MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 29, NO. 3, 2022
journal homepage: www.mais-journal.ru

THEORY OF DATA

Classification of Articles from Mass Media by Categories and Relevance of the Subject Area

V. D. Larionov[1], I. V. Paramonov[1]

DOI: 10.18255/1818-1015-2022-3-266-279

[1]P. G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.

MSC2020:
Research article
Full text in Russian

Received June 5, 2022
After revision August 23, 2022
Accepted August 26, 2022

The research is devoted to classification of news articles about P. G. Demidov Yaroslavl State University (YarSU) into 4 categories: “society”, “education”, “science and technologies”, “not relevant”.
The proposed approaches are based on using the BERT neural network and methods of machine learning: SVM, Logistic Regression, K-Neighbors, Random Forest, in combination of different embedding types: Word2Vec, FastText, TF-IDF, GPT-3. Also approaches of text preprocessing are considered to achieve higher quality of the classification. The experiments showed that the SVM classifier with TF-IDF embedding and trained on full article texts with titles achieved the best result. Its micro-F-measure and macro-F-measure are 0.8214 and 0.8308 respectively. The BERT neural network trained on fragments of paragraphs with YarSU mentions, from which the first 128 words and the last 384 words were taken, showed comparable results. The resulting micro-F-measure and macro-F-measure are 0.8304 and 0.8181 respectively. Thus, using paragraphs with the target organisation mentions is enough to classify text by categories efficiently.

Keywords: classification by categories; automatic text processing; subject area; Russian language; news articles

INFORMATION ABOUT THE AUTHORS

Vladislav Dmitrievich Larionov	orcid.org/0000-0002-5591-8332. E-mail: vladlarionov998@gmail.com Student.
Ilya Vyacheslavovich Paramonov correspondence author	orcid.org/0000-0003-3984-8423. E-mail: ilya.paramonov@fruct.org PhD, associate professor.

Funding: This work was supported by P. G. Demidov Yaroslavl State University Project No. VIP-016.

For citation: V. D. Larionov and I. V. Paramonov, “Classification of Articles from Mass Media by Categories and Relevance of the Subject Area”, *Modeling and analysis of information systems*, vol. 29, no. 3, pp. 266-279, 2022.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 29, № 3, 2022
сайт журнала: www.mais-journal.ru

THEORY OF DATA

Классификация статей из средств массовой информации по категориям и релевантности предметной области

В. Д. Ларионов[1], И. В. Парамонов[1]

DOI: 10.18255/1818-1015-2022-3-266-279

[1]Ярославский государственный университет им. П. Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 5 июня 2022 г.
После доработки 23 августа 2022 г.
Принята к публикации 26 августа 2022 г.

Исследование посвящено классификации новостных статей о Ярославском государственном университете им. П. Г. Демидова (ЯрГУ) на 4 категории: общество, образование, наука и технологии, нерелевантная.
Предложенные подходы основаны на нейронной сети BERT и методах машинного обучения SVM, Logistic Regression, K-Neighbors, Random Forest в сочетании с эмбеддингами различных видов: Word2Vec, FastText, TF-IDF, GPT-3. Также предложены способы предобработки текстов для достижения более высокого качества классификации. В ходе экспериментов установлено, что лучше всего с задачей справляется SVM-классификатор с эмбеддингом TF-IDF, обученный на полных текстах статей с заголовками. Его значения микро- и макро-F-меры достигают 0.8214 и 0.8308 соответственно. Сопоставимые результаты показывает нейронная сеть BERT, обученная на фрагментах абзацев с упоминанием ЯрГУ, из которых брались 128 слов из начала и 384 слова из конца. Её показатели микро- и макро-F-меры достигают 0.8304 и 0.8181 соответственно. Таким образом, установлено, что абзацев с упоминанием конкретной организации оказывается достаточно, чтобы классификация по категориям была эффективной.

Ключевые слова: классификация по категориям; автоматическая обработка текстов; предметная область; русский язык; новостные статьи

ИНФОРМАЦИЯ ОБ АВТОРАХ

Владислав Дмитриевич Ларионов
orcid.org/0000-0002-5591-8332. E-mail: vladlarionov998@gmail.com
студент.

Илья Вячеславович Парамонов
автор для корреспонденции
orcid.org/0000-0003-3984-8423. E-mail: ilya.paramonov@fruct.org
канд. физ.-мат. наук, доцент.

Финансирование: Работа выполнена в рамках инициативной НИР ЯрГУ им. П. Г. Демидова № VIP-016.

Для цитирования: V. D. Larionov and I. V. Paramonov, “Classification of Articles from Mass Media by Categories and Relevance of the Subject Area”, *Modeling and analysis of information systems*, vol. 29, no. 3, pp. 266-279, 2022.

Введение

Каждый день тысячи новостных ресурсов генерируют сотни тысяч статей о различных событиях в мире. Редакторы многих ресурсов перед публикацией вручную назначают статьям ключевые слова, теги и категории, чтобы пользователи могли легче фильтровать и ранжировать контент. Однако не всем потребителям контента подходят выставленные редакторами категории. Например, PR-специалисты, задача которых отслеживать информационный фон вокруг определённых компаний, как правило, формируют собственные категории и критерии для их назначения на основе не всей статьи в целом, а только фрагментов статьи, которые относятся к конкретной компании или человеку. При создании автоматизированных систем, ориентированных на таких специалистов, требуется назначать категории текстам автоматически, в связи с чем актуальной оказывается задача автоматической классификации с учётом предметной области.

В данной работе рассматриваются различные подходы по классификации русскоязычных новостных статей, которые посвящены Ярославскому государственному университету им. П.Г. Демидова (ЯрГУ), на 4 категории, определённые специалистом PR-службы: «общество», «образование», «наука и технологии», «нерелевантная». Текст относится к определённой категории, если содержание всего текста или его отдельных фрагментов относится непосредственно к ЯрГУ и соответствует критериям, сформулированным специалистом PR-службы.

Рассмотрим несколько примеров, проясняющих принцип классификации с учётом предметной области. Рассмотрим статью, в которой говорится о том, что рядом с главным зданием ЯрГУ прошёл парад. Статья посвящена непосредственно параду, а не ЯрГУ, главное здание которого в статье используется лишь в качестве ориентира для указания места проведения. Следовательно, статья не содержит ценной информации для PR-службы и будет считаться нерелевантной.

Рассмотрим ещё один пример статьи, в которой говорится, что число заболевших вирусом COVID-19 неуклонно растёт, вследствие чего многие вузы региона, среди которых ЯрГУ, закрываются на карантин и переходят на удалённое обучение. Несмотря на то, что эта статья о вирусе и о его распространении, ей будет назначена категория «образование», поскольку в статье говорится о смене формы обучения в вузе — с очной на удалённую.

Аналогичный ход рассуждения применяется к статье, в которой журналист обсуждает с сотрудником ЯрГУ некоторое экономическое явление или экономическое положение в стране и ссылается на него как на эксперта. Поскольку в статье сотрудника ЯрГУ называют экспертом, он является лидером мнений и, следовательно, упоминание его имени и аффилиации с вузом повышает рейтинг вуза в глазах общественности. Кроме того, PR-служба может отслеживать наличие и корректность аффилиации в качестве одного из критериев выбора СМИ для продвижения организации. Таким образом, данная статья будет представлять ценность для PR-службы и будет иметь категорию «общество».

1. Обзор связанных работ

Задача классификации новостных статей по категориям — весьма распространённая задача в компьютерной лингвистике. Больших успехов в решении данной задачи достигли A. Hussain с соавторами [1], которые классифицировали англоязычные новостные статьи, используя категории «спорт», «политика», «технологии», «бизнес», «развлечения», без учёта предметной области. В качестве инструмента классификации авторы использовали методы машинного обучения, среди которых: SVM-классификатор, метод K-ближайших соседей и другие. Эксперименты показали высокие значения точности и полноты — от 0.91 и выше.

Сопоставимые результаты получили G. Kaur с соавторами [2] при классификации англоязычных статей BBC без учёта предметной области на 4 категории: «спорт», «здоровье», «бизнес», «развлечения». Перед классификацией авторы предобрабатывали тексты следующим образом: в текст

статей добавляли их заголовки, разбивали тексты на отдельные слова, удаляли стоп-слова, добавляли в текст синонимы слов и стеммировали слова. Далее предобработанные тексты использовались для классификации с помощью SVM-классификатора. Результаты исследования показали достаточно высокое значение метрики accuracy для каждой категории — от 0.9129 до 0.9354.

Чуть худших результатов добились P. Semberecki и H. Maciejewski [3], которые классифицировали англоязычные статьи с Wikipedia по более широкому набору категорий: «искусство», «история», «юриспруденция», «медицина», «религия», «спорт», «технологии». При классификации авторы не выделяли конкретную предметную область, а классифицировали статьи целиком. Предварительно авторы переводили тексты статей в вектора с помощью Word2Vec, затем классифицировали их с помощью нейронных сетей LSTM и CNN. Средняя точность определения категорий составила 86.21 %.

Интересные результаты получил X. Luo [4], классифицируя англоязычные наборы текстов по категориям, причём каждый набор текстов имел собственный перечень категорий. Первый набор — «женщины», «спорт», «литература», «университет», второй набор — «спорт», «созвездие», «игра», «развлечения» и третий набор — «наука и техника», «мода», «текущие события». Для классификации автор использовал SVM-классификатор, наивный байесовский классификатор и логистическую регрессию. С помощью SVM-классификатора были получены высокие показатели F-меры для первых двух наборов категорий. Значения варьируется в пределах от 0.71 до 0.86. Для третьего набора показатели метрик значительно ниже — от 0.18 до 0.63. Это может быть связано с тем, что тексты категории «мода» и «текущие события» могут сильно различаться между собой даже внутри своей категории.

Таким образом, для английского языка задача достаточно хорошо решается. Применительно к текстам на русском языке также есть исследования, в которых достигаются высокие результаты. Например, в работе [5] сравниваются нейронная сеть RuBERT и методы машинного обучения в сочетании с TF-IDF в классификации русскоязычных новостных статей на 6 категорий: «спорт», «политика», «общество», «культура», «инциденты», «экономика». При классификации авторы не учитывали предметную область. Лучшие результаты показывают SVM-классификатор и нейронная сеть RuBERT, значение точности и полноты которых достигают 0.80 и выше для каждой из категорий, за исключением категории «общество». Значение метрик данной категории варьируется в пределах от 0.65 до 0.75.

Более высоких результатов удалось добиться в работе [6], где сравнивались методы машинного обучения в сочетании с TF-IDF в классификации русскоязычных новостных статей без учёта предметной области на 3 категории: «технологии и медиа», «политика», «экономика, финансы, бизнес». Для сравнения автор выбрал методы машинного обучения, среди которых наивный байесовский классификатор, Random Forest, классификатор на основе градиентного бустинга и классификатор на основе стохастического градиентного спуска (SGD-классификатор). Лучший результат показал SGD-классификатор, который определял категории «политика» и «экономика, финансы, бизнес» с показателями 0.9283 и 0.9124 F-меры соответственно. С категорией «технологии и медиа» классификатор справился значительно хуже и достиг значения 0.7546 F-меры. Такая разница в результатах между категориями может быть связана с дисбалансом обучающего корпуса, содержащего всего 51 857 текстов, из которых: 24 028 с категорией «экономика, финансы и бизнес», 23 066 с категорией «политика» и 4763 с категорией «технологии и медиа».

Аналогичных результатов добилась Е. Н. Каруна [7], сравнивая нейронные сети в классификации русскоязычных новостных статей без учёта предметной области на 10 категорий: «Украина», «расследования», «экономика», «инциденты», «футбол», «бизнес», «музыка», «люди», «наука», «фильмы». Перед классификацией, тексты предобрабатывались следующим образом: из текста удалялись все неалфавитные символы, текст разделялся на слова, которые лемматизировались,

удалялись стоп-слова. Для сравнения авторы выбрали такие нейронные сети, как FFNN, CNN, BiLSTM, LSTM. Для нейронной сети FFNN авторы также сравнивали влияние *n*-грамм, где *n* от 1 до 3. Лучший результат показала нейронная сеть BiLSTM и достигла значения 0.905 по метрике accuracy. Лучше всего классифицировались такие категории, как «Украина», «футбол», «наука», «фильмы». Значения метрики accuracy для данных категорий варьируется в пределах от 0.92 до 0.97. Хуже всего классифицируется категория «бизнес», её показатели изменяются в пределах от 0.66 до 0.79. Это может быть связано с тем, что содержание статей данной категории трудно отличать от статей категории «экономика».

Из представленных исследований видно, что высокие результаты достигаются на категориях, которые достаточно хорошо различаются и тексты которых имеют схожие ключевые слова и контекст. В тоже время более широкие категории, например, «общество», классифицируются заметно хуже. Кроме того, во всех исследованиях тексты классифицируются по категориям, которые не привязаны к предметной области.

2. Постановка задачи

Рассмотрим следующую задачу. Дана статья на русском языке, включающая в себя упоминание Ярославского государственного университета им. П. Г. Демидова (ЯрГУ), причём название может встречаться в полной или одной из сокращённых форм в любом падеже. Статья включает в себя заголовок и собственно текст статьи. Необходимо отнести данную статью к одной из следующих категорий: «общество», «образование», «наука и технологии», «нерелевантная».

Критерии принадлежности статьи каждой из категорий сформулированы специалистом PR-службы ЯрГУ и выглядят следующим образом.

Статья относится к категории «образование», если:

- в статье говорится о жизни в вузе, формате обучения, поступлении, стоимости обучения;
- в статье говорится о том, что сотрудники вуза проводят занятия, не связанные с вузом (лектории или вебинары);
- в статье говорится о достижениях и успехах студентов, сотрудников вуза или учеников сотрудников вуза;
- в статье говорится о связанном с учёбой сотрудничестве вуза с какой-то компанией или каким-то человеком;
- в статье говорится об участии вуза, сотрудника вуза или студента в мероприятии, программе или проекте, связанными с учёбой;
- в статье говорится о развитии образовательной программы;
- в статье говорится об обучении чему-то на базе вуза;
- в статье говорится об обучении чему-то студентов вуза где-то вне вуза;
- в статье говорится, что вуз вошёл в какой-то рейтинг.

Статья относится к категории «общество», если:

- в статье есть интервью или высказывание сотрудника вуза о каких-то общественных проблемах;
- в статье говорится об участии сотрудников вуза в качестве экспертов;
- в статье говорится о том, что сотрудник вуза разработал или предложил что-то, связанное с городом;
- в статье говорится о происшествии, связанном со студентами вуза;
- в статье говорится об участии вуза, сотрудника вуза или студента в мероприятии, программе или проекте, не связанном с учёбой;
- в статье говорится о назначении или кандидатуре сотрудника вуза на какую-то важную должность в Правительстве;
- в статье говорится о том, что вуз получил грант на что-то не связанное с наукой;

- в статье есть интервью или высказывание студента вуза.

Статья относится к категории «наука и технологии», если:

- в статье говорится о разработке или показе конкретных технологий, проектов;
- в статье говорится о том, что вуз или сотрудник вуза получил грант;
- в статье говорится о том, что вуз занял место в научном рейтинге;
- в статье говорится о планах создания новой цифровой технологии или программы;
- в статье говорится о созданной вузом технологии в любой сфере;
- в статье говорится о технологическом сотрудничестве вуза с компаниями или предприятиями.

Статья считается нерелевантной, если:

- в статье вуз упоминается в качестве учреждения, в котором кто-то учился или работал, но в настоящее время никак не связан с вузом;
- в статье вуз упоминается в качестве точки на карте, ориентира или помещения;
- в статье говорится о происшествии, связанном со студентами вуза, но к которым вуз не имеет отношения.

Несмотря на то, что потенциально статья может относиться к нескольким категориями одновременно (кроме категории «нерелевантная»), анализ доступных данных специалистом PR-службы показал, что на практике подобная ситуация возникает крайне редко, поэтому ей можно пренебречь. Таким образом, рассматриваемая в данной статье задача может быть рассмотрена как задача классификации статей с взаимоисключающими классами.

3. Подход с применением нейронной сети BERT

В качестве одного из инструментов решения поставленной задачи была выбрана нейронная сеть BERT, которая является весьма распространённым инструментом в решении задач компьютерной лингвистики. BERT (Bidirectional Encoder Representations from Transformers) — это двунаправленная нейронная сеть, построенная на архитектуре «трансформер» [8]. Модель на основе BERT заранее обучена на большом числе неразмеченных текстов для решения двух задач: генерации пропущенного токена и определения следующего предложения. Для решения конкретной задачи выполняется тонкая настройка предобученной модели.

Перед загрузкой в BERT входные тексты предобрабатывались и делились одним или несколькими способами, в частности:

- использовался только основной текст статьи;
- использовались только абзацы с упоминанием ЯрГУ;
- в текст статьи добавлялся её заголовок.

Для того, чтобы уложиться в ограничение BERT в 512 слов на один текст, входные тексты усекались одним из следующих способов:

- брались первые 512 слов;
- брались первые 128 слов и последние 384 слова;
- брались первые 256 слов и последние 256 слов.

Для тонкой настройки была использована модель *rubert_cased_L-12_H-768_A-12_v2*, обученная командой DeepPavlov на русскоязычных текстах. Параметр *learning rate* был установлен равным $3 \cdot 10^{-5}$, *train batch size* — 12, количество эпох — 3, максимальная длина последовательности — 512. Реализация алгоритма обучения и классификации была взята из официального репозитория BERT [8].

4. Подход с применением методов машинного обучения

Для решения задачи с применением методов машинного обучения входные тексты сначала предобрабатывались и делились с помощью одного или нескольких подходов:

- использовался только основной текст статьи;
- из текста удалялись стоп-слова;

- использовались только абзацы с упоминанием ЯрГУ;
- в текст статьи добавлялся её заголовок.

Далее тексты переводились в нижний регистр, разбивались на предложения, удалялись знаки препинания, предложения разбивались на слова, которые лемматизировались. Затем тексты преобразовывались в вектора с помощью одного из эмбеддингов: Word2Vec, FastText, TF-IDF, GPT-3. Для эмбединнгов Word2Vec и FastText предобработанные тексты загружались двумя способами: в виде отдельных предложений и в виде цельного текста, без разделения на предложения.

Полученные вектора использовались для обучения классификаторов: SVM, Logistic Regression, K-Neighbors, Random Forest. Реализация классификаторов была взята из библиотеки scikit-learn [9].

Рассмотрим алгоритм решения задачи с каждым видом эмбеддингов подробнее.

4.1. Word2Vec

Word2Vec [10] — это модель векторного представления слов, которая использует контекст соседних слов при формировании векторов. Таким образом, слова, которые используются в одном и том же контексте, будут иметь похожие вектора.

Принцип работы Word2Vec можно описать следующим образом:

1. Word2Vec принимает большой текстовый корпус в качестве входных данных и сопоставляет каждому слову вектор, выдавая координаты слов в тексте на выходе.
2. Генерируется словарь корпуса.
3. Вычисляется векторное представление слов в процессе обучения на входных текстах.

Обучаться Word2Vec может с помощью двух подходов: CBoW (Continuous Bag of Words) — подход, при котором контекст слова используется для предсказания самого слова, и Skip-gram — это подход, при котором текущее слово используется, чтобы угадать соседние слова.

При инициализации Word2Vec размер скользящего окна был установлен равным 8, размер словаря — 70, число итераций — 15, подход обучения — Skip-gram. Реализация Word2Vec взята из библиотеки Gensim [11].

Алгоритм решения задачи с использованием Word2Vec выглядит следующим образом:

1. Весь обучающий корпус в виде отдельных предложений или без деления на предложения загружается в Word2Vec.
2. Каждый текст разбивается на слова и для каждого слова берётся соответствующий ему вектор.
3. Вычисляется среднее значение векторов слов для каждого текста. Таким образом вычисляется характеристика текста.
4. Полученные вектора далее используются при обучении классификатора и собственно классификации текстов.

4.2. FastText

FastText [12] — это модель векторного представления слов, которая похожа на Word2Vec. Одним из отличий является возможность использования негативного сэмплирования при обучении с подходом Skip-gram. Негативное сэмплирование — это способ указать отрицательные примеры при обучении модели, другими словами, учесть при обучении модели пары слов, которые не являются соседними. Ещё одной особенностью является subword-модель — представление слова цепочками длиной от 3 до 6 символов от начала до конца слова плюс само слово целиком. Такое представление слов помогает модели работать со словами, которые ранее не встречались.

При инициализации FastText размер скользящего окна был установлен равным 8, размер словаря — 70, число итераций — 20, подход обучения — Skip-gram, размер subword-модели — 3. Реализация FastText взята из библиотеки Gensim [11].

Алгоритм решения задачи с FastText аналогичен алгоритму с Word2Vec:

1. Весь обучающий корпус в виде отдельных предложений или без деления на предложения загружается в FastText.
2. Каждый текст разбивается на слова и для каждого слова берётся соответствующий ему вектор.
3. Вычисляется среднее значение векторов слов для каждого текста. Таким образом вычисляется характеристика текста.
4. Полученные вектора далее используются при обучении классификатора и собственно классификации текстов.

4.3. TF-IDF

TF-IDF (term frequency-inverse document frequency) [13] — это статистическая мера, с помощью которой можно оценить важность слова или фразы в тексте или в наборе текстов. TF-IDF вычисляется как произведение двух компонент: TF и IDF. TF — это частота слова или фразы в тексте или в корпусе. IDF — это логарифм от обратной частоты документов с данным словом или фразой.

При вычислении TF-IDF максимальная частота слова или фразы была установлена равной 0.8, был использован учёт униграмм и биграмм. Реализация TF-IDF была взята из библиотеки scikit-learn [9].

4.4. GPT-3

GPT-3 (generative pre-trained transformer 3) [14] — это одна из наиболее детальных современных моделей компьютерного представления языка, построенная на архитектуре «трансформер».

Для перевода текста в вектора была использована предобученная компанией «Сбер» модель *rugpt3small_based_on_gpt2*. Модель обучена на последовательностях из 1024 токенов тремя эпохами. После этого к модели была применена тонкая настройка с последовательностями из 2048 токенов. Для взаимодействия с моделью использовалась библиотека transformers [15].

Алгоритм решения задачи с GPT-3 выглядит следующим образом:

1. Входной текст преобразовывается в список идентификаторов.
2. По списку идентификаторов вычисляется матрица размером 114×768.
3. Вычисляется среднее значение строк матрицы. Таким образом, получается вектор, описывающий текст.
4. Полученные вектора используются в обучении классификатора и собственно в классификации текстов.

5. Эксперименты

Для проведения экспериментов авторами работы был построен собственный корпус. В качестве источников текстов были выбраны новостные ресурсы: 76.ru, 1yar.tv, cnews.ru, yarnews.net, kommersant.ru, yarregion.ru, yarnovosti.com. В общей сложности корпус состоит из 1 119 текстов. Распределение по категориям выглядит следующим образом: «общество» — 447, «образование» — 307, «наука и технологии» — 108, «нерелевантная» — 257. Каждая статья содержит как минимум одно упоминание ЯрГУ. Разметка корпуса производилась вручную по подготовленным экспертом критериям, приведённым в разделе 2 данной статьи.

Для обучения классификаторов использовалось 80 % предложений корпуса, причём по 80 % от каждой категории; оставшаяся часть корпуса использовалась для тестирования.

Эксперименты проводились в три этапа. На первом этапе экспериментов выяснялось какой фрагмент текста лучше подходит для того, чтобы уложиться в ограничение BERT в 512 слов. На втором этапе в обучении и классификации использовались только полные тексты статей. На третьем этапе в обучении и классификации использовались только абзацы с упоминанием ЯрГУ.

Table 1. Choosing a fragment of articles for BERT **Таблица 1.** Выбор фрагмента статей для BERT

Текст	Фрагмент	*P*-micro	*P*-macro	*R*-micro	*R*-macro	*F*-micro	*F*-macro
абзацы с упоминанием ЯрГУ с заголовками	первые 512 слов	0.7902	0.7773	0.7902	0.7590	0.7902	0.7660
абзацы с упоминанием ЯрГУ с заголовками	первые 128 слов и последние 384	0.8304	0.8264	0.8304	0.8133	0.8304	0.8181
абзацы с упоминанием ЯрГУ с заголовками	первые 256 слов и последние 256	0.8170	0.8021	0.8170	0.8106	0.8170	0.8054
полные тексты статей с заголовками	первые 512 слов	0.7679	0.7677	0.7679	0.7620	0.7679	0.7604
полные тексты статей с заголовками	первые 128 слов и последние 384	0.7768	0.7801	0.7768	0.7865	0.7768	0.7826
полные тексты статей с заголовками	первые 256 слов и последние 256	0.7634	0.7557	0.7634	0.7764	0.7634	0.7625

Table 2. The best results by classifiers. Only full article texts are used for training

Таблица 2. Лучшие результаты по классификаторам. В обучении только полные тексты статей

Классификатор	Текст	*P*-micro	*P*-macro	*R*-micro	*R*-macro	*F*-micro	*F*-macro
SVM + TF-IDF	тексты статей с заголовками	0.8214	0.8432	0.8214	0.8209	0.8214	0.8308
Logistic Regression + TF-IDF	тексты статей без заголовков	0.8080	0.8402	0.8080	0.7912	0.8080	0.8117
Random Forest + TF-IDF	тексты статей без стоп-слов с заголовками	0.7946	0.8209	0.7946	0.7733	0.7946	0.7933
BERT	тексты статей с заголовками; из начала 128 слов и из конца 384	0.7768	0.7801	0.7768	0.7865	0.7768	0.7826
K-Neighbors + Word2vec	тексты статей без стоп-слов без деления на предложения с заголовками	0.7723	0.7836	0.7723	0.7755	0.7723	0.7793

Для оценки использовались стандартные статистические метрики: точность (*P*), полнота (*R*) и F-мера (*F*), а также их микро- и макро-версии [16]. Макро-версия метрики — это среднее арифметическое значение этой же метрики, посчитанное отдельно для каждой категории. При вычислении микро-версии метрики сначала осредняются входящие в состав метрики показатели True Positive, False Positive, False Negative, True Negative, а затем уже по ним вычисляется итоговая метрика.

Table 3. The best results by embeddings. Only full article texts are used for training

Таблица 3. Лучшие результаты по эмбеддингам. В обучении только полные тексты статей

Классификатор	Текст	*P*-micro	*P*-macro	*R*-micro	*R*-macro	*F*-micro	*F*-macro
TF-IDF + SVM	тексты статей с заголовками	0.8214	0.8432	0.8214	0.8209	0.8214	0.8308
Word2vec + SVM	тексты статей без стоп-слов с делением на предложения	0.8214	0.8082	0.8214	0.8381	0.8214	0.8205
FastText + SVM	тексты статей без деления на предложения	0.8125	0.7955	0.8125	0.8140	0.8125	0.8038
BERT	тексты статей с заголовками; из начала 128 слов и из конца 384	0.7768	0.7801	0.7768	0.7865	0.7768	0.7826
GPT-3 + SVM	тексты статей с заголовками	0.7768	0.8270	0.7768	0.7531	0.7768	0.7797

Table 4. The best results by embeddings. Only paragraphs with Demidov state university mentions are used for training

Таблица 4. Лучшие результаты по эмбеддингам. В обучении только абзацы с упоминаниями ЯрГУ

Классификатор	Текст	*P*-micro	*P*-macro	*R*-micro	*R*-macro	*F*-micro	*F*-macro
BERT	абзацы с заголовками; из начала 128 слов и из конца 384	0.8304	0.8264	0.8304	0.8133	0.8304	0.8181
TF-IDF + Logistic Regression	абзацы с заголовками без стоп-слов	0.7857	0.7762	0.7857	0.7848	0.7857	0.7803
Word2vec + SVM	абзацы с заголовками без стоп-слов без деления на предложения	0.7679	0.7659	0.7679	0.7788	0.7679	0.7718
FastText + SVM	абзацы с заголовками без стоп-слов без деления на предложения	0.7679	0.7674	0.7679	0.7681	0.7679	0.7678
GPT-3 + SVM	абзацы с заголовками	0.7500	0.7506	0.7500	0.7488	0.7500	0.7492

Table 5. The best results by classifiers. Only paragraphs with Demidov state university mentions are used for training

Таблица 5. Лучшие результаты по классификаторам. В обучении только абзацы с упоминаниями ЯрГУ

Классификатор	Текст	*P*-micro	*P*-macro	*R*-micro	*R*-macro	*F*-micro	*F*-macro
BERT	абзацы с заголовками. Из начала 128 слов и из конца 384	0.8304	0.8264	0.8304	0.8133	0.8304	0.8181
SVM + FastText	абзацы с заголовками и с делением на предложения	0.7812	0.7863	0.7812	0.7781	0.7812	0.7812
Logistic Regression + TF-IDF	абзацы с заголовками без стоп-слов	0.7857	0.7762	0.7857	0.7848	0.7857	0.7803
K-Neighbors + TF-IDF	абзацы с заголовками	0.7589	0.7835	0.7589	0.7600	0.7589	0.7700
Random Forest + FastText	абзацы с заголовками без стоп-слов и без деления на предложения	0.7545	0.8338	0.7545	0.7179	0.7545	0.7579

Из результатов первого этапа (таблица 1) видно, что какой бы текст ни использовался при обучении, наиболее информативные для BERT фрагменты находятся в первых 128 и последних 384 словах. Также видно, что результаты классификации лучше, когда обучение проводится на абзацах с упоминанием организации, а не на полных статьях.

По результатам второго этапа экспериментов (таблицы 2 и 3) видно, что лучше всего с задачей справляется SVM-классификатор с эмбеддингом TF-IDF, обученный на полных текстах статей с добавленными заголовками. Его значения микро- и макро-F-меры достигают 0.8214 и 0.8308 соответственно. Чуть хуже с задачей справился SVM-классификатор с эмбеддингом Word2Vec, обученный на полных текстах статей без стоп-слов и с делением текстов на предложения при обучении эмбеддинга. Его значения микро- и макро-F-меры достигают 0.8214 и 0.8205 соответственно.

При более подробном анализе эксперимента с лучшим результатом, в котором использовался SVM-классификатор в сочетании с эмбеддингом TF-IDF, обученный на полных текстах статей с заголовками, из таблицы 6 видно, что категория «наука и технологии» распознаётся лучше всего. Её значение F-меры достигает 0.8780. Хуже всего распознаются тексты категории «нерелевантная». Значение F-меры данной категории достигает 0.8077. Две оставшиеся категории распознаются на одном уровне: «образование» — 0.8197 и «общество» — 0.8177.

Из матрицы ошибок в таблице 7 видно, что классификатор чаще всего ошибался при определении категории «общество». Это может быть связано с тем, что данная категория очень широкая и её тексты могут иметь схожую тематику и общие ключевые слова с текстами других категорий. Меньше всего классификатор ошибался с определением текстов категории «наука и технологии». Это можно объяснить тем, что тексты данной категории похожи между собой, имеют общие ключевые слова и схожий контекст.

Table 6. The results of each category of the best result of the second stage of experiments

Таблица 6. Результаты по каждой категории лучшего результата второго этапа экспериментов

Категория	P	R	F
Нерелевантная	0.7925	0.8235	0.8077
Общество	0.8132	0.8222	0.8177
Образование	0.8197	0.8197	0.8197
Наука и технологии	0.9474	0.8182	0.8780

Table 7. The confusion matrix of the best result of the second stage of experiments

Таблица 7. Матрица ошибок для второго этапа экспериментов

Категория	Наука и технологии	Нерелевантная	Образование	Общество
Наука и технологии	18	0	3	1
Нерелевантная	0	42	1	8
Образование	1	2	50	8
Общество	0	9	7	74

По результатам третьего этапа экспериментов (таблицы 4 и 5) видно, что лучший результат достигает нейронная сеть BERT. Значения её микро- и макро-F-меры достигают 0.8304 и 0.8181 соответственно. Значения метрик достаточно высокие, поэтому можно говорить о том, что абзацев с упоминанием целевой организации достаточно для эффективной классификации по категориям.

При более детальном анализе результатов по категориям в таблице 8 видно, что хуже всего определяется категория «наука и технологии». Однако если посмотреть на матрицу ошибок в таблице 9, можно увидеть, общее число текстов в данной категории невелико, поэтому каждая ошибка сильно влияет на значения метрик. Лучше всего классификатор определяет категорию «образование» с показателем F-меры 0.8480. Больше всего классификатор ошибается с определением категорий «нерелевантная» и «общество». По-видимому, это связано с тем, что данные категории являются самыми широкими по тематике.

Table 8. The results of each category of the best result of the third stage of experiments

Таблица 8. Результаты по каждой категории лучшего результата третьего этапа экспериментов

Категория	P	R	F
Нерелевантная	0.8837	0.7451	0.8085
Общество	0.8211	0.8667	0.8432
Образование	0.8281	0.8689	0.8480
Наука и технологии	0.7727	0.7727	0.7727

Table 9. The confusion matrix of the best result of the third stage of experiments

Таблица 9. Матрица ошибок для третьего этапа экспериментов

Категория	Наука и технологии	Нерелевантная	Образование	Общество
Наука и технологии	17	0	4	1
Нерелевантная	1	38	0	12
Образование	3	1	53	4
Общество	1	4	7	78

Заключение

В работе проведены эксперименты по классификации русскоязычных новостных статей, которые посвящены Ярославскому государственному университету им. П. Г. Демидова, с использованием нейронной сети BERT и методов машинного обучения SVM, Logistic Regression, K-Neighbors, Random Forest в сочетании с эмбеддингами различных видов: Word2Vec, FastText, TF-IDF, GPT-3. Были предложены способы предобработки текстов для достижения более высокого качества классификации.

В ходе экспериментов установлено, что лучше всего с задачей справляется SVM-классификатор с эмбеддингом TF-IDF, обученный на полных текстах статей с заголовками. Его значения микро- и макро-F-меры достигают 0.8214 и 0.8308 соответственно. Сопоставимые результаты показывает нейронная сеть BERT, обученная на фрагментах абзацев с упоминанием ЯрГУ, из которых брались 128 слов из начала и 384 слов из конца. Её показатели микро- и макро-F-меры достигают 0.8304 и 0.8181 соответственно. Таким образом, абзацев с упоминанием конкретной организации оказывается достаточно, чтобы классификация по категориям была эффективной.

Хотя полученный результат уступает на 5 % в точности результатам исследования [5] о классификации русскоязычных новостных статей по категориям без привязки к предметной области, можно говорить о том, что предложенный подход эффективен, о чём свидетельствуют высокие показатели для достаточно широких категорий в числе которых «общество». Показатель F-меры данной категории достигает — 0.8177, что на 8 % выше, чем тот же показатель в аналогичной работе.

Авторы выражают благодарность Ю. А. Цофиной за обсуждение постановки задачи и помощь в разметке корпуса текстов.

References

[1] A. Hussain, G. Ali, F. Akhtar, Z. H. Khand, and A. Ali, “Design and analysis of news category predictor”, *Engineering, Technology & Applied Science Research*, vol. 10, no. 5, pp. 6380–6385, 2020.

[2] G. Kaur and K. Bajaj, “News classification using neural networks”, *Communications on applied electronics*, vol. 5, no. 1, pp. 42–45, 2016.

[3] P. Semberecki and H. Maciejewski, “Deep learning methods for subject text classification of articles”, in *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, IEEE, 2017, pp. 357–360.

[4] X. Luo, “Efficient English text classification using selected machine learning techniques”, *Alexandria Engineering Journal*, vol. 60, no. 3, pp. 3401–3409, 2021.

[5] S. Vychegzhanin, E. Kotelnikov, and V. Milov, “Comparative analysis of machine learning methods for news categorization in Russian”, in *CEUR Workshop Proceedings*, vol. 2922, 2021, pp. 100–108.

[6] N. A. Gordienko, “Klassifikaciya novostej s primeneniem metodov mashinnogo obucheniya i obrabotki estestvennogo yazyka”, in *Innovacionnye resheniya social’nyh, ekonomicheskih i tekhnologicheskih problem sovremennogo obshchestva*, in Russian, 2021, pp. 63–65.

[7] E. N. Karuna and P. V. Sokolov, “Comparison of methods for automatic classification of Russian-language texts”, in *Journal of Physics: Conference Series*, vol. 1864, 2021, p. 012 117.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, 2018. arXiv: 1810.04805 [cs.CL].

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, “Scikit-learn: Machine learning in Python”, *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, 2013. arXiv: 1301.3781v3 [cs.CL].

[11] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora", in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.

[12] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, *Fasttext.zip: Compressing text classification models*, 2016. arXiv: 1612.03651 [cs.CL].

[13] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval", *Journal of documentation*, vol. 28, no. 1, pp. 11–22, 1972.

[14] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners", *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[15] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-art natural language processing", in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[16] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks", *Information Processing & Management*, vol. 45, pp. 427–437, 2009.