УДК 004.8 **DOI** 10.15622/ia.24.6.4

А.М. Федоров, И.О. Датьев, И.Г. Вишняков

МЕТОД ИНТЕГРАЦИИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ В АЛГОРИТМЫ ФОКУСИРОВАННОГО МОНИТОРИНГА ОТКРЫТЫХ ДАННЫХ СОЦИАЛЬНЫХ МЕДИА

Федоров А.М., Датьев И.О., Вишняков И.Г. **Метод интеграции больших языковых моделей в алгоритмы фокусированного мониторинга открытых данных социальных медиа.**

Аннотация. Актуальность исследования определяется важностью и сложностью выполнения оперативных обобщений постоянно возрастающего массива пользовательских сообщений в социальных сетях. Предлагается уменьшить сложность решаемой задачи за счет использования роботизированных алгоритмов и их автоматизированной интеллектуальной фокусировки на целевые платформы, доступность данных и их объемы. Рассматривается способность больших языковых моделей (LLMs) генерировать высококачественные, связные и контекстно-зависимые аннотации (рефераты), подходящие для динамической природы неструктурированных, «зашумленных» данных социальных сетей. Представлены особенности технологии RAG LLM при реферировании публикаций социальных сетей. Основным недостатком языковых моделей является нестабильность и сложность прослеживания результатов для подтверждения фактической точности. Авторами предложен гибридный метод для обобщения сообщений в социальных сетях за определенный период времени. Метод заключается в комплексном и вариативном сочетании классических способов извлечений данных из их хранилищ, а также реферативных и генеративных возможностей больших языковых моделей. Большие языковые модели использованы для векторизации анализируемых данных. Применение алгоритмов кластеризации к полученным векторным представлениям позволило повысить стабильность и качество результатов. В рамках технологии RAG возможности больших языковых моделей расширены средствами интеллектуального поиска в используемой для хранения исходных данных базе MongoDB. В работе представлены три конвейера, каждый из которых является вариантом реализации метода и обладает преимуществами и недостатками в различных условиях применения. Приведены используемые для оценки конвейеров метрики и произведен сравнительный анализ. В целом, метод позволяет уменьшить конфабуляции большой языковой модели и получать обобщения публикаций за разные временные периоды в режиме реального времени. Предложенный метод применяется на практике в разработанной авторами системе мониторинга открытых данных социальных медиа.

Ключевые слова: социальные медиа, публикации, обобщение, большие языковые модели, генерация дополненная поиском, интеллектуальные агенты, гибридный метод.

1. Введение. В настоящее время активно развиваются средства интеллектуальной обработки данных. Одним из самых мощных и универсальных инструментов в этой области являются технологии искусственного интеллекта, реализующие их нейронные сети, в частности большие языковые модели (LLM – large language models). Флагманами разработки в этой сфере регулярно демонстрируются новые технологии, средства и получаемые с их помощью результаты.

Такое мощное развитие индустрии технологий искусственного интеллекта требует вложения очень больших ресурсов: материальных, энергетических, производственных, научных, кадровых и др. В свою очередь разработчики среднего звена используют доступные флагманские результаты, развивают их, в меру имеющихся сил совершенствуют и применяют на практике. Несмотря на все достижения, потенциал развития в области технологий искусственного интеллекта остается очень высоким. Многие эксперты сейчас говорят о том, что аббревиатуру ИИ скорее можно расшифровать как «имитация интеллекта», и что настоящий искусственный интеллект только еще впереди.

Тем не менее, доступный ИИ-инструментарий в виде больших языковых моделей ускорил решение задач обработки и генерирования текста. Для управления работой таких моделей уже не нужно быть программистом. Языком управления стал естественный язык, с помощью которого составляется запрос (промпт). Однако даже в этом случае следует понимать, что от способа формулирования, формирования, оформления, наполнения запросов зависит качество получаемого результата.

В настоящей работе авторы решают задачу обобщения публикаций социальных сетей на основе результатов мониторинга созданной авторами программной системы получения и анализа данных социальных медиа [1] с использованием программных веб-роботов (ботов) для сбора сообщений с платформ VK, Telegram, Whatsap. Обобщение или суммаризация публикаций – получение (синтез, генерация) тематических абстрактов (аннотаций, рефератов, дайджестов, тем), другими словами, получение сокращенного изложения содержания исходного текста. Практической целью данной процедуры является решение различных задач автоматизированной обработки текста, в том числе, в данной работе, сокращение объема текста для дальнейшей обработки оператором системы мониторинга.

Исходными данными являются тексты пользовательских сообщений, получаемые в процессе мониторинга открытых данных социальных медиа. Объем таких текстов в общем случае ничем не ограничен. В работе реализован методологический подход и рассматриваются все возможные варианты объема исследуемых сообщений: 1) помещаются в контекст, 2) не помещаются в контекст. Счетный, но бесконечный характер числа пользовательских сообщений позволяет выбрать такое их количество, что оно не будет помещаться в контекстное окно используемых языковых моделей, даже с учетом постоянного роста их возможностей.

На основе заданного временного интервала для текстовой выборки публикаций социальных сетей необходимо сгенерировать обобщения (резюме, рефераты, аннотации, дайджесты, темы) и минимизировать возможные ошибки, обеспечить стабильность результата и его прослеживаемость, т.е. подтверждение на основе исходных данных.

Формально данная задача поставлена так: для получения оперативных и качественных тематических обобщений над элементами массива текстов *ДТ* объемом *ОДТ* найти удовлетворяющий заданному критерию *К* вариант реализации конвейера обработки на основе интеллектуального компонента в виде большой языковой модели *ЯМ* с объемом контекста *ОЯМ*. Комплексный критерий *К* состоит из двух: 1) «*воспроизводимость*» и 2) «*прослеживаемость*». Обе составляющие *К* используются для оценки негативных эффектов – галюцинаций/конфабуляций, возникающих в результате использования генеративной функции *ЯМ*. Первый критерий позволяет оценить стабильность получения результата при неизменных начальных условиях, а второй – возможность восстановить исходные данные на основе полученного результата. Дополнительным условием является решение поставленной задачи с учетом разных соотношений *ОЯМ* и *ОДТ*.

Таким образом, основная обладающая научной новизной задача, поставленная в данном исследовании, заключается в выявлении, изучении, системном анализе и практической реализации способов получения генеративных обобщений при снижении конфабуляций языковых моделей, используемых в конвейерах решений.

2. Социальные сети и большие языковые модели. Большой объем пользовательского контента на платформах социальных сетей способствует повышению актуальности разработки методов резюмирования (аннотирования, реферирования, обобщения, суммаризации) сообщений за определенные временные периоды. Искусственный интеллект, в частности достижения в обработке естественного языка (NLP), привели к появлению сложных методов автоматического резюмирования, адаптированных к особенностям специфических данных социальных сетей. Автоматическое резюмирование текста (ATS – Automatic Text Summarization) значительно изменилось за время своего существования. Алгоритмы резюмирования могут быть извлекающими (выбор ключевых предложений), абстрактными (генерация новых предложений), а также гибридными. Ранние методы основывались на статистических моделях и системах на основе правил, в основном извлекая ключевые

предложения. Улучшению результатов аннотирования также способствовали методы тематического моделирования, в том числе метод LDA (Latent Dirichlet Allocation)[2], более подробный обзор которых и применение к текстам социальных сетей представлены в работе [3]. В настоящее время область решения сместилась в сторону машинного обучения, при этом предварительно обученные большие языковые модели (LLM), такие как GPT, BERT, DeepSeek, GigaChat, YandexGPT демонстрируют значительные улучшения в создании связных и контекстно-зависимых резюме на основе «зашумленных» публикаций социальных сетей. В работе [4] представлен сравнительный анализ методов суммаризации текста с использованием больших языковых моделей. Авторы работы [5] отмечают, что интеграция технологий ИИ и генеративного ИИ предлагает многообещающие способы решения задач исследовательского поиска и резюмирования.

Сообщения в социальных сетях представляют собой совокупность текста и различных метаданных, например, эмодзи, эмотиконы и мультимедиа. Нарушение последовательности слов и неформальный язык, использующиеся в социальных сетях, являются серьезным препятствием для использования только статистических методов, в том числе алгоритмов выделения ключевых слов, таких как Rake [6], YAke [7], TextRank [8]. Реферирование этого контента требует моделей, которые могут эффективно обрабатывать сообщения, обладающие этими особенностями. В исследовании [9] проведён обзор гибридных методов реферирования и предложена система, объединяющая семантические (основанные на значении) и статистические (основанные на частоте) методы для обобщения постов в Instagram, что позволило повысить точность суммаризации.

В работе [10] освещается использование представлений на основе векторов, методов снижения размерности и технологий генеративного ИИ (например, LLM с подсказками Chain of Thought (CoT) – цепочки рассуждений [11, 12]) для выявления тем в больших открытых данных социальных сетей.

LLM превосходит другие методы суммаризации по связности результата и адаптивности к различным предметным областям, включая социальные сети. Способность выполнять контекстное и малочастотное обучение позволяет LLM генерировать краткие и контекстно релевантные обобщения даже из динамического и неформального контента, найденного на этих платформах [4].

Для извлечения и обобщения доминирующих тем из данных социальных сетей за определенный временной отрезок также

используются агентные генеративные системы ИИ, что позволяет масштабировать и адаптировать тематический анализ, что представлено в работе [10].

Все больше внимания уделяется тому, чтобы сделать сгенерированные технологиями ИИ резюме более прозрачными. Концепция "отслеживаемого текста" представлена в [13], и направлена на связывание контента резюме с его первоначальными источниками для повышения доверия пользователей и интерпретируемости выходных данных ИИ.

3. Дополненная поиском генерация (RAG). Технология Retrieval-Augmented Generation (RAG) – это гибридный подход в обработке естественного языка, который сочетает в себе способности к генерации больших языковых моделей с внешними системами извлечения информации [14]. Этот подход помогает LLM преодолевать такие ограничения, как галлюцинации, устаревшие знания и отсутствие специфики предметной области, позволяя моделям получать доступ к актуальным и релевантным внешним данным и использовать их во время генерации. Системы RAG обычно состоят из двух основных компонентов: т.н. ретривер (компонент поиска) и генератор. Ретривер находит и извлекает релевантные документы или фрагменты данных из внешней базы знаний на основе входного запроса. Генератор формирует свой вывод, основываясь как на исходном запросе, так и на извлеченном контексте, чтобы сгенерировать окончательный ответ. RAG быстро набирает популярность в различных областях, включая ответы на вопросы из различных предметных областей и суммаризацию текстов. В компоненте поиска могут использоваться и сочетаться различные методы, например, разреженный поиск (традиционный поиск по ключевым словам, алгоритм BM25), плотный поиск (семантический поиск с использованием векторизации на основе модели Sentence Transformers).

Авторы работы [15] отмечают, что для эффективного извлечения релевантной информации из больших неструктурированных данных социальных сетей требуются продвинутые стратегии поиска, и предлагают использовать "sentence window retrieval" и "auto-merging", чтобы найти наиболее точные и контекстуально релевантные фрагменты.

В работе [16] авторы оценивали различные стратегии поиска для технологии RAG и пришли к выводу, что гибридный поисковик (объединяющий разреженные и плотные методы, использующий расширение запроса и обратное ранговое слияние) показал наилучшие

результаты. В качестве общих причин галлюцинаций указываются: поиск, не приведший к результату и недостаточная (неполноценная) генерация. При гибридном подходе улучшилась релевантность поиска документов, понизились показатели галлюцинаций и повысилась точность ответов при тестировании на таких бенчмарках, как HaluBench.

Многоагентные гибридные фреймворки помимо классического поиска, могут включать несколько специализированных «агентов» или модулей. Каждый агент может фокусироваться на другом аспекте рассуждения или интеграции знаний. Примером является фреймворк RAG-KG-IL [17], объединяющий RAG для ответов на основе доказательств, граф знаний (KG) для предоставления структурированного фактического контекста, а также итеративное обучение (IL) для постоянного улучшения результата и исправления ошибок. Это многоагентное сотрудничество помогает улучшить рассуждения, уменьшить галлюцинации и динамически выбирать и объединять источники знаний.

Технология RAG является ключевой при разработке больших языковых моделей (LLM). Надежность систем RAG находится в стадии изучения. Авторы [18] предлагают фреймворк, который оценивает системы RAG по шести ключевым измерениям: фактичность, надежность, справедливость, прозрачность, подотчетность и конфиденциальность. Следует отметить, что решение не является исчерпывающим, авторы лишь предпринимают попытку заложить структурированную основу для будущих исследований и предоставить практические идеи для повышения надежности систем RAG в реальных приложениях. Человеческая экспертная оценка до сих пор остается наиболее важной.

4. Особенности RAG для суммаризации постов в социальных сетях. Социальные сети содержат огромное количество разнообразной и постоянно меняющейся информации. RAG позволяет LLM извлекать наиболее релевантные данные, обеспечивая более точные и актуальные сводки. Например, в исследовании, касающемся ответов на медицинские вопросы с использованием данных социальной сети Reddit, двухслойная RAG-система показала сравнимую производительность с GPT-4 в релевантности и связности в условиях ограниченных ресурсов [19].

LLM могут "галлюцинировать", то есть генерировать ложную информацию. Благодаря RAG, модель получает фактические данные из базы знаний, что уменьшает вероятность таких ошибок и делает сводки более надежными [20].

RAG может учитывать специфический контекст социальных медиа, включая сленг, сокращения и быстро меняющиеся тренды. В работе [21] авторы продемонстрировали возможности технологии RAG для прогнозирования реакций сообщества на посты в социальных сетях, учитывая идеологическое и эмоциональное разнообразие.

Одной из важнейших тем в исследованиях ИИ является вопрос о том, какой из подходов – технология RAG, тонкая настройка (fine-tuning) [22, 23] моделей или гибридный подход – лучше? Каждый метод имеет свои преимущества и недостатки, поэтому необходимо понимать их применимость.

Тонкая настройка подразумевает обновление весов предварительно обученного LLM на специфическом для предметной области наборе данных и требует повторного обучения каждый раз при поступлении новой информации, что в свою очередь, для больших моделей занимает много времени и требует вычислительных затрат. RAG обеспечивает актуальные ответы за счет извлечения знаний в реальном времени, но обладает большей задержкой из-за предварительного этапа поиска. Гибридный подход сочетает тонкую настройку для адаптации к предметной области с RAG для обновлений в реальном времени. Таким образом, выбор между тонкой настройкой, RAG или гибридным подходом зависит от компромисса между адаптивностью в реальном времени, вычислительными затратами и специфичностью предметной области [24].

5. Многошаговое промптирование или цепочки рассуждений. При работе с языковыми моделями очень важным элементом является т.н. промптирование, т.е. использование правильных запросов-подсказок. В работе [25] авторы исследуют различные варианты подсказок и демонстрируют соответствующие улучшения выдачи языковой модели.

Наиболее перспективным для решения задач резюмирования представляется метод CoT (Chain-of-Thought) – цепочки рассуждений – это метод, который побуждает LLM генерировать промежуточные шаги рассуждения перед тем, как дать окончательный ответ. Это улучшает прозрачность и качество рассуждений модели. Работы, содержащие прямое использование CoT с RAG для суммаризации публикаций социальных сетей все ещё находятся на ранних этапах развития, однако представляются перспективными по следующим причинам. CoT может сделать процесс суммаризации более понятным, показывая, как модель сформировала резюме. Социальные сети могут содержать сложные дискуссии и взаимосвязи между постами. CoT, в сочетании с RAG, может помочь LLM выявить

эти сложные взаимодействия и создать более содержательные резюме. Например, в работе [26] авторы используют графовый поиск для многошагового рассуждения, что концептуально схоже с CoT. Путем извлечения и интеграции информации из взаимосвязанных документов (например, цепочек в социальных сетях), предлагаемый фреймворк GRAG (Graph Retrieval-Augmented Generation) позволяет LLM выполнять пошаговые рассуждения с использованием графов, улучшая аннотирование и качество выводов.

Обобщение сообщений в социальных сетях за определенные периоды времени с помощью технологий ИИ – это быстро развивающаяся область, значительно усиленная гибридными методами и большими языковыми моделями RAG LLM. Эти технологии позволяют создавать эффективные и контекстно-релевантные резюме, которые полезны для таких приложений, как мониторинг событий и динамический (временной) тематический анализ. Текущие исследования продолжают решать проблемы, связанные со спецификой данных социальных сетей, а также прозрачностью и прослеживаемостью сгенерированных резюме.

При применении методов извлекающего резюмирования может отсутствовать связность. Абстрактное резюмирование более связно, но требует больших вычислительных ресурсов. Тематическое извлечение на основе эмбеддингов хорошо работает для выявления тем, но требует больших наборов предварительно подготовленных данных. Временное тематическое моделирование эффективно для обнаружения событий и временного анализа, но может упускать нюансы контекста. Технологии генеративного ИИ с подсказками в виде цепочки рассуждений (CoT) создает более высококачественные резюме, но зависит от возможностей LLM.

Большие языковые модели (LLM) являются мощным средством для решения задачи резюмирования, но склонны к «галлюцинациям» или конфабуляциям – генерации результатов, которые фактически неверны или не подтверждаются доказательствами. Гибридные методы, которые объединяют LLM с другими стратегиями, такими как расширенные методы поиска (в частности технология RAG) или многоагентные системы, наиболее перспективны для повышения надежности LLM и уменьшения их подверженности галлюцинациям.

RAG LLM представляют собой значительный шаг вперед в суммаризации постов социальных сетей, предлагая улучшенную фактическую точность, контекстуальную привязку и адаптируемость к динамичному контенту. Методологические инновации в поиске и инженерии подсказок, включая развивающиеся исследования

промптирования цепочек рассуждения, расширяют возможности этих систем. Необходимы масштабируемые, адаптируемые фреймворки и надежные оценочные метрики для результатов резюмирования публикаций в социальных сетях. На сегодняшний день, человеческая экспертная оценка остается наиболее важной.

6. Метод интеграции больших языковых моделей в алгоритмы фокусированного мониторинга. Используемый в рамках системы мониторинга социальных медиа механизм фокусировки предполагает динамическую корректировку заданий на мониторинг с учетом поступающих данных. Процесс такой корректировки представляет собой состоящих из нескольких шагов конвейер, в рамках которого полученные на предыдущем шаге результаты передаются на следующий и т.д. Работоспособность конвейера зависит от качества получаемых на каждом шаге данных. Важным является повышение этого качества.

Идея метода заключается в создании надежных и стабильных конвейеров, обеспечивающих качественный результат работы.

В рамках решения задачи обобщения множества сообщений в понятие качества также включается гарантированная возможность проверки полученного результата выборкой исходных данных.

Получение результирующих обобщений может рассматриваться и в качестве самостоятельной цели, и в качестве промежуточного этапа процедуры фокусировки. В любом случае от результата в первую очередь требуется стабильная воспроизводимость и прослеживаемость. Первое свойство можно проверить серией экспериментов. Второе свойство требует работы с дополнительными конструкциями, позволяющими на основе полученных результатов найти исходные данные.

Последовательность действий (блок-схема алгоритма) метода для получения искомого результата представлена на рисунке 1 в виде стандартной UML-диаграммы деятельности. Однако изображение в таком виде менее выразительно, так как не позволяет продемонстрировать ряд важных специфических аспектов, к которым в том числе относятся и объектная модель реализации алгоритма, и временная последовательность действий, и параллельность взаимодействия серверных и клиентских акторов. По этой причине здесь и далее в статье для иллюстрации реализованных алгоритмов авторами используются диаграммы последовательности, как более подходящий инструмент из многообразной палитры унифицированного языка моделирования UML.

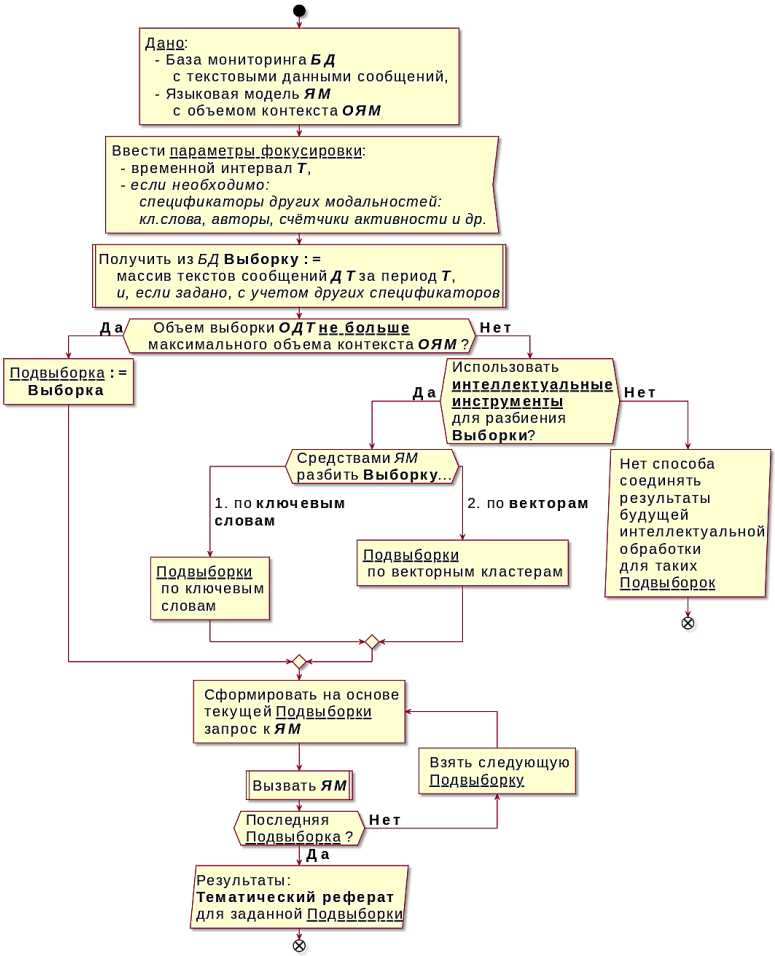

Рис. 1. Блок-схема метода получения обобщений для сообщений

Далее рассматриваются три конвейера, в которых несколькими способами решается поставленная задача.

6.1. Прямой запрос к большой языковой модели. Тривиальным вариантом получения генеративных обобщений (списка тем) является непосредственный запрос к большой языковой модели. Предварительно по заданным критериям в базе данных мониторинга производится поиск необходимых данных. Структурированный список полученных текстов сообщений включается в контекст запроса. Пример формулировки

запроса: «Выдели пять наиболее крупных тем в представленном ниже списке пользовательских сообщений: {список сообщений}».

В данном случае на основе технологии RAG большая языковая модель, используя свое умение работать с текстами на естественном языке и основываясь только на содержании запроса, в качестве результата выдаст список формулировок тем.

На рисунке 2 в виде UML диаграммы последовательности представлен порядок действий данного варианта реализации алгоритма получения тем для заданного списка сообщений.

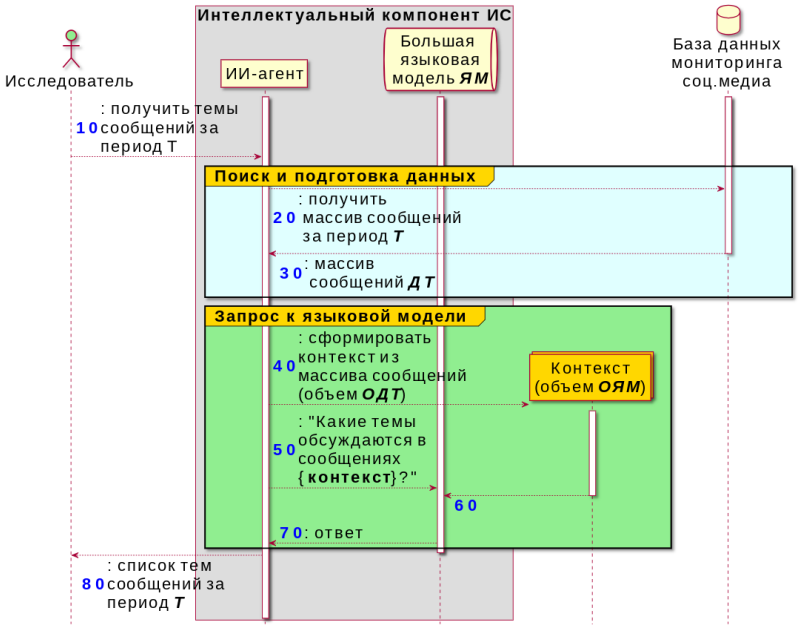

Рис. 1. Извлечение тем из списка сообщений путем прямого запроса к языковой модели

Особенностью данного подхода является то, что при повторении последовательности действий при одних и тех же исходных данных формулировки тем в целом семантически будут соответствовать ранее полученным вариантам, но синтаксически будут различаться. Эта нестабильность связана с вероятностными свойствами генеративной функции языковой модели. Снизить вероятность таких изменений можно с помощью конфигурационного параметра «температура модели», задав ему самое минимальное значение.

Преимуществом данного подхода является классическая для прикладных языковых моделей одношаговая функциональная реализация и зависящая лишь от объема контекста скорость выполнения. Для его реализации достаточно сформировать один запрос к большой языковой модели, разместив в нем список необходимых для анализа сообщений.

6.2. Предварительное определение ключевых слов. Для повышения стабильности результатов получения генеративных обобщений (тем) из данных социальных медиа предложен вариант, основанный на предварительном получении ключевых слов и последующем их использовании в фокусировке алгоритмов именования тем.

В отличие от предложенного в предыдущем пункте варианта здесь на первом этапе с помощью большой языковой модели на основании множества сообщений формируются списки наиболее релевантных ключевых слов по каждой потенциальной теме.

На следующем этапе списки полученных ключевых слов используются для поиска содержащих их сообщений. В рамках данного исследования для извлечения таких сообщений использовался интеллектуальный поиск, который является встроенным средством в систему управления базами данных MongoDB. В данном случае очень полезным оказывается свойство интеллектуального поиска по нахождению заданных ключевых слов с учетом их словоформ.

На финальном этапе полученные на основе отбора по ключевым словам множества сообщений отдельными запросами обрабатываются с помощью большой языковой модели. Результатом такой обработки являются формулировки тем для каждой группы сообщений.

Таким образом, представленный выше конвейер обработки данных позволяет получить искомый набор тем и дополнительно список ключевых слов, с помощью которых можно найти сообщения, относящиеся к каждой теме.

Усложнение представленного варианта обработки данных введением в него промежуточной стадии компенсируется более стабильными результатами. Дополнительным свойством получаемых результатов является прослеживаемость (прозрачность), которая в большинстве случаев требуется в качестве обязательной характеристики используемого на практике решения. Также данный подход позволяет работать с исходными данными, объемы которых превышают максимальный размер запроса к языковой модели. В этом случае ключевые слова для тем должны быть получены отдельными запросами для каждого из подмножеств исходных данных. В итоге

полученные частичные данные объединяются и уже в таком виде обрабатываются дальше.

На рисунке 3 в виде UML диаграммы последовательности представлен порядок действий варианта реализации алгоритма получения обобщений через предварительную обработку соответствующих ключевых слов.

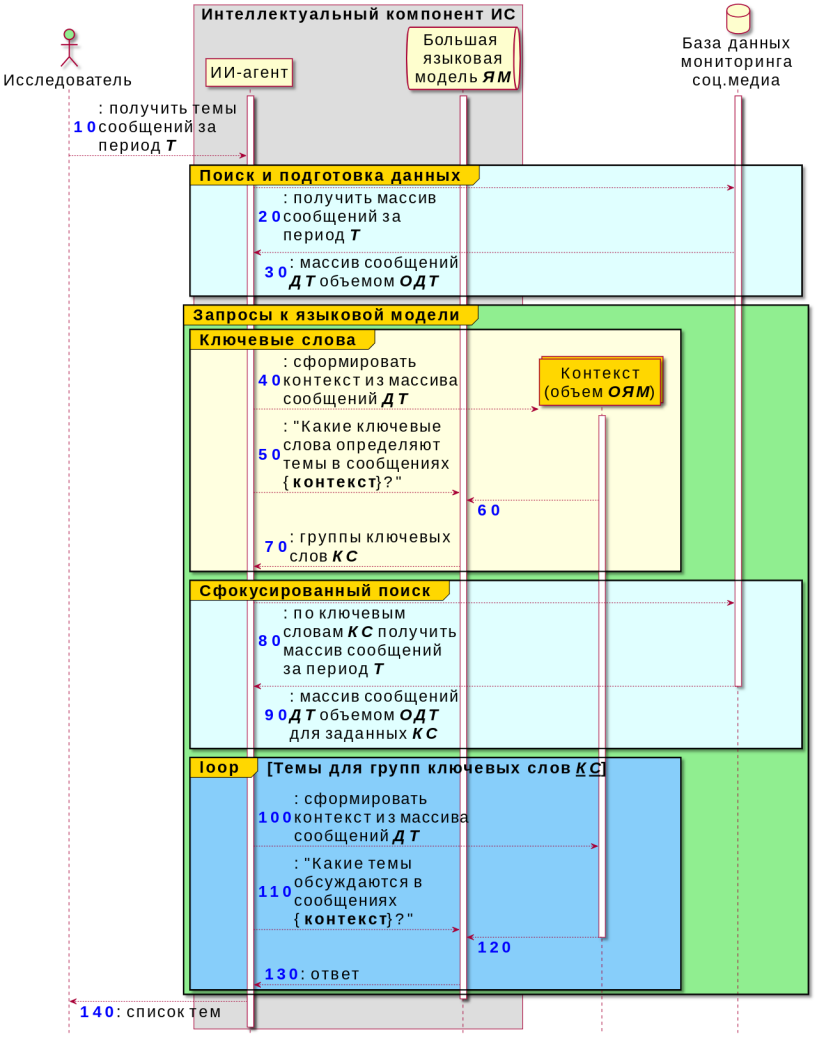

Рис. 2. Извлечение тем из списка сообщений на основе обработки ключевых слов

На рисунке 4 в виде UML диаграммы последовательности представлен вариативный этап работы конвейера, учитывающий ситуацию, когда объем исходных данных превышает максимальный размер контекста запроса к языковой модели.

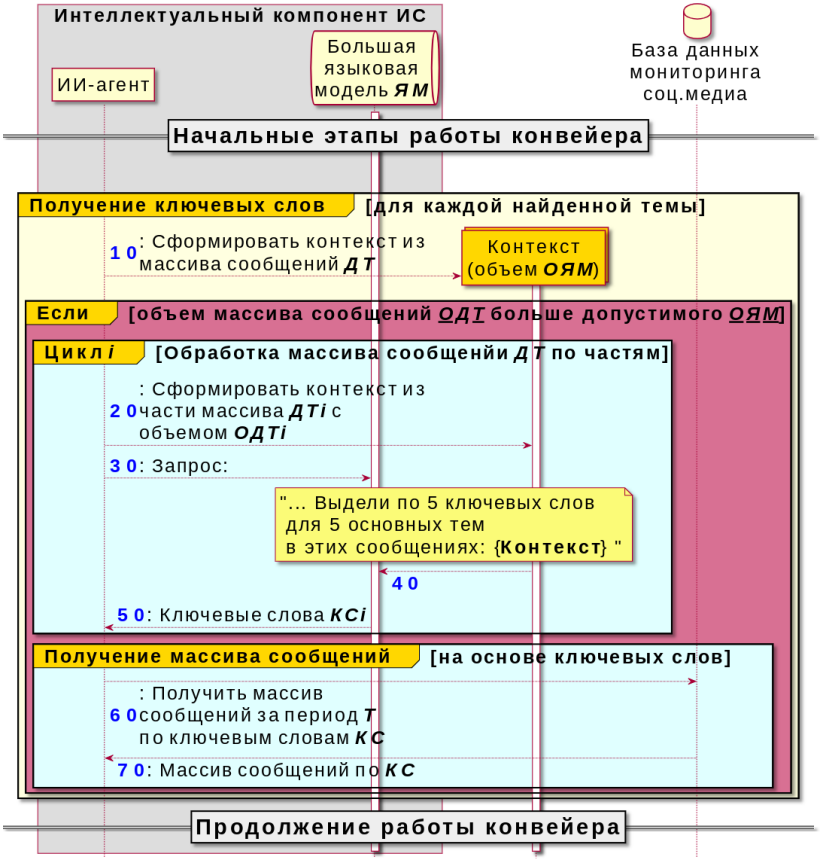

Рис. 3. Этап предварительного отбора ключевых сообщений

6.3. Векторизация и кластеризация. В данном варианте большая языковая модель главным образом используется в качестве инструмента для размещения текстов социальных медиа в своем векторном семантическом пространстве.

Целевые сообщения векторизуются и записываются в векторную базу данных. В данной работе используется векторная база данных Milvus.

Далее полученное множество векторов с помощью алгоритма HDBSCAN [27] разбивается на кластеры. Данный алгоритм кластеризации предлагается в качестве рабочего примера разработчиками базы Milvus. Однако для исследовательских целей в будущих работах планируется использовать различные алгоритмы кластеризации, т.к. в рамках решаемой задачи результат разбиения на кластеры играет большую роль.

Заключительным действием является генерация с помощью большой языковой модели обобщений (формулировок тем) для множества сообщений, входящих в каждый полученный кластер.

На рисунке 5 в виде UML диаграммы последовательности представлен порядок действий варианта реализации алгоритма получения тем через предварительную кластеризацию векторов исходных сообщений.

Особенностью данного варианта является то, что группировка сообщений по темам производится явно с помощью алгоритмов кластеризации. Это повышает точность и прослеживаемость результатов. Также преимуществом является то, что ограничение на количество анализируемых сообщений задается характеристиками векторной базы данных, а не максимальным количеством токенов, из которых формируется промпт для языковой модели. Фактически векторизации и последующей кластеризации можно подвергать всё множество сообщений из базы мониторинга, в то время как для формирования промпта количество таких сообщений будет значительно ограничено.

Основные компоненты предложенного метода, в виде фрагментов программного кода, зарегистрированы авторами в Роспатент [28] и доступны по ссылке https://git.iimm.ru/iimm-pub/2025-iimm-rid2-fed_dat_vish-soc_mon_kw_themes_gen.

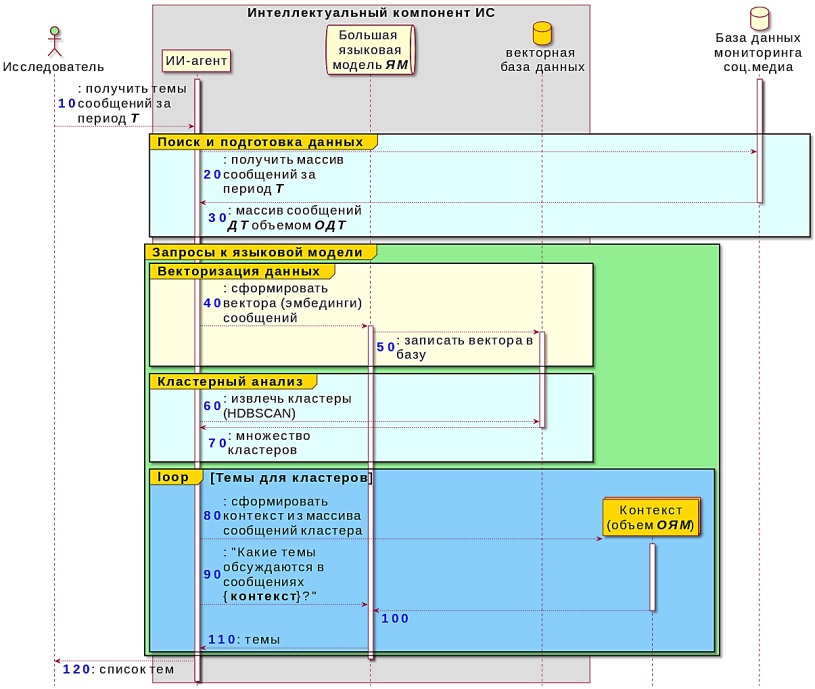

Рис. 5. Извлечение тем из списка сообщений с помощью кластеризации векторов

7. Комплекс метрик для оценки конвейеров метода

7.1. Стабильная воспроизводимость результата. Точность на а) синтаксическом б) семантическом уровне. Уровень стабильности определяется количеством повторений результата при повторной генерации при одинаковых начальных условиях. Более строгий синтаксический способ предполагает точное синтаксическое равенство получаемых результатов. Семантический вариант предполагает смысловое равенство результатов.

7.2. Прослеживаемость результата. Проверка а) синтаксическим б) интеллектуальным (например, MongoDB) поиском.

Факт прослеживаемости результата определяется возможностью на его основе найти в исходном наборе данных элементы, благодаря которым результат получен.

В данном случае хорошим примером является получение в качестве результата множества ключевых слов, поиск которых в

исходных данных приведет к множеству текстов, являющихся основой полученного результата. Такого рода операция нужна для проверки достоверности полученных результатов.

7.3. Объем обрабатываемых (исходных) данных. Ограничения на объем а) контекста языковой модели или б) хранилища данных для «обогащения» (дополнения) запроса.

Классическое использование технологии RAG накладывает ограничение на объем данных, на основе которого формируется результат. Ограничение зависит от измеряемого в токенах максимального размера текста запроса, в который включаются и управляющие инструкции и непосредственно сами данные. В режиме конвейеризации можно делать несколько таких запросов и тем самым каждый раз иметь возможность использовать запросы максимального объема. Однако, если использовать большую языковую модель в качестве векторизатора, то можно работать одновременно со всем объемом данных. Векторные хранилища имеют достаточно большой объем и позволяют разместить все необходимые для работы данные. Некоторые задачи успешно решаются с помощью непосредственной обработки векторов. Например, одним из этапов конвейера обработки данных может быть кластеризация, которую как раз и можно решать, напрямую обращаясь к векторному хранилищу.

8. Прогнозирование результатов экспериментов. Потенциальные результаты работы представленных конвейеров соответствуют особенностям их построения. Ожидаемые результаты запланированных экспериментов представлены в таблице 1, а ниже дано соответствующее обоснование сделанных прогнозов.

Таблица 1. Варианты реализации метода и соответствующие показатели

Вариант реализации метода / Метрика	Классический RAG LLM	Конвейерный RAG LLM	Кластерный анализ векторов LLM
«Воспроизводимость» (Стабильность)	-	-+	++
«Прослеживаемость»	-	++	++
«Объем данных»	-	-+	++

Одношаговый вариант конвейера с комплексным запросом к большой языковой модели позволяет сразу получить и список тем, и список соответствующих ключевых слов. Генеративный характер работы модели предопределяет синтаксическую вариативность результата и весьма вероятные эффекты проявления «конфабуляции».

Смысловая составляющая таких результатов сохраняется. Однако поиск подтверждений в исходных данных на основе ключевых слов не дает гарантированного результата.

Конвейер с предварительной генерацией ключевых слов и последующим поиском соответствующих им исходных данных имеет б**о**льшую стабильность. Это связано с тем, что запрос к языковой модели о получении генеративных обобщений формируется на основании отобранных по ключевым словам сообщений. Сами же ключевые слова проходят дополнительную проверку на присутствие в сообщениях, что гарантирует возможность использования их для нахождения исходных сообщений.

Третий вариант также основывается на предварительном этапе, в рамках которого в векторном пространстве выделяются кластеры сообщений. Каждый запрос к языковой модели ограничен только относящимися к определенному кластеру сообщениями, что значительно повышает стабильность получаемого результата. В этом случае начальная кластеризация гарантирует возможность однозначного сопоставления результата обобщения и соответствующих исходных данных.

Таким образом, более качественными вариантами реализации конвейеров генеративных обобщений можно считать те, в которых упрощается структура запроса при обращении к языковой модели, а на отдельном этапе обеспечивается сохранение признаков, позволяющих найти необходимые для контроля прослеживаемости исходные данные.

9. Результаты эксперимента. На основе представленных трех конвейеров проведена серия экспериментов, результаты которых оценены с помощью трех метрик, приведенных выше.

В экспериментах использованы два набора данных, сформированных на основе реальных данных из системы мониторинга домовых чатов:

– Набор данных №1 (Дс1): 280 сообщений, 18509 символов.
– Набор данных №2 (Дс2): 911 сообщений, 59211 символов.

Для запросов и векторизации использована языковая модель:

– bambucha/saiga-llama3:latest, температура модели 0.1.

Суть эксперимента состояла в том, чтобы на каждом наборе данных несколько раз проверить работу каждого конвейера и оценить полученные результаты. Каждый конвейер на каждом наборе данных отработал десять раз. Обобщенные результаты приведены в таблице 2.

Оценка по метрике «Объем данных» определяется характеристиками специально подобранных наборов данных. Первый

набор сформирован на основе данных суточного мониторинга домовых чатов. Объем сообщений позволяет загрузить их в контекст одного запроса к языковой модели. Таким образом становится возможна работа одноэтапного конвейера, в рамках которой все необходимые сведения о данных получаются в результате единого запроса. Второй набор данных сформирован на основе более продолжительного мониторингового периода. В этом случае суммарный объем текстов сообщений превышает ограничения на размер контекста запроса к языковой модели. В результате для обработки таких данных либо требуется предварительное их уплотнение, либо применение соответствующих таким условиям технологий. Для этих целей как раз подходят второй и третий вид описываемых в данной работе конвейеров. В таблице 2 отражены результаты экспериментов со всеми вариантами конвейеров в отношении первого (Дс1) и второго (Дс2) наборов данных.

Оценки по метрике «Воспроизводимость (стабильность)» сделаны с помощью экспертов. В результате десяти запусков конвейера для одного и того же набора данных получены десять наборов по пять обобщающих формулировок тем в каждом. Сформирована сравнительная таблица, в которую полученные наборы тем записаны и по строкам, и по столбцам, а на их пересечении (кроме диагональных элементов) эксперты выставили оценку соответствия. Для данной работы эксперты оценивали смысловую эквивалентность тематических наборов. Пара тематических наборов полностью (на 100%) соответствует друг другу, когда для каждой темы из одного набора найдена эквивалентная тема из сравниваемого набора. Если для темы не нашлось эквивалентной, то оценка соответствия пропорционально уменьшается (на 1/5, т.е. на 20%). В итоге, среднее значение по сравнительным таблицам всех экспертов дает искомую оценку по метрике «Воспроизводимость».

На первом наборе данных лучшие результаты продемонстрировал второй вариант. Это связано с тем, что обработка каждой темы в отдельном запросе дает более стабильные формулировки. Противоположный результат продемонстрировал первый вариант, так как использование одного запроса сразу для всех тем повышает уровень вариативности их формулировок. Третий вариант продемонстрировал самый плохой результат. Это связано с тем, что используемый набор данных не имеет достаточного объема, для того чтобы их качественно сформировать и обработать на кластерном уровне. В то же время, на более объемном наборе данных третий вариант продемонстрировал наилучший результат.

Оценки по метрике «Прослеживаемость» получены автоматически. Алгоритм расчета формально соответствует классической метрике из машинного обучения Precision (точность), которая рассчитывается как отношение количества правильно классифицированных положительных объектов (True Positives) к общему числу положительных предсказаний модели (True Positives + False Positives). В описываемых экспериментах положительными исходами стали полученные для темы ключевые слова, поиск по которым в наборе данных дал ненулевые результаты. Как и предполагалось до эксперимента, некоторая часть сгенерированных языковой моделью ключевых слов, возможно, по смыслу отражает суть темы, но по факту не позволила найти в наборе данных ни одного содержащего их сообщения. Оценка по данной метрике произведена как с помощью оригинальных ключевых слов, так и с помощью их нормализованных (стемминг) вариантов. В таблице 2 данная метрика имеет два соответствующих представления. Результаты экспериментов для всех рассматриваемых конвейеров показали в среднем примерно одинаковые результаты. Однако для наиболее реального на практике набора данных (Дс2) свое преимущество продемонстрировал третий вариант. Это подтверждает тот факт, что использование генеративных возможностей языковых моделей вносит дополнительную вариативность в получаемые результаты. С другой стороны, использование языковых моделей для формирования векторных представлений требует дополнительной специфической программной обработки результатов. Золотой серединой является конвейерное использование языковых моделей с подходящими вариантами предварительного уплотнения, фильтрации и фрагментирования исходных данных, включаемых в контекст запроса.

Таблица 2. Оценки результатов эксперимента

Вариант реализации метода	Классический RAG LLM		Конвейерный RAG LLM		Кластерный анализ векторов LLM	
Метрика	Дс1	Дс2	Дс1	Дс2	Дс1	Дс2
«Воспроизводимость» (стабильность)	60%	-	80%	62%	66%	100%
«Прослеживаемость»	68%	-	68%	63%	65%	89%
«Прослеживаемость» (стемминг)	81%	-	81%	70%	70%	94%
Объем данных	+	-	+	+	+	+

9. Заключение. Предложен и опробован на практике гибридный метод интеграции больших языковых моделей в алгоритмы фокусированного мониторинга открытых данных социальных медиа. В качестве примера практического применения метода реализованы три разнотипных конвейера для получения генеративных обобщений на основе множества сообщений социальных медиа. Каждый конвейер включает в себя этап обращения к большой языковой модели. С помощью сравнительного анализа показано, что особенности реализации конвейеров определяют качество получаемого с помощью них результата. В данном исследовании качество определено воспроизводимостью результатов, их прослеживаемостью, а также допустимым объемом исходных данных.

Литература

1. Датьев И.О., Федоров А.М., Ревякин А.А. Фокусированный сбор и обработка открытых данных социальных медиа // Онтология проектирования. 2024. Т. 14. № 4(54). С. 569–581. DOI: 10.18287/2223-9537-2024-14-4-569-581.
2. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. vol. 3. pp. 993–1022.
3. Федоров А.М., Датьев И.О., Вишняков И.Г. Проектирование информационной системы комплексного тематического анализа больших данных социальных медиа // Онтология проектирования. 2024. Т. 14. № 1(51). С. 55–70. DOI: 10.18287/2223-9537-2024-14-1-55-70.
4. Zhang Y., Jin H., Meng D., Wang J., Tan J. A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods. arXiv preprint arXiv:2403.02901. 2024.
5. Glickman M., Zhang Y. AI and Generative AI for Research Discovery and Summarization // Harvard Data Science Review. 2024. vol. 6(2). DOI: 10.1162/99608f92.7f9220ff.
6. Rose S., Engel D., Cramer N., Cowley W. Automatic Keyword Extraction from Individual Documents // Text mining: applications and theory. 2010. pp. 1–20. DOI: 10.1002/9780470689646.ch1.
7. Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. YAKE! Keyword extraction from single documents using multiple local features // Information Sciences. 2020. vol. 509. pp. 257–289. DOI: 10.1016/j.ins.2019.09.013.
8. Mihalcea R., Tarau P. TextRank: Bringing Order into Text // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: Association for Computational Linguistics, 2004. pp. 404–411.
9. Tabanmehr Z., Akhtarkavan E. Automatic summarization of Instagram social network posts by combining semantic and statistical approaches // 6th International Conference on Pattern Recognition and Image Analysis (IPRIA). 2023. pp. 1–6. DOI: 10.1109/IPRIA59240.2023.10147186.
10. Ghali M.K., Farrag A., Lam S., Won, D. Beyondwords is All You Need: Agentic Generative Ai Based Social Media Themes Extractor. URL: https://ssrn.com/abstract=5155101 (дата обращения: 26.10.2025)
11. Wei J., Wang X., Schuurmans D., Bosma M., Ichter B., Xia F., Chi E., Le Q., Zhou D. Chain-of-thought prompting elicits reasoning in large language models // Proceedings

of the 36th International Conference on Neural Information Processing Systems (NIPS'22). NY, USA: Curran Associates Inc., Red Hook, 2022. pp. 24824–24837.
12. Kojima T., Gu S.S., Reid M., Matsuo Y., Iwasawa Y. Large language models are zero-shot reasoners // In Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22). NY, USA: Curran Associates Inc., Red Hook, 2022. pp. 22199–22213.
13. Kambhamettu H., Flores J., Head A. Traceable Text: Deepening Reading of AI-Generated Summaries with Phrase-Level Provenance Links. arXiv preprint arXiv:2409.13099. 2024. DOI: 10.48550/arXiv.2409.13099.
14. Zhao S., Yang Y., Wang Z., He Z., Qiu L.K., Qiu L. Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely. arXiv preprint arXiv:2409.14924. 2024. DOI: 10.48550/arXiv.2409.14924.
15. Hidayaturrahman, Prawira I. Retrieval-Augmented Generation for Social Media Content Creation with Sentence Window and Auto-Merging Retrieval. International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS). 2024. pp. 608–613. DOI: 10.1109/ICIMCIS63449.2024.10956838.
16. Zhang W, Zhang J. Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review // Mathematics. 2025. vol. 13(5). DOI: 10.3390/math13050856.
17. Yu H.Q., McQuade F. RAG-KG-IL: A Multi-Agent Hybrid Framework for Reducing Hallucinations and Enhancing LLM Reasoning through RAG and Incremental Knowledge Graph Learning Integration // arXiv preprint arXiv:2503.13514. 2025.
18. Zhou Y., Liu Y., Li X., Jin J., Qian H., Liu Z., Li C., Dou Z., Ho T.-Y., Yu P. Trustworthiness in Retrieval-Augmented Generation Systems: A Survey. arXiv preprint arXiv:2409.10102. 2024. DOI: 10.48550/arXiv.2409.10102.
19. Das S., Ge Y., Guo Y., Rajwal S., Hairston J., Powell J., Walker D., Peddireddy S., Lakamana S., Bozkurt S., Reyna M., Sameni R., Xiao Y., Kim S., Chandler R., Hernandez N., Mowery D., Wightman R., Love J., Spadaro A., Perrone J., Sarker A. Two-Layer Retrieval-Augmented Generation Framework for Low-Resource Medical Question Answering Using Reddit Data: Proof-of-Concept Study // Journal of Medical Internet Research. 2025. vol. 27. DOI: 10.2196/66220.
20. Gupta S., Ranjan R., Singh S.N. A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. arXiv preprint arXiv:2410.12837. 2024. DOI: 10.48550/arXiv.2410.12837.
21. Sun D., Lyu Y., Li J., Chen Y., Wang T., Kimura T., Abdelzaher T. SCRAG: Social Computing-Based Retrieval Augmented Generation for Community Response Forecasting in Social Media Environments. arXiv preprint arXiv:2504.16947v1. 2025. DOI: 10.48550/arXiv.2504.16947.
22. Wu X.-K., Chen M., Li W., Wang R., Lu L., Liu J., Hwang K., Hao Y., Pan Y., Meng Q., et al. LLMFine-Tuning: Concepts, Opportunities,and Challenges // Big Data Cogn.Comput. 2025. vol. 9. DOI: 10.3390/bdcc9040087.
23. Song Y., Lv C., Zhu K., et al. LoRA fine-tuning of Llama3 large model for intelligent fishery field // Discov Computing. 2025. vol. 28. DOI: 10.1007/s10791-025-09663-6.
24. Ramachandran A. Advancing Retrieval-Augmented Generation (RAG): Innovations, Challenges, and the Future of AI Reasoning. 2025.
25. Bsharat S.M., Myrzakhan A., Shen Z. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4 // arXiv preprint arXiv:2312.16171. 2023.
26. Hu Y., Lei Z., Zhang Z., Pan B., Ling C., Zhao L. GRAG: Graph Retrieval-Augmented Generation // Findings of the Association for Computational Linguistics: NAACL 2025. 2025. pp. 4145–4157. DOI: 10.18653/v1/2025.findings-naacl.232.

27. Campello R.J.G.B., Moulavi D., Sander J. Density-Based Clustering Based on Hierarchical Density Estimates // Advances in Knowledge Discovery and Data Mining (PAKDD 2013). Lecture Notes in Computer Science. 2013. vol. 7819. pp. 160–172. DOI: 10.1007/978-3-642-37456-2_14.
28. Федоров А.М., Датьев И.О., Вишняков И.Г. Модуль системы мониторинга социальных медиа: извлечение ключевых слов и генерация названий тем // Роспатент: Свидетельство о государственной регистрации программы для ЭВМ № 2025668928 от 21 июля 2025 г.

Федоров Андрей Михайлович — канд. техн. наук, ведущий научный сотрудник, заместитель директора по научной работе, ИИММ КНЦ РАН; доцент кафедры, кафедра информатики и вычислительной техники, филиал МАУ в г. Апатиты. Область научных интересов: разработка моделей и технологий информационной поддержки для регионального управления, мониторинг и анализ данных социальных медиа. Число научных публикаций — 60. a.fedorov@ksc.ru; улица Ферсмана, 24а, 184209, Апатиты, Мурманская область, Россия; р.т.: +7(81555)740-50.

Датьев Игорь Олегович — канд. техн. наук, старший научный сотрудник, ученый секретарь, ИИММ КНЦ РАН. Область научных интересов: разработка моделей и технологий для обработки больших данных, мониторинг и анализ данных социальных медиа. Число научных публикаций — 87. i.datyev@ksc.ru; улица Ферсмана, 24а, 184209, Апатиты, Мурманская область, Россия; р.т.: +7(81555)740-50.

Вишняков Иван Геннадьевич — аспирант, системный администратор, ИИММ КНЦ РАН. Область научных интересов: разработка информационных систем анализа больших открытых данных социальных медиа. Число научных публикаций — 6. i.vishnyakov@ksc.ru; улица Ферсмана, 24а, 184209, Апатиты, Мурманская область, Россия; р.т.: +7(81555)740-50.

Поддержка исследований. Исследование выполнено в рамках государственного задания ИИММ КНЦ РАН Министерства науки и высшего образования РФ, тема НИР: «Методы и технологии создания интеллектуальных информационных систем для поддержки развития сложных динамических систем с региональной спецификой в условиях неопределённости и риска» (шифр темы FMEZ-2025-0053).

DOI 10.15622/ia.24.6.4

A. FEDOROV, I. DATYEV, I. VISHNYAKOV

THE METHOD FOR INTEGRATING LARGE LANGUAGE MODELS INTO ALGORITHMS FOR FOCUSED MONITORING OF OPEN SOCIAL MEDIA DATA

Fedorov A., Datyev I., Vishnyakov I. **The Method for Integrating Large Language Models into Algorithms for Focused Monitoring of Open Social Media Data.**

Abstract. The relevance of the study is determined by the importance and complexity of performing rapid summarization of a vast array of user-generated content on social networks. It is proposed to reduce the complexity of the problem by using robotic algorithms and their automated intelligent focusing on specific platforms, data availability, and data volumes. The paper examines the ability of large language models (LLMs) to generate high-quality, coherent, and context-sensitive annotations (summaries) that are suitable for the dynamic nature of unstructured, noisy social network data. The features of the RAG LLM technology for summarizing social network publications are presented. The main disadvantage of language models is instability and the difficulty of tracking the results to confirm factual accuracy. The authors propose a hybrid method for summarizing social media posts over a given period of time. The method involves a complex and variable combination of classical methods for extracting data from their repositories, as well as the abstractive and generative capabilities of large language models. Large language models are used to vectorize the analyzed data. The application of clustering algorithms to the obtained vector representations made it possible to increase the stability and quality of the results. Within the RAG technology, the capabilities of large language models are expanded by means of intelligent search in the MongoDB database used to store the original data. The paper presents three pipelines, each of which is a variant of the method implementation, and has advantages and disadvantages in various application conditions. The metrics used to evaluate the pipelines are given, and a comparative analysis is performed. Overall, the method allows us to reduce the confabulations of a large language model and obtain annotations of publications for different time periods in real-time. The proposed method is used in practice in the open social media data monitoring system developed by the authors.

Keywords: social media, posts, text summarization, LLMs, RAG, AI agents, hybrid method.

References

1. Datyev I.O., Fedorov A.M., Reviakin A.A. [Focused collection and processing of open social media data]. Ontologija proektirovanija – Ontology of designing. 2024. vol. 14. no. 4(54). pp. 569–581. DOI: 10.18287/2223-9537-2024-14-4-569-581. (In Russ.).
2. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003. vol. 3. pp. 993–1022.
3. Fedorov A.M., Datyev I.O., Vishnyakov I.G. [Designing an information system for integrated topic analysis of social media big data]. Ontologija proektirovanija – Ontology of designing. 2024. vol. 14. no. 1(51). pp. 55–70. DOI: 10.18287/2223-9537-2024-14-1-55-70. (In Russ.).
4. Zhang Y., Jin H., Meng D., Wang J., Tan J. A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods. arXiv preprint arXiv:2403.02901. 2024.

5. Glickman M., Zhang Y. AI and Generative AI for Research Discovery and Summarization. Harvard Data Science Review. 2024. vol. 6(2). DOI: 10.1162/99608f92.7f9220ff.
6. Rose S., Engel D., Cramer N., Cowley W. Automatic Keyword Extraction from Individual Documents. Text mining: applications and theory. 2010. pp. 1–20. DOI: 10.1002/9780470689646.ch1.
7. Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. YAKE! Keyword extraction from single documents using multiple local features. Information Sciences. 2020. vol. 509. pp. 257–289. DOI: 10.1016/j.ins.2019.09.013.
8. Mihalcea R., Tarau P. TextRank: Bringing Order into Text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: Association for Computational Linguistics, 2004. pp. 404–411.
9. Tabanmehr Z., Akhtarkavan E. Automatic summarization of Instagram social network posts by combining semantic and statistical approaches. 6th International Conference on Pattern Recognition and Image Analysis (IPRIA). 2023. pp. 1–6. DOI: 10.1109/IPRIA59240.2023.10147186.
10. Ghali M.K., Farrag A., Lam S., Won, D. Beyondwords is All You Need: Agentic Generative Ai Based Social Media Themes Extractor. Available at: https://ssrn.com/abstract=5155101 (accessed 26.10.2025)
11. Wei J., Wang X., Schuurmans D., Bosma M., Ichter B., Xia F., Chi E., Le Q., Zhou D. Chain-of-thought prompting elicits reasoning in large language models. Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS'22). NY, USA: Curran Associates Inc., Red Hook, 2022. pp. 24824–24837.
12. Kojima T., Gu S.S., Reid M., Matsuo Y., Iwasawa Y. Large language models are zero-shot reasoners. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22). NY, USA: Curran Associates Inc., Red Hook, 2022. pp. 22199–22213.
13. Kambhamettu H., Flores J., Head A. Traceable Text: Deepening Reading of AI-Generated Summaries with Phrase-Level Provenance Links. arXiv preprint arXiv:2409.13099. 2024. DOI: 10.48550/arXiv.2409.13099.
14. Zhao S., Yang Y., Wang Z., He Z., Qiu L.K., Qiu L. Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely. arXiv preprint arXiv:2409.14924. 2024. DOI: 10.48550/arXiv.2409.14924.
15. Hidayaturrahman, Prawira I. Retrieval-Augmented Generation for Social Media Content Creation with Sentence Window and Auto-Merging Retrieval. International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS). 2024. pp. 608–613. DOI: 10.1109/ICIMCIS63449.2024.10956838.
16. Zhang W, Zhang J. Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review. Mathematics. 2025. vol. 13(5). DOI: 10.3390/math13050856.
17. Yu H.Q., McQuade F. RAG-KG-IL: A Multi-Agent Hybrid Framework for Reducing Hallucinations and Enhancing LLM Reasoning through RAG and Incremental Knowledge Graph Learning Integration. arXiv preprint arXiv:2503.13514. 2025.
18. Zhou Y., Liu Y., Li X., Jin J., Qian H., Liu Z., Li C., Dou Z., Ho T.-Y., Yu P. Trustworthiness in Retrieval-Augmented Generation Systems: A Survey. arXiv preprint arXiv:2409.10102. 2024. DOI: 10.48550/arXiv.2409.10102.
19. Das S., Ge Y., Guo Y., Rajwal S., Hairston J., Powell J., Walker D., Peddireddy S., Lakamana S., Bozkurt S., Reyna M., Sameni R., Xiao Y., Kim S., Chandler R., Hernandez N., Mowery D., Wightman R., Love J., Spadaro A., Perrone J., Sarker A. Two-Layer Retrieval-Augmented Generation Framework for Low-Resource Medical

Question Answering Using Reddit Data: Proof-of-Concept Study. Journal of Medical Internet Research. 2025. vol. 27. DOI: 10.2196/66220.
20. Gupta S., Ranjan R., Singh S.N. A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. arXiv preprint arXiv:2410.12837. 2024. DOI: 10.48550/arXiv.2410.12837.
21. Sun D., Lyu Y., Li J., Chen Y., Wang T., Kimura T., Abdelzaher T. SCRAG: Social Computing-Based Retrieval Augmented Generation for Community Response Forecasting in Social Media Environments. arXiv preprint arXiv:2504.16947v1. 2025. DOI: 10.48550/arXiv.2504.16947.
22. Wu X.-K., Chen M., Li W., Wang R., Lu L., Liu J., Hwang K., Hao Y., Pan Y., Meng Q., et al. LLMFine-Tuning: Concepts, Opportunities,and Challenges. Big Data Cogn.Comput. 2025. vol. 9. DOI: 10.3390/bdcc9040087.
23. Song Y., Lv C., Zhu K., et al. LoRA fine-tuning of Llama3 large model for intelligent fishery field. Discov Computing. 2025. vol. 28. DOI: 10.1007/s10791-025-09663-6.
24. Ramachandran A. Advancing Retrieval-Augmented Generation (RAG): Innovations, Challenges, and the Future of AI Reasoning. 2025.
25. Bsharat S.M., Myrzakhan A., Shen Z. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. arXiv preprint arXiv:2312.16171. 2023.
26. Hu Y., Lei Z., Zhang Z., Pan B., Ling C., Zhao L. GRAG: Graph Retrieval-Augmented Generation. Findings of the Association for Computational Linguistics: NAACL 2025. 2025. pp. 4145–4157. DOI: 10.18653/v1/2025.findings-naacl.232.
27. Campello R.J.G.B., Moulavi D., Sander J. Density-Based Clustering Based on Hierarchical Density Estimates. Advances in Knowledge Discovery and Data Mining (PAKDD 2013). Lecture Notes in Computer Science. 2013. vol. 7819. pp. 160–172. DOI: 10.1007/978-3-642-37456-2_14.
28. Fedorov A.M., Datyev I.O., Vishnyakov I.G. [Social media monitoring system module: extracting keywords and generating topic titles]. Rospatent: Certificate of state registration of a computer program No. 2025668928 dated July 21, 2025. (In Russ.).

Fedorov Andrey — Ph.D., Leading researcher, deputy director for research, IIMM KSC RAS; Associate professor of the department, Department of informatics and computer engineering, Apatity branch of MAU. Research interests: development of models and technologies for information support for regional management, monitoring and analysis of social media data. The number of publications — 60. a.fedorov@ksc.ru; 24a, Fersmana St., 184209, Apatity, Murmansk region, Russia; office phone: +7(81555)740-50.

Datyev Igor — Ph.D., Senior researcher, scientific secretary, IIMM KSC RAS. Research interests: development of models and technologies for big data processing, monitoring and analysis of social media data. The number of publications — 87. i.datyev@ksc.ru; 24a, Fersmana St., 184209, Apatity, Murmansk region, Russia; office phone: +7(81555)740-50.

Vishnyakov Ivan — Postgraduate student, system administrator, IIMM KSC RAS. Research interests: development of information systems for analyzing big open data in social media. The number of publications — 6. i.vishnyakov@ksc.ru; 24a, Fersmana St., 184209, Apatity, Murmansk region, Russia; office phone: +7(81555)740-50.

Acknowledgements. The work is supported by the Ministry of Science and Higher Education of the Russian Federation. Topic title: Methods and technologies for creating intelligent information systems to support the development of complex dynamic systems with regional specificity in conditions of uncertainty and risk (No. FMEZ-2025-0053).