Научная статья
УДК 004.9; DOI: 10.61260/2218-13X-2025-1-160-173

АЛГОРИТМ ДЛЯ РЕАЛИЗАЦИИ МЕТОДА КОМПЛЕКСИРОВАННОЙ ОБРАБОТКИ ДАННЫХ С ЦЕЛЬЮ ФОРМИРОВАНИЯ И ПРЕДОСТАВЛЕНИЯ ДОСТОВЕРНОЙ ИНФОРМАЦИИ

✉Карманова Наталия Андреевна.
Национальный исследовательский университет ИТМО, Санкт-Петербург, Россия
✉karmanova.ifmo@gmail.com

Аннотация. Представлен разработанный алгоритм комплексированной обработки данных, направленный на формирование достоверного информационного потока из исходных данных, поступающих от разнородных и неоднородных по надёжности источников, и его программная реализация. Алгоритм базируется на концепции адаптивной оценки надёжности данных и их источников с учётом уровня согласованности, степени конфликтности и подтверждённости информации.

В современных условиях интенсивного развития цифровых технологий и увеличения объёма обрабатываемой информации особую актуальность приобретает задача формирования достоверного информационного потока из разнородных источников данных, которые могут различаться по уровню надёжности, структурным характеристикам и степени согласованности предоставляемой информации.

Предлагаемый алгоритм комплексированной обработки данных основывается на концепции адаптивной оценки надёжности как самих данных, так и их источников, с учётом таких критериев, как уровень согласованности, степень конфликтности и подтверждённости информации. Разработанный алгоритм направлен на обеспечение объективной фильтрации, интеграции и оптимального использования поступающих данных.

Ключевые слова: достоверная информация, ложная информация, вредоносная информация, сложный информационный поток, надежность источников, согласованность данных, степень конфликтности

Для цитирования: Карманова Н.А. Алгоритм для реализации метода комплексированной обработки данных с целью формирования и предоставления достоверной информации // Науч.-аналит. журн. «Вестник С.-Петерб. ун-та ГПС МЧС России». 2025. № 1. С. 160–173. DOI: 10.61260/2218-13X-2025-1-160-173.

Scientific article

ALGORITHM FOR REALIZATION OF THE METHOD OF COMPLEX DATA PROCESSING IN ORDER TO FORM AND PROVIDE RELIABLE INFORMATION

✉Karmanova Natalia A.
ITMO University, Saint-Petersburg, Russia
✉karmanova.ifmo@gmail.com

Abstract. This paper presents the developed algorithm of complex data processing, aimed at forming a reliable information flow from the initial data coming from heterogeneous and heterogeneous sources in terms of reliability and its software implementation. The algorithm is based on the concept of adaptive assessment of reliability of data and their sources, considering the level of consistency, the degree of conflict and confirmation of information.

In modern conditions of intensive development of digital technologies and increase in the volume of processed information the task of forming a reliable information flow from heterogeneous data sources, which may differ in the level of reliability, structural characteristics and the degree of consistency of the information provided, is of relevance.

© Санкт-Петербургский университет ГПС МЧС России, 2025

The proposed algorithm of complexized data processing is based on the concept of adaptive assessment of reliability of both the data and their sources, considering such criteria as the level of consistency, the degree of conflict and confirmation of information. The developed algorithm is aimed at ensuring objective filtering, integration and optimal use of incoming data.

Keywords: reliable information, false information, malicious information, complex information flow, source reliability, data consistency, degree of conflict

For citation: Karmanova N.A. Algorithm for realization of the method of complex data processing in order to form and provide reliable information // Scientific and analytical journal «Vestnik Saint-Petersburg university of State fire service of EMERCOM of Russia». 2025. № 1. P. 160–173. DOI: 10.61260/2218-13X-2025-1-160-173.

Введение

Современные информационные потоки характеризуются высоким уровнем сложности, включающим присутствие значительного количества ложной и вредоносной информации, что создает угрозу не только информационной безопасности, но и функционированию различных систем управления. В связи с этим актуальной становится задача обеспечения информационной безопасности еще до передачи данных в систему обработки информации. Выявление и классификация сообщений, основанных на их достоверности и деструктивности (например, ложные или вредоносные), представляет собой сложную задачу из-за многокомпонентного характера информации, изменчивости контекста, а также динамического взаимодействия между различными типами данных. В текущих условиях распространения ложной информации через цифровые каналы требуется разработка эффективных математических и вычислительных инструментов для анализа, классификации и управления информационными потоками, направленными на повышение надежности систем. Непроведение предварительной проверки информационных потоков до их попадания в систему является недостатком, который подчёркивается в ряде исследований, рассматривающих распространение ложной или вредоносной информации. Подобные упущения можно отметить, например, в исследованиях [1–5]. Однако увеличение количества каналов передачи информации и их разнородность представляет собой как вызов, так и уникальную возможность. Использование разнородных данных от n источников позволяет применить метод их комплексирования, целью которого становится нивелирование ошибок, устранение противоречий и фильтрация ложных или вредоносных сигналов.

Методы исследования

С целью борьбы с распространением ложной и вредоносной информации, а также для повышения достоверности входных данных представлен метод их комплексирования, направленный на повышение достоверности информации в условиях избыточности, разнородности и противоречивости потоков данных из множества источников. Метод комплексирования данных можно описать как процесс объединения информации от n различных источников или каналов с целью создания единого информационного потока повышенной достоверности. Такой метод позволяет снизить влияние ошибок, выбросов и ложных данных за счет перекрестной верификации данных, поступающих из разных источников.

Метод основан на вероятностных подходах, которые учитывают надёжность источников, согласованность данных и их характеристики, такие как временные метки и контекст. Построена модель взвешенного комбинирования вероятностей, позволяющая минимизировать влияние дезинформации через механизм перекрёстной верификации сообщений. Вводится динамическая адаптация весов источников на основе их исторической точности и согласованности данных, что позволяет системе адаптироваться к изменяющимся

условиям. Описаны методы активного выявления ложных и вредоносных сигналов, включая анализ метаданных, временной согласованности и аномалий. Применение данных подходов обеспечивает повышение устойчивости системы к информационным угрозам и улучшение качества итогового информационного потока для дальнейшего анализа.

В ходе работы был разработан алгоритм, выражающий последовательность действий при реализации метода комплексированной обработки данных. Он направлен не только на выявление достоверных данных, но и на детектирование ложных и вредоносных потоков, что минимизирует риски использования искажённой информации. Новизна предлагаемого алгоритма заключается в его способности одновременно учитывать множество факторов (надёжность, согласованность, конфликтность и подтвержденность), динамически адаптировать параметры в процессе работы и классифицировать данные по трём уровням. За счёт таких особенностей он существенно превосходит существующие аналоги по точности, устойчивости и универсальности, что в дальнейшем подтверждается результатами экспериментальных исследований [1–3] (табл. 1).

Таблица 1

Сравнение с другими алгоритмами

Характеристика	Существующие алгоритмы	Предложенный алгоритм
Надёжность источников	Фиксированная, редко обновляемая	Динамическая, адаптируемая на каждом этапе анализа
Многокритериальный анализ	Учитывается частично, чаще 1–2 критерия	Оценка по согласованности, конфликту, подтверждению и др.
Трёхуровневая классификация	Обычно только «достоверные» и «недостоверные»	Классификация на достоверные, ложные и вредоносные данные
Работа в реальном времени	Часто отсутствует	Поддержка поточного реального времени
Адаптация порогов	Жёсткие фиксированные значения	Гибкие пороги, адаптируемые к задаче
Стабильность к атакам	Частичная защита	Высокая устойчивость к искажениям

Блок-схема алгоритма для реализации метода комплексированной обработки данных с целью формирования и предоставления достоверной информации предоставлена на рисунке.

Рис. **Блок-схема алгоритма для реализации метода комплексированной обработки данных с целью формирования и предоставления достоверной информации**

Подробное описание этапов алгоритма представлено далее.

Начало обработки данных от n источников

Этап «Начало» – это первая и очень важная часть алгоритма комплексированной обработки данных. Его основная цель заключается в инициализации процесса, когда начинается сбор и первичная обработка информации от нескольких источников. Здесь происходит подготовка к дальнейшему анализу данных, их классификации и оценке достоверности. В результате выполнения этапа обеспечивается:

– упорядоченный поток данных, организованный в единый формат для последующих шагов алгоритма;

– установленные параметры временных интервалов для анализа;

– подготовленный список активных источников с их текущими характеристиками (включая начальную оценку надёжности);

– сообщения, очищенные от дубликатов, готовые к дальнейшему анализу.

Первичная обработка данных

Первичная обработка данных – это вторая ключевая часть алгоритма, в рамках которой происходит подготовка поступивших данных для их корректного использования на следующих этапах. На данном этапе данные, поступившие из различных источников, проходят проверку, структуризацию и очистку от несоответствий.

Основная цель этапа первичной обработки данных заключается в приведении разнородных данных, полученных от множества источников S_i, к единому формату путём выполнения следующих действий:

– проверка корректности переданных данных;
– фильтрация шумов, некорректных и избыточных данных;
– нормализация данных для обеспечения совместимости с последующими этапами алгоритма;
– дедупликация (удаление дублирующихся сообщений);
– структурирование данных в единый набор, пригодный для дальнейших расчётов.

Уникальные, коррелированные и структурированные сообщения объединяются в подготовленный поток данных, пригодный для объединения и анализа:

$$D_{prosessed} = \{d_1, d_2, ..., d_k\},$$

где k – количество сообщений после фильтрации и нормализации.

Расчёт надёжности источников

Этот этап алгоритма определяет и корректирует уровень надёжности Q_i каждого источника S_i, участвующего в предоставлении данных. Надёжность источников – это важный параметр, который влияет на вес каждой информации, поступающей от источника, при принятии решений на следующих этапах алгоритма.

Рассмотрим детально шаги, выполняемые при оценке надежности источников:

1. Инициализация надёжности источников

На этапе инициализации каждому источнику S_i присваивается начальное значение надёжности Q_i в диапазоне от 0 до 1. Значение Q_i^{init} может быть либо одинаковым для всех источников, либо основываться на априорных данных (например, предыдущем опыте работы с источником).

Значение по умолчанию для всех источников:

$$Q_i = Q_{init}, Q_{init} = 0{,}5,$$

где Q_{init} указывает, что источник нейтрален (50 % надёжности).

Источникам с известной историей надёжности присваиваются индивидуальные начальные значения:

$$Q_i = \frac{T_i^{hist}}{T_i^{hist} + F_i^{hist} + M_i^{hist}},$$

где $T_i^{hist}, F_i^{hist}, M_i^{hist}$ – статистика достоверных, ложных и вредоносных данных в прошлом.

2. Анализ достоверности данных источников

Для каждого источника рассчитываются показатели, характеризующие, насколько достоверна информация, поступающая от него в данном интервале времени.

Для каждого источника Si на основании поступивших сообщений Di определяется количество достоверных сообщений T_i , ложных сообщений F_i , вредоносных сообщений M_i.

Формула для статического расчета надёжности:

$$Q_i = \frac{T_i}{T_i + F_i + M_i}.$$

Если вся информация от источника отсутствует или она признана некорректной:

$$Q_i = 0.$$

Согласованность данных между источником Si и остальными источниками определяется на основе согласия информации [6–9].

Формула для согласованности:

$$K(S_i) = \frac{\sum_{j=1}^{n} Q_j * C_{ij}}{\sum_{j=1}^{n} Q_j},$$

где $C_{ij} = 1$, если источник S_j подтверждает данные S_i; $C_{ij} = 0$, если данные противоречат.

Надёжность источника пересчитывается следующим образом:

$$Q_i^{(t+1)} = \alpha * Q_i^t + (1 - \alpha) * K(S_i),$$

где Q_i^t – значение надёжности источника на предыдущей итерации; $K(S_i)$ – согласованность источника с другими, пересчитанная на текущей итерации; α – параметр сглаживания для уменьшения влияния резких изменений (обычно $\alpha \in [0{,}5;0{,}9]$).

Источники S_i, предоставляющие деструктивные (F_i) или вредоносные (M_i) сообщения, получают значительное снижение Q_i. Это реализуется через штрафной коэффициент:

$$Q_i = Q_i \cdot (1 - \beta \cdot P_{error}),$$

где β – коэффициент штрафа; $P_{error} = \frac{F_i + M_i}{T_i + F_i + M_i}$ – доля ошибок в сообщениях источника.

Если значение Q_i для источника падает ниже определенного порога $Q_{threshhold}$, источник исключается из дальнейшего анализа.

Формирование объединённого потока данных

На этапе формирования объединённого потока данных выполняется интеграция информации, поступающей от разных источников $S_1, S_2, \ldots, S_n$, с учётом их достоверности, определённой в предыдущем этапе (п. 3). На основе оценок надёжности Q_i источников генерируется единый поток данных D_{total}, в котором данные от более надёжных источников имеют больший вес, а информация из менее надёжных источников либо игнорируется, либо снижает своё влияние [10–12].

Проверка данных на подтверждение и противоречие

На данном этапе алгоритм выполняет анализ согласованности объединённого потока данных D_{total}, сформированного на предыдущем шаге.

Сначала каждое сообщение $d \in D_{total}$ содержит уникальные значения ключевых параметров (например, идентификатора объекта или временной метки). На этом этапе данные группируются и анализируются по уникальным параметрам: идентификатор объекта, временные характеристики, пространственные данные (если применимо, например, координаты).

Каждой группе записей, относящихся к одному и тому же событию, объекту или явлению, присваивается идентификатор G_k. Например:

$$G_1:\{d_{11},d_{21},d_{31}\}.$$

Это значит, что данные d_{11}, d_{21}, d_{31} от разных источников относятся к одному событию или объекту [13].

Данные считаются подтверждающими друг друга, если они:

– имеют схожие или идентичные значения ключевых полей (например, одинаковое значение объекта, дата или описание);

– принимаются от надёжных источников (с высокими значениями Q_i).

Суммарный вес подтверждения для каждой группы:

$$W_{confirm}(G_k) = \sum_{i \in S} Q_i \cdot C_{ik} ,$$

где Q_i – надёжность источника, передавшего данные; C_{ik} – бинарный коэффициент согласованности (1 – если данные подтверждены данным источником, 0 – если противоречат).

Если вес $W_{confirm(G_k)}$ превышает пороговое значение $W_{threshold}$, данные группы G_k считаются подтверждёнными.

При необходимости можно вычислить коэффициент согласованности $K_{data}(G_k)$, показывающий степень подтверждения:

$$K_{data}(G_k) = \frac{\text{количество согласованности данных внутри } Gk}{\text{общее количество данных в группе } Gk}.$$

Если $K_{data}(G_k) \approx 1$, данные внутри группы согласованы и имеют высокий уровень доверия.

Противоречия внутри группы G_k оцениваются через:

$$W_{conflict}(G_k) = \sum_{i \in S} Q_i \cdot (1 - C_{ik}),$$

где $1 - C_{ik}$ указывает на конфликты (отсутствие подтверждения).

Если $W_{conflict(G_k)}$ превышает заданный порог $W_{conflict,threshold}$, данные признаются противоречивыми. Такие данные либо исключаются на уровне системы, либо отмечаются для дальнейшего анализа.

Далее происходит анализ изолированных данных (предоставленных только одним источником без подтверждения другими). Для них:

– анализируется надёжность источника Q_i: если источник надежен, изолированные данные могут быть сохранены;

– если $Q_i \leq Q_{trust}$, данные признаются недостоверными и либо исключаются, либо помечаются как подозрительные.

На основании проведённого анализа формируется статистический отчёт о согласованности данных:

– процент согласованных данных:

$$\mathrm{P_{agree}} = \frac{\text{количество подтверждённых данных}}{\text{общее количество данных}} \cdot 100\ \% ;$$

– процент конфликтных данных:

$$P_{conflict} = \frac{\text{количество противоречивых данных}}{\text{общее количество данных}} \cdot 100\ \%\ .$$

Эти метрики могут использоваться для динамического контроля и настройки алгоритмов обработки данных.

Присвоение данным одной из категорий (достоверные, ложные, вредоносные)

На данном этапе обрабатывается подготовленный поток данных, полученный после этапа анализа согласованности (п. 5), и каждой записи присваивается одна из основных категорий (достоверные, ложные, вредоносные).

Для каждого сообщения $d_i \in D_{total}$ рассчитывается уровень доверия сообщения на основе надёжности источника Q_i, уровня согласованности данных $K_{data}(d_i)$, веса подтверждения $W_{confirm}(d_i)$, уровня конфликта $W_{conflict}(d_i)$.

Сравниваются значения параметров с заданными порогами:

– $K_{threshold,\ true}$ – порог для достоверных данных;

– $W_{threshold,\ conflict}$ – порог для конфликта;

– $Q_{low\text{-}trust}$ – минимальный уровень надёжности источника, ниже которого данные могут считаться ложными или вредоносными.

Сообщение классифицируется как достоверное, если:

– оно подтверждено несколькими надёжными источниками;

– уровень согласованности данных $K_{data}(d_i)$ приближается к 1 (почти все источники подтверждают данные);

– параметр $W_{confirm}(d_i)$ значительно выше порога согласованности, заданного системой;

– отсутствуют конфликты или противоречия ($W_{conflict}(d_i) \approx 0$).

Сообщение классифицируется как ложное, если:

– оно не подтверждено большинством источников или противоречит их информации;

– уровень согласованности данных $K_{data}(d_i)$ низок;

– параметр $W_{conflict}(d_i)$ превышает установленный порог;

– источник данных имеет крайне низкую надёжность.

Сообщение классифицируется как вредоносное, если:

– источник данных ранее предоставлял заведомо ложные данные (история ошибок источника);

– присутствуют специфические признаки преднамеренной дезинформации (например, отсутствие адекватных подтверждений от любых источников + явные признаки искажения в данных);

– источник имеет крайне низкий рейтинг доверия ($Q_i \approx 0$) и замечен в предоставлении некорректной информации.

Дополнительные аспекты классификации:

– учёт истории источника: источник, ранее замеченный в предоставлении ложных или вредоносных данных, получает штрафной коэффициент, который влияет на вес сообщений;

– работа с частично подтверждёнными данными: если данные подтверждены частично (например, только малой группой источников), они могут быть помечены как условно достоверные или требующие дальнейшей проверки;

– работа с изолированными данными: изолированные данные, предоставленные только одним источником, анализируются строго на основании надёжности источника и других признаков.

Оценка общего веса подтверждения сообщения:

$$W_{confirm}(d_i) = \sum_{j=1}^{n} Q_j * C_{ij},$$

где Q_j – надёжность источника S_j; C_{ij} – показатель согласованности данного сообщения с данными других источников (1 – подтверждено, 0 – не подтверждено).

Оценка уровня недоверия сообщения (по весу конфликта):

$$W_{confirm}(d_i) = \sum_{j=1}^{n} Q_i * (1 - C_{ij}).$$

Данные считаются ложными, если $W_{conflict}(d_i)$ превосходит порог недоверия $W_{threshold,\ conflict}$.

Итоговое решение:

$$\text{Категория } d_i \begin{cases} D_{true}, \text{если } K_{data}(d_i) \geq K_{threshold,true} \\ D_{false}, \text{если } W_{conflict}(d_i) \geq W_{threshold,conflict} \\ D_{malicious}, \text{если } Q_i \approx 0 \text{ и данные конфликтны.} \end{cases}$$

Формирование достоверного потока данных

Этап формирования достоверного потока данных завершает процесс обработки, фильтрации и анализа исходных данных. Он организует итоговый поток D_{output}, содержащий только подтверждённые (достоверные) данные, которые прошли все стадии проверки и классификации. Данный этап играет ключевую роль в построении финального набора информации, который будет использован конечным потребителем системы (другими модулями, приложениями, операторами или внешними системами). Он обеспечивает завершение логики обработки данных, гарантируя, что финальный массив информации соответствует ключевым требованиям: достоверность, структурированность и готовность к дальнейшему применению. Таким образом, финальный поток D_{output} содержит только достоверную информацию и организован таким образом, чтобы данные были полноценно структурированы, уникальны и надёжны. Потоки D_{false} и $D_{malicious}$ сохраняются для дальнейшего анализа (при наличии необходимости).

Корректировка надёжности источников

Этап корректировки надёжности источников является заключительным этапом обработки в рамках системы управления потоками данных. Его основная задача – на основании результатов анализа, классификации и формирования достоверного потока данных (на предыдущих этапах) динамически обновить оценки надёжности Q_i для каждого источника данных S_i. Этот блок работает в системе «обратной связи»: текущие результаты обработки данных влияют на будущие оценки источников. Это актуально в системах, где важна постоянная адаптация к изменяющимся условиям или злонамеренным действиям. В результате выполнения этапа система становится более эффективной в обработке данных на основе изменяющихся условий, минимизируя риски доверия вредоносным источникам за счет обновленных рейтингов источников Q_i и управления доверием.

Завершение обработки

Этап «Завершение» служит финальной стадией цикла обработки данных в системе. На этом этапе выполняются задачи, связанные с подведением итогов, консолидацией результатов всех предыдущих этапов, их сохранением и подготовкой системы к следующей итерации обработки данных. Основная цель этапа – корректное завершение текущей работы модуля, обеспечение целостности данных, освобождение ресурсов и подготовка среды для дальнейшего функционирования системы.

Результаты экспериментальных исследований по оценке формирования и предоставления достоверных данных

Целью экспериментальных исследований стало подтверждение гипотезы о том, что метод комплексирования данных от n-источников способен существенно повысить достоверность формируемой информации за счёт использования избыточности и перекрёстной корреляции данных между источниками и исключения низкодостоверных источников и сообщений с информацией, выходящей за рамки согласованных данных. Результаты проведённых экспериментов позволили выявить эффективность предложенного метода и подтвердить его преимущества.

Для проверки гипотезы были смоделированы различные сценарии объединения данных от n-источников, где источники имели разные уровни надёжности (Q_i). Данные могли быть достоверными (согласованными между большинством источников), ложными (содержащими погрешности или противоречащие достоверным данным), вредоносными (намеренно искажёнными).

Параметры оценивались при разном уровне избыточности данных (процент дублирующих или схожих сообщений в общей выборке).

Метрики оценки:

1. Достоверность объединённых данных (D): определялась как отношение количества достоверной информации в итоговом потоке D_{total} к общему объёму данных от всех источников.

$$D = \frac{\text{Количество достоверных сообщений}}{\text{Общее количество обработанных сообщений}}.$$

2. Согласованность (C): оценка доли данных, подтверждённых большинством источников (перекрёстных коррелированных данных).

$$C = \frac{\text{Количество подтверждённых данных}\left(\text{более}\,50\%\ \text{голосов}\right)}{\text{Общее количество обработанных сообщений}}.$$

3. Доля исключённых ненадёжных сообщений (E): процент сообщений, которые были исключены из итогового потока данных из-за выявленных несоответствий или низкой надёжности источника.

$$E = \frac{\text{Количество исключённых сообщений}}{\text{Общее количество обработанных сообщений}}.$$

В эксперименте участвовало 10 источников данных (n = 10) с разной начальной надёжностью, случайным образом установленной в пределах от 0,3 до 0,9. Всего было сгенерировано 10 000 сообщений, из которых 60 % – достоверные данные (описание одного и того же события от нескольких источников), 30 % – ошибочные данные (включая шумовые данные, незначительные расхождения), 10 % – вредоносные данные (полностью искажённые или противоречивые).

Избыточность данных (частота получения совпадающих данных от разных источников) варьировалась от 10 % до 90 %.

Этапы эксперимента:

1. Комбинация данных от всех источников: для каждого сообщения проверялось, предоставляло ли его несколько источников.

2. Перекрёстная корреляция: оценивалась согласованность сообщений (анализ временных меток, полей содержимого).

3. Исключение недостоверных сообщений: на основании расчёта надёжности источников (Q_i) и соответствия данных большинству источников.

4. Формирование итогового потока D_{total}.

5. Оценка метрик *(D, C, E)* для каждого уровня избыточности данных и уровня шума.

Результаты исследований и их обсуждение

1. Повышение достоверности информации (*D*).

При увеличении избыточности (наличии перекрытия сообщений между источниками) достоверность объединённых данных значительно увеличилась:

- при низкой избыточности (10 %): $D \approx 0,75$.
- при высокой избыточности (90 %): $D \approx 0,98$.

Это объясняется тем, что при высокой степени избыточности система могла легче идентифицировать ошибки и вредоносные данные путём анализа перекрёстной корреляции между источниками. Полученные результаты зависимости достоверности информации от избыточности данных представлены в табл. 2.

Таблица 2

Зависимость достоверности информации от избыточности данных

Уровень избыточности (%)	*D* (достоверность итогового потока)
10	0,75
30	0,85
50	0,90
70	0,95
90	0,98

2. Улучшение согласованности данных (C).

При использовании перекрёстной корреляции удалось достичь высокого уровня согласованности данных (табл. 3). Даже при низких значениях избыточности алгоритм обеспечивал согласованность $C \geq 0,8$. При средней и высокой степени избыточности (50 % − 90 %) согласованность достигала $C \geq 0,95$.

Таблица 3

Результаты использования перекрестной корреляции

Уровень избыточности (%)	*C* (согласованность итогового потока)
10	0,83
30	0,89
50	0,93
70	0,97
90	0,97

3. Исключение недостоверных данных (*E*).

На этапе обработки исключалось до 98 % вредоносных данных. При этом 100 % дублирующих сообщений также удалялись как избыточные. Ошибочные сообщения от ненадёжных источников или конфликтующие данные составляли основную часть исключённых элементов (при средней надёжности источников). Результаты исключения недостоверных данных представлены в табл. 4.

Таблица 4

Результаты исключения недостоверных данных

Уровень шума в данных (%)	*E* (исключённые данные)
10	0,12
30	0,22
50	0,36
70	0,51

4. Сравнение с некоррелированными данными (без комплексирования).

Для сравнения достоверность данных $D^{(без\ корреляции)}$, обработанных без перекрёстной корреляции, составила 0,65 при низкой избыточности данных и 0,81 – при высокой избыточности.

Таким образом, метод комплексирования данных обеспечил рост достоверности итогового потока *D* на 15 % – 20 % в сравнении.

Перспективы дальнейших исследований связаны с расширением подходов к учёту специфики различных типов данных и алгоритмическим совершенствованием для повышения точности и масштабируемости метода в сложных информационных системах.

Заключение

Методы, описанные в данной статье, позволяют не только улучшить качество обработки данных, но и развивать механизмы защиты от ложной или вредоносной информации в условиях информационной войны, кибератак или массовой дезинформации.

Преимущества предложенных методов повышения достоверности:

1. Гибкость и адаптивность:

– методы динамической корректировки надёжности источников позволяют системе адаптироваться к изменяющимся условиям. Например, если новый источник со временем продемонстрирует высокую достоверность, он сможет получить более значительный вес в последующей обработке данных;

– система адаптирует свои параметры (пороги, весовые коэффициенты и др.), чтобы реагировать на проявляющиеся типы ошибок обработки информации – упреждающий ответ на характерные схемы дезинформации.

2. Многослойность подхода.

В основу системы заложен многоуровневый анализ данных: от индивидуальных характеристик сообщений до глобальной оценки всей совокупности данных. Это делает подход более глубоким и проактивным за счёт перерасчёта показателей согласованности, вероятностей категорий данных и взаимодействия разных источников.

3. Обнаружение и блокировка вредоносной информации.

Одной из ключевых функций системы является активное выявление потенциально вредоносных сообщений и минимизация их влияния. Благодаря статистическим и вероятностным методам, становится возможным предотвращение сбоев системы из-за некорректной или враждебной информации.

4. Устойчивость к информационным атакам.

Критически важным аспектом является защита данных от внешних атак. Применение описанных алгоритмов обеспечивает снижение вероятности дезинформации или манипуляции данными на выходе системы.

5. Улучшение качества анализа на следующем уровне.

Итоговый поток данных, построенный на результатах комплексирования, классификации и повышения достоверности, представляет собой основу для дальнейших аналитических процессов. Эти данные являются согласованными и корректными, что даёт возможность улучшить предсказательные модели, аналитические системы и процедуры принятия решений.

Список источников

1. Fake news detector using deep learning / A. Akshansh [et al.] // International Journal of Advanced Research. 2024. № 11. Vol. 04. P. 1612–1621.

2. Revisiting Fake News Detection: Towards Temporality-aware Evaluation by Leveraging Engagement Earliness / J. Kim [et al.] // arXivLabs. 2024. № 2411. Vol. 12775. P. 1–11.

3. Zhou X., Zafarani R. A survey of fake news: Fundamental theories, detection methods and opportunities // ACM Computing Surveys. 2020. № 53 (5). P. 1–40.

4. Привалов А.Н., Смирнов В.А. Поиск фейковых сайтов с использованием метода определения визуального сходства страниц // Известия ТулГУ. Технические науки. 2022. № 9. С. 260–264.

5. Обеспечение целостности данных посредством частичной «фрагментации» данных / А.К. Куртов [и др.] // Современные научные исследования и инновации. 2023. № 9. URL: https://web.snauka.ru/issues/2023/09/100716 (дата обращения: 05.03.2025).

6. Zhou Yu. The Silent Saboteur: The Impact and Management of Malicious Word-Of-Mouth in The Digital Age // Highlights in Business Economics and Management. 2024. № 41. P. 381–386.

7. The impact of malicious nodes on the spreading of false information / Z. Ruan [et al.] // Chaos: An Interdisciplinary Journal of Nonlinear Science. 2020. № 30. P. 083101.

8. Creating and detecting fake reviews of online products / J. Salminen [et al.] // Journal of Retailing and Consumer Services. 2022. № 64. P. 102771.

9. Wesam H.A., Ragheed A., Yossra H.A. Opinion mining for fake recommendations in e-commerce: A machine learning approach using LightGBM // AIP Conference Proceedings. 2025. № 3169. Vol. 030015. P. 1–11.

10. Shilpa Yu., Gulbakshee Dharmela K.M. Fake Review Detection Using Machine Learning Techniques // Journal of Emerging Technologies and Innovative Research (JETIR). 2021. Vol. 8 (4).

11. Лебедев И.С. Адаптивное применение моделей машинного обучения на отдельных сегментах выборки в задачах регрессии и классификации // Информационно-управляющие системы. 2022. № 3 (118). С. 20–30.

12. Тымчук А.И. Информационная система контроля достоверности данных приборов учёта в автоматизированной информационно-измерительной системе контроля и учёта электроэнергии // МНИЖ. 2024. № 6 (144). С. 1–9.

13. Минаков С.С., Михайленко Н.В. Проблемы обеспечения достоверности технических данных и сведений, сопряжённых с выявлением и расследованием инцидентов и преступлений, совершённых с использованием информационно-телекоммуникационных технологий // Вестник экономической безопасности. 2023. № 6. С. 107–112.

References

1. Fake news detector using deep learning / A. Akshansh [et al.] // International Journal of Advanced Research. 2024. № 11. Vol. 04. P. 1612–1621.

2. Revisiting Fake News Detection: Towards Temporality-aware Evaluation by Leveraging Engagement Earliness / J. Kim [et al.] // arXivLabs. 2024. № 2411. Vol. 12775. P. 1–11.

3. Zhou X., Zafarani R. A survey of fake news: Fundamental theories, detection methods and opportunities // ACM Computing Surveys. 2020. № 53 (5). P. 1–40.

4. Privalov A.N., Smirnov V.A. Poisk fejkovyh sajtov s ispol'zovaniem metoda opredeleniya vizual'nogo skhodstva stranic // Izvestiya TulGU. Tekhnicheskie nauki. 2022. № 9. S. 260–264.

5. Obespechenie celostnosti dannyh posredstvom chastichnoj «fragmentacii» dannyh / A.K. Kurtov [i dr.] // Sovremennye nauchnye issledovaniya i innovacii. 2023. № 9. URL: https://web.snauka.ru/issues/2023/09/100716 (data obrashcheniya: 05.03.2025).

6. Zhou Yu. The Silent Saboteur: The Impact and Management of Malicious Word-Of-Mouth in The Digital Age // Highlights in Business Economics and Management. 2024. № 41. P. 381–386.

7. The impact of malicious nodes on the spreading of false information / Z. Ruan [et al.] // Chaos: An Interdisciplinary Journal of Nonlinear Science. 2020. № 30. P. 083101.

8. Creating and detecting fake reviews of online products / J. Salminen [et al.] // Journal of Retailing and Consumer Services. 2022. № 64. P. 102771.

9. Wesam H.A., Ragheed A., Yossra H.A. Opinion mining for fake recommendations in e-commerce: A machine learning approach using LightGBM // AIP Conference Proceedings. 2025. № 3169. Vol. 030015. P. 1–11.

10. Shilpa Yu., Gulbakshee Dharmela K.M. Fake Review Detection Using Machine Learning Techniques // Journal of Emerging Technologies and Innovative Research (JETIR). 2021. Vol. 8 (4).

11. Lebedev I.S. Adaptivnoe primenenie modelej mashinnogo obucheniya na otdel'nyh segmentah vyborki v zadachah regressii i klassifikacii // Informacionno-upravlyayushchie sistemy. 2022. № 3 (118). S. 20–30.

12. Tymchuk A.I. Informacionnaya sistema kontrolya dostovernosti dannyh priborov uchyota v avtomatizirovannoj informacionno-izmeritel'noj sisteme kontrolya i uchyota elektroenergii // MNIZH. 2024. № 6 (144). S. 1–9.

13. Minakov S.S., Mihajlenko N.V. Problemy obespecheniya dostovernosti tekhnicheskih dannyh i svedenij, sopryazhyonnyh s vyyavleniem i rassledovaniem incidentov i prestuplenij, sovershyonnyh s ispol'zovaniem informacionno-telekommunikacionnyh tekhnologij // Vestnik ekonomicheskoj bezopasnosti. 2023. № 6. S. 107–112.

Информация о статье:
Статья поступила в редакцию: 14.01.2025; одобрена после рецензирования: 19.03.2025;
принята к публикации: 22.03.2025
Information about the article:
The article was submitted to the editorial office: 14.01.2025; approved after review: 19.03.2025;
accepted for publication: 22.03.2025

Сведения об авторах:
Карманова Наталия Андреевна, аспирант Национального исследовательского университета ИТМО (197101, Санкт-Петербург, Кронверкский пр., д. 49), e-mail: karmanova.ifmo@gmail.com, https://orcid.org/0000-0002-7007-3120, SPIN-код: 3628-9988

Information about the authors:
Karmanova Natalia A., postgraduate student at the ITMO University (197101, Saint-Petersburg, Kronverksky ave., 49), e-mail: karmanova.ifmo@gmail.com, https://orcid.org/0000-0002-7007-3120, SPIN: 3628-9988