Искусственный интеллект в медиа-индустрии (на примере AI2Media)

В. П. Семенов
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
vps290446@mail.ru

В. Ю. Мяленка
ООО «НБМ-ИТ»
myalenkavadim@yandex.ru

А. И. Яковлев
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
alex.iakovlev@mail.ru

Д. Е. Мещеряков
ООО «НБМ-ИТ»
milkdimka@gmail.com

Аннотация. **Проект AI2Media заключается в разработке комплекса программного обеспечения с целью создания платформы для генерирования уникального медиа-контента и качественной бизнес-аналитики. AI2Media.com – медиа-платформа, на которой реализуются базовые потребности потенциального автора или редактора.**

Проект пишется на языке Python с использованием framework Django, а также NLTK (Natural Language ToolKit), и предполагает разработку комплекса из семи модулей. Анализ перечисленных модулей - основа формирования бизнес-плана научно-исследовательского start-up AI2Media, создаваемого для эффективной поддержки пользователя в сфере медиа-коммуникаций на базе гибридного искусственного интеллекта. Отличие предлагаемого проекта от других состоит в его системности. Проект охватывает всю цепочку создания добавленной стоимости медиа-продукта. Проект реализуется в сотрудничестве с Гуманитарным факультетом СПбГЭТУ «ЛЭТИ», включая студентов кафедры «Связи с общественностью», участвующих в анализе и разработке бизнес-процессов.

Ключевые слова: искусственный интеллект, гибридный интеллект, медиа-платформа, агрегатор, компьютерное зрение, бизнес-аналитика, контент, инфо-робот

I. Введение

В медиа-пространстве России формируется новая парадигма: на смену традиционным СМИ и журналистике как креативной профессии пришли медиа-коммуника-ции, дающие любому читателю возможность стать автором публикации, любому потребителю сформировать информационный поток. С 2015 года на медиа-рынке формируется принципиально новый этап: активное внедрение AI-технологий, искусственного интеллекта (AI) позволяет генерировать и создавать уникальный контент в разных форматах, что до сих пор считалось уделом творчества человека.

II. Цепочка создания добавленной стоимости проекта AI2Media

На практике имеют место следующие ключевые элементы цепочки создания добавленной стоимости, необходимые для эффективной монетизации проекта AI2Media.

A. Агрегатор источников информации

Первый элемент - формирование единого центра, куда должны регулярно (потоком) поступать все новости по выбранной заказчиком/пользователем тематике, его организация и управление.

Большинство корпоративных интернет-ресурсов создают собственный RSS-канал, формируют поток новостей, цель которого – продвижение бизнеса или PR, информирование пользователей о корпоративных событиях и новостях. Так, популярная в мире CMS WordPress устанавливает RSS-канал на создаваемый ею сайт по умолчанию. RSS, согласно RSS 2.0 Specification [1] – акроним фразы «Really Simple Syndication». На русский переводят и как «очень простое распространение», и как «очень простое получение», и даже как «простой доступ» к информации). RSS-канал – это как бы школьная доска, где на одной стороне – сайт или блог, на котором пишется информация, а на другой – все остальные сайты и блоги, с которых информацию можно списывать.

RSS-канал для рассылки новостей технически представляет собой интернет-сервис в XML-формате согласно протоколу TCP или IP. В тоже время RSS – инструмент, интегрирующий в одном канале всю нужную пользователю информацию. RSS её извлекает из других сайтов и блогов, проверяет на уникальность и собирает в едином удобном для пользователя формате.

Крупнейшие агрегаторы событийных баз данных на русском языке – поисковые системы Yandex, Rambler, Mail и Google, где преобладает агрегирование информации юридических лица. Агрегатор проекта AI2Media считывает любое требуемое клиентом число RSS-каналов, осуществляет мониторинг всех необходимых ему источников новостей, включая социальные сети, где преобладают физические лица. В итоге создается поток media-событий от юридических и физических лиц для их анализа и распространения в интересах пользователя.

Наиболее перспективна финансовая модель, где справочники услуг клиента объединены в одном издании с потоком медиа-событий и новостей соответствующей тематики, генерируемых проектом AI2Media.

Первый этап в цепочке создания проектом AI2Media добавленной стоимости технически прост, но принципиально важен. В интересах пректа клиента создается единая база данных RSS-источников, а также – по его желанию – социальных сетей и отдельныхе сайов.

B. Парсинг потока новостей на основе синтаксического анализа

Следующим звеном в цепочке проекта AI2Media является поиск, анализ и выборка актуальных для пользователя новостей. Это очень распространенное в сети явление называется парсинг, его популярное определение – «автоматизированный сбор информации» [2].

Парсинг в информатике и программировании – то, что в лингвистике определяется как синтаксический анализ. При чтении текста человек, сам того не подозревая, совершает синтаксический разбор или парсинг, сравнивая, лексемы (смысловые единицы языка) со словами из своего словарного запаса в соответствии с правилами формальной грамматики и синтаксиса, т. е. с образцом.

Для анализа естественного языка создается технически сложный программный комплекс, называемый «парсер», дословно анализатор, а процедура агрегирования источников и их анализа – «parsing», т. е. разбор, анализ. Парсер сравнивает шаблон на языке программирования с массивом информации, размещенным в сети, и производит выборку согласно заданному алгоритму.

Сегодня известны три взаимосвязанных между собой основных алгоритма парсинга, поэтому деление условно.

- Парсинг с использованием инструмента программирования «регулярные выражения».

Regular Expressions или RegExp – популярный уже много лет функционал по поиску текста или его элементов согласно шаблону, созданному по определенным правилам. Все основные языки программирования поддерживают этот шаблон, решая задачи по проверке, замещению или/и выборке из текстов и баз данных нужной информации, например, извлечение email-адресов, номеров телефонов, ip-адресов и т. д. Парсинг с использованием RegExp – надежный и доступный инструмент на стыке SEO и программирования.

- Парсинг с использованием инструмента «абстрактное синтаксическое дерево».

Парсинг с использованием AST (Abstract Syntax Tree) более сложен технически, но и более универсален, чем RegExp. Подавляющее большинство современных языков программирования, например, Python, JavaScript, Ruby – императивные. Поэтому типичный алгоритм программирования содержит три основных элемента синтаксического анализа: выражения, инструкции, объявления, которые в совокупности образуют его структуру, а по мере создания парсера или осуществления синтаксического анализа формируют AST как логическую схему парсинга.

- Парсинг с использованием инструмента программирования BNF.

В основе структуры естественного языка лежит грамматика или набор правил употребления отдельных слов. По аналогии с ним для описания языка из семейства контекстно-свободных грамматик Джон Бэкус создал первый метаязык или алгоритм описания синтаксиса, называемую Backus Normal Form (BNF). BNF - универсальный парсинг, где одни категории или единицы синтаксиса логически, последовательно определяют другие его единицы. Сегодня среди программистов популярна улучшенная версия Extended Backus–Naur Form.

У метаязыка BNF простой и стабильный набор правил формирования нетерминальных и терминальных символов (базовых смысловых единиц букв и слов), логически выстроенная и понятная форма записи языка, предсказуемое и адекватное время решения задачи, а также библиотека для большинства современных языков. Написать универсальный парсинг или синтаксический анализатор произвольной грамматики на метаязыке BNF намного сложнее, чем написать вручную лексический анализатор, но и выгоды его создания очевидны.

В рамках проекта AI2Media создается собственная версия алгоритма выборки новостей согласно шаблону на основе заказа клиента с выдачей в удобном для него формате. Парсер AI2Media пишется на языке Python с использованием SpeedParser (библиотеки для парсинга RSS), библиотеки Pandas для работы с datasets (таблицами структурированной информации), модуля CSV для сохранения результатов в csv-формате и модуля Re (Regular Expressions).

Созданный AI2Media алгоритм позволяет извлекать информацию из социальных сетей Twitter и Telegram. Парсинг сетей FaceBook и VK, а также текста из видео на канале YouTube – на высокой стадии реализации.

Гибридному AI на входе (по умолчанию) требуется релевантный набор ключевых слов и список RSS-ресурсов. В работе по их подбору и комплектации активно участвуют студенты гуманитарного факультета ЛЭТИ. Опыт работы с AI-функционалом позволяет им получить навыки, востребованные не только в медиа-индустрии, но также в маркетинге, бизнес-аналитике и даже в программировании.

C. Достоверность информации. Экспертиза/ выдача NFT – сертификата

Сегодня в медиа-пространстве, в социальных медиа особенно остро ощущается дефицит сервисов, которые смогли бы обеспечить автоматизированный анализ/ проверку на достоверность. В редакциях крупных изданий, в частности Deutsche Welle, применяются сервисы truly.media, weverify.eu. Но акцент пока делается на методике, которую можно назвать традиционным журналистским расследованием.Так, желательно иметь два независимых доказательства события, что напоминает показания свидетелей. Примером пока являются PolitiFact, который выставляет новости рейтинг по шестибальной шкале (от оценки «правда» до оценки «откровенная ложь») и Full Fact, использующий AI при сравнении данных в СМИ с цифрами официальной статистики

Одним из направлений развития AI2Media является не только использование AI при экспертизе достоверности, но и в перспективе выдача NFT – сертификата.

D. Компьютерное зрение в медиа-индустрии

Современные IT-технологии, в частности, мобильная связь, сделали основным форматом и источником информации видео-новости. «Computer Vision (CV) – это область искусственного интеллекта, связанная с анализом изображений и видео» [3]. CV – практическое направление AI, набор методов и алгоритмов для получения данных из разного рода изображений (видео, фото, рентгеновских снимков, рисунков и др.). CV в медиа – малая часть того, что применяется в этой области.

В медиа-индустрии два основных направления совершенствования CV. Во-первых, пользователю требуется транскрибация, т. е. преобразование/перевод информации из аудио – или видеопотока в традиционный текст, а во-вторых, наоборот, из текста как заготовки производителю контента требуется создать информативный аудио-видео поток. Если транскрибация как перевод из звука в текст – процесс, известный со времен изобретения магнитофона, то транскрибация текста на изображении как технология распознавания и перевода его в текстовой редактор – результат широкого использования AI.

В AI2Media поиск видео происходит по ключевым словам через Google API и предоставление доступа к субтитрам видео, которые затем пересылаются на платформу AI2Media для выделения смысловых единиц /предложений и расстановку знаков препинания. Алгоритм реализован на языке Python с использованием предварительно обученной модели.

Согласно прогнозу Cisco к «By 2022, online videos will make up more than 82 % of all consumer internet traffic - 15 times higher than it was in 2017» [4]. CV сегодня уже может узнавать изображение и понимать взаимосвязь между его элементами. Эта позволяет CV не только находить изображение, но и с помощью графического редактора рисовать новое согласно заданному алгоритму. CV также способно самостоятельно создавать видео, причем даже на основе текстовой ссылки на новость. Baudi запустил AI-powered video synthesis tool под названием VidPress.

Проект AI2Media рассматривает развитие CV как важнейший элемент AI2Media Value Chain. В проекте используются стандартные сервисы перевода медиа-файлов от Yandex «Перевод по фото on-line» и «Перевести все картинки». На этапе реализации сервисы «Получение текста видео-записи на YouTube» и «Нарезка по тематике заказчика видео-записей на YouTube».

E. Перевод новости и её адаптация

Для многих клиентов основным, а в некоторых случаях и единственным источником информации являются новости из-за рубежа, требующие качественного и быстрого перевода. Профессиональный перевод сегодня имеет устойчивый спрос, стоит дорого, по оценкам, «перевод одного слова человеком стоит в среднем 5–8 центов»,[5] и занимает много времени.

Имеют место два основных сегмента перевода с использованием AI: универсальный/повседневный и профессиональный. Универсальными системами перевода большинство людей пользуется практически ежедневно, такую услугу бесплатно предоставляют поисковые платформы Яндекс и Google.

Профессиональный перевод не всем доступен, да и нужен не всем. В России популярны две платформы профессионального перевода: Smartcat (США) и SDL Trados Studio Professional (Германия), где CAT tools – аббревиатура от «Computer- assisted translation tools».

Профессиональные платформы активно используют самообучение, что является признаком использования AI, путем формирования персональной базы памяти с нуля, где хранятся единицы перевода от слова до параграфа. База памяти заполняется во взаимодействии с терминологическими базами, по сути, огромными, тоже пополняемыми базами данных или словарями.

Основные направления развития AI в сфере перевода: создание приложения для CAT с глубокой специализацией, требующей профессиональной терминологии (например, фармакология, цифровые финансы и пр.) и совершенствование software с целью снизить затраты на редакцию уже переведенного текста. Во время эпидемии значительно вырос запрос на AI-технологии обработки языка общения – виртуальные ассистенты, «синтез и распознавание речи, клонирование голосов, речевая биометрия, голосовая активация и т. п.» [6].

Проект AI2Media осуществляет встроенный перевод, используя API эксклюзивной версии системы перевода Deep Learning на базе AI, изданной в 2020 году. On-line переводчик DeepL отличается высоким качеством и стоит особняком, занимая более 10 % рынка Google Translate. Клиенту предоставляется перевод письменного текста, файлов pdf, ауди-текста (новости, конференции, обучение). Предоставляется также видео-перевод, включая титры, timing и печатный текст для цитирования, перевод текста на изображениях, а также перевод телефонных звонков с транскрипцией.

Студенты ЛЭТИ со знанием языков в ходе обучения и углубленной специализации в рамках проекта AI2Media выделяют фразы, сравнивают их с текстом от DeepL и, при необходимости, заменяют на более подходящие, тем самым участвуя в формировании словаря проекта, его переводе, выборке и коррекции ключевых слов.

F. Генерация AI-новости на базе NLG

Использование AI-алгоритмов для производства новостей перешло в практическую фазу. Но, по оценкам, AI-технологии сегодня достигли такого уровня, что могут «взять на себя только около 15 % работы репортера и 9 % работы редактора» [7], т. е. совсем немного.

А потоки данных гигантские: по подсчетам IBM, уже в 2017 году каждый день человечество генерировало 2,5 квинтиллиона байт информации – 90 % всей информации в мире было создано за последние два года» [8], в связи с чем сформировалось целое перспективное AI-направление по обработке печатного, аудио и видео текста, называемое «генерация естественного языка».

NLG (natural language generation) – нейросеть, преобразующая набор данных в алгоритмы естественного языка – на порядок быстрее людей ищет новости, лучше структурирует архитектуру и редактирует содержание текста. NLG способна решать фундаментальную задачу логической экстраполяции информации: как бы «создавать» новость, хотя на самом деле это лишь её изложение в новой редакции и в удобном формате.

Основная проблема, с которой сталкивается сегодня AI при генерации контента – возможность его эффективного применения только там, где имеет место структурированный и плотно насыщенный сведениями (цифрами, фамилиями, фактами) поток данных, а тематика его пока узко ограничена, что предполагает построение шаблона. Отход от шаблона резко снижает качество генерации, а затраты на создание и поддержку алгоритма, наоборот, резко растут.

Проект AI2Media использует библиотеки SpaCy и Gensim, способные самостоятельно генерировать небольшие уникальные тексты (рецензия, аннотация). В данной работе активно участвуют студенты ЛЭТИ.

G. Дистрибьюция и публикация AI-новости

Есть две основные версии генерирования информационного потока в сети, в их основе – два различных подхода к производству контента, что непосредственно отражается на технологиях дистрибьюции, тем самым, что не тождественно, хотя, конечно, тесно связано, и на условиях его публикации.

Первая версия – у субъекта есть собственный бизнес, создаваемый ему платформой AI2Media информационный поток – средство продвижения и продаж его основного бизнеса, особая форма маркетинга и PR, порой качественная, как правило, полезная, добавляющая определенную ценность в цепочку создания стоимости, но по определению ограниченная. Информационный поток решает фундаментальную задачу - удерживает внимание потенциального покупателя. Но в этом случае достаточно создать собственную RSS-ленту (feed).

Вторая версия – когда информационный поток, создаваемый платформой AI2Media для клиента в интересах его развития, выступает альтернативой всем основным направлениям традиционного цифрового маркетинга. В этой версии субъект медиа-индустрии – уже производитель уникального информационного потока, где дистрибьюция новостей – эффективное завершение цепочки создания добавленной стоимости. Все инструменты в совокупности позволяют не только удерживать внимание клиентов, но и его монетизировать. У каждого субъекта своя индивидуальная «дорожная карта» дистрибьюции, т. е. распространения новости, это в принципе те же RSS-feeds и те же социальные сети, но с различной степенью интенсивности и в разных комбинациях.

Гибридный интеллект AI2Media способен участвовать в формировании повестки дня, подбирать наиболее trendy темы, определять режим публикации материалов (время выхода, порядок подачи), распределять контент издания по различным площадкам и каналам, то есть заниматься дистрибуцией созданных новостей.

H. Бизнес-аналитика

В последнее время активизировалось применение NLP для генерации развлекательного контента в единстве с новостным с целью расширения и удержания аудитории. В рамках этого бурно растущего направления на первый план выходит новая популярная профессия – блогер.

В качестве альтернативы индустрии развлечений проект AI2Media занят развитием бизнес-аналитики как важнейшего элемента цепочки создания добавленной стоимости. Акцент делается на анализе поведения пользователя с использованием AI. Пользователю демонстрируется именно тот информационный поток (smart-дистрибьюция), который ему интересен.

III. Заключение

Сегодня медиа-индустрия находится на той стадии развития, когда методы трансформации и генерации контента становятся промышленными, имеют высокую производительность, что позволяет надеяться на их быструю монетизацию.

Сегодня проект AI2Media находится на этапе создания MVP (Minimum Viable Product). MVP – в отличие от прототипа – процесс создания и тестирования скорее финансовой модели, монетизации продукта. Самая главная часть MVP – воронка привлечения пользователей, именно она сейчас тестируется.

Проект AI2Media не претендует на открытия в области AI и не содержит технических решений, достойных патента, его уникальное торговое предложение состоит в грамотном маркетинге и оригинальной комбинации доступных во многом даже рядовому пользователю AI-технологий, что позволяет поставить производство новостей на конвейер, внедрить промышленные методы организации труда, эффективно масштабировать и монетизировать создаваемый информационный поток.

Благодаря AI-технологиям в сети объективно происходит совмещение функций журналиста и читателя. В повседневной работе специалисту необходимо совмещать навыки уже не только журналиста, но редактора, маркетолога, включая, разумеется, PR, дизайнера и даже инженера-проектировщика. Главное отличие проекта AI2Media – в наборе дополнительных навыков, которые нужны специалисту, чтобы работать с большим объемом данных. навыков. Все эти навыки в совокупности формируют профессию будущего, которую будут получать и уже получают студенты ЛЭТИ.

Список литературы

[1] RSS 2.0 at Harvard Law // URL: https:// RSS 2.0 Specification (RSS 2.0 at Harvard Law) (дата обращения: 01.04.2022).

[2] Правда про парсинг сайтов // URL: https://habr.com/ru/post/446488 (дата обращения: 01.04.2022).

[3] Что такое компьютерное зрение // URL: https://trends.rbc.ru/trends/industry/5f1f007e9a794756fafbfa83 (дата обращения: 01.04.2022).

[4] Over 82 % of internet traffic will be online videos by 2022 // URL: https://medium.com/mavericks-thoughts/82-of-internet-traffic-will-be-video-by-2022-should-you-keep-blogging-d7dfe620882b (дата обращения: 01.04.2022).

[5] Рудак А. Потратил 1 000 000 $, чтобы сделать свой переводчик // URL: https://habr.com/ru/post/651345(дата обращения: 01.04.2022).

[6] Исследование. Рынок разговорного ИИ в России 2020-2025. // URL: https://just-ai.com/ru/blog/issledovanie-rynok-razgovornogo-ii-v-rossii-2020-2025 (дата обращения: 01.04.2022).

[7] Nicholas Diakopoulos. Artificial intelligence-enhanced journalism offers a glimpse of the future of the knowledge economy // URL: theconversation.com (дата обращения: 01.04.2022).

[8] Сальманов О. Поможет ли искусственный интеллект в борьбе за внимание аудитории // URL:https://www.vedomosti.ru/partner/articles/2018/06/27/773891-roboti-lyudei (дата обращения: 01.04.2022)