Мультимодальный анализ цифрового следа студентов: прогнозирование успеваемости и рисков отчисления

Горшков Сергей Сергеевич

Приглашенный преподаватель

Факультет компьютерных наук, Высшая школа экономики, Москва, Россия

E-mail: serggorsar@yandex.ru

Научный руководитель *— Игнатов Дмитрий Игоревич*

Современные образовательные учреждения нуждаются в новых подходах к прогнозированию академической успеваемости и выявлению студентов с высоким риском отчисления. Традиционные методы, основанные на академических показателях, не учитывают поведенческие и социальные аспекты. В данном исследовании представлен новый мультимодальный подход, использующий анализ цифрового следа студентов в социальной сети ВКонтакте, включая текстовые, визуальные и аудиоданные, а также информацию о графе дружбы и подписках студентов на сообщества. Метод интегрирует передовые модели машинного обучения, включая Sentence-BERT, CLIP, BERTopic, HDBSCAN, K-Means++ и CatBoost, а также включает механизм внимания и стекинг моделей для повышения точности предсказаний [1].

Анализ текстового контента осуществляется с помощью BERTopic, который позволяет выделить 87 тематических направлений, отражающих ключевые интересы студентов. Помимо тематического моделирования [2], были применены модель для выявления эмоционального состояния студентов и анализ тональности, а также анализ семантической сложности текстов. Для получения эмбеддингов BERT-подобная модель была дообучена на контенте из ВКонтакте. Визуальный анализ основан на моделях Places365 и CLIP, позволяющих классифицировать сцены и определять контекст изображений. Для анализа музыкальных предпочтений использовалась классификация по 20 основным жанрам, позволяющая связать музыкальные интересы с академической успеваемостью.

Одним из ключевых элементов исследования стало применение механизма внимания для динамического взвешивания признаков. Эта методология позволила автоматически определять значимость различных факторов при прогнозировании успеваемости. Веса, присваиваемые тематическим моделям, эмоциям, лексической сложности и визуальному контенту, варьировались в зависимости от модели

и целевой переменной. Абляционное исследование подтвердило, что удаление любого из типов данных (текста, изображений или графа связей) снижает точность предсказаний, что доказывает важность мультимодального подхода.

Предложенный подход продемонстрировал высокую точность прогнозирования риска отчисления (ROC-AUC = 0.802) и идентификации наиболее успевающих студентов. В частности, алгоритмы успешно предсказывали, относится ли студент к группе с идеальным GPA (5.0), с ROC-AUC около 0.78. Для классификации студентов с высокой академической успеваемостью использовался стекинг моделей с вниманием к важности отдельных признаков, где линейная регрессия, случайные леса и градиентные бустинг-алгоритмы дополнялись трансформерными архитектурами. SHAP-анализ показал, что наибольшее влияние оказывают семантические характеристики текстов, принадлежность к образовательным сообществам, а также сложность используемого языка.

Результаты исследования могут быть использованы университетами для раннего выявления студентов, находящихся в зоне риска, а также для персонализации образовательных траекторий. Помимо образовательного сектора, методология применима в финансовых организациях (оценка риска по образовательным кредитам) и HR-аналитике для выявления перспективных студентов. В дальнейшем возможны дополнения в виде лонгитюдного анализа цифрового следа студентов, включения данных с других платформ и расширения метода на более широкие социологические исследования. Таким образом, предложенная модель представляет собой мощный инструмент для образовательных, социальных и финансовых организаций, позволяя прогнозировать академическую успеваемость и оптимизировать образовательные стратегии.

Литература

1. Gorshkov S. S. et al. Identifying Top-Performing Students via VKontakte Social Media Communities Using Advanced NLP Techniques // IEEE Access. 2025. Vol. 13, pp. 962-979.
2. Gorshkov S. et al. Using topic modeling for communities clusterization in the VKontakte social network // International Journal of Open Information Technologies. 2021. Vol. 9, №. 5, pp. 12-17.