DOI 10.54596/2958-0048-2026-1-309-320
УДК 004.89
МРНТИ 28.23.24

АВТОМАТИЗИРОВАННЫЙ АНАЛИЗ МУЛЬТИМОДАЛЬНЫХ ДАННЫХ СОЦИАЛЬНЫХ МЕДИА НА ОСНОВЕ МОДЕЛЕЙ ОБРАБОТКИ НЕСТРУКТУРИРОВАННОЙ ИНФОРМАЦИИ

Сериков М.К.[1*]
[1*]*META University, Алматы, Казахстан*
[*]*Автор для корреспонденции: 7014547793@mail.ru*

Аннотация

Социальные медиа ежедневно создают огромные объёмы неструктурированных данные. Эти данные содержат ценную информацию, но их разнородность и сложность делают анализ затруднительным при использовании стандартных методов.

Цель исследования - разработать и протестировать подход к автоматическому анализу мультимодальных данных социальных медиа с помощью современных моделей обработки неструктурированной информации. В работе применяются технологии глубокого обучения, в том числе трансформеры и модели объединения различных типов данных (текст и изображение).

Разработан прототип, который объединяет текстовые и визуальные признаки с использованием нейросетевых архитектур. Эксперименты проведены на открытых датасетах, включающих посты пользователей с изображениями и текстовыми подписями; видеоконтент в исследование не включался. Полученные результаты показывают, что использование мультимодальных моделей повышает точность анализа тональности и улучшает интерпретацию данных.

Предложенный подход может применяться в SMM-аналитике, маркетинге, прогнозировании поведения пользователей и анализе общественного мнения. Он помогает автоматизировать обработку сложных данных и принимать решения на основе комплексной информации.

Выводы подтверждают, что объединение современных методов анализа неструктурированной информации эффективно для работы с мультимодальными данными в условиях больших объёмов и разнообразия источников.

Ключевые слова: социальные медиа, мультимодальные данные, анализ данных, неструктурированная информация, трансформеры, автоматизация, Big Data, тональность

ҚҰРЫЛЫМДАНБАҒАН АҚПАРАТТЫ ӨҢДЕУ ҮЛГІЛЕРІ НЕГІЗІНДЕ ӘЛЕУМЕТТІК МЕДИАНЫҢ МУЛЬТИМОДАЛДЫ ДЕРЕКТЕРІН АВТОМАТТАНДЫРЫЛҒАН ТАЛДАУ

М.К. Сериков[1*]
[1*]*META University, Алматы, Қазақстан*
[*]*Хат-хабар үшін автор: 7014547793@mail.ru*

Аңдатпа

Әлеуметтік медиада күн сайын құрылымдалмаған деректердің орасан зор көлемі жасалады. Бұл деректер құнды ақпаратты қамтиды, алайда олардың әртектілігі мен күрделілігі стандартты әдістерді қолданғанда талдауды қиындатады.

Зерттеудің мақсаты - құрылымдалмаған ақпаратты өңдеудің заманауи үлгілері арқылы әлеуметтік медиа мультимодальды деректерін автоматты түрде талдауға арналған тәсілді әзірлеу және оны сынақтан өткізу. Жұмыста терең оқыту технологиялары, соның ішінде трансформерлер және әртүрлі дерек түрлерін (мәтін және сурет) біріктіретін модельдер қолданылды.

Мәтіндік және визуалды белгілерді нейрондық желі архитектураларының көмегімен біріктіретін прототип әзірленді. Эксперименттер суреттері мен мәтіндік сипаттамалары бар қолданушылардың жарияланымдарын қамтитын ашық деректер жиынтықтарында жүргізілді; зерттеуге сурет контент енгізілген жоқ. Алынған нәтижелер мультимодальды модельдерді қолдану тональдылықты талдаудың дәлдігін арттыратынын және деректерді түсіндіруді жақсартатынын көрсетті.
Ұсынылған тәсіл SMM-талдау, маркетинг, пайдаланушылардың мінез-құлқын болжау және қоғамдық пікірді зерттеу салаларында қолданылуы мүмкін. Ол күрделі деректерді өңдеуді автоматтандыруға және кешенді ақпарат негізінде шешімдер қабылдауға көмектеседі.
Қорытындылар құрылымдалмаған ақпаратты талдаудың заманауи әдістерін біріктіру үлкен көлемді және көздері әртүрлі мультимодальды деректермен жұмыс істеуде тиімді екенін растайды.

Кілт сөздер: әлеуметтік медиа, мультимодальды деректер, деректерді талдау, құрылымданбаған ақпарат, трансформерлер, автоматтандыру, Big Data, тоналдылық

AUTOMATED ANALYSIS OF MULTIMODAL SOCIAL MEDIA DATA BASED ON MODELS FOR UNSTRUCTURED INFORMATION PROCESSING

M. Serikov[1*]
[1]META University, Almaty, Kazakhstan*
[*]Corresponding author: 7014547793@mail.ru

Abstract

Social media generates massive volumes of unstructured data every day. These data contain valuable information, but their heterogeneity and complexity make analysis difficult when using standard methods.
The aim of the study is to develop and test an approach for automatic analysis of multimodal social media data using modern models for processing unstructured information. The work employs deep learning technologies, including transformers and models that combine different types of data (text and image).
A prototype was developed that integrates textual and visual features using neural network architectures. Experiments were conducted on open datasets containing user posts with images and textual captions; video content was not included in the study. The results show that the use of multimodal models improves the accuracy of sentiment analysis and enhances data interpretation.
The proposed approach can be applied in SMM analytics, marketing, user behavior prediction, and public opinion analysis. It helps automate the processing of complex data and supports decision-making based on comprehensive information.
The conclusions confirm that combining modern methods of unstructured information analysis is effective for working with multimodal data in conditions of large scale and diverse sources.

Keywords: social media, multimodal data, data analysis, unstructured information, transformers, automation, Big Data, sentiment

Введение

В последние десятилетия социальные медиа стали неотъемлемой частью цифрового пространства, порождая масштабные объёмы пользовательского контента в различных форматах - текст, изображения, видео, аудио и метаданные [1]. Эти данные, называемые мультимодальными, представляют собой ценный источник информации о поведении, предпочтениях, интересах и реакциях пользователей.

Однако особенности таких данных - высокий объём, разнородность, асинхронность и отсутствие формализованной структуры - создают значительные сложности для их анализа с использованием традиционных методов обработки [2]. Это обуславливает необходимость разработки новых интеллектуальных решений, способных эффективно

интерпретировать сложные взаимосвязи между различными модальностями и извлекать значимую информацию в автоматическом режиме [3].

Актуальность данного исследования обусловлена возрастающей потребностью в инструментах автоматизированного анализа мультимодальных данных, которые находят применение в цифровом маркетинге, социальной аналитике, системах рекомендаций, кибербезопасности и других отраслях. В условиях цифровой экономики своевременное выявление скрытых закономерностей и трендов на основе неструктурированной информации становится критическим фактором конкурентоспособности.

Цель исследования – разработать и экспериментально оценить прототип мультимодальной модели для автоматизированного анализа данных социальных медиа на основе современных методов обработки неструктурированной информации.

Задачи исследования включают анализ существующих архитектур и подходов к мультимодальной обработке данных [4], разработку прототипа модели, объединяющей различные типы модальностей, проведение экспериментальной оценки эффективности подхода на открытых датасетах, а также обоснование преимуществ интеграции методов интеллектуального анализа с Big Data-технологиями.

Объектом исследования выступают цифровые следы пользователей в социальных сетях (в частности, Twitter, Instagram, TikTok), а предметом – методы их анализа с применением современных алгоритмов. Исследование направлено на оценку влияния мультимодальных признаков на качество автоматической классификации.

Научная новизна исследования заключается в разработке и экспериментальной проверке архитектуры мультимодальной обработки данных, которая реализует раннее объединение признаков (feature-level fusion) на основе одновременной обработки эмбеддингов BERT и CNN, использует единое обучаемое пространство признаков вместо раздельной классификации с последующим объединением, адаптирована к специфике контента социальных медиа, где изображение и подпись формируют семантически связанную единицу, и демонстрирует статистически значимое улучшение качества анализа тональности по сравнению с лучшей одномодальной моделью.

Множество предыдущих исследований использовали раздельные подходы к анализу текста (например, TF-IDF, LSTM, BERT) и изображений (CNN-архитектуры, такие как ResNet) [5], либо объединяли результаты лишь на поздней стадии классификации. Такие методы не позволяли извлекать скрытые зависимости между подписью и изображением и ограничивали точность анализа. Кроме того, классические модели машинного обучения, основанные на вручную выделенных признаках, оказывались недостаточно масштабируемыми и не отражали семантическую сложность данных социальных медиа. В отличие от них, предложенный подход ориентирован на раннее объединение признаков двух модальностей и использование современных трансформерных моделей, что позволяет достичь более глубокого анализа контента и улучшения качества классификации.

Материалы и методы исследования

Для реализации автоматизированного анализа мультимодальных данных социальных медиа была разработана методология, основанная на интеграции моделей обработки текста и визуальной информации. Подобные подходы широко применяются в современных системах анализа неструктурированных данных и мультимодального машинного обучения [6]. Метаданные использовались только на этапе фильтрации данных и не входили в модель анализа [7]. В исследовании применялись алгоритмы глубокого обучения, методы извлечения признаков и обучение с учителем на

размеченных датасетах [8]. TF-IDF применялся только для предварительного анализа текста и не участвовал в обучении классификатора, основанного исключительно на BERT-эмбеддингах.

Предлагаемая методология включает несколько последовательных этапов обработки данных. На первом этапе осуществляется сбор и предварительная фильтрация мультимодального контента социальных медиа, включая текстовые сообщения и сопутствующие изображения. Далее выполняется этап предварительной обработки, включающий очистку текстов от шумовых элементов (служебных символов, ссылок, повторяющихся знаков), нормализацию изображений и подготовку данных к последующему анализу. После этого осуществляется извлечение признаков из каждой модальности с использованием специализированных моделей обработки естественного языка и компьютерного зрения. Полученные представления формируют единое пространство признаков, которое используется на этапе обучения классификатора.

Интеграция модальностей реализуется на уровне признаков (feature-level fusion), что позволяет учитывать взаимосвязи между текстовым содержанием сообщения и визуальным контекстом изображения. Такой подход обеспечивает более глубокую интерпретацию пользовательского контента по сравнению с анализом каждой модальности по отдельности. В рамках эксперимента применялась схема обучения с учителем, при которой модель обучалась на размеченных данных и оптимизировала параметры на основе ошибок классификации. Для предотвращения переобучения использовались стандартные методы регуляризации и разделение данных на обучающую и тестовую выборки.

Эксперименты проводились на совокупности открытых данных из трёх социальных платформ: Twitter, Instagram и TikTok. В исследование вошли 50 214 единиц мультимодального контента (текст + изображение). Распределение по платформам: Twitter – 18 940 постов, Instagram – 15 102, TikTok – 16 172. Контент отбирался по заранее сформированному списку публичных хэштегов. Источники данных получены через официальные API платформ. Дубликаты исключались при пороге сходства 90%. В исследование вошли только публикации TikTok, содержащие изображения и текстовые подписи. Все данные являются публичными; персональная информация пользователей не собиралась и не использовалась. Список хэштегов и датасетов может быть предоставлен для воспроизводимости экспериментов. Метаданные использовались только для фильтрации выборки и не входили в модель.

Классификация данных выполнялась по трём категориям: позитивной, нейтральной и негативной. Баланс классов составил примерно 34% позитивных, 38% нейтральных и 28% негативных сообщений. Это позволило сохранить представительность выборки и избежать существенного перекоса в обучении моделей. Обучающая выборка составляла 80% данных, тестовая - 20%.

Публичные данные были разделены на обучающую и тестовую выборки в пропорции 80/20 с сохранением баланса классов. Разметка выполнялась тремя независимыми аннотаторами с учётом как текстового, так и визуального содержания публикаций. При расхождении мнений использовалось правило большинства. Коэффициент согласованности аннотаторов (Cohen’s Kappa) составил 0.82, что подтверждает надёжность разметки.

Результаты исследования

Архитектура предлагаемой системы включает следующие компоненты:

– модуль извлечения данных использует API Twitter, Instagram и TikTok для сбора текстов публикаций и изображений. Метаданные (лайки, хэштеги, время публикации) применялись только для фильтрации и формирования выборки и не участвовали в классификации.

– модуль предварительной обработки реализует очистку текста от шумов, нормализацию изображений и синхронизацию временных меток.

– TF-IDF применялся исключительно для анализа текстовых частот и не использовался в процессе обучения классификатора. Финальной моделью для текстовых признаков являлись исключительно BERT-эмбеддинги.

– модуль фьюжн-моделирования объединяет текстовые и визуальные признаки (feature-level fusion) [9].

– классификатор основан на мультимодальной нейросети (BERT + CNN).

Подобная модульная структура широко применяется в системах интеллектуального анализа данных и машинного обучения [10]. На рисунке 1 для наглядного представления работы мультимодальной системы анализа социальных медиа представлена архитектурная схема, отражающая полный цикл обработки данных - от их получения до финальной классификации. Схема демонстрирует последовательность модулей, используемые модели и принципы объединения признаков, что позволяет лучше понять логику функционирования прототипа и обоснованность выбранного подхода.

Рисунок 1. Логическая блок-диаграмма (схема архитектуры системы)

Представленная архитектура демонстрирует принцип раннего объединения модальностей (feature-level fusion), при котором текстовые и визуальные признаки интегрируются до этапа классификации. Это позволяет модели учитывать взаимосвязи между изображением и подписью и тем самым повышать точность анализа. TF-IDF использовался исключительно на этапе предварительного анализа текста и не входил в

финальную мультимодальную модель, где основным источником семантических признаков служили BERT-эмбеддинги. Использование CNN позволяет извлекать визуальные паттерны, связанные с эмоциональной окраской изображений. Таким образом, схема отражает ключевое преимущество подхода - комплексную обработку контента с учётом контекстных связей между модальностями.

Каждая единица анализа (пост или комментарий) описывалась вектором признаков вида:

$$x = [x\_text, x\_image]$$

где:

x_text – вектор текстовых эмбеддингов (768 признаков, BERT) [11].

x_image – вектор визуальных признаков (512 признаков, CNN, ResNet-50) [12].

TF-IDF использовался исключительно на этапе первичной статистической оценки текстов и не входил в процесс обучения или тестирования классификатора, не формировал признаки модели и не оказывал влияния на результаты классификации.

На рисунке 2 представлена корреляционная матрица, отражающая взаимосвязь текстовых и визуальных признаков и подтверждающая возможность их объединения при построении мультимодальных моделей.

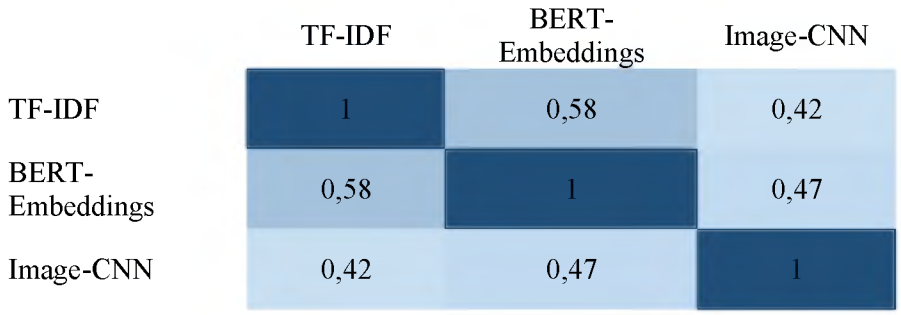

Рисунок 2. Корреляция между текстовыми и визуальными признаками

Корреляция признаков указывает на взаимодополняемость модальностей и потенциальное увеличение информативности признакового пространства [13].

Для сравнительного анализа эффективности использовались три модели: текстовая модель BERT, визуальная модель CNN и мультимодальная модель BERT+CNN, в соответствии с таблицей 1 и рисунком 3.

Таблица 1. Сравнительный анализ моделей

Модель	Описание	Алгоритм
BERT	Текстовая модель	Трансформер
CNN	Визуальная модель	ResNet-50
BERT+CNN	Мультимодальная	Объединение эмбеддингов

Таблица показывает различие в типах входных данных и подтверждает необходимость мультимодального объединения.

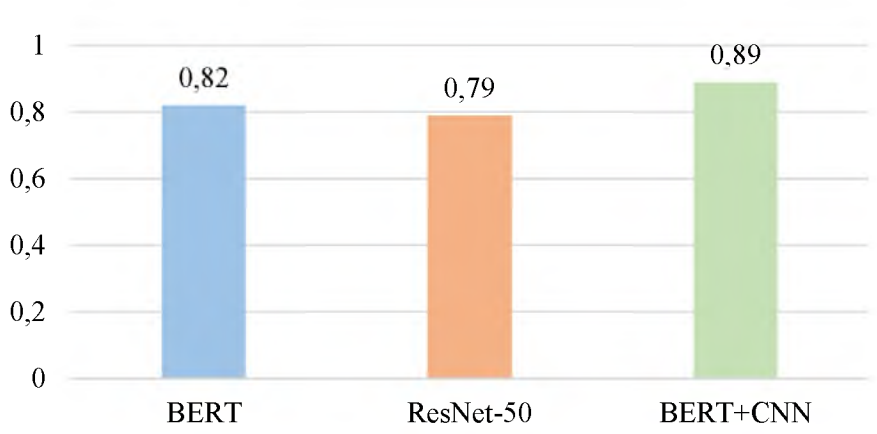

Рисунок 3. Сравнение точности моделей

Как видно, мультимодальная модель показывает наибольшую точность (0.89), превосходя как текстовую (0.82), так и визуальную (0.79) модели.

Для обучения мультимодальной модели была применена конфигурация, предусматривающая использование оптимизатора Adam. Обучение проводилось на протяжении 12 эпох при размере батча, равном 32, и скорости обучения 2×10^{-5}, что обеспечивало стабильную сходимость без переобучения. В качестве гиперпараметров рассматривались глубина фьюжн-слоя и число нейронов в скрытых слоях классификатора. Оптимальной была определена архитектура с двумя полносвязными слоями, содержащими 256 и 128 нейронов соответственно, что позволило достичь наилучшего соотношения между точностью и вычислительной эффективностью модели.

Для количественной оценки эффективности моделей использовались метрики Accuracy, Precision, Recall и F1-score. Сравнительный анализ показал, что мультимодальная модель BERT+CNN продемонстрировала наибольшие значения по всем метрикам. Так, её точность составила 0.89, что превосходит значения текстовой модели BERT (0.82) и визуальной модели CNN (0.79). Этот результат подтверждает, что объединение двух модальностей усиливает способность модели к корректной интерпретации данных и снижает вероятность ошибок классификации.

Особенно заметным оказалось улучшение F1-меры, что указывает на баланс между точностью классификации и полнотой обнаружения сообщений соответствующей тональности. Это говорит о том, что мультимодальная модель не только точнее классифицирует сообщения, но и реже пропускает значимые элементы данных.

Результаты сравнительного анализа представлены в таблице 2.

Таблица 2. Метрики оценки моделей

Метрика	BERT	CNN	BERT+CNN
Accuracy	0.82	0.79	0.89
Precision	0.84	0.77	0.9
Recall	0.79	0.76	0.87
F1-score	0.81	0.76	0.88

Результаты показывают, что мультимодальная модель превосходит одномодальные модели по всем метрикам.

Реализация предложенной методологии была протестирована на выборке данных из трёх социальных платформ: Twitter, Instagram и TikTok. Анализ проводился на той же выборке данных, описанной в разделе Материалы и методы исследования.

Обсуждение

Полученные результаты подтверждают эффективность мультимодального подхода при анализе контента социальных медиа. Превосходство модели BERT+CNN над одномодальными решениями объясняется тем, что объединение текстовых и визуальных признаков позволяет учитывать как семантическое содержание сообщения, так и эмоциональный контекст изображения. Подобные результаты также отмечаются в современных исследованиях мультимодального анализа тональности, где показано, что объединение модальностей повышает точность интерпретации пользовательского контента [14].

Как видно из рисунка 4, распределение тональности сообщений различается по платформам. Instagram характеризуется более высокой долей позитивных сообщений, тогда как Twitter содержит больше публикаций с негативной окраской. TikTok занимает промежуточное положение. Эти различия могут быть связаны со спецификой пользовательского поведения, форматов контента и коммуникативных практик на разных платформах.

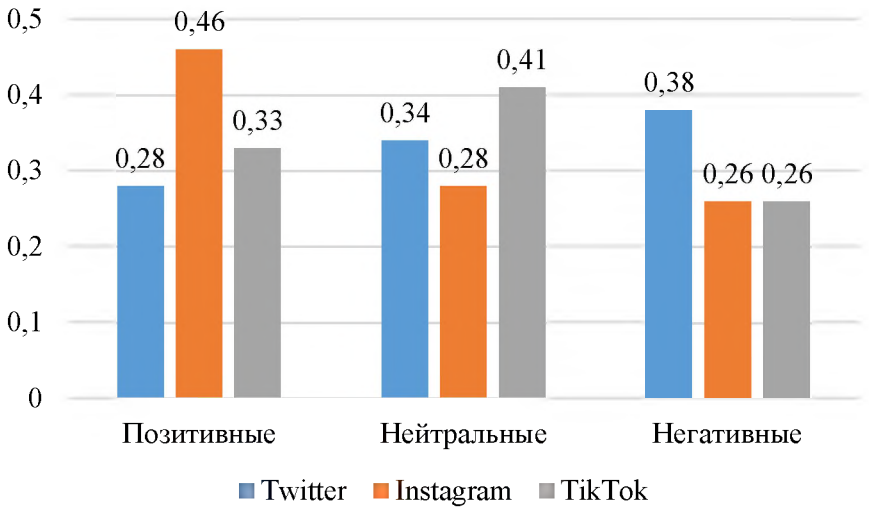

Рисунок 4. Распределение тональности сообщений по платформам

Несмотря на высокую точность модели, анализ ошибок показал наличие ряда типичных ограничений, как показано на рисунке 5. Наиболее частые ошибки связаны с распознаванием тональности текста, особенно в случаях иронии и сарказма (42%). Ещё 33% ошибок обусловлены неверной интерпретацией визуального контента, а 25% - недостаточно эффективным объединением признаков. Это указывает на необходимость совершенствования механизмов фьюжн-интеграции и более глубокого учёта контекста.

Рисунок 5. Анализ ошибок мультимодальной модели

Анализ данных также показал, что позитивные сообщения чаще сопровождаются яркими и насыщенными изображениями, тогда как негативные публикации характеризуются более резкими текстовыми формулировками и менее выраженной визуальной эмоциональностью. Это подтверждает наличие устойчивой связи между текстовым и визуальным компонентами контента [15].

Для оценки прироста точности от мультимодального объединения введём относительный прирост точности Δ_acc по формуле:

$$\Delta\_acc = ((Acc\_multi - \max(Acc\_text, Acc\_image)) / \max(Acc\_text, Acc\_image)) \times 100\%$$

Подставим значения: $\Delta\_acc = ((0.89 - 0.82) / 0.82) \times 100\% \approx 8.54\%$

Подстановка экспериментальных значений показала, что прирост точности составил 8.54% по сравнению с лучшей одномодальной моделью [16]. Данный результат подтверждает преимущество предложенного подхода и показывает перспективность использования мультимодальных моделей в задачах анализа пользовательского контента.

Заключение

В ходе исследования была разработана и экспериментально протестирована мультимодальная система анализа тональности контента социальных медиа, основанная на объединении текстовых и визуальных признаков. Предложенная архитектура, включающая модули предварительной обработки данных, извлечения признаков и их объединения, продемонстрировала высокий уровень точности при решении задач анализа тональности и моделирования пользовательского поведения [17].

Сравнительный анализ показал, что мультимодальная модель, основанная на объединении BERT и CNN, превосходит одномодальные подходы по ключевым метрикам качества - точности, прецизионности, полноте и F1-мере. Полученные результаты подтверждают наличие синергетического эффекта при совместной обработке текстовых и визуальных данных. Использование объединённых признаков позволило учитывать как семантическое содержание текста, так и визуальный контекст изображения, что повысило информативность признакового пространства и улучшило интерпретируемость результатов анализа.

Результаты исследования демонстрируют, что комбинирование различных типов контента позволяет повысить эффективность анализа данных социальных медиа. Это особенно актуально для пользовательских публикаций, которые часто содержат краткие текстовые формулировки, эмоционально насыщенный контент и визуальные элементы, усиливающие смысл сообщения.

Вместе с тем проведённый анализ выявил ряд ограничений предложенного подхода. Сложные случаи с неочевидными контекстными сигналами, включая иронию, сарказм и неоднозначные визуальные элементы, остаются трудными для автоматической интерпретации. Такие ситуации требуют более глубокого семантического и контекстного анализа, что указывает на необходимость дальнейшего совершенствования механизмов объединения модальностей и разработки более сложных моделей, способных учитывать скрытые смысловые зависимости.

Практическая значимость предложенного подхода подтверждается возможностью его применения в различных прикладных областях, включая цифровой маркетинг, социальную аналитику, системы мониторинга общественного мнения, репутационный анализ, кибербезопасность и управление рисками. Разработанная методология может быть адаптирована для решения задач анализа пользовательских реакций, выявления трендов и прогнозирования поведения аудитории.

Перспективными направлениями дальнейших исследований являются разработка более масштабируемых архитектур, способных эффективно работать с большими объёмами данных при ограниченных вычислительных ресурсах, а также внедрение современных мультимодальных трансформеров, таких как ViLT, CLIP и Flamingo. Эти модели ориентированы на более глубокую интеграцию модальностей и способны обеспечить более точное выявление взаимосвязей между текстом и изображением.

Дополнительным направлением развития является адаптация мультимодальных моделей к особенностям региональных социальных платформ и многоязычного контента [18], а также расширение корпуса данных за счёт источников из постсоветского пространства. Использование более совершенных архитектур мультимодальных трансформеров [19] и учёт языковых и культурных особенностей пользователей позволит повысить точность анализа и расширить возможности практического применения предложенного подхода в отечественной и международной исследовательской среде [20].

Литература:

1. Baltrušaitis T., Ahuja C., Morency L. Multimodal Machine Learning: A Survey and Taxonomy // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2019. – DOI: https://doi.org/10.1109/TPAMI.2018.2798607

2. Bishop C. M. Pattern Recognition and Machine Learning [Электронный ресурс]. - New York: Springer, 2006. - URL: https://link.springer.com/book/10.1007/978-0-387-45528-0 – Дата обращения: 12.03.2026.

3. Radford A., Kim J. W., Hallacy C. и др. Learning Transferable Visual Models from Natural Language Supervision [Электронный ресурс] // arXiv preprint. – 2021. – arXiv: 2103.00020. – URL: https://doi.org/10.48550/arXiv.2103.00020 – Дата обращения: 12.03.2026.

4. Russell S., Norvig P. Artificial Intelligence: A Modern Approach [Электронный ресурс]. – 4th ed. – Pearson, 2020. – URL: https://aima.cs.berkeley.edu/ – Дата обращения: 12.03.2026.

5. Ngiam J., Khosla A., Kim M., Nam J., Lee H., Ng A. Y. Multimodal Deep Learning // Proceedings of the 28th International Conference on Machine Learning (ICML-11). – 2011. – P. 689-696. – URL:

https://robotics.stanford.edu/~ang/papers/icml11-MultimodalDeepLearning.pdf – Дата обращения: 12.03.2026.
6. Li H., Zhang Y., Chen X. Multi-Modal Sentiment Analysis Based on Image and Text Using Cross-Attention Mechanism // Electronics. – 2024. – Vol. 13, № 11. – Art. 2069. – DOI: https://doi.org/10.3390/electronics13112069 - Дата обращения: 12.03.2026.
7. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Электронный ресурс] // arXiv preprint. – 2018. – URL: https://doi.org/10.48550/arXiv.1810.04805 – Дата обращения: 12.03.2026.
8. Aggarwal C. C. Neural Networks and Deep Learning [Электронный ресурс]. – Cham: Springer, 2018. – URL: https://link.springer.com/book/10.1007/978-3-319-94463-0 – Дата обращения: 12.03.2026.
9. Ясницкий Л. Н. Интеллектуальные системы: учебник / Л. Н. Ясницкий. – Москва: Лаборатория знаний, 2020. – 224 с. – ISBN 978-5-00101-897-1.
10. Абдрахманов Б. С., Омаров Б. Б. Интеллектуальный анализ данных: учебное пособие. – Алматы: Қазақ университеті, 2021. – 240 с.
11. Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition [Электронный ресурс] – 3rd ed. – Stanford University, 2026. – URL: https://web.stanford.edu/~jurafsky/slp3/ – Дата обращения: 12.03.2026.
12. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 770-778. – DOI: https://doi.org/10.1109/CVPR.2016.90
13. Урмашев Б. А. Information-communication technology: учебное пособие / Б. А. Урмашев. – Алматы: Қазақ университеті, 2017. – 336 с.
14. Gandhi A., Adhvaryu K., Poria S., Cambria E., Hussain A. Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions // Information Fusion. – 2023. – Vol. 91. – P. 424-444. – DOI: https://doi.org/10.1016/j.inffus.2022.09.025 – Дата обращения: 12.03.2026.
15. Хорошевский В. Ф. Искусственный интеллект: учебник / В. Ф. Хорошевский. – Москва: Питер, 2018. – 496 с.
16. Нуртазин К. К., Сагындыков Ж. С. Анализ данных и интеллектуальные системы: учебное пособие / К. К. Нуртазин, Ж. С. Сагындыков. – Алматы: КазНУ им. аль-Фараби, 2020. – 296 с.
17. Флах П. Машинное обучение: наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах; пер. с англ. – Москва: ДМК Пресс, 2015. – 400 с.
18. Гаврилов А. В. Машинное обучение: методы и алгоритмы: учебное пособие / А. В. Гаврилов. – Москва: МГТУ им. Н. Э. Баумана, 2019. – 312 с.
19. Омаров А. Т., Нурпеисов Д. М. Машинное обучение и анализ данных: учебное пособие / А. Т. Омаров, Д. М. Нурпеисов. – Астана: ЕНУ им. Л. Н. Гумилёва, 2021. – 284 с.
20. Кожахметов Е. Б. Обработка больших данных: учебное пособие / Е. Б. Кожахметов. – Алматы: КазНИТУ им. К. И. Сатпаева, 2022. – 300 с.

References:

1. Baltrušaitis T., Ahuja C., Morency L. Multimodal Machine Learning: A Survey and Taxonomy // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2019. – DOI: https://doi.org/10.1109/TPAMI.2018.2798607
2. Bishop C. M. Pattern Recognition and Machine Learning [Elektronnyj resurs]. – New York: Springer, 2006. – URL: https://link.springer.com/book/10.1007/978-0-387-45528-0 – Data obrashcheniya: 12.03.2026.
3. Radford A., Kim J. W., Hallacy C. i dr. Learning Transferable Visual Models from Natural Language Supervision [Elektronnyj resurs] // arXiv preprint. – 2021. – arXiv: 2103.00020. – URL: https://doi.org/10.48550/arXiv.2103.00020 – Data obrashcheniya: 12.03.2026.
4. Russell S., Norvig P. Artificial Intelligence: A Modern Approach [Elektronnyj resurs]. – 4th ed. – Pearson, 2020. - URL: https://aima.cs.berkeley.edu/ – Data obrashcheniya: 12.03.2026.
5. Ngiam J., Khosla A., Kim M., Nam J., Lee H., Ng A. Y. Multimodal Deep Learning // Proceedings of the 28th International Conference on Machine Learning (ICML-11). – 2011. – P. 689-696. – URL:

https://robotics.stanford.edu/~ang/papers/icml11-MultimodalDeepLearning.pdf – Data obrashcheniya: 12.03.2026.
6. Li H., Zhang Y., Chen X. Multi-Modal Sentiment Analysis Based on Image and Text Using Cross-Attention Mechanism // Electronics. – 2024. – Vol. 13, № 11. – Art. 2069. – DOI: https://doi.org/10.3390/electronics13112069 - Data obrashcheniya: 12.03.2026.
7. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Elektronnyj resurs] // arXiv preprint. – 2018. – URL: https://doi.org/10.48550/arXiv.1810.04805 – Data obrashcheniya: 12.03.2026.
8. Aggarwal C. C. Neural Networks and Deep Learning [Elektronnyj resurs]. – Cham: Springer, 2018. – URL: https://link.springer.com/book/10.1007/978-3-319-94463-0 – Data obrashcheniya: 12.03.2026.
9. YAsnickij L. N. Intellektual'nye sistemy: uchebnik / L. N. YAsnickij. – Moskva: Laboratoriya znanij, 2020. – 224 s. – ISBN 978-5-00101-897-1.
10. Abdrahmanov B. S., Omarov B. B. Intellektual'nyj analiz dannyh: uchebnoe posobie. – Almaty: Қазақ universiteti, 2021. – 240 s.
11. Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition [Elektronnyj resurs] – 3rd ed. – Stanford University, 2026. – URL: https://web.stanford.edu/~jurafsky/slp3/ – Data obrashcheniya: 12.03.2026.
12. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 770-778. – DOI: https://doi.org/10.1109/CVPR.2016.90
13. Urmashev B. A. Information-communication technology: uchebnoe posobie / B. A. Urmashev. – Almaty: Қазақ universiteti, 2017. – 336 s.
14. Gandhi A., Adhvaryu K., Poria S., Cambria E., Hussain A. Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions // Information Fusion. – 2023. – Vol. 91. – P. 424-444. – DOI: https://doi.org/10.1016/j.inffus.2022.09.025 – Data obrashcheniya: 12.03.2026.
15. Horoshevskij V. F. Iskusstvennyj intellekt: uchebnik / V. F. Horoshevskij. – Moskva: Piter, 2018. – 496 s.
16. Nurtazin K. K., Sagyndykov ZH. S. Analiz dannyh i intellektual'nye sistemy: uchebnoe posobie / K. K. Nurtazin, ZH. S. Sagyndykov. – Almaty: KazNU im. al'-Farabi, 2020. – 296 s.
17. Flah P. Mashinnoe obuchenie: nauka i iskusstvo postroeniya algoritmov, kotorye izvlekayut znaniya iz dannyh / P. Flah; per. s angl. – Moskva: DMK Press, 2015. – 400 s.
18. Gavrilov A. V. Mashinnoe obuchenie: metody i algoritmy: uchebnoe posobie / A. V. Gavrilov. – Moskva: MGTU im. N. E. Baumana, 2019. – 312 s.
19. Omarov A. T., Nurpeisov D. M. Mashinnoe obuchenie i analiz dannyh: uchebnoe posobie / A. T. Omarov, D. M. Nurpeisov. – Astana: ENU im. L. N. Gumilyova, 2021. – 284 s.
20. Kozhahmetov E. B. Obrabotka bol'shih dannyh: uchebnoe posobie / E. B. Kozhahmetov. – Almaty: KazNITU im. K. I. Satpaeva, 2022. – 300 s.

Information about the authors

M.K.Serikov – corresponding author, Senior Lecturer, School of Engineering and Information Technologies, Master of Technical Sciences, META University, Almaty, Kazakhstan; e-mail: 7014547793@mail.ru.