УДК 004.8
https://www.doi.org/10.47813/dnit.4.2025.3018
EDN
EQHVZT

Обработка естественного языка с использованием глубокого обучения

О.А. Митина*

МИРЭА – Российский технологический университет, Москва, Россия

*E-mail: alogmi@yandex.ru

Аннотация. Изучены модели и методы обработки естественного языка для создания алгоритма, способного с высокой точностью определять принадлежность литературного произведения к определённой группе на основе общих характеристик и элементов. Произведена оценка существующих методов обработки естественного языка, используемых в задачах классификации текстов. Спроектирована и разработана последовательность эффективных действий – алгоритм – для определения жанров книг на «домашнем» компьютерном оборудовании. Важной составляющей полученных алгоритмов и моделей является возможность выбора имеющихся вычислительных ресурсов. В современном мире экспоненциальный рост текстовых данных стал серьезной проблемой для организаций и компаний. С оцифровкой всего – от электронных писем до документов и электронной почты – постоянно увеличивается объем текстовых данных, которыми необходимо управлять, которые необходимо хранить и структурировать. Частично решить проблему можно с помощью машинной обработки естественного языка. Поскольку объем этой информации продолжает расти, возникла насущная проблема автоматизации рутинных человеческих задач, связанных с классификацией и структурированием этого огромного объема информации. Решение этой проблемы имеет решающее значение для обеспечения эффективного управления постоянно растущими объемами данных. С ростом важности обработки естественного языка в современной экономике, основанной на данных, крайне важно получить преставление о ее возможностях и ограничениях, что является одной из задач этой работы.

Ключевые слова: задача классификации, токенизация, стемминг, лемматизация, векторизация, нейронная сеть.

Natural language processing using deep learning

O.A. Mitina*

MIREA – Russian Technological University, Moscow, Russia

*E-mail: alogmi@yandex.ru

Abstract. This paper proposes an efficient and novel technique for assessment of the direction of switched capacitor bank as well as estimating its distance from the monitoring location in real distribution systems. At first, the proposed …. Models and methods of natural language processing are studied to create an algorithm capable of accurately determining the belonging of a literary work to a certain group based on common characteristics and elements. Existing methods of natural language processing used in text classification problems are assessed. A sequence of effective actions - an algorithm - for determining book genres on "home" computer equipment is designed and developed. An important component of the resulting algorithms and models is the ability to select available computing resources. In the modern world, the exponential growth of text data has become a serious problem for organizations and companies. With the digitization of everything - emails to documents and e-mail - the volume of text data that must be managed, stored and structured is constantly increasing. Partially, this problem can be solved using machine natural language processing. As the volume of this information continues to grow, an urgent problem has arisen of automating routine human tasks associated with the classification and structuring of this huge amount of information. Solving this problem is crucial to ensuring the effective management of ever-growing volumes of data. With the growing importance of natural language processing in today's data-driven economy, it is critical to gain insight into its capabilities and limitations, which is one of the objectives of this paper.

Keywords: classification problem, tokenization, stemming, lemmatization, vectorization, neural network.

1. Введение

Обработка естественного языка и искусственный интеллект в настоящее время переживают этап быстрой эволюции, открывая огромное количество возможностей для изучения моделей использования языка и расшифровки глубинных смыслов, заложенных в различных формах текста. Этот стремительный прогресс в обработке естественного языка (NLP – Natural Language Processing) и искусственного интеллекта прокладывает путь для научных прорывов в широком спектре областей, таких как здравоохранение, образование, развлечения и коммуникация. Эти достижения не только повышают эффективность и результативность этих отраслей, но и открывают новые пути для инноваций и открытий.

Эти научные прорывы в свою очередь способствуют генерированию еще большего количества информации. Сюда входит как искусственно созданный контент, например статьи и книги, написанные в соавторстве с искусственным интеллектом, так и традиционная информация, созданная человеком, включающая отчеты, статьи, обзоры, книги и другие формы письменных документов.

Результатом исследования является наиболее эффективный алгоритм определения жанров книг на основе методов обработки естественного языка, способный работать на «домашнем» компьютерном оборудовании. Это исследование потенциально может помочь книгоиздателям, библиотекарям и литературоведам более эффективно классифицировать и организовывать свои коллекции на основе жанров. Кроме того, это исследование может способствовать разработке новых методов NLP для задач классификации текстов.

2. Постановка задачи (Цель исследования)

Цель исследования – изучить существующие модели и методы обработки естественного языка для создания алгоритма, способного с высокой точностью классифицировать жанры книг, работающего на «домашнем» компьютерном оборудовании.

Исследование может помочь книгоиздателям, библиотекарям и литературоведам более эффективно классифицировать и организовывать свои коллекции на основе жанров, а также способствовать разработке новых методов NLP для задач классификации текстов.

3. Методы и материалы исследования

Обработка естественного языка – это наука о проектировании методов и алгоритмов, которые принимают и порождают неструктурированные данные естественного языка [1].

Использование методов NLP необходимо в связи с большим объемом неструктурированных данных (текст, аудио и видео), которые генерируются в современном мире и составляют около 80 % от всех. NLP помогает извлекать значимую информацию из неструктурированных текстовых данных, обеспечивая широкий спектр применения, таких как обслуживание клиентов, анализ настроений, классификация текстов, машинный перевод и распознавание голоса.

Следует отметить, что NLP – это важная область машинного обучения и сфере ИИ, которая позволяет компьютерам «понимать», обрабатывать и генерировать человеческий язык. Она необходима для извлечения основной информации из неструктурированных данных, и уже успешно решает множество задач, включая анализ настроений, машинный перевод и распознавание речи. В будущем NLP продолжит играть важнейшую роль в различных областях нашей жизни, все больше проникая в нее.

Классификация текстов по жанрам – это распространенная задача NLP, которая включает в себя присвоение жанровой метки конкретному тексту [2].

Этапы практической реализации схемы классификации текстов по жанрам показаны на рисунке 1.

Рисунок 1. Этапы практической реализации схемы классификации текстов.

Процесс классификации книг с помощью методов NLP включает в себя систематический подход, начиная со сбора и предварительной обработки данных и заканчивая обучением и тонкой настройкой моделей для точного предсказания жанра. Используя передовые методы векторизации текста и алгоритмы машинного обучения, системы классификации книг на основе NLP могут обеспечить множество практических приложений, таких как персонализированные рекомендации и эффективная каталогизация. Эти системы обладают огромным потенциалом для изменения способа

обнаружения, организации и взаимодействия с литературным контентом.

Для решения задач классификации существует множество моделей и методов машинного обучения, включая наивный Байесовский классификатор, метод опорных векторов, деревья решений, случайный лес, сверточные нейронные сети, рекуррентные нейронные сети и Long Short-Term Memory. Выбор подходящего алгоритма зависит от специфики проблемы, ее сложности и характеристик данных.

Производительность моделей машинного обучения, используемых для обработки естественного языка, должна строго оцениваться, а сами модели должны быть точно настроены для получения оптимальных результатов.

Оценка и точная настройка моделей машинного обучения в задачах классификации NLP имеет важное значение для достижения оптимальной производительности. Такие методы, как разделение обучения и тестирования, K-кратная кросс-валидация и стратифицированная K-кратная кросс-валидация, обеспечивают различные уровни надежности при оценке производительности, одновременно предотвращая перебор.

Оценивая модели с помощью соответствующих оценочных показателей, таких как accuracy, precision, recall, F1-score, ROC-AUC и PR-AUC, можно получить представление о производительности модели, что позволяет точно настроить гиперпараметры, архитектуру или процессы обучения для улучшения обобщения и достижения лучших результатов на невидимых данных [3].

Чтобы проверить работоспособность модели, оценив ее метрики, необходимо с помощью построенной модели спрогнозировать метки классов для данных, на которых модель не обучалась.

При классификации текстов для обработки естественного языка этап предсказания включает в себя присвоение меток классов невидимым текстам с помощью обученной модели машинного обучения. Он имеет ключевое значение для оценки способности модели к обобщению и применения ее к реальным задачам. Процесс включает предварительную обработку текстов, извлечение признаков, применение модели и интерпретацию результатов.

Перед подачей входного текста в обученную модель необходимо провести предварительную обработку текстовых данных таким же образом, как и обучающих

данных. Это гарантирует, что модель сможет точно обработать и классифицировать входной текст.

Этапы предварительной обработки обычно включают в себя:

- токенизация: разделение текста на отдельные слова или лексемы;
- понижение регистра: преобразование всех символов в строчные для обеспечения согласованности;
- удаление стоп-слов: устранение распространенных слов, которые не вносят существенного вклада в смысл текста;
- стемминг или лемматизация: сокращение слов до их корневой формы или инфинитивов [4-5];
- удаление специальных символов и пунктуации: удаление из текста символов, которые не могут способствовать решению задачи классификации.

В данной работе для предобработки текстов используется лемматизация.

После предварительной обработки входного текста следующим шагом является векторизация текстовых данных.

Векторизация текста – это процесс преобразования текстовых данных в числовые векторы, которые могут быть использованы алгоритмами машинного обучения. Векторизация текста – важный этап в таких задачах NLP, как анализ настроений, тематическое моделирование и классификация текстов.

Существуют различные методы векторизации текста в задачах NLP, каждый из которых обладает уникальными преимуществами и ограничениями: Bag-of-Words, TF-IDF, Word Embeddings и языковые модели [6].

Следует отметить, что целесообразно использовать Bag-of-Words, TF-IDF, Word2Vec и GloVe, поскольку они эффективны и менее требовательны к вычислениям.

После извлечения характеристик из входного текста они могут быть введены в обученную модель машинного обучения для создания прогнозов классов.

Результатом этапа прогнозирования обычно является распределение вероятностей по меткам класса, на основе которого можно получить окончательный прогноз класса.

Этап предсказания в задачах классификации текстов является важнейшим этапом. Предварительная обработка входного текста, извлечение релевантных признаков, применение обученной модели, интерпретация результатов и выполнение любой

необходимой постобработки позволяют получить точные и надежные прогнозы классов для ранее не встречавшихся экземпляров текста.

Более того, анализируя работу модели на этапе предсказания с помощью различных метрик, можно определить области улучшения и итеративно доработать модель для достижения лучших результатов в реальных приложениях.

Успех модели классификации текста в NLP зависит не только от процесса обучения, но и от ее способности обобщать и адаптироваться к новым и разнообразным данным, встречающимся на этапе предсказания.

В зависимости от специфики задачи классификации и распределения классов в наборе данных, для тщательной оценки работы классификатора можно использовать различные типы F1-коэффициентов, каждый из которых по-своему эффективен в зависимости от задачи классификации. При оценке результатов работы моделей используется средневзвешенная F1-score из-за дисбаланса классов.

В работе реализуется десятиклассовая классификация: деловая литература, детективы и триллеры, документальная литература, поэзия, любовные романы, наука, образование, приключения, проза, религия, духовность, эзотерика, фантастика.

Массив книг для обучения собран из открытых источников, имеет размер 109 Гб и содержит в себе 99283 книги с расширением «.epub» от 63 276 различных авторов (рисунок 2).

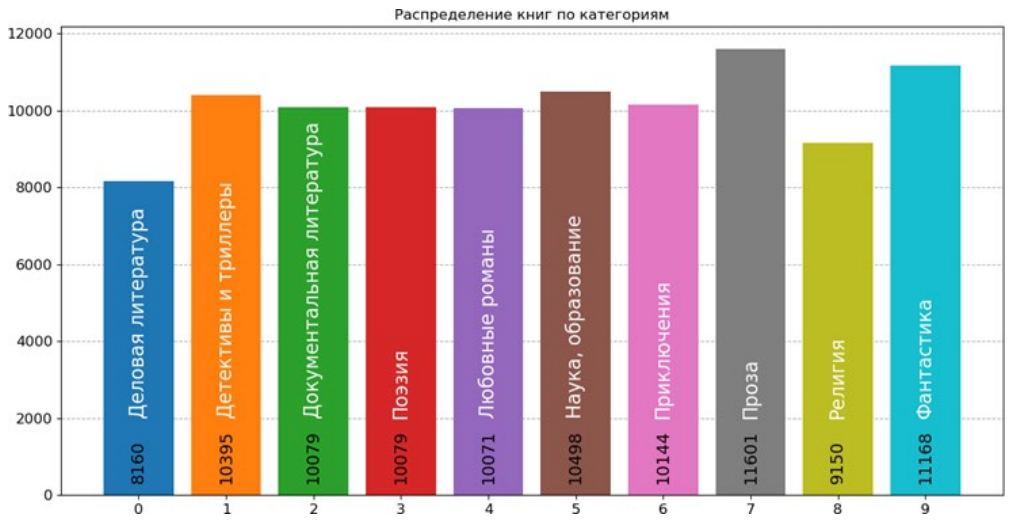

Рисунок 2. Исходное распределение книг по категориям.

В таблице 1 показаны итоговые результаты по двум лучшим парам «метод векторизации – метод обучения».

Таблица 1. Результаты машинного обучения.

Метод обучения	Векторизация	Точность, %	Полнота, %	F1-score, %
1	2	3	4	5
Градиентный бустинг	GloVe	86,21	85,01	85,37
«Комплексная» нейронная сеть	TF-IDF	87,23	87,05	87,05

В результате созданы два высокоэффективных алгоритма, предназначенных для задачи определения жанров книг. Лучшим алгоритмом для классификации текстов является комбинация TF-IDF и «комплексной» нейронной сети. С другой стороны, для ситуаций, когда вычислительные ресурсы ограничены, оптимальной парой является «GloVe – Градиентный бустинг». Эта комбинация оказывается более подходящей для ограниченных ресурсов и при этом дает достаточно хорошие результаты, обеспечивая баланс между производительностью и потреблением ресурсов.

4. Выводы

Результаты, полученные в рамках данной работы, имеют значительную практическую ценность и могут быть использованы различными заинтересованными сторонами в литературной и издательской экосистеме. Книгоиздатели, библиотекари и литературоведы, и многие другие, могут извлечь значительную пользу из этих достижений.

Доступность полученных алгоритмов, способность работать на маломощном «домашнем» компьютерном оборудовании, демократизирует доступ к передовым возможностям обработки естественного языка.

В целом, данная работа не только упрощает процесс организации и классификации книжных коллекций, но и позволяет улучшить библиотечное обслуживание, расширяет возможности издателей, тем самым способствуя эволюции литературного и издательского дел.

Список литературы

1. Гольдберг, Й. Нейросетевые методы в обработке естественного языка / Й. Гольдберг. – М.: ДМК Пресс, 2019. – 282 с.

2. Официальный сайт компании NLP Cloud: сайт. – URL: https://nlpcloud.com/ru/nlp-text-classification-api.html (дата обращения 25.02.2025).
3. Волосова, А.В. Технологии искусственного интеллекта в ULS-системах / А.В. Волосова. – СПб.: Лань, 2022. – 308 с.
4. Воронова, Л.И. Предобработка данных для нейросетевого управления / Л.И. Воронова, В. Р. Брус, В. И. Воронов, А. Н. Баширов. – М.: МТУСИ, 2021. – 49 с.
5. Рашка, С. Python и машинное обучение / С. Рашка. – М.: ДМК Пресс, 2017. – 418 с.
6. Паттерсон, Дж. Глубокое обучение с точки зрения практика / Дж. Паттерсон, А. Гибсон. – М.: ДМК Пресс, 2018. – 418 с.