АҚПАРАТТЫҚ-КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР
ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ
INFORMATION AND COMMUNICATION TECHNOLOGIES

АҚПАРАТТЫҚ, ЖҮЙЕЛЕР
ИНФОРМАЦИОННЫЕ СИСТЕМЫ
INFORMATION SYSTEMS

DOI 10.51885/1561-4212_2025_2_182
МРНТИ 28.23.37

Д.О. Оралбекова[1], О.Ж. Мамырбаев[1], А.Б. Имансакипова[2], А.Ж. Жигер[3], К.Ж. Мухсина[1]
[1]Институт информационных и вычислительных технологий, г. Алматы, Казахстан
*E-mail: dinaoral@mail.ru**
E-mail: morkenj@mail.ru
E-mail: email123@gmail.com
[2]Алматинский технологический университет, г. Алматы, Казахстан
E-mail: aimansakipova@bk.ru
[3]Университет Нархоз, г. Алматы, Казахстан
E-mail: alia_94-22@mail.ru

СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ МОДЕЛЕЙ LSTM И BERT ДЛЯ ЗАДАЧ МУЛЬТИКЛАССИФИКАЦИИ С ИСПОЛЬЗОВАНИЕМ НАБОРА ДАННЫХ NER

NER ДЕРЕКТЕР ЖИНАСЫН ПАЙДАЛАНУ НЕГІЗІНДЕ КӨПӨЛШЕМДІ ЖІКТЕЛУ МӘСЕЛЕЛЕРІНІҢ LSTM ЖӘНЕ BERT МОДЕЛІН САЛЫСТЫРМАЛЫ ЗЕРТТЕУІ

A COMPARATIVE STUDY OF LSTM AND BERT MODELS FOR MULTI-CLASSIFICATION TASKS USING NER DATASET

Аннотация. *В статье проведён сравнительный анализ моделей LSTM и BERT, применяемых к задачам мультиклассификации на казахском языке с использованием набора данных для распознавания именованных сущностей. Основной акцент исследования сделан на преодоление проблемы ограниченности ресурсов для обработки текстов на казахском языке посредством адаптации существующих методов машинного обучения для анализа многомерных классификационных задач. Оба подхода продемонстрировали свою эффективность в различных аспектах обработки текстовых данных, включая моделирование контекстных зависимостей и точную классификацию по множеству категорий. Модель LSTM показала высокую способность к учёту временных зависимостей в тексте, что делает её пригодной для решения задач классификации в условиях ограниченных языковых ресурсов. В то же время модель BERT, основанная на архитектуре Transformer, продемонстрировала конкурентоспособные результаты в области контекстного анализа и обработки сложных текстовых структур, что обеспечивает её более высокую производительность при мультиклассификации текстов на казахском языке. Результаты экспериментов свидетельствуют о том, что обе модели могут эффективно применяться для задач классификации текстов на казахском языке, однако модель BERT продемонстрировала более стабильные и надёжные результаты, обусловленные её способностью к более глубокому контекстуальному пониманию. Полученные данные подчёркивают значимость использования современных методов обработки естественного языка для языков с ограниченными ресурсами и открывают перспективы для их дальнейшего исследования и практического применения.*

Ключевые слова*: NLP, мультиклассификация, NER, LSTM, BERT, малоресурсный язык.*

Аңдатпа. *Мақалада аталған нысанды тану үшін деректер жиынтығын пайдалана отырып,*

қазақ тіліндегі көп классификациялық есептерге қолданылатын LSTM және BERT модельдерінің салыстырмалы талдауы берілген. Зерттеу жұмысының негізгі бағыты көпөлшемді жіктеу мәселелерін талдау үшін қолданыстағы машиналық оқыту әдістерін бейімдеу арқылы қазақ тіліндегі мәтіндерді өңдеу ресурстарының шектеулі мәселесін шешуге бағытталған. Екі модель де мәтіндік деректерді өңдеудің әртүрлі аспектілерінде, соның ішінде контекстік тәуелділіктерді модельдеуді және бірнеше санаттарға дәл жіктеуді қоса, олардың тиімділігін көрсетті. LSTM моделі мәтіндегі уақытша тәуелділіктерді есепке алудың жоғары мүмкіндігін көрсетті, бұл оны шектеулі тілдік ресурстар жағдайында жіктеу мәселелерін шешуге қолайлы етеді. Сонымен бірге, Transformer архитектурасына негізделген BERT моделі контекстік талдау және күрделі мәтін құрылымдарын өңдеу саласында өте жақсы нәтиже көрсетті, бұл қазақ тіліндегі мәтіндерді көпөлшемді жіктеуде оның жоғары өнімділігін қамтамасыз етеді. Эксперименттік нәтижелер екі модельді де қазақ тіліндегі мәтінді жіктеу тапсырмаларында тиімді пайдалануға болатынын көрсетеді, бірақ BERT моделі тереңірек контекстік түсінуді қамтамасыз ету қабілетінің арқасында тұрақты және сенімді нәтижелер көрсетті. Қорытындылар ресурстары шектеулі тілдер үшін табиғи тілді өңдеудің заманауи әдістерін қолданудың маңыздылығын көрсетеді және оларды әрі қарай зерттеу мен практикалық қолдану перспективаларын ашады.

Түйін сөздер: *NLP, көпөлшемді жіктелу, NER, LSTM, BERT, ресурсы аз тіл.*

Abstract. *The article presents a comparative analysis of LSTM and BERT models applied to multi-classification tasks in the Kazakh language using a named entity recognition dataset. The study primarily focuses on addressing the issue of limited resources for processing Kazakh text by adapting existing machine learning methods for the analysis of multidimensional classification tasks. Both approaches have demonstrated their effectiveness in various aspects of text data processing, including modeling contextual dependencies and accurately classifying multiple categories. The LSTM model exhibited a high capability for capturing temporal dependencies in text, making it suitable for classification tasks in low-resource language settings. Meanwhile, the BERT model, based on the Transformer architecture, showed superior results in contextual analysis and processing of complex text structures, ensuring higher performance in multi-classification of Kazakh text. The experimental results indicate that both models can be effectively employed for text classification tasks in the Kazakh language. However, the BERT model demonstrated more stable and reliable outcomes, attributed to its ability for deeper contextual understanding. The findings underscore the importance of applying modern natural language processing (NLP) methods to low-resource languages and open new avenues for further research and practical application.*

Keywords*: NLP, multiclassification, NER, LSTM, BERT, low-resource language.*

Введение. В сфере обработки естественного языка (NLP) (Oralbekova et al., 2023) классификация текста становится фундаментальной задачей с широким спектром применений: от анализа настроений до категоризации документов. Классификация текстов представляет собой процесс распределения текстов по конкретным категориям или классам, основываясь на их содержании и характеристиках (Dogra et al., 2022). Эта задача является ключевой в лингвистике и обработке естественного языка (NLP). Она имеет широкое применение в различных сферах, включая анализ настроений в текстах, определение их тематики, фильтрацию спама и другое. Целью классификации текстов является систематизация и структурирование больших массивов текстовой информации, что позволяет упростить их анализ и поиск. В целом классификация текстов способствует организации, анализу и извлечению информации из текстового контента, делая его более доступным и полезным для различных задач и целей. Например, тексты можно разделить по жанру, стилю, тематике или авторству, чтобы исследовать их различия и общие черты, а также выявить определенные закономерности.

Задача классификации становится особенно сложной для языков со скудными вычислительными ресурсами, таких как казахский, где нехватка доступных наборов данных для задач классификации создает серьезные препятствия. Несмотря на существование многочисленных исследований, подробно описывающих методологии предварительной обработки и подготовки данных для казахского языка (Mamyrbayev, Oralbekova, 2020), а также относительную простоту поиска необработанных данных, по-прежнему не хватает общедоступных наборов данных, подходящих для различных

практических приложений классификации.

Чтобы устранить этот пробел, наше исследование стремится использовать потенциал существующих ресурсов (Yeshpanov, Khassanov, Varol, 2022), в частности, адаптируя методологию и результаты предыдущих исследований. Этот подход не только использует богатые лингвистические особенности казахского языка, но также подчеркивает важность инновационных стратегий NLP в активизации усилий по классификации текстов для недостаточно представленных языков. Эта работа основана на новой схеме мультиклассификации, полученной на основе наборов данных распознавания именованных объектов (NER), что открывает путь для практических приложений, требующих детальной категоризации текста.

Благодаря этому начинанию мы стремимся внести свой вклад в расширение объема исследований NLP, ориентированных на казахский язык, демонстрируя жизнеспособность передовых и наиболее часто используемых моделей машинного обучения, таких как LSTM (Bai, 2018) и BERT (Zhang et al., 2020), в решении уникальных проблем, связанных с менее широко изучаемыми языками. Наша методология, основанная на адаптации существующих наборов данных для целей мультиклассификации, обещает предложить новый взгляд на практическую реализацию методов NLP в лингвистически разнообразных условиях.

Новизна нашего исследования заключается в использовании подходов, ранее применявшихся к языкам с богатыми ресурсами, для улучшения классификации текстов на казахском языке, который является языком с ограниченными ресурсами. Наша цель – разработать и внедрить эффективные методологии мультиклассификации, используя существующие наборы данных NER, и адаптировать их для более точной и глубокой категоризации текстов. Мы стремимся показать, что модели LSTM и BERT могут успешно решать задачи классификации для недостаточно изученных языков, обеспечивая тем самым расширение области применения NLP и улучшение качества обработки текстов на казахском языке.

Литературный обзор. Исследователи (Maheen et al., 2022) предлагают архитектуру, которая комбинирует слои CNN + BiLSTM + CNN для того, чтобы превзойти производительность моделей классификации текстов, связанных с BERT, в языках с ограниченными ресурсами. Проведенные эксперименты показывают, что предложенный подход показал конкурентоспособный результат в различных задачах NLP на бенгальском языке и в задаче обнаружения эмоций из шести классов для нового набора данных на бенгальском языке. Также предложенная модель превосходит реализацию BERT для вьетнамских языков и почти так же работает в задачах английского NLP, который тоже имеет искусственный дефицит данных.

В (Garrido-Merchan et al., 2023) исследовали эмпирическое поведение BERT на основе набора различных данных по сравнению со словарем TF-IDF. После добавления эмпирических доказательств в поддержку использования BERT результаты показали превосходство модели BERT над стандартными подходами и независимость от особенностей задачи NLP.

В (Yu et al., 2024) была построена квантово-классическая гибридная архитектура. Данный подход использует предварительно обученную многоязычную двунаправленную модель кодировщика из модели BERT для получения векторных представлений слов и объединяет предлагаемую квантово-рекуррентную нейронную сеть пакетной загрузки и квантовую рекуррентную нейронную сеть с необщей пакетной загрузкой параметров в качестве моделей извлечения признаков для анализа тональности на бенгальском языке, который является малоресурным языком. Проведенные эксперименты показали, что предложенная архитектура обеспечивает максимальное повышение точности на 0,993 % в

задачах классификации бенгальского текста при одновременном снижении средней сложности модели на 12 %.

Положение NLP для казахского языка постепенно обогащается благодаря различным целенаправленным исследованиям. Одной из основополагающих работ является исследование экспериментов по расширенному языковому моделированию казахского языка (Myrzakhmetov, Kozhirbayev, 2018), которое заложило основу для понимания сложностей, связанных с созданием вычислительных ресурсов для недостаточно представленных языков. Это исследование подчеркивает первоначальные шаги по адаптации передовых методов NLP к казахскому языку, подчеркивая потенциал сложных языковых моделей для улучшения обработки и понимания текста.

Основываясь на этих основополагающих моделях, недавние достижения продемонстрировали адаптируемость современных технологий, таких как BERT, к казахскому языку. В частности, исследование извлечения ключевых слов из наборов данных казахстанских новостей с использованием BERT (Abibullayeva, Kazbekova, Zhunissov, 2024) показало, как модели на основе Transformer могут эффективно использоваться для задач, требующих глубокого семантического понимания, таких как определение ключевых терминов и фраз, которые отражают суть новостных статей. Это приложение не только демонстрирует гибкость BERT, но и его потенциал для улучшения поиска информации и анализа контента на казахском языке.

Центральное место в текущем исследовании занимает набор данных KazNERD (Yeshpanov, Khassanov, Varol, 2022), который является значительным вкладом в эту область и предоставляет богато аннотированный корпус для NER на казахском языке. Создание и использование этого набора данных знаменует собой важнейший прогресс в исследованиях NLP для казахского языка, предлагая ценный ресурс для обучения и оценки моделей NLP для задач распознавания сущностей. Этот набор данных служит основой для нашей работы, позволяя исследовать проблемы мультиклассификации через призму NER.

Дальнейшее расширение сферы применения NLP на казахском языке, классификация научных документов с использованием глубоких нейронных сетей и слияние изображений и текста (Bogdanchikov, Ayazbayev, Varlamis, 2022) иллюстрируют новые подходы, изучаемые в сообществе. В этой работе особое внимание уделяется интеграции мультимодальных данных для целей классификации и предлагаются новые пути повышения производительности модели и ее применимости при обработке текстов на казахском языке.

Модели и методы. Для работы с текстом существует большое количество решений. Самая простая и популярная связка – TF-IDF + линейная модель. Данный подход позволяет обрабатывать и решать языковые задачи без особых затрат вычислительных ресурсов. Однако процесс использования такой связки требует дополнительных операций: чистка, лемматизация. Кроме того, модели, использующие LSTM-архитектуру, также требуют предобработки текста. В случае с BERT можно опустить препроцессинг и сразу перейти к токенизации и обучению. Помимо дополнительных шагов, линейные модели часто выдают некорректные результаты, так как не учитывают контекст слов. Понимание контекста является главным преимуществом Transformer.

LSTM. Одним из наиболее эффективных методов для этой задачи являются рекуррентные нейронные сети (RNN), и в частности их улучшенная версия – длинная краткосрочная память (LSTM).

Архитектура модели классификации текста на основе LSTM состоит из трех основных компонентов:

1) входной слой (embedding layer) – преобразует входной текст в векторные представления,

2) LSTM-слой – извлекает временную зависимость и характиристику из последовательности векторных представлений,

3) полносвязный слой (dense layer) – преобразует выходы LSTM-слоя в целевые классы.

Входной слой (Embedding Layer). Пусть $X = [x_1, x_2, \ldots, x_T]$ – входная последовательность текстовых токенов, где T – длина последовательности. Каждый токен x_T из словаря V представляется как вектор $e_{x_t} \in \mathbb{R}^d$, где d – размерность векторного пространства (1).

$$E = [e_{x_1}, e_{x_2}, \ldots, e_{x_t}], \quad (1)$$

где: $E \in \mathbb{R}^{T \times d}$ – матрица эмбеддингов.

Полносвязный слой (Dense Layer). Выходное скрытое состояние последнего временного шага h_T передается на вход полносвязного слоя, который выполняет классификацию (2):

$$y = softmax(W_d h_T + b_d). \quad (2)$$

Здесь $W_d \in \mathbb{R}^{h \times \mathrm{C}}$ – обучаемая весовая матрица, $b_d \in \mathbb{R}^{\mathrm{C}}$ – обучаемое смещение, C – количество классов.

Для обучения модели используется кросс-энтропийная функция потерь (3):

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\, log\left(\hat{y}_{i,c}\right), \quad (3)$$

где: N – количество обучающих примеров;

$y_{i,c}$ – истинная метка класса c для примера i,

$\hat{y}_{i,c}$ – предсказанная вероятность класса c для примера i.

Оптимизация параметров модели производится с использованием алгоритма Adam.

Метод классификации текста на основе LSTM хорошо учитывает временные зависимости и последовательные характеристики текста, что позволяет достичь высоких результатов в задачах классификации.

BERT. Модель BERT от Google является одной из самых эффективных для задач обработки естественного языка. Она предварительно обучена на текстах из открытых источников, таких как Wikipedia и BookCorpus. Архитектура модели основана на технологии Transformer, которая позволяет учитывать зависимости между словами в тексте. После дообучения на специализированных данных BERT достигает отличных результатов в решении задач.

Архитектура BERT включает компонент Encoder от Transformer. Существуют базовые и большие версии модели, отличающиеся количеством слоев и голов внимания: базовая версия имеет 12 слоев, большая – 24 (рис. 1) (Aswini, 2024). Общее количество параметров в модели достигает нескольких сотен миллионов, что требует значительных вычислительных ресурсов. Несмотря на то, что наибольшие успехи первоначально были достигнуты с 24-слойным BERT, часто возникали проблемы из-за ограниченных вычислительных мощностей. Однако данная проблема была решена с помощью улучшенных версий BERT (Chi et al., 2022; Araci, 2019; Liu et al., 2019; Lan et al., 2020; Yang et al., 2019; He et al., 2020; Sanh et al., 2019; Clark et al., 2020).

BERT использует контекстные вложения. Для каждого слова определяется фиксированное количество векторов вложения размером 768 измерений, которые служат входными данными модели. BERT выводит вектор с 768 параметрами для каждого слова или фразы на основе своих внутренних вычислений. Эти выходные последовательности становятся вложениями, представляющими предложение в контексте данных. Проще говоря, начальные входные данные в BERT представлены базовыми векторами, но после

обработки через модель создаются контекстные векторы для каждого слова.

Модель включает части для форматирования входных данных и извлечения вложений. Выделенные токены, такие как [CLS] и [SEP], используются для обозначения начала и конца предложения соответственно. Токен [PAD] используется для дополнения, а [UNK] обозначает неизвестное слово. В слоях вложения включены токенизация, сегментация и позиционирование. Для токенизации в BERT обычно используется метод WordPiece. Сегментационное вложение учитывает семантические отношения между двумя предложениями, а вложение позиции сохраняет информацию о положении слова для создания последовательности слов.

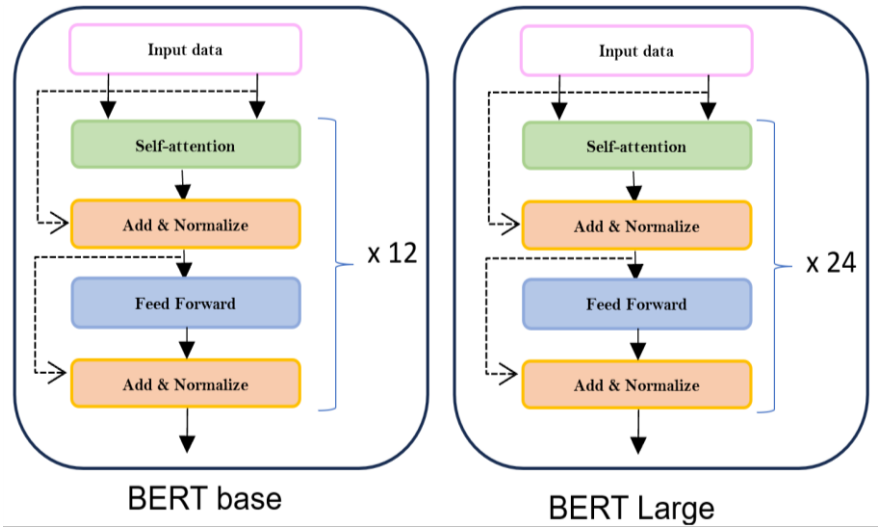


Рисунок 1. Архитектура BERT

Примечание – составлено автором на основе (Zhang, 2020)

Хотя BERT использует двунаправленное кодирование, его работа существенно отличается от рекуррентных нейронных сетей. Это происходит потому, что BERT отображает вложения позиций каждого токена на сегментационные вложения, чтобы представить свойства слова и его местоположение. Другими словами, достаточно форматировать входной текст с использованием вложенного слоя, и кодировщик предоставит соответствующий вывод.

Данные. Было решено создать сложную схему бинарной классификации, которая является основой для решения нашей задачи мультиклассификации. Этот подход позволил нам углубиться в нюансы категоризации текста, используя сильные стороны как сетей с длинной краткосрочной памятью (LSTM), так и моделей двунаправленного кодирования от трансформаторов (BERT), используемых в этой статье. Здесь мы подробно рассказываем о создании четырех бинарных меток классов и последующем процессе мультиклассификации, подчеркивая его значение и методологию.

Учитывая подробные описания именованных объектов (NE), мы разработали схему двоичной классификации, которая отражает аспект данных и разделяет их на две значимые категории. Принимая во внимание природу NE и контекст, в котором они, вероятно, будут использоваться, логическая бинарная классификация будет различать следующие виды:

1. Классификация чувствительности ко времени:

Содержит NE, которые обычно указывают на необходимость своевременного действия или внимания, например: 'DATE', 'TIME'.

2. Классификация организационного взаимодействия:

Предложения с NE, такими как 'ORGANISATION', 'CONTACT', 'LAW', 'MONEY' и 'FACILITY', которые предполагают взаимодействие с организациями или внутри них.

3. Классификация личной значимости:

Предложения, включающие личные имена ('PERSON'), должности ('POSITION') и контактную информацию ('CONTACT'), которые обычно встречаются в личных сообщениях.

4. Классификация коммерческих намерений:

Предложения, в которых упоминаются 'MONEY', 'PRODUCT', 'ORGANISATION' и 'CONTACT', указывают на потенциальную коммерческую деятельность или транзакции.

В рамках данного исследования были логически выбраны категории, созданные вручную, служащие в первую очередь для демонстрации функциональности метода. Однако наши будущие усилия направлены на автоматизацию создания классификационных тегов с использованием сложных алгоритмов, которые учитывают более широкий спектр NE внутри предложения и его контекстуальное значение. Это усовершенствование не только упростит процесс классификации, но также повысит точность и применимость задачи мультиклассификации при анализе текста на казахском языке.

В общей сложности используемый набор данных включает 110 675 последовательностей в наборах обучения, проверки и тестирования. Он богато аннотирован для детальной классификации текста, а его словарный запас превышает 72 000 слов. Примечательно, что в обучающем наборе обнаружено более 20 000 экземпляров, помеченных как «чувствительные ко времени», а также заметные значения для «организационного взаимодействия» (10 577), «личной значимости» (10 729) и «коммерческого намерения» (9 115). Такое справедливое распределение тегов имеет решающее значение для детального понимания и обработки NLP, предлагая прочную основу для продвижения задач классификации текста и распознавания именованных объектов в недостаточно представленных лингвистических исследованиях.

Препроцессинг. Для предварительной обработки текстовых данных было выполнено несколько шагов. Сначала данные были токенизированы с использованием WordPiece для модели BERT (из библиотеки Hugging Face Transformers). Для модели LSTM использовалась токенизация с применением NLTK. Это обеспечивало разделение текста на компоненты для дальнейшей обработки. Затем были удалены стоп-слова, лишняя пунктуация и специальные символы. Также была проведена лемматизация для приведения слов к их базовой форме. Этот шаг особенно важен для таких моделей, как LSTM, которые выигрывают от более стабильной структуры входных данных.

Результаты и их обсуждение. При оценке эффективности моделей NLP для классификации казахского текста мы использовали LSTM и BERT. Эти модели были протестированы на нашем подготовленном наборе данных с целью расшифровки сложной сети многомерных атрибутов, присущих тексту.

Параметры обучения моделей были выбраны для максимизации их производительности на валидационных данных и улучшения способности обрабатывать текстовые данные на казахском языке. Для модели LSTM использовались следующие параметры: количество эпох – 50, размер батча – 32, скорость обучения – 0.001, оптимизатор – Adam, функция потерь – кросс-энтропия, метод регуляризации – Dropout (0.5) и инициализация весов – He. Для модели BERT были выбраны следующие параметры: количество эпох – 5, размер батча – 16, скорость обучения – 2e-5, оптимизатор – AdamW, функция потерь – кросс-энтропия, а график изменения скорости обучения был настроен на линейный спад.

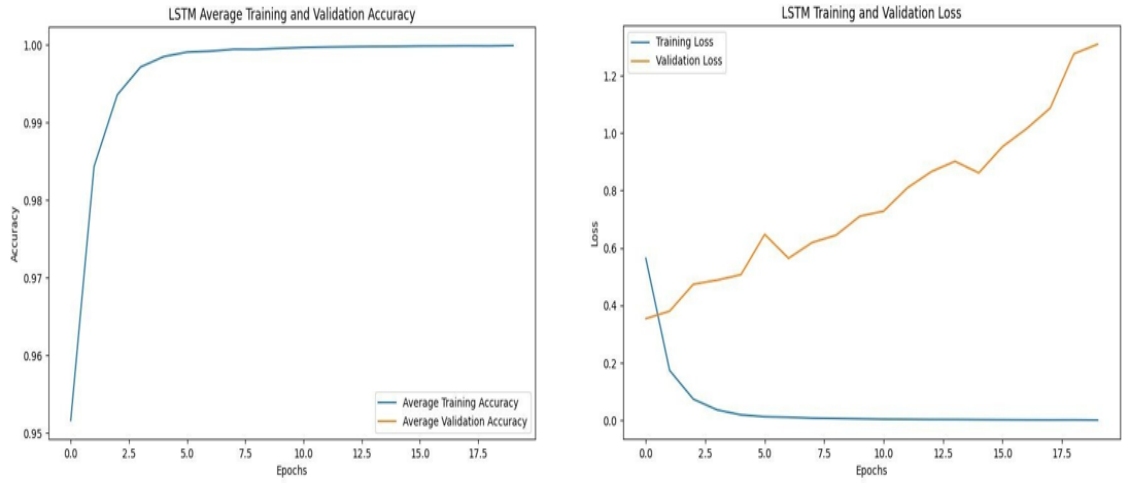
Для оценки производительности модели мы используем такие показатели, как точность и потери, которые легче интерпретировать в задачах классификации. Точность измеряет процент правильно классифицированных объектов, что позволяет легко понять, насколько

эффективна модель. Потери (например, кросс-энтропия) показывают разницу между предсказанными и реальными метками, что помогает оценить достоверность модели.

После того как мы установили двоичные метки для данных, исследование перешло к задаче мультиклассификации. В этой задаче тексту одновременно присваивается несколько меток. Это не просто определение одной категории, а понимание различных характеристик текста, которые могут быть полезны для классификации. Мы использовали модели LSTM и BERT для мультиклассификации, так как обе модели показали хорошие результаты в классификации текста. LSTM эффективно работает с долгосрочными зависимостями в данных, а BERT, основанный на архитектуре Transformer, лучше понимает контекст и связи между словами в тексте. Это сравнение стало ключевым для оценки их эффективности в обработке сложных текстов на казахском языке.

Результаты экспериментов, полученные с помощью графиков потерь и точности при обучении и проверке, дают убедительное представление о закономерностях обучения моделей:

1. Для LSTM потери при обучении резко снижались перед стабилизацией, в то время как потери при проверке испытывали колебания, что указывает на области, где обобщение модели можно улучшить (рис. 2). График точности для LSTM продемонстрировал достаточно заметный подъем к высоким уровням точности, что предполагает хорошее понимание задачи классификации.



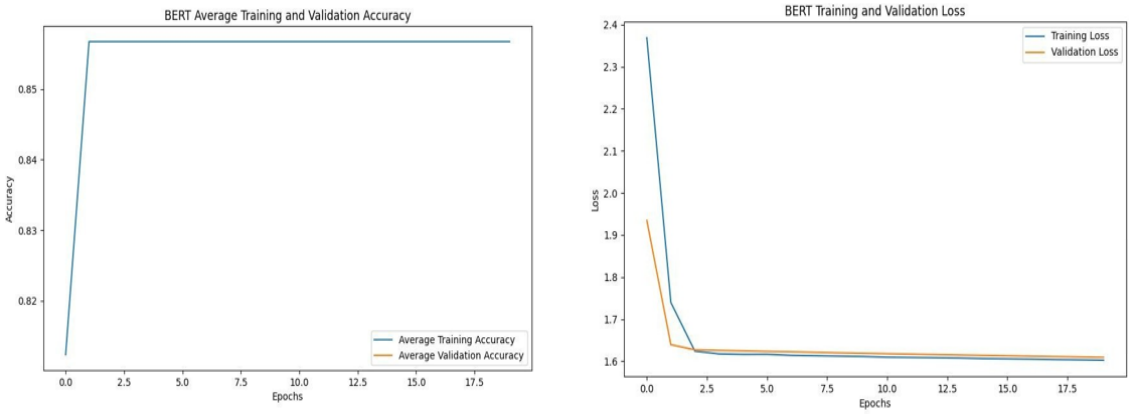
а *б*

Рисунок 2. а – средняя точность обучения и проверки LSTM;
б – потеря обучения и проверки LSTM

Примечание – составлено автором

2. Производительность модели BERT, представленная на графиках потерь и точности (рис. 3), показала быстрое снижение потерь в процессе обучения, которое стабилизировалось с увеличением числа эпох. Это подтверждает, что модель эффективно обучается на предоставленных данных. График точности показывал высокие результаты на протяжении всего обучения, что свидетельствует о высокой надежности BERT в задачах классификации текста на казахском языке. В нашем сравнении модель BERT показала лучшие результаты по сравнению с LSTM. BERT продемонстрировал отличные результаты в обработке сложных текстов казахском языке благодаря своей способности понимать контекст. Это подтверждается стабильным поведением на графиках потерь и точности.

Для более наглядного сравнения производительности моделей LSTM и BERT на обучающем и валидационном наборах данных результаты представлены в виде таблиц (табл. 1 и 2).



а *б*

Рисунок 3. а – средняя точность обучения и проверки BERT;
б – потеря обучения и проверки BERT
Примечание – составлено автором

Таблица 1. Производительность моделей на обучающих данных

Метрика	LSTM	BERT
Потери	0,1993	0,0477
Точность (Accuracy)	0,4674	0,9255
Точность (Precision)	0,8453	0,9682
Примечание – составлено автором		

Таблица 2. Производительность моделей на валидационных данных

Метрика	LSTM	BERT
Потери	0,1993	0,0481
Точность (Accuracy)	0,4681	0,9202
Точность (Precision)	0,8524	0,9712
Примечание – составлено автором		

В процессе экспериментов были выявлены случаи, в которых модель LSTM показывала ошибки классификации, особенно при обработке сложных текстов с неочевидными контекстами. LSTM иногда не могла точно распознать временные метки, особенно в сложных предложениях с несколькими временными ссылками. А при наличии неструктурированной информации (например, имен) модель иногда путала контекст, что приводило к ошибочной классификации.

Для улучшения данной модели и лучшего распознавания временных меток и имен можно добавить дополнительные шаги в предобработку, такие как использование внешних словарей для именованных сущностей. Можно использовать более мощные модели на основе трансформеров с улучшенной настройкой или подходы с использованием двухуровневого классификатора. Такие модели сначала детектируют сущности, а затем классифицируют их по категориям.

На графиках точности и потерь видно, что модель BERT стабильно улучшает точность в процессе обучения, что связано с его архитектурной особенностью, в то время как LSTM показывает некоторое колебание точности на валидационных данных. Это может указывать на проблему переобучения у LSTM, что проявляется в несоответствии между результатами на обучающем наборе и валидационном. BERT обладает двусторонним пониманием контекста, которое помогает модели правильно классифицировать сложные тексты.

Заключение. Внедрение системы из четырех двоичных меток и её применение в задаче мультиклассификации для анализа текстов на казахском языке представляет значительный прогресс в исследованиях NLP. Сравнивая модели LSTM и BERT в этом контексте, наша работа не только способствует пониманию эффективности моделей для недостаточно представленных языков, но и открывает новые возможности для практического применения NLP. Мы продемонстрировали, как модели LSTM и BERT могут быть эффективно развёрнуты в структуре мультиклассификации. Наши эксперименты подтвердили практичность и надёжность предложенной методологии в решении сложных задач многоклассовой классификации. Успех нашего подхода прокладывает путь для дальнейших улучшений, включая совершенствование модели BERT и развитие автоматической генерации меток классификации. Мы продолжаем работать над укреплением этого подхода, уделяя особое внимание созданию более надёжной системы генерации меток, способной максимально использовать потенциал BERT в различных приложениях NLP.

Благодарности: Работа выполнена при финансовой поддержке Комитета науки Министерства науки и высшего образования Республики Казахстан (Грант № AP19174298).

Список литературы

Abibullayeva A.A., Kazbekova G.N., Zhunissov N.M. (2024). Keyword extraction from kazakh text with machine learning algorithms. Bulletin of the Abai KazNPU, the series of Physical and Mathematical Sciences, vol. 1(85), 106-113.

Araci D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. ArXiv, abs/1908.10063, https://doi.org/10.48550/arXiv.1908.10063.

Aswini R. (2024, August 31). How Does BERT NLP Optimization Model Work? https://www.turing.com/kb/how-bert-nlp-optimization-model-works.

Bai X. (2018). Text classification based on LSTM and attention. 2018 Thirteenth International Conference on Digital Information Management (ICDIM), Berlin, Germany, 29-32, doi: 10.1109/ICDIM.2018.8847061.

Bogdanchikov A., Ayazbayev D., Varlamis I. (2022). Classification of Scientific Documents in the Kazakh Language Using Deep Neural Networks and a Fusion of Images and Text. Big Data and Cognitive Computing, vol. 6(123), doi: 10.3390/bdcc6040123.

Chi Z., Huang S., Dong L., Ma S., Zheng B., Singhal S., Bajaj P., Song X., Mao X.-L., Huang H., Wei F. (2022). XLM-E: Cross-lingual Language Model Pre-training via ELECTRA. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, vol. 1, 6170–6182, doi: 10.18653/v1/2022.acl-long.427.

Clark K., Luong M.-T., Le Q. V. Manning, C. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. ArXiv: 2003.10555, https://doi.org/10.48550/arXiv.2003.10555.

Dogra V., Verma S., Kavita Chatterjee, P., Shafi J., Choi J., Ijaz M. F. (2022). A Complete Process of Text Classification System Using State-of-the-Art NLP Models. Computational Intelligence and Neuroscience, Article 1883698. https://doi.org/10.1155/2022/1883698.

Garrido-Merchan, E. C., Gozalo-Brizuela, R., Gonzalez-Carvajal, S. (2023). Comparing BERT Against Traditional Machine Learning Models in Text Classification. Journal of Computational and Cognitive Engineering, vol. 2(4), 352–356. https://doi.org/10.47852/bonviewJCCE3202838.

He P., Liu X., Gao J., Chen W (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. ArXiv, abs/2006.03654.

Lan Z., Chen M., Goodman S., Gimpel K., Sharma P., Soricut R. (2020). ALBERT: A Lite BERT for Self-

supervised Learning of Language Representations. ArXiv, abs/1909.11942, https://doi.org/10.48550/arXiv.1909.11942.

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, arxiv:1907.11692, https://doi.org/10.48550/arXiv.1907.11692.

Maheen S. M., Faisal M. R., Rahman Md. R., Karim Md. S. (2022). Alternative non-BERT model choices for the textual classification in low-resource languages and environments. In Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing, 192–202, doi: 10.18653/v1/2022.deeplo-1.20.

Mamyrbayev O., Oralbekova D. (2020). Modern trends in the development of speech recognition systems. News of the National academy of sciences of the republic of Kazakhstan, vol. 4, no. 332, 42-51. https://doi.org/10.32014/2020.2518-1726.64.

Myrzakhmetov, B., Kozhirbayev, Zh. (2018). Extended Language Modeling Experiments for Kazakh. International Workshop on Computational Models in Language and Speech, vol. 2303, 35–43.

Oralbekova D., Mamyrbayev O., Othman M., Kassymova D., Mukhsina K. (2023). Contemporary Approaches in Evolving Language Models. Applied Sciences, vol. 13(23):12901. https://doi.org/10.3390/app132312901.

Sanh V., Debut L., Chaumond J., Wolf T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv, arXiv:1910.01108.

Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., Le Q.V. (2019). XLNet: generalized autoregressive pretraining for language understanding. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, vol. 517, 5753–5763, https://doi.org/10.48550/arXiv.1906.08237.

Yeshpanov R., Khassanov Y., Varol H. A. (2022). KazNERD: Kazakh Named Entity Recognition Dataset. Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), European Language Resources Association, 20-25 June 2022, 417-426, https://doi.org/10.48550/arXiv.2111.13419.

Yu W., Yin L., Zhang C., Chen Y., Liu A. X. (2024). Application of Quantum Recurrent Neural Network in Low-Resource Language Text Classification. IEEE Transactions on Quantum Engineering, vol. 5, no. 2100213,1-13, doi: 10.1109/TQE.2024.3373903.

Zhang A., Bohan L., Wang W., Wan S., Chen W. (2020). MII: A Novel Text Classification Model Combining Deep Active Learning with BERT. Computers, Materials & Continua, vol. 63, 1499-1514, https://doi.org/10.32604/cmc.2020.09962.

Information about authors

Oralbekova Dina – PhD, Institute of Information and Computational Technologies, Almaty, Kazakhstan, E-mail: dinaoral@mail.ru, ORCID: 0000-0003-4975-6493, +7 771 131 01 88

Mamyrbayev Orken – PhD, Institute of Information and Computational Technologies, Almaty, Kazakhstan, E-mail: morkenj@mail.ru

Imansakipova Ayagoz – master of technical sciences, Almaty technological university, Almaty, Kazakhstan, E-mail: aimansakipova@bk.ru

Zhiger Aliya – master of technical sciences, Narxoz University, Almaty, Kazakhstan, E-mail: alia_94-22@mail.ru

Mukhsina Kuralai – PhD, Institute of Information and Computational Technologies, Almaty, Kazakhstan, E-mail: email123@gmail.com