**Исследование и разработка интеллектуальной системы проверки текста на плагиат**

**Аймаганбетов Б.Д. Сулейменова Б.Б.**

**Каспийский университет технологий и инженерии имени Ш.Есенова, г.Актау, Казахстан**

**Исследование и разработка интеллектуальной системы проверки текста на плагиат**

Проблема выявления плагиата в текстах становится всё более актуальной в условиях стремительного роста информационного потока, легкости копирования и доступности текстового контента. Особенно остро данный вопрос стоит в образовательной и научной среде, где важны оригинальность и достоверность источников. Плагиат может принимать разные формы — от прямого копирования до перефразирования чужих идей без указания источника. Это делает задачу автоматической проверки на плагиат сложной и многогранной, требующей использования интеллектуальных методов.

Существует множество систем проверки текста на плагиат, однако большинство из них базируются на сравнении текста с уже известными источниками и не всегда эффективно выявляют переформулированные заимствования. Поэтому актуальной является задача разработки интеллектуальной системы, способной анализировать тексты с учетом семантики, структуры и контекста, выявляя даже скрытые формы плагиата.

**Подходы к построению интеллектуальных систем обнаружения плагиата**

Современные интеллектуальные системы обнаружения плагиата строятся на основе комбинации методов обработки естественного языка (NLP), машинного обучения, семантического анализа и лингвистических моделей. Рассмотрим ключевые этапы построения таких систем:

– **Предобработка текста**: включает очистку текста от стоп-слов, нормализацию (лемматизацию), удаление знаков препинания и токенизацию. Эта стадия критически важна для дальнейшего анализа.

– **Формирование признаков**: текст представляется в виде векторного пространства с использованием таких методов, как TF-IDF, Word2Vec, FastText или BERT. Это позволяет представить семантическое содержание текста в виде числовых признаков.

– **Сравнение текстов**: осуществляется с использованием различных метрик сходства — косинусного расстояния, евклидовой метрики, Jaccard-коэффициента и др. Для более точного анализа применяются нейронные сети, такие как трансформеры и LSTM-модели, способные улавливать смысловые связи между словами и предложениями.

– **Обнаружение перефразированного плагиата**: традиционные методы неэффективны при выявлении скрытого плагиата, поэтому применяются глубокие модели, способные понимать контекст и производить семантическое сопоставление. Такие модели обучаются на параллельных корпусах текстов с метками оригинальности.

**Применение методов машинного обучения и искусственного интеллекта**

Методы машинного обучения позволяют обучить модели на больших объемах текстов и выявлять закономерности, характерные для оригинального и заимствованного контента. Используются следующие подходы:

– **Классификация текстов**: модель обучается определять, является ли фрагмент оригинальным или заимствованным. Для этого используются алгоритмы SVM, логистическая регрессия, решающие деревья, а также более современные модели на базе глубоких нейронных сетей.

– **Семантическое сопоставление**: применяется для определения степени сходства между фрагментами текста. Здесь особенно эффективны предобученные языковые модели (например, BERT), позволяющие учитывать контекст и скрытое значение слов.

– **Обучение с учителем и без учителя**: в зависимости от наличия размеченных данных, применяются различные методы обучения. Обучение с учителем требует предварительной разметки текстов, тогда как обучение без учителя позволяет выявлять кластеры схожих фрагментов без явной метки.

**Гибридные подходы и интеграция баз данных**

Наиболее эффективными являются гибридные системы, сочетающие лексическое, синтаксическое и семантическое сравнение. Такие системы интегрируют:

– локальные базы данных ранее загруженных работ и научных статей, – глобальные источники информации (например, открытые электронные библиотеки и веб-контент), – алгоритмы поиска по графу или семантическому индексу.

Это позволяет не только повысить точность проверки, но и выявлять заимствования из различных источников, включая переводы, адаптированные и переформулированные тексты.

**Примеры успешных реализаций и исследований**

На практике интеллектуальные системы проверки текста на плагиат уже внедряются в университетах, научных издательствах и онлайн-платформах. Например:

– Система Turnitin использует собственную обширную базу данных и алгоритмы сопоставления для выявления сходств, но слабо справляется с парафразированием.

– Российская система "Антиплагиат" применяет как лексический, так и семантический анализ, расширяя функциональность за счет подключения к научным электронным библиотекам.

– Исследования, основанные на применении BERT и других трансформеров, показывают высокую точность в определении перефразированного плагиата и активно внедряются в научной сфере.

**Перспективы развития**

Дальнейшее развитие интеллектуальных систем обнаружения плагиата связано с:

– **Интеграцией моделей глубокого обучения**, способных учитывать не только текстовую структуру, но и логическую связность, интонационные и прагматические особенности.

– **Развитием многоязычных систем**, способных выявлять переведенный плагиат, что особенно важно в международной научной коммуникации.

– **Использованием онтологий и семантических графов** для более глубокого понимания смысла текста и установления тематических связей между фрагментами.

– **Самообучающимися системами**, способными накапливать знания и улучшать точность на основе обратной связи от пользователей.

**Заключение**

Разработка интеллектуальной системы проверки текста на плагиат требует комплексного подхода с использованием современных методов искусственного интеллекта и анализа естественного языка. Такие системы не только улучшают точность обнаружения заимствований, но и способствуют формированию академической честности, повышению качества научных и образовательных текстов. В условиях цифровой трансформации общества интеллектуальные системы становятся необходимым инструментом обеспечения оригинальности и авторской этики.

**Литература**

1.Alzahrani, S. M., Salim, N., & Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(2), 133–149.

2. Barrón-Cedeño, A. (2010). On the Language Specificity in Plagiarism Detection. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*.
 3. Potthast, M., Stein, B., & Anderka, M. (2008). A Wikipedia-Based Multilingual Retrieval Model. *ECIR 2008: Advances in Information Retrieval*, 522–530.
 4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*.
 5. Eissen, S. M., & Stein, B. (2006). Intrinsic Plagiarism Detection. *Advances in Information Retrieval, ECIR 2006*, 565–569.
 6. Kumar, N., & Vats, M. (2020). A Survey on Plagiarism Detection Techniques. *Journal of Information and Optimization Sciences*, 41(5), 1173–1187.

**Облачная платформа автоматизированного корпоративного документооборота: всесторонний обзор**

**Аймаганбетов Б.Д. Сулейменова Б.Б.**

**НАО «Каспийский университет технологий и инженерии имени Ш. Есенова», г. Актау, Казахстан**

### Резюме

В данной статье рассматриваются современные подходы к созданию интеллектуальных систем обнаружения текстового плагиата. Описаны ключевые этапы обработки текста, включая предобработку, формирование признаков и методы сравнения, основанные на векторных и семантических представлениях. Особое внимание уделено применению методов машинного обучения, включая обучение с учителем и без, а также использованию предобученных языковых моделей (например, BERT) для выявления скрытых и перефразированных заимствований. Рассматриваются гибридные системы, сочетающие семантический и лексический анализ, а также примеры существующих решений и направления дальнейшего развития в области многоязычного анализа, применения онтологий и глубокого обучения.

### Ключевые слова

плагиат, интеллектуальные системы, машинное обучение, семантический анализ, BERT, обработка естественного языка, сравнение текстов, векторизация текста, гибридные алгоритмы, проверка на заимствование