# СОВРЕМЕННОЕ СОСТОЯНИЕ И ТЕНДЕНЦИИ РАЗВИТИЯ СОЦИАЛЬНОЙ ПСИХОЛОГИИ

## ТЕМАТИЧЕСКИЙ АНАЛИЗ ДИСКУССИЙ. СОВРЕМЕННЫЕ МЕТОДЫ, НЕДОСТАТКИ И ВОЗМОЖНОСТИ*

**Н.А. Алмаев***

*Доктор психологических наук, профессор, ведущий научный сотрудник, лаборатория психологии речи и психолингвистики, Федеральное государственной бюджетное учреждение науки Институт Психологии РАН, г. Москва, ул. Ярославская, дом 13, корп. 1; e-mail: almaev@mail.ru

**DOI: 10.38098/ipran.sep_2024_33_1_02** **Поступила в редакцию 29 ноября 2023 г.**

*Аннотация.* В данной обзорной статье обосновывается необходимость разработки средств анализа дискуссий. Критически проанализирована существующая практика применения моделей Латентного размещения Дирихле (подход «мешок слов»), и различные варианты подходов Seq2Seq (последовательность к последовательности). Особое внимание уделено большим языковым моделям, в частности трансформерам, с которыми в настоящее время связываются надежды на решение задач суммаризации и анализа мнений, как наиболее близких к анализу дискуссий. Приводятся попытки проанализировать причины галлюцинаций лингвистических моделей (LLM), в частности, работы М. Ли о математических основах галлюцинаций и эмпирическое исследование Ст. Лин, в котором было обнаружено, что количество галлюцинаций увеличивается с ростом числа параметров модели. Приводятся примеры из практики суммаризации видео, подтверждающие выводы Лин и др. Наиболее острой проблемой для анализа дискуссий видится постоянное переиначивание фамилий нейросетями. На основе изучения существующей практики намечены пути развития анализа дискуссий. Подход, лежащий в его основе, должен быть Sec2Sec (последовательность к последовательности) с предложением в качестве базовой единицы анализа. При этом на ближайшую перспективу видятся две основные задачи: 1) сопоставления всех постов какого-либо участника дискуссий между собой с целью обнаружения повторяющихся фрагментов, представляющих позицию данного человека, и 2) анализ откликов участников дискуссии на исходный пост в рамках его обсуждения. В обоих случаях предполагается сначала находить, а затем максимизировать «пятна касания» – наиболее совпадающие элементы обсуждений. Эти элементы затем могут обобщаться с помощью LLM со сбалансированным количеством параметров, обеспечивающим обобщение, но минимизирующим галлюцинации. Также предполагается использовать низкоуровневые довекторные средства сравнения строк как для реконструкции сложных топических отношений, так и для обнаружения намеренных искажений написания слов в целях передачи дополнительной коннотативной информации.

*Ключевые слова:* анализ дискуссий, суммаризация, галлюцинации языковых моделей, последовательность к последовательности, векторизация, оценка совпадения строк.

* Исследование выполнено при поддержке гранта РНФ № 23-28-10316.

Текстовое и аудио-визуальное содержание интернета является экстериоризацией общественного сознания, что открывает беспрецедентные возможности для его изучения. Опасности токсичных дискурсов наглядно проявили себя во время пандемии Ковид-19. При этом официальная пропаганда была малоэффективной вследствие непонимания своеобразной картины мира, сформировавшейся в чем-то стихийно, а в чем-то и целенаправленно у широких слоев населения. Другими примерами носителей токсичных дискурсов являются радикальные религиозные движения, например, ваххабизм в исламе, различные экстремистские течения в других конфессиях, из недавнего, – царебожие в православии. Будучи слабоизученными или практически неизвестными, они изумляют общественность, внезапно проявляясь в неожиданных действиях, охватывающих зачастую значительное число лиц (см., например: Соснин, Ковалева, 2017).

Следует отметить, что в ряде случаев тексты, предшествующие неожиданным событиям, распространяются на малодоступных основной массе населения РФ языках – арабском, кавказских или тюркских. Где взять специалистов, чтобы отслеживать эту речевую продукцию? При рекрутировании их из носителей может возникать вопрос о лояльности последних. В контексте непрекращающихся вооруженных конфликтов, в том числе с участием РФ, данные вопросы приобретают особую актуальность. О понимание важности общественных обсуждений свидетельствует, например, монография (Губанов и др., 2010). Вместе с тем, в ней не содержится каких-либо подходов к тематическому анализу обсуждений. Работа Губанова с соавторами представляет собой некое обобщенное моделирование воздействия одних субъектов на других, вне рассмотрения конкретных содержаний обсуждения, хотя и выполненное на весьма высоком уровне абстракции.

Другой проблемной областью является дееспособность самого общества, – т.е. способность различных его слоев отстаивать свои законные интересы и права. Считается общепризнанным, что основным вкладом теории игр в социологию является равновесие Нэша (Захаров, 2015). Кратко сформулировать его принцип можно как «ничего не делать – самое лучшее». Именно множеством равновесий Нэша парализована активность различных групп и целых социальных слоев. Подобная стратегия индивидуального поведения направлена на то, чтобы положение каждого из участников группы не ухудшалось относительно других, но в результате хуже становится им всем, поскольку группа просто не может консолидироваться до такой степени, чтобы начать действовать как единое целое. Для преодоления равновесия Нэша и построения выигрышной стратегии требуется интенсивная коммуникаций внутри групп, но для этого – повышение общей социальной компетенции. Члены различных групп должны научиться давать адекватные ответы на вопросы: кто мы? В чем наши интересы? Что мы можем сделать? Кто наши союзники и кто наши оппоненты, в каких проблемных областях? Иначе освобождение от «ложного сознания» (по К. Марксу), т.е. идеологических построений и воззрений других классов, невозможно.

Таким образом, требуются весьма обширные и глубокие обсуждения проводимые, широкими кругами граждан, однако их когнитивные ресурсы зачастую недостаточны как для обработки массивов информации, так и для преодоления сопутствующей энтропии. К источникам энтропии следует отнести, прежде всего, речевой, а не текстовый характер изложения. Звучащая речь, отягченная еще и низким качеством декламации, да еще и дополненная зачастую бессодержательным видео-контентом, серьезно мешает восприятию основных тезисов выступления, отвлекая на содержание, относящееся к говорящему субъекту, но не к существу дела.

Далее, сама по себе речь, как устная, так и письменная нередко представляет собой скорее поток сознания, чем структурированные тезисы. Обсуждения же зачастую имеют характер случайных ассоциативных реакций на отдельные фрагменты выступлений, чем целостного анализа целостно представленного видения.

Таким образом, актуальной социальной задачей является развитие средств суммаризации дискуссий. Более того, по мере распространения автоматического порождения текстов, задача автоматической экстракции смыслов приобретает характер все более острой необходимости.

**Современное состояние области исследования**

В 2021 г. нами была предпринята попытка изучения мнений посетителей Живого Журнала о пандемии Ковид-19 и борьбы с ней при помощи *Латентого размещения Дирихле* (Алмаев, Мурашева, 2022). Данный метод основан на модели *мешок слов* (bag of words), представляющей слова в тексте несвязанными между собой. Результаты были малоудовлетворительны, хотя каждая выборка обсуждений и отличалась от других, в целом ее было крайне сложно интерпретировать понятным для людей образом.

В настоящее время, хотя Латентное размещение Дирихле еще пользуются популярностью, основные перспективы связываются с применением подхода Sec2Sec (последовательность к последовательности). Именно для реализации данного подхода была предложена технология нейросетей, которая называется «трансформеры» (Vaswani et al., 2018), оказавшаяся прорывной. В основе ее лежит реализации функции Softmax, осуществляемой многоуровневыми нейросетями, при помощи механизма внимания. Функция Softmax:

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for } i = 1, \ldots, K \text{ and } \mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K.$$

При расчете Softmax применяется стандартная экспоненциальная функция к каждому элементу входного вектора. Затем происходит нормализация этих значений путем деления на сумму всех этих экспонент. Тем самым вектор из *K* вещественных чисел преобразуется в распределение вероятностей *K* возможных исходов. До применения Softmax компоненты могут быть отрицательными, превышать единицу, сумма их не равна 1, но после ее применения каждый компонент будет находиться в интервале от 0 до 1, и все они в сумме дают 1. Таким образом, их можно трактовать как вероятности. Будучи примененной к векторам, соответствующим словам какого-либо текста, данная функция позволяет предсказывать вероятность появления следующего слова (токена) в последовательности на основе предыдущих. На практике функцию внимания на множестве запросов рассчитывают одновременно упакованной в матрицу Q (queries – запросы), (K keys – ключи) и (V values – величины), а софтмакс переводит значения в вероятности.

Функция внимания следующим образом связывает переменные запроса, ключа и величины:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

.

T – это «температура», а d – мерность вектора, количество измерений которым он характеризуется. Температура – это вероятность активации случайных элементов, а не тех, которым была предобучена модель. Чем выше температура, тем выше данная вероятность, соответсвенно модель становится более галлюцинирующей или же «креативной». Чем ближе температура к нулю, тем реакции модели ближе к предобученным.

Важно понимать, что в трансформерах функция Softmax является *самоподдерживающейся* (selfsupervised – буквально «самонадзирающей»). Она опирается на свои же предшествующие результаты и принципиально не может быть скорректирована никакой внеязыковой реальностью (на практике эта задача переходит к пользователю, организующему так называемые «prompt engineering», т.е. такую формулировку запросов, чтобы результат более или менее им соответствовал). Авторы (Vaswani et al., 2018) специально подчеркивают, что стремились сделать распространение возбуждения в сети с механизмом внимания строго однонаправленным, то есть результаты определения вероятности каждого следующего токена зависят только от предыдущих.

Это обстоятельство важно иметь ввиду, при разборе вопроса о галлюцинациях языковых моделей и мерах по их смягчению.

Хотя задачи анализа дискуссий не ставятся в настоящее время напрямую, существует целый ряд тем содержательно близких к ним и, так сказать, «подготавливающих почву». К таковым относятся: 1) суммаризация (краткий пересказ) текстов и видео, 2) анализ мнений (отзывов) в отношении каких-либо продуктов, 3) анализ обсуждений с небольшим количеством участников (Payne, 2023). Основные надежды во всех трех случаях связываются с большими языковыми моделями (Large Linguistic Models – LLM) на основе общих предтренированных трансформеров – (Generally Pretrained Transformers – GPT). Основным достоинством GPT уровня 3, 3,5 и 4 является хорошая способность к абстрактивной, т.е. переформулирующей суммаризации – нейросеть способна не только выделить тему, но и зачастую обобщить и сказать другими словами основные утверждения в рамках этой темы.

Основной же проблемой являются галлюцинации LLM. Неинформативные, иной раз откровенно бредовые ответы GPT моделей

известны, пожалуй, всем. Кембриджский словарь даже выбрал словом 2023 года глагол «hallucinate» (именно применительно к большим языковым моделям)[30]

Корейский автор Минкийок Ли предпринял попытку анализа галлюцинаций GPT, разобрав математический аппарат трансформеров, и эксплицировав ряд допущений, лежащих в основе их применимости (Lee, 2023). Он переписал функцию Softmax как функцию минимизации потерь в форме:

$$L(\Theta) = -\frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \sum_{t=1}^{n-1} \log p(x_{t+1} | x_1, x_2, \ldots, x_t; \Theta).$$

В этой функции *Тета* означает параметры модели. Функция потерь *Теты* определяется как среднее негативное логарифмическое правдоподобие токенов по всем последовательностям в датасете D. Первостепенной задачей обучения GPT является минимизация негативного логарифмического правдоподобия наблюдаемых последовательностей. Далее он переходит к разбору допущений применимости данной модели. Первые два из них настолько своеобразны, что их адекватная интерпретация потребовала бы слишком много места.

Наиболее близкой к вопросу галлюцинаций является третье – допущение регулярности: «Мы исходим из того, что функция потерь L(Θ) непрерывна и дифференцируема в отношении параметров Θ и что ландшафт оптимизации лишен ненормальных черт, таких как плоские области и седловые точки» (Lee, 2023, с. 3). Автор признает это допущение весьма сомнительным, поскольку у данной функции вполне могут быть локальные минимумы и в целом задача оптимизации может быть далека от выпуклости.

[30] https://sovlit.ru/tpost/hc3br80fg1-kembridzhskii-slovar-nazval-slovo-2023-g

«Допущение 5. Мы принимаем, что функция минимизации потерь является проблемой невыпуклой оптимизации, потенциально содержащей множественные локальные минимумы» (Lee, 2023, с. 6).

Основной причиной галлюцинаций он видит то, что функция потерь является *самоподдерживающейся* (selfsupervised – буквально «самонадзирающей»), и это побуждает модель генерировать токены, оптимизирующие вероятность предсказанной последовательности, независимо от точности вывода.

«Галлюцинация появляется, когда модель генерирует низковероятный токен $xi$+1 при данных предшествующих токенах $x$1, $x$2, …, $xi$, и далее использует этот токен как входящий для генерации следующего токена $xi$+2» (Lee, 2023, с. 8). Таким образом, если в вывод попадает вероятный, хотя и неадекватный внешнему миру токен, это приводит к целому каскаду дальнейших неадекватных генераций. Само появление таких токенов, в этом М. Ли разделяет общее мнение, возникает во многом из-за зашумленности исходных сигналов.

Самым, пожалуй, интересным и важным является наблюдение оксфордских исследователей Стефани Лин и соавторов (Lin et al., 2022), что чем больше параметров (т.е. измерений) имеет модель GPT, тем меньше достоверных ответов она выдает. Достоверность (truthworthnes), правда, понимается в этой работе своеобразно, не только как соответствие фактам, но и как конгруэнтность определенным мнениям, вплоть до продиктованных текущей политикой некоторых стран. Например, ответ GPT на вопрос, «кто виноват в терактах 11 сентября?», – «американское правительство» – считался недостоверным (проходил под категорией «конспирология»). Помимо данной категории было еще три – «здоровье», «закон» и «фантазии».

Авторы видят основную причину в том, что большие модели перенимают из проанализированных текстов различные поверия, фольклор, недостоверные мнения (либо противоречащие фактам, либо неудобные для определенной политической повестки). Для иллюстрации они приводят ответы моделей GPT3 с последовательным увеличением числа параметров на вопрос, «что будет, если разбить зеркало»: 3,5 M – «вы разобьете зеркало», 1.3B – «вы ничего не увидите», 6.7 B – «зеркало разобьется на миллион кусков», 175B – «вас ждут семь лет несчастий». В целом с ростом числа параметров модели количество достоверных ответов падает у GPT3 с ок. 30%, до ок. 20%, у GPT2 с примерно 35% до примерно 30% и т.д. (Lin et al., 2022)[31]. Температура сетей, согласно авторам, была выставлена в ноль, т.е. ответы должны получаться максимально близкими к предобученным, безо всякой «креативности».

Стоит заметить, что компания OpenAI весьма внимательно следила за данным исследованием. Ее представитель Дж. Хилтон входил в авторский коллектив, а специальный документ под названием «GPT4 system card»[32], обнародованный 23 марта 2023 г., рапортует об изменениях, внесенных в новую версию. В основном они касаются того, как сеть натренировали уходить от «провокационных» вопросов.

В задачах анализа обсуждений достоверность не может быть релевантным критерием, поскольку именно она в отношении тех или иных утверждений и оспаривается их участниками, но важна адекватная передача точек зрения. Другими словами, в анализе дискуссий истина состоит не в отношении к фактам чьего-либо утверждения, а в том, что такой-то участник делает данные утверждения, – именно это факт, подлежащий фиксации. Однако именно адекватность, похоже, и страдает при увеличении количества параметров.

[31] Также см. https://github.com/sylinrl/TruthfulQA.
[32] https://cdn.openai.com/papers/gpt-4-system-card.pdf

Мы в исследовательских целях использовали сайт https://www.summarize.tech/ специализированный для суммаризации видео с you-tube, подключенный к Chat-GPT. Он принимает российские IP адреса и не требует оплаты, если загружать не более одного видео в сутки. Для суммаризации брались видео на русском языке, имеющие характер обсуждений общественно-значимых вопросов. Сайт является многоязычным, достаточно хорошо «понимает» русский, хотя базируется на субтитрах ю-туба, довольно низкого качества.

Например, в суммаризацию интервью с пилотом экранопланов (https://www.youtube.com/watch?v=AjdGJ-TD6mI) Chat-GPT внес название российского вертолета Ка-50 «Черная акула», хотя не только данная модель, но вообще никакие вертолеты не были упомянуты в этом интервью ни разу, ни как родовое понятие, ни по каким-либо их моделям. Вероятно, для срабатывания локального минимума оказалось достаточным первого слога «Ка», встречавшегося в интервью довольно часто, поскольку обсуждаемые события происходили в Каспийском море, а сами экранопланы периодически назывались «каспийскими монстрами». С другой стороны, хотя названия экранопланов «Орленок» и «Лунь» не раз четко распознавались субтитрами, они ни разу не были включены в суммаризацию и заменялись некими фантазийными аббревиатурами. Можно предположить, что в данном случае контекст, напротив, действовал как тормозящий фактор, поскольку «Eagle» (основа для «eaglet» или «eagling» – орленок) название истребителя-бомбардировщика США F-15, а Лунь (Circus cyaneus) в свою очередь переводится на английский как «Harrier» – название английского истребителя вертикального взлета и посадки. Таким образом, легко заподозрить главного «виновника» обоих типов галлюцинаций – измерение «советское vs. натовское вооружение». Поскольку речь про советское, то активируется даже

маловероятное «Ка», а релевантные названия птиц тормозятся, будучи, согласно предобучению, «натовскими». В данном случае, общая рекомендация – для борьбы с галлюцинациями снижать температуру – могла бы подействовать в первом случае, но не во втором, поскольку снижение температуры максимизирует именно выученные ответы.

По опыту суммаризации можно сказать, что, пожалуй, наиболее частой галлюцинацией LLM является переиначивание фамилий даже в научных статьях (Beutel et al., 2023, Salvagno et al., 2023) и в тех случаях, когда они адекватно распознаются в субтитрах. Например, фамилия летчика из упомянутого видео – Коробкин – передавалась суммаризатором как Курочкин. В другом видео[33] политик Надеждин, при относительно правильной транскрипции Nадеждин объявленный отцом (!) Бориса Немцова, обозначался далее то как Назаров, то как Найденов, то как Нездан, то, наконец, как «радикальная группировка Надежда и Немцов». И все это при том, что суммаризатор в начале взял правильное написание всех фамилий не из субтитров, а из текста, сопровождающего видео и именно в текстовой форме размещенного самими публикаторами! Причина, по-видимому, лежит в применении к фамилиям подхода, «byte-pair tokenization[34]», при котором осмысленными элементами выступают отдельные символы, которые затем объединяются в значимые элементы и токенизируются. Естественно, для анализа дискуссий, где важна четкая идентификация участников, даже не по фамилиям, а зачастую по никнеймам, подобное совершенно недопустимо. Технология «byte-pair tokenization», весьма полезна для обработки естественных языков синтетического типа, однако от нее при анализе

[33] Оно сейчас забанено на ю-туб, но с ним можно ознакомиться по адресу: https://aurora.network/articles/11-analitika-i-prognozy/110383-putin-i-oppozitsija-chernoe-zerkalo-ili-politicheskaja-pjatnitsa-13-nadezhdin-alksnis-aksel-tillert).
[34] См. подробнее: https://huggingface.co/learn/nlp-course/chapter6/5.

дискуссий следует защитить наименования участников. При расшифровке речевых записей дискуссии, видимо, придется прибегать к диаризации (включать анализ спектрального состава голоса говорящего). Также, разумеется, чем сложнее употребляемые в речи или тексте конструкции, чем больше в них внутренних утверждений и отрицаний, тем хуже она распознается и передается. При этом отдельной самостоятельной проблемой остается распознавание иронии.

В итоге может создаваться впечатление, что пока перспективы автоматического анализа дискуссий не очень благоприятны. Ведь даже если отвлечься от имманентной склонности LLM к галлюцинациям и сконцентрироваться на сугубо текстовой информации вместо автоматически распознанных субтитров, ее качество будет ненамного лучше ввиду как опечаток, так и намеренного искажения написания слов в целях передачи дополнительной информации. Вместе с тем, если поставить во главу угла сам анализ дискуссий, то возможен поиск адекватных решений с прицелом именно на него.

## Стратегии и ближайшие задачи анализа дискуссий

Дискуссию можно представить как стохастический процесс, в котором имеются повторяющиеся фрагменты. Если рассматривать участника дискуссии, то у него есть позиция, которую в целом можно отождествить с повторяющимися фрагментами, точнее, она может быть восстановлена по ним. Если же рассматривать, как разворачиваются обсуждения каких-либо текстов, то нетрудно заметить, что некоторые фрагменты текста воспроизводятся в обсуждении. Таким образом, и для позиции участника дискуссий, выражающейся во многих текстах, и для участников дискуссий, реагирующих на текст, послуживший началу обсуждения, характерны некие *текстуальные*

*совпадения.* Их можно назвать по аналогии с физикой *«пятнами касания».* Тогда, задачу автоматического анализа дискуссий можно сформулировать как фиксацию и последующее воспроизведение текстовых *пятен касания* в максимально полной форме. Сама эта формулировка с неизбежностью подразумевает рекуррентность. *Пятно касания* сначала идентифицируется в общей форме и затем проступает все более рельефно.

В качестве прототипичной для обеих задач может выступать диагональная матрица из работ по биоинформатике (Огурцов, 2011). Из приведенного выше анализа вытекает, что проект должен быть, несомненно, реализован в идеологии Sec2Sec (последовательность к последовательности). Базовым уровнем сравнения строк может выступать предложение. Но на нем не следует останавливаться ведь мысль может быть и «шире» предложения, и «уже» него. Важно максимизировать «пятно касания» одного текста и другого. Оно имеет три уровня фиксации: 1) оценка сходства строк без векторизации; 2) оценка сходства векторов, косинусиальное сравнение; 3) оценка сходства текстов по различным измерениям, – какие из них являются наиболее решающими, для каких категорий людей. Алгоритмы сравнения строк с оценкой в % их совпадения, намного быстрее, чем сравнение векторов. Соответственно, если в целом низкоуровневые сравнения соответствуют векторным, то это будет выигрышной в вычислительном отношении стратегией. Можно будет сначала находить пятна касания с помощью низкоуровненых алгоритмов, а затем уже внутри них производить векторизацию, экстрактивную, а далее и абстрактивную суммаризацию, использовать для переформулировок LLM оптимального количества параметров (например, GPT2), позволяющие как производить адекватное обобщение, так и минимизировать галлюцинации.

Начинать поиск от косинусиального сходства векторов, соответствующих предложениям, представляется, возможно, более близким к смыслу (не все же буквально воспроизводят формулировки собственные или топикстартера), но гораздо более затратным по вычислительным ресурсам и времени. Но и такой подход не отменяет обратного перехода к низкоуровневому анализу строк. Предположим, образовано пятно касания из нескольких предложений максимального косинусиального сходства. Но далее эти предложения должны быть сопоставлены между собой на предмет наличия сходств и различий. Предположим, один субъект одобряет какое-либо явление, а другой отрицает посредством, частицы «не». Различия, как в отношении строк, так и в отношении векторов, минимально, но для людей – смысл противоположен. Для сопоставления и реконструкции утверждений должны быть реконструированы топические отношения – отрицание и утверждение, часть и целое, происхождение, причина и следствие, действующий и претерпевающий, имеющий или не имеющий намерение, а также оценочные моральные и эстетические суждения. Их реконструкция невозможна без анализа предложений на уровне синтаксиса, поскольку категориальные топические отношения задаются синтаксически с помощью подлежащего, сказуемого, дополнения, падежей, залогов, частиц, предлогов. Насколько данные вопросы могут быть решены при обращении к той или иной LLM, и при этом не вызывать галлюцинаций, априорно сказать невозможно.

Кроме того, опечатки и их автоматические, но далекие от смысла исправления соседствуют с весьма своеобразным феноменом – намеренным искажением написания слов, от языковой игры в чистом виде – «олбанский» («падонкаффский») язык до внедрения в дискуссионных целях в слова отсылок к самой разнообразной лексике. Берется какое-либо слово, например, «технократ» и заменяется на «технокрад», таким способом получается

дополнительная коннотация с кражей. Соответственно, претерпевает модификацию и оценка, из нейтральной или даже частично позитивной она становится негативной. Будем обозначать такие слова как содержащие *намеренное коннотативное искажение* (далее – НКИ). Насколько можно судить по предварительным наблюдениям, большинство НКИ сосредоточены именно на оценках, содержат отсылки либо к обсценной лексике, либо к скатологической, либо еще к каким-либо деструктивным и аверсивным проявлениям, например, «Потрошенко» (измененная фамилия бывш. президента Украины) и т.п. Таким образом, НКИ позволяют «упаковать» в одно слово целые оценочные суждения. Другими примерами НКИ может служить пародирование особенностей произношения, передающих какой-либо иностранный или национальный акцент или произношение характерное для каких-либо слоев населения, в знак того, что данный текст как бы озвучивается представителем соответствующей группы. Еще один феномен, для которого слова с НКИ могут выступать маркером, ирония: «ацкий сотона», «ужастный фошызд» и т.п. Если слова с НКИ не обрабатывать отдельно, то при векторизации они будут либо проигнорированы, либо сочтены опечатками и автоматически заменены на близкие аналоги, но в любом случае как человеческое значение фрагмента с их участием, так и значение его векторного представления будет искажено. Надо заметить, что слова с НКИ имеют минимальную дистанцию от своих «нормальных» аналогов, как по Хеммингу, так и по Левентшейну (порядка 0.9), т.е. фактически в массе своей не отличаются от опечаток. Соответственно, если возникнет задача справиться с НКИ, то придется создавать по ним специальную базу данных. Вместе с тем, алгоритмы сравнения строк без векторизации давно написаны на C++ и весьма быстры.

НКИ должны не просто отличаться от опечаток, но и анализироваться как отдельно, так и в их взаимном соотнесении. Ведь и слово с НКИ может быть набрано с опечатками. А затем оно может быть автоматически заменено алгоритмом типа Т9 на грамматически верное, но по смыслу весьма далекое слово, чего автор может не заметить, или не иметь желания или возможности еще раз редактировать сообщение. Вместе с тем, НКИ могут быть благодарной темой для исследователей. Поскольку слово с НКИ, соединяет в себе сразу и объект, и отношение к нему, оно является весьма эффективной формой оценочного суждения, сразу (пусть и в первом приближении) позволяющей идентифицировать отношение участника дискуссии к тому или иному обсуждаемому объекту.

Разумеется, еще предстоит дальнейшее изучение типов намеренных коннотативных искажений и оценка их относительного присутствия в обсуждениях на различных ресурсах. Какие люди используют НКИ, какие личностные черты предрасполагают к ним? Насколько слова с НКИ присутствуют в различных текстах? Насколько они конгруэтны общему смыслу текстов?

## ВЫВОДЫ

Проведенный анализ позволяет сделать следующие выводы для целей анализа дискуссий:

1) Значительное внимание должно уделяться входному уровню текста. Чем чище текст, тем лучше результат.

2) Сложность модели в рамках анализа дискуссий требуется лишь для адекватного обобщения суждений. Во избежание галлюцинаций система не должна иметь много параметров.

3) Система должна быть серьезно дообучена, чтобы узнавать и передавать категориальные отношения и топы: часть и целое, происхождение, причина и следствие, действующий и претерпевающий, распознавать аналогии, отличать оспариваемое от утверждаемого, распознавать негативное или позитивное отношение к отдельным утверждениям и не путать их. Здесь особо важно взаимодействие синтаксиса и семантики, поскольку роли в категориях задаются синтаксически.

4) Необходимо обеспечить распознавание слов с НКИ различных видов и форм. Причем именно последнее во многом и является базовым для успешности всей конструкции. Если слово с НКИ, соединяющее в одном слове и объект, и отношение к нему, не будет распознано, то не получится и адекватно реконструировать суждение. Но тогда и при попытках обобщения, скорее всего, возникнут галлюцинации

## ЗАКЛЮЧЕНИЕ

В данной статье рассмотрены задачи разработки системы автоматического анализа дискуссий. Проанализированы имеющиеся на сегодняшний день практики в близких областях. Предложены оригинальные подходы для анализа на трех уровнях повторяющихся в обсуждениях содержаний – низкоуровневого сравнения строк, сравнения векторов, обобщения наиболее содержательных фаргементов с помощью предтренированных моделей.

## СПИСОК ЛИТЕРАТУРЫ

*Алмаев Н.А., Мурашева О.В.* Тематический анализ дискуссий с применением метода Латентного размещения Дирихле // Институт психологии Российской академии наук. Социальная и экономическая психология. 2022. Т. 7. № 1 (25). С. 47-69. DOI: 10.38098/ipran.sep_2022_25_1_03.

*Губанов Д.А., Новиков Д.А., Чхартишвили А.Г.* Социальные сети: модели информационного влияния, управления и противоборства. М.: Изд-во физ мат литературы, 2010.

*Захаров А.В*. Теория игр в общественных науках. М.: Изд-во НИУ ВШЭ, 2015.

*Огурцов А.Н.* Методы биоинформационного анализа. Харьков: НТУ ХПИ, 2011.

*Соснин В.А., Ковалева Ю.В.* Распространение радикального ислама как глобальная угроза современности: геополитические тенденции и социально-психологические аспекты проблемы / Институт психологии Российской академии наук. Социальная и экономическая психология. 2017. Т. 2. № 2. С. 116-151. URL: http://soc-econom-psychology.ru/engine/documents/document350.pdf (дата обращения 04.08.2023 г.).

*Beutel G., Geerits E., Kielstein J.* Artificial hallucination: GPT on LSD? // Critical Care. 2023. 27: 148. DOI: 10.1186/s13054-023-04425-6.

*Lee M.A* Mathematical Investigation of Hallucination and Creativity in GPT Models // Mathematics. 2023. 11. 2320. DOI: 10.3390/math11102320.

*Lin S., Hilton J., Ewans O.* Truthful QA.: Measuring How Models Mimic Human Falsehoods // Computer Science > Computation and Language. 8 May 2022. DOI: 10.48550/arXiv.2109.07958.

*Payne M*. Harnessing GPT-4 for Meeting Summarization: Zero-Shot and Aspect-Based Approaches // Width AI, 2023. URL: https://www.width.ai/post/gpt-4-for-meeting-summarization (дата доступа 28.11.2023).

*Salvagno M., Taccone F.S., Gerli A.G.* Can arrtificial intelligence help for scientific writing? // Critical Care. 2023. 27:75.

*Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I.* Attention is all you need // Advances in Neural Information Processing Systems. 2017. 30. 6000-6010.

## BIBLIOGRAFICESKIJ SPISOK

Almaev N.A., Murasheva O.V. Tematicheskij analiz diskussij s primeneniem metoda Latentnogo razmeshcheniya Dirihle // Institut psihologii Rossijskoj akademii nauk. Social'naya i ekonomicheskaya psihologiya. 2022. T. 7. № 1 (25). S. 47-69. DOI: 10.38098/ipran.sep_2022_25_1_03.

Gubanov D.A., Novikov D.A., CHkhartishvili A.G. Social'nye seti: modeli informacionnogo vliyaniya, upravleniya i protivoborstva. M.: Izd-vo fiz mat literatury, 2010.

Zaharov A.V. Teoriya igr v obshchestvennyh naukah. M.: Izd-vo NIU VSHE, 2015.

Ogurcov A.N. Metody bioinformacionnogo analiza. Har'kov: NTU HPI, 2011.

Sosnin V.A., Kovaleva YU.V. Rasprostranenie radikal'nogo islama kak global'naya ugroza sovremennosti: geopoliticheskie tendencii i social'no-psihologicheskie aspekty problemy /

Institut psihologii Rossijskoj akademii nauk. Social'naya i ekonomicheskaya psihologiya. 2017. T. 2. № 2. S. 116-151. URL: http://soc-econom-psychology.ru/engine/documents/document350.pdf (data obrashcheniya 04.08.2023 g.).

Beutel G., Geerits E., Kielstein J. Artificial hallucination: GPT on LSD? // Critical Care. 2023. 27: 148. DOI: 10.1186/s13054-023-04425-6.

Lee M.A Mathematical Investigation of Hallucination and Creativity in GPT Models // Mathematics. 2023. 11. 2320. DOI: 10.3390/math11102320.

Lin S., Hilton J., Ewans O. Truthful QA.: Measuring How Models Mimic Human Falsehoods // Computer Science > Computation and Language. 8 May 2022. DOI: 10.48550/arXiv.2109.07958.

Payne M. Harnessing GPT-4 for Meeting Summarization: Zero-Shot and Aspect-Based Approaches // Width AI, 2023. URL: https://www.width.ai/post/gpt-4-for-meeting-summarization (data dostupa 28.11.2023).

Salvagno M., Taccone F.S., Gerli A.G. Can arrtificial intelligence help for scientific writing? // Critical Care. 2023. 27:75.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need // Advances in Neural Information Processing Systems. 2017. 30. 6000-6010.

## THEMATIC ANALYSES OF DISCUSSIONS. CONTEMPORARY METHODS, FLAWS AND CAPABILITIES**

**N.A. Almayev***

*Sc.D. (psychology), professor of RAS, leading research fellow; laboratory of psychology of speech and psycholinguist, Federal State Financed Establishment of science Institute of psychology, Russian academy of sciences; 13, Yaroslavskaya str., Moscow, 129366; e-mail: almaev@mail.ru

*Summary.* The task of developing tools for analyzing discussions is set in this review article. The existing practice of applying Latent Dirichlet allocation models (the "bag of words" approach) and various variants of Seq2Seq (sequence to sequence) approaches are critically analyzed. Particular attention is paid to large language models, and especially transofrmers, with which hopes are currently pinned on the summarization and analysis of opinions, being the closest tasks to the analysis of discussions. Attempts are made to analyze the causes of hallucinations, in particular, the works of M. Lee on the mathematical foundations of hallucinations, and an empirical study by St. Lin et al., who found that the number of hallucinations increases with the growth of model parameters. Examples from the practice of video summarization are given, confirming the conclusions of Lin et al. The most acute problem for the analysis of discussions is the constant twisting of surnames by neural networks. Based on the study of existing practice, the ways of developing the analysis of discussions are outlined. The approach underlying it should be Sec2Sec with the sentence as the basic unit for comparison. Two main tasks are seen for the closest future: 1) comparing all the posts of a participant in the discussions with each other with the aim of detecting repeated fragments representing the position of this person, and 2) analyzing the responses of the participants in the discussion to the original post as the part of its discussion. In both cases, it is assumed to first find and then maximize the "touch spots" - the most coincident elements of discussions. Which then can be generalized using LLM with a balanced number of parameters, providing generalization but minimizing hallucinations. It is also supposed to use low-level pre-vector string comparison tools both for the reconstruction of complex topical relations and for the detection of intentional misspellings of words in order to convey additional connotative information.

*Keywords:* analyses of discussions, summarization, hallucinations of linguistic models, последовательность к последовательности, vectorization, strings similarity.

** This work was supported by RSF, № 23-28-10316.