УДК 004.89:371.26

Метод автоматизированного выявления аномалий в учебных документах на основе нейросетевого анализа текста

А.А. Назаров, А.П. Преображенский

Воронежский институт высоких технологий, Воронеж, Россия

В статье рассматривается метод автоматизированного выявления аномалий в учебных документах на основе нейросетевого анализа текста. Предложенный подход использует семантический, структурный и стилистический анализ для оценки качества учебных работ. Для этих целей вводится интегральный индекс аномальности, позволяющий количественно оценивать степень отклонения документа от нормативных требований. Рассматриваются принципы вычисления индекса, включая весовые коэффициенты и методы нормализации результатов. Показано, что использование sigmoid-нормализации повышает интерпретируемость итоговых оценок. Предложенный метод может быть применён в системах автоматизированного контроля результатов обучения и интеллектуальной поддержки преподавателей при проверке учебных работ.

Ключевые слова: нейросетевой анализ текста, индекс аномальности документа, sigmoid-нормализация, весовые коэффициенты, функция потерь.

A Method for the Automated Detection of Anomalies in Educational Documents Based on Neural Network Text Analysis

A.A. Nazarov, A.P. Preobrazhenskiy

Voronezh Institute of High Technologies, Voronezh, Russia

This article examines a method for the automated detection of anomalies in educational documents based on neural network text analysis. The proposed approach employs semantic, structural, and stylistic analysis to evaluate the quality of academic assignments. To this end, an integral anomaly index is introduced, enabling a quantitative assessment of the extent to which a document deviates from standard requirements. The principles of calculating this index are discussed, including weighting coefficients and result normalization methods. It is demonstrated that the use of sigmoid normalization enhances the interpretability of the final scores. The proposed method can be applied in systems for the automated monitoring of learning outcomes and for providing intelligent support to instructors during the assessment of academic work.

Keywords: neural network text analysis, document anomaly index, sigmoid normalization, weighting coefficients, loss function.

В рамках задач автоматизации контроля результатов обучения особый интерес представляет метод выявления аномалий в учебных документах, основанный на использовании нейросетевого анализа текстовой информации. Под аномалиями в данном контексте понимаются отклонения учебного документа от ожидаемых структурных, семантических или стилистических характеристик, которые могут свидетельствовать о низком качестве работы, нарушении требований оформления либо наличии несоответствий содержательному наполнению.

В отличие от традиционных подходов, ориентированных на формальные критерии проверки, данный метод предполагает анализ документа на более глубоком уровне – с учётом смысловых и контекстных особенностей текста. Это позволяет

рассматривать учебный документ не как набор отдельных элементов, а как целостную текстовую структуру, обладающую внутренней логикой и связностью.

Основой метода является преобразование текстового документа в векторное представление с использованием нейросетевых моделей обработки естественного языка. Каждый документ отображается в многомерное пространство признаков, где учитываются семантические связи между словами, структура предложений и общая тематическая направленность текста. В результате формируется «нормативный профиль» учебных работ, соответствующих установленным требованиям [1].

Далее осуществляется сравнение анализируемого документа с эталонным распределением. Исследуется резкое изменение стиля повествования, слишком краткие или чересчур объёмные разделы документа.

Отдельное внимание уделяется соответствию содержания текста озвученной теме. Нейросетевые модели способны выявлять такие расхождения, анализируя смысловое соответствие отдельных фрагментов текста.

Практическая реализация рассматриваемого метода возможна с применением предварительно обученной трансформерной нейросети, такой как prajjwal1/bert-tiny. Такая модель формирует эмбеддинги, то есть векторные представления текста, и сравнивает их с эмбеддингом эталона. Работать данный трансформер может локально на компьютере при ограниченных вычислительных ресурсах [2].

Результатом применения метода является формирование интегральной оценки аномальности учебного документа, которая может использоваться как дополнительный критерий при автоматизированной проверке. Такая оценка позволяет выделять документы, требующие повышенного внимания со стороны преподавателя, тем самым оптимизируя распределение его рабочей нагрузки.

Таким образом, метод автоматизированного выявления аномалий в учебных документах обеспечивает переход от формальной проверки к интеллектуальному анализу содержания, что способствует повышению эффективности управления функцией контроля результатов образовательной деятельности.

Для количественной оценки степени отклонения учебного документа от нормативного представления вводится интегральный индекс аномальности, отражающий совокупное влияние семантических, структурных и стилистических характеристик текста [3].

$$AI(D) = \alpha \cdot Ssem(D) + \beta \cdot Sstr(D) + \gamma \cdot Ssty(D), \quad (1)$$

где AI(D) – индекс аномальности документа, Ssem(D) – семантическое отклонение документа, Sstr(D) – структурное отклонение документа, Ssty(D) – стилистическое отклонение, α, β, γ – весовые коэффициенты значимости $(\alpha + \beta + \gamma = 1)$.

Данный показатель позволяет формализовать процесс выявления отклонений и использовать его в задачах автоматизированного принятия управленческих решений в рамках системы контроля результатов обучения.

При вычислении данного показателя могут получиться различные значения, как очень большие, так и очень маленькие. Для приведения их к виду, удобному для дальнейшей интерпретации, производят нормализацию или масштабирование этих данных. Нормализация может проводиться различными методами. При работе с нейросетями для вероятностной интерпретации чаще всего используется sigmoid-нормализация:

$$AI(D) = \frac{1}{1+e^{-AI(D)}}, \quad (2)$$

где $AI(D)$ – нормализация или порог принятия решений.

При вычислении порога принятия решений по формуле (2) получается число в диапазоне от нуля до единицы. При значении, меньшем 0,3, проверяемый документ считается соответствующим норме. Если полученное значение попадает в интервал от 0,3 до 0,7, то текст требует дополнительной проверки. Значение порога большее 0,7, считается высокой аномалией. Документ должен быть кардинально переработан.

Весовые коэффициенты α, β, γ в формуле (1) могут определяться на основе экспертной оценки значимости семантических, структурных и стилистических характеристик документа с соблюдением условия нормировки. Например, если семантика важнее, то значение $\alpha = 0{,}5$. Тогда коэффициент структуры документа $\beta = 0{,}3$, а весовой коэффициент стиля $\gamma = 0{,}2$. В случае наличия обучающей выборки (например, оценок преподавателей) параметры могут быть оптимизированы с использованием методов машинного обучения, направленных на минимизацию функции ошибки между рассчитанным индексом аномальности и эталонной оценкой качества документа. В этом случае используется линейная регрессия или градиентный спуск.

Другим научно обоснованным методом подбора весовых коэффициентов является метод оптимизации, при использовании которого происходит минимизации функции потерь:

$$L = (AI(D) - y)^2, \quad (3)$$

где y – эталонная оценка преподавателя.

Процесс обучения модели с минимизацией функции потерь можно описать как непрерывный цикл взаимодействия между предсказанием и корректировкой параметров. Сначала модель принимает входной учебный документ и преобразует его во внутреннее представление, на основе которого формируется прогнозное значение индекса аномальности $AI(D)$. Затем это значение сравнивается с эталонной оценкой y, заданной либо экспертом, либо полученной из обучающей выборки, после чего вычисляется величина ошибки, отражающая степень расхождения между предсказанием и фактическим результатом. На основе этой ошибки определяется направление изменения параметров модели, то есть устанавливается, какие внутренние веса необходимо скорректировать для уменьшения отклонения в будущем. Далее параметры модели обновляются с учётом вычисленных поправок, и процесс повторяется многократно на большом количестве примеров, благодаря чему система постепенно «обучается» более точно оценивать учебные документы и снижать величину ошибки.

Обновление параметров происходит по формуле обучения:

$$\theta := \theta - \eta \cdot \frac{\partial L}{\partial \theta}, \quad (4)$$

где θ – параметры модели, η – скорость обучения. В процессе обучения параметр θ обновляется и становится равным старому значению минус шаг корректировки. Скорость обучения определяет величину корректировки параметров модели на каждом шаге оптимизации функции потерь. Малые значения обеспечивают устойчивое, но медленное обучение, тогда как большие значения ускоряют процесс, но могут привести к нестабильности сходимости. В современных нейросетевых системах часто

применяются адаптивные методы оптимизации, позволяющие динамически изменять скорость обучения в зависимости от характера функции потерь.

Ниже представлена архитектура процесса вычисления индекса аномальности документа в виде блок-схемы:

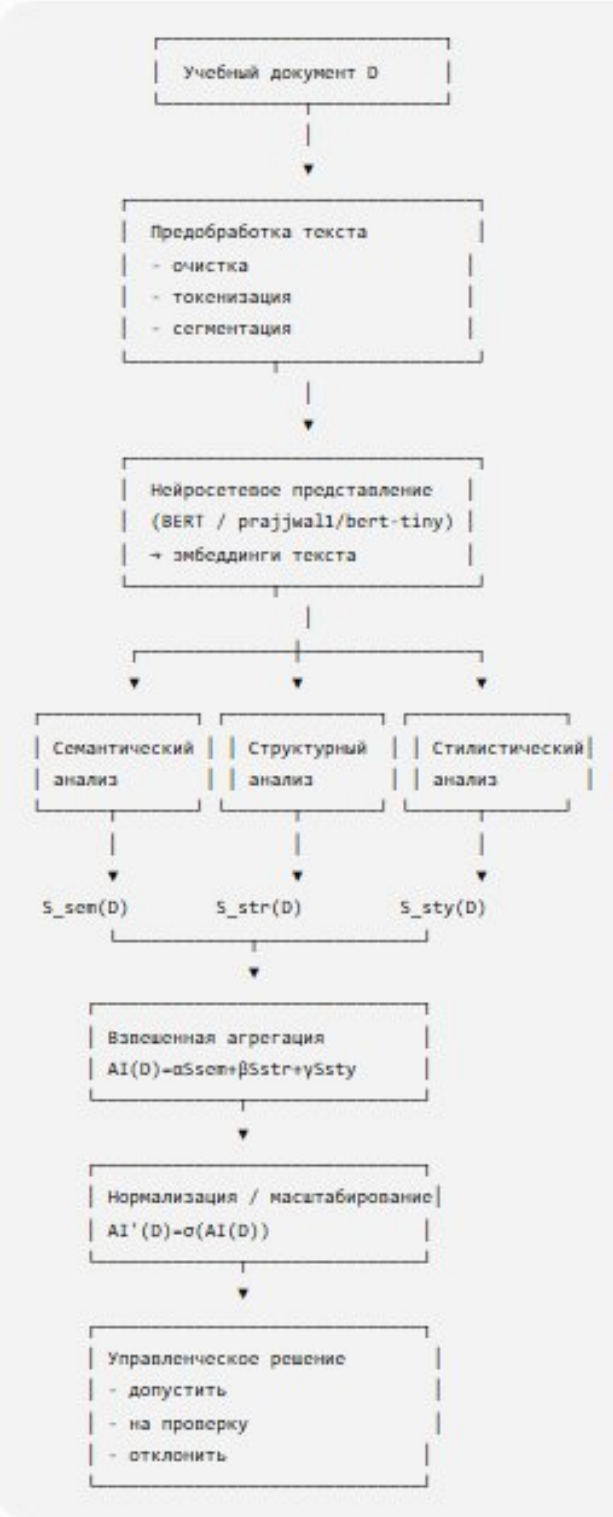

Рисунок. Общая структура процесса вычисления индекса аномальности

Сначала учебный документ проходит этап предварительной обработки, включающий очистку текста, токенизацию и разбиение на смысловые сегменты. Далее полученные токены преобразуются в эмбеддинги в многомерном пространстве признаков. В рассматриваемой нейросетевой модели используется 128-мерное пространство [2].

На следующем этапе текст анализируется параллельно по трём направлениям. Первый компонент анализа – семантический, второй – структурный и, наконец, третий – стилистический. Для каждого вида анализа вычисляется частный показатель отклонения от нормы. Затем все полученные значения умножаются на весовой коэффициент и суммируются между собой. Использование весовых коэффициентов позволяет учитывать различную значимость каждого типа отклонения. Более существенный вид несоответствия имеет больший вес.

Следующим этапом выполняется нормировка полученного значения. Этот показатель уже используется для принятия решения о соответствии текста установленным нормам [4].

Предложенный подход позволяет учитывать не только соответствие текста установленным требованиям, но и устанавливает логическую целостность, связность, а также его связь с заявленной темой. Рассчитанный индекс аномальности может применяться в автоматизированных системах принятия решений.

СПИСОК ИСТОЧНИКОВ

1. Шамарина Е.А. Сентимент-анализ на основе нейросетей-трансформеров / Е.А. Шамарина, А.И. Гусева, В.С. Киреев // XXV Международная научно-техническая конференция «Нейроинформатика-2023»: Сборник научных трудов, Москва, 23–27 октября 2023 года. – М.: Национальный исследовательский ядерный университет «МИФИ», 2023. – С. 292–301.

2. Усен Е.Б. Применение нейросетевых моделей для семантического анализа публикаций в социальных сетях / Е.Б. Усен // Вестник науки. – 2025. – Т. 4, № 2 (83). – С. 480–485.

3. Иванов В.М. Интеллектуальные системы: учебное пособие для вузов / В.М. Иванов; под научной редакцией А.Н. Сесекина. – М.: Издательство Юрайт, 2025. – 88 с.

4. Кудрявцев В.Б. Интеллектуальные системы: учебник и практикум для вузов / В.Б. Кудрявцев, Э.Э. Гасанов, А.С. Подколзин. – 2-е изд., испр. и доп. – М.: Издательство Юрайт, 2025. – 165 с.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Назаров Антон Андреевич, аспирант, Воронежский институт высоких технологий, Воронеж, Россия.

Преображенский Андрей Петрович, доктор технических наук, профессор, Воронежский институт высоких технологий, Воронеж, Россия.