Advanced Implementation of a Multilevel Model for Text Summarization in Kazakh Using Pretrained Models

Dina Oralbekova
Institute of Information and Computational Technologies, Almaty, Kazakhstan | International Engineering and Technological University, Almaty, Kazakhstan
d.oralbekova@ipic.kz (corresponding author)

Orken Mamyrbayev
Institute of Information and Computational Technologies, Almaty, Kazakhstan
morkenj@gmail.com

Mohamed Othman
Malaysia Department of Communication Technology and Networks, Universiti Putra Malaysia, Serdang, Malaysia | Laboratory of Computational Science and Mathematical Physics, Institute for Mathematical Research, Universiti Putra Malaysia, Serdang, Malaysia
mothman@upm.edu.my

Sholpan Zhumagulova
Al-Farabi Kazakh National University, Almaty, Kazakhstan
sh.zhumagulovakz@gmail.com

Received: 17 June 2025 | Revised: 11 July 2025 and 17 July 2025 | Accepted: 20 July 2025

 DOI: https://doi.org/10.48084/etasr.12799

ABSTRACT

This study investigates transformer models for the task of hybrid text summarization in the Kazakh language. Using mBART, mT5, and XLM-RoBERTa models, a multilevel architecture was developed that processes text at the character, subword, word, and contextual levels. The proposed system performs feature fusion across multiple linguistic layers, enabling the model to capture both fine-grained lexical variation and broader contextual dependencies. The architecture also allows flexible integration with various transformer models, supporting both encoder-decoder and hybrid configurations. This approach significantly improved the quality of generated summaries by effectively accounting for the morphological and semantic features of the Kazakh language. The experimental results showed that mBART achieved the best performance in terms of ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore-F1 metrics, confirming the high effectiveness of the proposed multilevel transformer architecture. This is the first implementation of such an architecture for hybrid summarization in Kazakh, which is a low-resource and morphologically rich language.

Keywords-multilevel modeling; Kazakh language; hybrid summarization; transformer models; mBART; mT5; XLM-RoBERTa

I. INTRODUCTION

Language Modeling (LM) is one of the central tasks in Natural Language Processing (NLP), which involves the creation and development of algorithms capable of processing, understanding, and generating textual data in human languages. In recent decades, this process has gained significant popularity due to advances in deep learning and the emergence of transformer models [1-2]. These models can take into account context, sentence structure, and text semantics. Today, LM serves as the foundation for a wide range of applications, including machine translation, text summarization, question answering, sentiment analysis, and more [3-4]. One of the most important tasks in the field of LM is text summarization, which can be classified into two types: extractive summarization (extracting sentences from the original text) and abstractive summarization (where a new text is generated, briefly summarizing the main ideas of the original) [5]. In recent years,

transformer models such as BERT [6], GPT [7], T5 [8], and BART [9] have shown significant progress in text generation tasks, including summarization. BART and T5 are advanced models that have been specifically optimized for text generation tasks, such as abstractive summarization, due to their encoder-decoder architecture. However, in addition to purely extractive and abstractive summarization, there is also a hybrid approach that combines elements of both methods [10]. Hybrid summarization involves extracting key fragments of text and then generating a summary based on them, combining extraction accuracy with the flexibility of generation. This approach can be especially useful for improving the quality of summaries. Moreover, it eliminates the drawbacks of purely abstractive methods, which are associated with the risk of information distortion, while preserving the creative aspect of text formation, characteristic of abstractive summarization. In the context of the Kazakh language, with a limited volume of data and resources, hybrid models can significantly improve the accuracy and coherence of summaries.

Multilingual models, such as mBART and mT5, can work with many languages, including those with limited resources [11]. These models allow overcoming language barriers and applying common algorithms for working with various languages. mBART (Multilingual BART) and mT5 (multilingual T5) have demonstrated high performance in multitask scenarios, such as translation, text analysis, and summary generation. XLM-RoBERTa [12], a multilingual version of RoBERTa [13], has also been shown to be effective for various tasks in LM, including text classification and information extraction. However, despite significant advances in language models for widely spoken languages such as English, French, and Chinese, low-resource languages remain a major challenge in the field of LM. Kazakh is one such language, with its agglutinative structure, rich morphology, and unique features that require specialized approaches for processing. However, the availability of corpora and pretrained models for Kazakh is limited, which significantly restricts the application of modern LM algorithms. Developments in text summarization for Kazakh are still in an early stage, and existing solutions have shown unsatisfactory results, failing to achieve high performance metrics [14-16].

Unlike previous multilevel approaches that combine features mainly at the word or subword level, the proposed model includes four different types of linguistic features, character, word, subword, and contextual, which are especially suitable for agglutinative and low-resource languages such as Kazakh. The proposed architecture also supports flexible use with encoder-decoder transformer models, such as mBART and mT5, making it possible to combine detailed linguistic information with the strong generative abilities of large pretrained models. In contrast to previous approaches, this study introduces a novel multilevel summarization framework that fuses character, subword, word, and contextual-level features into a unified architecture. This is the first known implementation of such a detailed multilevel model specifically adapted for the Kazakh language. Experimental results demonstrate significant performance gains over both hybrid and traditional transformer-based summarization methods.

In recent years, significant efforts have focused on the development of hybrid summarization models, adopted for both rich- and low-resource languages. These works serve as the foundation for the present study and highlight the ongoing developments of multilingual and cross-lingual summarization techniques. In [17], an abstractive summarization model was developed for the Indonesian language, based on an embedding stack as the encoder and a transformer-based decoder. The embeddings used included BERT, Byte Pair Encoding (BPE), Character Embedding (CE), and FastText. This study investigated the impact of the selection of BERT layers and embedding configurations on summary quality. The model was trained on Liputan6 sub-corpora (50K and 75K articles). The highest scores for the ROUGE-1 (37.18), ROUGE-2 (18.19), and ROUGE-L (34.28) metrics were achieved when using all BERT layers. The proposed architecture demonstrated high performance even with a limited training dataset.

In [18], a hybrid summarization system was proposed for the Urdu language, combining extractive methods (TF-IDF, sentence weighting, word frequency) with a BERT-based abstractive model. Summary quality was evaluated by Urdu language experts. The corpus included articles from sources such as Express, BBC Urdu, and Dawn covering diverse topics. This was one of the first abstractive summarization studies for Urdu, addressing its complex morphology and lexical diversity. In [19], an integrated structure was proposed for abstractive summarization, combining semantic generalization methods, Word Sense Disambiguation (WSD), and a neural seq2seq model with attention. Text vectorization was performed using Word2Vec, and the architecture included bidirectional LSTMs, custom attention, and a TimeDistributed layer. The model was trained and tested on datasets such as Gigaword, DUC-2004, and CNN/DailyMail. The use of WSD and post-processing improved the consistency and coverage of rare words, improving summary quality compared to the baseline seq2seq models without semantic augmentations.

In [20], a hybrid architecture was developed for patent summarization by combining LexRank and BART, with fine-tuning via Low-Rank Adaptation (LoRA). Meta-learning was used to generalize the model to new patent domains. Tested on long, legally complex texts, the approach reduced computational costs while maintaining quality, proving effective for abstractive summarization in intellectual property. In [21], the Discrete Diffusion Language Model (DDLM) used Semantic-Awareness routing and a CrossMamba architecture, an adaptation of the Mamba model customized for long-text summarization tasks. This model outperformed previously introduced discrete diffusion models and even outperformed autoregressive models in generation speed on datasets such as CNN/DailyMail, ArXiv, and Gigaword. The key contribution of this study lies in shifting from random to semantically guided noise, which enables the model to prioritize the retention of important tokens and better control the structure of the generated output.

In [22], a two-stage summarization system was developed, combining Graph Neural Networks (GNN) to select key discourse units (EDUs) and the BART model to generate the final text. The approach relied on initial text segmentation into

elementary units, which were then evaluated based on a graph structure. Experiments were conducted on the CNN/DailyMail dataset, and the proposed method demonstrated higher ROUGE scores compared to baseline transformers. This method showcases the effectiveness of integrating graph structures with generative transformers. In [23], a method for abstractive summarization of business reports was proposed, which integrated the processing of both tabular and textual data. A transformer architecture incorporating elements of the Switch Transformer was used to enhance accuracy when working with multi-format financial content. The main focus was on adapting the model to tabular structures in financial reports. The research was conducted on several business datasets, including Reuters Financial and Bloomberg. The model's performance was evaluated using ROUGE and F-measure metrics, showing improved results compared to the classical transformer model.

In [24], a comparative analysis was performed on various summarization models, including RNN, LSTM, Seq2Seq, BART, T5, and Pegasus. The study focused on news summarization tasks, evaluating the accuracy, coherence, and overall suitability of the models. Several corpora, including CNN/DailyMail and BBC News, were used. The study highlighted the advantages of the Pegasus and BART models, particularly in generating contextually meaningful summaries. Additionally, the limitations of ROUGE metrics and the need for more complex evaluations (e.g., coherence and factual accuracy) were discussed. In [25], a new evaluation metric for abstractive summarization was proposed, based on the semantic similarity between the reference and the generated summaries. The Universal Sentence Encoder (USE) was used as the base model, allowing comparison of meanings at the sentence level. The proposed metric was shown to correlate better with human evaluations than ROUGE, especially in cases of low n-gram overlap. Experiments were conducted on CNN/DailyMail and other datasets. The study also explored other alternative metrics, including BERTScore, MoverScore, and ROUGE-G. In [26], the ARLED model was proposed for abstractive summarization of long texts in Persian. The architecture combined the ARMAN model (based on Longformer) and the LED (Longformer Encoder-Decoder), allowing it to handle inputs of up to 8192 tokens. A new corpus of 49,457 Persian scientific articles was created from ensani.ir, with extensive preprocessing. During training, HuggingFace AutoTokenizer was used, and specialized strategies for semantic sentence reordering were implemented.

In [27], an automated system was introduced to summarise abstracts and titles of multiple documents using large language models (LLMs), including GPT-4. The system employed a hybrid approach: first, extraction of key fragments was performed, followed by abstractive summarization using the LLM. The evaluation used ROUGE, METEOR, and BLEU metrics, along with the introduction of new qualitative metrics. The system was shown to be particularly effective in the technical and medical domains. In [28], the open Hungarian corpus HunSum-2 was presented, which was created from Common Crawl news texts and is intended for both extractive and abstractive summarization tasks. FastText was used for language detection, and quntoken was applied for tokenization and sentence segmentation during data preprocessing. In abstractive experiments, models based on mT5 and the Bert2Bert architecture were used, with an input sequence length limitation of 512 tokens. In the extractive approach, BertSum models were implemented using huBERT and a simple classifier, where sentence labels were generated based on the cosine distance between their embeddings. The results showed that extractive models outperformed abstractive ones in terms of ROUGE and BERTScore metrics. Among abstractive systems, the mT5 model achieved the best results but still lagged behind extractive solutions in terms of accuracy and completeness. The main drawbacks of abstractive models were related to factual inaccuracies and discrepancies with the content of the original text, highlighting the need for further improvements.

II. METHODOLOGY

This section formalizes the proposed multilevel modeling framework for hybrid summarization of Kazakh texts. The approach integrates multiple linguistic levels: character, subword, word, and contextual into a unified architecture. Each level captures different granularities of linguistic features and contributes complementary representations for improved summarization quality.

A. Problem Formulation

Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ be a dataset of N Kazakh text-summary pairs, where x_i is the input document and y_i is the reference summary. The objective is to learn a conditional probability distribution $P(y|x)$ that generates coherent and semantically faithful summaries $\hat{y}$, such that $\hat{y} = argmax_y\ P(y|x;\ \theta)$, where θ are the model parameters. To improve the model's generalization in a low-resource setting, a multilevel encoder pipeline fuses representations at four linguistic levels.

B. Multilevel Encoding Architecture

Multilevel modeling is based on the idea that successful text processing, particularly for languages with rich morphology and complex structures, requires consideration of various levels of text units. In this method, each model level plays a crucial role (see Algorithm 1).

- Character level encoding (h^{char}) helps the model understand the structure of words, which is important for the Kazakh language, where changes in word forms can dramatically alter their meaning. Models such as character-level embeddings enable text processing at the level of individual characters [29]. Character embeddings $e_c \in \mathbb{R}^d$ are learned using a shallow BiLSTM or convolutional encoder to capture sub-morphemic variations, especially relevant in agglutinative structures.
- Subword level (h^{sub}) allows efficient handling of compound words and agglutinative languages, where a single word can be complex and multifaceted. Text is tokenized using SentencePiece or BPE to create a subword sequence $s = \{s_1, \dots, s_T\}$ [30]. These tokens are passed through pretrained transformer encoders to obtain contextualized subword representations.

- Word-Level Encoding (h^{word}) FastText embeddings $e_w \in \mathbb{R}^d$ are used to encode semantic properties at the word level, including representations for rare or morphologically rich forms. Optionally, POS tags from the Stanza toolkit are embedded and concatenated to word vectors to enhance syntactic awareness [31].
- Contextual level (h^{ctx}). Full-sequence embeddings are computed using contextual encoders such as XLM-RoBERTa, producing representations that capture long-range dependencies. For XLM-RoBERTa, outputs $h^{ctx} = XLM - R(x)$ are used as the base for generation when paired with a separate decoder.

The fused representation at each token position t is defined as:

$$h_t = Fuse(h_t^{char}, h_t^{sub}, h_t^{word}, h_t^{ctx})$$

where $Fuse()$ can be:

- concatenation+projection: a linear transformation W_f applied to concatenated vectors,
- attention-based fusion: learnable weights over levels,
- gated mechanisms: such as Highway layers or gating functions to control contribution from each level.

```
Algorithm 1: Hybrid Text Summarization
Using Pretrained Models
Goal: Utilize models for generating
  summaries of Kazakh texts with a
  multilevel approach, including
  processing at the character, subword,
  word, and contextual levels.
Inputs: A corpus of Kazakh texts with
  corresponding summaries.
Outputs: Generated summaries for
  representative Kazakh texts.
Load the Kazakh text dataset and
corresponding summaries.
Preprocess the input texts:
- Add the prefix "summarize:" to each
  input text.
- Perform character-level tokenization
  using models such as character-level
  embeddings to capture word structure,
  particularly for Kazakh, where
  morphological changes are critical.
- Perform subword tokenization using
  methods such as BPE or SentencePiece to
  handle complex compound words typical
  for the Kazakh language.
- Tokenize the corresponding summaries.
Initialize a pretrained model for
  sequence-to-sequence tasks.
Fine-tune the model on the prepared
  dataset (Kazakh texts with summaries)
  using appropriate training parameters,
  such as learning rate, batch size, and
  number of epochs.
Compute evaluation metrics during training
  to monitor model performance.
Generate summaries for each input text by
  feeding it into the trained model. The
  model should ensure proper processing of
  contextual dependencies, including both
  local and global dependencies in the
  text.
Output the generated summaries and assess
  model performance based on the computed
  metrics
```

C. Model Components

1) mT5 Model

mT5 is a multilingual version of the T5 (Text-to-Text Transfer Transformer) model, which was trained on more than 100 languages, including Kazakh. The mT5 model uses an encoder-decoder architecture and is designed to perform various NLP tasks in the universal "text in - text out" format. This allows it to effectively address tasks such as machine translation, text generation, summarization, and many others.

At the core of mT5 is a transformer, where the encoder processes the input text, and the decoder generates the corresponding output text. This makes the model suitable for sequence-to-sequence tasks, such as hybrid summarization, where the goal is not only to extract text, but also to generate new text that conveys the main ideas of the original. mT5 uses SentencePiece tokenization to convert text into sequences of tokens. In the architecture, the Encoder breaks the input text into tokens and passes them through several layers of the self-attention mechanism to compute hidden states. Each layer consists of multi-head attention and a feed-forward neural network. After passing through all the layers, the output representations become contextual embeddings, which are then passed to the Decoder. The Decoder receives the hidden states from the Encoder and uses them to generate the output text. The Decoder also uses self-attention, allowing it to take into account previously generated tokens when producing subsequent ones. Thus, the model predicts each successive token based on the previously generated ones and the hidden states.

2) mBART Model

mBART (Multilingual BART) is a multilingual version of the BART (Bidirectional and Auto-Regressive Transformers) model, developed for text generation tasks such as machine translation, text summarization, and other sequence-processing tasks. Like mT5, the mBART model uses an encoder-decoder architecture. mBART is trained on multiple languages using a multilingual corpus, enabling it to handle texts in various languages, including Kazakh. It uses SentencePiece for tokenization. The model has demonstrated excellent results in translation, generation, and text summarization tasks, particularly when working with low-resource languages. mBART combines the advantages of BERT (through the bidirectional context in the encoder) and GPT (through the autoregressive decoder).

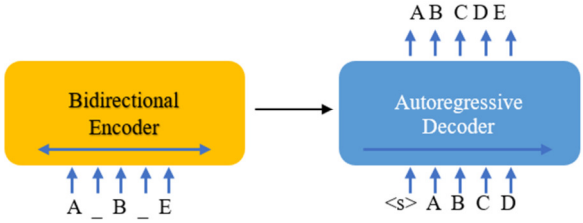

Fig. 1. BART architecture.

3) XLM-RoBERTa Model

XLM-RoBERTa (Cross-lingual RoBERTa) is an improved version of the RoBERTa model, designed to work with multilingual texts. It uses a transformer architecture trained on a massive multilingual corpus, which includes Kazakh. Like RoBERTa, XLM-RoBERTa employs a transformer architecture that considers the context to the left and the right of each token. This allows the model to capture long-term dependencies and improve text understanding. XLM-RoBERTa was trained on a large multilingual corpus that includes more than 100 languages, making it particularly well-suited for working with low-resource languages. This model is capable of extracting contextual features and relationships between words, even when they appear in rare or complex forms. XLM-RoBERTa uses enhanced pretraining methods, including masked token prediction and improved representations for various languages.

The model uses an encoder to process input texts. Each token in the text is transformed into contextual embeddings that capture both semantic and syntactic dependencies. Self-attention is employed, allowing the model to consider both the context to the left and right of each token in the sentence. The output of the encoder consists of a set of hidden states that represent the input text in a more abstract form. These hidden states can then be used to extract features at the word- and contextual levels. These features contain important information on the meaning and context of words in the sentence for further analysis and text generation. XLM-RoBERTa uses the BPE tokenizer, which allows it to handle languages with a large number of rare or unknown words. BPE splits words into subwords, which helps in processing rare or compound words and adapting to new linguistic data.

D. Embedding Multilevel Language Model

Embedding a multilevel language model into the text summarization task involves processing text at multiple levels, which is especially important for languages with rich morphology. Each level of the model plays a key role in extracting and processing information from the text. This study uses a multilevel language model that includes the character, subword, word, and contextual levels. Character-level embeddings are used at the character level. This enables the model to analyze and understand the structure of words. This is particularly important for Kazakh, where a word can change significantly depending on its grammatical form, and the same morpheme can be written in different variations. The character level helps the model work effectively with changing word forms, enhancing its ability to handle the lexical diversity of the language.

For Kazakh, with its rich affixation, subword level tokenization is performed using SentencePiece. This allows the model to handle compound words and affixed forms, which are frequently encountered in Kazakh texts. SentencePiece splits words into smaller units (subwords), enabling the model to deal effectively with unknown words and morphemes.

At the word level, FastText is used for word vectorization. This helps the model capture semantic dependencies between words and their meanings. FastText considers the context of each word and creates representations that may not have appeared in the training corpus. This is especially useful for Kazakh, where many words may be rare or compound, and FastText helps the model to understand their meanings.

At the contextual level, the XLM-RoBERTa model or other transformer models are used to extract contextual features. This step is crucial for capturing dependencies between words within a sentence and across sentences in the text. XLM-RoBERTa considers both local and long-term dependencies in the text, which helps the model to better understand the meaning of a sentence and the context in which each word is used. This context is critical for generating a summary, as it is important not only to extract information but also to paraphrase it into a concise form.

Once the text has been processed at all levels, BART generates the summary using the contextual features obtained from XLM-RoBERTa. BART uses an encoder-decoder architecture, where the encoder extracts information from the text, and the decoder creates a new summary by paraphrasing the main ideas of the original text (Figure 2). This enables the generation of high-quality summaries that retain the essence of the original material while presenting it in a shorter and more concise form.

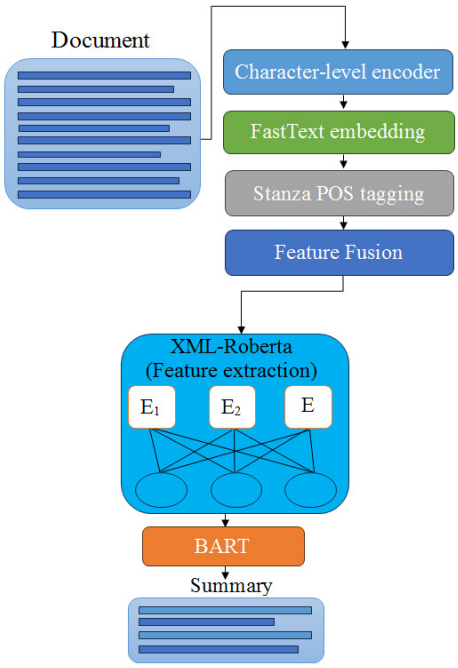

Fig. 2. Multilevel language modeling with XLM-RoBERTa and BART.

However, for the mBART and mT5 models, the use of the BART model is not required. Both of these approaches already incorporate the encoder-decoder architecture. Therefore, the mBART and mT5 models can directly generate summaries, as they use a similar structure for sequence-to-sequence tasks, including text generation. For XLM-RoBERTa, the use of BART is necessary because XLM-RoBERTa is a model based solely on the transformer encoder. It does not have a text generation mechanism, and to create a summary, BART is additionally used to decode the extracted features and generate the final text.

The proposed model adopts a multilevel architecture for hybrid summarization of Kazakh texts, integrating heterogeneous linguistic features through four distinct encoding levels: character, word, subword, and contextual. The input document is initially processed via a character-level encoder to capture morphological variations inherent in agglutinative languages. These representations are enriched with word-level FastText embeddings and POS tags extracted using the Stanza toolkit. Concurrently, contextual features are derived through a pretrained XLM-RoBERTa encoder, capturing long-range dependencies across the text. All representations are aggregated in a dedicated Feature Fusion module, where fusion is performed via concatenation, followed by a learnable linear projection. The resulting unified embedding is then passed to a BART decoder for summary generation. In an alternative configuration, the model supports direct use of encoder-decoder transformer architectures such as mBART or mT5, where the multilevel input is aligned to the model's subword tokenization. This dual-path architecture enhances the model's flexibility, enabling it to benefit from both hierarchical linguistic cues and the generative capabilities of large pretrained transformers.

III. EXPERIMENTS

A. Dataset

To fine-tune and evaluate the summarization model, a domain-specific dataset was created, consisting of 2,000 news articles collected from four major web resources: informburo.kz, baq.kz, kaz.tengrinews.kz, and kaz.nur.kz. These sources were selected for their credibility and thematic diversity, covering news reports, expert opinions, social topics, and entertainment. To improve consistency and stability during training, the texts were formatted in a standardized way: most articles contain no more than 250 words, while the corresponding summaries average around 40 words (Figure 3). This helped balance the dataset in terms of length and avoid strong biases between short and long materials, while maintaining sufficient diversity in terms of topics and structure.

The average length of an article was 112 words, with a median length of 98 words. The summary length ranged from 5 to 50 words, with an average of approximately 21 words. The thematic and source-wise distribution of the dataset was analyzed, which is presented in Table I. The data were balanced in terms of topic diversity and article length to avoid overfitting the model to specific content types or lengths. Such a distribution also helps to ensure a more robust generalization during inference.

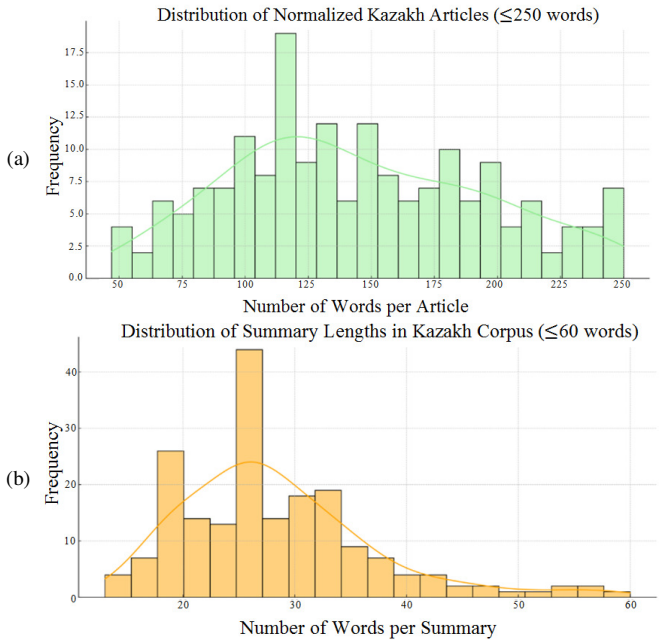

Fig. 3. Distributions of article (a) and summary (b) lengths.

TABLE I. DISTRIBUTION OF ARTICLES BY WEB RESOURCE AND CATEGORY

Web resource	Analytics	Interview	Sports	Culture & Entertainment	Health
baq.kz	276	143	97	126	152
informburo.kz	117	73	62	86	59
kaz.tengrinews.kz	98	84	85	77	61
kaz.nur.kz	95	72	102	101	34
Total	586	372	346	390	306

The dataset was stored in CSV format, and each record consisted of two fields: text for the full news article and summary for the corresponding human-written abstract. To facilitate training and evaluation, the dataset was randomly divided into three subsets in an 80/10/10 ratio: 1,600 samples for training, 200 for validation, and 200 for testing. The dataset is not publicly available due to institutional policy, but can be provided upon reasonable request.

B. Data Preprocessing

Several key preprocessing steps were performed to prepare the data for training. During the normalization phase, the text was converted to lowercase, extraneous spaces were removed, and the words were brought into their standard forms (e.g., converting to the lemma). This helped improve data quality and increased model efficiency. Non-informative characters, such as special symbols, unnecessary spaces, and other elements irrelevant to the summarization task, such as links and HTML tags, were removed. At the Character-level tokenization stage, an RNN-based model was used to analyze the text at the character level. FastText was used to obtain word vector representations. The FastText parameters included a vector size of 300, a context window of 5, and 10 epochs to train the model on the preprocessed data. POS tagging was performed using Stanza after tokenization and before passing the text to the model for feature extraction.

For text tokenization during preprocessing, various methods were used for different models. For mT5 and mBART, the SentencePiece tokenizer was used, which splits the text into subwords and helps efficiently handle rare or compound words. For XLM-RoBERTa, the standard XLM-RoBERTa tokenizer, based on BPE, was used. For the mT5 and mBART models, the vocabulary size was set to 32,000 tokens. For XLM-RoBERTa, the vocabulary size was set to 250,000 subwords. The mT5 model was configured with 12 layers and a hidden size of 768. The mBART model used 12 layers with a hidden size of 1024. Both XLM-RoBERTa and BART models utilized 12 layers with a hidden size of 1024. Training was conducted for 10 epochs with a batch size of 16.

All stages of implementation were carried out using the HuggingFace Transformers library, including the datasets, tokenizers, and Trainer modules, as well as gensim for FastText and Stanza for POS annotation. During fine-tuning, best practices for training transformers for text generation tasks were applied: the AdamW optimizer was used with a learning rate of 3e-5, regularization was achieved through Dropout and weight decay with a coefficient of 0.01, and warmup steps accounted for 10% of the total steps to stabilize gradients smoothly. Gradient accumulation was applied to increase the effective batch size, and the gradient norm was capped at 1.0 (gradient clipping).

A set of metrics was used to evaluate the performance of the summarization models, covering both surface-level matches and deep semantic correspondences between the generated and reference summaries.

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): This is a set of metrics that measure the similarity between n-grams, substrings, and word sequences in the generated text and the reference summary. ROUGE-1 measures unigram matches (individual words), ROUGE-2 measures bigram matches (word pairs), and ROUGE-L measures the longest common subsequence (LCS) matches. Each of the ROUGE metrics is calculated using the following formula:

$$ROUGE - N = \frac{\sum_{overlap\ of\ N-grams} matches}{\sum_{N-grams\ in\ reference}} \quad (1)$$

  where N is the size of the n-grams, and $matches$ refers to the number of matching fragments between the generated and reference summary.

- BLEU (Bilingual Evaluation Understudy) measures precision between the generated and reference text by comparing matches in n-grams up to the 4th order (2):

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} \omega_n \log\ p_n) \quad (2)$$

  where p_n is the proportion of matching n-grams, ω_n is the weight (usually equal), and BP is the penalty for overly short sentences.

- METEOR (Metric for Evaluation of Translation with Explicit ORdering) considers word matches, synonyms, stemming, and paraphrases. METEOR is considered a more flexible metric because it uses semantic similarity and penalizes word order.

$$METEOR = F_{mean} \cdot (1 - Penalty) \quad (3)$$

  where F_{mean} is the harmonic mean of precision and recall, and $Penalty$ depends on the number of word order swaps.

- BERTScore is a metric based on contextual embeddings. Instead of counting surface-level matches between tokens, BERTScore measures semantic similarity between word vector representations, extracted from a pretrained model. It is computed as the average maximum cosine similarity between tokens:

$$BERTScore\ F1 = \frac{1}{N} \sum_{i=1}^{N} \max_j\ cos(\mathrm{v}_i, \mathrm{v}_j) \quad (4)$$

  where N is the number of tokens in the generated text, v_i is the embedding of the i-th token in the generated text, v_j is the embedding of the j-th token in the reference text, and $cos(\mathrm{v}_i, \mathrm{v}_j)$ is the cosine similarity between the embeddings of the tokens. This metric is particularly effective for evaluating the quality of summaries when the generated text paraphrases the original without repeating it verbatim.

C. Experimental Results and Ablation Study

1) Ablation Study

To evaluate the contribution of each linguistic level in the multilevel architecture, an ablation study was conducted across all three models: mBART, mT5, and XLM-R+BART. In each configuration, one level (character, subword, word, or contextual) was removed, and the model was re-trained on the same dataset of 2,000 Kazakh articles (Table I). The results showed that removing the character-level input (Char Level) had only a minor impact on the overall performance metrics. This suggests that character-level features play a supporting role when more abstract representations, namely subword- and word-level embeddings, are present. In contrast, the exclusion of the subword level results in the most significant drop in performance across all models. Subword segmentation is essential for representing rare, compound, and morphologically rich words. Its absence considerably limits the model's ability to generate coherent and accurate summaries, particularly in low-resource settings. Therefore, the subword level can be considered a critical component of the proposed architecture.

Removing the word-level input leads to a moderate decline in performance. Static word embeddings, such as those from FastText, help reinforce lexical semantics and act as anchors in constructing meaningful representations. Without them, the model loses access to a broader lexical context and must rely solely on more granular features. Nevertheless, the impact on performance is less pronounced than with the subword level, likely due to partial compensation by other input modalities.

The most substantial degradation is observed when the contextual level (Contextual Level) is removed, which is obtained from large pretrained models such as XLM-R. This level captures inter-sentential dependencies and global textual coherence. Specifically, in the XLM-R+BART architecture, removing the contextual component resulted in the sharpest decline across all key metrics, highlighting its fundamental role in semantic representation [32].

TABLE II. METRIC RESULTS FOR FULL AND ABLATION VARIANTS

Model	Configuration	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR	BERTScore F1
mBART	Full Multilevel	0.60	0.43	0.52	0.49	0.62	92.40
	Char Level	0.59	0.42	0.51	0.48	0.61	91.90
	Subword Level	0.54	0.40	0.45	0.43	0.55	87.90
	Word Level	0.57	0.42	0.47	0.46	0.60	88.30
	Contextual Level	0.56	0.42	0.49	0.47	0.59	88.00
mT5	Full Multilevel	0.52	0.37	0.48	0.32	0.37	91.20
	Char Level	0.51	0.36	0.47	0.31	0.36	90.60
	Subword Level	0.47	0.32	0.41	0.27	0.34	83.80
	Word Level	0.48	0.34	0.43	0.29	0.35	84.20
	Contextual Level	0.49	0.34	0.44	0.29	0.35	84.00
XLM-R + BART	Full Multilevel	0.57	0.34	0.47	0.33	0.38	90.50
	Char Level	0.56	0.33	0.46	0.32	0.37	90.00
	Subword Level	0.53	0.31	0.43	0.30	0.35	85.50
	Word Level	0.55	0.32	0.45	0.31	0.36	86.70
	Contextual Level	0.52	0.30	0.42	0.28	0.34	84.00

Among the three models, mBART with all levels active outperformed all others, showing the highest scores on all metrics. This supports the hypothesis that multilevel modeling is especially effective in encoder-decoder architectures trained on multilingual corpora.

D. Comparison of Results

To compare the results, other studies on Kazakh text summarization were reviewed. Currently, the number of studies dedicated to the automatic summarization of Kazakh texts remains limited. Most existing approaches rely on extractive methods, such as TF-IDF and fuzzy logic. Research on hybrid summarization has emerged more recently, mainly involving transformer architectures and transfer learning strategies (Table III).

In [14], an extractive model based on fuzzy logic with preliminary pronoun resolution and morphological analysis was used, along with a corpus of 100 news articles. The results showed ROUGE-1 F1 at 0.44 and ROUGE-2 at 0.34, which is relatively high for a classical method without a generative architecture. In [15], a classical extractive approach based on TF-IDF was implemented, using a corpus of 1,422 news articles collected from the inform.kz news portal. The use of a specially compiled stopword list for Kazakh allowed the model to achieve ROUGE-L F1 at 0.35, despite the lack of abstractive generation. The model in [16] applied sequential sentence extraction using TF-IDF and their simplification through Seq2Seq, trained using transfer learning on the Simple English Wikipedia corpus. For the Kazakh language component, the Kaz-skaz corpus was employed. The final BLEU score was 7.0, although the ROUGE scores were not directly reported.

Based on the experiments conducted and the metrics obtained, several key observations can be made that highlight the performance differences between the mBART, mT5, and XLM-RoBERTa models (Figure 4). mBART demonstrated significant improvements in the ROUGE-1 metric, reaching 0.60. This indicates the high efficiency of the model in generating hybrid summaries, which is further supported by the steady increase during the training process.

TABLE III. METRIC RESULTS FOR DIFFERENT MODELS

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR	BERT-Score F1
[14] (Fuzzy Logic, with pronoun resolution)	0.44	0.34	0.4	-	-	-
[15] (Extractive TF-IDF)	0.4	0.31	0.35	-	-	-
[16] (TF-IDF + Seq2Seq via Transfer Learning)	-	-	-	7	-	-
mBART with Multilevel Architecture	0.60	0.43	0.52	0.49	0.62	92.40
mT5-small with Multilevel Architecture	0.52	0.37	0.48	0.32	0.37	91.20
XLM-R + BART with Multilevel Architecture	0.57	0.34	0.47	0.33	0.38	90.50

mT5 also showed a steady improvement in the ROUGE-1 and ROUGE-L metrics, reaching a peak of 0.52 for ROUGE-1 and 0.48 for ROUGE-L. Despite these gains, its performance remains lower than that of mBART, especially in the ROUGE-2 metric, where it reached 0.37. While these results are noteworthy, they still fall behind the performance of mBART and XLM-R+BART. This suggests that mT5 has a lower capacity to generate high-quality summaries compared to the other models. XLM-R+BART showed consistent improvement in the ROUGE-2 metric, with a peak value of 0.34, as well as in ROUGE-L, achieving 0.47. It should be noted that although XLM-R+BART significantly improved its results in the later epochs, it did not reach the same levels of ROUGE-1 and ROUGE-2 as the other models. However, its ability to handle context and generate more accurate summaries remains commendable.

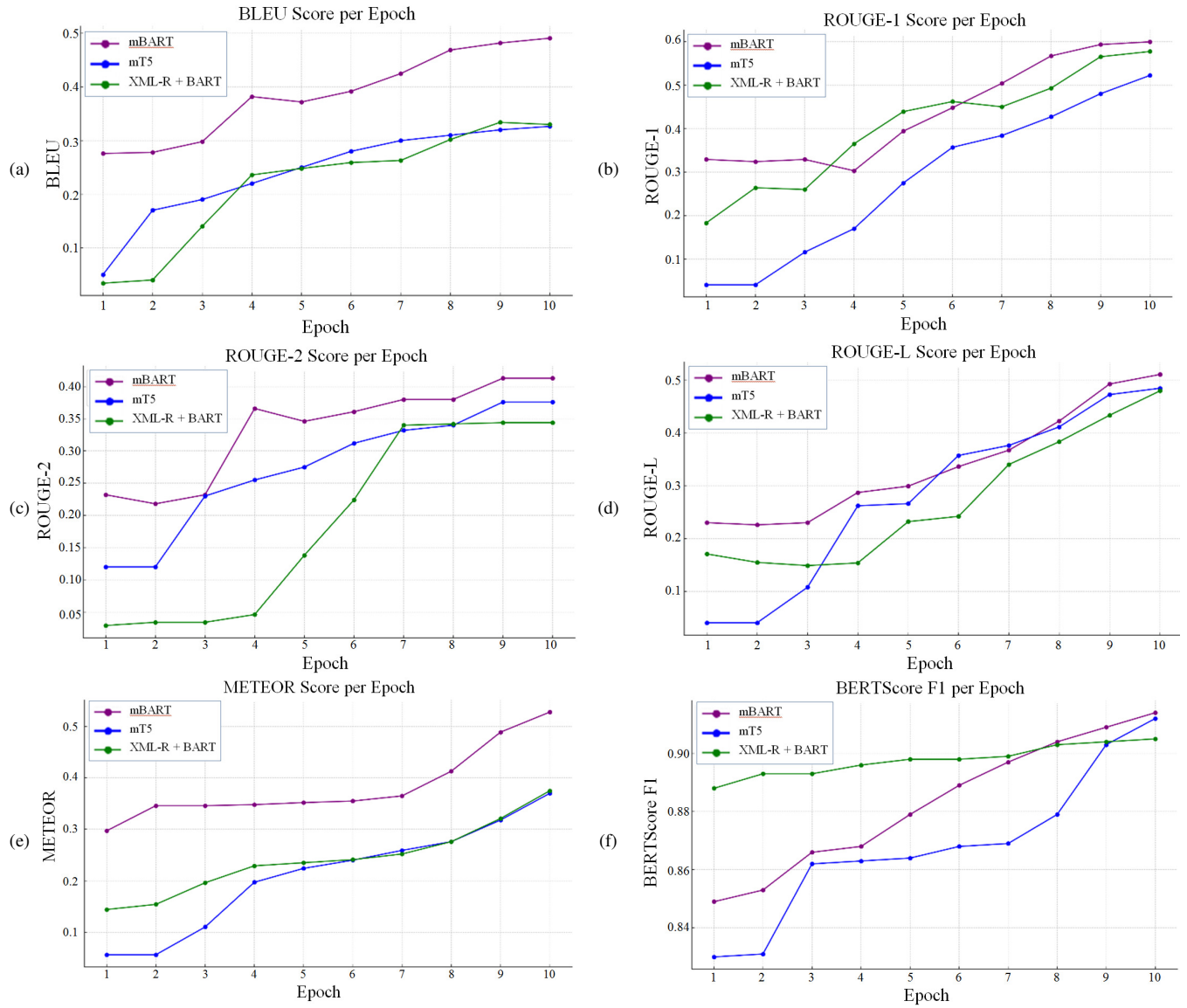

Fig. 4. Metric graphs by training epochs for mBART, mT5, and XLM-RoBERTa Models: (a) BLEU, (b) ROUGE-1, (c) ROUGE-2, (d) ROUGE-L, (e) METEOR, (f) BERTScore F1.

METEOR also highlights the clear advantage of mBART in generating hybrid summaries. mBART consistently showed high results in METEOR, reaching 0.62. This underscores its ability to consider not only word matches, but also synonyms, stems, and paraphrases. For mT5 and XLM-R+BART, the METEOR results were lower, but still showed improvements compared to previous studies: mT5 reached 0.37 and XLM-R+BART reached 0.38. This shows that these models are capable of generating texts with a high level of semantic similarity [33]. In the BERTScore-F1 metric, which measures the semantic similarity between the generated and reference texts, mBART also achieved the best results, consistently reaching high values of 92.40. This confirms its ability to generate not only accurate but also contextually meaningful summaries. This is crucial for hybrid summarization, where maintaining the meaning of the text is paramount.

All considered models (mBART, mT5, XLM-R+BART) significantly outperformed other studies [14-16], whose ROUGE results did not exceed 0.44 for ROUGE-1 [14] and 0.35 for ROUGE-L [15]. These results confirm that modern transformer models, such as mBART, mT5, and XLM-R+BART, have substantially improved performance.

In addition to automated evaluation metrics, a preliminary human assessment was conducted by a native Kazakh linguist. The generated summaries were subjectively reviewed for fluency, coherence, and semantic faithfulness to the source content. In most cases, the summaries were found to be grammatically correct, well-structured, and accurately reflective of the main ideas. Although a formal multi-annotator evaluation was not performed, this initial expert validation supports the quantitative results and confirms the applicability of the model to practical scenarios.

IV. CONCLUSION

This work introduced a novel multilevel summarization framework, the first of its kind applied to the Kazakh language, that integrates character, subword, word, and contextual-level features. This architecture was developed and evaluated using state-of-the-art transformer models, including mBART, mT5, and XLM-RoBERTa. The novel framework processes input data at four distinct levels, helping to better capture the morphological richness and semantic structure of the Kazakh language. This multilevel design proved particularly effective for handling agglutinative features and improving summary fluency and informativeness in a low-resource setting. The experimental results demonstrated that the mBART model, utilizing a multilevel architecture, achieved the best results in terms of the ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore-F1 metrics.

Future work may include further optimization of the fusion strategy, training on larger or domain-specific Kazakh corpora, and extending the model to handle other low-resource agglutinative languages. Additional research could also explore semi-supervised or data augmentation techniques to further improve generalization, as well as human evaluation studies to better assess summary coherence, factual accuracy, and readability across diverse genres. As part of future work, the authors plan a structured human evaluation with multiple annotators to systematically assess summary fluency, coherence, and factual consistency.

ACKNOWLEDGMENT

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19174298).

REFERENCES

[1] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[2] A. Rahali and M. A. Akhloufi, "End-to-End Transformer-Based Models in Textual-Based NLP," *AI*, vol. 4, no. 1, pp. 54–110, Jan. 2023, https://doi.org/10.3390/ai4010004.

[3] K. Rani Narejo, H. Zan, D. Oralbekova, K. Parkash Dharmani, M. Orken, and K. Mukhsina, "Enhancing Emoji-Based Sentiment Classification in Urdu Tweets: Fusion Strategies With Multilingual BERT and Emoji Embeddings," *IEEE Access*, vol. 12, pp. 126587–126600, 2024, https://doi.org/10.1109/access.2024.3446897.

[4] D. Oralbekova, O. Mamyrbayev, S. Zhumagulova, and N. Zhumazhan, "A Comparative Analysis of LSTM and BERT Models for Named Entity Recognition in Kazakh Language: A Multi-classification Approach," in *Modeling and Simulation of Social-Behavioral Phenomena in Creative Societies*, 2024, pp. 116–128, https://doi.org/10.1007/978-3-031-72260-8_10.

[5] N. Giarelis, C. Mastrokostas, and N. Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," *Applied Sciences*, vol. 13, no. 13, Jan. 2023, Art. no. 7620, https://doi.org/10.3390/app13137620.

[6] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing." arXiv, Aug. 28, 2021, https://doi.org/10.48550/arXiv.2108.05542.

[7] X. Liu *et al.*, "GPT understands, too," *AI Open*, vol. 5, pp. 208–215, Jan. 2024, https://doi.org/10.1016/j.aiopen.2023.08.012.

[8] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[9] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." arXiv, Oct. 29, 2019, https://doi.org/10.48550/arXiv.1910.13461.

[10] M. Kirmani, N. M. Hakak, M. Mohd, and M. Mohd, "Hybrid Text Summarization: A Survey," in *Soft Computing: Theories and Applications*, 2019, pp. 63–73, https://doi.org/10.1007/978-981-13-0589-4_7.

[11] L. Qin *et al.*, "A survey of multilingual large language models," *Patterns*, vol. 6, no. 1, Jan. 2025, https://doi.org/10.1016/j.patter.2024.101118.

[12] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale." arXiv, Apr. 08, 2020, https://doi.org/10.48550/arXiv.1911.02116.

[13] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv, Jul. 26, 2019, https://doi.org/10.48550/arXiv.1907.11692.

[14] A. Zulkhazhav *et al.*, "Kazakh Text Summarization using Fuzzy Logic," *Computación y Sistemas*, vol. 23, no. 3, pp. 851–859, Sep. 2019, https://doi.org/10.13053/cys-23-3-3239.

[15] B. Kynabay, A. Aldabergen, and A. Zhamanov, "Automatic Summarizing the News from Inform.kz by Using Natural Language Processing Tools," in *2021 IEEE International Conference on Smart Information Systems and Technologies (SIST)*, Nur-Sultan, Kazakhstan, Apr. 2021, pp. 1–4, https://doi.org/10.1109/sist50301.2021.9465885.

[16] T. Zhabayev and U. Tukeyev, "Development of Technology for Summarization of Kazakh Text," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 9, 2021, https://doi.org/10.14569/ijacsa.2021.0120914.

[17] E. Winarko, L. Tanoto, and M. H. Reza, "Indonesian Abstractive Text Summarization Using Stacked Embeddings and Transformer Decoder.," *IAENG International Journal of Computer Science*, vol. 52, no. 4, 2025.

[18] A. Raza, M. H. Soomro, Salahuddin, I. Shahzad, and S. Batool, "Abstractive Text Summarization for Urdu Language," *Journal of Computing & Biomedical Informatics*, vol. 7, no. 02, Sep. 2024.

[19] B. C. Challagundla and C. Peddavenkatagari, "Neural Sequence-to-Sequence Modeling with Attention by Leveraging Deep Learning Architectures for Enhanced Contextual Understanding in Abstractive Text Summarization." arXiv, Apr. 08, 2024, https://doi.org/10.48550/arXiv.2404.08685.

[20] N. Jayatilleke and R. Weerasinghe, "A Hybrid Architecture with Efficient Fine Tuning for Abstractive Patent Document Summarization," in *2025 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, Colombo, Sri Lanka, Apr. 2025, pp. 1–6, https://doi.org/10.1109/scse65633.2025.11030964.

[21] D. H. Dat, D. D. Anh, A. T. Luu, and W. Buntine, "Discrete Diffusion Language Model for Efficient Text Summarization." arXiv, Mar. 10, 2025, https://doi.org/10.48550/arXiv.2407.10998.

[22] Y. R. Gogireddy, A. N. Bandaru, and V. Sumanth, "Synergy of Graph-Based Sentence Selection and Transformer Fusion Techniques For Enhanced Text Summarization Performance," *Journal of Computer Engineering and Technology*, vol. 7, no. 1, pp. 33–41, Jun. 2024.

[23] B. Faizal, S. Abraham, and S. Thomas, "Automated Business Report Summarization Using Transformer Model," in *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, Mar. 2024, pp. 254–258, https://doi.org/10.1109/icaccs60874.2024.10716930.

[24] R. Chakraborti, R. Banerjee, and S. Das, "Evaluating the Efficacy of Text Summarization Models: A Comparison of NLP Algorithms," in *2025 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, Kolkata, India, Jan. 2025, pp. 1–5, https://doi.org/10.1109/iementech65115.2025.10959463.

[25] A. B. Rao, S. G. Aithal, and S. Singh, "An Evaluation Metric for Assessing Summary-Level Semantic Similarity in Abstractive Text Summarization," in *2025 International Conference on Artificial*

Intelligence and Data Engineering (AIDE), Nitte, India, Feb. 2025, pp. 602–607, https://doi.org/10.1109/aide64228.2025.10987460.

[26] S. Zangooei, A. Darmani, H. F. Nezhad, and L. Mahmoudi, "ARLED: Leveraging LED-Based ARMAN Model for Abstractive Summarization of Persian Long Documents," in *2025 11th International Conference on Web Research (ICWR)*, Tehran, Iran, Apr. 2025, pp. 25–32, https://doi.org/10.1109/ICWR65219.2025.11006228.

[27] O. Langston and B. Ashford, "Automated Summarization of Multiple Document Abstracts and Contents Using Large Language Models." TechRxiv, https://doi.org/10.36227/techrxiv.172262754.45577350/v1.

[28] B. Barta, D. Lakatos, A. Nagy, M. K. Nyist, and J. Ács, "From News to Summaries: Building a Hungarian Corpus for Extractive and Abstractive Summarization," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 7503–7509.

[29] S. Jebbara and P. Cimiano, "Improving Opinion-Target Extraction with Character-Level Word Embeddings," in *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, 2017, pp. 159–167.

[30] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018, pp. 66–71.

[31] R. A. Stein, P. A. Jaques, and J. F. Valiati, "An analysis of hierarchical text classification using word embeddings," *Information Sciences*, vol. 471, pp. 216–232, Jan. 2019, https://doi.org/10.1016/j.ins.2018.09.001.

[32] G. Aguilar, B. McCann, T. Niu, N. Rajani, N. S. Keskar, and T. Solorio, "Char2Subword: Extending the Subword Embedding Space Using Robust Character Compositionality," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 1640–1651.

[33] S. R. Basha, J. K. Rani, and J. J. C. P. Yadav, "A Novel Summarization-based Approach for Feature Reduction Enhancing Text Classification Accuracy," *Engineering, Technology & Applied Science Research*, vol. 9, no. 6, pp. 5001–5005, Dec. 2019, https://doi.org/10.48084/etasr.3173.