# Нейросетевой подход к анализу необработанных данных научных статей на примере оптических свойств наночастиц нитрида бора

А.Р. Резникова, А.В. Фролова

Тольяттинский государственный университет, Тольятти, Россия

**Обоснование.** Современные научные исследования сопровождаются значительным объемом данных, включая необработанные спектры и текстовые описания экспериментов. Традиционные методы ручного анализа требуют больших временных затрат и часто страдают от субъективности оценок. Исследования показывают, что в 72 % публикаций отсутствуют полные описания экспериментальных условий, что существенно затрудняет воспроизводимость результатов. Эти проблемы особенно актуальны для исследований в области материаловедения, таких как изучение оптических свойств нано- и микрочастиц нитрида бора.

**Цель** — разработка нейросетевого метода, позволяющего автоматизировать анализ научных статей, сочетая обработку текстовой информации и данных инфракрасной спектроскопии. Метод должен обеспечивать высокую точность извлечения ключевых параметров экспериментов и эффективную интеграцию разнородных данных.

**Методы.** В исследовании использован комплексный подход, объединяющий современные методы обработки естественного языка и анализа спектральных данных. Для семантического анализа текстов применялись специализированные модели SciBERT и ChemBERTa, предобученные на научных и химических текстах соответственно. Обработка спектров инфракрасного поглощения осуществлялась с помощью сверточных нейронных сетей, показавших высокую эффективность в распознавании спектральных паттернов. Особое внимание было уделено разработке гибридной архитектуры, сочетающей возможности трансформерных моделей для работы с текстами и CNN для анализа спектров. В работе использовались такие инструменты, как ChemDataExtractor для извлечения химических данных и RDKit для работы с молекулярными структурами. Данные собирались через Semantic Scholar, а их тематическая кластеризация выполнялась методом k-means.

**Результаты.** Разработанная система продемонстрировала высокую эффективность при анализе научных публикаций. На тестовой выборке из 100 статей достигнута точность 89 % при полноте данных 84 %. Временные затраты на обработку сократились с 3,5 часов при ручном анализе до менее 15 минут при использовании автоматизированной системы. Для анализа спектральных данных точность распознавания ключевых паттернов составила 91 %. В табл. 1 представлено сравнение эффективности различных NLP-моделей, используемых в работе.

**Таблица 1.** Сравнительный анализ эффективности NLP-моделей

| Параметр | BERT | SciBERT | ChemBERTa |
|---|---|---|---|
| Точность | 75 % | 89 % | 91 % |
| Область применения | Общие тексты | Научные статьи | Химические данные |

**Выводы.** Разработанный нейросетевой метод обеспечивает существенное повышение эффективности анализа научных публикаций за счет автоматизации процессов извлечения и интеграции данных. Система позволяет стандартизировать анализ экспериментальных условий, повысить воспроизводимость исследований и значительно сократить временные затраты. Перспективными направлениями дальнейших исследований являются развитие мультимодальных моделей и интеграция системы с электронными лабораторными журналами.

**Ключевые слова:** машинное обучение; инфракрасная спектроскопия; нитрид бора; обработка естественного языка; мультимодальный анализ.

*Сведения об авторах:*

**Анастасия Романовна Резникова** — студентка, ПИб-2106а; Тольяттинский государственный университет, Тольятти, Россия. E-mail: stasyrez@gmai.com

**Анастасия Валерьевна Фролова** — студентка, ПИб-2106а; Тольяттинский государственный университет, Тольятти, Россия. E-mail: fro1owa.anas7@yandex.ru

*Сведения о научном руководителе:*

**Илья Михайлович Соснин** — кандидат физико-математических наук, старший научный сотрудник; Тольяттинский государственный университет, Тольятти, Россия. E-mail: i.sosnin@tltsu.ru