Датасет для анализа русскоязычных отзывов на MOOK, извлеченных с платформы Stepik

Юлия Дюличева

Статья поступила в редакцию в июне 2022 г.

Дюличева Юлия Юрьевна — кандидат физико-математических наук, доцент кафедры прикладной математики, ФГАОУ ВО «Крымский федеральный университет имени В.И. Вернадского». Адрес: 295007, Симферополь, просп. Академика Вернадского, 4. E-mail: dyulicheva_yu@mail.ru, ORCID: https://orcid.org/0000-0003-1314-5367

Аннотация

В статье приведен обзор направлений исследований в области анализа образовательных данных на основе методов обработки естественного языка и соответствующих датасетов, из которого, в частности, становится очевиден недостаток датасетов для анализа русскоязычных отзывов на MOOK. На основе скрапинга отзывов с платформы *Stepik* сформирован датасет из 5721 русскоязычного отзыва на MOOK по математике, программированию, биологии, химии и физике. Выполнено исследование русскоязычных отзывов из датасета на основе описательной статистики, частотного анализа униграмм и биграмм, сентиментного анализа с помощью *python*-библиотеки *dostoevsky*, продемонстрировавшего 74%-ную точность классификации по классам тональности на основе взвешенной метрики *F1-score*. С помощью анализа униграмм выявлены описательные характеристики курсов с учетом тональности, а анализ биграмм позволил получить описания различных аспектов учебного контента и трудностей, с которыми столкнулись слушатели при изучении MOOK. По результатам сентиментного анализа можно судить о преобладании в изучаемом датасете позитивных и нейтральных отзывов на MOOK. Датасет размещен в открытом доступе на платформе *Mendeley Data* и будет полезен специалистам в области анализа текстовых данных и разработки инструментов учебной аналитики.

Ключевые слова

MOOK, датасет, частотный анализ униграмм и биграмм, сентиментный анализ, *python*-библиотека *dostoevsky*, nltk, *pymorphy2.*

Для цитирования

Дюличева Ю.Ю. (2022) Датасет для анализа русскоязычных отзывов на MOOK, извлеченных с платформы Stepik. *Вопросы образования / Educational Studies Moscow,* № 4, сс. 298–321. https://doi.org/10.17323/1814-9545-2022-4-298-321

Dataset for Analysis of Russian-Language Reviews on MOOCs Extracted from Stepik

Yulia Dyulicheva

Yulia Yu. Dyulicheva, Candidate of Sciences in Physics and Mathematics, Associate Professor of the Department of Applied Mathematics, V.I. Vernadsky Crimean Federal University. Address: 4 Akademika Vernadskogo Ave, 295007 Simferopol, Russian Federation. E-mail: dyulicheva_yu@mail.ru, ORCID: https://orcid.org/0000-0003-1314-5367

Abstract The article provides an overview of datasets and research areas in the field of educational data analysis based on natural language processing methods. The overview demonstrates the lack of datasets for the analysis of Russian-language reviews on MOOCs. Based on the scraping of reviews from the Stepik platform, a dataset of 5721 Russian-language reviews for MOOCs in mathematics, programming, biology, chemistry and physics was formed. A study of Russian-language reviews from the dataset was carried out based on descriptive statistics, frequency analysis of unigrams and bigrams, sentiment analysis using the dostoevsky python library with weighted F1-score for estimation accuracy of classification by sentiment as 74%. The descriptive characteristics of courses with respect to sentiments were detected based on unigrams analysis, the description of different aspects of learning content and difficulties encountered by students in learning MOOCs were detected based on bigrams analysis. The results of the sentiment analysis demonstrate the predominance of positive and neutral reviews of MOOCs in the studied dataset. The dataset is placed in the public domain Mendeley Data and will be useful to specialists in the field of text data analysis and the development of learning analytics tools.

Keywords MOOC, dataset, frequency analysis of unigrams and bigrams, setiment analysis, python-library dostoevsky, nltk, pymorphy2

For citing Dyulicheva Yu.Yu. (2022) Dataset dlya analiza russkoyazychnykh otzyvov na MOOK, izvlechennykh s platformy Stepik [Dataset for Analysis of Russian-Language Reviews on MOOCs Extracted from Stepik]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 4, pp. 298–321. https://doi.org/10.17323/1814-9545-2022-3-298–321

Стремительный рост количества массовых открытых онлайн-курсов (МООК) и систем управления обучением, активное использование социальных сетей и мессенджеров для организации дистанционного обучения приводят к накоплению больших данных как о самом обучающемся и его успеваемости, так и о его настроениях, мнениях и предпочтениях. Численность слушателей МООК по всему миру превышает 100 млн человек, а накапливаемые данные оцениваются терабайтами [Shah, 2019]. Для обработки и анализа больших объемов образовательных данных требуются специальные подходы — от краудсорсинга для создания размеченных датасетов до методов машинного обучения.

Анализ образовательных данных и учебная аналитика — два наиболее динамично развивающихся сегодня направления исследований, активно использующих методы машинного обучения и искусственного интеллекта. Современные методы анализа образовательных данных и сервисы учебной аналитики позволяют эффективно решать задачи анализа и прогнозирования успеваемости обучающихся, их настроений, получения обратной связи и моделирования поведения обучающихся. На основе этих данных разрабатываются стратегии персонализации учебного процесса, совершенствуются методики обучения и системы поддержки обучающихся [Sarıyalçınkaya et al., 2021]. Накопление больших данных в образовании выдвигает на повестку дня необходимость разработки новых методов предобработки и анализа данных, в том числе с использованием глубоких нейронных сетей и методов анализа естественного языка и изображений.

Анализ данных MOOK — сложная задача, поскольку требует не только обработки данных о контенте курсов, отзывов и комментариев, но и понимания социальных условий использования MOOK и особенностей поведения обучающихся. Создание инструментов аналитики MOOK направлено на выявление закономерностей на основе анализа данных, поступающих из разных источников обратной связи с обучающимися и на получение своевременного отклика от инструктора курса для предотвращения оттока обучающихся. Например, методология 3S учебной аналитики основана на изучении сентимента (*sentiment*), социального аспекта (*social aspect*) и навыков (*skills*) при анализе сообщений и взаимодействия между обучающимися на форумах MOOK [Moreno-Marcos et al., 2019]. Для анализа сложного датасета OULAD, содержащего разнородные образовательные данные, предлагается применять фреймверк *Flask*, модель распределенных вычислений *MapReduce* и *python*-библиотеки [Siddique, 2020]. Система учебной аналитики OXALIC разработана для мониторинга лог-файлов платформы edX и построения графов взаимодействия обучающихся на основе анализа потока кликов, просмотра видео, взаимодействия на форумах и т.п. [Khalil, Belokrys, 2020]. Существует методология учебной аналитики на основе выявления закономерностей из данных о просмотре видеоконтента MOOK [Shridharan et al., 2018]. На основе модели ARIMA строятся прогнозы продолжительности обучения на онлайн-курсах [Sun et al., 2021]. Фреймверк MORF разработан для исследования общих и уникальных закономерностей различных MOOK, представленных в виде продукционных правил — если «условие», то «событие» — и позволяющих анализировать поведение слушателей курсов на основе времени, проведенного на форумах, и особенностей

сообщений, включающих анализ конкретных слов, биграмм и триграмм, значимых и сложных слов и т.п. [Andres et al., 2018]. Аналитическая система *MOOCad* направлена на выявление аномалий в образовательных данных и поведении обучающихся и обеспечивает визуализацию кластеров данных МООК с учетом аномалий [Mu et al., 2019]. Визуальные системы учебной аналитики позволяют наглядно представить выявленные закономерности — от влияния на образовательные результаты демографических факторов до роли личностных характеристик слушателей МООК в успешности освоения курса [Li et al., 2017]. В частности, с помощью инструмента визуальной учебной аналитики *MessageLens* инструкторы получают представление о взаимодействии обучающихся на форумах МООК и об их отношении к курсам [Wong, Zhang, 2018]. При разработке сервисов учебной аналитики широко используется анализ последовательностей кликов, просмотров видео, тех или иных наборов активностей обучающихся на платформах МООК. Инструмент визуальной учебной аналитики ViSeq позволяет получить представление о том, в какой последовательности выстраивают освоение курса как разные группы обучающихся, так и отдельные слушатели МООК [Chen et al., 2020]. Инструмент учебной аналитики MOOC-ASV для визуализации результатов взаимодействия слушателей МООК с обучающими видео разработан с целью улучшения качества обучающих видео [Mubarak, Ahmed, Cao, 2021]. Инструмент учебной аналитики *PeakVizor* определяет пики просмотров видеоконтента для разных групп слушателей онлайн-курсов [Chen et al., 2016]. Для мониторинга платформ МООК и данных об обновлении онлайн-курсов разработана система *MOOClink*, она помогает слушателям в поиске актуальных МООК [Dhekne, Bansal, 2018].

Одним из инновационных направлений в развитии МООК является применение учебного контента на основе иммерсивных технологий. Инструмент учебной аналитики *GazeMOOC* позволяет оценивать вовлеченность слушателей МООК в учебный процесс за счет интеграции учебного контента в расширенной реальности (XR) [Wang et al., 2021].

Интеллектуальный анализ данных применяется также для прогнозирования успеваемости обучающихся на основе данных об их кликах по разным типам контента в *Moodle* с помощью таких методов машинного обучения, как метод ближайших соседей, метод опорных векторов, решающее дерево, случайный лес и т.п. [Shrestha, Pokharel, 2021]. Создание такого рода аналитических инструментов позволяет разрабатывать инновационные подходы в образовании, основанные на анализе больших данных и выявлении скрытых закономерностей.

1. Обзор датасетов в области анализа образовательных данных МООК

Количество датасетов, содержащих образовательные данные, в последнее время заметно растет. На платформе *Kaggle* (https://www.kaggle.com/) можно найти более 4 тыс. датасетов по теме «образование» — от исследования успеваемости и влияния пандемии COVID-19 на уровень адаптации к онлайн-образованию до решения задачи распознавания лиц обучающихся в масках, вопросов, относящихся к сфере ментального здоровья, и анализа влияния игр на подростков. Специальная платформа-репозиторий *MOOCCube* содержит данные более чем о 700 МООК, о слушателях этих курсов и их контенте [Yu et al., 2020]. *Python*-пакет с открытым исходным кодом *edx2bigquery* позволяет извлекать с помощью запросов различные данные о курсах МООК на платформе edX [Lopez et al., 2017].

Как видно из табл. 1, число датасетов, предоставляющих для анализа открытые данные о курсах МООК и доступных для скачивания в полном объеме, ограничено, а на русском языке они практически отсутствуют. Постановка конкретной исследовательской задачи в области анализа образовательных данных влечет за собой поиск необходимого датасета, а при его отсутствии — скрапинг открытых данных с платформ МООК по тематике исследования.

Таблица 1. **Примеры некоторых датасетов о МООК**

Датасет	Платформа	Ссылка	Краткое описание
Данные ЭЭГ обучающихся	*Kaggle*	https://www.kaggle.com/datasets/wanghaohan/confused-eeg	Датасет содержит результаты ЭЭГ 10 студентов колледжа после просмотра видеофрагментов различных МООК, в том числе по незнакомым для обучающихся дисциплинам, при этом степень их замешательства при просмотре видео оценивалась по 7-балльной шкале
Данные о курсах *Udemy*	*Kaggle*	https://www.kaggle.com/code/andrewmvd/udemy-courses-getting-started/comments	Датасет содержит следующие данные о 3678 курсов: id курсов, их названия, число подписчиков, количество отзывов, лекций, длительность видеоконтента, цена, тематика (web-разработка, бизнес, финансы, музыкальные инструменты, графический дизайн) и т.п.
Данные о курсах *Coursera* и *Udacity*	*Kaggle*	https://www.kaggle.com/datasets/ayushbatra/online-mooc	Датасет содержит следующие данные: название курса, url-адрес, данные о партнерах курса, краткое описание курса, уровень сложности, навыки
Данные о курсах *Courcera*	*Kaggle*	https://www.kaggle.com/datasets/imuhammad/course-reviews-on-coursera	Датасет содержит следующие данные о 622 курсах на платформе *Coursera*: название курса, url курса, id курса, отзывы, автор и дата создания курса, рейтинг

Датасет	Платформа	Ссылка	Краткое описание
Данные о курсах на Edx	*Kaggle*	https://www.kaggle.com/datasets/imuhammad/edx-courses	Датасет содержит следующие данные о 976 курсах на платформе Edx: название курса, учебная программа и краткое содержание курса, тип и число записавшихся обучающихся, уровень, язык, наличие субтитров
Размеченные данные о курсах MOOK	*Kaggle*	https://www.kaggle.com/datasets/hongliyuan/mooc-review1	Датасет содержит 14 774 отзыва на курсы MOOK с указанием сентимента (отрицательный, положительный, нейтральный)
Данные стенограмм видеолекций MOOK с платформы *Coursera*	*Mendeley Data*	https://data.mendeley.com/datasets/xknjp8pxbj/1	Датасет содержит стенограммы 12 032 видеолекций к MOOK

Целью данной работы является создание датасета для анализа образовательных текстовых данных на русском языке и анализ датасета методами описательной статистики и обработки естественного языка. Для достижения цели исследования реализован скрапинг русскоязычных отзывов на MOOK с платформы Stepik, предоставлен открытый доступ к полученному датасету на платформе *Mendeley Data*, выполнено исследование текстовых образовательных данных на основе анализа униграмм и биграмм с учетом их частотности и тональности и предложены направления дальнейшего изучения датасета.

2. Подходы к анализу образовательных данных MOOK

Платформы MOOK предоставляют большие объемы текстовых образовательных данных — от сообщений на форумах до описания содержания онлайн-курсов, учебных программ, субтитров к обучающим видео и т.п. В частности, представляет интерес задача анализа потоков кликов слушателей курсов и других данных в лог-файлах и соотношения субтитров к обучающим видео и комментариев на форумах MOOK на основе каскадной модели [Jiang et al., 2017], а также применение нейронных сетей для извлечения последовательностей ресурсов, упоминаемых в комментариях форумов [An et al., 2019]. С анализом текстов непосредственно связано и такое направление технологических разработок, как создание интеллектуальных чат-ботов, автоматизирующих процесс ведения диалога со слушателями MOOK. Этапы разработки интеллектуального чат-бота для MOOK с возможностью распознавания и генерации речи обсуждаются в [Lim, Goh, 2016].

Выделим основные направления анализа образовательных текстовых данных.

2.1. Кластеризация образовательных текстовых данных

Кластеризация позволяет выделить группы близких по содержанию онлайн-курсов, схожих лекций или видеофрагментов, группы студентов с одинаковыми предпочтениями, сходными интересами или проблемами при изучении MOOK и т.п. Для извлечения кластеров лекций со схожим учебным контентом предлагается построение графа предшествования и просмотр альтернативных путей в графе при выборе индивидуальной траектории обучения [Alsaad, Alawini, 2020]. Анализ сообщений на форумах позволяет реализовать таксономию Блума для выделения ключевых слов и определения основных типов взаимодействий между слушателями MOOK [Wong et al., 2015]. Кластеризацию учебных ресурсов MOOK и выработку рекомендаций в отношении учебных ресурсов предлагается осуществлять на основе анализа ассоциативных правил [Koffi et al., 2021]. Схожие посты на дискуссионных форумах MOOK кластеризованы с помощью метода *k*-медиан, каждый кластер описан вручную, а затем построены описания на основе тематического моделирования [Ezen-Can et al., 2015]. В частности, авторы описали кластеры постов, включающих отношение к системе образования, к идеям слушателей, а также одобрение слушателями тех или иных содержательных и технологических решений или несогласие с ними.

2.2. Анализ тональности образовательных текстовых данных и задача классификации текстов

Исследование настроений как общества в целом, так и отдельных обучающихся, разработка методов противодействия буллингу и троллингу на основе анализа комментариев и сообщений в социальных медиа позволяют получить обратную связь от обучающихся и сформировать благоприятную атмосферу общения в образовательных онлайн-сообществах. Анализ тональности создает основу для улучшения взаимодействия между участниками на асинхронных дискуссионных досках онлайн [Thoms et al., 2017], понять настроения обучающихся и их мнения об инструкторе курса и обучающем контенте и т.п. Задачу классификации отзывов на MOOK предлагается решать, например, на основе векторизации с помощью TF-IDF и извлечения N-грамм с учетом тональности извлекаемых ключевых слов [Singh et al., 2021]. При разработке дизайна и формировании контента MOOK учитываются предпочтения пользователей, выявленные с помощью сентиментного анализа отзывов [Dina, Yunardi, Firdaus, 2021]. Эмпирически обоснована эффективность векторизации текстов отзывов MOOK на основе подхода GloVe и сетей LSTM для классификации текстов сообщений по тональности, достигнутая точность распознавания составила 95,8% [Onan, 2020]. Сентиментный анализ отзывов обучающихся на уровне аспектов — характеристик онлайн-курсов, к которым

слушатели выражают свое отношение, — как правило, выполняется в два этапа. На первом этапе решается задача обучения без учителя и выделяются аспекты. На втором этапе решается задача классификации по тональности на основе аспектов. Выделяются, например, кластеры отзывов, относящихся к трем аспектам: курс, инструктор и оценивание, а затем решается задача классификации по двум классам (положительные и отрицательные отзывы) с помощью сети долгой краткосрочной памяти (LSTM) и сверточной нейронной сети (CNN) [Kastrati, Imran, Kurti, 2020].

2.3. Прогнозирование успеваемости на основе анализа текстовых образовательных данных

Для предсказания успешности окончания MOOK исследователи из американских университетов использовали гибридный подход к выбору признаков: они сочетали характеристики, описывающие активность обучающихся, и данные, извлеченные с помощью анализа комментариев форума, и достигли 78%-ной точности прогноза [Crossley et al., 2016]. На форумах MOOK содержится множество сообщений и отзывов на MOOK, в этом потоке очень трудно отследить реплики слушателей, испытывающих трудности, нуждающихся в поддержке, теряющих интерес к курсу. Для своевременного выявления обучающихся, у которых возникли проблемы в процессе прохождения курса, предложен формальный подход к анализу текстов сообщений на основе байесовского глубокого обучения [Yu et al., 2021].

2.4. Тематическое моделирование образовательных текстовых данных

Применение гибридных подходов для извлечения признаков из комментариев форума способствует повышению качества классификации данных. В частности, протестирована возможность извлечения признаков на основе латентного размещения Дирихле (LDA) при анализе тем из сообщений форума [Yu et al., 2021], предложен способ извлечения тем из неструктурированных отзывов для понимания интересов слушателей MOOK и разработки персонализированной системы рекомендаций курса или контента [Liu et al., 2017]. Для изучения причин, по которым слушатели выбирали те или иные курсы по виртуальной реальности на платформе *FutureLearn*, использовано тематическое моделирование с помощью LDA [Onah, Pang, 2021]. Еще одним примером применения латентного размещения Дирихле является фреймверк для автоматической генерации тем и их разметки с учетом схожести термов из тем и ключевых слов, извлекаемых из дискуссий и контента курса [Atapattu, Falkner, 2016]. Возможности применения тематического моделирования для понимания того, какие слушатели MOOK «выживут» до конца курса, обсуждаются в [Yao et al., 2021]. Тематически

ческое моделирование используется для сопоставления тем, извлеченных из материалов курса и обсуждений на форумах, с его помощью, например, удается выявить темы для классификации видов активностей слушателей курсов на форумах [Ramesh et al., 2014]. Структурное тематическое моделирование дает возможность оценить качество МООК на основе комментариев форума [Reich et al., 2014].

2.5. Анализ состояния ментального здоровья слушателей МООК

Выявление проблем, возникающих у слушателей при изучении онлайн-курсов, важно как с точки зрения успешности освоения курса, так и в интересах сохранения ментального здоровья обучающихся: например, при отсутствии своевременного реагирования на тревожность и страхи, связанные с изучением учебной дисциплины, у обучающегося могут развиться фобии, препятствующие изучению этой дисциплины и освоению профессий, связанных с ней. Определить причины математической тревожности помогает кластеризация отзывов на МООК по математике [Дюличева, 2021].

2.6. Разработка рекомендательных систем МООК на основе анализа образовательных текстовых данных

Рекомендательные системы МООК представляют собой алгоритмы выработки персонализированных предложений курсов на основе анализа предпочтений, степени удовлетворенности курсом, типичных трудностей и т.п. Примером рекомендательной системы, позволяющей на основе анализа отзывов на МООК получить представление о доступности онлайн-среды обучения, является *YourMOOC4all* [Iniesto, Rodrigo, 2019]. Разработана система рекомендаций МООК на основе выделения слушателей со схожими интересами и предпочтениями с помощью тематического моделирования средствами LDA сообщений на форумах [Zarra и др., 2018].

3. Описание датасета русскоязычных отзывов на МООК и результаты его анализа

Обучающиеся часто сталкиваются с трудностями при изучении естественнонаучных дисциплин: они могут испытывать страх перед этими дисциплинами, считая их очень сложными, изучаемый материал может быть для них скучным, они могут не видеть связи между изучаемой теорией и практикой. Для исследования на основе отзывов слушателей трудностей, с которыми сталкиваются обучающиеся при изучении МООК, причин их неудовлетворенности определенным МООК, для выявления их отношения к тем или иным аспектам МООК необходимо создать датасет с отзывами на МООК по исследуемой тематике.

В данном исследовании сформирован, а затем проанализирован на основе сентиментного анализа и частотного ана-

лиза униграмм и биграмм датасет русскоязычных отзывов на МООК, извлеченных с платформы *Stepik*. Скрапинг отзывов частично автоматизирован на основе библиотеки *scrapy* и *stepik api*, частично выполнялся вручную. Всего с платформы *Stepik* извлечен 5721 отзыв на 102 массовых открытых онлайн-курса, как показано в табл. 2. Сформированный датасет текстов отзывов на МООК находится в открытом доступе на платформе *Mendeley Data* [Dyulicheva, 2022].

Таблица 2. **Ключевые слова запросов для скрапинга данных с платформы *Stepik***

Ключевые слова, которые использовались в запросах	Число извлеченных курсов на платформе *Stepik*	Число извлеченных отзывов на платформе *Stepik*
Математика, теория вероятностей, статистика	35	2290
Физика, оптика, механика	13	58
Биология, генетика, ботаника, анатомия, филогенетика	22	856
Химия	18	237
Программирование, *Python*, C++, C#	14	2280

В табл. 3 приведены некоторые статистические показатели, описывающие среднее число токенов (слов) в комментариях по разным дисциплинам, среднеквадратичное отклонение, минимальное и максимальное число токенов после удаления знаков пунктуации, 25-процентиль, 50-процентиль (медиану) и 75-процентиль. Из табл. 3 видно, например, что по математическим дисциплинам в исследуемом датасете минимальное число слов в отзывах — 1, а максимальное — 398.

Таблица 3. **Описательная статистика по извлеченным комментариям**

Дисциплины	Показатели						
	mean	std	min	25%	50%	75%	max
Математика	19,420914	29,075746	1	4	10	24	398
Программирование	23,767473	36,552573	1	5	13	27	695
Физика	34,5	87,693895	1	4	11,5	26,75	538
Биология	18,776995	24,743327	1	4	9,5	24	203
Химия	15,5	19,696743	1	3	8	20	113

Примеры случайно извлеченных отзывов по исследуемым дисциплинам представлены в табл. 4.

Таблица 4. **Примеры случайно извлеченных отзывов**

Дисциплины	Примеры отзывов (сохранена авторская орфография отзывов)
Математика	Самый лучший курс по профильной математике! Огромное спасибо Тимуру за полученные знания!!!
Программирование	Отличный курс. Помимо интересных логических задач, много упражнений на освоение библиотек numpy и pandas. Спасибо
Физика	Мертвый курс. Учащихся мало. Косяков в задачах не много, но есть. Много проблем с представлением ответа (далеко не всегда ответы даются в СИ). Прошло курс очень мало народа. Научиться чему либо вряд ли возможно
Биология	Хороший курс, особенно для биологов, крайне важные вещи для современной биологии рассказывают
Химия	Курс хорош для школьников начальных классов (тех, кто не изучал неорганическую химию) тем, что дает легкий старт в предмет

Сравнительный анализ библиотек сентиментного анализа (анализа тональности текстов) *textblob* и *dostoevsky* на примере оценки тональности русскоязычных сообщений в *Facebook*[1] приводится в [Нугуманова и др., 2021]. Авторы отмечают, что библиотека *dostoevsky* продемонстрировала незначительные преимущества по сравнению с *textblob* с точки зрения точности распознавания тональности, однако при применении *textblob* используются не оригинальные русскоязычные отзывы, а их перевод на английский язык. По этой причине для анализа текстов в данной работе была выбрана *python*-библиотека *dostoevsky*. Пример работы *python*-библиотеки *dostoevsky* представлен в табл. 5.

Как видно из табл. 5, отзывы могут содержать предложения как с негативной, так и с позитивной тональностью, поэтому для дальнейшего исследования использовалась токенизация отзывов на предложения с помощью *python*-библиотеки nltk и оценка их тональности с помощью *python*-библиотеки *dostoevsky*, как показано в табл. 6. *Python*-библиотека *dostoevsky* основана на использовании предобученной модели нейронной сети и *ruSentiment* — корпуса размеченных комментариев по классам тональности. *Dostoevsky* позволяет определять следующие классы тональности: позитивная, негативная и нейтральная тональность, а также «речь» и «пропуск». К классу тональности «речь» относятся тексты, выражающие благодарность, а к классу тональности «пропуск» — тексты, тональность которых нейронная сеть не смогла определить. Примером «пропуска» в исследуемом датасете является предложение «А то за суперлегкую задачу 1 балл и за головоломную когда не один час решаешь тоже 1 балл».

[1] Деятельность социальной сети признана экстремистской и запрещена на территории РФ, данные используются в исследовательских целях и не направлены на одобрение экстремистской деятельности.

Таблица 5. **Примеры отзывов, размеченных с помощью *python*-библиотеки *dostoevsky***

Пример отзыва после удаления пунктуации и приведения символов к нижнему регистру	Оценка тональности с помощью библиотеки *dostoevsky*
отличный курс большое вам спасибо артем	Отзыв отнесен к категории «речь» (благодарность) с оценкой 0,99683732
по русски читать невозможно ошибка на опечатке но для тех кому химия никак не дается курс будет полезен в задачках непосредственно показывается откуда что брать и куда подставлять	Отзыв отнесен к категории «негативный» с оценкой 0,44553956
лучший курс для подготовки к егэ по математике с отличным учителем ever учеба с тимуром соточка баллов за экзамен отлично проведенное время	Отзыв отнесен к категории «позитивный» с оценкой 0,65842754
курс хорош и не только для начинающих для тех кто знаком с *python* не первый год так же есть немало полезного я бы особенно отметил некоторые полуолимпиадные задачки по программированию а так же минипроекты по *matplotlib seaborn* и *plotly* которые хоть и не простые и многие не любят их делать но приносят реальную пользу	Отзыв отнесен к категории «нейтральный» с оценкой 0,62978464

Для оценки точности классификации предложений по трем классам тональности — позитивная, негативная, нейтральная — из датасета извлечены 983 предложения. Такая подвыборка получена после удаления из отзывов по биологии (1080 предложений) тех, которые на основе *dostoevsky* были отнесены к классам «речь» и «пропуск». Для получения истинных значений тональности далее 983 предложения размечены вручную по трем классам тональности, при этом к классу нейтральных отнесены предложения, содержащие констатацию факта, например «модули курса плавно идут один за другим создавая целостность картины», или пожелания и рекомендации, например «единственное хотелось бы более подробно узнать о том как правильно и качественно отбирать данные для филогенетических деревьев модуль iv» (примеры приведены после удаления знаков пунктуации из предложений с сохранением авторской орфографии). К негативным предложениям, помимо предложений со словами, имеющими негативную эмоциональную окраску, отнесены также предложения, описывающие тревожности и страхи обучающегося, например «каждый раз когда я сталкиваюсь с новой программой для своей научной работы я боюсь что не подружусь с ней».

В результате разметки вручную в подвыборке, содержащей 983 предложения, выделены 393 позитивных, 62 негативных и

528 нейтральных предложений. Поскольку подвыборка не сбалансирована по числу элементов в классах тональности, для исследования качества классификации по тональности выбрана взвешенная метрика *F1-score.* Метрика *F1-score* вычисляется по формуле:

$$F1-score=\frac{2\cdot Precision\cdot Recall}{Precision+Recall},\ \text{где}$$

$$Precision=\frac{TP}{TP+FP}\ \text{и}\ Recall=\frac{TP}{TP+FN},$$

TP — верно положительные значения, FP, FN — ложно положительные и ложно отрицательные значения соответственно. В табл. 6 приведены результаты вычисления метрик качества классификации *Precision*, *Recall*, *F1-score* и поддержка (*Support*) — количество предложений.

Таблица 6. **Данные после выполнения токенизации отзывов на предложения и классификации по классам тональности**

Дисциплина	Количество отзывов	Количество предложений в отзывах после токенизации и удаления пунктуации и «пустых» предложений	Распределение предложений из отзывов по тональности на основе *dostoevsky*
Математика	2290	3655	Позитивная — 909, негативная — 133, нейтральная — 2247, речь — 308, пропуск — 58
Программирование	2280	4160	Позитивная — 875, негативная — 169, нейтральная — 2865, речь — 210, пропуск — 41
Физика	58	132	Позитивная — 21, негативная — 12, нейтральная — 91, речь — 3, пропуск — 5
Биология	856	1080	Позитивная — 263, негативная — 47, нейтральная — 673, речь — 89, пропуск — 8
Химия	237	242	Позитивная — 53, негативная — 16, нейтральная — 153, речь — 17, пропуск — 3

Взвешенная метрика *F1-score* вычисляется с учетом доли поддержки предложений разной тональности и для несбалансированной подвыборки используется в качестве основной характеристики качества классификации. Так, для результатов из табл. 7 взвешенная метрика *F1-score* вычисляется как

$$weighted\ F1-score = \frac{0,29 \cdot 62}{983} + \frac{0,8 \cdot 528}{983} + \frac{0,73 \cdot 393}{983} \approx 0,74.$$

Таблица 7. **Метрики качества классификации по классам тональности, вычисленные с помощью *python*-библиотеки *sklearn***

	Precision	Recall	F1-score	Support
негативная	0,34	0,26	0,29	62
нейтральная	0,71	0,91	0,80	528
позитивная	0,92	0,61	0,73	393
accuracy			0,75	983
macro avg	0,66	0,59	0,61	983
weighted avg	0,71	0,75	**0,74**	983

Таким образом, точность классификации по классам тональности на выделенной подвыборке составила 74%. Как видно из табл. 6, снижение точности наблюдается на негативных предложениях из-за недостаточного количества предложений этой тональности и отнесения предложений с описанием трудностей и тревожностей к классу нейтральных предложений. Некоторые примеры ошибочного отнесения предложений к классам тональности с помощью библиотеки *dostoevsky* приведены в табл. 8.

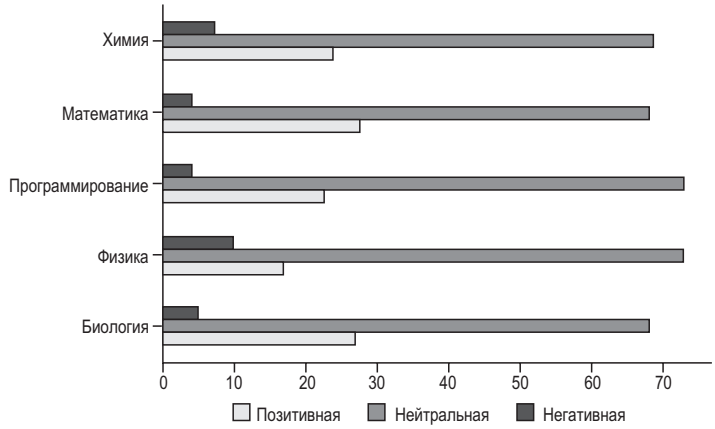
Таблица 8. **Примеры ошибочного определения тональности предложений на основе *dostoevsky***

Предложения с авторской орфографией (после удаления знаков пунктуации и приведения к нижнему регистру)	Тональность на основе *dostoevsky*	Истинное значение после разметки вручную
мне очень понравилось содержание курса вся информация очень подробно изложена	Нейтральная	Позитивная
лично для себя я вновь убедилась что наука необъятная	Негативная	Нейтральная
мне эти программы не показались простыми и инструкций по фото было не достаточно	Нейтральная	Негативная

Распределение предложений из отзывов по разным дисциплинам по классам тональности на основе библиотеки *dostoevsky* приведено на рис. 1.

Из рис. 1 видно, что в датасете преобладают предложения с позитивной эмоциональной окраской и предложения с нейтральной тональностью, которые преимущественно содержат описание структуры курсов и их контента без ярко выраженного эмоционального отношения, а также пожелания авторам MOOK.

Рис. 1. **Распределение предложений из отзывов на МООК по классам тональности на основе *python*-библиотеки *dostoevsky*, %**



Извлечение наиболее часто встречающихся униграмм (отдельных слов) и биграмм (словосочетаний из двух слов) позволяет описать различные аспекты МООК, их контента и качества преподавания. В табл. 9 приведены результаты извлечения наиболее часто встречающихся униграмм, связанных с ключевым словом «курс». В позитивных предложениях слушатели МООК наиболее часто используют такие характеристики курсов, как «отличный», «хороший», «прекрасный», «замечательный», «интересный», при этом положительные характеристики не зависят от изучаемых дисциплин. В немногочисленных отрицательных предложениях в исследуемом датасете либо отсутствовали описания отрицательных характеристик, либо курс характеризовался как «сложный», «школьный» или «университетский».

Таблица 9. **Извлечение описательных характеристик курсов на основе анализа униграмм с учетом тональности предложений**

Дисциплины	Наиболее часто встречающиеся униграммы с учетом тональности в формате «униграмма: количество» для ключевого слова «курс»	
	Позитивная	Негативная
Математика	отличный: 279, хороший: 167, замечательный: 74, прекрасный: 58, понравиться: 49	—
Программирование	отличный: 293, хороший: 178, понравиться: 40, замечательный: 32, начинающий: 27, прекрасный: 20	—
Физика	хороший: 3	университетский: 3
Биология	отличный: 45, интересный: 33, хороший: 27, замечательный: 22, понравиться: 12	сложный: 3, школьный: 3
Химия	хороший: 17, понравиться: 5, отличный: 5, прекрасный: 4	забросить: 3

Из табл. 9 и 10 видно, что в предложениях, извлеченных из комментариев, преобладали описательные характеристики курсов. После удаления биграмм с ключевым словом «курс» частотность биграмм существенно уменьшилась, но были извлечены биграммы, характерные для конкретной предметной области. Например, среди биграмм по биологии преобладала «молекулярная биология», встречавшаяся в позитивных предложениях отзывов, что может свидетельствовать о направленности МООК на изучение этого раздела биологии. В позитивных предложениях отзывов на МООК по программированию слушатели наиболее часто отмечали подачу материала и интересные задачи. Анализ биграмм в негативных предложениях показывает, что трудности, которые возникают у слушателей при освоении курсов, они связывают с уровнем сложности и количеством рассматриваемых тем и задач и работу над некоторыми из них считают напрасно потраченным временем. Предварительная обработка предложений на этапе анализа униграмм и биграмм включала удаление стоп-слов и приведение к нормальной форме с помощью *python*-библиотек nltk и *pymorphy2* соответственно. Более глубокие методы анализа биграмм и триграмм позволяют выявить причины трудностей и типы тревожностей, которые возникают у слушателей при изучении МООК, а также исследовать качество учебного контента и методик преподавания.

Таблица 10. **Результаты анализа биграмм, не относящихся к слову «курс», с учетом тональности предложений**

Дисциплины	Наиболее часто встречающиеся биграммы с учетом тональности в формате «биграмма: количество» после приведения к нормальной форме	
	Позитивная	Негативная
Математика	(подача, материал): 12, (прекрасный, преподаватель): 7, (научиться, решать): 5, (преподаватель, отличный): 5	(количество, задача): 4, (огромный, количество): 2, (большой, количество): 2
Программирование	(подача, материал): 15, (задача, отличный): 9, (интересный, задача): 8, (хотеть, поблагодарить): 2	(задание, сложный): 2, (тратить, время): 2, (ломать, голова): 2
Физика	(приятный, впечатление): 2, (интересный, наука): 2, (помогать, увидеть): 2, (хороший, осмысление): 2	(время, тратить): 2, (стать, тупик): 2, (непреодолимый, препятствие): 2
Биология	(быть, интересно): 8, (молекулярный, биология): 6	(сложный, тема): 2, (даться, нелегко): 2, (отчаяться, задание): 2
Химия	(подача, материал): 2, (подача, информация): 2, (хороший, подача): 2	(химия, бояться): 2

4. Обсуждение результатов и направления дальнейших исследований

Комментарии и отзывы на платформах МООК оставляют далеко не все пользователи, но тексты отзывов и комментарии на форумах часто являются единственным источником обратной связи от обучающихся. Отслеживание эмоционального отношения обучающихся к курсу и анализ причин оттока слушателей — важные задачи учебной аналитики МООК.

Одной из доступных библиотек для исследования эмоциональной окраски русскоязычных текстов является библиотека *dostoevsky*. В процессе исследования датасета и классификации по тональности русскоязычных предложений из отзывов датасета библиотека *dostoevsky* продемонстрировала 74%-ную точность классификации по тональности. В дальнейшем необходимо разрабатывать специальные инструменты учебной аналитики, основанные на методах и алгоритмах для классификации предложений по тональности с учетом особенностей предметной области, в том числе для выявления причин и типов трудностей и тревожностей, которые возникают при обучении на платформах МООК в условиях ограниченной обратной связи от слушателей и тьютора или ее отсутствия.

Предложим некоторые направления дальнейших исследований рассматриваемого датасета на основе методов обработки естественного языка:

1) разработка методов и алгоритмов для повышения качества классификации предложений по тональности с учетом особенностей предметной области;
2) исследование тональности русскоязычных предложений с учетом проявления более сложных эмоций, таких как агрессия, наслаждение, удовлетворенность, ненависть и т.п.;
3) выявление причин и типов тревожностей обучающихся на основе анализа текстов отзывов на платформах МООК;
4) анализ деструктивного контента в текстах отзывов обучающихся;
5) оценка влияния тональности отзывов на платформах МООК на академическую успешность обучающихся;
6) выявление причин неудовлетворенности МООК на основе анализа текстов отзывов;
7) оценивание качества учебного контента и качества преподавания на основе анализа текстов отзывов на МООК.

5. Заключение

Сформированный датасет состоит из 5721 русскоязычного отзыва на МООК и может быть использован в своих исследованиях специалистами в области компьютерной лингвистики, анализа тональности текстов и извлечения мнений, анализа образовательных данных и разработки инструментов учебной анали-

тики. В частности, данные отзывов с платформы *Stepik* позволят разрабатывать инновационные инструменты учебной аналитики для выявления отношения обучающихся к курсам, тьютору, качеству видеолекций, конспектов, интерактивных тестов, обратной связи от слушателей, причин трудностей при обучении на курсах и т.п., а также универсальные критерии для оценивания качества МООК и разработки систем рекомендаций МООК на основе формальных подходов. Применение анализа униграмм и биграмм с учетом эмоциональной окраски позволило выделить описательные характеристики курсов, трудности (например, сложность и объем заданий) и наиболее понравившиеся аспекты обучения (например, подача материала, интересные задания, прекрасный лектор). Таким образом, обработка естественного языка является одним из перспективных направлений для изучения мнений и настроений обучающихся.

Литература

1. Дюличева Ю. (2021) Учебная аналитика МООК как инструмент анализа математической тревожности. *Вопросы образования / Educational Studies Moscow*, № 4, сс. 243–265. https://doi.org/10.17323/1814-9545-2021-4-243-265
2. Нугуманова А.Б., Ахмед-Заки Д.Ж., Байбурин Е.М., Апаев К.С. (2021) Сентмент-анализ отзывов пользователей в Фейсбуке: сравнение библиотек Textblob и Dostoevsky. *Вестник Национальной инженерной академии Республики Казахстан*, № 4 (82), сс. 97–104. https://doi.org/10.47533/2020.1606-146X.120
3. Alsaad F., Alawini A. (2020) Unsupervised Approach for Modeling Content Structures of MOOCs. Proceedings of the *13th International Conference on Educational Data Mining (EDM 2020) (online, 2020, 10–13 July)*, pp. 18–28.
4. An Y.-H., Pan L., Kan M.-Y., Dong Q., Fu Y. (2019) Resource Mention Extraction for MOOC Discussion Forums. *IEEE Access*, vol. *7*, pp. 87887–87900. https://doi.org/10.1109/access.2019.2924250
5. Andres J.M.L., Baker R.S., Gašević D., Siemens G., Crossley S.A., Joksimović S. (2018) Studying MOOC Completion at Scale Using the MOOC Replication Framework. Proceedings of the *8th International Conference on Learning Analytics and Knowledge (LAK '18) (Sydney, Australia, 2018, 07–09 March)*, pp. 71–78. https://doi.org/10.1145/3170358.3170369
6. Atapattu T., Falkner K. (2016) A Framework for Topic Generation and Labeling from MOOC Discussions. Proceedings of the *Third (2016) ACM Conference on Learning @ Scale (Edinburgh, United Kingdom, 2016, 25–29 April)*, pp. 201–204. https://doi.org/10.1145/2876034.2893414
7. Chen Q., Chen Y., Liu D., Shi C., Wu Y., Qu H. (2016) PeakVizor: Visual Analytics of Peaks in Video Clickstreams from Massive Open Online Courses. *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no 10, pp. 2315–2330. https://doi.org/10.1109/tvcg.2015.2505305
8. Chen Q., Yue X., Plantaz X., Chen Y., Shi C., Pong T., Qu H. (2020) ViSeq: Visual Analytics of Learning Sequence in Massive Open Online Courses. *IEEE Transactions on Visualization and Computer Graphi*cs, vol. 26, no 3, pp. 1622–1636. https://doi.org/10.1109/tvcg.2018.2872961
9. Crossley S., Paquette L., Dascalu M., McNamara D.S., Baker R.S. (2016) Combining Click-Stream Data with NLP Tools to Better Understand MOOC Completion. Proceedings of the *Sixth International Conference on Learning Analytics*

& Knowledge (Edinburgh, United Kingdom, 2016, 25–29 April), pp. 6–14. https://doi.org/10.1145/2883851.2883931
10. Dhekne C., Bansal S.K. (2018) MOOClink: An Aggregator for MOOC Offerings from Various Providers. *Journal of Engineering Education Transformations*, vol. 31, January, Special issue, Article no eISSN 2394-1707. https://doi.org/10.16920/jeet/2018/v0i0/120907
11. Dina N.Z., Yunardi R.T., Firdaus A.A. (2021) Utilizing Text Mining and Feature-Sentiment-Pairs to Support Data-Driven Design Automation Massive Open Online Course. *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no 1, pp. 134–151. https://doi.org/10.3991/ijet.v16i01.17095
12. Dyulicheva Yu. (2022) Dataset of MOOCs' Reviews from Stepik on Russian Language, Mendeley Data, V1. https://doi.org/10.17632/8rwpvrw4hw.1 Available at: https://data.mendeley.com/datasets/8rwpvrw4hw/1 (accessed 20 November 2022).
13. Ezen-Can A., Boyer K.E., Kellogg S., Booth S. (2015) Unsupervised Modeling for Understanding MOOC Discussion Forums. Proceedings of the *Fifth International Conference on Learning Analytics and Knowledge (Poughkeepsie, NY, 2015, 16–20 March),* pp. 146–150. https://doi.org/10.1145/2723576.2723589
14. Iniesto F., Rodrigo C. (2019) YourMOOC4all: A Recommender System for MOOCs Based on Collaborative Filtering Implementing UDL. *Transforming Learning with Meaningful Technologies. EC-TEL 2019. Lecture Notes in Computer Science* (eds M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, J. Schneider), Cham: Springer, vol. 11722, pp. 746–750. https://doi.org/10.1007/978-3-030-29736-7_80
15. Jiang Z., Feng S., Cong G., Miao C., Li X. (2017) A Novel Cascade Model for Learning Latent Similarity from Heterogeneous Sequential Data of MOOC. Proceedings of the *2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen, Denmark, 2017, 07–11 September), pp. 2768–2773. https://doi.org/10.18653/v1/d17-1293
16. Kastrati Z., Imran A.S., Kurti A. (2020) Weakly Supervised Framework for Aspect-Based Sentiment Analysis on Students' Reviews of MOOCs. *IEEE Access*, vol. 8, pp. 106799–106810. https://doi.org/10.1109/access.2020.3000739
17. Khalil M., Belokrys G. (2020) OXALIC: An Open edX Advanced Learning Analytics Tool. Proceedings of the *2020 IEEE Learning with MOOCS (LWMOOCS) (Antigua Guatemala, Guatemala*, *2020, 29 September — 02 October),* pp. 185–190. https://doi.org/10.1109/lwmoocs50143.2020.9234322
18. Koffi D.D.A.S., Ouattara N., Mambe D.M., Oumtanaga S., Assohoun A.D.J.E. (2021) Cources Recommendation Algorithm Based on Performance Prediction in E-learning. *IJCSNS International Journal of Computer Science and Network Security,* vol. 21, no 2, pp. 148–158. https://doi.org/10.22937/IJCSNS.2021.21.2.17
19. Li X., Men C., Zhang F., Du Z. (2017) A Smart Visual Analysis Solution for MOOC Data. Proceedings of the *2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 15th International Conference on Pervasive Intelligence and Computing, 3rd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech) (Orlando, FL, 2017, 06–10 November)*, pp. 101–106. https://doi.org/10.1109/dasc-picom-datacom-cyberscitec.2017.31
20. Lim S.L., Goh O.S. (2016) *Intelligent Conversational Bot for Massive Online Open Courses (MOOCs)*. arxiv.org/abs/1601.07065. https://doi.org/10.48550/arXiv.1601.07065
21. Liu S., Ni C., Liu Z., Peng X., Cheng H.N. (2017) Mining Individual Learning Topics in Course Reviews Based on Author Topic Model. *International Journal of Distance Education Technolo*gies, vol. 15, no 3, pp. 1–14. https://doi.org/10.4018/ijdet.2017070101

22. Lopez G., Seaton D.T., Ang A., Tingley D., Chuang I. (2017) Google BigQuery for Education: Framework for Parsing and Analyzing edX MOOC Data. Proceedings of the *Fourth (2017) ACM Conference on Learning @ Scale (Cambridge, MA, 2017, held on 20–21 April),* pp. 181–184. https://doi.org/10.1145/3051457.3053980
23. Moreno-Marcos P.M., Alario-Hoyos C., Muñoz Merino P.J., Estevez-Ayres I., Kloos C.D. (2019) A Learning Analytics Methodology for Understanding Social Interactions in MOOCs. *IEEE Transactions on Learning Technologies*, vol. 12, no 4, pp. 442–455. https://doi.org/10.1109/tlt.2018.2883419
24. Mu X., Xu K., Chen Q., Du F., Wang Y., Qu H. (2019) MOOCad: Visual Analysis of Anomalous Learning Activities in Massive Open Online Courses. Proceedings of the *21st Eurographics Conference on Visualization, EuroVis 2019 — Short Papers (Porto, Portugal, 2019, 03–07 June)* (eds J. Johansson, F. Sadlo, G.E. Marai), Porto: The Eurographics Association. https://doi.org/10.2312/evs.20191176
25. Mubarak A.A., Ahmed S.A., Cao H. (2021) MOOC-ASV: Analytical Statistical Visual Model of Learners' Interaction in Videos of MOOC Courses. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2021.1916768
26. Onah D., Pang E. (2021) MOOC Design Principles: Topic Modelling-Pyldavis Visualization & Summarization of Learners' Engagement. Proceedings of the *13th Annual International Conference on Education and New Learning Technologies (online, 2021, 05–06 July)*, pp. 1082–1088. https://doi.org/10.21125/edulearn.2021.0282
27. Onan A. (2020) Sentiment Analysis on Massive Open Online Course Evaluations: A Text Mining and Deep Learning Approach.*Computer Applications in Engineering Education*, vol. 29, no 3, pp. 572–589. https://doi.org/10.1002/cae.22253
28. Ramesh A., Goldwasser D., Huang B., Daume H., Getoor L. (2014) Understanding MOOC Discussion Forums Using Seeded LDA. Proceedings of the *Ninth Workshop on Innovative Use of NLP for Building Educational Applications* (Baltimore, ME, 2014, 26 June), pp. 28–33. https://doi.org/10.3115/v1/w14-1804
29. Reich J., Tingley D.H., Leder-Luis J., Roberts M.E., Stewart B. (2014) Computer-Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses. *SSRN Electronic Journal*, vol. 2, no 1, pp. 156–184. https://doi.org/10.2139/ssrn.2499725
30. Sarıyalçınkaya A.D., Karal H., Altinay F., Altinay Z. (2021) Reflections on Adaptive Learning Analytics: Adaptive Learning Analytics. *Advancing the Power of Learning Analytics and Big Data in Education* (eds A. Azevedo, J. Azevedo, J. Onohuome Uhomoibhi, E. Ossiannilsson), Hershey, PA: IGI Global, pp. 61–84. https://doi.org/10.4018/978-1-7998-7103-3.ch003
31. *Shah D. (2019) Year of MOOC-Based Degrees: A Review of MOOC Stats and Trends in 2018*. Available at: https://www.classcentral.com/report/moocs-stats-and-trends-2018/ (accessed 08 November 2022).
32. Shrestha S., Pokharel M. (2021) Educational Data Mining in Moodle Data. *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 10, no 1, pp. 9–18. https://doi.org/10.11591/ijict.v10i1.pp9-18
33. Shridharan M., Willingham A., Spencer J., Yang T., Brinton C. (2018) Predictive Learning Analytics for Video-Watching Behavior in MOOCs. Proceedings of the *52nd Annual Conference on Information Sciences and Systems (CISS) (Princeton, NJ, 2018. 21–23 March)*, pp. 1–6. https://doi.org/10.1109/ciss.2018.8362323
34. Siddique S.A. (2020) Improvement of Online Course Content Using MapReduce Big Data Analytics. *International Research Journal of Engineering and Technology (IRJET),* vol. 7, no 8, pp. 50–56.
35. Singh A.K., Kumar S., Bhushan S., Kumar P., Vashishtha A. (2021) A Proportional Sentiment Analysis of MOOCs Course Reviews Using Supervised Lear-

ning Algorithms. *Ingénierie des systèmes d information*, vol. 26, no 5, pp. 501–506. https://doi.org/10.18280/isi.260510

36. Sun D., Li T., You F., Hu M., Li Z. (2021) Prediction of Learning Behavior Characters of MOOC's Data Based on Time Series Analysis. *Journal of Physics: Conference Series*, vol. 1994, no 1, Article no 012009. https://doi.org/10.1088/1742-6596/1994/1/012009
37. Thoms B., Eryilmaz E., Mercado G., Ramirez B., Rodriguez J. (2017) Towards a Sentiment Analyzing Discussion-Board. Proceedings of the *50th Hawaii International Conference on System Sciences (2017) (Hilton Waikoloa Village, Hawaii, 2017, 04–07 January)*, pp. 184–193. https://doi.org/10.24251/hicss.2017.021
38. Yao J., Wang L., Liu Y., Kui Y. (2021) Research on the Data Analysis System of Student Stress in English MOOC Based on Fuzzy C-Means Algorithm. *Journal of Intelligent & Fuzzy Systems*, May, pp. 1–11. https://doi.org/10.3233/jifs-219048
39. Yu J., Alrajhi L., Harit A., Sun Z., Cristea A.I., Shi L. (2021) Exploring Bayesian Deep Learning for Urgent Instructor Intervention Need in MOOC Forums. Proceedings of the *Intelligent Tutoring Systems: 17th International Conference, ITS 2021 (online, 2021, 07–11 June)*, pp. 78–90. https://doi.org/10.1007/978-3-030-80421-3_10
40. Yu J., Luo G., Xiao T., Zhong Q. et al. (2020) MOOCCube: A Large-Scale Data Repository for NLP Applications in MOOCs. Proceedings of the *58th Annual Meeting of the Association for Computational Linguistics (online, 2020, 05–10 July)*, pp. 3135–3142. https://doi.org/10.18653/v1/2020.acl-main.285
41. Wang H., Xie Y., Wen M., Yang Z. (2021) GazeMOOC: A Gaze Data Driven Visual Analytics System for MOOC with XR Content. Proceedings of the *27th ACM Symposium on Virtual Reality Software and Technology (Osaka, Japan, 2021, 08–10 December)*, Article no 74. https://doi.org/10.1145/3489849.3489923
42. Wong J., Pursel B., Divinsky A., Jansen B.J. (2015) Analyzing MOOC Discussion Forum Messages to Identify Cognitive Learning Information Exchanges. Proceedings of the *Association for Information Science and Technology*, vol. 52, no 1, pp. 1–10. https://doi.org/10.1002/pra2.2015.145052010023
43. Wong J., Zhang X. (2018) MessageLens: A Visual Analytics System to Support Multifaceted Exploration of MOOC Forum Discussions. *Visual Informatics*, vol. 2, no 1, pp. 37–49. https://doi.org/10.1016/j.visinf.2018.04.005
44. Zarra T., Chiheb R., Faizi R., El Afia A. (2018) MOOCs' Recommendation Based on Forum Latent Dirichlet Allocation. Proceedings of the *2nd International Conference on Smart Digital Environment (Rabat, Morocco, 2018, 18–20 October)*, pp. 88–93. https://doi.org/10.1145/3289100.3289115

References

Alsaad F., Alawini A. (2020) Unsupervised Approach for Modeling Content Structures of MOOCs. Proceedings of the *13th International Conference on Educational Data Mining (EDM 2020) (online, 2020, 10–13 July),* pp. 18–28.

An Y.-H., Pan L., Kan M.-Y., Dong Q., Fu Y. (2019) Resource Mention Extraction for MOOC Discussion Forums. *IEEE Access*, vol. 7, pp. 87887–87900. https://doi.org/10.1109/access.2019.2924250

Andres J.M.L., Baker R.S., Gašević D., Siemens G., Crossley S.A., Joksimović S. (2018) Studying MOOC Completion at Scale Using the MOOC Replication Framework. Proceedings of the *8th International Conference on Learning Analytics and Knowledge (LAK '18) (Sydney, Australia, 2018, 07–09 March)*, pp. 71–78. https://doi.org/10.1145/3170358.3170369

Atapattu T., Falkner K. (2016) A Framework for Topic Generation and Labeling from MOOC Discussions. Proceedings of the *Third (2016) ACM Conference on Learning @ Scale (Edinburgh, United Kingdom, 2016, 25–29 April)*, pp. 201–204. https://doi.org/10.1145/2876034.2893414

Chen Q., Chen Y., Liu D., Shi C., Wu Y., Qu H. (2016) PeakVizor: Visual Analytics of Peaks in Video Clickstreams from Massive Open Online Courses. *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no 10, pp. 2315–2330. https://doi.org/10.1109/tvcg.2015.2505305

Chen Q., Yue X., Plantaz X., Chen Y., Shi C., Pong T., Qu H. (2020) ViSeq: Visual Analytics of Learning Sequence in Massive Open Online Courses. *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no 3, pp. 1622–1636. https://doi.org/10.1109/tvcg.2018.2872961

Crossley S., Paquette L., Dascalu M., McNamara D.S., Baker R.S. (2016) Combining Click-Stream Data with NLP Tools to Better Understand MOOC Completion. Proceedings of the *Sixth International Conference on Learning Analytics & Knowledge* (Edinburgh, United Kingdom, 2016, 25–29 April), pp. 6–14. https://doi.org/10.1145/2883851.2883931

Dhekne C., Bansal S.K. (2018) MOOClink: An Aggregator for MOOC Offerings from Various Providers. *Journal of Engineering Education Transformations*, vol. 31, January, Special issue, Article no eISSN 2394-1707. https://doi.org/10.16920/jeet/2018/v0i0/120907

Dina N.Z., Yunardi R.T., Firdaus A.A. (2021) Utilizing Text Mining and Feature-Sentiment-Pairs to Support Data-Driven Design Automation Massive Open Online Course. *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no 1, pp. 134–151. https://doi.org/10.3991/ijet.v16i01.17095

Dyulicheva Yu. (2022) Dataset of MOOCs' Reviews from Stepik on Russian Language, Mendeley Data, V1, https://doi.org/ 10.17632/8rwpvrw4hw.1 Available at: https://data.mendeley.com/datasets/8rwpvrw4hw/1 (accessed 20 November 2022).

Dyulicheva Y.Y. (2021) Uchebnaya analitika MOOK kak instrument analiza matematicheskoy trevozhnosti [Learning Analytics in MOOCs as an Instrument for Measuring Math Anxiety]. Voprosy obrazovaniya / Educational Studies Moscow, no 4, pp. 243–265. https://doi.org/10.17323/1814-9545-2021-4-243-265

Ezen-Can A., Boyer K.E., Kellogg S., Booth S. (2015) Unsupervised Modeling for Understanding MOOC Discussion Forums. Proceedings of the *Fifth International Conference on Learning Analytics and Knowledge (Poughkeepsie, NY, 2015, 16–20 March),* pp. 146–150. https://doi.org/10.1145/2723576.2723589

Iniesto F., Rodrigo C. (2019) YourMOOC4all: A Recommender System for MOOCs Based on Collaborative Filtering Implementing UDL. *Transforming Learning with Meaningful Technologies. EC-TEL 2019. Lecture Notes in Computer Science* (eds M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, J. Schneider), Cham: Springer, vol. 11722, pp. 746–750. https://doi.org/10.1007/978-3-030-29736-7_80

Jiang Z., Feng S., Cong G., Miao C., Li X. (2017) A Novel Cascade Model for Learning Latent Similarity from Heterogeneous Sequential Data of MOOC. Proceedings of the *2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen, Denmark, 2017, 07–11 September), pp. 2768–2773. https://doi.org/10.18653/v1/d17-1293

Kastrati Z., Imran A.S., Kurti A. (2020) Weakly Supervised Framework for Aspect-Based Sentiment Analysis on Students' Reviews of MOOCs. *IEEE Access*, vol. 8, pp. 106799–106810. https://doi.org/10.1109/access.2020.3000739

Khalil M., Belokrys G. (2020) OXALIC: An Open edX Advanced Learning Analytics Tool. Proceedings of the *2020 IEEE Learning with MOOCS (LWMOOCS) (Antigua Guatemala, Guatemala*, *2020, 29 September — 02 October),* pp. 185–190. https://doi.org/10.1109/lwmoocs50143.2020.9234322

Koffi D.D.A.S., Ouattara N., Mambe D.M., Oumtanaga S., Assohoun A.D.J.E. (2021) Cources Recommendation Algorithm Based on Performance Prediction in E-learning. *IJCSNS International Journal of Computer Science and Network Security,* vol. 21, no 2, pp. 148–158. https://doi.org/10.22937/IJCSNS.2021.21.2.17

Li X., Men C., Zhang F., Du Z. (2017) A Smart Visual Analysis Solution for MOOC Data. Proceedings of the *2017 IEEE 15th International Conference on Depend-*

able, Autonomic and Secure Computing, 15th International Conference on Pervasive Intelligence and Computing, 3rd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech) (Orlando, FL, 2017, 06–10 November), pp. 101–106. https://doi.org/10.1109/dasc-picom-datacom-cyberscitec.2017.31

Lim S.L., Goh O.S. (2016) *Intelligent Conversational Bot for Massive Online Open Courses (MOOCs)*. arxiv.org/abs/1601.07065. https://doi.org/10.48550/arXiv.1601.07065

Liu S., Ni C., Liu Z., Peng X., Cheng H.N. (2017) Mining Individual Learning Topics in Course Reviews Based on Author Topic Model. *International Journal of Distance Education Technologies*, vol. 15, no 3, pp. 1–14. https://doi.org/10.4018/ijdet.2017070101

Lopez G., Seaton D.T., Ang A., Tingley D., Chuang I. (2017) Google BigQuery for Education: Framework for Parsing and Analyzing edX MOOC Data. Proceedings of the *Fourth (2017) ACM Conference on Learning @ Scale (Cambridge, MA, 2017, held on 20–21 April),* pp. 181–184. https://doi.org/10.1145/3051457.3053980

Moreno-Marcos P.M., Alario-Hoyos C., Muñoz Merino P.J., Estevez-Ayres I., Kloos C.D. (2019) A Learning Analytics Methodology for Understanding Social Interactions in MOOCs. *IEEE Transactions on Learning Technologies*, vol. 12, no 4, pp. 442–455. https://doi.org/10.1109/tlt.2018.2883419

Mu X., Xu K., Chen Q., Du F., Wang Y., Qu H. (2019) MOOCad: Visual Analysis of Anomalous Learning Activities in Massive Open Online Courses. Proceedings of the *21st Eurographics Conference on Visualization, EuroVis 2019 — Short Papers (Porto, Portugal, 2019, 03–07 June)* (eds J. Johansson, F. Sadlo, G.E. Marai), Porto: The Eurographics Association. https://doi.org/10.2312/evs.20191176

Mubarak A.A., Ahmed S.A., Cao H. (2021) MOOC-ASV: Analytical Statistical Visual Model of Learners' Interaction in Videos of MOOC Courses. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2021.1916768

Nugumanova A.B., Akhmed-Zaki D.Zh., Bayburin E.M., Apaev K.S. (2021) Sentiment-analiz otzyvov pol'zovatelej v Fejsbuke: sravnenie bibliotek Textblob i Dostoevsky [Sentiment Analysis of Users Reviews in Facebook: Comparison of Textblob and Dostoevsky Libraries]. *Bulletin of the National Engineering Academy of the Republic of Kazakhstan*, no 4 (82), pp. 97–104. https://doi.org/10.47533/2020.1606-146X.120

Onah D., Pang E. (2021) MOOC Design Principles: Topic Modelling-Pyldavis Visualization & Summarization of Learners' Engagement. Proceedings of the *13th Annual International Conference on Education and New Learning Technologies (online, 2021, 05–06 July)*, pp. 1082–1088. https://doi.org/10.21125/edulearn.2021.0282

Onan A. (2020) Sentiment Analysis on Massive Open Online Course Evaluations: A Text Mining and Deep Learning Approach. *Computer Applications in Engineering Education*, vol. 29, no 3, pp. 572–589. https://doi.org/10.1002/cae.22253

Ramesh A., Goldwasser D., Huang B., Daume H., Getoor L. (2014) Understanding MOOC Discussion Forums Using Seeded LDA. Proceedings of the *Ninth Workshop on Innovative Use of NLP for Building Educational Applications* (Baltimore, ME, 2014, 26 June), pp. 28–33. https://doi.org/10.3115/v1/w14-1804

Reich J., Tingley D.H., Leder-Luis J., Roberts M.E., Stewart B. (2014) Computer-Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses. *SSRN Electronic Journal*, vol. 2, no 1, pp. 156–184. https://doi.org/10.2139/ssrn.2499725

Sarıyalçınkaya A.D., Karal H., Altinay F., Altinay Z. (2021) Reflections on Adaptive Learning Analytics: Adaptive Learning Analytics. *Advancing the Power of Learning Analytics and Big Data in Education* (eds A. Azevedo, J. Azevedo, J. Onohuome Uhomoibhi, E. Ossiannilsson), Hershey, PA: IGI Global, pp. 61–84. https://doi.org/10.4018/978-1-7998-7103-3.ch003

Shah D. (2019) *Year of MOOC-Based Degrees: A Review of MOOC Stats and Trends in 2018*. Available at: https://www.classcentral.com/report/moocs-stats-and-trends-2018/ (accessed 08 November 2022).

Shrestha S., Pokharel M. (2021) Educational Data Mining in Moodle Data. *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 10, no 1, pp. 9–18. https://doi.org/10.11591/ijict.v10i1.pp9-18

Shridharan M., Willingham A., Spencer J., Yang T., Brinton C. (2018) Predictive Learning Analytics for Video-Watching Behavior in MOOCs. Proceedings of the *52nd Annual Conference on Information Sciences and Systems (CISS) (Princeton, NJ, 2018. 21–23 March)*, pp. 1–6. https://doi.org/10.1109/ciss.2018.8362323

Siddique S.A. (2020) Improvement of Online Course Content Using MapReduce Big Data Analytics. *International Research Journal of Engineering and Technology (IRJET),* vol. 7, no 8, pp. 50–56.

Singh A.K., Kumar S., Bhushan S., Kumar P., Vashishtha A. (2021) A Proportional Sentiment Analysis of MOOCs Course Reviews Using Supervised Learning Algorithms. *Ingénierie des systèmes d information*, vol. 26, no 5, pp. 501–506. https://doi.org/10.18280/isi.260510

Sun D., Li T., You F., Hu M., Li Z. (2021) Prediction of Learning Behavior Characters of MOOC's Data Based on Time Series Analysis. *Journal of Physics: Conference Series*, vol. 1994, no 1, Article no 012009. https://doi.org/10.1088/1742-6596/1994/1/012009

Thoms B., Eryilmaz E., Mercado G., Ramirez B., Rodriguez J. (2017) Towards a Sentiment Analyzing Discussion-Board. Proceedings of the *50th Hawaii International Conference on System Sciences (2017) (Hilton Waikoloa Village, Hawaii, 2017, 04–07 January)*, pp. 184–193. https://doi.org/10.24251/hicss.2017.021

Yao J., Wang L., Liu Y., Kui Y. (2021) Research on the Data Analysis System of Student Stress in English MOOC Based on Fuzzy C-Means Algorithm. *Journal of Intelligent & Fuzzy Systems*, May, pp. 1–11. https://doi.org/10.3233/jifs-219048

Yu J., Alrajhi L., Harit A., Sun Z., Cristea A.I., Shi L. (2021) Exploring Bayesian Deep Learning for Urgent Instructor Intervention Need in MOOC Forums. Proceedings *of the Intelligent Tutoring Systems: 17th International Conference, ITS 2021 (online, 2021, 07–11 June)*, pp. 78–90. https://doi.org/10.1007/978-3-030-80421-3_10

Yu J., Luo G., Xiao T., Zhong Q. et al. (2020) MOOCCube: A Large-Scale Data Repository for NLP Applications in MOOCs. Proceedings of the *58th Annual Meeting of the Association for Computational Linguistics (online, 2020, 05–10 July)*, pp. 3135–3142. https://doi.org/10.18653/v1/2020.acl-main.285

Wang H., Xie Y., Wen M., Yang Z. (2021) GazeMOOC: A Gaze Data Driven Visual Analytics System for MOOC with XR Content. Proceedings of the *27th ACM Symposium on Virtual Reality Software and Technology (Osaka, Japan, 2021, 08–10 December)*, Article no 74. https://doi.org/10.1145/3489849.3489923

Wong J., Pursel B., Divinsky A., Jansen B.J. (2015) Analyzing MOOC Discussion Forum Messages to Identify Cognitive Learning Information Exchanges. Proceedings of the *Association for Information Science and Technology*, vol. 52, no 1, pp. 1–10. https://doi.org/10.1002/pra2.2015.145052010023

Wong J., Zhang X. (2018) MessageLens: A Visual Analytics System to Support Multifaceted Exploration of MOOC Forum Discussions. *Visual Informatics*, vol. 2, no 1, pp. 37–49. https://doi.org/10.1016/j.visinf.2018.04.005

Zarra T., Chiheb R., Faizi R., El Afia A. (2018) MOOCs' Recommendation Based on Forum Latent Dirichlet Allocation. Proceedings of the *2nd International Conference on Smart Digital Environment (Rabat, Morocco, 2018, 18–20 October)*, pp. 88–93. https://doi.org/10.1145/3289100.3289115