DOI 10.37882/2223-2966.2024.04.08

МЕТОДИКА ОПРЕДЕЛЕНИЯ ОПТИМАЛЬНЫХ КАТЕГОРИЙ ДЛЯ КЛАССИФИКАЦИИ НОВОСТНОГО МАССИВА

METHODOLOGY FOR DETERMINING OPTIMAL CATEGORIES FOR CLASSIFYING A NEWS ARRAY

B. Goryachkin
T. Korenkova
Yu. Chernykh

Summary. One of the most important sources of analysis of social events and processes is news, since they reflect almost all their aspects and afford the opportunity to build a complete picture of social reality. To do this, it is necessary to carry out a preliminary classification of news by social topics, and the original news headings in various news resources are not well suited for this task. Therefore, in this paper it was developed and tested in practice a methodology for determining the optimal categories for classifying news texts, in particular, for the social sphere. The methodology includes the definition of new preliminary news categories by the Word2Vec algorithm, multiple thematic modeling using Zero-Shot classification and semi-automatic modification of categories until the desired thresholds of the derived metric are reached. As a result, an optimal list of categories reflecting social reality was obtained, and its advantage over the initial categories was proved.

Keywords: News topic modeling, news classification, social modeling, Word2Vec, Zero-Shot classification, NLI.

Горячкин Борис Сергеевич
кандидат технических наук, доцент,
Московский государственный технический университет им. Н.Э. Баумана
bsgor@mail.ru
Коренькова Татьяна Вячеславовна
Московский государственный технический университет им. Н.Э. Баумана
korenkova.tanya@mail.ru
Черных Юлия Сергеевна
Московский государственный технический университет им. Н.Э. Баумана
chernyh_julia@mail.ru

Аннотация. Одним из важнейших источников анализа социальных событий и процессов являются новости, поскольку они отражают практически все их аспекты и позволяют выстраивать полноценную картину социальной реальности. Для этого необходимо проводить предварительную классификацию новостей по социальным тематикам, а исходные рубрики новостей в различных новостных ресурсах недостаточно хорошо подходят для данной задачи. Поэтому в работе была разработана и проверена на практике методика определения оптимальных категорий для классификации новостных текстов, в частности, для социальной сферы. Методика включает в себя определение новых предварительных категорий новостей алгоритмом Word2Vec, многократное тематическое моделирование с помощью Zero-Shot классификации и полуавтоматическую модификацию категорий до достижения нужных порогов производной метрики. В результате был получен оптимальный список категорий, отражающих социальную реальность, а также доказано его преимущество по сравнению с исходными категориями.

Ключевые слова: тематическое моделирование новостей, классификация новостей, социальное моделирование, Word2Vec, Zero-Shot классификация, NLI.

Введение

Одним из важнейших источников изучения социальных явлений и процессов являются новости, поскольку они практически всесторонне отражают социальную реальность. Однако для проведения различного рода анализа, установления взаимосвязей между социальными явлениями и событиями и дальнейшего построения социальных прогнозов необходимо соотносить новости с классами, каждый из которых как раз и представляет себя определенную область социальной сферы.

Основная проблема в том, что готовые классы-рубрики новостей, выделенные экспертами в новостных ресурсах, плохо подходят для данной задачи, так как являются мало информативными и недостаточно отражающими социальную тематику. Поэтому возникает необходимость разработки собственной методики определения оптимальных социальных категорий и проверки качества соотнесения новостей к ним.

В настоящей статье будут рассмотрены вопросы разработки и валидации качества методики определения оптимальных социальных категорий для классификации новостей. Задачами исследования будут следующие:

— Сбор массива и категорий новостей с выбранного новостного ресурса.
— Разработка теоретической методики определения оптимальных социальных категорий.

— Формирование метрик и ограничений для определения качества методики.
— Апробация разработанной методики на практике.
— Представление итоговых результатов в виде графиков.

Классификация первичного новостного массива

За период 2020–2023 годов с российского новостного ресурса Lenta.ru была извлечена выборка из почти 400000 российских новостей и их метаданных, таких как дата публикации, название и тематика [1]. Этот источник был выбран из-за большого объема репрезентативных новостей, удобной структуры и возможности проведения web scraping. Сбор данных был осуществлен с использованием библиотек «BeautifulSoup» и «requests» на языке программирования Python [2]. Фрагмент собранного новостного массива представлен в табл. 1.

Цель следующего шага — определить первоначальный список новых категорий путем расширения исходных категорий, собранных с новостного ресурса. Необходимость определения новых категорий обосновывается тем, что исходные категории, собранные с новостного ресурса, недостаточно полно отражают суть новостей, особенно их социальную направленность. На рис. 1 показан график частотного распределения исходных новостных тематик. Можно увидеть, что некоторые тематики очень широкие, например, Россия и Мир, а некоторые формулировки слабо отражают смысловую суть представленных категорий.

Таблица 1.

Фрагмент новостного массива

Дата	Новостной заголовок	Исходная категория
2023/03/17	В США призвали Швейцарию заблокировать больше российских активов	Мир
2023/03/17	Созданы устойчивые к повреждениям крыльев роботы-шмели	Наука и техника
2023/03/17	В Венгрии обвинили Европарламент в эскалации конфликта на Украине	Мир
2023/03/17	В России раскрыли срок подготовки предложений по системе высшего образования	Россия
2023/03/17	США одобрили продажу Польше 800 ракет Hellfire	Мир

В контексте задачи классификации первичного новостного массива использована модель Word2Vec для обучения на массиве новостных заголовков с целью формирования таблицы из всех слов из данного массива и списка максимально контекстно-близких к ним. Модель Word2Vec, созданная Google, представляет собой нейронную сеть, которая обрабатывает текстовые данные [3]. Word2Vec –включает в себя две модели обучения: «Continuous Bag of Words» (CBOW) [4] и Skip-gram [5]. CBOW — «непрерывный мешок со словами», архитектура, которая предсказывает текущее слово, исходя из окружающего его контекста. Архитектура типа Skip-gram действует иначе: она использует текущее слово,

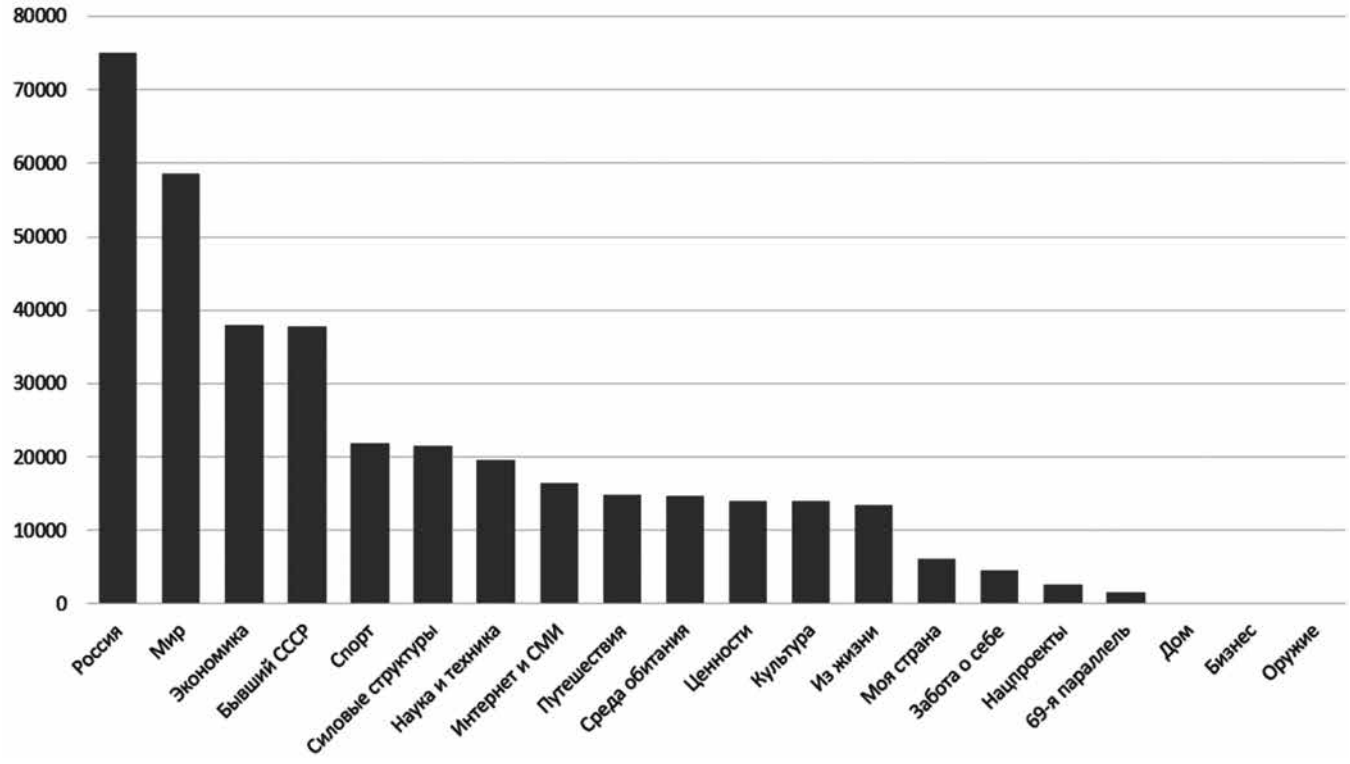

Рис 1. Частотное распределение исходных новостных тематик

чтобы предугадывать окружающие его слова. На вход в обучающую модель Word2Vec подается текстовый массив данных, а на выходе генерируются векторы слов. Кроме того, Word2Vec имеет возможность вычислять косинусное расстояние между каждым словом. Это значит, что для каждого слова из обучающей выборки можно найти список самых близких к нему слов, то есть таких, которые чаще всего упоминаются в одном контексте, на основе схожести их векторов. Модель Word2Vec обучена на языке программирования Python с использованием библиотеки «genism» [6].

Необходимость данного этапа заключается в том, что определение нового списка категорий вручную требует глубокого погружения в каждую из тематик, а также неизвестно, какие из них в принципе присутствуют в новостях. С помощью модели можно как раз получить предварительный список новых вероятных тематик, а вручную остается лишь отфильтровать нужные и подходящие для данного исследования.

После получения полной таблицы необходимо оставить только те слова, которые представляют из себя исходные категории. В списке самых контекстно-близких слов в полуавтоматическом режиме фильтруются те потенциальные слова-категории, относящиеся к социальной тематике, которые расширяют исходные. Например, одна из исходных категорий — экономика. В результате применения модели Word2Vec к слову «экономика» был сформирован список самых контекстно-близких слов, одним из которых было «безработица». Его можно добавить к списку новых потенциальных категорий, так как оно расширяет исходную категорию. В табл. 2 представлены примеры результатов работы модели Word2Vec, а, именно, список исходных тематик с самыми контекстно-близкими словами, в порядке убывания векторной близости слов. На основе анализа этих контекстно-близких слов и был составлен первоначальный список новых категорий.

Таблица 2.

Результаты работы модели Word2Vec

Исходная категория	Список самых контекстно-близких слов
Экономика	кризис, отрасль, инфляция, нефть, промышленность, энергетика, дефолт, рынок, энергокризис, бюджет, безработица, ввп, госдолг, бедность, неурожай, газ, инвестор, рецессия, экспорт, рождаемость
Политика	стратегия, разногласие, демократия, союзник, кризис, санкции, русофобия, суверенитет, импичмент, конституция, альянс, распад, реформа, заговор, дефолт, революция, война, государство, преступление
СМИ	телеканал, журналист, издание, олигарх, хакер, политик, спецслужба, дипломат, роскомнадзор, пропаганда, telegram, facebook, оппозиция, цензура, сайт, meta, минюст, иноагент, блокировка, соцсеть
Технологии	инновации, интеллект, наука, алгоритм, импортозамещение, корпорация, нейросеть, медицина, промышленность, инвестиции, туризм, космос, исследование, образование, мониторинг, экосистема, сколково, модернизация, виртуальный, микроэлектроника
Культура	национальный, народный, студенческий, музыкальный, молодёжный, патриотический, архитектура, искусство, фестиваль, литература, наследие, общество, театральный, музей, кинофестиваль, язык, возрождение, концертный, фотовыставка, цифровизация

Тематическое моделирование новостей

Тематическое моделирование — одно из современных приложений машинного обучения к анализу текстов, которое позволяет определить, к каким тематикам относится каждый документ и какие слова образуют каждую из них [7]. При этом одна из главных трудностей — формирование обучающих данных под каждую категорию, что в рамках данного исследования очень затруднительно, так как целью является многократная апробация алгоритма на разных вариантах списков новых категорий, а сделать качественную разметку под каждую из них не представляется возможным. В связи с этим был выбран подход Zero-shot classification для тематического моделирования, который позволяет обойти все эти ограничения.

Zero-shot классификация текста — задача классификации, в которой модели могут классифицировать текст, не обучаясь при этом на наборе данных, созданном для этой задачи классификации. Модель способна предсказать, к какому из предложенных классов вероятнее всего относится текст на основе анализа ключевых слов и контекста [8].

Такое возможно благодаря задаче NLI (Natural Language Inference — вывод по тексту) [9]. С помощью предобученных мультиязычных моделей, решающих задачу NLI, можно переформулировать задачу определения класса следующим образом: пусть есть множество Y, состоящее из n классов, заданных текстовой строкой, к которым может относиться текстовый документ d. Тогда с помощью модели NLI, определяющей вероятность $P(h \mid p)$ того, что пара высказывание (p) — гипотеза (h) относится к логическому следствию, для классификации достаточно найти [10]:

$$argmax_{y \in Y} P(y \mid d)$$

Тогда класс *y*, для которого будет достигаться максимум вероятности, можно считать классом, к которому относится текстовый документ.

В данном исследовании была взята предобученная мультиязычная модель для Zero-Shot классификации текстов mDeBERTa-v3-base-mnli-xnli [11], код написан на языке программирования Python с использованием библиотеки «transformers» [12].

В разработанной методике целью применения тематического моделирования является классификация новостей по исходным и новым категориям. Ставится гипотеза, что исходные категории малоинформативны для классификации новостей по социальным тематикам, но это можно доказать, только применив модель тематического моделирования для них и сравнив их с помощью метрик. Для новых категорий цель применения тематического моделирования немного отличается — необходимо многократно запустить модель на небольших выборках новостей объемом 300 наблюдений и модифицировать категории до тех пор, пока метрика качества по каждой из них не достигнет определенного значения, и можно будет считать, что был сформирован оптимальный и достаточно информативный список новых категорий для последующей классификации новостей по социальным тематикам.

Описание существующих и создание производной метрики качества тематического моделирования

В результате применения модели тематического моделирования для классификации новостей каждое наблюдение получает вероятность принадлежности к каждой из категорий — пусть эта вероятность будет называться *PB*. Тогда можно считать, что чем выше *PB*, тем больше модель «уверена» в своем выборе, и, значит, тем лучше контекст новостного заголовка и ключевые слова в нем представляют определенную категорию. В качестве примера можно взять одну и ту же новость, проклассифицированную моделью на исходном и новом списке категорий. Сравнение первых трех полученных категорий в порядке убывания значения PB представлено в табл. 3.

Выбрав самую вероятную категорию для каждой новости и соответствующую ей вероятность *PB*, можно рассчитать средний *PB* как для всей выборки, так и для каждой категории в отдельности. Логично предположить, что если средний *PB* для старых категорий при многократном запуске будет во всех случаях ниже *PB* для новых категорий, то можно считать новые категории более информативными и репрезентативными, чем старые.

Однако нужен критерий качества завершения цикла запуска алгоритма и модификации новых категорий.

Таблица 3.

Сравнение PB на исходных и новых категориях

Новость	Топ категорий		PB	
	Исходные	Новые	Исходные	Новые
ВОЗ рассказала об итогах испытаний российской вакцины от коронавируса	Россия Мир Забота о себе	Болезнь Наука Медицина	0.88 0.07 0.07	0.98 0.87 0.94
Лукашенко пообещал организатору госпереворота ответ «на всю катушку»	Силовые структуры Оружие Ценности	Политика Преступление Армия	0.69 0.54 0.48	0.95 0.73 0.54
В Германии возобновят работу угольной ТЭС из-за проблем с газом	Экономика Нацпроекты Ценности	Энергетика Кризис Промышленность	0.78 0.77 0.71	0.99 0.98 0.97

Пусть *PB* для каждой из новых категорий будет выше 0.8. Это будет обозначать, что итоговый список новых категорий достаточно оптимален в рамках данной задачи и дальнейшие модификации категорий не принесут сильного прироста в качестве.

Экспериментальные исследования вновь смоделированных категорий новостей и их оценка

В результате пятикратного запуска процесса тематического моделирования был достигнут порог метрики *PB* = 0.8 для каждой из новых категорий, после которого алгоритм был завершен. Финальный список категорий, подходящий под критерии оптимальности, оказался следующим: алкоголизм, армия, банкротство, бедность, безработица, болезнь, вакцина, война, голод, жилищно-коммунальное хозяйство, инвестиции, инновации, интернет, инфляция, информационные технологии, искусственный интеллект, искусство, катастрофа, кибератака, коронавирус, космос, кража, кризис, культура, медицина, наука, образование, оружие, политика, преступление, промышленность, развлечения, религия, рождаемость, санкции, социальная сфера, спорт, терроризм, туризм, экология, экономика, энергетика.

На рис. 2 можно увидеть график изменения среднего *PB* для новых и старых категорий на каждом из запусков. Полученные результаты доказывают, что новые категории во всех случаях репрезентативнее старых, так как для них *PB* на всех итерациях сильно выше.

На рисунках 3 и 4 представлены графики изменения *PB* для самых нерепрезентативных категорий на каждой из итераций. Можно увидеть, что разброс «неуверенности» для новых тематик относительно стабилен, и те категории, для которых *PB* низкий, подвергались дополнительному ручному анализу и модификации, то есть

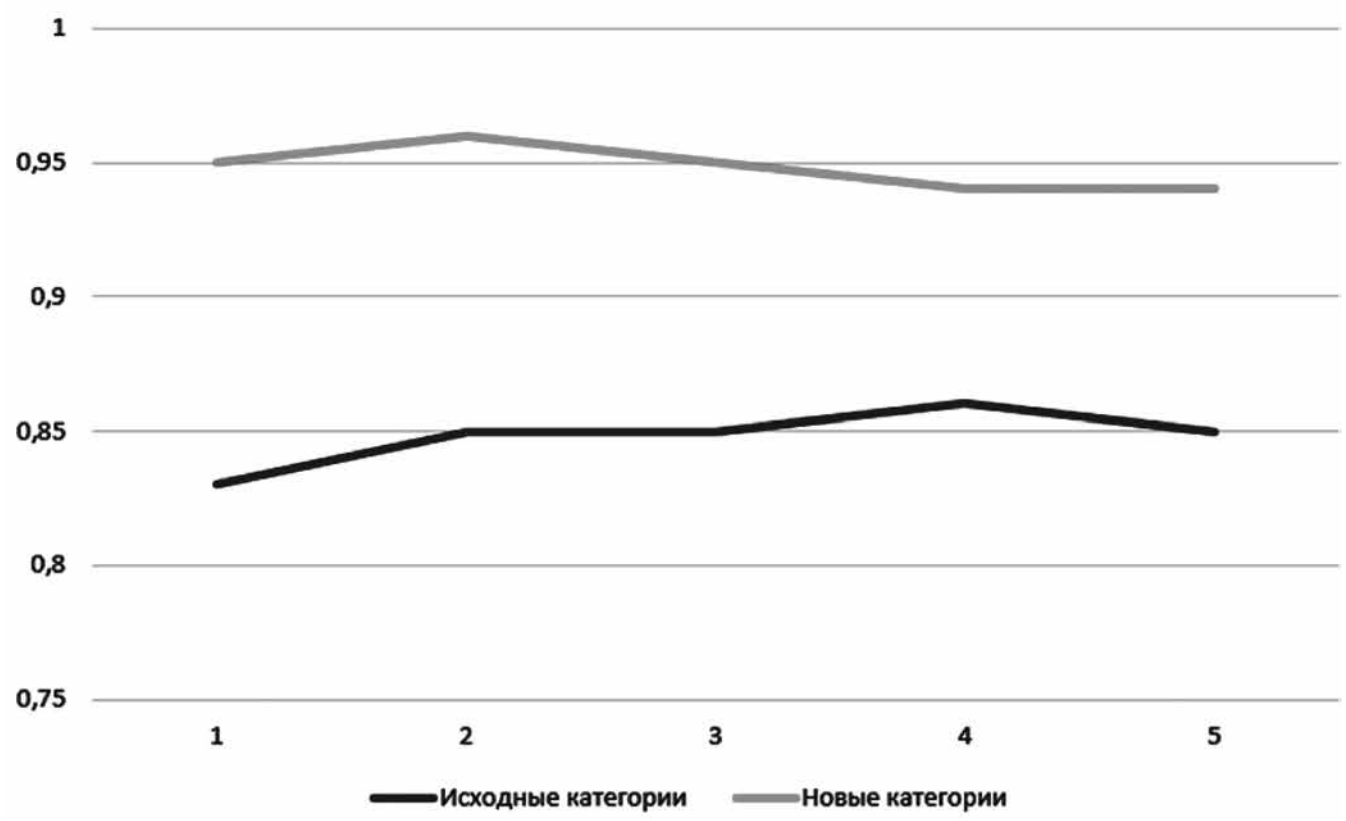

Рис. 2. График изменения среднего *PB* для новых и старых категорий по итерациям

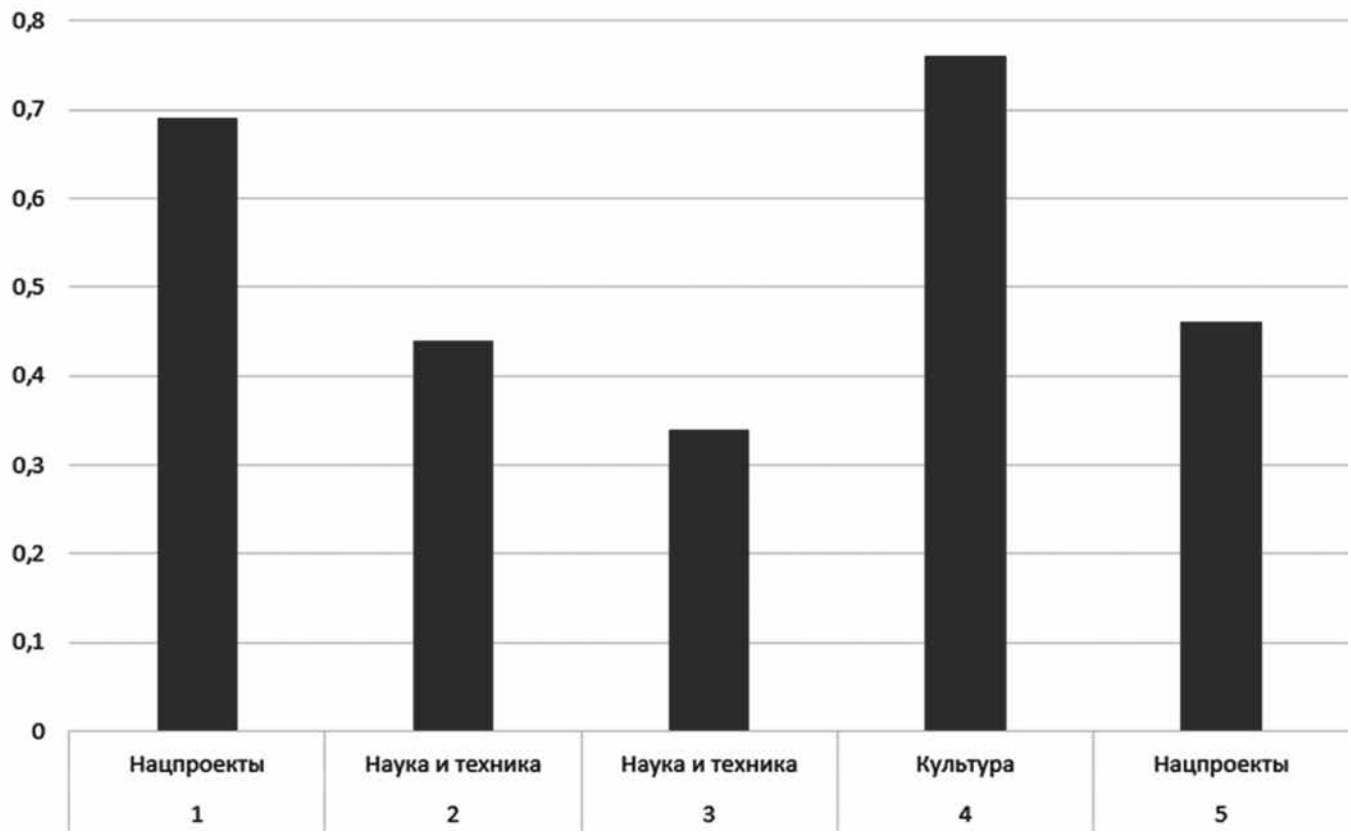

Рис. 3. График изменения *PB* для наименее репрезентативных старых категорий по итерациям

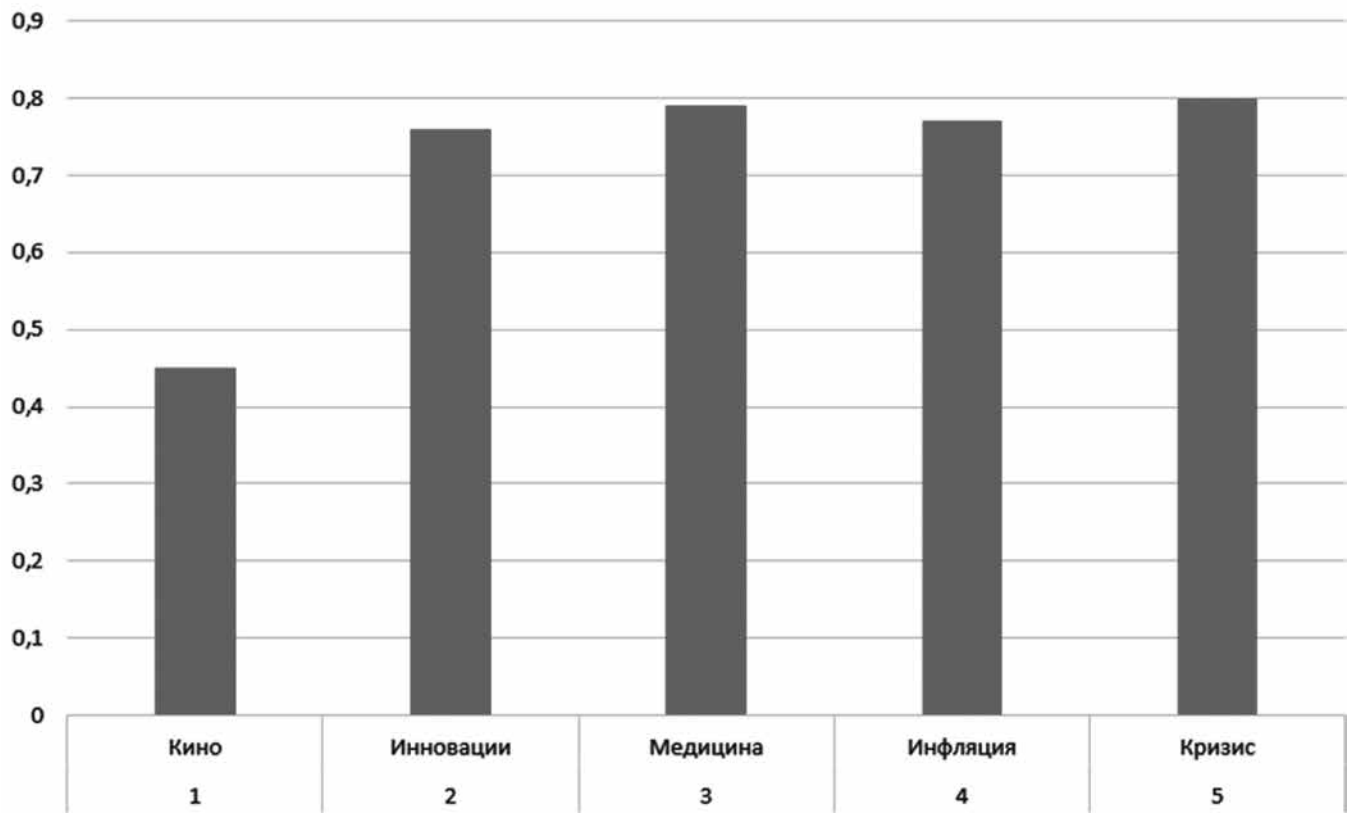

Рис. 4. График изменения *PB* для наименее репрезентативных новых категорий по итерациям

расширению, переформулированию или удалению, и в результате график *PB* стал практически стабильно возрастающим. Для старых же тематик *PB* ведет себя очень нестабильно на одних и тех же категориях.

Заключение

В результате проведенного исследования была достигнута поставленная цель — разработана и апробирована на практике методика определения оптимальных социальных категорий для классификации новостей. Он состоит из этапов сбора данных, определения первоначальных категорий с помощью модели Word2Vec, многократного запуска тематического моделирования новостей с подходом Zero-Shot classification, ручной модификации категорий на основании количественной метрики и сравнения итоговых результатов. С помощью метрики, отражающей вероятность и информативность определения новостей к категориям, было доказано, что исходные готовые категории, взятые с новостного ресурса, недостаточно репрезентативны для отражения социальной реальности через новости. В результате многократного повторения алгоритма был определен новый оптимальный список категорий, которые могут отображать те или иные социальные явления и процессы и позволять проводить дальнейший анализ и прогнозы.

Разработанную методику можно использовать для классификации новостей и изучения социальной обстановки в различных мониторинговых службах, государственных структурах, с целью установления взаимосвязей между социально-значимыми событиями и явлениями. Также его можно использовать в новостных агентствах для улучшения качества и большей автоматизации рубрикации новостей по категориям.

ЛИТЕРАТУРА

1. Lenta.ru (2023). Available at: https://lenta.ru/ (accessed 17 December 2023).
2. J.M. Patel, «Web scraping in python using beautiful soup library», Getting Structured Data from the Internet: Running Web Crawlers // Scrapers on a Big Data Production Scale, 2020. — Pp. 31–84.
3. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // Proc. Workshop at ICLR. — 2013.
4. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J., Distributed Representations of Words and Phrases and their Compositionality // Proc. NIPS. — 2013.
5. Mikolov T., Yih W., Zweig G., Linguistic Regularities in Continuous Space Word Representations // Proc. NAACL HLT. — 2013.

6. Haider M. M. et al. Automatic text summarization using gensim word2vec and k-means clustering algorithm // 2020 IEEE Region 10 Symposium (TENSYMP). — IEEE, 2020. — Pp. 283–286.
7. Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China. — 2010. — Vol. 4, no. 2. — Pp. 280–301.
8. Ji Z. et al. Zero-shot classification with unseen prototype learning //Neural computing and applications. — 2021. — Pp. 1–11.
9. Kim N. et al. Probing what different NLP tasks teach machines about function word comprehension // arXiv preprint arXiv:1904.11544. — 2019.
10. Dmitrievic B. U. Automatic creation of video presentation from text. — 2021.
11. Huggingface.co (2023). Available at: huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli (accessed 17 December 2023).
12. Tunstall L., Von Werra L., Wolf T. Natural language processing with transformers. — «O'Reilly Media, Inc.». — 2022.