СЕКЦИЯ 3
МЕТОДЫ, ТЕХНОЛОГИИ И ПРОГРАММНЫЕ СРЕДСТВА ОБРАБОТКИ И АНАЛИЗА ДАННЫХ

А. Н. Ходырев, Д. В. Шушарин, И. Г. Захарова
Тюменский государственный университет, г. Тюмень
УДК 004.9

РАЗРАБОТКА И ИССЛЕДОВАНИЕ МЕТОДОВ ВЫЯВЛЕНИЯ НЕДОСТОВЕРНЫХ ОТЗЫВОВ НА ЭЛЕКТИВНЫЕ ДИСЦИПЛИНЫ

Аннотация. В работе рассмотрены вопросы выявления «недостоверных» онлайн-отзывов на элективные дисциплины на основе методов глубокого обучения. Представлены этапы предварительной обработки данных, разработаны и исследованы модели классификации отзывов. Для обучения моделей использованы данные, полученные с помощью веб-сервиса «Отзывус».

Ключевые слова: классификация, глубокое обучение, трансформеры, отзывы, недостоверная информация, элективные дисциплины.

Введение. В сети интернет уже достаточно долго существуют сервисы по покупке/продаже товаров и услуг на различные области, также у пользователей некоторых сервисов есть доступ к перечню различных заведений. Одним из инструментов ориентира среди всех объектов, выставленных на «витрину» интернет-полок, является обратная связь. Обратная связь необходима для нескольких категорий пользователей: будущих потребителей товаров, производителей товаров и их поставщиков. Для первой категории пользователей, как одной из причин необходимости обратной связи, является определение качества приобретаемого товара, тогда как для второй категории пользователей важно знать о мнении на свой производимый товар, а для третьей категории нужно иметь понимание о необходимости поставки товара от данного производителя. Одной из областей, для которой также очень важно получение обратной связи, может являться и образование, где этой теме посвящено немало исследований.

Так, например, в своей работе [1] А. А. Белоглазов в результате анализа обратной связи выявил перечень проблем, связанных с рассматриваемыми онлайн-курсами. На основании отзывов можно анализировать их эмоциональную окраску, как показано в работе Кириной М. А. и Дюличевой Ю. Ю. [2, 3]. За рубежом анализу обратной связи посвящен целый ряд исследований [4-6], в них авторы также изучают эмоциональность отзывов и формируют выводы на основании анализа соответствующих текстов.

Для администрирования и управления образовательными треками в Тюменском государственном университете Управление индивидуальных образовательных траекторий осуществляет добровольный сбор обратной связи студентов о пройденных ими элективных дисциплинах с помощью интеллектуального веб-сервиса для сбора отзывов на дисциплины «Отзывус» [7]. Однако для формирования представления о процессах, связанных с реализацией дисциплины, и принятия системных решений по управлению этими процессами нужно не только собирать, но и обрабатывать, анализировать данную обратную связь, на что сейчас

тратятся ощутимые временные и человеческие ресурсы. Появляется необходимость в отсутствующем на сервисе в данный момент аналитическом инструменте, частью которого было бы выявление недостоверных отзывов. Именно решению этой проблемы и посвящено настоящее исследование.

Таким образом, целью работы является разработка и исследование методов автоматического выявления недостоверных отзывов.

Задачи исследования:

1. Исследовать имеющиеся решения анализа обратной связи.
2. Провести анализ собранных данных, извлечь признаки из отзывов.
3. Разработать и исследовать методы выявления недостоверных отзывов.

Разработанные модели и методы будут программно реализованы и включены в аналитический инструмент, что позволит на основании собранных отзывов получить необходимые данные, что, в свою очередь, позволит по результатам их обработки и анализа формировать выводы об элективном образовательном пространстве.

Обзор литературы. В качестве одного из направлений исследований обратной связи на онлайн-курсы является сентимент-анализ, или анализ эмоциональной окраски, отражающий настроения текстов и помогающий выявить негативные и позитивные аспекты дисциплины. В работах [8, 9] анализируется эмоциональная окраска на основе нейросетевых моделей, таких как CNN, LSTM, BiLSTM, ALBERT. В качестве источника данных для исследования Ванга брались китайские информационные курсы, полученные с помощью краулера, а в качестве метрики рассматривалась точность (accuracy). Итоговая метрика составила 0.916 при использовании архитектуры трансформера ALBERT для получения векторного представления текста с BiLSTM для извлечения эмоциональной окраски. Для работы З. Кастрати использовался датасет англоязычных онлайн-дисциплин Coursera, хоть в исследовании рассматривались более традиционные методы векторизации, такие как FastText, GloVe, Word2Vec, а также искусственные нейронные сети CNN, LSTM, наивысший показатель F1-меры показал результат в 0.821 (CNN) и 0.819 (LSTM).

В последние годы также уделялось внимание проблеме метода выявления фейковых отзывов, что иллюстрируется следующими работами [10-12]. Первые два исследования берут за основу источник данных "YelpCHI" с отзывами на рестораны. В научных трудах М. Элмоги и С. Ванга рассматриваются классические подходы векторизации текстов с помощью TF-IDF, Doc2Vec и классификации с помощью KNN, Naive bayes, SVM. Отличия данных работ заключается не только в итоговых метриках F1-меры (KNN, Naive bayes, SVM соответственно): 0.813; 0.804; 0.808 у первого автора, 0.793; 0.697; 0.789 у второго автора, но и в более углубленном подходе описания раздела извлечения признаков из отзывов в работе С. Ванга. Результаты исследования Р. Мохавеша выделяется рассмотрением несбалансированного датасета "Yelp Consumer Electronic" и датасета, построенного с помощью краулера сервисов "TripAdvisor" и "Amazon Mechanical Turk". Вдобавок автор взял во внимание модели архитектуры трансформер (BERT, DistilBERT, RoBERTa), имея высшую F1-меру среди исследуемых моделей: у первого датасета — 0.615, у второго датасета — 0.905.

Результаты работ, представленных выше, не могут быть в предложенном виде применены для решения проблемы выявления фейковых отзывов, так как, рассматриваемая в текущем

исследовании, предметная область (образование), в силу своих особенностей, не позволяет полагаться на часть из признаков, приведенных в этих статьях.

Основными технологиями, используемыми для анализа отзывов в данной работе являются методы векторизации, кластеризации и классификации. Упомянутые методы были доступны для работы благодаря фундаментальному труду [13].

Материалы и методы. Исследование базировалось на анализе 1096 объектов «отзыв». Объект «отзыв» характеризуется множеством атрибутов (признаков), включающем в себя:

- Дисциплина, на которую оставлен отзыв (идентификатор дисциплины, на которую оставлен отзыв, в текстовом формате).
- Дата оставления отзыва в формате (в формате ДД:ММ:ГГГГ).
- Кортеж значения оценок «Полезность», «Интересность», «Легкость» (по пятибалльной порядковой шкале).
- Текст отзыва на естественном языке (неструктурированный).

По результатам предварительного анализа для каждого объекта «отзыв» выделены следующие атрибуты (признаки):

- Длина отзыва (число символьных токенов).
- Количество предложений, содержащих положительную, нейтральную или негативную эмоциональную окраску, в тексте отзыва.

Для обучения модели классификации была получена размеченная выборка. Такая выборка является результатом кластеризации, которая провалидирована ручной разметкой исходных данных. Для кластеризации в свою очередь потребовались векторные представления объектов. Векторные представления получены с помощью различных методов векторизации текстов отзывов, а также с помощью получения векторов из прочих (нетекстовых) признаков.

Для векторизации текстов использовались классические методы, основанные на прямом частотном анализе токенов: TF-IDF[14], Bag of words[15]; и на основе нейросетевых моделей: Word2Vec[16], GloVe[17], FastText[18].

Для кластеризации использовались следующие методы: KMeans, MiniBatchKMeans, DBSCAN, Agglomerative clustering[19]. Первые два метода использовались по причине их универсальности, получения сбалансированных кластеров и ограничений работы с небольшим количеством кластеров. Третий метод имеет неплоскую геометрию в своих результатах, может работать с несбалансированными кластерами, а также имеет особый подход к выбросам. Последний метод может работать с большим количеством кластеров и также с несбалансированными размерами кластеров. Для оценки числа кластеров использовался метод «силуэт» [20].

Для построения классификаторов использовались нейросетевые модели.

Анализ данных проводился с помощью инструментов свободно распространяемых библиотек SciKit Learn, Natural Language Tool Kit, PyTorch.

Результаты. Решаемая проблема может быть поставлена в следующем формальном виде:

Имеется конечное множество отзывов из N элементов, каждому из которых соответствует один класс из конечного множества всех возможных классов (из M элементов) для объекта «отзыв». В нашем случае по признаку недостоверности. Задача заключается в поиске единственного класса для каждого отзыва.

$F = \{ f_i \}_{i=1}^{N}$ — множество объектов «отзыв»

$C = \{ c_j \}_{j=1}^{M}$ — множество меток возможных классов отзыва f_i

Таким образом, каждому объекту «отзыв» соответствует только одна метка класса. Представить это можно следующим образом $F^* = \{(f_i, c_j)\}_{i=1}^{N}$.

Существует целевая зависимость $c^*: F \rightarrow C$, определенная только для обучающей выборки: $train = \{(f_t, c_j)\}_{t=1}^{T} \in F^*$, где $T < N$.

Требуется построить функцию-классификатор $a: F \rightarrow C$ произвольных объектов «отзыв» $f \in F$, минимизирующую функцию потерь (для бинарного классификатора — это перекрестная энтропия):

$$CE = -\sum_{j=1}^{M}(c_j * log(c_j{}^*)),$$

где c_j — фактические метки классов; $c_j{}^*$— предсказанные вероятности.

Для получения множества F был проведен этап векторизации, на котором решались две подзадачи: векторизация текстов и векторизация нетекстовых признаков.

Способ векторизации текстов заключается в рассмотрении в исходных данных только одного признака — текста отзыва. В тексте отзыва могут содержаться аспекты, которые могли бы явно отличать недостоверные отзывы от остальных.

Получение векторов из остальных (нетекстовых) признаков сводится к отбору наиболее значимых признаков, а также составлению или получению из различных признаков новых. Далее проводилась нормализация всех признаков с целью приведения к общей шкале, при этом с сохранением их распределения.

Размерности векторов, полученные с помощью векторизации текста, имеют разные значения, так у Bag of words и TF-IDF она составила 11518, у Word2Vec и GloVe — 300, у FastText — 16.

По итогам векторизации объектов «отзыв» способом, не включающим в себя текст отзыва, получилось векторное представление размерностью 12, содержащее как атрибуты, описанные выше, так и следующие атрибуты:

- Сумма трех оценок «Полезность», «Интересность», «Легкость».
- Бинарный признак, отражающий крайние значения суммы выставленных оценок (сумма оценок = 15 или сумма оценок = 3).
- Отношение количества предложений с позитивной эмоциональной окраской к количеству предложений во всем тексте отзыва.
- Отношение количества предложений с негативной эмоциональной окраской к количеству предложений во всем тексте отзыва.
- Бинарный признак, отражающий характер периода оставления отзыва (отзыв оставлен в «обычный» период или нет).

Характер распределения признаков, представленных на рис. 1, показывает целесообразность их использования в следующем этапе исследования — кластеризации.

При кластеризации полученные наивысшие метрики «силуэта» изображены в табл. 1.

Метрики «силуэта» на табл. 1 были получены путем перебора значений классов от 2 до 10 для соответствующего метода, кроме DBSCAN, так как этот метод сам определяет количество кластеров в зависимости от гиперпараметров. Также стоит отметить, что в таблице все метрики посчитаны на векторах, чьи признаки были сокращены до 2 методом главных

компонент, так как при подсчете на исходных векторных представлениях значения были слишком малы для их рассмотрения. Вектора, построенные на нетекстовых признаках, показали низкие значения из-за характера их распределения на плоскости, что представлено на рис. 2.

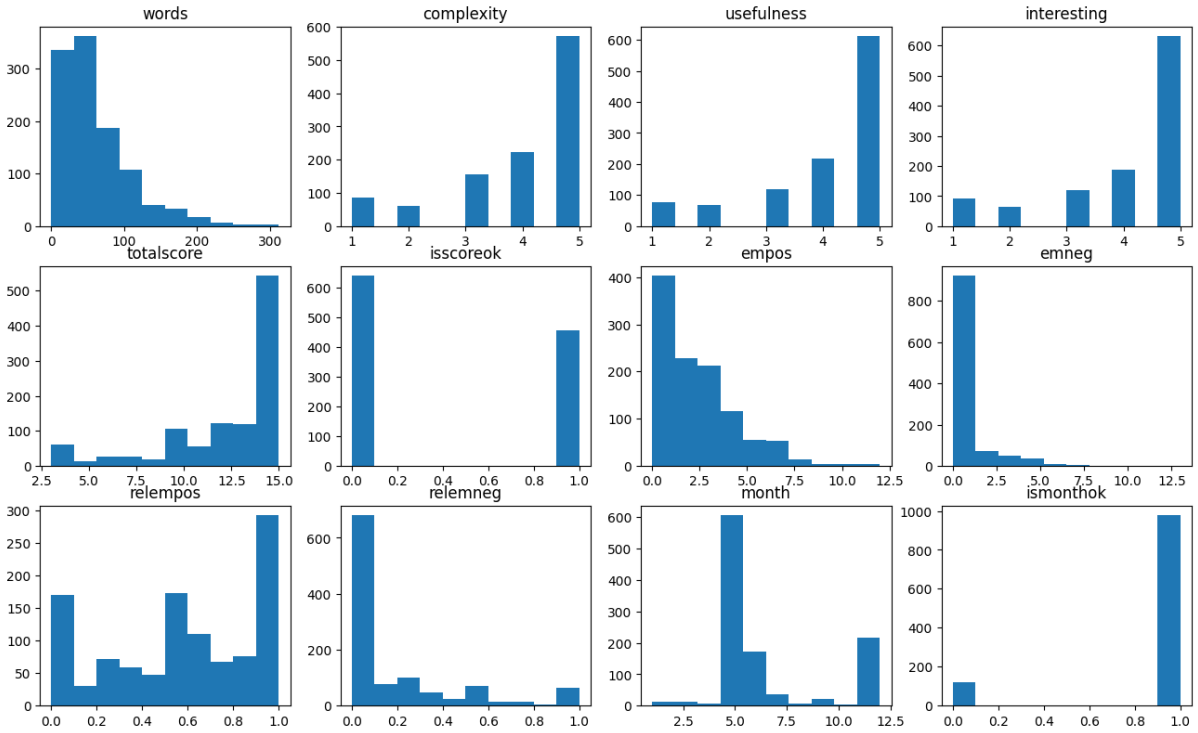

Рис. 1. Плотность распределения признаков векторного представления «отзыв»

Для оценки качества кластеризации была проведена ручная разметка 1096 текстов отзывов, по результатам которой было выявлено 994 обычных и 102 недостоверных отзыва.

Таблица 1

Значение метрик «силуэт» и оптимальное количество кластеров в зависимости от методов кластеризации и векторизации

Метод векторизации	*Метод кластеризации*			
	KMeans	*MiniBatchKMeans*	*DBSCAN*	*Agglomerative clustering*
	Значение метрики «силуэт» / Количество кластеров			
Bag of words	0.986 / 2	0.668 / 2	0.958 / 2	0.667 / 2
TF-IDF	0.912 / 2	0.729 / 4	0.866 / 2	0.867 / 2
Word2Vec	0.840 / 2	0.500 / 2	0.715 / 2	0.721 / 2
GloVe	0.517 / 3	0.517 / 3	0.152 / 2	0.317 / 2
FastText	0.306 / 3	0.314 / 2	0.733 / 2	0.792 / 2
Нетекстовые признаки	0.547 / 2	0.350 / 2	0.200 / 3	0.392 / 2

В сравнении со всеми методами кластеризации больше всего выделяется DBSCAN с векторизацией FastText, где 1078 обычных и 18 недостоверных отзывов. На рис. 3 также видна близость результатов ручной разметки и применения данного метода (на левом графике —

применение метода DBSCAN с FastText, на правом графике — результат ручной разметки). Можно сформулировать вывод о том, что можно обходиться при кластеризации без дополнительных признаков и опираться только на текст.

Итоговый состав кластеров с векторными представлениями, составленными из нетекстовых признаков, степень различий варьируется в зависимости от методов кластеризации, однако общая тенденция присутствует в разнице следующих показателей: трех оценок «Полезность», «Интересность», «Легкость»; отношение положительных и негативных предложений ко всем предложениям; бинарный признак, отражающий характер периода оставления.

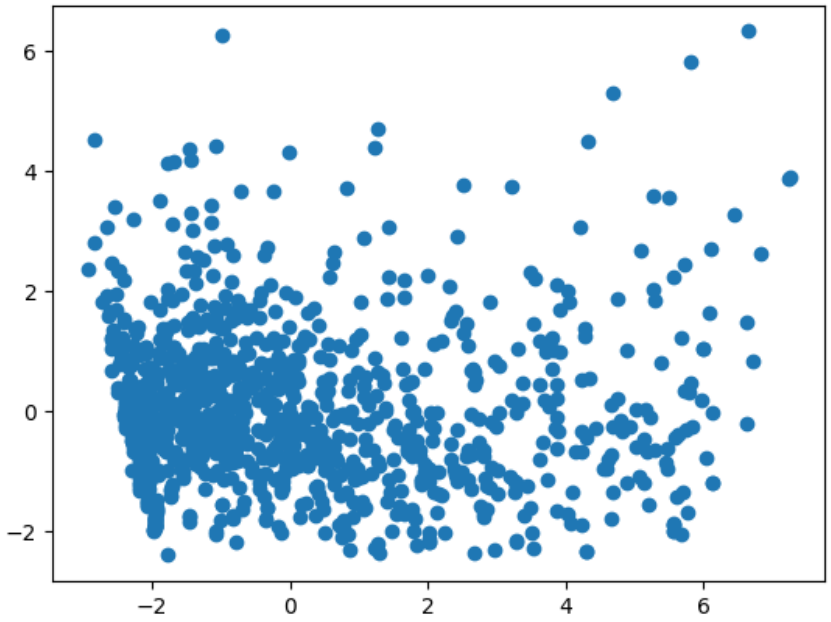

Рис. 2. Распределение векторных представлений нетекстовых признаков отзыва на плоскости

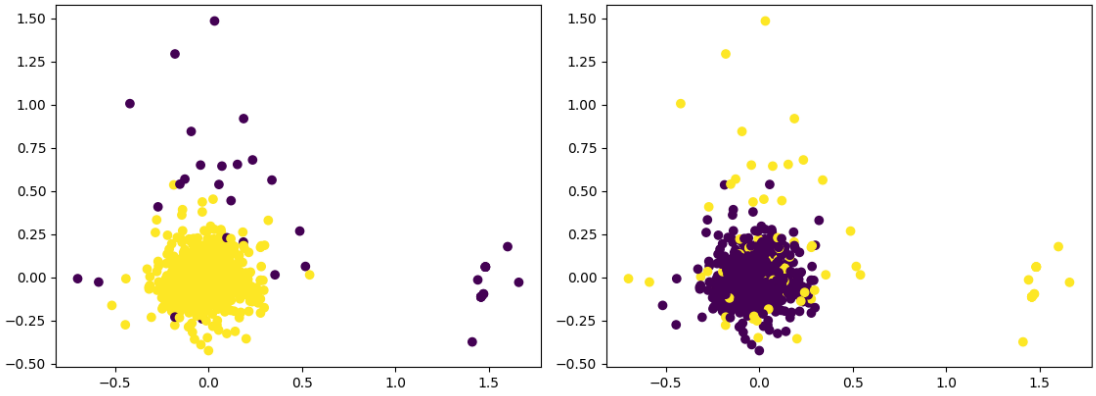

Рис. 3. Распределение векторных представлений текстов отзыва с ручной разметкой и с методом кластеризации DBSCAN с методом векторизации FastText

На заключительном этапе разработки классификаторов были использованы нейросетевые модели CNN и RNN, а также модели, основывающиеся на архитектуре Transformer: RuBERT, RuRoBERTa. Первые две были использованы для оценки их результативности в задаче поиска недостоверных отзывов при малом размере датасета. Модели на основе

архитектуры Transformer были выбраны для проверки гипотезы о том, что крайне значимым для обнаружения недостоверных отзывов является глубокое понимание контекста, а также по причине оптимизации в этих моделях процессов вычислений (за счет их распараллеливания), в отличие от RNN.

Для оценки качества классификации была выбрана метрика macro f1-score, потому что этот способ вычисления f1-score учитывает пропорции распределения меток классов в наборе данных. В результате обучения сверточной и рекуррентной нейросетей были получены значения f1-score в 0.166 и 0.786 соответственно, что показывает большую приспособленность RNN к работе с классификацией текстов.

В табл. 2 представлены метрики качества предобученных моделей, основанных на архитектуре Transformers, показывающие, что неспециализированные модели без дообучения нельзя использовать для классификации «недостоверных» отзывов.

Таблица 2

Результаты классификации отзывов по признаку фейковости предобученными моделями на архитектуре Transformers

Модель	*Precision*	*Recall*	*F1-score*
DeepPavlov/rubert-base-cased-sentence	0.549	0.638	0.508
DeepPavlov/distilrubert-tiny-cased-conversational	0.718	0.558	0.578
DeepPavlov/distilrubert-tiny-cased-conversational-v1	0.905	0.611	0.660
DeepPavlov/rubert-base-cased	0.610	0.530	0.535
DeepPavlov/rubert-base-cased-conversational	0.450	0.437	0.169
DeepPavlov/distilrubert-small-cased-conversational	0.339	0.454	0.093
DeepPavlov/distilrubert-base-cased-conversational	0.437	0.363	0.227

Для дообучения была выбрана модель DeepPavlov/distilrubert-tiny-cased-conversational-v1, показавшая наилучшие результаты f1-score. В течение 6 эпох дообучения было зафиксировано состояние модели с итоговым показателем f1-score = 0.922, при precision = 0.988, recall = = 0.875.

Таким образом, по результатам исследования был разработан метод выявления «недостоверных» текстов отзывов с помощью модели классификации на базе Transformer с итоговым показателем f1-score = 0.922. Метки обучающей выборки базируются на методе кластеризации DBSCAN, которая была провалидирована ручной разметкой, и векторизована с помощью метода FastText.

Заключение. В ходе исследования был проведен обзор существующих работ по анализу обратной связи, по результатам которого была подтверждена актуальность данной работы в рамках изучения обратной связи. В рамках работы выполнен анализ имеющегося набора данных на веб-сервисе «Отзывус», для дальнейшей разработки методов выявления «недостоверных» отзывов уделено внимание задаче по выделению признаков из объектов «отзыв». Результатом работы стала модель классификации, полученная путем применения комплексного метода, включающего этапы векторизации, кластеризации и дообучении модели Transformer.

В дальнейшем для повышения качества значения итоговых метрик классификации будет дополнена обучающая выборка отзывов, а также рассмотрен более широкий спектр методов и подходов основных этапов настоящего исследования.

СПИСОК ЛИТЕРАТУРЫ

1. Образовательные технологии онлайн-обучения: анализ массовых открытых онлайн-курсов российских вузов / А.А. Белоглазов, Л.Б. Белоглазова, И.А. Белоглазова [и др.]. — Текст: электронный // Вестник Московского городского педагогического университета. — 2018. — № 4 (46). — С. 50-57. — URL: https://elibrary.ru/item.asp?id=36527220 (дата обращения: 16.02.2024).
2. Кирина М.А. Автоматическая оценка впечатлений обучающихся методами анализа тональности (на материале отзывов на онлайн-курсы на русском и английском) / М.А. Кирина, Л.Д. Тельнина; под ред. В.В. Рубцова, М.Г. Сороковой, Н.П. Радчиковой. — Текст: непосредственный // Цифровая гуманитаристика и технологии в образовании (DHTE 2022): сб. статей III Всероссийской научно-практической конференции с международным участием. 17–18 ноября 2022 г. — Москва: Издательство ФГБОУ ВО МГППУ, 2022. — 355–374 с. — ISBN 978-5-94051-275-2.
3. Дюличева Ю.Ю. Учебная аналитика МООК как инструмент анализа математической тревожности / Ю.Ю. Дюличева. — Текст: электронный // Вопросы образования. — 2021. — № 4. — С. 243-265. — URL: https://doi.org/10.17323/1814-9545-2022-4-298-321 (дата обращения: 16.02.2024).
4. Adamopoulos P. What makes a great MOOC? An interdisciplinary analysis of student retention in online courses. / P. Adamopoulos. — Текст: электронный. — 2013. — URL: https://core.ac.uk/reader/301361240 (дата обращения: 16.02.2024).
5. Almatrafi O. Systematic review of discussion forums in massive open online courses (MOOCs) / O. Almatrafi, A. Johri. — Текст: электронный // IEEE Transactions on Learning Technologies. — 2018. — Т. 12, № 3. — С. 413-428. URL: https://ieeexplore.ieee.org/document/8418792 (дата обращения: 18.02.2024).
6. Onan A. Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach / A. Onan. — Текст: электронный // Computer Applications in Engineering Education. — 2021. — Т. 29. — № 3. — С. 572-589. — URL: https://onlinelibrary.wiley.com/doi/10.1002/cae.22253 (дата обращения: 18.02.2024).
7. Разработка сервиса для сбора и анализа отзывов на элективные дисциплины / Д.Д. Криворогов, Т.Д. Низамов, А.А. Фазлыев [и др.]. — Текст: электронный // Вестник НГУ. Серия: Информационные технологии. — 2023. — № 21 (3). — С. 5-19. — URL: https://doi.org/10.25205/1818-7900-2023-21-3-5-19 (дата обращения: 16.02.2024).
8. Kastrati Z. Weakly supervised framework for aspect-based sentiment analysis on students' reviews of MOOCs / Z. Kastrati, A.S. Imran, A. Kurti. — Текст: электронный // IEEE Access. — 2020. — Т. 8. — С. 106799-106810. — URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9110884&tag=1 (дата обращения: 19.02.2024).
9. Wang C. Sentiment analysis of MOOC reviews via ALBERT-BiLSTM model / C. Wang, S. Huang, Y. Zhou. — Текст: электронный // MATEC Web of Conferences. — EDP Sciences, 2021. — Т. 336. — С. 05008. URL: https://doi.org/10.1051/matecconf/202133605008 (дата обращения: 19.02.2024).
10. Fake reviews detection: A survey / R. Mohawesh, S. Xu, SN. Tran [et al.]. — Текст: электронный // IEEE Access. — 2021. — Т. 9. — С. 65771-65802. — URL: https://ieeexplore.ieee.org/abstract/document/9416474 (дата обращения: 16.02.2024).

11. Fake reviews detection using supervised machine learning / A. M. Elmogy, U. Tariq, A. Mohammed, A. Ibrahim. — Текст: электронный // International Journal of Advanced Computer Science and Applications. — 2021. — Т. 12. — № 1. URL: https://dx.doi.org/10.14569/IJACSA.2021.0120169 (дата обращения: 18.02.2024).
12. Fake review detection based on multiple feature fusion and rolling collaborative training / J. Wang, H. Kan, F. Meng [et al.]. — Текст: электронный // IEEE access. — 2020. — Т. 8. — С. 182625-182639. URL: https://doi.org/10.1109/ACCESS.2020.3028588 (дата обращения: 18.02.2024).
13. Scikit-learn: Machine learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort [et al.]. — Текст: электронный // the Journal of machine Learning research. — 2011. — Т. 12. — С. 2825-2830. URL: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https:/ (дата обращения: 09.03.2024).
14. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval / K. Sparck Jones. — Текст: электронный // Journal of documentation. — 1972. — Т. 28. — № 1. — С. 11-21. URL: https://doi.org/10.1108/eb026526 (дата обращения: 10.03.2024).
15. Sivic J. Efficient visual search of videos cast as text retrieval / J. Sivic, A. Zisserman. — Текст: электронный // IEEE transactions on pattern analysis and machine intelligence. — 2008. — Т. 31. — № 4. — С. 591-606. URL: https://doi.org/10.1109/tpami.2008.111 (дата обращения: 10.03.2024).
16. Efficient estimation of word representations in vector space / T. Mikolov, K. Chen, G. Corrado, J. Dean. — Текст: электронный // arXiv preprint arXiv:1301.3781. — 2013. URL: https://doi.org/10.48550/arXiv.1301.3781 (дата обращения: 12.03.2024).
17. Pennington J. Glove: Global vectors for word representation / J. Pennington, R. Socher, C. D. Manning. — Текст: электронный // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). — 2014. — С. 1532-1543. URL: https://nlp.stanford.edu/pubs/glove.pdf (дата обращения: 10.03.2024).
18. Enriching word vectors with subword information / P. Bojanowski, E. Grave, A. Joulin, T. Mikolov. — Текст: электронный // Transactions of the association for computational linguistics. — 2017. — Т. 5. — С. 135-146. URL: https://arxiv.org/pdf/1607.04606.pdf (дата обращения: 10.03.2024).
19. Watson J. NLP with Python: 3 Books in 1 — "From Beginner to Advanced: The Future Frontier and Next-Gen Solutions": учебное пособие / J. Watson. — 2023. — 424 с. — ICBN B0CN459GWB. — Текст: непосредственный.
20. Multilingual neural machine translation with language clustering X. Tan, J. Chen, D. He [et al.]. — Текст: электронный // arXiv preprint arXiv:1908.09324. — 2019. — URL: https://arxiv.org/pdf/1908.09324 (дата обращения: 10.03.2024).