DOI 10.58351/2949-2041.2025.29.12.019

Шевкунова Татьяна Петровна, соискатель
ФГАОУ ВО «Российский университет
дружбы народов им. Патриса Лумумбы»
Shevkunova Tatiana, RUDN University

ТОКСИЧНОСТЬ МАССОВОЙ КОММУНИКАЦИИ: СОЗДАНИЕ ОБУЧАЮЩИХ ДАТАСЕТОВ ДЛЯ ИИ-СИСТЕМ НА ОСНОВЕ ЛИНГВИСТИЧЕСКИХ МАРКЕРОВ ОГРАНИЧИВАЮЩИХ УБЕЖДЕНИЙ
TOXICITY OF MASS COMMUNICATION: CREATING TRAINING DATASETS FOR AI SYSTEMS BASED ON LINGUISTIC MARKERS OF LIMITING BELIEFS

Аннотация. В статье представлена методология формирования обучающих датасетов для ИИ-систем, основанная на выявлении лингвистических маркеров ограничивающих убеждений в текстах.

Разработанная система позволяет эффективно анализировать и классифицировать токсичный контент с учётом контекстуальной специфики и культурных особенностей.

Abstract. The article introduces an innovative methodology for creating training datasets for artificial intelligence systems, which is based on identifying linguistic markers of limiting beliefs in texts.

The developed system enables effective analysis and classification of toxic content while taking into account contextual specifics and cultural characteristics.

Ключевые слова: Токсичный контент, обучающий датасет, лингвистические маркеры, ограничивающие убеждения, токсичный контент, контекстуальный анализ.

Keywords: Toxic content, training dataset, linguistic markers, limiting beliefs, toxic content, contextual analysis.

В современных условиях стремительного роста пользовательского контента традиционные методы ручной модерации уже не справляются с объемом информации, требующим проверки, что создает серьезную проблему для поддержания безопасной цифровой среды.

Экспоненциальный рост количества публикуемых сообщений делает невозможным их полный просмотр модераторами, поэтому возникает острая необходимость в разработке интеллектуальных систем фильтрации, способных оперативно обрабатывать большие массивы данных.

Современные алгоритмы должны уметь распознавать не только явные формы агрессии, но и более тонкие, скрытые проявления токсичности в тексте, включая сарказм, манипулятивные конструкции и завуалированные оскорбления, которые часто остаются незамеченными при поверхностном анализе.

Решение этой задачи требует создания комплексных систем искусственного интеллекта, способных анализировать контекст сообщений, выявлять скрытые паттерны токсичного поведения и оценивать потенциальную опасность контента для пользователей социальных платформ.

Существенной проблемой современного развития систем искусственного интеллекта является отсутствие стандартизированных наборов данных (датасетов), необходимых для эффективного обучения алгоритмов распознаванию ограничивающих убеждений как особой формы детерминирующей лингвистическую токсичность коммуникации.

Это затрудняет создание надёжных моделей машинного обучения, способных точно идентифицировать подобные паттерны в пользовательском контенте, поскольку без качественных обучающих материалов невозможно обеспечить корректное понимание и

классификацию различных проявлений ограничивающих убеждений в текстах, которые являются внутренней причиной появления негативно окрашенных комментариев к публикациям в социальных сетях.

В результате разработчики сталкиваются с серьёзными ограничениями в создании эффективных инструментов модерации, способных не только обнаруживать явные формы токсичности, но и распознавать более тонкие, скрытые проявления ограничивающих убеждений в пользовательских сообщениях.

Цель данной работы: разработать методологию формирования обучающих датасетов на основе лингвистических маркеров ограничивающих убеждений, детерминирующих токсичность массовой коммуникации. Для достижения поставленной цели необходимо решить следующие задачи:

1. выделить ключевые лингвистические маркеры ограничивающих убеждений;
2. определить принципы отбора и разметки текстов;
3. предложить структуру аннотации для ИИ-обучения.

Материалом для исследования стали тексты, расположенные в Национальном корпусе русского языка [9]. Исследование проводилось при помощи методов интерпретативного и функционального анализа, синтеза, наблюдения, описания, корпусного метода.

Осуществлён комплексный подход к интеграции лингвистического анализа токсичности в задачи обработки естественного языка (NLP). Это позволяет создать принципиально новую методологию выявления токсичного контента, основанную на систематизации лингвистических маркеров ограничивающих убеждений и их автоматической обработке. Что позволит применить разработанные алгоритмы в реальных системах модерации контента. Интеграция лингвистического анализа с технологиями NLP открывает перспективы для создания более эффективных инструментов автоматического выявления токсичных сообщений, что особенно важно в условиях постоянного роста объёма пользовательского контента в социальных сетях и мессенджерах.

Разработанный подход позволяет не только обнаруживать явные формы токсичности, но и распознавать скрытые проявления ограничивающих убеждений через специфические языковые маркеры, что существенно повышает качество автоматической модерации и делает её более целенаправленной.

Ограничивающие убеждения определяются в психологии как «убеждения, которые включают в себя негативный ресурс и ограничивают личность по причине неадекватной самооценки и неадекватной оценки происходящего» [6]. Они формируются под влиянием социального окружения, воспитания, культуры и личного опыта, часто начиная с детства.

Ограничивающие убеждения в контексте общественного сознания представляют собой глубоко укоренившиеся представления о мире, обществе и социальных отношениях, которые формируют искаженное восприятие реальности и препятствуют конструктивному взаимодействию между людьми. В отличие от индивидуальных ограничивающих убеждений, социальные ограничивающие убеждения имеют более масштабное влияние, поскольку формируют коллективные представления и могут распространяться через медиаканалы, усиливая негативное восприятие реальности широкими группами населения.

В лингвистике ограничивающие убеждения исследуются через языковые средства их конструирования. В современном медиапространстве такие убеждения проявляются через специфические языковые конструкции, отражающие негативное отношение не только к себе, но и к социальным группам, институтам власти, экономическим и политическим процессам [10]. Они могут выражаться в форме категоричных обобщений о социальной несправедливости, пессимистических прогнозов развития общества, отрицания возможности позитивных изменений.

Лингвистические маркеры ограничивающих убеждений в токсичном общественном дискурсе включают лексические, синтаксические и прагматические маркеры токсичности.

Лексические маркеры:

- абсолютные обобщения о социальных явлениях (определительные местоимения и наречия «все», «всегда», «постоянно», «любой» и др.);
- отрицательные местоимения и наречия («никто», «никогда», «нигде» и др.);
- негативные прогнозы, в том числе общественного развития;
- отрицание возможностей изменений;
- категоричные оценки как отдельных личностей, так и социальных институтов;
- модальные слова («должен», «невозможно», «недопустимо» и др.).
- использование гиперболы и катастрофизация;
- упрощенные дихотомии “свой-чужой” [8, 10].

Синтаксические маркеры:

- императивы, выражающие требование, приказ или угрозу;
- риторические вопросы с негативной окраской, используемые для провокации или унижения оппонента;
- перегруженные синтаксические конструкции, затрудняющие понимание и усиливающие давление;

Прагматические маркеры:

- манипулятивные тактики (газлайтинг (отрицание реальности), виктимблейминг (обвинение жертвы), социальное подтверждение («все так делают», «это знают все», «никто не сомневается» и т.д.), манипулятивные вопросы и др.;
- обесценивание (понижение чувств, мнений или достижений);
- неправомерные обобщения, когда распространяется частный случай на всю реальность («все в офисе тебя ненавидят, в тобой никто не хочет работать»);

Гибридные и дополнительные маркеры:

- сарказм – как самоподтверждение правильности своих убеждений и способ восприятия реальности через призму негатива;
- пассивная агрессия – выражение негативных чувств в завуалированной форме;
- эмоционально окрашенная лексика (ненависть, ужас, катастрофа).

В лингвистике эти убеждения изучаются через лексические и синтаксические средства их выражения, тогда как в психологии акцент делается на их функции и влияние на поведение.

Функции ограничивающих убеждений:

1. информационный фильтр – отбор информации, соответствующей имеющимся убеждениям.
2. демотивация – предоставление аргументов для бездействия или действий в рамках существующих установок.
3. формирование личности – влияние на характер, систему ценностей и поведение.
4. определение выбора – воздействие на решения в различных жизненных ситуациях [7].

В отличие от явной агрессии, которая направлена на прямое оскорбление оппонентов, и троллинга как формы провокационного поведения, ограничивающие убеждения формируют целостную картину мира, основанную на негативных установках и препятствующую конструктивному диалогу. Они создают барьеры в общественном дискурсе, затрудняют поиск компромиссов и препятствуют социальному развитию.

Мы можем выделить следующие отличия между ограничивающими убеждениями, явной агрессией и троллингом.

Таблица 1

Отличия между ограничивающими убеждениями, явной агрессией и троллингом.

Критерий	Ограничивающие убеждения	Явная агрессия	Троллинг
Цель	Внутренние ограничения, саботаж. Формирование негативного восприятия реальности, создание барьеров для социальных изменений.	Нанесение вреда, унижение. Демонстрация доминирования, получение эмоциональной разрядки.	Получение удовольствия от реакции жертвы и реакции других участников коммуникации, провокация.
Форма проявления	Скрытые, часто не осознанные установки.	Открытые оскорбления, угрозы.	Ироничные, провокационные комментарии, маскирующиеся под «невинный юмор».
Мотив	Результат прошлого опыта восприятия.	Эмоциональная разрядка, доминирование	Эгоцентричные гедонис-тические устремления, желание развлечься.
Механизм воздействия	Создаёт искаженную картину мира через обобщения и негативные прогнозы, влияет на массовое сознание.	Нацелен на немедленное причинение эмоционального вреда оппоненту.	Использует манипулятивные техники для дестабилизации коммуникации.
Сущность явления	Система негативных убеждений о мире и обществе, формирующая пессимистичный взгляд на социальные процессы.	Открытое выражение враждебности, направленное на конкретных участ-ников коммуникации.	Осознанная провокация с целью вызвать эмоциональную реакцию у других участников общения.
Влияние на коммуникацию	Мешают конструктивному диалогу, создают барьеры. Формируют негативный социальный дискурс через категоричные обобщения и пессимистические прогнозы.	Разрушают коммуникацию, вызывают конфликт. Выражается в прямых оскорблениях, угрозах и вербальной агрессии.	Блокируют диалог. Превращают общение в бессмысленную дискуссию. Проявляется в форме провокационных сообщений и ироничных комментариев.
Социальный эффект	Способствует поляризации общества, усиливает социальные конфликты, снижает доверие к инструментам власти, формирует атмосферу всеобщего пессимизма.	Разрушает конструктивный диалог, создает токсичную атмосферу в сообществе.	Дестабилизирует коммуникацию, отвлекает от конструктивного обсуждения.
Влияние на общество	Замедляет социальные изменения, препятствует позитивным трансформациям.	Разрушает механизмы конструктивного взаимодействия.	Подрывает доверие к онлайн-коммуникации, формирует атмосферу враждебности.

Современные системы обработки естественного языка (NLP) используют для автоматической детекции токсичности. Алгоритмы анализируют частоту употребления определенных конструкций, сочетают их с другими признаками (например, эмоциональная окраска, контекст) и классифицируют текст как токсичный или нет.

Таким образом, лингвистические маркеры токсичности помогают выявить не только явную агрессию, но и скрытые формы деструктивного общения. Их анализ важен как для понимания психологии взаимодействия, так и для разработки инструментов модерации контента.

Детекция токсичности в текстах является активно развивающейся областью исследований, и создано множество открытых датасетов, позволяющих обучать и тестировать соответствующие модели. Рассмотрим наиболее известные из них:

1. Toxic Comment Classification Challenge (Kaggle)

Этот датасет стал результатом конкурса Kaggle в 2017 году, организованного совместно с компаниями Google Jigsaw и Civil Comments. Задача состояла в обучении моделей для классификации комментариев по пяти категориям токсичности: toxic, severe_toxic, obscene, threat, insult и identity_hate. Набор данных включает почти 300 тыс. аннотированных комментариев, собранных с сайта Civil Comments. Отличается разнообразием токсичных выражений и простотой формата [5].

2. Wikipedia Detox Dataset

Собран сотрудниками Google Jigsaw на основе архивов дискуссионных страниц Wikipedia. Включает два набора данных: один с комментариями, помеченными вручную добровольцами, второй – автоматически помеченный системой автоматического обнаружения токсичности. Многоязычный датасет. Данные состоят примерно из 1 млн. аннотированных комментариев [1].

3. ru_paradetox_toxicity

Важной особенностью RuParatext является его ориентация на русскоязычные тексты, что отличает его от большинства международных датасетов, разрабатываемых преимущественно на английском языке. Возможность обучения и тестирования моделей именно на русскоязычных данных делает RuParatext уникальным и востребованным ресурсом для специалистов, работающих с русским языком. Собран через Yandex.Toloka и имеет объем около 6,5 тысячи данных.

4. Multilingual Toxicity Dataset (Hugging Face)

Крупный многоязычный датасет, разработанный для задач детекции токсичности в текстах на множестве языков. Содержит аннотированные примеры (более 70 тыс.) токсичных и нетоксичных высказываний на английском, французском, немецком, итальянском, польском, португальском, румынском, русском и испанском языках. Идеален для кросс-лингвистических исследований и проверки устойчивости моделей к языку и культурным особенностям [2].

Выбор датасета зависит от задачи: для исследований на английском языке часто используют Toxic Comment Classification Challenge, для русскоязычных моделей – ru_paradetox_toxicity, а для многоязычных систем – multilingual_toxicity_dataset.

Определим критерии отбора текстов для обучающих датасетов для ИИ систем на основе лингвистических маркеров ограничивающих убеждений.

В качестве источников данных, источников реальных примеров коммуникации возьмем социальную сеть ВКонтакте, и мессенджер Telegram. Новостные комментарии, форумы и блоги – площадки с развернутыми дискуссиями и выражениями личных убеждений, а также неформальная коммуникация с естественным проявлением ограничивающих убеждений в чатах мессенджера.

Определим жанровые рамки:

- Диалогическая коммуникация (комментарии, диалоги, обсуждения);
- Монологические тексты (посты, статьи, блоги);
- Микротексты (короткие сообщения);
- Развернутые тексты (обзоры, аналитические материалы).

Тематическая репрезентативность должна быть представлена отбором текстов по социальным вопросам (темы, вызывающие острые дискуссии), личностным и профессиональным сферам (высказывания о себе и других, обсуждения), общественно-политическим темам (области с высоким потенциалом конфликтности)

Представление текстов с разной степенью выраженности ограничивающих убеждений – от слабых намеков до явных проявлений. Для баланса выборки по степени выраженности токсичности будем учитывать нейтральные тексты.

Существенные критерии отбора: актуальность данных во временном периоде, исключение текстов с чрезмерным количеством шума, полнота контекста коммуникации и репрезентативность, разнообразие по демографическим характеристикам авторов.

При отборе текстов необходимо исключить фейковые и сгенерированные генеративным искусственным интеллектом тексты, сохранять исходное окружения текста и включать в выборку тексты разной длины и структуры.

Схема аннотации текста выстраивается на основе общей информации о документе, контент анализе текстового сообщения, определении уровня маркеров по специфическим признакам, градации интенсивности и категориальным признакам. Учитывается контекст публикации, наличие провокаций и реакция аудитории.

Рассмотрим схему аннотации.

Таблица 2

Схема аннотации с примерами.

Документ	Комментарий к публикации	Пост в социальной сети	Комментарий к публикации
ID			
Источник	Социальная сеть …	Социальная сеть …	Социальная сеть …
Дата публикации			
Анализируемый текст	«В нашем обществе нет места честным людям»	«Все чиновники только и думают о своей выгоде»	«Система создана для того, чтобы обычные люди страдали»
Тематика	Социальная критика, общественные отношения	Критика государственного управления, коррупция	Критика социальных институтов, государ-ственного устройства.
Жанровая принадлежность	Публицистический дискурс с элементами социальной критики. Оценочное суждение с негативной коннотацией.	Общественно-политический дискурс. Обобщающая характеристика социальной группы.	Социально-политический памфлет, конспирологическое повествование.
Целевая аудитория	Широкая общественность.	Широкая общественность.	Широкая общественность.
Эмоциональный тон	Негативный.	Пессимистичный	Негативный.

Контекст	Дискуссия о социальных проблемах.	Дискуссия о социальных проблемах.	Дискуссия о социальных проблемах.
Тип убеждения	Социальное отчуждение.	Обобщение о социальной группе.	Конспирологическое мышление.
Интенсивность	Сильная.	Умеренная.	Сильная.
Маркеры	Абсолютизация.	Категоричность, негативная оценка.	Обобщение, намеренное зло.
Лингвистические маркеры	«нет места» – маркер отсутствия возможностей.	«все» – абсолютизация негативного личного опыта. «только» – намеренное утрирование негативных качеств.	«все страдали» – намеренное зло, обобщение.
Рекомендации	Требуется модерация.	Требуется модерация.	Требуется модерация.

Данная таблица может быть легко адаптирована под конкретные требования и дополнена новыми параметрами при необходимости. Все данные структурированы по уровням анализа для удобства обработки и машинного обучения.

Опираясь на предложенную схему, необходимо обеспечить процедуру валидации аннотированных данных путем двойного кодирования. Несколько, не менее двух, независимых экспертов параллельно аннотируют один и тот же набор данных, используя единую инструкцию по аннотации. При этом, эксперты не должны иметь доступ к результатам друг друга. Затем результаты сравниваются для выявления расхождений.

Количественная оценка межэкспертной согласованности вычисляется с помощью коэффициента Cohen’s kappa:

$\kappa = (Po - Pe)/ (1 - Pe)$,

где Po – наблюдаемое согласие, Pe – ожидаемое случайное согласие.

При $\kappa < 0.4$ – низкая согласованность, 0.4-0.6 – умеренная согласованность, от 0.6 и выше – высокая согласованность.

Опираясь на репрезентативную выборку небольшого размера (50-100 примеров), проводится предварительная аннотация, где определяется среднее время аннотации, сложности инструкции и частота расхождения между экспертами. Вносятся необходимые корректировки в инструкцию и схему аннотации. При необходимости меняется размер выборки. После корректировки проводится повторная аннотация, рассчитывается итоговый показатель согласованности и формируется финальная версия инструкции. Данные считаются валидными при достижении приемлемого уровня kappa ≥ 0.6

Данные для машинного обучения обычно сохраняются в форматах JSON или CSV. В JSON можно хранить не только текст, но и дополнительную информацию о каждом примере: метки, категории, результаты аннотирования. JSON удобен для хранения сложных структур с метаданными. CSV подходит для простых случаев, когда нужно сохранить текст и основные характеристики, удобно для хранения табличных данных.

Для обучения базовой модели требуется от пяти тысяч размеченных примеров.

При этом важно, чтобы данные были сбалансированными по маркерам (лингвистическим, синтаксическим и прагматическим) и представляли все возможные случаи, которые должна уметь распознавать модель.

Предложенная методология имеет несколько важных преимуществ. Во-первых, она учитывает контекст, в котором появляются ограничивающие убеждения. Это значит, что система может точнее определять, когда и почему человек выражает такие убеждения, и лучше понимать их смысл.

Во-вторых, методология гибкая и может быть адаптирована под разные языки и культуры. Это позволяет использовать её в разных странах и для разных групп людей, учитывая их особенности и традиции.

В-третьих, такой подход помогает создавать более точные и эффективные инструменты для анализа текстов, поскольку учитывает не только сами убеждения, но и то, как они проявляются в разных контекстах.

При этом, стоит учитывать, что важную роль играет субъективный фактор: разные эксперты могут по-разному интерпретировать одни и те же тексты, что влияет на качество разметки. А также необходимость обновления наборов данных, чтобы они оставались актуальными и отражали текущие тенденции в языке и коммуникации.

Разработанная система может быть использована двумя основными способами. Можно «научить» лучше понимать тексты с ограничивающими убеждениями, дообучив современные языковые модели, такие как Multilingual BERT или RuBERT. Это поможет им точнее распознавать подобные паттерны мышления в текстах [3]. Или такие обученные модели можно внедрить прямо в социальные сети (ВКонтакте, Telegram, Одноклассники), форумы и мессенджеры, чтобы они автоматически помогали модераторам находить потенциально проблемные сообщения и фильтровать токсичный контент.

Это позволит создать более безопасную и здоровую онлайн-среду для пользователей. По сути, это установка умного фильтра, который помогает поддерживать здоровую атмосферу в онлайн-пространстве и защищает пользователей от нежелательного контента.

В результате проведенного исследования была разработана комплексная методология формирования обучающих датасетов на основе лингвистических маркеров ограничивающих убеждений, которая учитывает контекстуальную специфику и может быть адаптирована под различные языки и культуры.

Предложенный подход позволяет эффективно выявлять токсичный контент через систематизацию языковых маркеров и их автоматическую обработку, что открывает новые возможности для создания более совершенных инструментов модерации в социальных сетях и мессенджерах.

Несмотря на определенные ограничения, связанные с субъективностью экспертной разметки и необходимостью регулярного обновления данных, разработанная методология имеет значительный потенциал для практического применения, в частности, в области дообучения трансформерных моделей и интеграции в системы автоматической модерации контента.

Перспективными направлениями дальнейшей работы являются автоматизация процесса разметки с помощью слабых сигналов и проведение кросс-культурных исследований маркеров токсичности для повышения эффективности систем выявления деструктивного контента.

Список литературы:

1. Dixon L., Li J., Sorensen J., Thain N., Vasserman L. Measuring and Mitigating Unintended Bias in Text Classification // Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. – New Orleans, LA, USA: ACM, 2018. – P. 67–73

2. Hugging Face. Datasets Hub [Электронный ресурс] // Hugging Face Hub. – URL: huggingface.co/datasets (дата обращения: 15.11.25)

3. Rothman, D. Transformers for Natural Language Processing: Build Innovative Deep Neural Network Architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and More. Birmingham: Packt Publishing, Limited, 2021. 384 p

4. Toxic-comments-detector: система обнаружения токсичных комментариев / Barsukov N.S [Электронный ресурс]. – Лицензия: MIT. – URL: https://github.com/nsbarsukov/toxic-comments-detector (дата обращения: 10.11.25).

5. Wiegand M., Siegel M., Ruppenhofer J. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language // Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018). – Vienna, Austria: Austrian Academy of Sciences, 2018. – P. 1–10.

6. Авдеев П.С. Понятие, функции и виды убеждений // Психология, социология и педагогика. – 2014. – № 12 (39). – С. 72-75

7. Келли Дж. А. Теория личности: Психология личностных конструктов / Пер. с англ. и науч. ред. А. А. Алексеева. – СПб.: Речь, 2000. – 248 с. – (Мастерская психологии и психотерапии). ISBN 5-9268-0007-2.

8. Крылова М.Н. Средства художественной выразительности. Тропы: Учебное пособие. – М.: Директ-Медиа, 2014. – 101 с.

9. Национальный корпус русского языка. – URL: https://ruscorpora.ru (дата обращения: 12.11.2025).

10. Шевкунова Т.П. Основные лексические средства конструирования ограничивающих убеждений//Современная наука: актуальные проблемы теории и практики. – 2024. – 9-2 С. 208-210.