текстовых данных и концепции NLP

Обработка текстовой информации тесно связана с задачами обработки естественного языка (NLP). Главная цель — преобразовать тексты на естественном языке в машиночитаемую форму и обеспечить их интеллектуальный анализ. Обычно системы NLP включают следующие этапы: **1)** сбор и очистка данных (токенизация, нормализация), **2)** выделение признаков (векторизация слов, эмбеддинги), **3)** моделирование (классификация, кластеризация, машинный перевод и др.), **4)** визуализация и представление результатов для принятия решений.

К числу основных задач NLP относятся: **текстовая классификация** (распределение по заранее заданным категориям), **анализ тональности** (определение эмоциональной окраски текста), **распознавание сущностей (NER)** (выделение имен, локаций, организаций), **машинный перевод**, а также построение **диалоговых систем (чат-ботов)**. Классификация является одной из наиболее распространенных задач NLP, позволяя, например, разделять сообщения на «спам» и «безопасные», «положительные» и «отрицательные».

- **Задачи обработки текста:** тематическая категоризация, определение тональности, выявление намерений пользователя (например, для чат-ботов), детекция токсичного контента.
- **Классификационные алгоритмы:** традиционно применялись наивный Байес, SVM, логистическая регрессия, Random Forest. Сегодня лидирующие позиции занимают методы глубокого обучения на базе архитектуры Transformer (BERT, GPT и др.), обеспечивающие высокую точность при решении задач классификации и суммаризации текста.

Роль моделей BERT и GPT

Ключевую роль в развитии NLP сыграли модели BERT и GPT.

- **BERT** (Bidirectional Encoder Representations from Transformers) основан на двунаправленном механизме внимания, что позволяет учитывать контекст как слева, так и справа от целевого слова. Это обеспечивает высокую эффективность в задачах понимания контекста (вопрос-ответ, NER, контекстная семантика).
- **GPT** (Generative Pre-trained Transformer), напротив, использует авторегрессию, предсказывая следующее слово в последовательности. Благодаря этому GPT особенно силен в генерации текста, построении диалоговых систем и автоматизированном контент-создании.

Таким образом, BERT показывает преимущества в анализе и понимании текста, а GPT — в генерации и диалоговом взаимодействии.

Применение ИИ в NLP-задачах

Методы ИИ применяются в следующих направлениях:

- **Классификация текста:** распределение сообщений по категориям (например, спам-фильтры, тематическая классификация, тональность) .
- **Анализ тональности:** выявление эмоциональной окраски сообщений

(положительная, отрицательная, нейтральная) .

• **Распознавание сущностей (NER):** извлечение имен, географических объектов, организаций .

• **Машинный перевод и суммаризация:** использование трансформер-архитектур (BERT, GPT) для автоматического перевода и генерации кратких резюме .

• **Чат-боты и диалоговые системы:** GPT-модели позволяют строить интеллектуальных ассистентов и генеративных чат-ботов, имитирующих естественный диалог .

Концептуальные рамки и интеграция методов

Современные концептуальные рамки (например, DAFIM) включают:

• сбор данных (API, веб-скрапинг) ,

• предварительную обработку (очистка, токенизация, нормализация),

• обогащение (NER, семантическая разметка, анализ тональности) ,

• аналитическую обработку (кластеризация, выявление трендов, визуализация).

Эти этапы интегрируются с ИИ-модулями в единую систему, позволяющую получать структурированные знания из неструктурированных текстов.

Технические проблемы и пути их решения

Интеграция методов ИИ в обработку текстов сопровождается рядом сложностей:

• **Вычислительные ресурсы:** обучение крупных моделей (GPT-4) требует кластеров GPU, в то время как облегчённые версии (DistilBERT, MobileBERT) адаптированы для работы на ограниченных устройствах.

• **Объем и качество данных:** для языков с ограниченными ресурсами (например, узбекского) применимы методы трансфер-обучения.

• **Морфологическая сложность:** агглютинативные языки требуют специальных морфологических анализаторов.

• **Семантическая неоднозначность:** решается использованием контекстных моделей (BERT).

• **Смещение данных (bias):** корректируется балансировкой корпусов и методами верификации.

Заключение

Интеграция методов искусственного интеллекта в обработку текстовых данных является стратегическим направлением исследований. Модели BERT и GPT обеспечивают новый уровень понимания и генерации текста. Их применение в классификации, анализе тональности, машинном переводе и чат-ботах подтверждает эффективность NLP-интеграций. Однако остаются вызовы, связанные с вычислительными затратами, качеством данных и языковой спецификой. Решение этих проблем возможно благодаря трансфер-обучению, облегчённым моделям и созданию специализированных архитектур.

Таким образом, концептуальные рамки интеграции ИИ и NLP обеспечивают формирование многоуровневых систем обработки текста, способных повысить эффективность анализа больших массивов информации и качество цифровых сервисов

Литература:

1. Vaswani A., Shazeer N., Parmar N., et al. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017.

2. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL, 2019.
3. Brown T., Mann B., Ryder N., et al. Language Models are Few-Shot Learners. NeurIPS, 2020.
4. Jurafsky D., Martin J.H. Speech and Language Processing. Pearson, 2023.
5. Goldberg Y. Neural Network Methods for Natural Language Processing. Morgan & Claypool Publishers, 2017.
6. RN Usmanov, KK Seitnazarov. The problem of information model development for the relationship between hydrogeological object and its fuzzy-deterministic model // The Advanced Science Journal. USA - 2014. С. 67-73
7. KK Seitnazarov. Integration of gis technology for fuzzy deterministic simulation of conditions of operation and maintenance Kegeyli groundwater is abstracted // IJRET» Volum 4. С. 727-735
8. KK Seitnazarov. Dosımbetov AM, Aytanov AK, Omaraov X./Software Principles for Mapping the Relative State of Groundwater // European Journal of Molecular & Clinical Medicine ISSN. С. 2515-8260
9. KK Seitnazarov, D Turdishov, A Dosimbetov. Knowledge base of algorithmic software complex for providing agricultural fields with water resources // AIP Conference Proceedings - 2024/5/6.
10. KK Seytnazarov, Turdyshov D Kh, GP Aymurzaeva. Formation of geospatial data for information support of agricultural land monitoring // Мухаммад Ал-хоразмий авлодлари - 2019.
11. KK Seitnazarov. Dosımbetov AM, Aytanov AK/Strategy for Organization of Computational Experiments of the Functioning of Underground Water Inlets Using a Fuzzy Multiple Approach // International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan – 2020 С 1-4
12. РН Усманов, КК Сеитназаров. Об организации параллельных вычислений в процессе решения геофильтрационных задач // Вестник ТУИТ - 2014. С 101-106
13. КК Сеитназаров, ДХ Турдышов, БК Туремуратова, НС Мухиятдинов. ОБЗОР МЕТОДОВ ПОЛУЧЕНИЯ КОСМИЧЕСКИХ ИЗОБРАЖЕНИЙ С ВЫСОКИМ РАЗРЕШЕНИЕМ // НАУКА и ОБЩЕСТВО. С. 28
14. KK Seytnazarov, AA Kidirbayevich, DA Muxambetmustapayevich, XS Omarova. Software principles for mapping the relative state of groundwater // European Journal of Molecular and Clinical Medicine - 2020 С. 319-323