Обзор информационных технологий для автоматической классификации информации в сети интернет

А. Ю. Гжималаускас
АО «НИЦ СПб ЭТУ»
alifa.stdn@gmail.com

Н. С. Кравчук
СПбГЭТУ «ЛЭТИ»
nskravchuk@stud.etu.ru

А. М. Кирсанов
Unirock Partners
andkirsanov@yandex.ru

Н. О. Шошков
СПбГЭТУ «ЛЭТИ»
noshoshkov@etu.ru

А. М. Скатков
СПбГЭТУ «ЛЭТИ»
sktkv@yandex.ru

Аннотация. **В настоящее время идёт активное развитие глобальной сети интернет. Она является хранилищем огромного объема совершенно разной информации, и продолжает пополняться. Использование ручного поиска информации уже является неэффективным и трудоемким процессом. Решением этой проблемы является использование технологий, направленных на сбор и классификацию информации автоматически. Были рассмотрены современные информационные технологии: Palantir, Perplexity, CatBoost, и представлены их преимущества и недостатки.**

Ключевые слова: методы классификации информации, Palantir, Perplexity, CatBoost

I. Введение

Глобальная сеть интернет – это большой источник разной информации. С течением времени она всё сильнее проникает во все сферы деятельности человека. На первый план выходит умение работать с ней, собирать и обрабатывать информацию. Однако из-за, буквально, огромного числа различных источников данных, ручная работа становится трудоемким процессом, которая будет и дальше усложняться. По этой причине на помощь приходят информационные технологии, направленные на автоматическую классификацию информации в сети интернет. Одними из таких средств является программное обеспечение (ПО) Palantir [1] и Perplexity AI [2]. Palantir известен своими технологиями анализа данных для различных целей, включая разведку и борьбу с преступностью. Perplexity AI, согласно данным, является информационно-справочной системой, которая использует большие языковые GPT-модели (Generative Pre-trained Transformer – авторегрессионные генеративные языковые модели на архитектуре трансформер) для точных ответов на сложные вопросы. Обе платформы представляют собой инновационные решения в области автоматической классификации информации. В их основе лежат различные алгоритмы, направленные на перевод подаваемого текста на другие языки, а также на обработку и анализ данных с использованием методов машинного обучения и искусственного интеллекта. Одним из примеров методов машинного обучения выступает библиотека CatBoost [3], разработанная российской компанией Яндекс, которая позволяет эффективно работать с категориальными данными, что делает ее особенно полезной в задачах, где присутствуют как числовые, так и категориальные признаки [4]. В настоящее время она уже имеет широкое применение в различных областях, включая финансы, медицину, маркетинг и другие. Совокупность этих и подобных технологий, поможет осуществлять быстрый сбор и обработку информации для дальнейшего её использования. Однако, на данный момент, они не лишены и недостатков, о которых будет указано далее.

II. Основная часть

A. Библиотека CatBoost

Рассмотрим сначала библиотеку CatBoost. Она предоставляется компанией Яндекс, которая предлагает методы машинного обучения деревьев решений. Эти деревья помогают решать задачи классификации, регрессии и ранжирования. На текущий момент это одни из самых распространенных алгоритмов, необходимых для работы с сетью интернет и информацией в целом. Основной особенностью данной технологии является работа с категориальными признаками – признаками с фиксированным набором уникальных значений [5]. CatBoost автоматически обрабатывает категориальные признаки, не требуя предварительного кодирования или преобразования данных. Это позволяет избежать множества проблем, связанных с обработкой, которые часто возникают в других библиотеках. Помимо быстрого обучения, кросс-валидации и других, в целом типичных для подобных библиотек, параметров и очередных преимуществ, стоит отметить визуализацию результатов, большой спектр задач и широкие настройки процесса обучения и качества модели. Однако, как и у всех библиотек, у CatBoost есть и минусы. Поскольку используются сложные методы обработки данных, стоит вопрос в вычислительных ресурсах и времени для обучения, особенно при больших наборах данных. Также как и у большинства алгоритмов машинного обучения, эффективность CatBoost сильно зависит от выбора параметров, таких как скорость обучения, глубина деревьев и т. д. [6]. Неправильные настройки могут

ухудшить конечный результат. Стоит отметить и выбор задач, в которых планируется использоваться библиотека, если задача простая или мы не используем категориальные признаки, то подобный выбор может быть излишним.

По краткому описанию библиотеки CatBoost, можно сделать два вывода. Во-первых, это далеко не идеальный алгоритм, который имеет как значительные плюсы, так и минусы. Преимуществам выступают: отличные результаты при настройке по умолчанию; самостоятельная обработка пропущенных значений; поддержка языков Python, R и интерфейса командной строки. А недостатками: преобразование категориальных функций в числовые значения, может исказить их суть и снизить точность модели. Во-вторых, использование в паре с ним противоположного алгоритма, который покроет минусы, может дать хорошую эффективность для универсальной обработки данных.

B. Perplexity AI

Следующая из технологий, Perplexity AI – разговорная поисковая система, которая отвечает на запросы, используя предикативный текст на естественном языке. Она помогает обычным пользователям упростить работу с поиском информации в интернете, обобщая результаты, чтобы дать как можно более прямой ответ на поставленный вопрос. Perplexity AI в своей основе использует различные архитектуры нейронных сетей: GPT – для генерации текста и предоставления структурированных ответов на запросы пользователя [7]; BERT (Bidirectional Encoder Representations from Transformers) – для понимания смысла запроса пользователя и обеспечения более точного ответа [8]; RAG (Retrieval Augmented Generation) – сочетая методы извлечения и генерации информации для создания более релевантных ответов [9]. По этой причине, недостатки и преимущества также будут связаны с ними. Основными преимуществами являются: простота использования, технология не требуют каких-либо настроек; предоставляет прямой доступ без сторонних инструментов; возможность задавать запросы на любом языке и получать ответ на нем же; предварительный анализ собранной информации и предоставление исчерпывающего ответа, который может включать изображение, ссылки, видео, код и т. д. С другой стороны, основным недостатком можно было бы посчитать так называемые «галлюцинации» ИИ [10–11]. Однако, по сравнению с другими нейронными сетями, Perplexity имеет их меньше. Но стоит учитывать, что поскольку информация берется из сети интернет, то она заведомо может быть недостоверной по разным причинам. Наиболее вероятно, что алгоритмы Perplexity AI не смогут определить, корректен ли полученный результат запроса и есть ли там, так называемый «информационный шум». Этот недостаток не противоречит основным задачам технологии – сбор, обработки и предоставление краткой информации по запросу.

Подводя общий итог по Perplexity AI – это новый удобный инструмент для поиска информации в сети интернет, который использует современные возможности GPT, BERT и REG. Дальнейшее развитие этого продукта может привести к его более широкому применению, а возможно и к возможности интеграции в другие продукты, для получения уже обработанных данных из интернета.

C. Palantir

Ещё одними технологиями автоматической классификации информации являются продукты компании Palantir [12]. По легендариуму Дж. Р. Р. Толкина Палантир – камень, с помощью которого можно было видеть происходящее или произошедшее в другом месте. Это хорошо характеризует ПО Palantir, примерно тем же оно и занимается, заглядывает в сеть интернет, получает разную информацию и обрабатывает её. Выделяют четыре основных пласта [13] того, что делает Palantir: интеграция данных; поиск и исследование; менеджмент знаний; совместная работа. Каждый из этих пластов проделывает большую и важную работу для получения конечного результата и возможности его использовать. Интеграция данных состоит в том, что Palantir способен обрабатывать разные источники информации – файлы, видео, изображения, GPS и т. д. Дальше идёт «Поиск и исследование», это возможность получить информации человеком обычным поиском. Данные связываются между собой, и на основании этой связи и осуществляется выдача результата по запросу. Следующий пласт – менеджмент знаний, иногда какой-то информации недостаточно для использования, нужна дополнительная: когда и кем создана, как изменялась и т. д. Последний пласт отчасти является опциональным, но поскольку анализ больших данных зачастую ведет не один человек, то совместная работа становится актуальна. Она схожа с системой Github, когда множество людей имеют доступ к одному ресурсу информации и работают с ним – изменяют, дополняют, анализируют. Это в итоге упрощает общий анализ, полученной информации, группой людей. Встает вопрос: «Какие же технологии используются для всего этого?», самые краткий ответ – все. На основании google patents у компании Palantir зарегистрировано почти 4 тысячи различных патентов [14]. Они касаются разных вещей: искусственный интеллект, генетический алгоритмы, алгоритмы машинного обучения для прогнозной аналитики, теория графов, алгоритмы обработки естественного языка для анализа текста, динамические онтологии, геотегирование данных, различная пересылка и хранение больших данных, алгоритмы сетевого анализа для понимания сложных взаимосвязей внутри данных [15]. Список достаточно большой. Но поскольку компания является закрытой, то подробностей практически нет.

Посмотрим на преимущества и недостатки использования Palantir. Очевидной является эффективность ПО, продукт спустя много лет всё ещё популярен среди госструктур США и ведущих западных стран. Это показывает, что используемые технологии актуальны и действенны. Поскольку Palantir разработан так, чтобы визуализировать информацию в более понятной форме, – удобство является одним из плюсов. Также преимуществом будет безопасность – доступ к данным имеют только те, у кого на них есть доступ. Что же до недостатков, то это закрытый исходный код – не

ясно, есть ли наличие вредоносный алгоритмов для пользователя; патентная защита на отдельные используемые компоненты – что ведет к торможению из развития и развитию монополии на рынке, поскольку они находится в руках одной компании. Проблема искажения реальности, как и в Perplexity AI, также немаловажная вещь, поскольку информацию в первую очередь обрабатывает не человек, то ошибка программы может сказаться на дальнейших действиях. Стоит упомянуть и высокую стоимость применения. Клиентам выгодно быть лояльными к Palantir и продолжать пользоваться их продукцией, чем перейти на какое-то другое решение.

III. Заключение

Представленные технологии для автоматической классификации информации: ПО Palantir, Perplexity AI и библиотека CatBoost, в сети интернет, ещё далеки от совершенства и имеют ряд значительных недостатков, однако сейчас они уже имеют качественно собирать и обрабатывать большие объемы информации. Дальнейшее развитие этих, и похожих, технологий позволит значительно улучшить скорость и точность поиска информации в интернете, особенно в условиях, когда существует множество схожих сайтов с однотипной информации. Основываясь на них, вполне возможно создать собственные отечественные аналоги. В качестве примера приведём классификатор информационных систем [16], структура которого должна состоять из несколько частей или модулей, каждый из которых будет выполнять свою задачу и применять собственные технологи для достижения своей цели по примеру. Для поиска и классификации информации предлагается использовать библиотеку CatBoost, как продукт компании Яндекс, а для анализа и визуализации, стоит рассмотреть уже различные архитектуры нейронных сетей, как например, Perplexity AI. Такая структура может быть применена при разработке различных систем классификации информации в целом.

Список литературы

[1] Palantir [электронный ресурс] // URL: https://www.palantir.com/ (дата обращения: 26.02.2024).

[2] Perplexity AI [электронный ресурс] // URL: https://www.perplexity.ai/ (дата обращения: 04.03.2024).

[3] CatBoost is a high-performance open-source library for gradient boosting on decision trees [электронный ресурс] // URL: https://catboost.ai/ (дата обращения: 02.03.2024).

[4] CatBoost [электронный ресурс] // URL: https://habr.com/ru/companies/otus/articles/778714/ (дата обращения: 12.03.2024).

[5] Категориальные признаки [электронный ресурс] // URL: https://habr.com/ru/articles/666234/ (дата обращения: 12.03.2024).

[6] 9 проблем машинного обучения [электронный ресурс] // URL: https://www.kaspersky.ru/blog/machine-learning-ten-challenges/21193/ (дата обращения: 04.03.2024).

[7] Кумратова А.М., Морозова Н.В., Василенко А.И., Когай И.Е. Анализ возможностей нейронной сети на основе языковой модели gpt-3 и способы ее применения на производстве // Вестник Адыгейского государственного университета. Серия 4: Естественно-математические и технические науки. 2023. №1 (316).

[8] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [электронный ресурс] // URL: https://arxiv.org/abs/1810.04805 (дата обращения: 09.03.2024).

[9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela University Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // College London; New York University

[10] Галлюцинации нейросетей: что это такое, почему они возникают и что с ними делать [электронный ресурс] // URL: https://habr.com/ru/companies/altcraft/articles/764464/ (дата обращения: 12.03.2024).

[11] Можно ли раз и навсегда устранить галлюцинации искусственного интеллекта? [электронный ресурс] // URL: https://ai.sber.ru/post/mojno-li-raz-i-navsegda-ustranit-gallucinacii-iskusstvennogo-intellekta (дата обращения: 10.03.2024).

[12] Лугачев Михаил Иванович, Скрипкин Кирилл Георгиевич. Информационная революция: средства анализа и прогнозирования. Инструменты прикладного анализа информационной революции и некоторые результаты их использования // Современные информационные технологии и ИТ-образование. 2017. №1.

[13] Palantir 101. Что позволено знать простым смертным о второй по крутости частной компании в Кремниевой Долине [электронный ресурс] // URL: https://habr.com/ru/articles/271883/ (дата обращения: 26.02.2024).

[14] Google patents [электронный ресурс] // URL: https://patents.google.com/?assignee=Palantir+Technologies+Inc&oq=assignee:(Palantir+Technologies+Inc)+ (дата обращения: 29.03.2024).

[15] Завьялов Иван Александрович. Зарубежный опыт использования искусственного интеллекта в раскрытии преступлений // Вестник Московского университета МВД России. 2021. №3.

[16] Гжималаускас А.Ю., Шошков Н.О. Сравнительный анализ существующих подходов к классификации информационных систем // Международная конференция «Проектирование и обеспечение качества информационных процессов и систем» (подана на публикацию).