Анализ алгоритмов машинного обучения, используемых для обработки текстовых документов

Н.В. Беспалова, С.А. Корчагин, Д.В. Сердечный

Финансовый университет при Правительстве Российской Федерации, Москва

Аннотация: Использование машинного обучения при работе с текстовыми документами существенно повышает эффективность работы и расширяет диапазон решаемых задач. В работе приведен анализ основных методов представления данных в цифровой формат и алгоритмов машинного обучения, сделан вывод об оптимальном решении для генеративных и дискриминативных задач.
Ключевые слова: машинное обучение, обработка естественного языка, модели архитектуры трансформер, градиентный бустинг, большие языковые модели.

Введение

Развитие и повсеместное внедрение цифровых технологий приводит к лавинообразному увеличению объемов информации, циркулирующих в современном сообществе. По оценке экспертов, в 2024 году ежедневно создавалось 328,77 млн терабайт информации, только за последние 2 года было создано более 90% всех мировых данных. По статистике за последние два года и прогнозам на 2025 год мировой объём данных можно представить следующим образом (Рисунок 1) [1]:

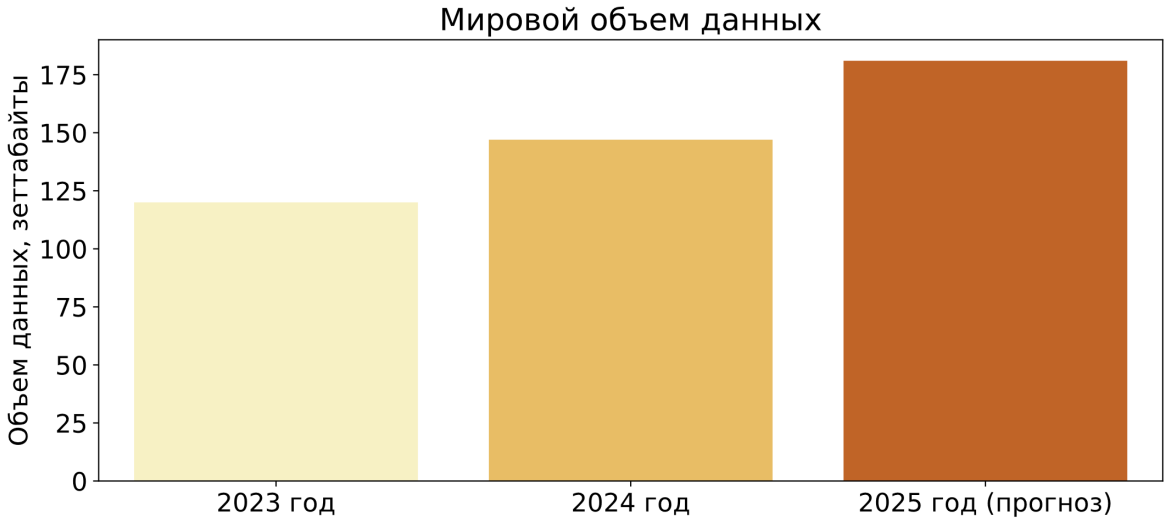

Рисунок 1. Мировой объем данных по статистике за последние два года и прогнозам на 2025 год.

Одной из актуальных задач становится обработка и классификация текстовых данных. Масштабы данных требуют использования современных инструментов обработка текстов, с целью автоматизации рутинных процессов и повышения эффективность работы. Использование машинного обучения при работе с документами приводит к созданию алгоритмов и моделей, которые способны автоматически извлекать знания из данных и решать задачи на их основе [2-4]. Подраздел информатики и искусственного интеллекта, который применяет алгоритмы машинного обучения для обработки текста и речи - Обработка естественного языка (Natural Language Processing - NLP). Область применения NLP достаточно обширна и не ограничивается распознаванием содержимого документов, также эти подходы могут использоваться для анализа документов с дальнейшим выделением областей интереса для обобщения, классификации информации, выявления спама, определение тональности текста и др. [5-6].

К преимуществам алгоритмов обработки текстов с использованием технологии NLP можно отнести:

- снижение нагрузки на человеческий ресурс;
- повышение эффективности работы с документами;
- снижение ошибок по причине человеческого фактора.

Обработка естественного языка

Этапы работы NLP можно описать в виде схемы, представленной на Рисунке 2.

Сбор данных обычно осуществляется из открытых источников, либо используется информация, собранная компанией.

Рисунок 2. Этапы работы NLP [7].

Предварительная обработка данных необходима для приведения их в понятную модели форму. Процесс проходит в несколько этапов: очистка данных (устранение дублирующей и нерелевантной информации), токенизация (разбивка текста на отдельные предложения, слова, символы, нужная степень детализации зависит от задачи и представленных данных), лемматизация (приведение слов к нормальной форме) и стемминг (нахождение основы слова), удаление стоп-слов [8].

Представление данных в цифровой формат может осуществляться различными методами, так, например, при анализе текста часто используется метод извлечения признаков из текстовых данных (Term Frequency-Inverse Document Frequency - TF-IDF), учитывающий частоту встречаемости слов как в отдельно взятом документе, так и во всем корпусе данных сразу, метод векторного представления слов (Word embedding - Word2Vec) – подход, переводящий отдельные слова в вектора, которые расположены в

пространстве согласно своей семантической близости, модели архитектуры трансформер (двунаправленный кодировщик из трансформеров (Bidirectional Encoder Representations from Transformers – BERT), оптимизированный двунаправленный кодировщик (Robustly Optimized BERT Pretraining Approach - RoBERTa), обработанная версия BERT (distilled version of BERT - DistillBERT и др.), отличительной особенностью которых является двунаправленность и возможность учитывать контекст (в отличие от Word2Vec) [9-10].

Выбор представления требует учета их специфики:

- метод TF-IDF ориентируется только на частоту встречаемости и никак не учитывает смысл слов;
- Word2Vec не учитывает контекст слова (слова-омонимы будут иметь одинаковое представление)
- Модели-трансформеры являются наиболее предпочтительными, т.к. на данный момент являются наиболее универсальными, однако требует учитывать, что представления из этих моделей обладают высокой размерностью, что понижает эффективность их использования в случаях малых объемов данных.

Наука о данных не стоит на месте и новые подходы разрабатываются и внедряются каждый день.

Алгоритмы машинного обучения для обработки текстовых документов

Выбор алгоритма машинного обучения напрямую зависит от решаемой задачи, в Таблице 1. представлены актуальные модели и их основные характеристики.

Таблица 1.

Модель	Область применения	Достоинства	Недостатки
Метод опорных векторов (Support Vector Machine - SVM)	Задачи классификации и регрессии	• обработка данных высокой размерности; • высокая точность работы; • единственность решения.	• сложность в работе с большими объемами данных; • понижение эффективность в условиях высокого шума в данных.
Наивный байесовский классификатор (Naive Bayes classifier)	Задачи классификации, выявление спама и анализ тональности данных	• высокая скорость работы; • устойчивость к шуму и выбросам.	• существенные искажения результатов при работе с несбалансированными данными; • ограничение предположением о независимости признаков, что крайне редко для реальных данных.
Случайный лес (Random Forest)	Задачи классификации и регрессии	• высокая скорость работы; • устойчивость к переобучению;	• ступенчатая форма предсказываемой зависимости в режиме регрессии

Модель	Область применения	Достоинства	Недостатки
		• возможность работы с пропущенными данными.	(и решающей гиперповерхности в режиме классификации).
Градиентный бустинг (Gradient Boosting)	Задачи классификации и регрессии	• высокая точность; • гибкость; • работа с большими объемами данных.	• более долгое обучение модели по сравнению со случайным лесом
Большие языковые модели (Large language models - LLM)	Обобщение и классификация текста и анализ тональности данных, генерация текстовых данных	• доступность моделей высокого уровня по интерфе́йсу программи́рования приложе́ний (Application programming interface – API); • гибкость.	• высокие вычислительные затраты при дообучении модели под свои задачи

Заключение

Анализ методов представления данных и алгоритмов машинного обучения позволяет сделать вывод, что для решения генеративных задач наиболее оптимальным является применение больших языковых моделей,

которые объединяют в себе оба этапа и не требуют предварительного выбора представления данных, для дискриминативных задач наилучшее решение – сочетание модели трансформера с градиентным бустингом.

Статья подготовлена по результатам исследований, выполненных за счет бюджетных средств по государственному заданию Финуниверситета.

Литература

1. Tadviser Большие данные (Big Data) мировой рынок. URL: tadviser.ru/index.php (дата обращения 13.02.2025)
2. Дементьев В. Е., Киреев С. Х. Выбор алгоритмов машинного обучения для классификации текстовых документов // Техника средств связи. – 2022. – №. 2 (158). – С. 22-52.
3. Барахнин В. Б. Кожемякина, О. Ю., Мухамедиев, Р. И., Борзилова, Ю. С., Якунин, К. О. Проектирование структуры программной системы обработки корпусов текстовых документов //Бизнес-информатика. – 2019. – Т. 13. – №. 4. – С. 60-72.
4. Джумабаева М. Ш., Бурнашев Р. Ф. Информационные технологии в обработке лингвистической информации //Science and Education. – 2023. – Т. 4. – №. 4. – С. 643-653.
5. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных // Litres. - 2022. - С. 400.
6. Азаренко Н. Ю., Казаков О. Д. NLP в задачах анализа научного текста // Глобальная нестабильность и цифровые технологии: реалии XXI века. – 2020. – С. 273-276.
7. Боженко В. В., Клюканов В. К. Применение алгоритмов машинного обучения в задачах классификации и кластеризации // Обработка, передача и защита информации в компьютерных системах 22. – 2022. – С. 28-33.

8. Ткаченко А. Л. Обзор методов интеллектуального анализа документов // Информационные технологии и автоматизация управления. – 2020. – С. 218-227.
9. Салып Б. Ю., Смирнов А. А. Анализ модели BERT как инструмента определения меры смысловой близости предложений естественного языка // StudNet. – 2022. – Т. 5. – №. 5. – С. 3509-3518.
10. Когай И. Е., Маслакова П. И., Попова Е. В. Особенности нейронной сети BERT и способы ее использования // Цифровизация экономики: направления, методы, инструменты. – 2023. – С. 365-367.

References

1. Tadviser Bol`shie danny`e (Big Data) mirovoj ry`nok [Big data world market]. tadviser.ru/index.php (date assessed 13.02.2025)
2. Dement`ev V. E., Kireev S. X. Texnika sredstv svyazi. 2022. №. 2 (158). pp. 22-52.
3. Baraxnin V. B., Kozhemyakina, O. Yu., Muxamediev, R. I., Borzilova, Yu. S., Yakunin, K. O. Biznes-informatika. 2019. T. 13. №. 4. pp. 60-72.
4. Dzhumabaeva M. Sh., Burnashev R. F. Science and Education. 2023. T. №. 4. pp. 643-653.
5. Flax P. Mashinnoe obuchenie. Nauka i iskusstvo postroeniya algoritmov, kotory`e izvlekayut znaniya iz danny`x [Machine learning. The science and art of building algorithms that extract knowledge from data]. Litres. 2022. p. 400.
6. Azarenko N. Yu., Kazakov O. D. Global`naya nestabil`nost` i cifrovy`e texnologii: realii XXI veka. 2020. pp. 273-276.
7. Bozhenko V. V., Klyukanov V. K. Obrabotka, peredacha i zashhita informacii v komp`yuterny`x sistemax'22. 2022. pp. 28-33.
8. Tkachenko A. L. Obzor metodov intellektual`nogo analiza dokumentov. 2020. pp. 218-227.
9. Saly`p B. Yu., Smirnov A. A. StudNet. 2022. T. 5. №. 5. pp. 3509-3518.

10. Kogaj I. E., Maslakova P. I., Popova E. V. Cifrovizaciya ekonomiki: napravleniya, metody, instrumenty. 2023. pp. 365-367.

Дата поступления: 18.02.2025

Дата публикации: 25.04.2025