УДК 004.9

ИНТЕЛЛЕКТУАЛЬНАЯ ПЛАТФОРМА АНАЛИЗА И ОБРАБОТКИ АКАДЕМИЧЕСКИХ ТЕКСТОВ НА ОСНОВЕ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ

Д. А. Сьянов

Белорусский государственный университет, Беларусь, Минск, syanov@bsu.by

В статье представлена архитектура платформы, интегрированной в Open Journal Systems (OJS), которая использует нейросетевые модели для интеллектуальной обработки текстов. Рассмотрены методы предобработки данных, оценки уникальности на основе KL-дивергенции, анализа стиля с использованием трансформеров и проверки структуры текста через распределение Бернулли.

Ключевые слова: цифровая трансформация; нейросетевая модель; OJS; интеллектуальная обработка текстов.

INTELLIGENT PLATFORM FOR ANALYSIS AND PROCESSING OF ACADEMIC TEXTS BASED ON NEURAL NETWORK MODELS

D. A. Syanov

Belarusian State University, Belarus, Minsk, syanov@bsu.by

This article presents the architecture of a platform integrated into Open Journal Systems (OJS), which uses neural network models for intelligent text processing. The methods discussed include data preprocessing, uniqueness assessment based on KL-divergence, style analysis using transformers, and text structure verification through Bernoulli distribution.

Keywords: digital transformation; neural network model; OJS; intelligent text processing.

Введение

Организаторы и программные комитеты научных конференций сталкиваются с большим количеством задач при организации конференций. Зачастую применяются устаревшие формы работы, что приводит к значительной трате ресурсов. Open Journal System (OJS), изначально разработанная для управления журналами, была доработана для проведения конференций и выполнения базовых функций, таких как регистрация участников, сбор докладов, формирование программ и сборников [1].

Однако недавние достижения в области методов обработки естественного языка (Natural Language Processing, NLP [2]) позволяют расширить список задач, решаемых полученной системой, и автоматизировать критически важные этапы: классификацию заявок по тематическим трекам, оценку соответствия требованиям к материалам, выявление заимствований, генерацию предварительных рекомендаций для рецензентов и т.п.

Интеграция интеллектуальных систем на основе нейросетевых моделей в OJS позволит оптимизировать и упростить процесс проведения научных конференций и повысить качество материалов. В данной статье рассматривается архитектура разработанной платформы, предлагающая инновационный подход к автоматизированной обработке тезисов, аннотаций и полнотекстовых материалов, призванный упростить ручную проверку сотен заявок на соответствие тематике, на оригинальность, структуру, заимствования.

Архитектура платформы

Для достижения максимальной эффективности в обработке текстов разработанная платформа позволяет выявлять заимствования, производить анализ стилистики и структуры текста, а также генерацию персонализированной обратной связи, что не предусматривается существующими на данный момент системами. Платформа реализована на основе многопрофильной архитектуры, сочетающей в себе методы глубокого обучения (KL-дивергенция, трансформеры) с классическими алгоритмами обработки текстов, решающая задачу комплексного подхода к оценке текстов. Платформа строится на четырёх взаимосвязанных компонентах, каждый из которых выполняет специфические задачи, обеспечивая сквозной цикл обработки данных — от загрузки текста до формирования обратной связи.

Первый компонент обеспечивает непосредственную интеграцию с OJS через модуль, который выступает связующим звеном между OJS и остальной частью системы. Данный компонент обрабатывает входящие запросы, включая аутентификацию пользователей (редакторов, авторов) и маршрутизацию данных. Например, при загрузке статьи в OJS система через API автоматически перенаправляет файл в модуль обработки, исключая необходимость ручного вмешательства. Для работы с разнородными форматами (PDF, DOCX, Markdown) используется компонент, который преобразует документы в стандартизированный текстовый формат, сохраняя структурные элементы: заголовки, списки, формулы и т.п.

Второй компонент платформы производит обработку данных. На данном этапе необработанные тексты подготавливаются для последую-

щего анализа платформой. Процесс включает несколько уровней. Сначала текст разделяется на слова и предложения с учётом особенностей научной лексики, проходит процесс токенизации и фильтрации. При необходимости данные подвергаются структурному анализу: текст сегментируется на логические блоки структурных частей. Система также проводит автоматическую идентификацию автора, дисциплины и ключевых слов с использованием алгоритмов Named Entity Recognition (NER) на базе трансформеров, извлекая метаданные из отформатированного текстового файла. Кроме этого в данном компоненте спроектирован и реализован метод Representational State Transfer и Application programming interface (REST API), выступающий связующим звеном всей платформы.

Далее обработанный текст поступает в третий компонент платформы, реализованный на адаптированных NLP-моделей (BERT [3], GPT [4]). Происходит содержательный анализ текста, который включает три специализированных модуля: модуль уникальности (оценка оригинальности текста и выявление заимствований) [5, 6], модуль анализа стиля (оценка грамматики, связности и соответствия научному стилю), модуль соответствия формату (проверка структуры текста на соответствие требованиям), модуль проверки качества текста [7].

Последний четвёртый компонент обрабатывает результаты анализа и генерирует интерактивные отчёты, которые помогают организаторам и программным комитетам оценить представленную работу. На начальном этапе формулируются рекомендации для изменения текстовых данных на основе выявленных проблем, например, при необходимости исправления грамматических или стилистических ошибок в текстовом файле. С помощью различных python-библиотек осуществляется визуализация аналитического процесса, формируются диаграммы и графики процессов соотношения категорий в исследуемой работе с показателями уникальности, стиля и структуры. На основе выявленных особенностей текста система предлагает адаптированные рекомендации по форматированию текстового файла персонально для каждого автора. Отчёты автоматически загружаются в интерфейс OJS, где становятся доступны автору и рецензентам. Это позволяет редакторам быстро принимать решения о публикации.

Компоненты в предлагаемой системе взаимодействуют по следующему алгоритму:

автор загружает текст в OJS;

• OJS передаёт данные в адаптер, который конвертирует файл в текст;

• компонент обработки очищает и сегментирует текст, извлекает метаданные;

• компонент анализа последовательно анализирует на наличие заимствований и корректность оформления;

• компонент обратной связи генерирует отчёт и возвращает его в OJS.

Предложенная архитектура обеспечивает сквозную автоматизацию обработки академических текстов, снижая нагрузку на организаторов и программный комитет конференции.

Заключение

Разработанная система интеллектуальной обработки академических текстов на основе NLP-моделей демонстрирует значительный потенциал для цифровой трансформации образовательного процесса в условиях цифровизации. Положительные эффекты для образовательного процесса заключаются в сокращении времени проверки и анализа работ за счёт автоматизации процессов анализа текста в области его уникальности и структуры, минимизации субъективности при помощи использования алгоритмов машинного анализа на основе KL-дивергенции и трансформеров. Визуализация результатов и текстовые рекомендации помогают авторам текстов идентифицировать недочеты и погрешности в данных и целенаправленно работать над их устранением. Все вместе эти факторы приводят к повышению качества научных публикаций и возможности быстро выявлять случаи плагиата в академической среде.

Библиографические ссылки

1. *Сьянов Д. А.* Цифровая инфраструктура системы проведения научных конференций: проблемы автоматизации редакционных процессов / Д. А. Сьянов // Веб-программирование и интернет-технологии (WebConf2024) : материалы 6-й Междунар. науч.-практ. конф., Минск, 15–16 мая 2024 г. / Белорус. гос. ун-т ; редкол.: И. М. Галкин (гл. ред.) [и др.]. Минск : БГУ, 2024. С. 68-71.

2. *Vaswani A., Shazeer N., Parmap N., Uszkoreit J., Jones L., Gomez A., Kaiser L., Polosukhin I.* (2017) Attention Is All You Need [Электронный ресурс]. Режим доступа: https://doi.org/10.48550/arXiv.1706.03762. (дата обращения: 11.04.2025).

3. *Devlin J., Chang M.-W., Lee K., Toutanova K.*(2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Электронный ресурс]. Режим доступа: https://doi.org/10.48550/arXiv.1810.04805 . (дата обращения: 11.04.2025).

4. *Brown* et al. (2018, 2019, 2020) Language Models are Few-Shot Learners [Электронный ресурс]. Режим доступа: https://doi.org/10.48550/arXiv.2005.14165 (дата обращения: 11.04.2025).

5. *Beltagy I., Lo K., Cohan A.* SciBERT: A Pretrained Language Model for Scientific Text [Электронный ресурс]. Режим доступа: https://doi.org/10.48550/arXiv.1903.10676. (дата обращения: 11.04.2025).

6. *You Zhou and Jie Wang.* 2024. Detecting AI-Generated Texts in CrossDomains . In ACM Symposium on Document Engineering 2024 (DocEng ’24), August 20–23, 2024, San Jose, CA, USA. ACM, New York, NY, USA, 4 pages [Электронный ресурс]. Режим доступа: https://doi.org/10.1145/3685650.3685673/ (дата обращения: 11.04.2025).

7. *Ali, N.F. Mosharoff, Sh. Mohtasim M. Krishna G.* Automated Literature Review Using NLP Techniques and LLM-Based Retrieval-Augmented Generation [Электронный ресурс]. Режим доступа: https://arxiv.org/pdf/2411.18583 (дата обращения: 11.04.2025).