Н. И. Блинов, К. О. Москвина, А. В. Глазкова
Тюменский государственный университет, г. Тюмень
УДК 004.912

РАЗРАБОТКА СЕРВИСА ДЛЯ ПОИСКА УПОМИНАНИЙ ЭКОЛОГИЧЕСКИХ ПРАКТИК В ПОСТАХ СОЦИАЛЬНЫХ СЕТЕЙ

Аннотация. Для получения информации о распространенности экологических практик в России было предложено создать web-сервис, способный проверять посты социальной сети ВКонтакте на содержание практик и предоставлять информацию об их распространенности. Был обучен multi-label классификатор, решена проблема несбалансированности имеющегося датасета, и получена модель на основе BERT с точностью от 0,7 по метрике F1-score.

Ключевые слова: NLP (обработка естественного языка), машинное обучение, анализ социальных сетей, классификация текстов, multi-label классификация, аугментация данных, token classification.

Введение. На сегодняшний день экологическая повестка составляет важную часть программы устойчивого развития в России [1]. Одной из целей экологической повестки является распространение экологических (зеленых) практик среди населения. По определению, экологические практики — это «повседневные действия, направленные на гармонизацию отношения человека и его окружающей среды».

Несмотря на постепенное развитие зеленых практик [2], сводный индекс экологичного поведения России равен 19% из 100%, что говорит о низком распространении экологических практик. Чтобы это исправить, необходимо проанализировать содержание и распространенность уже существующих практик.

Собрать информацию о распространенности тех или иных практик с помощью традиционных социологических методов не представляется возможным, но в социальных сетях в настоящее время сформирован значительный объем неструктурированной текстовой информации, связанной с экологической тематикой [3]. Более того, большую часть информации о состоянии окружающей среды опрошенные получают из социальных сетей [4], что подчеркивает

важность их роли в распространении идей экологии среди населения, а доступность и разнообразие текстовых данных, размещенных в социальных сетях, предоставляет большие возможности для изучения общественного мнения и позволяет анализировать пути распространения информации в интернет-источниках [5].

Методы машинного обучения, в частности классификация и анализ текстов социальных сетей, в последние годы часто используются при исследовании контента, наполняющего социальные сети. При этом применяются как традиционные методы машинного обучения, например, логистическая регрессия для анализа тональности постов в Twitter [7], так и более продвинутые, основанные на нейронных сетях методы — многослойный перцептрон, который часто используются в исследованиях с ограниченным набором помеченных данных [8]. Среди всех методов наиболее точными считаются методы нейронных сетей, основанные на архитектуре Transformer, в частности модель BERT и ее модификации [9].

Отдельно стоит отметить методы многоклассовой классификации текстов. В статье [10] успешно используется модифицированные методы k ближайших соседей для маркировки и классификации постов Twitter.

Помимо этого, при решении задачи классификации, поднимался вопрос об устранении несбалансированности датасета текстов, когда нет возможности просто дополнить его — задача, которая часто встречается при классификации текстов [11].

Среди методов устранения несбалансированности датасетов есть те, которые не затрагивают сам датасет — взвешивание классов, подбор порога. И те, которые изменяют его размер в большую или меньшую сторону. К ним относится ресэмплинг — undersampling (замена большого класса подвыборкой по мощности равной малому классу), oversampling (увеличение в размерах малого класса), генерация синтетических записей, схожих с реальными, аугментация. В случае с текстовым датасетом находят применение такие методы, как Back Translation [12], заключающийся в переводе текстов на иностранный язык и обратно, Easy Data Augmentation [13], состоящий из работы с синонимами, удалением, повтором и перемещением слов, и использование больших языковый моделей, например ChatGPT [14].

Проблема исследования. Главной является задача multi-label (многозначной) классификации: дано множество текстов $T = \{t1, t2, \ldots, tn\}$ и множество классов (зеленых практик) $P = \{p1, p2, \ldots, pm\}$, каждому тексту $ti \in T, 1 \leq i \leq n$ соответствует некое подмножество практик $Pi \subseteq P$. Требуется найти решающую функцию, приближающую неизвестную целевую зависимость $F : T \longrightarrow P$, значения которой известны только на обучающей выборке.

Материалы и методы. Работа ведется с постами русскоязычной социальной сети ВКонтакте. Главная их часть — текст, помимо этого имеются количество просмотров, количество репостов, количество лайков, количество комментариев. Каждый пост взят из одной из групп социальной сети. Был создан датасет размером в 1768 постов следующим образом: с помощью VKApi по запросу берутся посты из заданного списка групп, при этом экспертами отмечались фрагменты текста, которые относятся к одной из практик. В одном тексте может быть один, или несколько, или ни одного фрагмента, относящихся к практикам. Представленность практик в данном датасете показана на гистограмме (рис. 1).

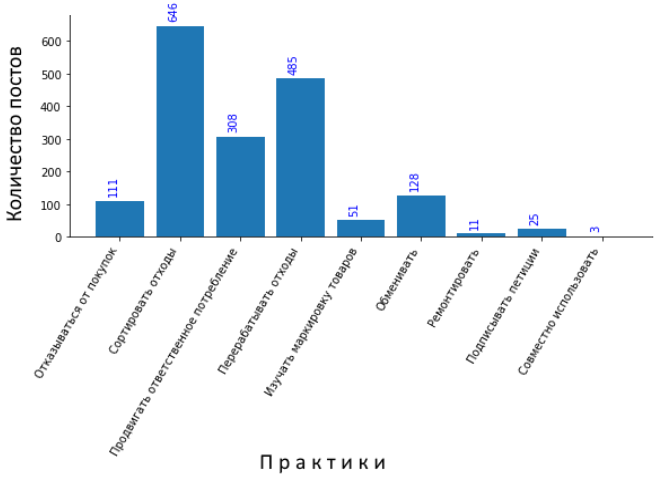

Рис. 1. Количество постов, связанных с зелеными практиками, в исходном датасете

Можно заметить, что ряд практик очень слабо представлены в датасете («Изучать маркировку товаров», «Ремонтировать», «Подписывать петиции», «Совместно использовать»).

Для дальнейшей работы тексты проходят очистку. Для ряда методов они векторизовывались методом TF-IDF.

Методы классификации

Решается данная задача разными методами. Первый — стратегия one-against-all (один-против-всех) при котором задача multi-label классификации разделяется на P задач бинарной классификации. Каждый классификатор p обучается на оригинальном датасете определению того, относится ли текст к одной конкретной практике, после чего вердикты всех классификаторов объединяются в один набор практик P_i , вменяемых тексту t_i. Преимуществами данного подхода являются вычислительная эффективность за счет относительной простоты бинарных моделей и бóльшая гибкость, выражающаяся в возможности работать отдельно с каждой практикой. К минусам же относится игнорирование корреляций между практиками, таким образом точность классификатора может снизиться из-за незамеченности возможных связей. Инструменты для создания бинарных классификаторов имеются в библиотеке scikit-learn (версия 1.2.2). В экспериментах в основном использовались: логистическая регрессия (LR), многослойный перцептрон (MLP), языковая модель rubert-tiny2.

Иной подход — сведение к задаче multi-class classification (многоклассовая классификация). Делается это за счет определения каждого набора подмножества практик P_i как отдельного класса, таким образом каждый пост относится к одному из $2^{|P|}$ классов. Преимуществом данного подхода является учет корреляций между метками. Однако при этом может проявить себя серьезная несбалансированность классов, многие из которых будут относиться к очень малому количеству экземпляров. В scikit-learn [15] инструментов для реализации такого подхода нет, они есть в расширении scikit-multilearn (использовалась версия 0.2.0), однако она для корректной работы требует scikit-learn версии 0.24.1. Использовались: модификация метода k ближайших соседей MLkNN, нейронная сеть архитектуры MLARAM.

Для проверки результатов по каждой отдельной практике (что возможно, даже если речь идет о модели, не разделяющей задачу классификации на ряд бинарных) использовалась метрика F1-score, для проверки по всем практикам в целом рассматривалась метрика F1-score macro average.

Методы аугментации

Для решения проблемы несбалансированности было решено увеличивать набор данных. Ручная разметка новых постов — эффективный способ, однако требует больших трудозатрат со стороны эксперта. Потому было решено провести эксперименты по компьютерной генерации новых текстов на основе имеющихся. Проверялись следующие методы: BackTranslation с помощью библиотеки googletrans, Easy Data Augmentation и использование языковой модели RuGPT3.

Методы выделения фрагментов

Последняя задача — выделение значимых фрагментов текста — таких, которые позволяют отнести его к конкретной практике. Для этого текст разделяется на предложения средствами библиотеки nltk [16], после чего рассматриваются два варианта: на каждом предложении по отдельности применяется готовый классификатор, относимые им к какому-либо классу считаются значимыми. Второй вариант — наоборот, классификатор применяется на тексте, из которого удаляется одно предложение. Если при удалении предложения текст перестает относится к какому-либо классу — оно признается значимым.

Результаты. Первоначально для определения возможности применения тех или иных классификаторов проводились эксперименты по обучению на изначальном датасете. 25% отводилось под тестовую выборку, в которой были представлены все классы. Данная выборка для корректности сравнения применялась для всех экспериментов. Результаты для показавших наилучший результат классификаторов представлены в табл. 1.

Для слабопредставленных практик результаты классификации по метрике F1-score ниже. Потому в дальнейшем проводились эксперименты по аугментации данных для уменьшения влияния несбалансированности датасета. Первый — с применением метода

BackTranslation. Посты переводились на иностранный язык (в большинстве случаев — английский), результат переводился обратно на русский. Для постов, связанных со слабопредставленными практиками, генерировалось несколько текстов (такой же принцип применялся и в дальнейших экспериментах по аугментации) путем перевода на несколько разных языков. Результаты представлены в табл. 2.

Таблица 1

Сравнение классификаторов на исходном датасете

	LR	*MLP*	*MLkNN*	*MLARAM*	*BERT*
Сортировка отходов	0,78	**0,8**	0,72	0,43	0,77
Изучение маркировки	0,49	0,55	0,52	0,49	**0,65**
Переработка отходов	0,73	**0,77**	0,65	0,68	0,7
Подписание петиций	0,5	**0,64**	0,6	**0,64**	0,53
Отказ от покупок	0,59	**0,62**	0,58	0,51	0,55
Обмен	0,64	**0,8**	0,7	0,69	0,5
Совместное использование	0,5	0,5	0,5	0,5	0,5
Продвижение отв. потребления	0,51	**0,71**	0,64	0,62	0,68
Ремонт	0,5	0,5	0,5	0,5	0,5
F1-score macro average	0,582	**0,654**	0,601	0,562	0,598

Таблица 2

Сравнение классификаторов на датасете, дополненном методом BackTranslation

	MLP	*MLkNN*	*BERT*	*Количество постов*
1	*2*	*3*	*4*	*5*
Сортировка отходов	**0,79**	0,7	**0,79**	906
Изучение маркировки	**0,68**	0,51	0,66	107
Переработка отходов	**0,79**	0,63	0,74	677
Подписание петиций	**0,7**	0,6	0,54	58
Отказ от покупок	**0,67**	0,59	0,55	224

Окончание табл. 2

1	*2*	*3*	*4*	*5*
Обмен	**0,83**	0,71	0,57	188
Совместное использование	0,5	0,5	0,5	18
Продвижение отв. потребления	**0,69**	0,66	0,65	617
Ремонт	0,5	0,5	0,5	44
F1-score macro average	**0,683**	0,6	0,611	

Как видно, результаты по F1-score macro average метрике могут увеличиться на 0,01-0,03. Дальше пробовался метод Easy Data Augmentation, при котором настраиваемое количество слов заменялось синонимами, удалялось, дублировалось или перемещались. Наилучшие результаты были получены при замене в 30% случаев, когда это возможно, остальные изменения в 10%. Однако и при таких параметрах результат выходит таким, что для каждого протестированного классификатора F1-score MA метрика меньше, чем на исходном датасете (табл. 3).

Таблица 3

Сравнение классификаторов на датасете, д ополненном методом Easy Data Augmentation

	MLP	*MLkNN*	*BERT*	*Количество постов*
Сортировка отходов	**0,73**	0,65	0,7	1272
Изучение маркировки	**0,68**	0,51	0,61	140
Переработка отходов	0,66	0,61	**0,68**	680
Подписание петиций	0,5	**0,56**	0,53	69
Отказ от покупок	0,48	**0,58**	0,52	279
Обмен	**0,72**	0,67	0,5	330
Совместное использование	0,5	0,5	0,5	23
Продвижение отв. потребления	0,56	0,62	**0,67**	774
Ремонт	0,5	0,5	0,5	47
F1-score macro average	**0,592**	0,577	0,578	

Последний рассматриваемый способ аугментации искусственными данными — генерация новых текстов посредством RuGPT3. Предварительно обученная имитировать тексты постов модель дополняла их в конце либо непосредственно после части, отмеченной как связанная с практикой. Данный способ показал наилучший результат в сочетании с языковой моделью BERT, все результаты представлены в табл. 4.

Таблица 4

Сравнение классификаторов на датасете, дополненном посредством RuGPT3

	MLP	*MLkNN*	*BERT*	*Количество постов*
Сортировка отходов	0,81	0,73	**0,82**	769
Изучение маркировки	0,69	0,58	**0,77**	102
Переработка отходов	0,77	0,64	**0,82**	523
Подписание петиций	0,62	0,6	**0,7**	64
Отказ от покупок	0,55	0,6	**0,79**	198
Обмен	0,81	0,72	**0,89**	190
Совместное использование	0,5	0,5	**1**	22
Продвижение отв. потребления	0,66	0,62	**0,77**	395
Ремонт	**0,75**	0,5	**0,75**	50
F1-score macro average	0,684	0,61	**0,812**	

Выделение значимых фрагментов производилось с помощью деления текстов на предложения средствами библиотеки nltk. Деление зачастую происходило с ошибками, в частности это связано с частыми нарушениями пунктуации в постах. Частично проблема решалась обработкой наиболее частых ошибок (удаление слишком коротких фрагментов, объединение фрагментов, один из которых начинается со строчной буквы). Для определения классов использовался обученный на полных текстах классификатор на основе BERT. В табл. 5 представлены результаты с обработкой ошибок и без нее, с рассмотрением отдельного предложения, либо текста с удаленным предложением.

Таблица 5

Сравнение способов определения значимых фрагментов текста с помощью BERT

	Удаление предложения, без обр. ошибок	*Удаление предложения, с обр. ошибок*	*Одно предложение, без обр. ошибок*	*Одно предложение, с обр. ошибок*
Сортировка отходов	0,56	0,56	0,63	**0,65**
Изучение маркировки	**0,72**	**0,72**	0,69	0,71
Переработка отходов	0,6	0,61	0,65	**0,66**
Подписание петиций	0,61	0,65	0,79	**0,82**
Отказ от покупок	0,56	0,56	**0,65**	**0,65**
Обмен	0,58	0,58	0,73	**0,74**
Совместное использование	**0,67**	**0,67**	0,62	0,66
Продвижение отв. потребления	0,56	0,56	0,67	**0,68**
Ремонт	**0,71**	**0,71**	0,61	0,65
F1-score macro average	0,618	0,625	0,672	**0,689**

Качество классификации для большей части практик выше для варианта с рассмотрением одного предложения отдельно, однако в некоторых случаях классификация текста с вырезанным предложением дает лучший результат.

Заключение. Как видно из табл. 1-3, при ограниченном наборе размеченных данных, MLP показывает лучшую точность, чем другие классификаторы. Однако при дополнении данных с помощью RuGPT3, хотя точность возрастает у всех методов, именно модель BERT дает лучшую точность.

Были выполнены задачи по обучению классификатора, были изучены различные способы решения проблемы несбалансированности. В итоге была получена модель классификатора, определяющая практики в постах с точностью в 0,812 по метрике F1-score macro average.

Также был разработан инструмент по выявлению связанных с практиками фрагментов текстов. Планируется его дальнейшее улучшение с целью увеличения точности работы. Планируется реализация пользовательского интерфейса на web-сервисе.

СПИСОК ЛИТЕРАТУРЫ

1. Национальные проекты России : сайт. — URL: https://xn--80aapampemcchfmo7a3c9ehj.xn--p1ai/ (дата обращения: 11.04.2023).
2. Шабанова М. А. Раздельный сбор бытовых отходов в России: уровень, факторы и потенциал включения населения / М. А. Шабанова // Мир России. Социология. Этнология. — 2019. — Т. 28, № 3. — С. 88-112.
3. Glazkova A. V. Detecting Mentions of Green Practices in Social Media Based on Text Classification / A. V. Glazkova [et al.] // Modeling and Analysis of Information Systems. — 2022. — Т. 29, № 4. — С. 316-332.
4. Экологическая повестка: за десять месяцев до выборов в Госдуму // Вицом новости : сайт. — URL: https://wciom.ru/analytical-reports/analiticheskii-doklad/ehkologicheskaja-povestka-za-desjat-mesjacev-do-vyborov-v-gosdumu (дата обращения: 11.04.2023).
5. Li Q. et al. A survey on text classification: From traditional to deep learning / Q. Li [et al.] // ACM Transactions on Intelligent Systems and Technology (TIST). — 2022. — Т. 13, № 2. — С. 1-41.
6. Мамонова Н. В. Классификация постов в англоязычной социальной сети Инстаграм (лингвосинергетический аспект) / Н. В. Мамонова // Вестник Челябинского государственного университета. — 2019. — № 4 (426). — С. 137-143.
7. Permana F. C. Naive Bayes as opinion classifier to evaluate students satisfaction based on student sentiment in Twitter Social Media / F. C. Permana, Y. Rosmansyah, A. S. Abdullah // Journal of Physics: Conference Series. — IOP Publishing, 2017. — Т. 893, № 1. — С. 012051.
8. Yantseva V. Stance Classification of Social Media Texts for Under-Resourced Scenarios in Social Sciences / V. Yantseva, K. Kucher // Swedish Workshop on Data Science : электронный журнал.

9. Vaswani A. Attention is all you need / A. Vaswani [et al.] // Advances in neural information processing systems. — 2017. — T. 30.
10. Srivastava, S. K. Multi-label Classification of Twitter Data Using Modified ML-KNN / S. K. Srivastava, S. K. Singh // Advances in Data and Information Sciences : электронный журнал.
11. Sun A. On strategies for imbalanced text classification using SVM: A comparative study / A. Sun, E. P. Lim, Y. Liu // Decision Support Systems. — 2009. — T. 48, № 1. — C. 191-201.
12. Ciolino M. Back Translation Survey for Improving Text Augmentation / M. Ciolino, D. Noever, J. Kalin // Proceedings of the Fourth Workshop on Discourse in Machine Translation. — 2019. — C. 35-44.
13. Wei J. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks / J. Wei, K. Zou // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — 2019. — C. 6382-6388.
14. Kumar V. Data Augmentation using Pre-trained Transformer Models / V. Kumar, A. Choudhary, E. Cho // Proceedings of the Second Workshop on Life-long Learning for Spoken Language Systems. — 2020. — C. 18-26.
15. Pedregosa F. Scikit-learn: Machine learning in Python / F. Pedregosa [et al.] // the Journal of machine Learning research. — 2011. — T. 12. — C. 2825-2830.
16. Bird S. NLTK: the natural language toolkit / S. Bird // Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. — 2006. — C. 69-72.