УДК 004.89
МРНТИ 28.23.29

https://doi.org/10.55452/1998-6688-2026-23-3-336-346

[1]Марламбеков Д.,
докторант, ORCID ID: 0009-0005-3266-2126,
e-mail: dmarlambekov@gmail.com
[1]Ахметова А.,
докторант, ORCID: ID0000-0002-7718-6478,
e-mail: baltabekova1994@gmail.com
[1]Төрекул С.,
докторант, ORCID ID: 0009-0003-8161-533X,
e-mail: saule.torekul@gmail.com
[1*]Уалиева И. М.,
к.ф.-м.н., ассоциированный профессор, ORCID ID: 0000-0003-3853-8896,
*e-mail: i.ualiyeva@gmail.com

[1]Казахский национальный университет им. аль-Фараби, г. Алматы, Казахстан

БЕНЧМАРКИНГ МОДЕЛЕЙ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ КАЗАХСКИХ НОВОСТНЫХ ТЕКСТОВ В УСЛОВИЯХ ОГРАНИЧЕННОЙ РАЗМЕТКИ

Аннотация

Стремительное развитие методов глубокого обучения и появление трансформерных архитектур кардинально изменили ландшафт обработки естественного языка, однако эти достижения остаются неравномерно распределенными. В то время как для английского и других высокоресурсных языков существуют масштабные бенчмарки, для низкоресурсных языков, таких как казахский, проблема качественной классификации текстов остается острым вызовом. Авторами проведен сравнительный анализ классической модели эмбеддингов Word2Vec, рекуррентных моделей BiLSTM, трансформеров BERT и XLM-R, генеративной модели mT5, а также классических и статистических бейзлайнов для задачи few-shot классификации казахских новостных текстов в условиях ограниченной разметки. Эксперименты проведены на датасете KazNews для few-shot классификации, собранном из статей tengrinews.kz и inform.kz и параллельных корпусов HuffPost. Датасеты содержат пять категорий (спорт, политика, бизнес, путешествия и др.). Эксперименты в режимах 1-shot и 5-shot продемонстрировали, что ни одна из моделей не доминирует одновременно на всех датасетах и режимах: Word2Vec показал лучшие результаты на датасетах на корпусах HuffPost (0.48 и 0.53) при 5-shot режиме; mT5-Seq2Seq лучший на KazNews при 5-shot (0.71); BERT в целом конкурентоспособен при 5-shot на датасете KazNews с длинными текстами. Авторы также выявили резкий скачок качества для модели mT5-Seq2Seq в 5-shot. Для режима 1-shot ни одна из моделей не показала приемлемого результата качества.

Ключевые слова: обработка естественного языка, низкоресурсные языки, классификация текстов, few-shot learning, рекуррентные модели, трансформерные модели, генеративные модели.

Введение

Проблема низкоресурсных языков, к которым относится и казахский язык, состоит в отсутствии огромных объемов данных, к тому же он обладает высокой морфологической вариативностью. Все вышеназванное усложняет обработку казахского языка методами глубокого обучения. Подходы, основанные на методах обучения с малым количеством данных (few-shot learning), могут обойти эти проблемы.

В данном исследовании рассмотрен few-shot learning (FSL) подход для задачи классификации текстов на датасете с ограниченным количеством обучающих примеров. Авторы взяли принципиально разные классы, начиная от простых эмбеддингов до генеративных трансфор-

меров, таких как Word2vec, BiLSTM, BERT, XLM-RoBERTa-base, mT5-Seq2seq, и сравнили с Random baseline и Majority baseline, чтобы показать, что предобученные модели действительно могут качественно классифицировать тексты на казахском языке с малым количеством примеров.

Для проведения исследования авторы собрали оригинальный корпус новостных текстов tengrinews.kz и inform.kz на казахском языке и параллельные датасеты: HuffPost ENG на оригинальном английском языке и его перевод на казахский HuffPost KAZ. Параллельные датасеты на английском и казахском языках включены, чтобы сравнить поведение моделей на ресурсном и низкоресурсном языках в идентичных условиях.

Авторы выявили критическую границу применимости генеративных моделей (эффект «нулевого» качества mT5 в 1-shot и резкий скачок в 5-shot) и вывели рекомендации для разработчиков NLP-систем для казахского языка.

Исследования, направленные на обработку текстов на казахском языке, постоянно растут. Однако их количество несопоставимо низко по сравнению с высокоресурсными языками, каким являются английский, испанский или турецкий [1]. Казахский язык относится к агглютинативным языкам, что говорит об их богатой морфологии, что накладывает отпечаток на методологию работы с ним.

Ранние подходы, которые и сейчас распространены, основывались на классических методах машинного обучения, как наивный Байес, метод опорных векторов SVM, логистическая регрессия [2-5] и пр., которые использовали подходы вроде «мешка слов» (BOW) или TF-IDF [6, 7]. Эти подходы относятся к частотным, и в случае сложной морфологии созданные словари становились слишком разреженными, что снижало обобщающую способность.

Ситуация заметно изменилась с приходом рекуррентных нейросетей, в частности LSTM и BiLSTM [8, 9]. Исследователи нашли, что двунаправленная обработка текста позволяет гораздо точнее улавливать внутреннюю структуру слова и локальные связи в предложении. Это был явный шаг вперед по сравнению со старой статистикой [1]. Тем не менее, серьезная нехватка огромных размеченных корпусов остается большой проблемой. Даже современные эмбеддинги, такие как Word2Vec или FastText [10, 11], не спасают рекуррентные модели от нестабильности, когда количество обучающих примеров слишком мало.

Значительные достижения связаны с появлением трансформеров, архитектур типа BERT, XLM-RoBERTa, mT5, KazSim [12–14]. За счет того, что они предобучены на огромном количестве текстов, эти модели научились обходить дефицит данных. Большое количество исследований также посвящено обучению на малых выборках (few-shot learning), особенно для высокоресурных языков [15, 16]. Для казахского же языка такие методы встречаются в задаче упрощения текстов [17], но для задач классификации текстов такие исследования не найдены.

Материалы и методы

Датасет

Для исследования были собраны три датасета: оригинальный корпус новостных текстов на казахском языке, опубликованных на tengrinews.kz и inform.kz; готовый датасет HuffPost ENG (huff_5cat_en.json) на английском языке и HuffPost KAZ и параллельный датасет, переведенный с HuffPost ENG на казахский язык (huff_5cat_translated_kk.json).

Тематически корпус охватывает пять рубрик. «Спорт» включает репортажи о соревнованиях, турнирах и спортивных достижениях. Рубрика «Политика» представлена материалами о государственных решениях и международных отношениях. «Бизнес» объединяет публикации об экономике, финансовых рынках и банковском секторе. «Путешествия» содержат туристические и страноведческие тексты. Пятая категория зависит от версии датасета: в корпусе HuffPost ей соответствуют разделы Entertainment или Technology. Такой диапазон тематик создает реалистичную классификационную задачу: часть рубрик семантически смежные и модели приходится работать с неоднозначностью. На рисунке 1 визуализация t-SNE показана

высокая разделимость некоторых классов и высокая степень смешивания признаков у остальных категорий новостей.

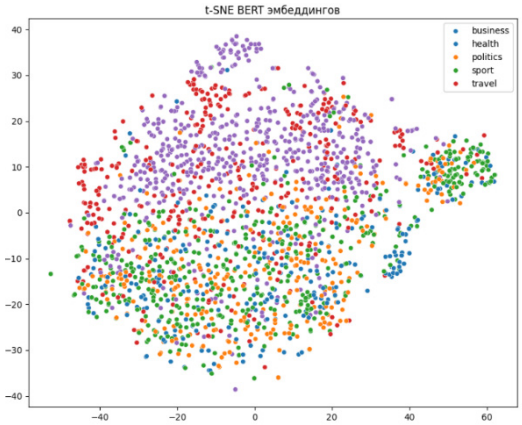

Рисунок 1 – t-SNE проекция BERT-эмбеддингов:
кластерная структура пяти тематических категорий

Количественные характеристики использованных датасетов обобщены в таблице 1.

Таблица 1 – Общий объем датасетов и средняя длина предложения

Датасет	Всего статей	Всего предложений	Средняя длина предложения (слов)	Медиана (слов)
KazNews	7520	87 428	13.6	12
HuffPost KAZ	65 406	102 618	10.7	10
HuffPost EN	65 406	156 481	11.7	10

Приведенные показатели средней длины предложения демонстрируют существенные различия между корпусами: тексты KazNews в среднем на 20–27% длиннее по числу слов в предложении, чем тексты HuffPost, что согласуется с жанровыми особенностями источников: новостные статьи tengrinews.kz и inform.kz написаны в более развернутом стиле, тогда как заголовочно-ориентированный формат HuffPost предполагает более короткие и лаконичные конструкции. Это обстоятельство существенно для интерпретации результатов Word2Vec, представленных в разделе «Результаты и обсуждение»: усреднение векторных представлений по короткому предложению дает менее устойчивый семантический сигнал, чем по длинному, чем отчасти объясняются более высокие показатели Word2Vec именно на датасете KazNews.

Распределение статей по рубрикам для каждого датасета представлено в таблице 2.

Таблица 2 – Количество статей по рубрикам

Датасет	Рубрика	Количество статей	Доля от датасета
KazNews	Путешествия	2266	30.1%
KazNews	Политика	1706	22.7%
KazNews	Спорт	1335	17.8%
KazNews	Здоровье	1149	15.3%
KazNews	Бизнес	1064	14.1%
HuffPost KAZ / EN	Politics	37 449	57.3%
HuffPost KAZ / EN	Travel	9936	15.2%
HuffPost KAZ / EN	Health	6807	10.4%
HuffPost KAZ / EN	Business	6059	9.3%
HuffPost KAZ / EN	Sports	5155	7.9%

Как следует из таблицы 2, датасет KazNews характеризуется относительно равномерным распределением классов: доля крупнейшей рубрики («Путешествия») не превышает 30%, а доля наименьшей («Бизнес») составляет 14%, то есть отношение крупнейшего класса к наименьшему не превышает 2.1. Датасеты HuffPost, напротив, демонстрируют выраженный дисбаланс: рубрика «Politics» охватывает 57.3% всех статей, тогда как «Sports» лишь 7.9%, что дает отношение более 7 к 1. Именно эта диспропорция определяет эффект, описанный в разделе «Результаты и обсуждение» как вырождение XLM-RoBERTa-base в константный предиктор мажоритарного класса: при столь значительном дисбалансе классификатору, не успевшему обучиться по 1–5 примерам на класс, статистически выгоднее «угадывать» доминирующую рубрику, чем распределять предсказания равномерно между всеми пятью категориями.

Для иллюстрации тематического разнообразия корпуса ниже приведены примеры заголовков статей датасета KazNews по каждой из пяти рубрик (таблица 3).

Таблица 3 – Примеры заголовков статей KazNews по рубрикам

Рубрика	Пример заголовка
Спорт	Спорт саласында бірнеше шенеунік тәртіптік жауапкершілікке тартылды
Путешествия	Қазақстаннан бірнеше елге жаңа тікелей рейстер ашылады
Здоровье	Диабет әлеуметтік маңызды аурулардың тізімінен шығарылды: енді не өзгереді?
Политика	Қасым-Жомарт Тоқаевқа Өзбекстанның ең жоғары мемлекеттік наградасы табысталды
Бизнес	Ел экономикасына құйылған инвестиция көлемі 5 жылда 18 трлн теңгеге жетті

Архитектуры моделей

Random baseline предсказывает метку случайным образом из пяти классов и задает абсолютный минимум производительности. Majority baseline всегда выбирает наиболее часто встречающийся класс тренировочного набора, что позволяет проверить, реагируют ли модели на реальную структуру данных или только эксплуатируют дисбаланс классов.

Word2Vec + нейронный классификатор [11]. Статические векторные представления Word2Vec, обученные на казахских текстах, усреднялись до уровня предложения и подавались в простой полносвязный классификатор. Эта конфигурация служит мостом между классическими методами и контекстуальными моделями и, как выяснилось, ведет себя неожиданно.

Рекуррентный блок представлен архитектурой BiLSTM с глубиной в 1,2 и 3 слоя [9].

Мультиязычная версия BERT mBERT [12] стала главным трансформерным бейзлайном. Архитектура включает 12 слоев self-attention с 12 головами внимания и скрытое пространство размером 768. При дообучении параметры энкодера обновлялись с пониженным learning rate, чтобы не «стереть» предобученные представления, тогда как классификационный слой тренировался с нуля.

XLM-RoBERTa-base [13] – еще один представитель мультиязычных трансформеров, модель, прошедшая предобучение на 2,5 ТБ разнородных текстов. В основе архитектуры лежат принципы RoBERTa: динамическое маскирование и расширенная субсловная токенизация SentencePiece. XLM-R была включена в эксперимент для того, чтобы оценить, насколько расширенные ресурсы предобучения помогают при работе с морфологически богатыми тюркскими языками.

mT5-Seq2Seq [14]. Генеративная модель mT5 реализует нестандартный подход: задача классификации формулируется как задача порождения текста. На входе: текст статьи, на выходе: слово с названием категории. Такой text-to-label формат любопытен тем, что полностью меняет способ, которым модель «думает» о классах, и позволяет выяснить, насколько генеративные архитектуры чувствительны к объему обучающей выборки.

Параметры обучения и программно-аппаратная среда

С целью повышения воспроизводимости результатов приводим детали экспериментальной настройки, восстановленные из исходного кода эксперимента. Bi-LSTM (1/2/3 слоя): обучаемые (не предобученные) эмбеддинги размерности 100, скрытый слой 64 единицы на направление, словарь строится по объединению обучающей и тестовой выборок с токенизацией по пробелам; обучение – 15 эпох, batch size = 16, оптимизатор Adam с learning rate = 0.001, Dropout = 0.3, критерий остановки – фиксированное число эпох (без early stopping). BERT: чекпоинт bert-base-multilingual-cased (Hugging Face Transformers), токенизация WordPiece, максимальная длина последовательности 128 токенов; обучение – 10 эпох, batch size = 8 (для теста – 16), оптимизатор AdamW с learning rate = 2e-5, линейный warmup-шедулер (get_linear_schedule_with_warmup, num_warmup_steps = 0), gradient clipping (max_norm = 1.0). mT5-Seq2Seq: чекпоинт google/mt5-small, токенизация SentencePiece, входной промпт вида «classify: {текст}», максимальная длина входа 128 токенов, длина целевой последовательности 16 токенов; обучение – 50 эпох, batch size = 2 (для теста – 8), оптимизатор AdamW с learning rate = 5e-4, линейный warmup-шедулер (num_warmup_steps = 0); генерация ответа – жадный поиск (num_beams = 1, max_length = 16). Во всех экспериментах фиксировался seed = 42. Библиотеки: PyTorch, Hugging Face Transformers, datasets, accelerate, sentencepiece, protobuf, scikit-learn устанавливались через pip без явной фиксации номеров версий в коде эксперимента.

Исходные эксперименты (accuracy_results.json / seq2seq_results.json) выполнялись в среде Google Colab с GPU-ускорением. Независимая переоценка метрик macro-F1/weighted-F1/precision/recall для BERT, XLM-RoBERTa-base и mT5-Seq2Seq (таблицы 4–6) выполнена локально на NVIDIA GeForce RTX 5080 Laptop GPU (CUDA 13.1, PyTorch с поддержкой sm_120). Для XLM-RoBERTa-base оригинальный код обучения и гиперпараметры в доступных материалах найти не удалось; при переоценке использованы гиперпараметры по аналогии с BERT (10 эпох, batch size = 8, learning rate = 2e-5, максимальная длина последовательности 128 токенов), и это приближение оригинальной настройки.

Результаты и обсуждение

Для оценки моделей few-shot классификации в условиях выраженного дисбаланса классов и предельно малой обучающей выборки (1 и 5 примеров на класс) одной метрики accuracy недостаточно: она может маскировать вырождение классификатора в константный предиктор мажоритарного класса. Поэтому все результаты приводятся сразу по пяти метрикам: Accuracy, Macro-F1, Weighted-F1, Precision (macro) и Recall (macro) (таблицы 4–6 для трех датасетов (KazNews, HuffPost KAZ, HuffPost EN), обоих режимов обучения (1-shot и 5-shot) и девяти моделей: двух формальных бейзлайнов (Random, Majority), классического подхода на статических эмбеддингах (Word2Vec + полносвязный классификатор), рекуррентных сетей (BiLSTM, 1–3 слоя) и трех предобученных языковых моделей (BERT, XLM-RoBERTa-base, mT5-Seq2Seq). Тестовая выборка в каждом случае – это вся оставшаяся после отбора few-shot примеров часть датасета без искусственного балансирования классов, поэтому распределение классов в тесте отражает естественный дисбаланс исходных данных.

Таблица 4 – Метрики качества few-shot классификации (Accuracy, macro-F1, weighted-F1, Precision, Recall) на датасете KazNews

	Модель	Режим	Accuracy	Macro-F1	Weighted-F1	Precision	Recall
Baseline	Random baseline	1-shot	0.196	0.192	0.201	0.196	0.195
		5-shot	0.197	0.192	0.201	0.196	0.195
	Majority baseline	1-shot	0.141	0.050	0.035	0.028	0.200
		5-shot	0.141	0.050	0.035	0.028	0.200
Эмбеддинги	Word2Vec	1-shot	0.491	0.421	0.440	0.488	0.480
		5-shot	0.643	0.625	0.657	0.638	0.623

Продолжение таблицы 4

Рекуррентные модели	1-layer BiLSTM	1-shot	0.333	0.324	0.330	0.359	0.342
		5-shot	0.487	0.450	0.493	0.463	0.461
	2-layer BiLSTM	1-shot	0.321	0.315	0.318	0.337	0.338
		5-shot	0.464	0.413	0.446	0.450	0.442
	3-layer BiLSTM	1-shot	0.308	0.292	0.295	0.346	0.331
		5-shot	0.426	0.330	0.370	0.415	0.385
Трансформеры	BERT	1-shot	0.241	0.163	0.174	0.152	0.214
		5-shot	0.567	0.477	0.528	0.607	0.513
	XLM-RoBERTa-base	1-shot	0.227	0.074	0.084	0.045	0.200
		5-shot	0.485	0.378	0.411	0.543	0.498
Генеративные модели	mT5-Seq2Seq	1-shot	0.000	0.000	0.000	0.000	0.000
		5-shot	0.710	0.680	0.682	0.738	0.724

Таблица 5 – Метрики качества few-shot классификации (Accuracy, macro-F1, weighted-F1, Precision, Recall) на датасете HuffPost KAZ

	Модель	Режим	Accuracy	Macro-F1	Weighted-F1	Precision	Recall
Baseline	Random baseline	1-shot	0.198	0.168	0.228	0.199	0.199
		5-shot	0.198	0.168	0.228	0.199	0.199
	Majority baseline	1-shot	0.093	0.034	0.016	0.019	0.200
		5-shot	0.093	0.034	0.016	0.019	0.200
Эмбеддинги	Word2Vec	1-shot	0.164	0.144	0.154	0.268	0.236
		5-shot	0.484	0.353	0.518	0.357	0.374
Рекуррентные модели	1-layer BiLSTM	1-shot	0.165	0.145	0.186	0.194	0.195
		5-shot	0.179	0.160	0.195	0.204	0.212
	2-layer BiLSTM	1-shot	0.337	0.191	0.353	0.201	0.202
		5-shot	0.282	0.197	0.312	0.213	0.215
	3-layer BiLSTM	1-shot	0.233	0.169	0.260	0.204	0.211
		5-shot	0.179	0.146	0.171	0.226	0.227
Трансформеры	BERT	1-shot	0.183	0.111	0.121	0.233	0.214
		5-shot	0.420	0.243	0.406	0.328	0.310
	XLM-RoBERTa-base	1-shot	0.571	0.149	0.418	0.242	0.201
		5-shot	0.415	0.224	0.380	0.328	0.273
Генеративные модели	mT5-Seq2Seq	1-shot	0.000	0.000	0.000	0.000	0.000
		5-shot	0.339	0.044	0.299	0.055	0.064

Таблица 6 – Метрики качества few-shot классификации (Accuracy, macro-F1, weighted-F1, Precision, Recall) на датасете HuffPost EN

	Модель	Режим	Accuracy	Macro-F1	Weighted-F1	Precision	Recall
Baseline	Random baseline	1-shot	0.198	0.168	0.228	0.199	0.199
		5-shot	0.198	0.168	0.228	0.199	0.199
	Majority baseline	1-shot	0.093	0.034	0.016	0.019	0.200
		5-shot	0.093	0.034	0.016	0.019	0.200
Эмбеддинги	Word2Vec	1-shot	0.352	0.274	0.377	0.333	0.342
		5-shot	0.533	0.442	0.563	0.443	0.493
Рекуррентные модели	1-layer BiLSTM	1-shot	0.222	0.174	0.257	0.201	0.196
		5-shot	0.251	0.190	0.281	0.212	0.216
	2-layer BiLSTM	1-shot	0.207	0.165	0.237	0.209	0.209
		5-shot	0.236	0.191	0.261	0.218	0.228
	3-layer BiLSTM	1-shot	0.305	0.203	0.336	0.215	0.214
		5-shot	0.294	0.212	0.324	0.233	0.236

Продолжение таблицы 6

Трансформеры	BERT	1-shot	0.197	0.162	0.132	0.264	0.287
		5-shot	0.269	0.251	0.269	0.336	0.359
	XLM-RoBERTa-base	1-shot	0.552	0.193	0.425	0.173	0.232
		5-shot	0.115	0.098	0.048	0.296	0.259
Генеративные модели	mT5-Seq2Seq	1-shot	0.000	0.000	0.000	0.048	0.000
		5-shot	0.251	0.230	0.248	0.356	0.242

Для всех датасетов случайный классификатор Random baseline показывает схожие значения accuracy в 0.196-0.198, что указывает на случайное угадывание классов. Мажоритарный классификатор Majority baseline показывает более низкие значения accuracy, 0.093 для датасетов HuffPost EN и HuffPost KAZ и 0.141 для KazNews, для всех 1-shot 5-shot режимов. Мажоритарный классификатор всегда выбирает наиболее частый класс, что повлияло на значения Macro-F1, равные 0.034 для датасетов HuffPost EN и HuffPost KAZ и 0.05 для KazNews, и это указывает на несбалансированность классов в выборке.

На датасете KazNews модель Word2Vec показала наиболее высокие результаты, достигнув accuracy 0.491 и Macro-F1 0.421 для режима 1-shot. В режиме 5-shot Word2Vec достигла хороших результатов, accuracy 0.643 и Macro-F1 0.625, уступив только генеративной модели mT5.

Рекуррентные сети (BiLSTM, 1–3 слоя) показывают результаты, близкие к бейзлайнам, на всех датасетах: accuracy в диапазоне 0.16–0.49, macro-F1 0.15–0.45. Увеличение числа слоев не дает систематического выигрыша: например, на HuffPost KAZ 2-слойная конфигурация неожиданно обгоняет 1- и 3-слойную при 1-shot (accuracy 0.337 против 0.165 и 0.233), но проигрывает им же при 5-shot (0.282). Такая немонотонность по числу слоев является ожидаемым следствием крайне малого объема обучающих данных: при 1–5 примерах на класс сеть не столько обучается обобщающим признакам, сколько подстраивается под конкретную случайную выборку, и результат сильно зависит от конкретных примеров, а не от архитектуры.

BERT (bert-base-multilingual-cased) при 1-shot показывает слабые и низкие по всем метрикам результаты на всех трех датасетах (accuracy 0.18–0.24, macro-F1 0.11–0.17), что ожидаемо для одного примера на класс. При переходе к 5-shot качество заметно растет, особенно на KazNews (accuracy 0.567, macro-F1 0.477), и это третий результат на этом датасете при 5-shot, после моделей mT5 и Word2Vec). На HuffPost KAZ и HuffPost EN рост при 5-shot более скромный (accuracy 0.420 и 0.269 соответственно). Важно, что во всех случаях accuracy и macro-F1 у BERT остаются в пределах одного порядка друг друга, модель не демонстрирует признаков вырождения в константный предиктор.

XLM-RoBERTa-base демонстрирует самый показательный случай расхождения метрик во всей таблице. На KazNews модель ведет себя предсказуемо: низкие accuracy и macro-F1 при 1-shot (0.227 / 0.074) и заметный рост при 5-shot (0.485 / 0.378), без аномальных расхождений между метриками. Но на обоих датасетах HuffPost при 1-shot возникает резкий разрыв между accuracy и macro-F1: 0.571 против 0.149 на HuffPost KAZ и 0.552 против 0.193 на HuffPost EN. Мы проверили это напрямую, построив confusion matrix по сохраненным предсказаниям модели на HuffPost KAZ (1-shot) (таблица 7):

Таблица 7 – Confusion Matrix для датасета HuffPost KAZ

Истинный класс	Business	Health	Politics	Sports	Travel
Business	0	0	6025	33	0
Health	0	1	6720	85	0
Politics	0	1	37265	182	0
Sports	0	0	5102	52	0
Travel	0	0	9912	23	0

в 65 024 из 65 401 случая (99.4%) модель предсказывала класс «Politics», то есть практически всегда один и тот же ответ, независимо от текста. Recall по классу Politics при этом составил 0.995, а по остальным четырем классам – от 0.000 до 0.010 (таблицы 8, 9). Поскольку доля класса Politics в тестовой выборке – около 57%, такой константный предиктор и дает accuracy ≈ 0.57 при формально нулевой полезности для четырех из пяти категорий, что и отражает низкий macro-F1 (0.149) и теоретическое значение macro-recall ровно 1/5 = 0.20 (при факте 0.201). Это ровно тот случай, когда одна лишь accuracy вводит в заблуждение: на датасете с сильным дисбалансом классов (HuffPost, где Politics ≈ 57%) XLM-RoBERTa при 1-shot фактически не научилась ничему, кроме «чаще всего это Politics», тогда как на более сбалансированном KazNews (максимальная доля класса около 30%) аналогичного вырождения не происходит. При 5-shot на HuffPost KAZ разрыв между метриками сокращается (0.415 / 0.224), а вот на HuffPost EN 5-shot неожиданно дает худший результат, чем 1-shot (accuracy 0.115, macro-F1 0.098), здесь случайность конкретной обучающей подвыборки, по-видимому, играет более значимую роль, чем само число примеров.

Таблица 8 – True Positives по каждому классу

Класс	TP	support	recall
Business	0	6058	0.000
Health	1	6806	0.000
Politics	37265	37448	0.995
Sports	52	5154	0.010
Travel	0	9935	0.000

Таблица 9 – Распределение предсказанных классов

Класс	Количество предсказаний	Доля (%)
Business	0	0.0
Health	2	0.0
Politics	65024	99.4
Sports	375	0.6
Travel	0	0.0

mT5-Seq2Seq демонстрирует наиболее контрастное поведение между режимами обучения. При 1-shot метрики равны нулю на всех трех датасетах: генеративная модель не в состоянии сформировать устойчивое отображение «текст – метка» по единственному примеру и не воспроизводит ни одной корректной метки. При 5-shot картина резко меняется: на KazNews mT5 показывает лучший результат среди всех протестированных моделей и конфигураций (accuracy 0.710, macro-F1 0.680, precision 0.738, recall 0.724), причем accuracy и macro-F1 здесь близки друг к другу, что говорит о содержательном, а не вырожденном предсказании. На HuffPost EN рост при 5-shot тоже сбалансирован по метрикам (0.251 / 0.230). А вот на HuffPost KAZ 5-shot вновь проявляется знакомый паттерн расхождения: accuracy 0.339 при macro-F1 всего 0.044, судя по всему, еще один случай смещения предсказаний в сторону одного-двух частых классов, хотя и не такой полный, как у XLM-RoBERTa (иначе macro-F1 был бы еще ниже).

Ограничения

Отметим методологические ограничения приведенных результатов. Во-первых, каждое значение в таблицах 4–6 получено за один запуск с фиксированным seed = 42, а не усреднением по нескольким случайным обучающим подвыборкам; при таком объеме обучающих данных (1–5 примеров на класс) результат может существенно меняться в зависимости от кон-

кретно отобранных примеров, что и наблюдается местами в виде немонотонной зависимости качества от числа shot'ов или числа слоев. Во-вторых, для Word2Vec использована собственная реализация (усредненный вектор + полносвязный классификатор по описанию в тексте статьи), а не оригинальный код, отдельного скрипта для Word2Vec в рабочих материалах не обнаружено. В-третьих, для XLM-RoBERTa-base использованы гиперпараметры по аналогии с BERT (10 эпох, batch size = 8, learning rate = 2e-5), поскольку оригинальный код обучения этой модели не найден, и это приближение, а не воспроизведение исходной настройки. К тому же близкие результаты (например, между 1-layer BiLSTM () и XLM-RoBERTa-base () в режиме 5-shot) находятся в пределах погрешности и требуют дальнейшей валидации.

Заключение

В целом результаты в данном эксперименте показывают следующие связанные выводы. Во-первых, ни одна модель не доминирует одновременно на всех датасетах и режимах: mT5-Seq2Seq лучший на KazNews при 5-shot, BERT в целом конкурентоспособен при 5-shot на датасете с длинными текстами, Word2Vec неожиданно силен на длинных текстах уже при 1-shot и при 5-shot на всех датасетах. Во-вторых, глубина архитектуры (число слоев BiLSTM, переход от BERT к более крупным моделям) не гарантирует улучшения в условиях экстремального few-shot: результат определяется в первую очередь конкретной случайной обучающей подвыборкой. В-третьих, именно случаи наибольшего расхождения между Accuracy и macro-F1 (XLM-RoBERTa на HuffPost при 1-shot, mT5 на HuffPost KAZ при 5-shot) являются самыми показательными: здесь accuracy в одиночку создала бы ложное впечатление приемлемого качества модели, тогда как совокупность метрик в таблицах 4–6 обнажает фактическое вырождение классификатора в константный предиктор мажоритарного класса. Также стоит отметить схожие результаты точности моделей на параллельных датасетах HuffPost.

ЛИТЕРАТУРА

1 Pakray, P., Gelbukh, A., & Bandyopadhyay, S. (2025). Natural language processing applications for low-resource languages. Natural Language Processing, 31(2), 183–197. https://doi.org/10.1017/nlp.2024.33

2 Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150. https://doi.org/10.3390/info10040150

3 McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. AAAI-98 Workshop on Learning for Text Categorization. https://aaai.org/papers/041-ws98-05-007/

4 Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), Machine Learning: ECML-98 (pp. 137–142). Springer. https://doi.org/10.1007/BFb0026683

5 Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–242.

6 Salton, G., Wong, A., & Yang, C.S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613–620. https://doi.org/10.1145/361219.361220

7 Spärck Jones, K. (2004). A statistical interpretation of term specificity in retrieval. Journal of Documentation, 60(5), 493–502. https://doi.org/10.1108/00220410410560573

8 Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

9 Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

10 Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv. https://doi.org/10.48550/arXiv.1607.01759

11 Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv. https://doi.org/10.48550/arXiv.1301.3781

12 Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

13 Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Zettlemoyer, L. (2019). Unsupervised cross-lingual representation learning at scale. arXiv. https://arxiv.org/abs/1911.02116

14 Xue, L., Constant, N., Roberts, A., Kale, K., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 483–498). https://doi.org/10.18653/v1/2021.naacl-main.41

15 Latief, A. D., Jarin, A., Yuyun, Hidayati, N. N., Afra, D. I. N., & Riza, H. (2025). A systematic review of few-shot and zero-shot learning for NLP in low-resource languages: Insights and challenges. IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/11325582

16 Latief, A. D., Jarin, A., Yuyun, Hidayati, N. N., Afra, D. I. N., & Riza, H. (2025). A systematic review of few-shot and zero-shot learning for NLP in low-resource languages: Insights and challenges. In 2025 International Conference on Computer, Control, Informatics and its Applications (IC3INA) (pp. 358–363). IEEE. https://doi.org/10.1109/IC3INA68387.2025.11325582

17 Toleu, A., Tolegen, G., & Ualiyeva, I. (2025). Fine-Tuning Large Language Models for Kazakh Text Simplification. Applied Sciences, 15(15), 8344. https://doi.org/10.3390/app15158344

[1]Марламбеков Д.,
докторант, ORCID ID: 0009-0005-3266-2126,
e-mail: dmarlambekov@gmail.com
[1]Ахметова Ә.,
докторант, ORCID ID: 0000-0002-7718-6478,
e-mail: baltabekova1994@gmail.com
[1]Төрекул С.,
докторант, ORCID ID: 0009-0003-8161-533X,
e-mail: saule.torekul@gmail.com
[1]*Уалиева И.М.,
ф.-м.-ғ.к, , қауымдастырылған профессор, ORCID ID: 0000-0003-3853-8896,
*e-mail: i.ualiyeva@gmail.com

[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан

ШЕКТЕУЛІ ТАҢБАЛАУ ЖАҒДАЙЫНДА ҚАЗАҚ ЖАҢАЛЫҚ МӘТІНДЕРІН ЖІКТЕУГЕ АРНАЛҒАН ТЕРЕҢ ОҚЫТУ МОДЕЛЬДЕРІНІҢ САЛЫСТЫРМАЛЫ ТАЛДАУЫ

Аңдатпа

Терең оқыту әдістерінің қарқынды дамуы мен трансформерлік архитектуралардың пайда болуы табиғи тілді өңдеу саласын түбегейлі өзгертті, дегенмен бұл жетістіктер бірқалыпты таралмаған. Ағылшын және басқа да ресурсы мол тілдер үшін ауқымды бенчмарктер бар болғанымен, қазақ тілі сияқты ресурсы аз тілдер үшін мәтіндерді сапалы классификациялау мәселесі әлі де өзекті мәселе болып отыр. Авторлар шектеулі белгілеу жағдайында қазақша жаңалықтар мәтіндерін few-shot классификациялау тапсырмасы үшін классикалық Word2Vec эмбеддингіне, BiLSTM рекурренттік моделдеріне, BERT және XLM-R трансформерлеріне, генеративті mT5 моделіне, сондай-ақ классикалық және статистикалық бейзлайндарға салыстырмалы анализ жүргізді. Эксперименттер tengrinews.kz, inform.kz мақалаларынан және HuffPost параллель корпустарынан жиналған, few-shot классификациялауға арналған KazNews датасетінде орындалды. Датасеттер бес категорияны қамтиды (спорт, саясат, бизнес, саяхат т.б.).1-shot және 5-shot режимдеріндегі эксперименттер бірде-бір модельдің барлық датасеттер мен режимдерде бірдей үстемдік етпейтінін көрсетті: Word2Vec 5-shot режимінде HuffPost корпусындағы датасеттерде ең жоғары нәтиже көрсетті (0.48 және 0.53); mT5-Seq2Seq 5-shot режимінде KazNews датасетінде үздік нәтижеге қол жеткізді (0.71); ал BERT ұзын

мәтіндерден тұратын KazNews датасетінде 5-shot режимінде жалпы бәсекеге қабілетті болды. Авторлар сонымен қатар 5-shot режимінде mT5-Seq2Seq моделі үшін сапаның кенет артуын анықтады. 1-shot режимі үшін моделдердің ешқайсысы қанағаттанарлық сапа нәтижесін көрсете алмады.

Түйін сөздер: few-shot learning, мәтіндерді жіктеу, қазақ тілі, ресурсы шектеулі тілдер, трансформер модельдері, BERT, XLM-R, BiLSTM, Word2Vec, терең оқыту, табиғи тілдерді өңдеу.

[1]Marlambekov D.,
PhD-student, ORCID ID: 0009-0005-3266-2126,
e-mail: dmarlambekov@gmail.com
[1]Akhmetova A.,
PhD-student, ORCID ID: 0000-0002-7718-6478,
e-mail:baltabekova1994@gmail.com
[1]Torekul S.,
PhD-student, ORCID ID: 0009-0003-8161-533X,
e-mail: saule.torekul@gmail.com
[1]*Ualiyeva I.M.,
Cand. Phys.-Math. Sc., Associate Professor, ORCID ID: 0000-0003-3853-8896,
*e-mail: i.ualiyeva@gmail.com

[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan

BENCHMARKING DEEP LEARNING MODELS FOR FEW-SHOT CLASSIFICATION OF KAZAKH NEWS TEXTS UNDER LIMITED ANNOTATION

Abstract

Deep learning methods and transformer architectures have fundamentally reshaped the field of natural language processing; however, these advancements remain unevenly distributed. While high-resource languages like English benefit from large-scale benchmarks, robust text classification for low-resource languages, such as Kazakh, continues to pose a significant challenge. This paper presents a comparative analysis of the classical Word2Vec embedding, BiLSTM recurrent models, BERT and XLM-R transformers, the generative mT5 model, alongside classical and statistical baselines for few-shot Kazakh news text classification under limited annotation constraints. Experiments were conducted on the KazNews dataset, curated for few-shot classification from tengrinews.kz and inform.kz articles, as well as parallel HuffPost corpora. The datasets comprise five categories (sports, politics, business, travel, etc.). The experimental findings in 1-shot and 5-shot settings demonstrate that no single model consistently dominates across all datasets and regimes: Word2Vec achieved the best performance on the HuffPost corpora in the 5-shot regime (0.48 and 0.53); mT5-Seq2Seq outperformed others on KazNews under 5-shot (0.71); while BERT remained competitive on the long-text KazNews dataset in the 5-shot setup. The authors also observed a sharp performance gain for the mT5-Seq2Seq model in the 5-shot setup. Conversely, none of the evaluated models yielded acceptable quality metrics in the 1-shot regime.

Keywords: few-shot learning, text classification, Kazakh language, low-resource languages, transformer models, BERT, XLM-R, BiLSTM, Word2Vec, deep learning, natural language processing.

Received July 15, 2026; revised August 7, September 3, 2026; accepted September 8, 2026.