

*Новиков Иван Владимирович, магистрант, Белгородский государственный
национальный исследовательский университет, Россия, г. Белгород*

**РАЗРАБОТКА МНОГОАГЕНТНОГО АЛГОРИТМА ВЕРИФИКАЦИИ
НЕДОСТОВЕРНЫХ НОВОСТЕЙ С АГРЕГАЦИЕЙ ПРИЗНАКОВ
МОДЕЛЮ-АРБИТРОМ**

АННОТАЦИЯ

В статье рассматривается проблема автоматизации выявления недостоверных новостей в условиях стремительного роста цифрового контента и высокой скорости его распространения через интернет-СМИ и социальные платформы. Предлагается многоагентный алгоритм верификации, построенный по принципу кооперации автономных программных агентов, каждый из которых специализируется на определённом аспекте анализа контента. Центральным координирующим элементом выступает Арбитр, выполняющий оркестрацию потоков информации и финальную агрегацию признаков с помощью модели CatBoost. Архитектура включает OCR-агент для извлечения текста из графических вложений, агент лингвостилистического анализа на основе BERT, детектор машинной генерации текста, анализатор фактологических противоречий и модуль оценки кликбейтных заголовков. Параллельная работа агентов обеспечивает высокую скорость обработки, а применение мета-классификатора на финальном этапе гарантирует интерпретируемость результатов. Разработанный алгоритм позволяет предоставлять пользователю не только вероятностную оценку достоверности, но и детализированное обоснование выявленных аномалий, что соответствует современным требованиям к системам автоматизированного фактчекинга.

ABSTRACT

This paper addresses the problem of automated fake news detection in the context of rapid digital content growth and high-speed information dissemination via

online media and social platforms. A multi-agent verification algorithm is proposed, built on the principle of cooperation between autonomous software agents, each specializing in a specific aspect of content analysis. The central coordinating element is an Arbiter, which orchestrates information flows and performs final feature aggregation using a CatBoost model. The architecture includes an OCR agent for text extraction from images, a BERT-based linguistic-stylistic analysis agent, a machine-generated text detector, a fact-checking module, and a clickbait assessment module. Parallel agent operation ensures high processing speed, while the meta-classifier at the final stage guarantees interpretability of results. The developed algorithm provides users not only with a probabilistic veracity score but also with a detailed justification of detected anomalies, meeting modern requirements for automated fact-checking systems.

Ключевые слова: многоагентная архитектура, выявление фейковых новостей, верификация контента, модель-арбитр, CatBoost, обработка естественного языка

Keywords: multi-agent architecture, fake news detection, content verification, arbiter model, CatBoost, natural language processing

Современное информационное пространство характеризуется лавинообразным ростом объёмов цифрового контента, значительная часть которого распространяется через социальные сети и интернет-СМИ. Наряду с оперативным информированием общества эта среда создаёт благоприятные условия для генерации и вирусного распространения недостоверных, искажённых или полностью вымышленных сообщений. Фейковые новости способны оказывать существенное влияние на общественное мнение, формировать ложные представления о событиях, а в условиях политической или социальной нестабильности — провоцировать напряжённость. Традиционные методы ручной верификации не успевают за скоростью распространения контента, что делает актуальной задачу автоматизации выявления недостоверных

новостей с использованием методов машинного обучения и обработки естественного языка.

Предлагаемый в работе подход базируется на модульной архитектуре, спроектированной по принципу кооперации автономных программных агентов, каждый из которых специализируется на определённом аспекте верификации контента и обладает собственной изолированной логикой обработки данных. В основу построения заложена концепция многоагентного взаимодействия, где центральным координирующим элементом выступает Арбитр. Модуль управляет потоками информации, пользовательскими сессиями через асинхронный чат-бот и агрегирует признаки для итогового вердикта. Процесс верификации инициируется при получении входных данных, которые могут быть представлены как в текстовом виде, так и в формате графических изображений или скриншотов новостных лент. Для обеспечения универсальности входного интерфейса в архитектуру интегрирован агент оптического распознавания символов (OCR).

Внутренний цикл обработки контента реализуется через параллельное функционирование специализированных агентов-аналитиков, использующих современные методы глубокого обучения и естественного языка. Агент лингвостилистического анализа производит оценку эмоциональной валентности сообщения с помощью дообученной нейросетевой модели семейства BERT, выявляя аномально высокий уровень экспрессии и наличие специфических лексических маркеров манипуляции, таких как кликбейтные заголовки или скрытые призывы к действию. Одновременно с этим агент детекции машинной генерации выполняет статистический мониторинг структуры текста на предмет выявления характерных признаков работы больших языковых моделей (LLM). Данная проверка основывается на анализе распределения вероятностей токенов и выявлении специфической «поверхностной глубины», которая в современных реалиях медиапространства является значимым индикатором синтетического происхождения новости.

Завершающий этап работы алгоритма связан с функционированием агента-Арбитра, который аккумулирует структурированные выводы всех аналитических модулей. Итоговое решение о степени достоверности контента формируется на основе мета-классификатора, использующего взвешенную оценку вклада каждого агента в общий результат. Применение подобной стратегии агрегирования позволяет реализовать принципы интерпретируемого искусственного интеллекта, предоставляя пользователю не только вероятностную оценку «фейк» или «не фейк», но и детализированное обоснование, отражающее конкретные аномалии, выявленные в ходе лингвистического и логического анализа. Графическое представление архитектуры представлено на рисунке 1.

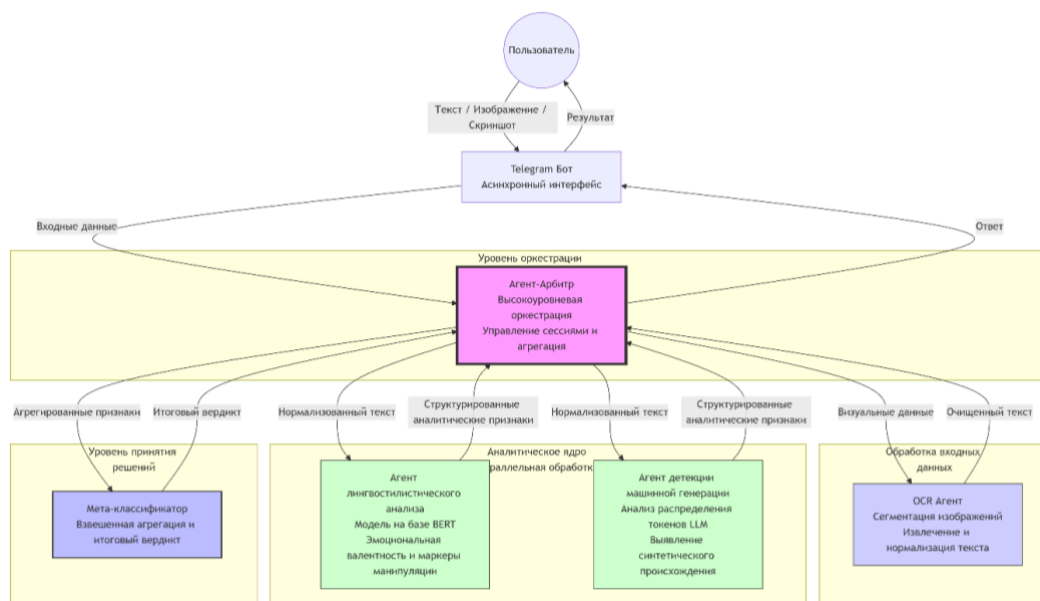


Рисунок 1. Графическое представление архитектуры

Представленная иерархическая модель демонстрирует разделение зон ответственности между интерфейсным слоем, уровнем оркестрации и аналитическим ядром. Ключевой особенностью является параллелизм работы агентов-аналитиков, что позволяет минимизировать общее время инференса. Использование мета-классификатора на финальном этапе гарантирует адаптивность к различным типам входных данных.

В рамках практического этапа исследования был спроектирован, реализован и подвергнут комплексному тестированию многоагентный алгоритм, архитектура которого базируется на принципах децентрализованной обработки данных и финальной агрегации признаков специализированной моделью-арбитром.

Техническая реализация исследования требует внедрения многоуровневого конвейера обработки информации, начинающегося с модуля оптического распознавания символов (OCR) для извлечения текстовых данных из графических вложений, что позволит анализировать не только прямые сообщения, но и скриншоты публикаций из социальных сетей. Ключевой особенностью разработанного алгоритма стало функционирование пяти автономных агентов: модуля сентимент-анализа для оценки эмоциональной экспрессивности, детектора машинной генерации текста, поисковика фактологических противоречий, анализатора кликбейтных заголовков и модуля оценки структуры текста. Каждый из этих компонентов генерирует собственный вектор признаков, который передается на вход финальному классификатору – модели-арбитру. Такой подход позволил в ходе исследования детально изучить вклад каждого отдельного параметра в формирование итогового вердикта о достоверности, обеспечивая при этом высокую интерпретируемость, так как итоговый индекс подозрительности формируется на основе прозрачного взвешивания результатов работы всех вовлеченных агентов. Алгоритм выявления недостоверных новостей в средствах массовой информации представлен на рисунке 2.

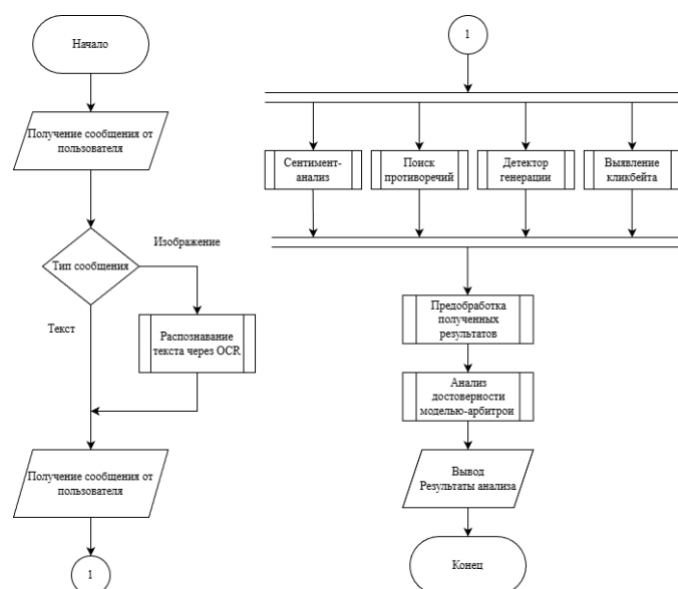


Рисунок 2. Алгоритм выявления недостоверных новостей в СМИ

При получении запроса объект управления (диспетчер) осуществляет первичную декомпозицию данных и активирует параллельный цикл опроса пула интеллектуальных агентов. На этом этапе взаимодействие объектов характеризуется асинхронной передачей информационных сигналов: каждый агент, будь то модуль сентимент-анализа или детектор фактологических противоречий, получает независимую копию данных для обработки. Взаимодействие между агентами в данной архитектуре исключено намеренно, чтобы обеспечить чистоту экспертной оценки и избежать взаимного влияния алгоритмов на промежуточные результаты. Основной поток управления на данной фазе сосредоточен на синхронизации ответов: диспетчер ожидает завершения работы всех активных модулей, после чего формирует агрегированный вектор признаков, инкапсулирующий в себе частные вердикты и числовые метрики, полученные от каждого функционального узла.

Заключительная фаза взаимодействия объектов связана с передачей сформированного вектора признаков на вход модели-арбитра, которая выполняет роль центрального процессора принятия решений. Взаимодействие между арбитром и нижестоящими агентами носит иерархический характер: арбитр сопоставляет поступившие сигналы с весами, определенными в ходе предварительного обучения, и разрешает возможные конфликты между

экспертными оценками. После завершения этапа классификации объект-арбитр транслирует финальный результат обратно диспетчеру, который инициирует построение визуального отчета для интерфейсного объекта.

В ходе проектирования алгоритма была разработана модульная многоагентная архитектура, обеспечивающая комплексную проверку медиаконтента через параллельное функционирование специализированных аналитических модулей. Предложенный алгоритм, объединяющий лингвистический анализ, детекцию машинной генерации и поиск логических противоречий под управлением модели-арбитра CatBoost, позволяет достичь высокой интерпретируемости и точности выносимых вердиктов. Реализация алгоритма в виде асинхронного чат-бота с поддержкой мультимодального ввода соответствует современным требованиям к производительности и удобству верификации информации в реальном времени.

Литература

1. Алюшин А. В. Многоагентный подход к анализу нейросетей большой размерности // Естественные и технические науки. – 2010. – № 1(45). – С. 313-315.
2. Асирян А. К. Оценка тональности новостных сообщений // Научный взгляд в будущее. – 2018. – Т. 1, № 9. – С. 27-34.
3. Гончаренко А. И. Разработка программного комплекса, выявляющего фейковые новости // Прикладная математика и фундаментальная информатика. – 2022. – Т. 9, № 1. – С. 32-37
4. Соболев А. Анализ тональности новостей о криптовалюте при помощи модели BERT // ГРААЛЬ НАУКИ. – 2021. – Р. 272-276. – DOI 10.36074/grail-of-science.19.11.2021.051.