**Ершов Иван Иванович,**
студент магистратуры 2 курса гр. ИСТм-41,
ФГОБУ ВО «Поволжский государственный
университет телекоммуникаций и информатики»
Ershov Ivan Ivanovich,
2nd year master's student gr. ISTm-41,
FGOBU in «Volga State University
of Telecommunications, and Informatics»

**Захарова Оксана Игоревна,** к.т.н, доцент,
ФГОБУ ВО «Поволжский государственный
университет телекоммуникаций и информатики»
Zakharova Oksana Igorevna, k.t.s, associate,
FGOBU in «Volga State University
of Telecommunications and Informatics»

# ЭВОЛЮЦИЯ МЕТОДОВ АНАЛИЗА БОЛЬШИХ ДАННЫХ В КОНТЕКСТЕ ОБРАБОТКИ ТЕКСТОВЫХ ДАННЫХ
# THE EVOLUTION OF BIG DATA ANALYSIS METHODS IN THE CONTEXT OF TEXT DATA PROCESSING

**Аннотация.** Статья прослеживает эволюцию методов анализа текстовых данных – от статистических моделей (мешок слов, TF-IDF) и классического машинного обучения к векторным представлениям, рекуррентным сетям и архитектуре трансформеров. Рассмотрена роль распределённых платформ (Hadoop, Spark) в обработке больших данных, а также прорыв генеративных моделей (GPT). Отмечены перспективы и этические вызовы современных технологий текстовой аналитики.

**Abstract.** This article traces the evolution of text data analysis methods–from statistical models (bag-of-words, TF-IDF) and classical machine learning to vector representations, recurrent networks, and transformer architecture. The role of distributed platforms (Hadoop, Spark) in big data processing is considered, as well as the breakthrough of generative models (GPT). The prospects and ethical challenges of modern text analytics technologies are highlighted.

**Ключевые слова:** Большие данные, обработка естественного языка, трансформеры, генеративные модели, машинное обучение, текстовая аналитика.

**Keywords:** Big data, natural language processing, transformers, generative models, machine learning, text analytics.

## Введение

Современный мир порождает огромные объемы данных, значительная часть которых представлена текстовой информацией. Эта информация, содержащаяся в электронных письмах, сообщениях в социальных сетях, новостных статьях, научных публикациях, отзывах о продуктах, записях в блогах и многих других источниках, представляет собой ценный ресурс для бизнеса, науки и общества в целом. Обработка таких данных требует сложных вычислительных методов и интеллектуальных систем. Исторически подходы к анализу текстовых данных эволюционировали от примитивных статистических моделей к современным глубоким нейронным сетям и генеративным моделям. Развитие технологий больших данных (Big Data) сделало возможным обработку информации в масштабах, немыслимых ранее, и открыло новые горизонты для извлечения знаний из текстовых

массивов. Актуальность анализа текстовых данных растет с каждым годом, так как текстовые данные генерируются в социальных сетях, новостных платформах, блогах и других онлайн-ресурсах, и понимание этих данных становится ключом к пониманию трендов, настроений и потребностей пользователей. В этой статье рассматривается эволюция методов анализа текстовых данных, начиная с традиционных статистических подходов и заканчивая трансформерами и самообучающимися системами, а также освещаются будущие направления развития этой области, включая вызовы и этические аспекты.

**Ранние подходы к обработке текстовых данных**

На ранних этапах автоматизации обработки текстов основное внимание уделялось простым статистическим методам. Эти методы были обусловлены ограниченными вычислительными ресурсами и зачастую требовали ручной разметки данных. Одним из первых методов была модель «мешок слов» (bag-of-words), использовавшая слова в тексте как независимые элементы, игнорируя порядок слов и контекст [1]. Этот подход позволял строить частотные таблицы слов для задач классификации и поиска, но был ограничен из-за неспособности уловить семантические связи между словами и зависимости в синтаксической структуре предложений.

Кроме того, ранние методы включали использование частотных словарей для определения значимости отдельных терминов в документе. Для улучшения качества анализа был разработан подход TF-IDF (term frequency-inverse document frequency), который учитывал важность слов в конкретных документах относительно их встречаемости в корпусе. TF-IDF позволил не только определить частоту слова в документе, но и учесть его редкость в масштабах всего корпуса, что повышало точность ранжирования. Этот метод позволял ранжировать документы по релевантности к запросу, но не решал задачи более глубокого анализа, такие как определение контекста или анализ настроений [2].

**Развитие машинного обучения**

В 1990-х годах произошел значительный прорыв благодаря внедрению методов машинного обучения. Это был период, когда вычислительные мощности начали расти, а доступ к данным упростился. Алгоритмы наивного байеса, деревья решений и линейные классификаторы, такие как метод опорных векторов (SVM), стали основными инструментами для построения первых систем автоматической категоризации текста. Эти алгоритмы использовали вероятностные модели для предсказания классов текста и могли применяться в фильтрации спама, анализе отзывов и категоризации новостных статей. SVM, например, показал себя особенно эффективным в задачах классификации текста благодаря способности находить оптимальную гиперплоскость, разделяющую классы. Тем не менее, они имели ограниченные возможности для работы с длинными текстовыми зависимостями, а качество работы сильно зависело от качества предобработки данных и выбора признаков [3].

**Латентное размещение Дирихле (LDA) и тематическое моделирование**

Латентное размещение Дирихле (LDA) предложило новый способ анализа текстов, выделяя скрытые темы на основе совместной встречаемости слов. Тематическое моделирование открыло возможности для анализа больших массивов текстовых данных, таких как научные публикации или обзоры пользователей, позволяя выявлять скрытые закономерности и кластеризовать документы по тематикам. LDA рассматривает каждый документ как смесь тем, а каждое слово в документе как принадлежащее одной из этих тем. Однако сложность интерпретации результатов и высокая вычислительная нагрузка ограничивали его использование. Кроме того, для эффективной работы LDA требовалось предварительно задавать количество тем, что часто было неочевидной задачей [4].

**Скрытые марковские модели**

Скрытые марковские модели (HMM) предоставили инструменты для анализа последовательностей, где каждая единица связана с предыдущими. HMM предполагают, что наблюдаемые данные (например, слова в тексте) генерируются скрытыми состояниями, которые представляют собой части речи или другие лингвистические единицы. Эти модели нашли применение в задачах морфологического анализа, разметки частей речи и даже в первых версиях систем машинного перевода. Они были особенно полезны для задач, где важен порядок слов и контекст. Основным недостатком HMM было ограничение на работу с длинными зависимостями из-за экспоненциального роста вычислительных потребностей и предположения о том, что текущее состояние зависит только от предыдущего, что не всегда верно для естественного языка.

**Современные подходы к обработке текстов**

Эпоха векторных представлений слов, или Word Embeddings, ознаменовала собой революционный прорыв в области обработки естественного языка**.** До этого момента, модели, оперирующие словами как дискретными, атомарными символами, сталкивались с непреодолимыми ограничениями в понимании тонкостей семантических отношений. Появление же концепции **Word Embeddings**, позволившей отображать слова в виде многомерных векторов в семантическом пространстве, открыло совершенно новые горизонты для анализа текста [5].

Одним из ключевых преимуществ векторных представлений стало резкое повышение точности моделей классификации и других задач обработки текста. Теперь стало возможным вычислять семантическую близость между словами и целыми контекстами, используя математические операции над векторами, например, **косинусную меру**, определяющую угол между векторами. Чем меньше угол, тем семантически ближе друг к другу слова или тексты.

Показательным примером эффективности такого подхода является знаменитая аналогия: «король» минус «мужчина» плюс «женщина» дает результат, удивительно близкий к вектору «королева». Этот пример наглядно демонстрирует, как Word2Vec улавливает и кодирует в векторной форме сложные семантические отношения между словами, такие как пол и социальный статус. По сути, модель смогла "понять", что отношения между "королем" и "мужчиной" аналогичны отношениям между "королевой" и "женщиной", и выразить это понимание в математической форме.

Вдохновленные успехом Word2Vec, исследователи продолжили совершенствовать методы построения векторных представлений. В последующие годы были разработаны и другие передовые модели, такие как **GloVe (Global Vectors for Word Representation) и FastText. GloVe**, разработанная в Стэнфордском университете, пошла дальше, учитывая не только локальный контекст, в котором встречается слово, но и глобальную статистику совместной встречаемости слов во всем корпусе текстов. Это позволило еще точнее отразить семантические нюансы слов [6].

**Глубокие нейронные сети и их развитие: RNN, LSTM, GRU**

Глубокие нейронные сети (DNN) и их специализации, такие как рекуррентные нейронные сети (RNN) и их улучшенные модификации LSTM и GRU, дали новый импульс текстовой аналитике. RNN были разработаны специально для обработки последовательных данных, таких как текст, учитывая предыдущие слова при обработке текущего. Эти сети могут моделировать зависимости между словами во временном контексте, что особенно важно для задач машинного перевода и создания текстов. Однако проблемы исчезающих и взрывающихся градиентов ограничивали их производительность, особенно при работе с длинными последовательностями. Позже эти трудности были преодолены благодаря улучшенным механизмам управления градиентами и введению архитектур LSTM (Long Short-

Term Memory) и GRU (Gated Recurrent Unit), которые позволяли сети "запоминать" информацию на более длительные периоды.

**Революция трансформеров: Attention is All You Need**

Архитектура трансформеров, впервые предложенная в статье «Attention is All You Need» в 2017 году, полностью изменила парадигму обработки текстов. До этого момента доминировали RNN и их вариации, но трансформеры предложили принципиально новый подход, основанный на механизме внимания. Трансформеры используют механизмы внимания, которые позволяют моделям фокусироваться на наиболее значимых частях текста независимо от их расстояния друг от друга. Это решало проблему обработки длинных зависимостей, с которой сталкивались RNN. Модели BERT (Bidirectional Encoder Representations from Transformers) и GPT (Generative Pre-trained Transformer) стали символами этой революции. BERT впервые применил двунаправленный анализ текста, позволяя моделям учитывать контекст слева и справа от слова, что значительно повысило качество решения задач понимания естественного языка. GPT-семейство, включая GPT-4, продвинулось в создании генеративных текстов, что сделало их популярными для создания чат-ботов, систем поддержки пользователей, генерации контента и решения других задач, требующих генерации связного и осмысленного текста [7].

**Обработка больших данных и распределенные системы: Hadoop, Spark**

С ростом объемов данных ключевым фактором успеха стало использование распределенных систем обработки. Традиционные подходы, основанные на обработке данных на одном сервере, перестали справляться с петабайтами информации. Платформы, такие как Apache Hadoop и Spark, обеспечивают параллельные вычисления, позволяя анализировать массивы текстов размером в терабайты и даже петабайты. Hadoop, с его распределенной файловой системой HDFS и моделью программирования MapReduce, стал пионером в области обработки больших данных. Spark, в свою очередь, предложил более быструю и гибкую модель обработки данных в оперативной памяти. Потоковая обработка (streaming) стала важным элементом для мониторинга данных в реальном времени, например, в анализе социальных медиа или новостных потоков. В сочетании с облачными сервисами, предоставляющими вычислительные мощности по запросу, эти технологии формируют основу современных систем анализа больших данных, позволяя обрабатывать огромные массивы текстовой информации с высокой скоростью и эффективностью [8].

**Самообучающиеся и генеративные модели: Few-Shot Learning, GPT-3, GPT-4**

Генеративные модели, представляющие собой передовой край искусственного интеллекта, демонстрируют поразительные способности в области обработки естественного языка. К числу наиболее известных относятся такие модели, как GPT-3 и её преемница, еще более мощная GPT-4, разработанные компанией OpenAI. Эти модели обладают уникальной способностью генерировать тексты высочайшего качества, опираясь на минимальное количество предоставленных примеров, что известно как обучение на нескольких примерах (few-shot learning). Впечатляет и то, что они могут функционировать даже в режиме обучения без примеров (zero-shot learning), то есть генерировать тексты на темы, которые не были представлены в обучающих данных напрямую [9].

В основе этих выдающихся возможностей лежит процесс обучения на колоссальных массивах текстовых данных. Модели анализируют и усваивают закономерности, структуру и нюансы языка из миллиардов слов, предложений и абзацев. Благодаря этому глубокому погружению в лингвистическое пространство, они приобретают способность генерировать тексты, которые по стилю, грамматике и содержанию практически неотличимы от текстов, написанных человеком. Более того, они способны адаптироваться к различным стилям и жанрам, демонстрируя гибкость и понимание контекста [10].

**Заключение**

Эволюция методов анализа больших данных демонстрирует стремительное развитие – от простых статистических моделей до сложных нейросетевых систем и генеративных моделей. Текущие достижения позволяют решать задачи, которые еще недавно казались невозможными, такие как автоматический перевод на уровне человека, генерация реалистичных текстов и глубокое понимание контекста. Эти технологии формируют основу для дальнейших инноваций в области текстовой аналитики и открывают новые возможности для бизнеса, науки и общества в целом. Однако важно помнить о вызовах и этических аспектах, связанных с разработкой и применением этих технологий, и стремиться к созданию ответственного и безопасного искусственного интеллекта.

*Список литературы:*

1. OpenAI. GPT-4 Technical Report [Электронный ресурс] / 2023. – Режим доступа: https://arxiv.org/pdf/2303.08774.pdf, свободный. – Загл. с экрана.
2. Chowdhery A. et al. PaLM: Scaling Language Modeling with Pathways [Электронный ресурс] / 2022. – Режим доступа: https://arxiv.org/pdf/2204.02311.pdf, свободный. – Загл. с экрана.
3. Anil R. et al. Pathways Language Model (PaLM) 2: Technical Report [Электронный ресурс] / 2023. – Режим доступа: https://arxiv.org/pdf/2307.07163.pdf, свободный. – Загл. с экрана.
4. Touvron H. et al. LLaMA: Open and Efficient Foundation Language Models [Электронный ресурс] / 2023. – Режим доступа: https://arxiv.org/pdf/2302.13971.pdf, свободный. – Загл. с экрана.
5. Scao T.L. et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model [Электронный ресурс] / 2022. – Режим доступа: https://arxiv.org/pdf/2211.05100.pdf, свободный. – Загл. с экрана.
6. Kaddour J. et al. Challenges and Applications of Large Language Models [Электронный ресурс] / 2023. – Режим доступа: https://arxiv.org/pdf/2307.10169.pdf, свободный. – Загл. с экрана.
7. Wei J. et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [Электронный ресурс] / 2022. – Режим доступа: https://arxiv.org/pdf/2201.11903.pdf, свободный. – Загл. с экрана.
8. Zhang S. et al. OPT: Open Pre-trained Transformer Language Models [Электронный ресурс] / 2022. – Режим доступа: https://arxiv.org/pdf/2205.01068.pdf, свободный. – Загл. с экрана.
9. Raffel C. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [Электронный ресурс] / 2023. – Режим доступа: https://arxiv.org/pdf/1910.10683v9.pdf, свободный. – Загл. с экрана.
10. Zeng A. et al. GLM-130B: An Open Bilingual Pre-trained Model [Электронный ресурс] / 2022. – Режим доступа: https://arxiv.org/pdf/2210.02414.pdf, свободный. – Загл. с экрана.