МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра социальной информатики

ОЦЕНКА ЭФФЕКТИВНОСТИ ИСПОЛЬЗОВАНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ВЫЯВЛЕНИИ ФЕЙКОВ И ДЕЗИНФОРМАЦИИ В МЕССЕНДЖЕРЕ TELEGRAM И СОЦИАЛЬНОЙ СЕТИ ВКОНТАКТЕ

(автореферат бакалаврской работы)

студента 5 курса 531 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Митяева Алексея Дмитриевича

Научный руководитель
профессор, доктор социологических наук,
профессор _____ О.А. Романовская
подпись, дата

Зав. кафедрой
кандидат социологических наук, доцент _____ И.Г. Малинский
подпись, дата

Саратов 2026

ВВЕДЕНИЕ

Актуальность темы исследования связана с ростом роли социальных сетей и мессенджеров в распространении новостей и формировании общественного мнения. Telegram и «ВКонтакте» занимают значимое место в российском информационном пространстве, однако высокая скорость передачи сообщений, децентрализованное производство контента, пересылки, репосты и комментарии создают условия для быстрого распространения фейков и дезинформации. Рост объема цифровых данных ограничивает возможности ручной проверки, что делает необходимой оценку эффективности технологий искусственного интеллекта, машинного обучения и обработки естественного языка при первичном выявлении недостоверных текстовых сообщений.

Степень научной разработанности проблемы. Проблематика дезинформации и манипуляции общественным сознанием рассматривается в работах У. Липпмана[1] и Г. Лассуэлла[2], заложивших основы анализа информационного воздействия и формирования общественного мнения. Социологическое осмысление связи информации, идеологии и социальных интересов представлено в трудах К. Мангейма[3], а критический анализ медиасистемы как механизма управления вниманием - в работах М. Хоркхаймера и Т. Адорно[4]. Политико-философское понимание разрушения границ между истиной и вымыслом раскрывается в исследованиях Х. Арендт[5].

Переход к информационному и сетевому обществу получил отражение в трудах Д. Белла[6], Э. Тоффлера[7] и М. Кастельса[8], где информация

---

[1] Липпман, У. Общественное мнение / У. Липпман. – Москва: АСТ, 2023. – 448 с.
[2] Лассуэлл, Г. Техника пропаганды в мировой войне / Г. Лассуэлл. – Москва: ИНИОН РАН, 2022. – 252 с.
[3] Мангейм, К. Идеология и утопия: [в 2 томах] / К. Мангейм. – Москва: АН СССР, 1976. – 399 с.
[4] Хоркхаймер М. Диалектика просвещения / М. Хоркхаймер, Т. Адорно. – Санкт-Петербург: Медиум — Ювента, 1997. – 312 с.
[5] Арендт, Х. Истоки тоталитаризма / Х. Арендт. – Москва: Скрипториум, 2020. – 672 с.
[6] Белл, Д. Грядущее постиндустриальное общество: Опыт социального прогнозирования / Д. Белл. – Москва: Academia, 1999. – 783 с.
[7] Тоффлер, Э. Третья волна / Э. Тоффлер. – Москва: АСТ, 2004. – 345 с.
[8] Кастельс, М. Информационная эпоха. Экономика, общество и культура / М. Кастельс. –

рассматривается как ключевой ресурс социальной организации. Вместе с тем значительная часть исследований ориентирована на традиционные медиа, тогда как цифровая среда, Telegram, «ВКонтакте» и применение искусственного интеллекта для выявления фейков требуют дальнейшего анализа. Недостаточно разработанным остается вопрос оценки эффективности машинного обучения при выявлении дезинформации в русскоязычном цифровом контенте.

Объект исследования - фейковый контент в цифровой среде.

Предмет исследования - методическая специфика моделей выявления фейков и дезинформации в мессенджере Telegram и социальной сети «ВКонтакте» при помощи технологий искусственного интеллекта.

Цель исследования: оценка эффективности методов и моделей выявления фейков и дезинформации на основе технологий искусственного интеллекта в мессенджере Telegram и социальной сети «ВКонтакте».

Реализация заявленной цели обусловливает необходимость постановки и решения следующих **задач**:

1. Раскрыть основные научные подходы к изучению феномена дезинформации и фейков;
2. Определить особенности распространения фейков и дезинформации в цифровой информационной среде, значимые для их автоматизированного выявления;
3. Сформулировать критерии и показатели, позволяющие выявлять фейки и дезинформацию в цифровом контенте с использованием технологий искусственного интеллекта;
4. Обосновать выбор архитектуры и принципов построения модели выявления фейков и дезинформации на основе технологий искусственного интеллекта;
5. Разработать модель выявления фейков и дезинформации на основе технологий искусственного интеллекта;

Москва: ИД ВШЭ, 1996. – 609 с.

6. Осуществить оценку эффективности разработанной модели выявления фейков и дезинформации на материалах мессенджера Telegram и социальной сети «ВКонтакте».

Эмпирическая база исследования состоит из первичной социологической информации, включающей массив коротких новостных публикаций, отобранных из мессенджера Telegram и социальной сети «ВКонтакте» для последующей автоматизированной оценки в программе «Детектор Фейк-Новостей».

В состав эмпирической базы вошли 5758 текстовых записей, сформированных на основе новостного контента Telegram и «ВКонтакте». Корпус включает две равные категории: 2879 фейковых новостных сообщений и 2879 достоверных новостных сообщений. Основной единицей анализа выступает отдельный текстовый информационный блок, представляющий собой законченную новостную публикацию, направляемую в программу для классификации. На данном массиве осуществлялись обучение, проверка и оценка эффективности модели машинного обучения при выявлении фейков и дезинформации в цифровой медиасреде.

Структура работы. ВКР состоит из введения, трех глав по три параграфа, заключения и списка использованных источников.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе «Дезинформация как социальный феномен в условиях цифрового социального пространства» дезинформация рассматривается как устойчивый механизм воздействия на общественное сознание, связанный с развитием массовых и цифровых коммуникаций. Теоретической основой анализа выступают работы У. Липпмана и Г. Лассуэлла, в которых массовая коммуникация раскрывается как инструмент формирования общественного мнения, политической мобилизации и управления коллективным восприятием. Подходы К. Мангейма, М. Хоркхаймера, Т. Адорно и Х. Арендт позволяют рассматривать искажение информации не только как отдельный акт

обмана, но и как элемент идеологического влияния, культурной индустрии, политического контроля и разрушения рационального публичного обсуждения.

Историческое развитие дезинформации связано с переходом от централизованных форм пропагандистского воздействия к сетевым и распределенным моделям коммуникации. В традиционной медиасреде ключевую роль играли государственные институты, редакции, политические структуры и профессиональные посредники. В цифровой среде круг участников значительно расширяется: пользователь не только получает информацию, но и распространяет, комментирует, переупаковывает и эмоционально усиливает ее. Концепции Д. Белла, Э. Тоффлера и М. Кастельса связывают данный процесс с переходом к информационному и сетевому обществу, где информация становится самостоятельным ресурсом социальной организации, а коммуникационные структуры приобретают решающее значение для формирования общественных представлений.

Механизмы информационного воздействия проявляются через управление контекстом, фреймирование событий, эмоциональное насыщение сообщения, повторение устойчивых смысловых конструкций и изменение доверия к источникам. Манипуляция в медиасреде не всегда строится на прямом распространении ложных фактов. Нередко она возникает через отбор отдельных фрагментов реальности, смещение акцентов, создание ощущения срочности, апелляцию к страху, возмущению или групповым ожиданиям. В цифровой среде данные механизмы усиливаются алгоритмической персонализацией, эхо-камерами, репостами, комментариями и видимыми реакциями других пользователей.

Цифровизация меняет саму структуру распространения фейков. Telegram и «ВКонтакте» формируют разные модели обращения информации: в Telegram важны скорость публикации, пересылки и канальная логика распространения, во «ВКонтакте» - социальные связи, сообщества, комментарии и групповое обсуждение. В результате фейк существует не только как отдельный текст, но и как часть сетевого процесса, где значение имеют источник, формат подачи,

повторяемость сообщения, эмоциональная реакция аудитории и траектория распространения.

Дезинформация в цифровом обществе выступает не только информационной проблемой, но и социальным феноменом, встроенным в механизмы доверия, восприятия и сетевого взаимодействия. Ее распространение определяется не только содержанием сообщения, но и архитектурой цифровой среды, скоростью коммуникации, алгоритмической персонализацией и реакцией аудитории. Поэтому выявление фейков требует сочетания социологического анализа цифровых коммуникаций и методов автоматизированной обработки текстового контента.

Во втором разделе «Особенности фейков и дезинформации в цифровой медиасреде» Telegram и «ВКонтакте» представлены как разные площадки распространения новостного контента. В Telegram сообщение быстро проходит через каналы, пересылки, группы и личные чаты. Один и тот же информационный повод может многократно появляться в разных каналах, менять заголовок, эмоциональную окраску или ссылку на источник. Для фейков в данной среде важны скорость публикации, краткость сообщения, эффект срочности и возможность быстрого тиражирования.

Во «ВКонтакте» публикация размещается через профиль, группу или публичное сообщество, попадает в ленту и включается в обсуждение через комментарии, репосты и реакции пользователей. Недостоверное сообщение здесь может распространяться медленнее, чем в Telegram, но дольше сохраняться внутри групповой коммуникации. На восприятие новости влияет не только содержание, но и тип сообщества, комментарии пользователей, видимая реакция аудитории и повторное появление записи в ленте.

Фейковый контент в цифровой среде имеет ряд признаков, значимых для анализа. Среди них эмоционально окрашенная лексика, категоричные утверждения, неопределенные ссылки на источник, маркеры срочности, упрощенные объяснения причин и последствий, повторяемость формулировок и имитация обычного новостного стиля. Значение имеют и параметры

распространения: резкий рост внимания к публикации, активные пересылки, репосты, комментарии, появление похожих сообщений в разных источниках. Данные признаки не доказывают ложность сообщения напрямую, но помогают выделить публикации, нуждающиеся в дополнительной проверке.

Фейки и дезинформация в Telegram и «ВКонтакте» имеют не только содержательные, но и платформенные признаки. Telegram усиливает скорость и повторяемость сообщения, а «ВКонтакте» закрепляет публикацию через сообщества, комментарии и репосты. Для первичного автоматизированного анализа наиболее значимой остается текстовая часть новости, поскольку в ней фиксируются лексические и структурные признаки, пригодные для машинной классификации. Выделенные характеристики позволяют рассматривать короткий новостной текст как основную единицу дальнейшего анализа.

Третий раздел «Применение технологий искусственного интеллекта для выявления фейков и дезинформации» связан с практической реализацией модели машинного обучения и оценкой ее эффективности. Автоматическое распознавание недостоверного контента рассматривается как задача бинарной классификации, где текстовое сообщение должно быть отнесено к одной из двух категорий: «фейк» или «правда». Ручная проверка больших массивов новостных публикаций требует значительных временных и экспертных ресурсов, поэтому в работе используются методы машинного обучения и обработки естественного языка, позволяющие выявлять статистические закономерности в текстах.

Для разработки программного решения были определены требования к автономности, корректной обработке русскоязычного текста, скорости предсказаний, возможности дообучения и простоте использования. Приложение FakeNewsDetector реализовано как настольная Windows-программа с графическим интерфейсом. В качестве технологической основы использованы C#, .NET Framework 4.8 и библиотека ML.NET. Архитектура построена по модульному принципу: отдельные блоки отвечают за загрузку CSV-файлов, распознавание кодировки, извлечение текстов, предобработку, обучение модели,

сохранение результата, проверку пользовательского сообщения и дополнительную проверку по известным фейковым источникам.

Работа модели строится через последовательный цикл обработки данных. Сначала загружаются размеченные CSV-файлы, затем тексты очищаются от ссылок, доменных имен, специальных символов, знаков препинания и лишних пробелов. После нормализации текст преобразуется в числовые признаки, пригодные для машинного анализа. Для классификации выбран алгоритм LightGBM, а выборка разделяется на обучающую и тестовую части в соотношении 80% и 20%. После обучения модель сохраняется в файл model.zip, что позволяет использовать ее повторно без нового обучения при каждой проверке сообщения.

Пользовательский интерфейс включает две вкладки: «Обучение модели» и «Анализ текста». Первая вкладка предназначена для выбора CSV-файлов, запуска обучения и отображения метрик качества. Вторая вкладка используется для проверки новостного текста: пользователь вводит сообщение, запускает анализ и получает результат классификации с процентом уверенности. Дополнительно предусмотрен префильтр по источникам: при наличии в тексте домена известного фейкового ресурса сообщение может быть сразу отмечено как фейковое без обращения к модели машинного обучения.

Эффективность модели оценивалась по метрикам Accuracy, Precision, Recall, F1-score и ROC-AUC. По итогам тестирования на подготовленной выборке общая точность классификации составила 87,2%, Precision по классу «фейк» - 85,1%, Recall - 90,3%, F1-score - 87,6%, ROC-AUC - 93,4%. Полученные показатели свидетельствуют о применимости модели для первичной классификации коротких новостных текстов. При этом программа не устанавливает достоверность сообщения окончательно: ее назначение состоит в предварительном выделении публикаций, которые обладают признаками фейка и требуют дальнейшей проверки.

Разработанная модель показала применимость машинного обучения для первичной классификации коротких новостных текстов, однако качество

результата зависит от состава обучающей базы и выраженности текстовых признаков. Ошибки возможны при очень коротких сообщениях, нейтральном официальном стиле фейка, ироничных оборотах, нестандартной лексике или цитатах из сомнительных источников. Дальнейшее повышение качества связано с расширением русскоязычной обучающей базы, добавлением признаков эмоциональной окраски, учетом частотности редких слов и применением более сложных языковых моделей.

ЗАКЛЮЧЕНИЕ

Фейки и дезинформация в Telegram и «ВКонтакте» функционируют не как случайные ошибки информационного обмена, а как устойчивые цифровые сообщения с повторяемой структурой. Их воздействие обеспечивается сочетанием нескольких элементов: категоричного утверждения, слабого или неопределенного источника, эмоциональной рамки, имитации доказательности и возможности быстрого распространения. Ложность сообщения в цифровой среде часто маскируется не сложной аргументацией, а внешними признаками достоверности: ссылкой на «очевидцев», скриншотом, фрагментом документа, указанием на ведомство, датой, географической привязкой или тревожной практической рекомендацией.

Telegram и «ВКонтакте» формируют разные механизмы закрепления недостоверной информации. В Telegram фейк получает силу за счет скорости, краткости, пересылок и многократного появления близких формулировок в разных каналах. Во «ВКонтакте» более значимыми становятся комментарии, репосты, видимость реакции аудитории и включенность публикации в социальное окружение пользователя. Поэтому выявление фейков не может ограничиваться анализом текста в отрыве от платформенной среды: один и тот же тезис в Telegram чаще действует как оперативный информационный сигнал, а во «ВКонтакте» - как элемент группового обсуждения и социального подтверждения.

Главный аналитический результат состоит в том, что отдельный признак не позволяет надежно определить фейковый характер публикации.

Эмоциональная лексика, срочность, визуальный материал, большое число реакций или отсутствие ссылки могут встречаться и в достоверных сообщениях. Значение имеет именно сочетание признаков: неопределенный источник, категоричная форма, тревожная тема, слабая проверяемость, псевдодоказательство, призыв к распространению и повторяемость смысловой конструкции. Сочетание признаков позволяет рассматривать фейк не как оценочное суждение, а как цифровой объект, который может быть частично формализован для машинной обработки.

Разработанная модель машинного обучения показала применимость автоматизированной текстовой классификации для первичной оценки коротких сообщений. Ее результат не устанавливает фактическую истинность или ложность публикации, а показывает близость текста к признакам, сформированным на обучающем корпусе. Наиболее значимой оказалась не только возможность получить ответы «Правда» и «Вероятно фейк», но и наличие промежуточной метки «Неопределенно». Она фиксирует ситуации, в которых сообщение содержит отдельные признаки оценочной или новостной подачи, но без источника и контекста не дает основания для уверенной классификации.

Эффективность модели проявляется в ее способности быстро выделять тексты повышенного риска среди коротких публикаций, заголовков и сжатых новостных формулировок. При этом границы применения остаются принципиальными: текстовая модель не анализирует изображения, видео, ссылки, происхождение публикации, историю распространения и пользовательские реакции. Следовательно, фейки, построенные на подмене визуального контекста, ложной подписи к реальному изображению или манипулятивном использовании старого документа, требуют более сложной проверки. Машинное обучение в данном случае повышает скорость первичного отбора, но не заменяет содержательный фактчекинг.

Искусственный интеллект может быть эффективен в выявлении фейков и дезинформации только как часть более широкой аналитической процедуры. Его практическая ценность состоит в формализации признаков риска, сокращении

объема ручной проверки и выделении сообщений, требующих дополнительного анализа. Наиболее обоснованной является комбинированная модель работы: автоматизированная классификация текста, затем проверка источника, контекста, визуальных материалов и траектории распространения. При таком формате модель машинного обучения становится не универсальным детектором истины, а прикладным инструментом анализа цифровой дезинформации в Telegram и «ВКонтакте».