УДК 004.93

АВТОМАТИЗИРОВАННАЯ ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА В ГУМАНИТАРНЫХ ИССЛЕДОВАНИЯХ

Устименко Э.Р. (СПБГУТ)
Научный руководитель – начальник отдела развития профессиональных компетенций Кривоносова Н.В.
(СПБГУТ)

Введение. В эру научных исследований, динамично развивающихся в направлении информационных технологий, автоматизированная обработка естественного языка (NLP) представляет собой неотъемлемую часть методологического арсенала, применяемого в гуманитарных науках. Этот подход, являющийся совокупностью алгоритмов и моделей машинного обучения, не только значительно повышает эффективность анализа текстовой информации, но и открывает новые перспективы для глубокого исследования культурных, литературных и исторических явлений.

Современный ученый, сталкивающийся с множеством текстовых данных, становится свидетелем уникального влияния NLP на традиционные методы гуманитарного анализа. Большие данные пришли в гуманитарные науки благодаря инициативам по оцифровке исторических документов, таких как собрание газет в Библиотеке Конгресса. Для исследователей это одновременно и проблема, и новые возможности. Информации стало так много, что обработать ее без компьютерных технологий просто невозможно [1].

Основная часть. Автоматизированная обработка естественного языка – область в науке, объединяющая два направления: гуманитарную лингвистику и инновационные технологии искусственного интеллекта. Задача NLP – создать условия для понимания компьютером смысла речи человека [2]. В науках, где текстовые данные играют ключевую роль, NLP предоставляет исследователям новые возможности для анализа разнообразных текстовых корпусов.

Преимущества использования NLP:

1) быстрый и эффективный анализ больших объемов текстовой информации. Алгоритмы NLP способны обрабатывать миллионы слов в сжатые сроки, обеспечивая ценной информацией для дальнейшего анализа.
2) способность выявлять тематические и семантические паттерны в текстах. Автоматически определяются ключевые слова, выделяются темы и идентифицируются связи между понятиями.
3) предоставление более обоснованных и полных данных для поддержки принятия решений. Это особенно важно при исследованиях, направленных на выявление культурных и социальных тенденций.
4) экономия времени и ресурсов, которые ранее требовались для ручной обработки текстов.

Вопреки значительным преимуществам существуют ряд ограничений и трудностей, которые могут влиять на точность анализа текстовых данных. Основными проблемами NLP являются сложность достижения высокой семантической точности, трудности в учете контекстуальных нюансов, сложности в обработке разнообразных текстовых форматов и возможная алгоритмическая предвзятость.

В настоящее время существует множество успешных моделей NLP. Модели, такие как BERT, GPT (Generative Pre-trained Transformer), их модификации и другие предоставляют ученым мощные инструменты для анализа и интерпретации текстов. Эти модели часто основаны на архитектурах глубокого обучения и обеспечивают высокий уровень производительности в выделении смысловых структур в тексте.

Известные проекты и примеры их применения:

1) Google Cloud Natural Language API – оценка общей эмоциональной окраски

произведения.

2) Microsoft Azure Text Analytics – анализ общественного мнения по актуальным культурным вопросам.

3) spaCy – анализ литературных текстов на предмет использования стилистических приемов и лингвистических особенностей.

4) Natural Language Toolkit – широкие возможности для анализа текстов и выделения лингвистических особенностей.

В разработке и применении алгоритмов NLP ключевую роль играет сотрудничество между научными сотрудниками и программистами. Один из ключевых аспектов этого взаимодействия – адаптация моделей к конкретным задачам. Научные сотрудники обеспечивают программистов контекстом и требованиями для создания более точных и адаптированных моделей. Это включает в себя не только выбор соответствующих архитектур и алгоритмов, но и определение того, какие особенности текстовых данных важны для конкретного исследования. Программисты в свою очередь проводят перенастройку архитектуры моделей, учитывая специфические требования. Обучение моделей на специальных корпусах текстов становится неотъемлемой частью процесса, обеспечивая моделям необходимые знания и контекст для успешного анализа. Интеграция с базами данных обеспечивает доступ к большим объемам текстовых данных, необходимых для обучения и тестирования моделей.

Тесное взаимодействие между специалистами создает синергию, позволяя преодолевать технические и методологические вызовы, стоящие перед применением NLP. Результатом этой совместной работы становятся более точные, адаптированные к контексту модели, способные эффективно решать конкретные задачи и вносить ценный вклад в области науки.

Важным этапом в анализе эффективности NLP является сравнение полученных результатов с теми, которые были получены с использованием традиционных методов. Помимо технических аспектов, такое сравнение также включает в себя оценку точности, полноты и интерпретируемости данных.

Выводы. NLP не только ускоряет процессы гуманитарного анализа, но также открывает новые горизонты для глубокого изучения культурных, литературных и исторических явлений, делая его более доступным и перспективным в контексте современных технологий.

Список использованных источников:

1. Донован М., Созыкин П. Большие данные из глубины веков. Как искусственный интеллект помогает историкам узнать правду о прошлом // Нож. — URL: https://knife.media/ai-and-history/ (дата обращения: 02.02.2024).

2. Методы обработки естественного языка // Sber Developer. — URL: https://developers.sber.ru/help/ml/natural-language-processing-techniques/ (дата обращения: 02.02.2024).