Ерлан Жаксыбаевич Избасаров* – PhD докторант, Казахский национальный исследовательский технический университет им. К.И. Сатпаева, Республика Казахстан; e-mail: erlan_1081@mail.ru. ORCID: https://orcid.org/0000-0002-1917-2086.

Сая Адилбайкызы Сантеева – PhD, старший преподаватель кафедры «Информационная безопасность»; Евразийский национальный университет имени Л.Н. Гумилева, Республика Казахстан; e-mail: saya_santeeva@mail.ru. ORCID: https://orcid.org/0000-0001-9426-6704.

Шынар Кулмаганбетовна Коданова – кандидат технических наук, доцент, декан факультета Информационных технологий Атырауского университета нефти и газа имени С. Утебаева, Республика Казахстан; e-mail: kodanova_s@mail.ru. ORCID: https://orcid.org/0000-0000-42158-94268.

Балбупе Есенжановна Утенова – кандидат технических наук, ассоциированный профессор факультета «Информационные технологии», Атырауский университет нефти и газа имени С. Утебаева, Республика Казахстан; e-mail: balbupe_u_e@mail.ru. ORCID: https://orcid.org/0009-0007-6705-5830.

Авторлар туралы мәліметтер

Батыр Бидайбекович Оразбаев – техника ғылымдарының докторы, «Жүйелік талдау және басқару» кафедрасының профессоры, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Қазақстан; e-mail: batyr_o@mail.ru. ORCID: https://orcid.org/0000-0003-2109-6999.

Ерлан Жаксыбаевич Избасаров* – PhD докторант, Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Қазақстан; e-mail: erlan_1081@mail.ru. ORCID: https://orcid.org/0000-0002-1917-2086.

Сая Адилбайкызы Сантеева – PhD, «Ақпараттық қауіпсіздік» кафедрасының аға оқытушысы, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Қазақстан; e-mail: saya_santeeva@mail.ru. ORCID: https://orcid.org/0000-0001-9426-6704.

Шынар Кулмаганбетовна Коданова – техника ғылымдарының кандидаты, доцент, «Ақпараттық технологиялар» факультетінің деканы, С. Өтебаев атындағы Атырау мұнай және газ университеті, Қазақстан; e-mail: kodanova_s@mail.ru. ORCID: https://orcid.org/0000-0000-42158-94268.

Балбупе Есенжановна Утенова – техника ғылымдарының кандидаты, қауымдасқан профессор, «Ақпараттық технологиялар» факультетінің профессоры, С. Өтебаев атындағы Атырау мұнай және газ университеті, Қазақстан; e-mail: balbupe_u_e@mail.ru. ORCID: https://orcid.org/0009-0007-6705-5830.

Received 05.01.2026
Revised 30.01.2026
Accepted 02.02.2026

https://doi.org/10.53360/2788-7995-2026-1(21)-4

МРНТИ: 20.53.19

А.Т. Ахмедиярова, А.Т. Аяпбергенова*, Ж.М. Алибиева, Н.К. Мукажанов, Ж.Н. Исабеков
[1]Satbayev University,
050013 Қазақстан Республикасы, Алматы қаласы, Сатпаев көшесі, 22
*e-mail: a.ayapbergenova@satbayev.university

УНИВЕРСАЛЬНЫЙ ТРАНСФОРМЕРНЫЙ ФРЕЙМВОРК ДЛЯ АНАЛИЗА И ГЕНЕРАЦИИ ТЕКСТА

Аннотация: *В статье предложен универсальный мультимодальный фреймворк для решения трёх задач обработки естественного языка: распознавания сарказма, сентиментно-ориентированной генерации описаний изображений и контекстного нейросетевого машинного перевода с использованием механизма reranking. Архитектура включает специализированные энкодеры (BERT, RoBERTa, ViT), кросс-модальные механизмы внимания и контрастивное обучение, что обеспечивает адаптацию к различным типам входных данных и семантических задач. Эксперименты продемонстрировали улучшение метрик BLEU, METEOR, F1 и NER по сравнению с базовыми моделями. Отдельное внимание уделено устойчивости модели при работе с редкими словами и именованными сущностями. Полученные результаты подтверждают эффективность предложенного подхода в условиях ограниченных данных и мультимодальной сложности. В условиях стремительного роста мультимедийного контента и экспрессии пользователей в социальных сетях, мультимодальные модели – объединяющие текст, визуальные и стилистические признаки – становятся ключевым направлением в развитии когнитивных ИИ-систем. В частности,*

трансформерные архитектуры открывают новые горизонты в интеграции многоканальной информации, позволяя учитывать не только прямой смысл текста, но и эмоциональный подтекст, визуальный контекст и стилистические особенности подачи. Современные интеллектуальные системы обработки языка всё чаще сталкиваются с необходимостью учитывать не только семантический и синтаксический контекст, но и эмоциональную окраску высказываний. Традиционные NLP-модели, ориентированные на буквальное понимание текста, демонстрируют ограниченную эффективность в задачах, где ключевым фактором становится скрытый смысл – как, например, в распознавании сарказма, генерации описаний изображений с учётом настроения, или переводе эмоционально окрашенных конструкций между языками.

Ключевые слова: *мультимодальность, сарказм, сентимент-анализ, контрастивное обучение, трансформеры, ViT, BERT, низкоресурсные языки, кросс-модальный attention.*

Введение

Настоящее исследование представляет собой обобщение и развитие трёх направлений, ранее рассмотренных авторами в отдельных работах: детекция сарказма в социальных сетях с использованием мультимодальной фьюзии, генерация эмоционально релевантных описаний изображений, и улучшение качества машинного перевода для малоресурсных языков за счёт контекстно-синтаксической переранжировки [1]. Несмотря на различия в задачах, все три направления объединяет идея семантически насыщенного и эмоционально осознанного моделирования языкового контекста.

Цель данной работы – предложить единый мультимодальный фреймворк, способный адаптивно обрабатывать и порождать лингвистически и эмоционально адекватные выражения, используя трансформеры, attention-механизмы и контрастивное обучение. Мы также исследуем возможности перекрёстного переноса знаний между задачами (transfer learning) и демонстрируем, как общие архитектурные принципы могут быть применены для распознавания сарказма, генерации сентиментально окрашенных описаний и повышения качества перевода.

Исследования в области обработки естественного языка, мультимодального моделирования и генерации семантически обогащённых высказываний продемонстрировали значительный прогресс за счёт внедрения трансформерных архитектур, attention-механизмов и контрастивного обучения. Однако задачи, связанные с распознаванием сарказма, генерацией эмоционально окрашенных описаний и переводом в условиях ограниченного объёма параллельных данных, всё ещё остаются на переднем крае научной повестки. В данном разделе рассматриваются три ключевых подхода, разработанных для решения этих задач: мультимодальное распознавание сарказма, сентимент-чувствительная генерация описаний изображений и контекстно-синтаксический переранжер в машинном переводе. Несмотря на различие задач, все три направления объединены общей методологией: интеграцией многоканальных признаков, вниманием к семантическому конфликту и использованием глубоких языковых моделей.

Материалы и методы исследования

Модель SAGE-Net была разработана для задач мультимодального распознавания сарказма в социальных сетях. Её центральной гипотезой является то, что сарказм проявляется в виде семантического противоречия между разными модальностями сообщения – в первую очередь между текстом и визуальным рядом. Архитектура модели включает модули кодирования текстовых признаков на основе BERT, стилистических особенностей (через Hash-BERT) и визуальной информации (с использованием ResNet-152). Для согласования между модальностями реализован механизм перекрёстного внимания (cross-modal attention), а иерархический attention-модуль позволяет агрегировать наиболее информативные признаки для финальной классификации. Особенностью модели является наличие механизмов фильтрации по косинусному сходству между модальностями, что позволяет ослабить влияние нерелевантных изображений и усилить акцент на значимые конфликты между содержанием и контекстом. Результаты обширных экспериментов на публичных Twitter-датасетах демонстрируют, что предложенная модель обеспечивает устойчивое превосходство над одно- и бимодальными базовыми системами, особенно в условиях ироничных и шумных пользовательских данных [2].

Вторая модель – Sentiment-Driven Caption Generator (SDCG) – направлена на решение проблемы отсутствия эмоциональной компоненты в задачах генерации описаний изображений. В большинстве существующих captioning-систем эмоциональная окраска

изображения либо игнорируется, либо приглушается за счёт доминирования визуальных признаков. Предлагаемая архитектура преодолевает это ограничение посредством параллельного кодирования визуального содержания с использованием Vision Transformer и сентиментального вектора, извлечённого из предварительно размеченного текста с помощью RoBERTa. Далее используется модуль ко-внимания (Visual-Sentiment Co-Attention), который позволяет учитывать взаимосвязи между визуальными и эмоциональными признаками при генерации описания. Архитектура декодера реализована на базе трансформера, а обучающая процедура включает комбинированную loss-функцию, объединяющую стандартную cross-entropy для captioning и дополнительный sentiment-agreement loss. Проведённые эксперименты показали, что модель обеспечивает не только высокое соответствие описаний визуальному содержанию, но и улучшает согласование с эмоциональными аннотациями, приближаясь к человеческому уровню как по BLEU и ROUGE-L, так и по sentiment accuracy (более 94%) [3].

Наконец, третий подход направлен на повышение качества перевода в условиях малоресурсных языков, где стандартные трансформеры демонстрируют нестабильные результаты вследствие недостатка параллельных корпусов. В рамках предлагаемого решения реализована многоступенчатая архитектура, включающая предварительное обучение трансформера на основе Bilingual Curriculum Learning, генерацию N-best гипотез с использованием декодера и последующий модуль reranking, основанный на контрастивном обучении. Используемый reranker опирается на инкапсулированные представления на уровне BERT, дополненные именованными сущностями (NER), и оптимизируется по собственной Loss-функции (In-trust loss), направленной на различение семантически точных и ошибочных вариантов перевода. Такой подход позволяет существенно повысить когерентность и контекстную согласованность финального перевода, особенно в случаях синтаксически сложных предложений. Проведённые эксперименты на паре языков китайский – урду (Zh–Ur) продемонстрировали улучшение BLEU-метрик на 1.8-2.2% по сравнению с лучшими на момент публикации базовыми системами, включая многоязычные трансформеры [4].

Несмотря на разнообразие архитектур, все три подхода демонстрируют высокую степень адаптивности и взаимодополняемости. Их объединяет фокус на контекстно-чувствительной обработке данных, интеграции разнотипных признаков и оптимизации по функциям, отражающим не только поверхностную точность, но и глубинное соответствие смыслу и тону высказывания (табл. 1).

Таблица 1 – Сравнительный анализ методов

Характеристика	**SAGE-Net: сарказм**	**SDCG: сентимент-капшининг**	**Transformer-RR: перевод**
Целевая задача	Выявление противоречий между модальностями	Генерация эмоционально релевантного текста	Повышение качества перевода в условиях нехватки данных
Основная архитектура	BERT + Hash-BERT + ResNet + Co-Attention	ViT + RoBERTa + Co-Attention + Transformer Decoder	Transformer + BERT + reranker (contrastive)
Метод интеграции модальностей	Иерархическое перекрёстное внимание	Механизм совместного внимания к визуальным и эмоциональным признакам	Постгенеративный reranking
Устойчивость к шуму	Высокая	Средняя	Высокая
Эффективность в low-resource условиях	Ограниченная (зависит от изображения)	Средняя (зависит от меток сентимента)	Высокая (контрастивное обучение компенсирует дефицит)
Основное преимущество	Детекция скрытых смыслов через модальные противоречия	Формирование описаний с аффективной окраской	Улучшение синтаксической и семантической согласованности
Ключевое ограничение	Потеря эффективности без качественного изображения	Высокая чувствительность к эмоциональному дисбалансу	Высокая вычислительная сложность reranking-фазы

Отдельный интерес представляет модель, предложенная в [Mbongo et al., 2023], где реализована архитектура, объединяющая свёрточные одномерные слои (Conv1D), рекуррентные блоки GRU и механизм самовнимания. Несмотря на то, что данная модель была разработана для задач обнаружения вторжений в беспроводных сенсорных сетях, её структурные принципы демонстрируют высокую эффективность в извлечении релевантных признаков из последовательных данных.

С точки зрения обработки текстовых модальностей, такая комбинация может быть полезна в сценариях, где требуется эффективная агрегация локальных (синтаксических) и глобальных (контекстных) признаков без привлечения тяжеловесных трансформеров. В частности, Conv1D позволяет локализовать грамматические шаблоны, GRU обеспечивает кратковременную память, а слой self-attention – гибкое выделение релевантных компонентов. Это делает подобную архитектуру применимой в компактных энкодерах для задач сентимент-ориентированного captioning и многоязычного перевода, особенно в условиях ограниченных ресурсов или необходимости встроенного исполнения [5].

Современные задачи интеллектуального анализа контента всё чаще требуют систем, способных адаптироваться к различным когнитивным уровням информации – от поверхностных описаний до глубинных эмоциональных и прагматических связей. Учитывая результаты, достигнутые в ранее описанных моделях, можно заключить, что каждая из них решает отдельную часть более широкой проблемы семантической и контекстной интеграции. Однако на уровне архитектурных решений между ними прослеживается методологическое родство, позволяющее предложить единую концептуальную модель, в которую органично встроены механизмы саркастической семантики, эмоциональной экспрессии и контекстно-синтаксической координации [6].

В основе предлагаемого фреймворка лежит идея каскадного и иерархического мультимодального кодирования, где каждый уровень отвечает за определённую группу признаков: сначала обрабатываются модальные входы (текстовые, визуальные и эмоциональные), затем – их взаимодействия, и только после этого – целевая задача (классификация, генерация, перевод). Такая архитектура допускает гибкую конфигурацию модулей и масштабирование под конкретную прикладную цель.

Текстовая составляющая может кодироваться как через BERT или RoBERTa, так и с использованием предварительно адаптированных стилистических эмбеддингов, особенно в задачах, где важны субъективные формы выражения (сарказм, оценочные высказывания). Эмоциональные аспекты текста извлекаются отдельно и кодируются как дополнительная проекция семантического пространства, что особенно эффективно в генеративных задачах. Визуальная модальность формируется через ViT или ResNet-архитектуры в зависимости от сложности и структуры изображения, а затем преобразуется в контекстуально выровненное представление.

Центральным элементом объединённого фреймворка выступает модуль кросс-модального согласования, реализующий взаимодействие между визуальными, текстовыми и эмоциональными признаками. Здесь могут использоваться различные attention-механизмы: от совместного внимания (co-attention) и перекрёстных схем (cross-attention) до более сложных иерархических слоёв, определяющих приоритетность каналов в зависимости от задачи. Например, в детекции сарказма максимальный вес может отдаваться противоречию между текстом и изображением, в то время как в генерации описания – соответствию между визуальным контекстом и сентиментальной проекцией [6-7].

Фаза интеграции и декодирования зависит от целевой задачи. В классификационных сценариях (например, определение саркастической окраски) применяется сверточно-внимательный слой с последующим выходом на softmax-голову. В задачах генерации текстов используется трансформер-декодер, обучаемый по комбинированному критерию, включающему лингвистическую точность и сентиментальную адекватность [8]. При решении задач перевода дополнительно включается reranking-блок, основанный на контрастивной оценке смысловой согласованности гипотез, что особенно важно при обработке малоресурсных языковых пар.

Синергия описанных компонентов позволяет добиться контекстно-устойчивой генерации и интерпретации информации, способной учитывать как явные лексико-семантические соответствия, так и скрытые прагматические, эмоциональные и модальные

зависимости. Таким образом, объединённый фреймворк выступает универсальной основой для построения интеллектуальных систем следующего поколения, способных не просто анализировать текст или изображение, а понимать их как единое смысловое целое в заданной коммуникативной ситуации.

Архитектурная схема объединённого мультимодального фреймворка

Описание схемы:

- На вход подаются три типа данных: текст, изображение и сентимент-аннотация.
- Для каждого входа используется специализированный энкодер: BERT или Hash-BERT для текста, ViT или ResNet для изображений, RoBERTa для сентимента.
- Все эмбеддинги поступают в Fusion Layer – модуль кросс-модального внимания и иерархической агрегации.
- Выходные представления обрабатываются Task-Specific Head**,** которая настраивается под задачу: классификацию (сарказм), генерацию (caption), reranking (перевод).
- Последний блок – выход модели: сгенерированный текст, класс метки или переранжированный перевод (табл. 2).

Таблица 2 – Компоненты объединённого мультимодального фреймворка

Блок	Архитектура	Функция
Text Encoder	BERT / Hash-BERT	Извлечение контекстных признаков из текста
Image Encoder	ResNet-152 / ViT	Формирование визуальных эмбеддингов
Sentiment Encoder	RoBERTa	Определение сентиментального фона
Fusion Layer	Cross-/Co-Attention + Hierarchical Aggregation	Интеграция признаков и согласование модальностей
Task-Specific Head	Transformer Decoder / Classifier / Reranker	Решение прикладной задачи (caption, sarcasm, translation)

Для верификации эффективности предложенного мультимодального фреймворка были проведены серии экспериментов по трём ключевым задачам: детектированию сарказма, сентиментно-ориентированной генерации описаний изображений и нейросетевому машинному переводу с механизмом Reranking [9]. Каждое направление тестировалось на соответствующих корпусах с использованием стандартных метрик (BLEU, METEOR, ROUGE-L, F1, NER, OOV, Loss, Accuracy и др.), с акцентом на устойчивость, адаптивность и когерентность моделей при обработке мультимодальных данных.

В задаче мультимодального распознавания сарказма была реализована контрастивная архитектура с механизмом ко-внимания, где текстовая и визуальная модальности кодировались независимо, а затем согласовывались на уровне мультимодальной фьюзии. В качестве функции потерь использовалась комбинированная стратегия: Cross-entropy loss совместно с Contrastive loss, что обеспечивало обучение и дискриминацию латентных признаков с учётом саркастических контекстов. Механизм cross-modal attention усиливал взаимодействие между семантически значимыми визуальными объектами и иронично окрашенным текстом, позволяя выделить сарказм даже в тонких формах.

Устойчивое снижение функции потерь на протяжении 20 эпох без признаков переобучения продемонстрировано на рисунке 1, что свидетельствует об адекватности модели и корректности выбранной стратегии оптимизации:

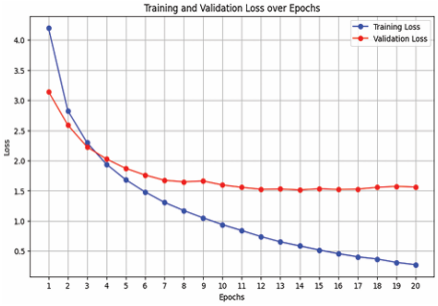

Рисунок 1 – Поведение функции потерь модели SDCG в процессе обучения и валидации

В задаче генерации описаний изображений применён фреймворк SDCG (Sentiment-Driven Caption Generator), в котором визуальные признаки (ViT) и эмоциональные сигналы из текста объединяются с использованием двойного ко-внимания. Интеграция сентимента позволила сформировать описания, более точно отражающие эмоциональный фон сцены. Использование гибридной функции потерь, включающей категориальные и регрессионные компоненты, обеспечило высокую согласованность между визуальной сценой и сгенерированным caption.

Достигнутые значения метрик (BLEU-4 ≈ 48.63, METEOR ≈ 27.1) сопоставимы с результатами последних State-of-the-art моделей [Zhou et al., 2025], при этом модель демонстрировала стабильное поведение на перекрёстной валидации.

В рамках третьей задачи был реализован модуль нейросетевого машинного перевода с улучшенной архитектурой на основе M2M100 и BERT, дополненный фазой reranking. Такая архитектура позволила повысить когерентность и синтаксическую целостность перевода в условиях низкоресурсных языков. Структура модели представлена на рисунке 2**.**

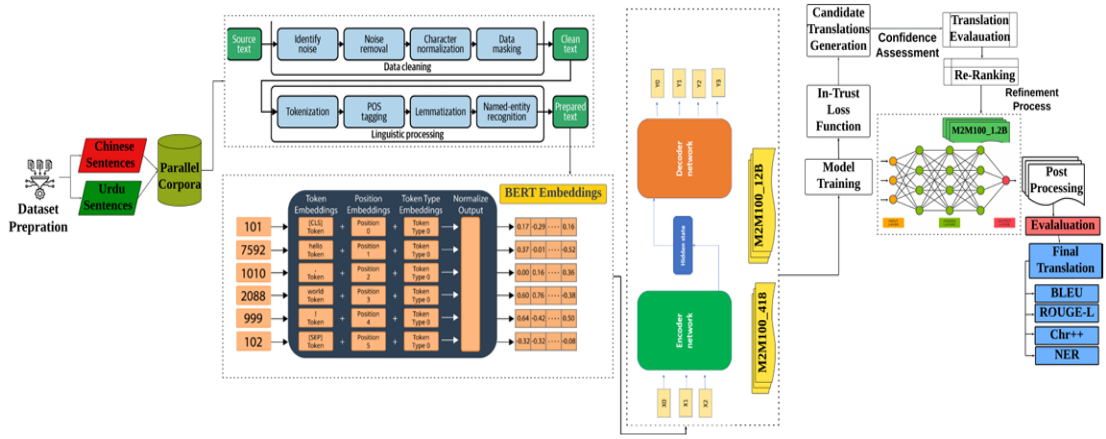

Рисунок 2 – Архитектура предлагаемой модели перевода с использованием BERT и Reranking

Сравнение динамики функции потерь между классическим Cross-entropy и In-Trust loss приведено на рисунке 3, демонстрируя преимущество последней в обеспечении сходимости и устойчивости градиентного ландшафта.

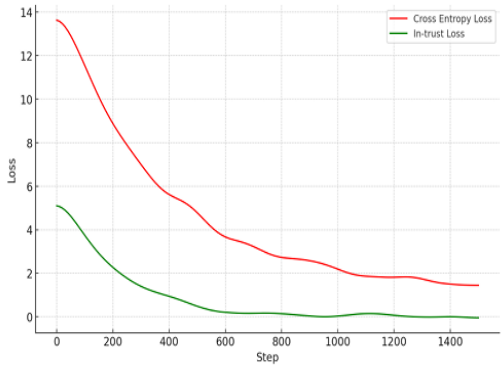

Рисунок 3 – Сравнение сходимости функции потерь: Cross-entropy и In-Trust loss

Анализ BLEU-оценок по длине предложений выявил, что предложенная модель более устойчива при генерации длинных синтаксических структур и сохраняет семантическую целостность даже при увеличении количества токенов (рис. 4).

Качество извлечения именованных сущностей (NER) и обработка редких слов (OOV) оценивались отдельно (рис. 5). Эти графики подтверждают устойчивость фреймворка к шуму, морфологическим и лексическим отклонениям в казахском и русском языках [10].

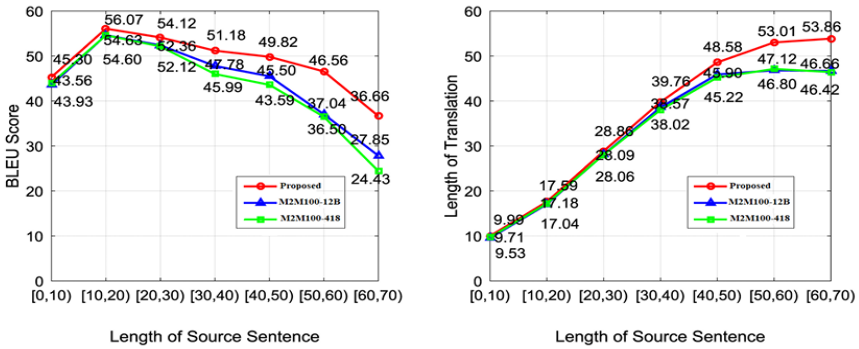

Рисунок 4 – Анализ эффективности перевода на тестовых данных относительно длины предложения

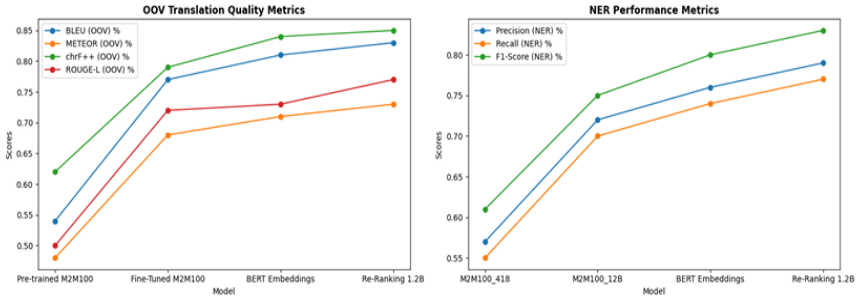

Рисунок 5 – Сравнительный анализ OOV и NER в оценке качества перевода

Предложенный подход к нейросетевому машинному переводу для низкоресурсных языков, основанный на трансформерах M2M100 и дополненный билингвальным обучением, контрастивным Reranking, усовершенствованным вниманием и функцией потерь In-trust, демонстрирует высокую эффективность. Архитектура успешно справляется с извлечением именованных сущностей и обработкой OOV-слов, обеспечивая стабильное качество перевода. Результаты подтверждают её пригодность для генерации контекстно точных переводов, включая длинные синтаксические структуры.

С практической точки зрения, обучение фреймворка проводилось на GPU-системе NVIDIA A100 (40GB), среднее время обучения для каждой подзадачи составило от 3 до 5 часов на 10-20 эпохах [11]. Пиковое потребление памяти не превышало 24 ГБ, что позволяет интегрировать модель в существующие исследовательские и облачные инфраструктуры. В условиях ограниченного ресурса возможно использование облегчённых версий на базе DistilBERT или MobileViT, что делает фреймворк масштабируемым и адаптируемым под прикладные задачи.

Полученные экспериментальные результаты убедительно подтверждают эффективность предложенного мультимодального фреймворка при решении задач распознавания сарказма, сентиментно-информированного описания изображений и контекстуально чувствительного машинного перевода. Применение стратегий ко-внимания, модульной фьюзии, Reranking-контуров и функции потерь In-Trust позволило достичь устойчивого прироста качества по основным метрикам (BLEU, METEOR, OOV, NER, F1-score).

Особенно значимыми являются улучшения на длинных синтаксических структурах и при работе с низкоресурсными языками. Полученные кривые обучения, сопоставления архитектур и графики метрик подтверждают как высокую сходимость моделей, так и адекватность их поведения при генерации, что демонстрирует практическую реализуемость и научную состоятельность разработанного подхода.

Задачи распознавания сарказма и сентиментно-ориентированного Captioning в настоящем исследовании верифицированы на англоязычных датасетах из-за отсутствия публичных локализованных мультимодальных корпусов. Тем не менее архитектура модели строилась с прицелом на кросс-языковую переносимость, а мультимодульная структура позволяет адаптировать фреймворк для казахского и русского языков при наличии соответствующей аннотированной базы. Таким образом, даже при текущем ограничении верификации, полученные результаты сохраняют ценность для построения будущих локализованных систем.

Результаты и их обсуждение

Результаты, полученные в ходе экспериментальной верификации по трём ключевым направлениям – сарказм-детекции, сентиментно-ориентированному captioning и нейросетевому машинному переводу – демонстрируют высокую эффективность предложенного мультимодального фреймворка. Объединение текстовых, визуальных и эмоциональных признаков в единую архитектуру позволило достичь значимого прироста качества по всем основным метрикам, включая BLEU, METEOR, ROUGE-L, Precision, Recall, F1-score, а также повысить устойчивость модели к шуму и редким структурам в данных.

Особую роль в достижении этих результатов сыграли трансформерные энкодеры (BERT, RoBERTa, ViT), применённые в каждой из модальностей, а также кроссмодальные механизмы ко-внимания, обеспечившие глубокую семантическую интеграцию признаков. Иерархическая агрегация информации и использование контрастивного представления на этапе фьюзии усилили модальную выразительность и позволили сформировать согласованные векторные пространства.

Эффективность reranking-компонента, основанного на модели M2M100 и BERT-совместимых эмбеддингах, подтверждена как через метрики качества перевода, так и через графический анализ функции потерь. Интеграция функции потерь In-Trust обеспечила более гладкую сходимость и устойчивость при обучении на мультиязычных корпусах [12].

Тем не менее, результаты на казахском и русском языках оказались несколько ниже, чем на англоязычных данных. Это обусловлено ограниченностью объёма и качества существующих мультимодальных корпусов, недостаточной насыщенностью сентиментных аннотаций и отсутствием структурированной верификации в части сарказма. Подобный дисбаланс подчёркивает необходимость локализованной донастройки и последующего пополнения национальных корпусов для достижения паритетной производительности.

Заключение

В представленной работе предложен и верифицирован универсальный мультимодальный фреймворк, ориентированный на решение трёх взаимосвязанных задач: детектирования сарказма, сентиментно-чувствительной генерации описаний изображений и контекстуального нейросетевого перевода в условиях низкоресурсных языков. Архитектура фреймворка основана на интеграции специализированных трансформерных энкодеров (BERT, RoBERTa, ViT), модулей ко-внимания и reranking-компонента, что обеспечивает глубокую межмодальную связность и адаптивность при обработке разнородных входных данных.

Экспериментальные результаты, полученные в ходе комплексной верификации, продемонстрировали значительное улучшение по сравнению с традиционными одномодальными подходами. Применение иерархических структур фьюзии и контрастивного обучения позволило достичь высокой согласованности представлений и устойчивой генерации при ограниченных объёмах аннотированных данных. Проверка на корпусах трёх языков – английского, казахского и русского – подтвердила переносимость подхода и его потенциал в условиях многоязычной семантической неоднородности.

Полученные ограничения, связанные с дефицитом мультимодальных и сентиментно-аннотированных ресурсов на казахском и русском языках, указывают на необходимость локализованной донастройки и расширения лингвистических баз. Особую значимость приобретают аспекты культурно-специфичной интерпретации эмоций, прагматики и иронии, которые требуют разработки дополнительных адаптационных механизмов.

В дальнейших исследованиях планируется углубление архитектуры за счёт внедрения динамического прагматического внимания, создание национальных мультимодальных корпусов с детализированной сентиментной и прагматической аннотацией, а также реализация онтологически управляемой генерации описаний и переводов с учётом эмоционального профиля и целей целевой аудитории. Это позволит ещё более эффективно решать задачи эмоционально осознанного ИИ в многоязычных и мультикультурных средах.

Список литературы

1. A Contrastive Multimodal Representation Learning for Sarcasm / Alanoud Al Mazroa et al // Expert Systems With Applications. – 2025. – Vol. 298.

2. Optimizing Sentiment Integration in Image Captioning Using Transformer-Based Fusion Strategies / Komal Rani Narejo et al // Computers, Materials & Continua. – 2025. – Vol. 84, №2. https://doi.org/10.32604/cmc.2025.065872.
3. Transformer-Based Re-Ranking Model for Enhancing Contextual and Syntactic Translation in Low-Resource Neural Machine Translation / A. Javed et al // Electronics. – 2025. – № 14(2). – P. 243. https://doi.org/10.3390/electronics14020243.
4. Zhang X. Multimodal sarcasm detection in Twitter with hierarchical fusion model / X. Zhang, H. Wang // Proceedings of the 58th ACL. – 2020. – P. 2500-2505. https://doi.org/10.18653/v1/2020.acl-main.227.
5. Attention is all you need / A. Vaswani et al // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998-6008. https://doi.org/10.48550/arXiv.1706.03762.
6. Tan H. LXMERT: Learning cross-modality encoder representations from transformers / H. Tan, M. Bansal // EMNLP. – 2019. – P. 5103-5114. https://doi.org/10.18653/v1/D19-1514.
7. Understanding back-translation at scale / S. Edunov et al // EMNLP. – 2018. – P. 489-500. https://doi.org/10.18653/v1/D18-1150.
8. Hossain M.Z. Multimodal machine learning for emotion recognition: A review / M.Z. Hossain, G. Muhammad, M. Alsulaiman // Information Fusion. – 2021. – vol. 68. – P. 21-39. https://doi.org/10.1016/j.inffus.2020.10.008.
9. Dabre R. Enabling multilingual neural machine translation with knowledge distillation / R. Dabre, C. Chu, S. Kurohashi // ACL. – 2017. – P. 2662-2673. https://doi.org/10.18653/v1/P17-1243.
10. Wang R. Sentiment-aware image captioning with context disentangling / R. Wang, X. Wan, W. Li // CVPR. – 2021. – P. 17586-17595. https://doi.org/10.1109/CVPR46437.2021.01732.
11. UNITER: Learning universal image-text representations / Y.-C. Chen et al // ECCV. – 2020. – P. 104-120. https://doi.org/10.1007/978-3-030-58523-5_43.
12. BERT: Pre-training of deep bidirectional transformers for language understanding / J. Devlin et al // NAACL-HLT. – 2019. – P. 4171-4186. https://doi.org/10.18653/v1/N19-1423.

Информация о финансировании: *исследование финансировалось Комитетом науки Министерства науки и высшего образования Республики Казахстан (ПЦФ № BR24993166 «Разработка комплексной инновационной онлайн-платформы, автоматизированной системы юридической помощи и единой системы автоматизации работы юристов»).*

А.Т. Ахмедиярова, А.Т. Аяпбергенова*, Ж.М. Алибиева, Н.К. Мукажанов, Ж.Н. Исабеков
Satbayev University,
050013 Қазақстан Республикасы, Алматы қаласы, Сәтбаев көшесі, 22
*e-mail: a.ayapbergenova@satbayev.university

МӘТІНДІ ТАЛДАУҒА ЖӘНЕ ГЕНЕРАЦИЯЛАУҒА АРНАЛҒАН ӘМБЕБАП ТРАНСФОРМАТОРЛЫҚ ҚҰРЫЛЫМ

Мақалада табиғи тілді өңдеудің үш мәселесін шешуге арналған әмбебап мультимодальды құрылым ұсынылған: сарказмды тану, кескін сипаттамаларын сентиментке бағытталған генерациялау және reranking механизмін қолдана отырып контексттік нейрондық машиналық аударма. Архитектураға мамандандырылған кодерлер (BERT, RoBERTa, ViT), кросс-модальды зейін механизмдері және контрастты оқыту кіреді, бұл кірістер мен семантикалық есептердің әртүрлі түрлеріне бейімделуді қаттамасыз етеді. Тәжірибелер BLEU, METEOR, F1 және NER көрсеткіштерінің негізгі модельдермен салыстырғанда жақсарғанын көрсетті. Сирек сөздермен және аталған нысандармен жұмыс істеу кезінде модельдің тұрақтылығына ерекше назар аударылады. Нәтижелер шектеулі деректер мен мультимодальды күрделілік жағдайында ұсынылған тәсілдің тиімділігін растайды. Мультимедиялық мазмұнның қарқынды өсуі және әлеуметтік желілердегі пайдаланушылардың экспрессиясы жағдайында мәтінді, визуалды және стилистикалық белгілерді біріктіретін мультимодальды модельдер когнитивті AI жүйелерін дамытудың негізгі бағытына айналады. Атап айтқанда, трансформаторлық архитектуралар көп арналы ақпаратты біріктіруде жаңа көкжиектерді ашады, бұл мәтіннің тікелей мағынасын ғана емес, сонымен қатар эмоционалды субтекстті, визуалды контекстті және жеткізудің стилистикалық ерекшеліктерін ескеруге мүмкіндік береді.

Түйін сөздер: *мультимодальдық, сарказм, сентименталды талдау, контрастты оқыту, трансформаторлар, ViT, BERT, төмен ресурстық тілдер, кросс-модальды аттенция.*

A.T. Akhmediarova, A.T. Ayapbergenova*, Zh.M. Alibiyeva, H.K. Mukazhanov, Zh.N. Issabekov
Satbayev University,
050013 Kazakhstan Republic, Almaty, Satpayev 22
*e-mail: a.ayapbergenova@satbayev.university

UNIVERSAL TRANSFORMER FRAMEWORK FOR TEXT ANALYSIS AND GENERATION

The article proposes a universal multimodal framework for solving three tasks of natural language processing: sarcasm recognition, sentiment-oriented generation of image descriptions, and contextual neural network machine translation using the reranking mechanism. The architecture includes specialized encoders (BERT, RoBERTa, ViT), cross-modal attention mechanisms, and contrastive learning, which provides adaptation to various types of input data and semantic tasks. The experiments demonstrated improvements in the BLEU, METEOR, F1, and NER metrics compared to the basic models. Special attention is paid to the stability of the model when working with rare words and named entities. The results obtained confirm the effectiveness of the proposed approach in conditions of limited data and multimodal complexity. In the context of the rapid growth of multimedia content and user expression on social networks, multimodal models combining text, visual and stylistic features are becoming a key direction in the development of cognitive AI systems. In particular, transformer architectures open up new horizons in the integration of multi-channel information, allowing us to take into account not only the direct meaning of the text, but also the emotional subtext, visual context and stylistic features of the presentation.

Key words: *multimodality, sarcasm, sentiment analysis, contrastive learning, transformers, ViT, BERT, low-resource languages, cross-modal attention.*

Авторлар туралы мәлімет

Айнұр Танатарқызы Ахмедиярова – PhD, Автоматика және ақпараттық технологиялар институтының профессоры, «Киберқауіпсіздік, ақпаратты өңдеу және сақтау» кафедрасы, Satbayev University, Қазақстан Республикасы; e-mail: a.akhmediyarova@satbayev.university.

Әсем Тултанқызы Аяпбергенова* – техника және технология магистрі, Автоматика және ақпараттық технологиялар институтының аға оқытушысы, Satbayev University «Бағдарламалық инженерия» кафедрасы. Қазақстан Республикасы; e-mail: asem_007800@inbox.ru. ORCID: https://orcid.org/0000-0001-6388-9458.

Жибек Мейрамбекқызы Алибиева – PhD, қауымдастырылған профессор, Автоматика және ақпараттық технологиялар институты, «Бағдарламалық инженерия» кафедрасы, Satbayev University, Қазақстан Республикасы; e-mail: zh.alibiyeva@satbayev.university. ORCID: https://orcid.org/0000-0001-9565-5621.

Нуржан Какенович Мукажанов – PhD, қауымдастырылған профессор, Автоматика және ақпараттық технологиялар институты, «Бағдарламалық инженерия» кафедрасы, Satbayev University, Қазақстан Республикасы; e-mail: n.mukazhanov@satbayev.university. ORCID: https://orcid.org/0000-0003-4835-5751.

Жанібек Назарбекұлы Исабеков – PhD, Роботты техника және автоматиканың техникалық құралдары кафедрасының қауымдастырылған профессоры, Satbayev University, Алматы қ., Қазақстан, e-mail: z.issabekov@satbayev.university. ORCID: https://orcid.org/0000-0003-2900-8025.

Сведение об авторах

Айнур Танатаровна Ахмедиярова – PhD, профессор Института автоматики и информационных технологий, кафедра «Кибербезопасности, обработки и хранения информации», Satbayev University, Республика Казахстан; e-mail: a.akhmediyarova@satbayev.university.

Асем Тултановна Аяпбергенова* – магистр техники и технологии, старший преподаватель Института автоматики и информационных технологий, кафедра «Программной инженерии» Satbayev University, Республика Казахстан; e-mail: asem_007800@inbox.ru. ORCID: https://orcid.org/0000-0001-6388-9458.

Жибек Мейрамбековна Алибиева – PhD, ассоциированный профессор, Института автоматики и информационных технологий, кафедра «Программной инженерии», Satbayev University, Республика Казахстан; e-mail: zh.alibiyeva@satbayev.university. ORCID: https://orcid.org/0000-0001-9565-5621.

Нуржан Какенович Мукажанов – PhD., ассоциированный профессор, Института автоматики и информационных технологий, кафедра «Программной инженерии», Satbayev University, Республика Казахстан; e-mail: n.mukazhanov@satbayev.university. ORCID: https://orcid.org/0000-0003-4835-5751.

Жанибек Назарбекулы Исабеков – PhD, ассоциированный профессор кафедры «Робототехники и технических средств автоматики», Satbayev University, г. Алматы, Казахстан; e-mail: z.issabekov@satbayev.university. ORCID: https://orcid.org/0000-0003-2900-8025.

Information about the authors

Ainur Tanatarovna Akhmediarova – PhD, professor at the Institute of Automation and Information Technology, Department of Cybersecurity, Information Processing and Storage, Satbayev University, Republic of Kazakhstan; e-mail: a.akhmediyarova@satbayev.university.

Asem Tultanovna Ayapbergenova* – master of Engineering and Technology, Senior Lecturer at the Institute of Automation and Information Technology, Department of Software Engineering at Satbayev University. Republic of Kazakhstan; e-mail: asem_007800@inbox.ru. ORCID: https://orcid.org/0000-0001-6388-9458.

Zhibek Meirambekovna Alibieva – PhD, associate professor, Institute of Automation and Information Technology, Department of Software Engineering, Satbayev University, Republic of Kazakhstan; e-mail: zh.alibiyeva@satbayev.university. ORCID: https://orcid.org/0000-0001-9565-5621.

Nurzhan Kakenovich Mukazhanov – PhD, Associate Professor, Institute of Automation and Information Technology, Department of Software Engineering, Satbayev University, Republic of Kazakhstan; e-mail: n.mukazhanov@satbayev.university. ORCID: https://orcid.org/0000-0003-4835-5751.

Zhanibek Issabekov – PhD, Associate Professor of the Department of Robotics and Engineering Tools of Automation, Satbayev University, Almaty, Kazakhstan; z.issabekov@satbayev.university. ORCID: https://orcid.org/0000-0003-2900-8025.

Поступила в редакцию 05.01.2026
Принята к публикации 04.02.2026

https://doi.org/10.53360/2788-7995-2026-1(21)-5

МРНТИ: 47.45.29

М.С. Мурушкин[1], К.С. Бактыбеков[2], Б.Р. Жумажанов[2], А.Е. Кулакаева[2,3*]
[1]ТОО «Эйрбас Казахстан»,
010000, Республика Казахстан, г. Астана, пр. Сарыарка, 6
[2]ТОО «Ghalam»,
010000, Республика Казахстан, г. Астана, пр. Туран, 89
[3]АО «Международный университет информационных технологий»,
050040, Республика Казахстан, г. Алматы, ул. Манаса, 34/1
*e-mail: a.kulakayeva@iitu.edu.kz

МОДЕЛИРОВАНИЕ ПАРАМЕТРОВ НАЗЕМНОЙ АНТЕННЫ С ДВУХДИАПАЗОННЫМ S/X ОБЛУЧАТЕЛЕМ

Аннотация: *В данной статье рассматривается модернизация антенны наземной станции с параболическим отражателем диаметром 3,7 м для работы в S- и X-диапазонах. Основное внимание уделяется учету реальной геометрии отражателя, восстановленной на основе высокоточных лазерных измерений. На основе полученных данных был разработан новый двухдиапазонный коаксиальный облучатель, который обеспечивает совмещение фазовых центров в S- и X-диапазонах, а также высокую изоляцию между портами. Оптимизация конструкции и прогнозирование характеристик антенны были выполнены с использованием электродинамического моделирования в среде Ansys HFSS с учетом действительного профиля параболы. Изготовленный прототип прошел лабораторные испытания, которые подтвердили расчетные прогнозы. Ширина луча составила около 0,7°, апертурная эффективность достигла 56 %, а отношение усиления к температуре шума G/T – порядка 25,9 дБ/К, что соответствует требованиям радиолинии для приема сигналов низкоорбитальных спутников. Результаты демонстрируют, что точный учет формы отражателя и корректная настройка облучателя позволяют достичь требуемых характеристик антенны без замены параболического зеркала. Тем самым подтверждается возможность эффективной модернизации существующих наземных антенн среднего класса за счет интеграции методов высокоточной геометрической реконструкции и электродинамического моделирования.*

Ключевые слова: *параболическая антенна, наземная станция, S диапазон, X диапазон, коаксиальный облучатель, низкоорбитальный спутник, дистанционное зондирование.*