УДК 004.85
DOI: 10.21122/2309-4923-2023-3-56-59

ПАРАМОНОВ А.И., ТРУХАНОВИЧ И.А.

МЕТОДЫ ИДЕНТИФИКАЦИИ АВТОРСТВА В ОПРЕДЕЛЕНИИ СТУДЕНЧЕСКОГО ПЛАГИАТА

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

В современном образовательном контексте проблема плагиата является актуальной и требует разработки эффективных методов обнаружения и предотвращения данного явления. Рассмотрено применение методов идентификации авторства в области обнаружения студенческого плагиата. Исследованы различные подходы, используемые для проверки, обнаружения и анализа плагиата в различных работах. Рассмотрены как классические методы, в числе которых сравнение текстов и поиск сходства, так и современные методы, основанные на алгоритмах машинного обучения, а также их комбинирование и потенциальные модификации. Также обсуждены преимущества и ограничения каждого метода и даны рекомендации по выбору того или иного подхода в соответствии с конкретными требованиями исследования.

Особое внимание уделено таким современным методам, как анализ метаданных и применение нейронных сетей. Стилистический анализ позволяет выявить авторские особенности, такие как выбор слов, предпочтительные формулировки и даже пунктуация. Лексические и синтаксические модели используются для выявления повторяющихся фраз и структур, которые могут указывать на плагиат. Статистические методы позволяют выявить аномалии в употреблении слов и фраз, а машинное обучение – создать модели, позволяющие рассчитать вероятность плагиата на основе большого количества данных.

В конечном итоге предоставлено сравнение методов идентификации авторства в области определения студенческого плагиата, что имеет целью дать ценную информацию о различных подходах и их применимости, а также поможет исследователям и преподавателям разработать эффективные стратегии выявления и предотвращения плагиата в образовательной среде.

Ключевые слова: *машинное обучение, плагиат, идентификация автора*

Введение

Плагиат студентов – серьезная проблема, которая продолжает оставаться актуальной в сфере образования. Использование чужих слов, идей или работ без должного указания называется плагиатом. Он ставит под угрозу такие ценности, как честность, оригинальность и академическая добросовестность.

В современную цифровую эпоху у студентов может возникнуть соблазн копировать и вставлять материалы из различных источников, поскольку информация так легко доступна. Однако они могут не осознавать, чем это чревато. Примером тому может служить использование чужого реферата или задания в качестве своего, копирование целых абзацев из статей в Интернете или даже покупка готовых работ на онлайн-площадках [1].

Студенческий плагиат имеет последствия, выходящие за рамки одного студента. Он нарушает целостность образовательного процесса, подрывает доверие к учебным заведениям и их репутацию, сводит к минимуму усилия прилежных студентов. Плагиат препятствует развитию критического мышления, исследовательских способностей, творческого потенциала – всего того, что необходимо для академического и профессионального роста.

Учебным заведениям и преподавателям следует работать с этой проблемой. Им следует пропагандировать принципы академической честности и давать четкие инструкции по правильному цитированию и оформлению ссылок.

Помимо ручных методов проверки и профилактических мер, учебные заведения могут использовать технологии для выявления и предотвращения плагиата.

Программное обеспечение для обнаружения плагиата позволяет выявлять случаи копирования материалов, сравнивая их с обширной базой данных научных источников, статей и студенческих работ.

Такое программное обеспечение может быть основано как на общепринятых статистических методах, так и на методах машинного обучения.

Классификация методов

В общем случае такие методы идентификации могут быть разделена на две большие группы: автоматизированные и неавтоматизированные [2].

К неавтоматизированным методам идентификации относятся:

1. Экспертная оценка.
2. Сравнение вручную.
3. Сравнение инструментами.

Экспертная оценка подразумевает ручное сравнение текстов особыми консультантами в определённой области. Они анализируют тексты на наличие схожих шаблонов и стилистических особенностей для оценки оригинальности авторства.

Следующий способ включает в себя ручное сравнение текстов путём их сопоставления и поиска схожих предложений и структур.

Третий способ подразумевает собой проверку с использованием особых инструментов сопоставления текстовых фрагментов, которые позволяют облегчить обнаружение заимствований и определить авторов.

К автоматизированным методам идентификации относятся:

1. Анализаторы стиля.
2. Антиплагиатные системы.
3. Методы машинного обучения.

Анализаторы стиля, как правило, представляют собой автоматизированные инструменты, которые в большинстве своём основаны на сопоставлении статистических показателей (слова, фразы, пунктуация). Кроме того, может быть использовано разделения текстов на адаптированные составляющие, статистические показатели которых также составляют авторский стиль. Такими адаптированными составляющими являются *n*-граммы.

Антиплагиатые системы обнаружения представляют собой ПО, которое на основе сформированной текстовой базы (в том числе онлайн) выполняет поиск схожих фрагментов. Анализ использования точных или видоизменённых фрагментов позволяет сделать вывод о предполагаемых заимствованиях [3].

Третья группа подразумевает собой использование методов машинного обучения для создания моделей, позволяющих классифицировать текст на основе закономерностей. Модели обучаются на больших объемах данных, включая оригинальные работы и образцы плагиата. Примерами методов машинного обучения является классификация на основе таких алгоритмов, как метод наивного Байеса, случайный лес и нейронные сети [4,5].

Кроме отдельного использования, данные методы могут комбинироваться различными способами. Благодаря этому, несмотря на увеличение затрат ресурсов, может достигаться более высокая точность за счёт многокритериального подхода.

Поиск таких комбинированных (ансамблевых) методов является одним из направлений исследований.

Ансамблевый метод

Рассмотрим комбинированный метод, позволяющий увеличить точность определения авторства.

Предлагаемый метод включает преобразование выделенных признаков текста в несколько групп, каждая из которых затем исследуется в отдельном классификаторе. Решение о выбранной принадлежности текста будет приниматься после изучения результатов итогового голосования. Данный метод анализирует текст с разных точек зрения, что повышает точность определения авторства.

Для повышения эффективности в качестве составляющих можно использовать значительно отличающиеся друг от друга компоненты, каждый из которых, анализируя текст, предоставляет отдельный независимый результат.

Компоненты могут отличаться как в методе обработки схожих признаков, так и в остальных аспектах.

В качестве примера одним из составляющих данного ансамбля может быть квантовый компонент, который использует отличающиеся от привычных методов машинного обучения признаки и модели. Они основаны на законах квантовой механики с учётом применения к привычным исходным данным в виде текстов.

Для итогового голосование можно использовать мажоритарную схему с весами. Благодарю их конфигурированию можно адаптировать модель к тем или иным условиям, отдавая приоритет определённым компонентам ансамбля.

Схема предлагаемого метода приведена на рисунке 1.

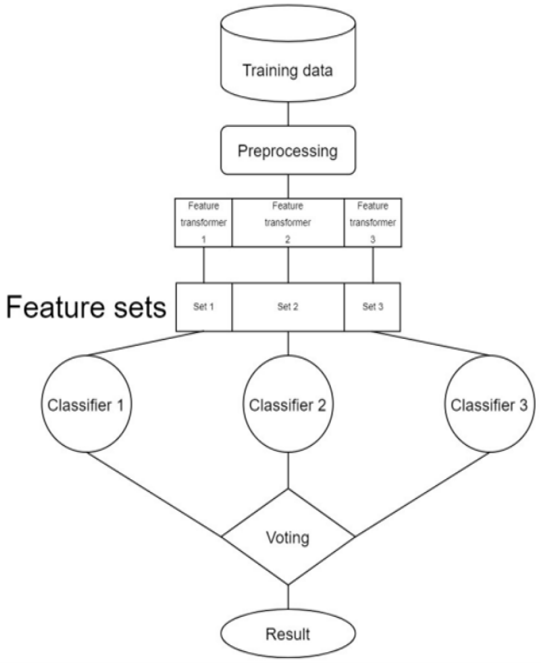

Рисунок 1. Архитектура метода

Эксперимент

Рассмотрим применение различных методов идентификации на практике.

Для этого возьмём данные, представляющих собой различные наборы студенческих работ.

Первый и второй набор представляют собой исправленные работы студентов, в которых намеренно убраны те или иные аспекты, связанные с заимствованиями (изменена лексика, синтаксис и пр.).

Третий набор представляет собой работы студентов в исходном виде.

На каждом из трёх наборов протестируем автоматизированные и неавтоматизированные методы. В качестве результата будет служить степень точность обнаружения плагиата.

Результаты эксперимента приведены в таблице.

Таблица

Результаты эксперимента

Название метода	Точность в наборе 1	Точность в наборе 2	Точность в наборе 3	Автоматизирован?
Ручная проверка	70	72	73	Нет
Экспертная оценка	75	73	78	Нет
Антиплагиатная система 1	52	41	68	Да
Антиплагиатная система 2	48	47	62	Да
Метод ближайших соседей	21	20	24	Да
Случайный лес	53	55	72	Да
Ансамблевый метод	68	71	82	Да

Анализ

Результаты данного описанного эксперимента подтверждают теоретические предположения.

Неавтоматизированные методы (ручная проверка и экспертная оценка) выделяются сравнительно стабильным и высоким уровнем точности. Тем не менее, использование этого метода требует больших временных затрат и накладывает значительные ограничения в масштабируемости.

Автоматизированные методы в силу своей сути могут масштабироваться и обрабатывать значительные объёмы данных в короткие сроки. Тем не менее, как показывает эксперимент, их точность заметно снижается при использовании методов скрытия плагиата.

Тем не менее, ансамблевый метод также показывает высокий уровень точности при сохранении остальных преимуществ автоматизированных методов в скорости и масштабируемости. Вместе с тем ансамблевые методы требует гораздо больше ресурсов, чем остальные методы из их группы.

Заключение

Рассмотренные методы и эксперименты в работе дают представление о текущем инструментарии в вопросе методов идентификации авторства в определении студенческого плагиата.

Выделенные преимущества и недостатки могут служить основой для надлежащего ситуативного применения и направления модификаций.

ЛИТЕРАТУРА

1. **Effects of Plagiarism on Education** [Электронный ресурс]. – Режим доступа: https://classroom.synonym.com/effects-plagiarism-education-6075742.html – Дата доступа: 12.07.2023.

2. **Батура, Т.В.** Формальные методы определения авторства текстов / Т.В. Батура // Вестник НГУ. Сер. Информационные технологии. – 2012. – № 6. – С. 2-3.

3. **How Do Plagiarism Checkers Work?** [Электронный ресурс]. – Режим доступа: https://www.scribbr.com/plagiarism/how-do-plagiarism-checkers-work/ – Дата доступа: 15.07.2023.

4. **M. Khonji, Y. Iraqi and L. Mekouar.** Authorship Identification of Electronic Texts, in IEEE Access, vol. 9, pp. 101124-101146, 2021, doi: 10.1109/ACCESS.2021.3098192.

5. **Authorship Identification Using Neural Networks** [Электронный ресурс]. – Режим доступа: https://community.wolfram.com/groups/-/m/t/1374319 – Дата доступа: 20.07.2023.

REFERENCES

1. **Effects of Plagiarism on Education** [Electronic resourse]. – Access: https://classroom.synonym.com/effects-plagiarism-education-6075742.html – Access date: 12.07.2023.

2. **Batura, T.V.** Formal'nye metody opredeleniya avtorstva tekstov. Vestnik NGU. Ser. Informacionnye tekhnologii, 2012, № 6, pp. 2-3.

3. **How Do Plagiarism Checkers Work?** [Electronic resourse]. – Access: https://www.scribbr.com/plagiarism/how-do-plagiarism-checkers-work/ – Access date: 15.07.2023.

4. **M. Khonji, Y. Iraqi and L. Mekouar,** "Authorship Identification of Electronic Texts," in IEEE Access, vol. 9, pp. 101124-101146, 2021, doi: 10.1109/ACCESS.2021.3098192

5. **Authorship Identification Using Neural Networks** [Electronic resourse]. – Access: https://community.wolfram.com/groups/-/m/t/1374319 – Access date: 20.07.2023.

PARAMONOV A., TRUKHANOVICH I.

AUTHORSHIP IDENTIFICATION METHODS IN STUDENT PLAGIARISM DETECTION

Belarusian state University of Informatics and Radioelectronics
Minsk, Republic of Belarus

In the modern educational context the problem of plagiarism is urgent and requires the development of effective methods of detection and prevention of this phenomenon. The application of authorship identification methods in the field of student plagiarism detection is considered. Different check, detect and analyze plagiarism approaches in various works are investigated. Both classical methods, which include text comparison and similarity search, and modern methods based on machine learning algorithms, as well as their combination and potential modifications, are considered. The advantages and limitations of each method are also discussed, and recommendations are given for choosing one or another approach according to the specific requirements of the research.

Special attention is paid to such modern methods as metadata analysis and the application of neural networks. Stylistic analysis reveals authorial peculiarities such as word choice, preferred wording, and even punctuation. Lexical and syntactic models are used to identify repetitive phrases and structures that may indicate plagiarism. Statistical methods can identify anomalies in the use of words and phrases, and machine learning can create models to calculate the probability of plagiarism based on large amounts of data.

Ultimately, an comparison of authorship identification techniques in the field of student plagiarism detection is provided, which aims to provide valuable information about different approaches and their applicability, and to help researchers and educators develop effective strategies for detecting and preventing plagiarism in educational environments.

Keywords: *machine learning, plagiarism, authorship identification*

Парамонов Антон Иванович, кандидат технических наук, доцент, заведующий кафедрой информационных систем и технологий Института информационных технологий БГУИР.

Anton Paramonov, PhD in Technical Sciences, Head of The Department of Information Systems and Technologies of Institute of Information Technologies BSUIR.

E-mail: a.paramonov@bsuir.by

Труханович Илья Александрович, аспирант Белорусского государственного университета информатики и радиоэлектроники, магистр технических наук.

Ilya Trukhanovich, PhD student of Belarusian State University of Informatics and Radioelectronics, Master of technical sciences.

E-mail: ilya.trukhanovich@gmail.com