38. ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА ДЛЯ КЛАССИФИКАЦИИ ТЕКСТОВ

Столбун Е. А., студент гр. 272301, Полоско Е. И., ассистент кафедры ЭИ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Ефремов А.А. – канд. эк. наук, доцент кафедры ЭИ

Аннотация. Данная работа посвящена анализу процесса обработки естественного языка для классификации текстов. Уделяется внимание способам применения искусственного интеллекта для решения прикладных задач и его актуальности в современном мире. В работе раскрываются процессы, позволяющие искусственному интеллекту понимать естественный язык. Также на примере популярного чат-бота ChatGPT описывается польза и вред использования инструментов искусственного интеллекта, тенденции их развития в будущем.

На сегодняшний день одной из актуальных проблем является работа с огромным количеством неструктурированных данных. Обработка большого объема информации и формирование выводов на ее основе требует продолжительного периода времени, которое люди не в состоянии уделить подобной задаче. Именно поэтому возникла необходимость в использовании искусственного интеллекта (ИИ), в частности, технологии обработки естественного языка.

Обработка естественного языка (англ. Natural Language Processing) совмещает компьютерную лингвистику, глубокое обучение и модели машинного обучения для эффективной обработки человеческой речи и принятия решений на основе ее содержания [1]. Natural Language Processing (NLP) играет важную роль в принятии бизнес-решений, повышает продуктивность работников ИТ-компаний и упрощает бизнес-процессы.

Одним из наиболее распространенных способов применения NLP является классификация текста. Технология обработки естественного языка для классификации текста подразделяется на несколько этапов [1]:

- токенизация;
- удаление стоп-слов;
- лемматизация;
- стемминг;
- маркировка частями речи.

Обработка естественного языка происходит следующим образом: во время токенизации текст разделяется на небольшие части под названием «токены», и из него удаляются знаки пунктуации. Затем из текста удаляются неактуальные для понимания его смысла слова. После определяется, к какой части речи относится каждое слово и на что оно указывает. В итоге чат-бот, обрабатывающий речь, генерирует ряд подходящих ответов на основе полученной информации.

Каждой из вышеперечисленных фаз соответствуют определенные методы. Ниже приведены основные задачи NLP [2]:

- распознавание речи;
- определение части речи;
- анализ настроений.

Чат-боты, работающие за счет технологий искусственного интеллекта, используют базы данных с необходимой информацией, глубокое обучение и NLP. Наиболее популярные алгоритмы, обеспечивающие работу чат-ботов включают в себя поиск и изучение шаблонов по большому количеству данных (англ. pattern matching), наивный байесовский алгоритм, рекуррентные нейронные сети, LSTM-сети и многие другие [3].

Как уже было отмечено ранее, NLP широко используется во многих областях. Наиболее частый способ применения обработки естественного языка – автоматизация рутинных задач. Современные чат-боты опираются на технологии NLP, следовательно, они способны распознавать запросы пользователей и формулировать подходящие ответы, генерировать тексты и многое другое. Делегируя монотонную работу чат-ботам, у людей высвобождается время и ресурсы для более креативных и сложных задач.

Кроме того, ИИ способен совершенствовать поиск информации. При поиске по документам и часто задаваемым вопросам NLP способно ускорить поиск по ключевым словам за счет снятия неоднозначности слов на основе контекста, сопоставления синонимов и учета морфологических вариантов.

Также с помощью обработки естественного языка можно осуществлять анализ комментариев и отзывов пользователей в социальных сетях. Анализ тональности комментариев позволяет определить реакцию покупателей на продукт и изменить стратегию компании в соответствии с этим.

Стоит упомянуть пользу NLP для анализа и упорядочивания больших коллекций документов. Некоторые методы обработки естественного языка, такие как кластеризация документов и тематическое моделирование, делают разнообразные материалы (например, отчеты компании и научные документы) более легкими для понимания.

Способы применения NLP не ограничиваются вышеописанными. На сегодняшний день обработка естественного языка уже используется в здравоохранении, юриспруденции, обслуживании покупателей и т.д.

Из вышесказанного становится очевидной польза технологий обработки естественного языка для различных сфер жизни. Но несмотря на множество достоинств NLP, внедрение ИИ в жизнь людей способно вызвать определенные проблемы.

Немаловажную роль в работе ChatGPT, наиболее продвинутого чат-бота на сегодняшний день, играет обработка естественного языка. По своей сути, ChatGPT является большой языковой моделью. Поэтому на его примере рассмотрим возможные негативные и позитивные последствия использования ИИ людьми, а также определим, какие изменения привносит в нашу жизнь NLP.

По причине того, что ChatGPT обрабатывает персональные данные пользователей, для того чтобы выполнять их запросы, возникает проблема утечки персональных данных. Кроме того, навыки обработки естественного языка, которыми обладает данный чат-бот, не являются совершенными. Соответственно, иногда ChatGPT может генерировать неточные или неподобающие ответы на вопросы. Компании, намеревающиеся внедрить ИИ в свою работу, должны следовать этическим принципам и осуществлять человеческий контроль над подобными технологиями во избежание вышеупомянутых проблем [3].

Тем не менее, появление ChatGPT поспособствовало развитию обработки естественного языка. К примеру, новейшая его версия под названием GPT-4 генерирует намного более естественный и читабельный текст по сравнению со всеми предыдущими NLP-моделями. Более того, данный чат-бот обучен автодополнению текста. Оно используется ChatGPT в мессенджерах и поисковых системах, чтобы коммуникация между людьми и поиск нужной информации онлайн происходили быстрее и эффективнее. Продвинутые навыки предсказывания, которыми обладает данная модель, позволяют поисковым системам улучшить пользовательский опыт, делая поисковые подсказки более точными. Также следует упомянуть, что до появления ChatGPT перевод, осуществлявшийся искусственным интеллектом, был довольно неточен. GPT-4 же предлагает своим пользователям машинный перевод более высокого качества, во многом благодаря NLP. ChatGPT не только точно переводит тексты с одного языка на другой, но и учитывает тональность переводимого текста и различные нюансы речи [3].

Проанализировав различные свойства ChatGPT, можно сформировать представление о современных трендах в сфере ИИ и в особенности NLP. Во-первых, доля использования ИИ людьми как в повседневной жизни, так и в работе возрастет в ближайшие годы, поскольку современные инструменты ИИ достаточно развиты, чтобы значительно облегчать человеческую жизнь. Во-вторых, можно ожидать, что благодаря обработке естественного языка взаимодействия между людьми и машинами станут более эффективными. Причина в том, что управлять компьютером при помощи команд на естественном языке интуитивно и просто. Кроме того, NLP коренным образом поменяет то, как мы работаем с информацией. Современные чат-боты способны быстро находить ответы на заданные им вопросы, обобщая большое количество данных. Поэтому стоит ожидать, что обработка естественного языка сделает нашу работу с информацией более продуктивной, и поможет нам экономить время.

Таким образом, обработка естественного языка — мощный инструмент для работы с информацией в современном мире. Она совмещает в себе машинное обучение, глубокое обучение и лингвистику. NLP находит применение во множестве сфер нашей жизни, делая работу специалистов более эффективной. Одно из популярнейших применений NLP — классификация текста. Инструменты классификации текста позволяют упорядочивать данные по теме, настроению, намерению и т. д. Они автоматизируют и совершенствуют трудоемкие процессы. Внедрение обработки естественного языка в жизнь людей способствует развитию всех сфер нашей жизни. Компании, которые грамотно применяют инструменты ИИ в своей работе, имеют все необходимое для того, чтобы преуспеть на рынке труда в ближайшие годы.

Список использованных источников:

1. Natural Language Processing /A. Chopra [et.al.] // International Journal of Technology Enhancements and Emerging Engineering Research, vol. 1, issue 4, 2013 – P. 1 – 3.

2. Parts of Speech Tagging: Rule-Based / Bao Pham // Harrisburg University of Science and Technology, 2020 – P. 4-5.

3. ChatGPT and Other Large Language Models Are Double-edged Swords / Y. Shen [et.al.] // New York University, Center for Data Science, 2023 – P. 5-9.