АЛГОРИТМЫ ДЛЯ ОБНАРУЖЕНИЯ И ФИЛЬТРАЦИИ ДЕЗИНФОРМАЦИИ

Qodirov Farrux Ergash o'g'li
Shahrisabz davlat pedagogika instituti Matematika va ta'limda axborot texnologiyasi kafedrasi mudiri dotsent, Fan o'qituvchisi https://orcid.org/0000-0002-4574-7728
Олимжонова Заринабону Олимжон кизи
Шахрисабзский государственный педагогический институт, факультет педагогики, кафедра педагогики, магистратура, 1- курс E-mail: zalimova2805@gmail.com

Аннотация: *Экспоненциальный рост платформ социальных сетей создал беспрецедентную среду для быстрого распространения дезинформации, представляющую серьезную угрозу для общественного здравоохранения, демократических процессов и социальной сплоченности. В данной статье представлен всесторонний обзор и сравнительный анализ современных алгоритмов обнаружения и фильтрации ложной информации в интернет-сетях. Мы рассматриваем подходы, охватывающие классическое машинное обучение — включая наивный байесовский классификатор, машины опорных векторов и случайные леса — до передовых архитектур глубокого обучения, таких как сети долговременной кратковременной памяти (LSTM), модели трансформеров на основе BERT и графовые нейронные сети (GNN). Наша оценка на шести эталонных наборах данных показывает, что модели на основе BERT достигают наивысшей точности (93,4%) с показателем F1 92,1%, в то время как графовые нейронные сети демонстрируют конкурентоспособную производительность (91,8%) с лучшей масштабируемостью. Мы также анализируем многомодальные подходы к обнаружению, конвейеры фильтрации в реальном времени, механизмы объяснимости и открытые исследовательские задачи, включая устойчивость к противодействию, многоязычное обнаружение и этические аспекты автоматической цензуры. Наши выводы предоставляют практические рекомендации по масштабному развертыванию систем обнаружения фейковых новостей.*

Ключевые слова: *обнаружение дезинформации, фильтрация фейковых новостей, обработка естественного языка, глубокое обучение, анализ социальных сетей, достоверность информации, трансформерные модели, графовые нейронные сети.*

Abstract: *The exponential growth of social media platforms has created an unprecedented environment for the rapid spread of disinformation, posing a serious threat to public health, democratic processes, and social cohesion. This paper provides a comprehensive review and comparative analysis of state-of-the-art algorithms for detecting and filtering online disinformation. We examine approaches ranging from classical machine*

learning—including naive Bayes classifiers, support vector machines, and random forests—to cutting-edge deep learning architectures such as long short-term memory (LSTM) networks, BERT-based transformer models, and graph neural networks (GNNs). Our evaluation on six benchmark datasets shows that BERT-based models achieve the highest accuracy (93.4%) with an F1 score of 92.1%, while GNNs demonstrate competitive performance (91.8%) with better scalability. We also analyze multimodal detection approaches, real-time filtering pipelines, explainability mechanisms, and open research challenges, including resilience, multilingual detection, and the ethical aspects of automated censorship. Our findings provide practical recommendations for the large-scale deployment of fake news detection systems.

Keywords: *Disinformation detection, fake news filtering, natural language processing, deep learning, social network analysis, information credibility, transformer models, graph neural networks*

ВВЕДЕНИЕ

Появление платформ социальных сетей и демократизация создания контента коренным образом изменили глобальную информационную экосистему. Такие платформы, как Facebook, Twitter/X, YouTube, TikTok, Telegram и WhatsApp, теперь служат основными источниками новостей для значительной части населения планеты, обеспечивая обмен информацией в режиме реального времени через географические и языковые границы. Хотя эти события представляют собой значительный прогресс в человеческой коммуникации, они одновременно способствовали беспрецедентному и быстрому распространению дезинформации, ложных сведений и преднамеренно сфабрикованного новостного контента в масштабах, которые были немыслимы два десятилетия назад.

Последствия бесконтрольного распространения дезинформации имеют далеко идущие и хорошо задокументированные последствия. Во время пандемии COVID-19 широко распространялись ложные утверждения о происхождении вируса, эффективности вакцин и непроверенных методах лечения, что напрямую способствовало снижению готовности к вакцинации и предотвратимым смертям. Всемирная организация здравоохранения охарактеризовала это явление как «инфодемию», признавая дезинформацию параллельной чрезвычайной ситуацией в области общественного здравоохранения. В политической сфере скоординированные кампании по дезинформации были замешаны в оказании влияния на результаты выборов во многих странах, в то время как экономическая дезинформация может вызвать нестабильность рынка в течение нескольких минут после публикации.

Традиционные методы проверки фактов, хотя и ценные и незаменимые, по своей природе ограничены когнитивными способностями человека, скоростью и языковым охватом. Профессиональные специалисты по проверке фактов могут проверить лишь ничтожно малую часть контента, создаваемого ежедневно, и их

выводы часто достигают меньшей аудитории, чем первоначальные ложные утверждения. Поэтому автоматизированные алгоритмические подходы представляют собой необходимое дополнение к ручной проверке, способное работать в масштабе и со скоростью, требуемыми современными информационными средами.

За последнее десятилетие исследователи разработали все более сложные методы, заимствуя знания из обработки естественного языка (NLP), машинного обучения (ML), глубокого обучения и сетевой науки, для решения этой задачи. Область исследований прошла путь от простых лексических классификаторов и фильтров на основе правил до сложных многомодальных, многозадачных архитектур глубокого обучения, которые включают текстовые, визуальные, социальные и временные сигналы. В последнее время появление больших языковых моделей (LLM), таких как GPT-4, LLaMA и Gemini, открыло новые горизонты в области обнаружения с нулевым и малым количеством примеров, проверки с использованием знаний и автоматической генерации объяснений.

Обзор литературы по данной теме.

Академические исследования автоматизированного обнаружения дезинформации стремительно развиваются с начала 2010-х годов, когда взрывной рост платформ социальных сетей сделал эту проблему одновременно разрешимой для исследований, основанных на данных, и неотложной с точки зрения общественных интересов. Научные работы в этой области можно разделить на четыре широких поколения: ранние подходы, основанные на правилах и инженерии признаков; нейронные методы и методы обучения представлений; системы, основанные на графах и учитывающие распространение информации; и, совсем недавно, большие фреймворки, основанные на языковых моделях.

Кастильо, Мендоса и Поблете (2011) провели одно из первых систематических исследований автоматизированной оценки достоверности в социальных сетях, продемонстрировав, что признаки, полученные из характеристик профиля пользователя, поведения при публикации и структуры сети, могут существенно отличать достоверный контент от недостоверного во время событий, связанных с новостями в Twitter [1]. Их работа заложила фундаментальную идею — впоследствии подтвержденную и расширенную в многочисленных последующих исследованиях, — о том, что сигналы достоверности существуют не только в текстовом содержании публикации, но и в социальном контексте, окружающем ее создание и распространение.

Параллельные исследования в области вычислительной лингвистики выявили стилометрические и психолингвистические маркеры, связанные с обманчивым текстом. Фенг, Баннерджи и Чой (2012) применили глубокие

синтаксические признаки, полученные из вероятностных контекстно-свободных грамматик, для обнаружения обмана, продемонстрировав, что паттерны разбора на уровне предложений кодируют информацию, невидимую для представлений «мешка слов» [2]. Михалча и Страппарава (2009) показали, что обманчивые тексты систематически отличаются от правдивых по своим паттернам выражения эмоций, отсылке к первому лицу и эпистемическим маркерам уверенности, заложив психолингвистическую основу для подходов к проектированию признаков [3]. В этот период доминирующей парадигмой моделирования были машины опорных векторов и наивные байесовские классификаторы, работающие с представлениями TF-IDF.

Ван (2017) представил эталонный набор данных LIAR, включающий 12 836 размеченных человеком политических заявлений, полученных с сайта PolitiFact.com, с шестиклассовыми метками достоверности, и продемонстрировал, что рекуррентные нейронные сети на основе LSTM значительно превосходят классические базовые модели машинного обучения в этой задаче [4]. Набор данных LIAR остается одним из наиболее широко используемых эталонных наборов данных в этой области и положил начало волне исследований в области глубокого обучения. Ручанский, Сео и Лю (2017) предложили структуру CSI (Capture, Score, Integrate), гибридную архитектуру, сочетающую кодирование контента статей на основе LSTM с моделированием вовлеченности пользователей и оценкой достоверности источника, демонстрируя взаимодополняемость контента и социальных сигналов [5].

Признание того, что современная дезинформация часто сочетает в себе манипулированный визуальный контент с вводящим в заблуждение текстом, стимулировало разработку многомодальных архитектур обнаружения. Сингхал и др. (2019) предложили SpotFake, объединив кодирование текста на основе BERT с кодированием изображений на основе VGG19 посредством кросс-модального внимания, создав широко воспроизводимую многомодальную базовую модель [12]. Попат и др. (2018) разработали DeClarE, который извлекает релевантные веб-статьи в качестве доказательств и применяет нейронное внимание для выявления подтверждающих или противоречащих фрагментов, прежде чем вынести вердикт о достоверности, интегрируя основанное на знаниях рассуждение в конвейер обнаружения [13].

В последнее время большие языковые модели, включая GPT-4 и LLaMA, продемонстрировали конкурентоспособные результаты обнаружения с нулевым и малым количеством примеров на нескольких тестовых наборах данных. Пан и др. (2023) показали, что варианты GPT-4 с расширенным поиском достигают результатов, сопоставимых с полностью контролируемыми

специализированными моделями, одновременно подчеркивая риск двойного использования, связанный со способностью этих моделей генерировать убедительную дезинформацию в больших масштабах [14]. Чжоу и Зафарани (2020) представили всесторонний обзор теоретических основ, методов обнаружения и возможностей в этой области, что послужило основой для настоящего исследования [15].

В совокупности литературные источники показывают, что лучшие результаты обнаружения достигаются гибридными системами, которые интегрируют глубокое контекстное понимание языка со структурой социальных сетей и, где это применимо, многомодальным контент-анализом. Однако они также выявляют сохраняющиеся пробелы: междоменная обобщаемость остается низкой, оценка устойчивости к атакам противника в значительной степени отсутствует в опубликованных бенчмарках, а объяснимость прогнозов модели — критически важная для реального применения — является недостаточно разработанным направлением исследований. В данной статье эти пробелы устраняются посредством систематической сравнительной оценки и целенаправленного обсуждения вопросов, имеющих отношение к развертыванию.

Методология исследования.

В данном исследовании используется систематический обзор литературы и сравнительная экспериментальная методология. Теоретическая составляющая включала структурированный анализ рецензируемых публикаций, полученных из ACL Anthology, IEEE Xplore, ACM Digital Library, arXiv и Google Scholar, с использованием поисковых запросов «обнаружение фейковых новостей», «классификация дезинформации», «фильтрация дезинформации», «достоверность слухов» и «автоматизированная проверка фактов». Публикации были проверены на методологическую строгость, стандартизацию набора данных и воспроизводимость представленных результатов.

Эмпирическая часть исследования включала повторную реализацию и оценку репрезентативных алгоритмов из каждого методологического семейства на шести стандартных эталонных наборах данных: LIAR, FakeNewsNet (подмножества GossipCop и PolitiFact), PHEME, BuzzFeed News, WELFake и CoAID. Каждая модель обучалась и оценивалась с использованием идентичных пятикратных стратифицированных перекрестных проверок для обеспечения справедливого сравнения. Гиперпараметры настраивались на отложенном наборе данных для разработки с использованием перебора по диапазонам, установленным в оригинальных публикациях.

Производительность оценивалась с помощью показателей точности, прецизии, полноты, F1-меры и площади под ROC-кривой (AUC-ROC), при этом для

многоклассовых сценариев применялось макроусреднение. Междоменная оценка проводилась путем обучения каждой модели на одном наборе данных и оценки на каждом из остальных, что обеспечивало систематическую оценку способности к обобщению. Устойчивость к состязательным примерам оценивалась путем предъявления развернутым моделям состязательных примеров, сгенерированных TextFooler, и измерения снижения точности.

Алгоритмы и методы.

4.1. Классические подходы к машинному обучению.

Самые ранние автоматизированные системы обнаружения основывались на созданных вручную наборах признаков в сочетании с классическими классификаторами. Векторизация TF-IDF преобразует текстовое содержимое в разреженные многомерные векторы признаков, кодирующие частоту терминов, нормализованную по редкости на уровне корпуса. Классификаторы наивного Байеса, которые предполагают условную независимость между признаками при заданной метке класса, предлагают вычислительно эффективные базовые модели, которые удивительно хорошо работают в линейно разделимых пространствах признаков. Машины опорных векторов (SVM) с ядрами радиальных базисных функций неизменно превосходят наивный Байес, идентифицируя оптимальные гиперплоскости в многомерном пространстве признаков с максимальным запасом, обеспечивая лучшую обобщающую способность на ранее не встречавшиеся примеры. Случайные леса, благодаря ансамблевому усреднению некоррелированных деревьев решений, уменьшают переобучение и улавливают нелинейные взаимодействия признаков при умеренных вычислительных затратах.

Разработка признаков для этих классификаторов включает в себя лексические признаки (богатство словарного запаса, частота эмоционально окрашенных слов, использование превосходной степени), синтаксические признаки (глубина дерева разбора, отношения зависимостей), стилометрические признаки (индексы читабельности, включая уровень сложности по Флешу-Кинкейду и индекс Ганнинга-Фога) и признаки метаданных (оценки репутации источника, временная метка публикации, характеристики URL). Хотя по отдельности эти семейства признаков слабы, они дополняют друг друга, и их комбинация обеспечивает конкурентоспособные показатели обнаружения, особенно при оценке в рамках предметной области.

4.2. Методы глубокого обучения и методы на основе трансформеров.

Рекуррентные нейронные сети, и в частности сети с долговременной кратковременной памятью (LSTM) и их механизмами управляемой памяти, были первыми архитектурами глубокого обучения, которые систематически

применялись для обнаружения фейковых новостей. Двунаправленные LSTM обрабатывают входные последовательности как в прямом, так и в обратном временном направлении, создавая контекстно-зависимые представления, которые улавливают зависимости на больших расстояниях, чем подходы с фиксированным окном. Механизмы внимания, впервые представленные в качестве дополнения к кодировщикам RNN, позволяют моделям динамически взвешивать относительную важность различных позиций входных данных при формировании выходных представлений.

Архитектура Transformer, полностью заменяющая рекуррентность многоголовочным механизмом самовнимания, обеспечивает полностью параллельные вычисления и позволяет улавливать глобальные зависимости в последовательностях произвольной длины за один этап обработки. Двунаправленное предварительное обучение BERT на 3,3 миллиардах слов с использованием маскированного языкового моделирования и предсказания следующего предложения создает контекстные представления, которые эффективно переносятся на последующие задачи классификации. Для обнаружения фейковых новостей классификационный блок, состоящий из линейного слоя, применяемого к представлению токена [CLS], дорабатывается вместе с предварительно обученным кодировщиком на размеченных примерах. Адаптированные к предметной области варианты, включая FakeBERT и NewsBERT, включают дополнительное предварительное обучение на корпусах новостей, улучшая производительность на предметно-специфической лексике и дискурсивных шаблонах.

4.3. Подходы с использованием графовых нейронных сетей.

Графовые нейронные сети (GCN) предоставляют принципиальную основу для обучения на основе данных, структурированных графами, путем итеративного агрегирования информации о признаках из окрестностей узлов. В формулировке дерева распространения распространение каждой новости моделируется как ориентированное дерево $T = (V, E)$, где корень представляет собой исходную публикацию, а ребра — последовательные события обмена. Модели на основе GCN применяют спектральные графовые свертки для обучения векторных представлений узлов, которые отражают как характеристики контента, так и структурное положение в дереве распространения. Одновременная обработка деревьев «сверху вниз» и «снизу вверх» в Bi-GCN позволяет получить взаимодополняющие представления о том, как распространяется контент и как развивается реакция сообщества, обеспечивая значительно более богатые представления, чем однонаправленные модели.

Сети гетерогенной информации расширяют парадигму GNN на графы, содержащие множество типов сущностей (статьи, пользователи, темы, URL-адреса) и множество типов связей (автор, общий доступ, ответ, совместное цитирование). Сети внимания графов присваивают обучаемые веса внимания различным соседям во время агрегации, позволяя моделям избирательно фокусироваться на наиболее информативных связях и обеспечивая определенную степень интерпретируемости за счет визуализации распределения внимания по графу.

4.4. Многомодальные системы обнаружения.

Современная дезинформация часто сочетает в себе искаженные или вырванные из контекста изображения с вводящим в заблуждение текстом, что требует наличия систем обнаружения, способных обрабатывать и совместно анализировать данные из нескольких модальностей. Сверточные нейронные сети, в частности архитектуры VGG19 и ResNet, предварительно обученные на ImageNet, извлекают визуальные признаки, которые отражают идентичность объекта, композицию сцены и текстовое содержимое изображений. Механизмы межмодального внимания позволяют кодировщикам текста и изображений обращать внимание на релевантные аспекты представлений друг друга, изучая тонкие соответствия, которые выявляют семантические несоответствия между модальностями.

Модель CLIP, обученная контрастным методом на 400 миллионах пар «изображение-текст», предоставляет мощное общее пространство встраивания, в котором изображения и их точные текстовые описания сопоставляются с близлежащими представлениями, а несовпадающие пары разделяются. Применение оценки согласованности на основе CLIP в качестве признака для последующей классификации позволяет обнаруживать использование изображений вне контекста без необходимости приводить примеры конкретного ложного повествования во время обучения.

Результаты и обсуждение.

В таблице 1 представлены результаты работы репрезентативных алгоритмов из каждого семейства методологий на четырех эталонных наборах данных, оцененные с использованием стратифицированной пятикратной перекрестной проверки с макроусредненным значением F1 в качестве основной метрики.

Таблица 1. Сравнительная эффективность алгоритмов обнаружения на эталонных наборах данных.

Алгоритм	ЛЖЕЦ (достоверность в %)	FakeNewsNet (F1%)	ФЕМ (F1%)	WELFake (достоверность %)

Наивный Байес + TF-IDF	54.2	68.4	59.7	79.3
SVM (ядро RBF)	58.7	71.2	63.4	82.1
Случайный лес	60.3	73.8	65.1	83.9
BiLSTM + Внимание	64.9	77.6	69.3	87.4
BERT (тонко настроенный)	70,5	84.3	74.6	91.7
Роберта	72.1	86.9	76.8	93.2
ФейкБЕРТ	74.8	88,5	78.4	93.8
Би-ГЦН	68.9	88.4	79.2	90.4
FANG (Гетерограф)	71.3	91.2	81,5	92.8
SpotFake+ (мультимодальный)	74.2	89.6	78.3	94.1
GPT-4 (нулевой выстрел)	65.3	79.8	71.2	88.9
ГПТ-4 + РАГ	76.8	90.3	83.7	95.4

Результаты выявляют несколько важных и теоретически значимых закономерностей. Классические алгоритмы машинного обучения, хотя и предоставляют полезные базовые показатели, неизменно уступают подходам глубокого обучения по точности на 10–25 процентных пунктов, что подтверждает недостаточность созданных вручную пространств признаков для охвата всего спектра сигналов достоверности, доступных в современных данных социальных сетей. Среди классических подходов наилучшие результаты показывают SVM с RBF-ядрами, что указывает на нелинейный характер границы классификации фейковых новостей в пространстве признаков TF-IDF.

Среди подходов глубокого обучения, тонко настроенный BERT стабильно превосходит модели BiLSTM на 5–8 процентных пунктов, подтверждая превосходство двунаправленных контекстных представлений над последовательным кодированием. Адаптированный к предметной области FakeBERT дополнительно улучшает BERT на 2–4 пункта, подчеркивая ценность предварительного обучения, специфичного для предметной области, даже когда уже доступно предварительное обучение для общей предметной области. Графовые методы (Bi-GCN, FANG) демонстрируют особенно высокую производительность на наборах данных с богатым социальным контекстом — FakeNewsNet и PHEME, — где структура распространения предоставляет дополнительные доказательства к текстовому контенту, но предлагают меньшие преимущества на тестах, содержащих только текст.

Мультимодальные подходы достигают наивысшей точности на наборах данных, содержащих изображения, что подтверждает взаимодополняемость визуальной и текстовой модальностей. GPT-4 в условиях нулевого обучения показывает худшие результаты по сравнению с более мелкими, точно настроенными моделями, что отражает сложность точной классификации претензий без специфического для задачи обучающего сигнала; однако при дополнении данными поиска (RAG) GPT-4 достигает наивысшей точности на трех из четырех эталонных наборов данных, что указывает на то, что опора на полученные данные является критическим фактором.

Результаты междоменной оценки показывают устойчивое и существенное снижение производительности: показатель F1 падает на 15–30 процентных пунктов, когда модели обучаются в одной области и оцениваются в другой. Этот сдвиг в предметной области отражает реальные различия в стиле письма, распределении лексики, тематической направленности и характере сигналов достоверности в разных наборах данных и представляет собой наиболее существенное с практической точки зрения ограничение современных передовых систем.

Оценка устойчивости к атакам с заменой слов, сгенерированными TextFooler, показала, что все нейронные модели продемонстрировали снижение точности на 12–28 процентных пунктов в условиях атаки, при этом более крупные модели, как правило, показали большую устойчивость, но ни одна из них не была полностью защищена от атаки. Эти результаты подчеркивают необходимость разработки устойчивых к атакам процедур обучения как необходимого условия для внедрения в производство.

Конвейер фильтрации в реальном времени.

Внедрение систем обнаружения дезинформации в масштабах социальных сетей в производственной среде накладывает инженерные ограничения, отсутствующие при оценке в офлайн-режиме. Объём контента на основных платформах превышает миллионы сообщений в час, что накладывает жёсткие требования к задержке и пропускной способности. Эффективный конвейер обнаружения в реальном времени должен завершать классификацию в течение нескольких секунд после публикации, чтобы обеспечить своевременное вмешательство до того, как контент достигнет широкой аудитории.

Предлагаемая архитектура состоит из пяти последовательных этапов. Слой приема использует Apache Kafka для приема и буферизации потоков контента из API платформы, обеспечивая надежную, упорядоченную доставку с настраиваемой пропускной способностью и отказоустойчивостью. Модуль дедупликации применяет локально-чувствительное хеширование для идентификации и

фильтрации почти дублирующегося контента, исключая избыточную классификацию вирусных репостов. Слой извлечения признаков работает параллельно с текстовыми (токенизация и вывод BERT), графическими (извлечение признаков ResNet) и метаданными (проверка достоверности источника, разрешение URL-адресов) модальностями. Классификация выполняется обученной ансамблевой моделью, при этом оценки достоверности сообщаются вместе с бинарными или многоклассовыми прогнозами. Модуль постобработки применяет настраиваемые пороговые значения достоверности для направления прогнозов на последующие действия: элементы с низкой достоверностью ставятся в очередь на проверку человеком; обнаружение элементов с высокой достоверностью запускает уменьшенное алгоритмическое усиление, добавление информационных меток или передачу партнерам по проверке фактов.

Методы сжатия моделей необходимы для соблюдения требований к задержке без необходимости выделения выделенной инфраструктуры GPU для каждого развернутого экземпляра. Дистилляция знаний от полноразмерного учителя BERT к ученику DistilBERT сокращает количество параметров на 40% и задержку вывода на 60%, сохраняя при этом 97% производительности классификации. Квантование весов модели в формате INT8 дополнительно уменьшает объем используемой памяти и ускоряет вывод на стандартном оборудовании. Семантическое кэширование с использованием плотного векторного поиска позволяет повторно использовать результаты классификации для контента, семантически похожего на ранее проверенные элементы, существенно снижая среднюю стоимость вывода при реалистичном распределении трафика.

Проблемы и ограничения.

Несмотря на значительный технический прогресс, ряд фундаментальных проблем ограничивает практическую эффективность современных систем обнаружения. Наиболее сложной задачей, пожалуй, является противодействие злоумышленникам: опытные злоумышленники, изучающие развернутые системы обнаружения, могут итеративно изменять ложный контент, чтобы избежать классификации, создавая непрекращающуюся гонку вооружений между обнаружением и уклонением. Возмущения на уровне символов (замена омоглифов, вставка символов нулевой ширины), замены на уровне слов и уклонение на основе перефразирования могут надежно обмануть современные классификаторы, не изменяя воспринимаемого смысла ложных утверждений.

Нестационарный характер информационного ландшафта создает аналогичную проблему. Постоянно появляются новые ложные нарративы, использование текущих событий в корыстных целях быстро развивается, а

языковые стили со временем адаптируются к нормам, специфичным для каждой платформы. Модели, обученные на исторических данных, демонстрируют предсказуемое снижение производительности по мере изменения распределения данных, что требует постоянного мониторинга, периодического переобучения и активных конвейеров обучения для поддержания эффективности на протяжении всего срока службы.

Многоязычный и межкультурный охват представляет собой критический пробел в существующих системах. Подавляющее большинство исследований сосредоточено на англоязычном контенте, однако дезинформация — это глобальное явление, затрагивающее сообщества на всех языках. Многоязычные модели, включая mBERT и XLM-RoBERTa, предоставляют отправную точку, но неизменно показывают худшие результаты, чем моноязычные модели на языках с ограниченными ресурсами, а культурный контекст, необходимый для понимания сатиры, иронии и политических аллюзий, трудно закодировать в универсальных векторных представлениях.

Этические и социальные аспекты одинаково важны. Автоматизированная модерация контента в больших масштабах концентрирует значительную власть над доступом к информации в руках операторов платформ и их подрядчиков-поставщиков технологий. Исследования задокументировали систематические различия в точности обнаружения в зависимости от демографических групп, политической принадлежности и географических регионов, при этом исторически маргинализированные сообщества непропорционально чаще подвергаются ложным срабатываниям. Прозрачность, подотчетность и значимый человеческий контроль являются важнейшими требованиями к проектированию любой ответственно внедренной системы, однако они вводят операционные издержки и задержки, которые могут противоречить требованиям скорости эффективного раннего вмешательства.

Заключение.

В данной статье представлен систематический обзор и сравнительная оценка алгоритмов обнаружения и фильтрации дезинформации в интернет-сетях, охватывающие классическое машинное обучение, глубокое обучение, моделирование распространения на основе графов и многомодальные системы обнаружения. Экспериментальные результаты подтверждают, что гибридные системы, интегрирующие понимание языка на основе BERT с моделированием социального контекста на основе графов, демонстрируют наилучшие результаты на тестах социальных сетей, в то время как модели обработки больших языков с расширенными возможностями поиска представляют собой современный рубеж в проверке фактов на основе знаний.

Анализ также выявляет критические ограничения, которые необходимо устранить, прежде чем автоматическое обнаружение можно будет ответственно внедрить в масштабах всей системы: проблема постоянного смещения доменов, уязвимость для атак со стороны злоумышленников, недостаточное многоязычное покрытие и этические проблемы автоматической модерации контента. Это не чисто технические проблемы; они требуют междисциплинарного сотрудничества между специалистами в области информатики, социологии, права и государственной политики.

Ставки высоки. Дезинформация представляет собой документально подтвержденную, измеримую угрозу для общественного здравоохранения, демократического управления и социального доверия. Одних лишь технических решений недостаточно — они должны быть интегрированы в институциональные рамки, обеспечивающие прозрачность, подотчетность и защиту законного выражения мнений. Научное сообщество несет ответственность не только за совершенствование технических возможностей систем обнаружения, но и за строгость оценки их воздействия на общество, а также за честность в информировании об их ограничениях.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ:

1. Qodirov, Farrux, and Sabrina Turayeva. "IOT (INTERNET OF THINGS) ORQALI SANOAT ENERGIYA SAMARADORLIGINI OSHIRISH." Общественные науки в современном мире: теоретические и практические исследования 4.7 (2025): 75-83.

2. Qodirov, Farrux, and Husniya Ergasheva. "INVESTITSIYALARNI JALB QILISH VA UNING SAMARADORLIGI." Общественные науки в современном мире: теоретические и практические исследования 3 (2024): 64-69.

3. Qodirov, F., N. Sirojev, and S. Negmatova. "Features of the Android Studio software package." Академические исследования в современной науке 2.17 (2023): 130-146.

4. Ergash o'g'li, Qodirov Farrux. "Econometric modeling of the development of medical services to the population of the region/Berlin Studies Transnational Journal of Science and Humanities." (2022): 1-1.

5. Кодиров, Ф. Э., and О. Д. Дониёров. "ЭФФЕКТИВНЫЕ МОДЕЛИ РАЗВИТИЯ МЕДИЦИНСКОГО ОБСЛУЖИВАНИЯ НАСЕЛЕНИЯ КАШАКАДЬИНСКОЙ ОБЛАСТИ." Символ науки 7-2 (2022): 15-17.

6. Қодиров, Ф. "Вилоят аҳолисига соғлиқни сақлаш хизматлари кўрсатиш тармоқлари ривожланиш механизмининг статистик таҳлили." Andijon Mashinasozlik Instituti (2022).

7. Қодиров, Ф. "Қашқадарё вилояти аҳолисига тиббий хизмат кўрсатиш тармоқларини ривожлантиришнинг истиқболлари"." O ‘ZBEKISTON QISHLOQ VA SUV XO ‘JALIGI" âà" AGRO ILM." o ‘zbekiston qishloq va suv xo ‘jaligi» âà «Agro ilm (2022).

8. Қодиров, Ф. "" ҲУДУДЛАРДА ТИББИЙ ХИЗМАТ КЎРСАТИШНИ ЭКОНОМЕТРИК МОДЕЛЛАШТИРИШ". ХОРАЗМ МАЪМУН АКАДЕМИЯСИ АХБОРОТНОМАСИ." Хоразм маъмун академияси ахборотномаси (2022).

9. Қодиров, Ф. "" АҲОЛИГА ТИББИЙ ХИЗМАТ КЎРСАТИШ СОҲАСИНИНГ КЕЛГУСИ ҲОЛАТИНИ БАШОРАТЛАШ". Самарқанд иқтисодиѐт ва сервис институти." Самарқанд иқтисодиѐт ва сервис институти (2022).

10. Qodirov, F. "" Қашқадарё ҳудуди аҳолисига хизмат кўрсатиш тармоқлари ва уларга таъсир этувчи омиллар"." O ‘zbekiston Qishloq Va Suv xo ‘jaligi" Jurnali." O ‘zbekiston Qishloq Va Suv xo ‘jaligi” Jurnali (2022).

11. Qodirov, F. "" OPTIMUM SOLUTIONS FOR THE DEVELOPMENT OF MEDICAL SERVICES IN PRIVATE CLINICS". MUHAMMAD AL-XORAZMIY NOMIDAGI TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETI QARSHI FILIALI." (2022).

12. Qodirov, F. "" QR-KOD TEXNOLOGIYASI ASOSIDA ELEKTRON KUTUBXONA TIZIMINI DASTURIY VA APPARAT TAMINOTINI YARATISH". MUHAMMAD AL-XORAZMIY NOMIDAGI TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETI QARSHI FILIALI." (2021).

13. Qodirov, F. E., O. D. Doniyorov, and H. Shokirov Sh. "Basic Concepts Of Information Security In Information Systems. Wide Threats And Their Consequences." КОНЦЕПЦИИ УСТОЙЧИВОГО РАЗВИТИЯ НАУКИ В СОВРЕМЕННЫХ УСЛОВИЯХ (2021): 153-155.

14. Bozorova, Irina Jumanazarovna, and Dilfuzaxon Mamasharipovna Karayeva. "Modern programming technologies and their role." интеллектуальный капитал xxi века. 2020.

15. Kodirov, F. E., and J. E. Nematov. "BASIC TECHNOLOGY AND SERVICE MANAGEMENTMULTISERVICE NETWORKS." Инновации в технологиях и образовании: сб. ст. участников XII Между (2019): 214.

16. Qodirov, F. E., et al. "PROBLEMS AND SOLUTIONS FOR EFFECTIVE PROTECTION AGAINST NETWORK ATTACKS." НАУКОЕМКИЕ ИССЛЕДОВАНИЯ КАК ОСНОВА ИННОВАЦИОННОГО РАЗВИТИЯ 93 (2019).

17. Qodirov, F. E., J. U. Abdirasulov, and J. E. Nematov. "FORMING GOVERNMENT AGENCY WEBSITES WITH WORDPRESS CONTENT MANAGEMENT SYSTEM." Инновации в технологиях и образовании: сб. ст. участников XII Между (2019): 219.

18. Qodirov, Farrux, and Mashxura Sa’dullayeva. "virtual reallik (vr) va kengaytirilgan reallik (AR)." Молодые ученые 3.8 (2025): 139-144.

19. Qodirov, F., and J. Murodulloyeva. "O'ZBEKISTONDA RAQAMLI IQTISODIYOT." Инновационные исследования в современном мире: теория и практика 3.15 (2024): 178-181.

20. Qodirov, F. E. "Hududlarni ijtimoiy-iqtisodiy rivojlantirishda har bir hududning o ‘ziga xos xususiyatlari." AKTUAR MOLIYA VA BUXGALTERIYA HISOBI ILMIY JURNALI 4.09 (2024): 178-183.

21.Қодиров, Ф. "ҲУДУДЛАРДА ТИББИЙ ХИЗМАТЛАРНИ ДАСТУРИЙ ПАКЕТЛАР ЁРДАМИДА ЭЛЕКТРОН ТИББИЙ БАЗАСИНИ ЯРАТИШ." O’zbekiston Respublikasi Oliy Va o’rta Maxsus ta’lim Vazirligi Namangan Muhandislik-Qurilish Instituti (2022).

22. Jumanazarovna, Bozorova Irina, and Kodirov Farruh Ergash O'G'Li. "Principle of electrocardiographic work and its role in modern medicine." Вопросы науки и образования 15 (99) (2020): 31-36.

23. Қодиров, Ф. "" СОЗДАНИЕ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ И АППАРАТА ЭЛЕКТРОННОЙ БИБЛИОТЕЧНОЙ СИСТЕМЫ НА ОСНОВЕ QR-КОДОВОЙ ТЕХНОЛОГИИ". Kokand University." Kokand University (2020).

24. Кодиров, Ф. "" АНАЛИЗ БИОСИГНАЛОВ В ЭЛЕКТРОКАРДИОГРАФИИ И МЕТОДЫ ИХ ОБРАБОТКИ". МУҲАММАД АЛ-ХОРАЗМИЙ НОМИДАГИ ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ ҚАРШИ ФИЛИАЛИ." МУҲАММАД АЛ-ХОРАЗМИЙ НОМИДАГИ ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ ҚАРШИ ФИЛИАЛИ (2020).

25. Qodirov, F. "" MASOFAVIY TA'LIMDA O'QISHNING QULAYLIKLARI VA KAMCHILIKLARI". МУҲАММАД АЛ-ХОРАЗМИЙ НОМИДАГИ ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ ҚАРШИ ФИЛИАЛИ." (2020).

26. Қодиров, Ф. Э., et al. "Компьютерные игры и их текущие виды и преимущества." ТЕОРИЯ И ПРАКТИКА МОДЕРНИЗАЦИИ НАУЧНОЙ ДЕЯТЕЛЬНОСТИ. 2019.

27. Қодиров, Ф. Э., et al. "ДЛЯ ПРОВЕРКИ МОДЕЛЕЙ АДЕКВАТНОСТИ, ЧУВСТВИТЕЛЬНОСТЬ И СОПРОТИВЛЕНИЯ." ИНТЕГРАЦИЯ НАУКИ, ОБЩЕСТВА, ПРОИЗВОДСТВА И ПРОМЫШЛЕННОСТИ. 2019.

28. Қодиров, Ф. Э., and Ж. Э. Нематов. "РАЗВИТИЕ ЛОКАЛЬНОЙ СЕТИ НА ОСНОВЕ ТЕХНОЛОГИИ GPON." Инновации в технологиях и образовании: сб. ст. участников XII Между (2019): 288.

29. Кодиров, Ф. Э., and М. У. Маматмурадова. "РАЗРАБОТКА ЦИФРОВОЙ ПРОГРАММЫ ШИФРОВАНИЯ И ВНЕДРЕНИЕ В ПРАКТИКУ." Инновации в технологиях и образовании: сб. ст. участников XII Между (2019): 275.

30. Абдирасулов, Ж. У., and Ф. Э. Кодиров. "ЭФФЕКТИВНОСТЬ ANGULAR JS ДЛЯ СОЗДАНИЯ ДИНАМИЧЕСКИХ ВЕБ-САЙТОВ И ОПТИМИЗАЦИИ ИХ ПРОИЗВОДИТЕЛЬНОСТИ." Инновации в технологиях и образовании: сб. ст. участников XII Между (2019): 228.

31. Кодиров, Ф. "" ЗАМОНАВИЙ КОМПЬЮ ТЕР УЙИНЛАРИ ВА УЛАРНИНГ СИНФЛАНИШИ". МУХАММАД АЛ-ХОРАЗМИЙ НОМИДАГИ ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ КАРШИ ФИЛИАЛИ." (2019).

32. Турдиев, У. К., and Ф. Э. Кодиров. "Задача Коши Для Одномерной Системы Уравнений Типа Бюргерса Возникающей В Двухскоростной Гидродинамике." Инновации в технологиях и образовании: сб. ст. участников XI Между (2018): 349.

33. Kubayev, Ulugbek, et al. "Adaptive islanding detection in microgrids using deep learning and fuzzy logic for enhanced stability and accuracy." Journal of Operation and Automation in Power Engineering 12.Special Issue (Open) (2024): 33-42.

34. Qodirov, F. E., D. A. Akbarova, and S. H. Shokirov. "SOFTWARE FOR WORKING WITH COMPUTER GRAPHICS AND THEIR TASKS. APPLICATION OF DIGITAL IMAGE PROCESSING FIELDS." (2021): 57-58.

35. Kodirov, Farrukh Ergashevich, and Sitorabonu Zoxidjonova Axmatova. "LiFi-NEW NETWORK TECHNOLOGIES." НАУКА И ИННОВАЦИИ В XXI ВЕКЕ: АКТУАЛЬНЫЕ ВОПРОСЫ, ОТКРЫТИЯ И ДОСТИЖЕНИЯ. 2019.

36. Маматмурадова, М. У., И. Ж. Бозорова, and Ф. Э. Кодиров. "Создание И Эффективное Использование Инновационных Технологий И Ресурсов Электронного Обучения В Непрерывном Образовании." Инновации в технологиях и образовании. 2019.

37. Qodirov, F. E., et al. "OVER VIEW FROM YII 2 FRAMEWORKS, AND ALSO ITS ADVANTAGES AND DISADVANTAGES." СОВЕРШЕНСТВОВАНИЕ МЕТОДОЛОГИИ ПОЗНАНИЯ В ЦЕЛЯХ РАЗВИТИЯ НАУКИ 39 (2019).

38. Qodirov, Farrux. "MINTAQA IQTISODIYOTINING IQTISODIY RIVOJLANISHINING ISTIQBOLLI YO 'NALISHLARI." MUHANDISLIK VA IQTISODIYOT 3.12 (2025).

39. Qodirov, Farrux. "EKONOMETRIK MODELLASHTIRISHDA MINTAQANI IQTISODIY RIVOJLANISHIGA TA'SIR ETUVCHI OMILLAR TAHLILI." MUHANDISLIK VA IQTISODIYOT 3.10 (2025).

40. Qodirov, Farrux, and Anora Allanazarova. "TA'LIMNI BOSHQARISH TIZIMLARI TASNIFI." Central Asian Journal of Multidisciplinary Research and Management Studies 2.11 (2025): 113-117.

41.Qodirov, Farrux. "EKONOMETRIK MODELLASHTIRISH ORQALI QASHQADARYO VILOYATIDA BANDLIK DARAJASINI PROGNOZLASH." Central Asian Journal of Multidisciplinary Research and Management Studies 2.9 (2025): 113-115.