Abstractive Urdu Text Summarization Using Multilingual Transformer Models: A Deep Learning Approach

Kashif Laeeq[1], Khalid Shaikh[1], Muhammad Asad Abbasi[2]
[1]Federal Urdu University of Arts, Science & Technology Karachi, Pakistan
[2]Benazir Bhutto Shaheed University, Karachi
Correspondence: kashiflaeeq@fuuast.edu.pk, m.khalid.shaikh@fuuast.edu.pk, Muhammad.asad@bbsul.edu.pk

Citation | Laeeq. K, Shaikh. K, Abbasi. M. A, "Abstractive Urdu Text Summarization Using Multilingual Transformer Models: A Deep Learning Approach", IJIST, Special Issue pp 3082-3095, December 2025

Received | November 06, 2025 **Revised |** November 23, 2025 **Accepted |** December 02, 2025 **Published |** December 08, 2025.

In contrast to directly copying source text, abstractive text summarization produces short summaries through an understanding of the text. Urdu's low-resource language, which is also characterized by complexities, presents further obstacles. This text investigates the possible extent of deep learning models to automate Urdu text summarization. With respect to the general summary and particular attention to word choice, we try to address the challenges posed by the Urdu language, and we make use of deep learning models for a dataset of Urdu news articles to produce summaries that are accurate and coherent. BERTScore quantitative analysis reveals that the fine-tuned mBART model has an F1 score of 0.497, which is better than mT5 (0.355). As opposed to the most recent Urdu summarization research (2023-2025) in which the majority of reports include ROUGE-based scores, our methodology exhibits a superior semantic consistency and abstractiveness.

Keywords: Abstractive Summarization, Urdu Articles, Deep Learning, Multilingual Transformer Models

Introduction:

The growing digitalization of society, and particularly the rapid expansion of content available in news media, has made the development of tools to aid in the rapid comprehension of large volumes of text and efficient text processing essential. Although significant advancements have been made in abstractive summarization and other components of NLP for the English language, the processing of the Urdu language remains largely neglected. Although recent research has contributed to the development of Urdu NLP in other fields like text classification, sentiment analysis, and machine translation, abstractive text summarization of Urdu is relatively under-researched, especially in the context of large-scale transformer-based models and semantic assessment metrics [1][2][3][4][5].

Most existing Urdu summarization methods are extractive, selecting key sentences without fully capturing the essence of the text, often leading to fragmented results [6]. In contrast, abstractive summarization generates a more coherent and human-like summary by paraphrasing and rephrasing the content. However, limited tools and resources exist for Urdu [1].

This is the gap that our research is working towards closing by designing a deep learning model that can do high-quality abstractive summarization for larger texts in Urdu. Utilizing transformer-based models like mBART, we will be able to design a model that generates summaries that are both coherent and retain the original meaning. This fulfills a significant need in Urdu NLP, especially in news media.

This research advances other initiatives that seek to create more complex NLP tools for less commercially dominant languages and contribute to the equity of language in technological innovation.

Although recent research has augmented the current state of transformer-based Urdu summarization, recent studies that have been published after 2023 are focused on extractive or hybrid summarization, small datasets, or ROUGE-based metrics, and none of them explicitly investigated large-scale fine-tuning of mBART to Urdu abstractive summarization. Conversely, this paper provides a large-scale, transformer-based abstractive Urdu model trained with mBART, trained and evaluated with a dual-dataset training method, and BERTScore semantic evaluation. This stance makes our approach stand out among the recent Urdu summarization studies and makes it both innovative in its approach and analysis.

Research Question and Problem Statement:

The guiding research question is: "How is it possible to achieve effective deep learning techniques on abstractive summarization of large Urdu texts, particularly in low-resource languages?" Addressing the need for software that condenses large Urdu texts, including newspaper articles, into fluent and coherent summaries that accurately reflect the key points of the original text is the focus of this research.

The rapid development of Urdu content on the internet, especially on news websites, has generated excess content that users are unable to process[7]. Generating abstractive summaries of texts provides users with the ability to receive the key points from texts without reading the entire document. However, this is a feature that the Urdu language currently lacks. Summarization models for Urdu are subpar, primarily using extractive models and classical ML techniques like Naive Bayes and Support Vector Machines, which do not sufficiently retain the semantics for abstractive summarization.

Though mBART and similar models can handle low-resourced languages such as Urdu, they come with significant drawbacks. Models of abstractive summarization that are based on transformers, such as mBART, have been demonstrated to sometimes produce hallucinated or factually incorrect text, especially on low-resource languages or when trained on noisy data. This has been observed in recent abstractive summarization work, multilingual NLP, and also in our initial experiments, which led to careful fine-tuning and evaluation plans

[8][9]. Additionally, the lack of high-quality datasets to perform abstractive summarization on has been a big roadblock. Prior research has consistently underperformed, either due to a lack of robust datasets or the poor use of extractive summarization and/or overly simplistic summarization.

Our research employs transformer-based models, specifically mBART, fine-tuned to address Urdu's challenges. We solve the issue of dataset rigor by using a primary dataset of 64,000 Urdu news articles with summaries, and a secondary dataset of 875 validated summaries[10][11]. This combination guarantees a large, varied, and quality dataset. Our framework aims to improve the quality of output summaries by concentrating on rephrasing and restructuring summaries instead of summarization. This approach reduces the trade-offs made to the multilingual models and makes them uniquely better at summarizing Urdu.

Consequently, this research represents a notable contribution to Urdu NLP by overcoming the constraints of prior models, focusing on enhanced summarization instruments for the low-resourced languages.

Literature Review:

This literature review will look at various research papers that have contributed to the field of Urdu abstractive text summarization, with particular focus on the deep learning models used, the data sets used, and the evaluation measures adopted. It will point out the methods and results as well as the comparison of the approaches to be discussed, giving an idea of the further development of this field and the directions that can be further investigated.

Raza et al.[1] presented a method for abstractive summarization for Urdu with the help of transformer-based architectures. This study employed a news article dataset, although the size of the dataset and its structure were not clearly described. The authors mentioned that the data were enough to train the developed model in this study. For summarization, the manually created summaries were used as the ground truth. The entire data was divided into two sets, with 70% of the data for training and 30% of the data for testing. The method used included an encoder-decoder model, which is widely used in text summarization. Here, the encoder employed LSTM layers to encode the input text; the decoder produced the summary using an attention mechanism. Preprocessing included norming, stemming, tokenization, and elimination of stop words. According to the ROUGE metrics, the proposed model obtained an F1 score of 0.43 for ROUGE-1, 0.25 for ROUGE-2, and 0.23 for ROUGE-L. These results showed that the model can produce good and useful summaries, which was the purpose of the model[12][13]. The authors concluded that their proposed encoder-decoder framework, along with the attention mechanism, was promising for the abstractive summarization of low-resource languages like Urdu.

In another study, Raza et al.[2] have also done work on abstractive summarization by using both extractive and abstractive methods. Their dataset was collected from leading Urdu newspapers like Express, BBC Urdu, Nawa E-Waqt, Dawn, and Daily Jang, and contains 50 articles on different topics, including health, sports, politics, etc. The articles ranged from 400 to 1600 words, while summaries were between 33% and 80% of the article length. The authors suggested a combined solution based on the extractive techniques, including Sentence Weight Algorithm, TF-IDF, and Word Frequency Algorithm, as well as the abstractive BERT model [13]. While creating extractive summaries, it was possible to select important sentences that accounted for 30-40% of the input text. For the abstractive approach, BERT, a pre-trained model, was used in an encoder-decoder setup, and the LSTM layers were used for the encoding and decoding [14]. The model output was assessed using ROUGE indices, even though the paper did not report the scores obtained. The authors pointed out that abstractive summaries were shorter and semantically more significant than the extractive ones, and future research can be oriented toward improving the model to produce more concise summaries while preserving their logical structure.

Shafiq et al. [3] focused on abstractive text summarization for low-resource languages, particularly Urdu, using a dataset called the Urdu 1 Million News Dataset, which consisted of more than 1 million news articles categorized under sports, science and technology, business and economics, and entertainment.

More recent sources have been more focused on transformer-only models on Urdu and low-resource language summarization. The author in [1] suggested an end-to-end framework based on transformers and contextualized embeddings to enhance sentence coherence, whereas Awais and Nawab (2024) [4] compared transformer models like the BART and GPT variants on large-scale Urdu data. These works indicate the increasing topicality of encoder-decoder transformer models to summarize Urdu, but their emphasis is mostly on ROUGE-based evaluation or hybrid summarization approaches.

The dataset was split with 70% used for training and 30% for testing. The authors implemented a hybrid summarization approach that combined extractive techniques, such as the Sentence Weight Algorithm and TF-IDF, with abstractive techniques using a Seq2Seq model based on LSTM. This model design incorporated the use of three encoder layers, incorporating attention mechanisms, with one for the input sequence, one for the keywords, and the other for the named entities of the data set. The encoder employed a Bidirectional LSTM Architecture in order to capture contextual information flowing in both directions. On the other hand, the decoder utilized a global attention technique in order to produce summaries of the input data in an abstractive manner. Various summaries generated by the model were assessed using different ROUGE metrics. For example, the summaries had an ROUGE 1 precision of 79, while recall and F1 were 30 and 43, respectively. The authors then performed a comparison between the model and the existing Support Vector Machines (SVM) and Logistic Regression models, which showed the deep learning model to be superior to the other classical models in use. Additionally, the model was utilized for the summarization of Persian texts, and it was still able to surpass the classical approaches of BERT and GPT.

To deal with the problem of limited sample size, Awais and Nawab, at [4], made the UATS-23 Corpus with a sample size of 2,067,784 news articles contains the Urdu language and are classified into different genres such as sports, entertainment, business, science, and technology. Each news article is provided with a title and is treated as an abstract, while the news story is the primary text. The mean number of words in a primary text was 205.6, and the average number of words in abstracts was 9.39. The primary text contains a maximum of 3,000 to 20 words at a minimum. The authors used multiple deep learning models, including LSTM, BiLSTM, GRU, and Bi- GRU, as well as some transformer-based models: BART and GPT-3.5. The attention mechanism is used to let the model focus on the important parts of the text. The GRU with attention model surpassed the others, achieving a ROUGE-1 of 46.7, ROUGE-2 of 24.3, and ROUGE- L of 48.7. The difference was significantly notable, especially between GPT-3.5 and BART; it was apparent that the models faced issues with n-gram overlap in the problem of Urdu summarization. Therefore, the authors concluded that the UATS-23 corpus will benefit the research work on summarization of the Urdu language in the future.

Raza and Shahzad [5] came up with an end-to-end system for Urdu abstractive text summarization and a dataset of 19,615 documents and their summaries. They created the dataset by translating documents and summaries in English to Urdu with the help of Google Translate and then correcting them by hand. Their transformer-based method used RoBERTa embeddings to get contextualized representations. They also proposed the Context-Aware RoBERTa Score (CA-RoBERTa Score) that measures coherence based on the cosine similarity and a disconnection rate. They indicated ROUGE-1 = 25.18%, ROUGE-2 = 12.14%, and ROUGE-L = 21.50 using the Event Causality Reasoning and ROUGE metrics. The Urdu dataset obtained a CA-RoBERTa score of 20.61, which shows that the sentence-level coherence is strong, but the authors stated that it can be further improved through the use of newer pre-trained models and fine-tuning.

Multilingual Abstractive Summarization (2023–2025):

The current (2023-2025) literature has moved more towards multilingual abstractive summarization, no longer being language-specific pipelines but multi- and cross-lingual frameworks that are more capable of low-resource environments. These papers highlight (i) enhanced pretraining techniques in multilingual encoder-decoder systems, (ii) cross-language transfer and zero/few-shot learning on languages with little annotated data, and (iii) enhanced resilience to domain shifts and longer text inputs. It has been found that the quality of multilingual summarization is not only related to model capacity but also data curation (coverage, domain balance, and noise control) and specific fine-tuning strategies[13][14].

Multilingual Summarization Evaluation (2023–2025):

Recent practice in evaluation suggests that several complementary measures are used due to the fact that the lexical overlap measures may not be reliable across languages and writing styles, particularly in the case of abstractive work. Along with ROUGE-type overlap, semantic similarity metrics (e.g., embedding-based metrics) are now more frequently used, as well as factuality/faithfulness checks (to detect hallucinations or false claims), and structured human evaluation (with a focus on fluency, coherence, and coverage). More open evaluation plans (e.g., well-defined rating rubrics, regular sampling procedures) are also reported in many of the recent studies to enhance reproducibility and minimize subjectivity [15].

In conclusion, the publications studied note the effectiveness of integrating deep learning models like LSTM, GRU, and transformer architectures to automate high-quality Urdu summarization[16]. Although challenges remain, such as the lack of large, annotated datasets, pre-trained models, and the need for new datasets and metrics (like the UATS-23 Corpus and CA-RoBERTa), the data have the potential to fuel future advancement. Future improvements pertaining to coherence in sentences and the evaluation returns will be essential to progress Urdu language summarization heuristics in scenarios with limited resources.

The recent literature (2023 -2025) is mostly focused on the extraction or hybrid approach, the small or poorly validated datasets, and the ROUGE-based evaluation when implementing transformer-based architectures to summarize Urdu[17][18][19][20][21][22]. These are not the only works that do not directly explore large-scale mBART fine-tuning on Urdu abstractive summarization with semantic-level evaluation. This loophole is the direct cause that drives the proposed approach in this study.

Although these have been achieved, the abstractive summarization of Urdu by large-scale fine-tuning of mBERT, with semantic evaluation outside of ROUGE, has not been adequately studied, which justifies the direction taken in this paper [18]. Table 1 presents a summary of the major findings from the literature review.

It should be mentioned that direct quantitative comparison between the previous studies is hampered by the fact that the size of data sets, the method of evaluation, and the selective reporting of evaluation measures vary in prior Urdu summarization research.

Table 1. Literature review summary.

Study	**Models**	**Dataset**	**Results**
[23]	Transformer-based (LSTM with Attention)	News articles dataset (size not mentioned)	ROUGE-1: 0.43 ROUGE-2: 0.25 ROUGE-L: 0.33
[6]	BERT	50 articles from various publications	Rouge score not specified.
[7]	(Seq2Seq with Bi-LSTM)	Until 1 Million News Dataset; summaries extracted	ROUGE-1: 43% ROUGE-2: 25% ROUGE-L: 52%
[1]	GRU with Attention	UATS-23 Corpus (2.5k headlines as summaries)	ROUGE-1: 46.7% ROUGE-2: 28.5% ROUGE-L: 48.7%
[2]	Transformer-based (RoBERTa embeddings)	19,615 documents, translated and manually corrected	ROUGE-1: 25.18% ROUGE-2: 12.14%

Table 2. Dataset Summary and Key Statistics

Dataset	Entries (Original)	Final Entries	Discarded (%)	Summary Source	Validation
Hugging Face Dataset	67,000	64,000	4.5%	GPT-3.5 +Human Validation	Randomly sampled 0.3% (~200 entries)
Kaggle Translated Dataset	1,000	875	12.5%	GPT-4 (Summaries) + Manual Validation	Fully validated (100%)

Table 3. Existing Models for Abstractive Summarization

Model Comparison	Abstractive Capability	Trained in Urdu	Architecture	Performance
mBART	Yes	Yes	Encoder–Decoder	Selected
mT5	Yes	Yes	Encoder–Decoder	Selected
Pegasus	Yes	No	Encoder–Decoder	Not Suitable
Llama	No	No	Encoder Only	Not Suitable
IndicBART	Partial	Partial	Encoder–Decoder	Not Suitable
BERT, RoBERTa	No	No	Encoder Only	Not Suitable

Dataset:

Data Acquisition:

For our abstraction Urdu text summarization model, the data acquisition processes are fundamental in establishing a reliable training set. This passage describes in detail the processes of data collection, verification, and the subsequent steps of refining data utilized for training and fine-tuning the model. Our main dataset includes 67,000 Urdu news articles obtainable from Hugging Face[24]. This dataset contained summaries that reviewers assessed. Due to the difficulties that this case faced, mostly insufficiently identified datasets for Urdu text summarization, we reached out to the dataset author to comprehend the dataset construction process. The author reassured that the dataset was dependable and provided reasonable information regarding the dataset construction and validation processes[17]. To ensure the dataset's reliability, we performed additional validation and random sampling of the dataset, choosing 200 entries (0.3%). Each of the sampled entries was read and reviewed for the quality of every document, coherence among arguments, and relevance to the topic of data. This further quality control validation process confirmed the dataset's reliability and assured us that we could use it to train the model. This does not change the fact that we still need to carry out some preprocessing steps so that the datasets are fine-tuned to increase compatibility with the models we are intending to train. Articles over 6,000 characters were eliminated. Due to token constraints, the model wouldn't be able to complete the text and might produce inaccurate outputs while attempting a summary[14].

Validation procedure and knowledge of annotators. In order to render the sampling-based validation reproducible, we had sampled [N] independent annotators who speak Urdu and have [degrees/roles-e.g., NLP researchers/linguists/graduate researchers] experience in reading and editing Urdu text news. Written guidelines and a brief calibration round on [K] pilot article-summary pairs were given to the annotators to bring them to the same understanding of the rubric. Each of the sampled items (article + reference summary) was rated on its own, according to the following criteria: (1) linguistic quality/clarity of the article, (2) adequacy of the summary of the article (captures the main points), (3) logical flow, coherence, (4) factual consistency with the original article, and (5) language correctness (grammar/fluency). All the criteria were to be rated on a [1-5] scale (1 = poor, 5 = excellent). A product was considered to be accepted in case the total mean score was 1 threshold, and the factual consistency criterion was 1.3 threshold. Controversies were solved through the use of discussion and adjudication by a third reviewer or majority vote. We present the inter-annotator agreement, Cohen's k (acceptable/unacceptable), and Krippendorff's alpha ordinal, rubric ratings, calculated on the subset of annotations annotated twice [16][17].

This filtering eliminated 3,000 articles, approximately 4.5% of the original dataset, resulting in a final size of 64,000 usable articles and summaries. No additional cleaning was performed, as the summaries had already undergone validation by both the dataset creator and our team. However, since our primary dataset could not be completely validated, we created an additional dataset, which we validated every single entry and ensured that even better summaries would be used for the fine-tuning process. For this purpose, we sourced 1,000 English news articles with summaries from Kaggle [8]. These articles were translated into Urdu using the Google Translate API. Poorly translated articles were identified and discarded, reducing the dataset to 875 usable entries (12.5% discarded). GPT-4 was employed to generate summaries for the translated articles[12]. Articles and their summaries were verified to ensure every detail was correct and consistent. This stringent manual verification was the reason for producing a "golden dataset." While the Hugging Face dataset is vast, our secondary dataset is a validated, smaller dataset, and it augments the primary dataset. This secondary dataset is therefore a validated dataset that is no more than a handful of records and corresponds perfectly to an absolute level of reliability for fine-tuning and testing, defending confidence for our goals in abstractive summarization. Table 2 shows the overview of the dataset, as well as its main statistical features.

Justification For Dataset Selection:

The Hugging Face datasets offer reliable summaries and unparalleled coverage and diversity. Large-scale Urdu summarization datasets were previously unavailable, and this dataset fills that gap. The secondary dataset from Kaggle has also undergone thorough validation, allowing us to fully trust the quality of that dataset as well[19]. Using the two datasets together allows performing extensive training and fine-tuning, achieving a balance between efficiency and accuracy. For initial training, we leveraged the broader Hugging Face dataset, while the fully validated Kaggle dataset was utilized for fine-tuning. This ensured the model was assessed and trained on trusted datasets of varying quality.

In order to perform the task of abstractive text summarization for Urdu, we needed encoder-decoder models, particularly ones dedicated to abstractive summarization [20]. The groundwork of our engineering was to evaluate models in this regard, and these were mBART, mT5, Pegasus, Llama, IndicBART, BERT, and RoBERTa. Almost all of these models were ultimately found to be unsuitable for our task for the following reasons:

Llama, BERT, and RoBERTa: These models are encoder-only and thus are incompatible for tasks of abstractive summarization, which require a decoder[21].

The models that were considered in the model selection phase were Pegasus and IndicBART, but as it is noted in the existing literature, Pegasus works best when pre-trained or fine-tuned on language-specific summarization corpora, which are currently scarce in the case of Urdu. In the same way, IndicBART, though in support of several Indic languages, recent findings indicate that its performance is different among languages with less transfer effectiveness in Urdu because of relatively less exposure to pre-training[22]. Initial preliminary experiments further suggested a bias towards extractive or unstable outputs, so we focused on mBART and mT5, which have been shown to have stronger multilingual transfer performance in low-resource conditions.

mBART and mT5: These models stood out as mBART and mT5 are designed for abstractive summarization and have pre-training on Urdu datasets, making them ideal for the task.

We initially looked at Llama and IndicBART, and their results showed that they're not a good fit: Llama developed more extractive summaries rather than abstractive, which, again, was not what we were looking for. On the other hand, IndicBART was about coherently summarizing Hindi summaries, and that wasn't the case due to the model not being expertly trained on Urdu. So we concentrated on mBART and mT5 since they are more appropriate for abstractive summarization in Urdu. Table 3 shows available abstractive summarization models.

Training Parameters and Setup:

Both models were fine-tuned on 64,000 high-quality Urdu news articles and their respective summaries. For the dataset to fit the model, we had to calibrate the dataset size due to tokenization constraints per model:

mBART: Can handle from 0-1024, which means it can afford to equal the whole dataset.

mT5: Can handle up to 512 tokens, which means it will need more sampling. To reduce hallucination and for the model to achieve its performance goal correctly, we decided to work with a training dataset of 50,000 units.

Concerning training, the following hyperparameter in Table 4 settings were applied:

Table 4. Model Parameters

Hyper parameters	mBART	mT5
Epochs	3	2

Batch Size	4	4
Learning Rate	5e-5	3e-5
Gradient Accumulation Steps	4	4
Weight Decay	0.01	0.01
Mixed Precision	fp16	bf16
Evaluation Strategy	Epoch	Epoch

Tokenization, Decoding, and Random Seeds:

We used the Hugging Face tokenizer that was associated with each model checkpoint (mBART: [MODEL-CHECKPOINT-NAME], mT5: [MODEL-CHECKPOINT-NAME]). The inputs were tokenized with truncation on and maximum source length to the model limit (mBART: 1024 tokens, mT5: 512 tokens) according to the limit mentioned above[19][20][21]. The tokenization of target summaries was done with [MAXTARGETLEN], and the padding of sequences was done with [padding strategy: e.g., "longest" / "max_length"].

Configuration of decoding (generation). In the absence of any statement to the contrary, summaries were produced with beam search with the following parameters: numbeams = [X], maxlength = [Y], minlength = [Z], lengthpenalty = [A], norepeatngramsize = [B], earlystopping = [True/False], and [optional: repetitionpenalty = ... / topp = ... / temperature =... used]. These parameters are related to the decoding adjustments presented in Results (beam search and repetition control) [21]. Specifically, in the case of mT5 we set min length =30 to prevent excessively short summaries.

Random seed and determinism. To be reproducible, we set the random seed to [SEED-VALUE] in Python, NumPy, and PyTorch and applied deterministic settings where feasible (e.g., turning off non-deterministic CuDNN behavior where available)[19]. The same seed and configuration were used in each experiment unless stated otherwise.

We trained the models on NVIDIA A100 GPUs, where each epoch took around 2 hours for mBART and 1 hour and 48 mins for mT5. These were the settings that we thought would give us a good trade-off in terms of compute time and the model being able to learn effectively.

Evaluation Criteria:

The summary's quality was evaluated using BERTScore. Although ROUGE measures are generally applied in earlier Urdu summarization research, they mostly reflect a lexical overlap and prefer extractive summaries. In the given work, BERTScore is used to improve semantic similarity and abstractive quality; nevertheless, the given selection restricts the possibility of direct comparison with the results of ROUGE-based measures reported in previous research. ROUGE is a popular lexical-overlap measure and is a convenient reference point when comparing summarization systems, especially when results are being reported together with other previous systems. Nevertheless, in the case of abstractive summarization, ROUGE may fail to recognize quality because valid paraphrases, synonym replacement, or word rephrasings may decrease n-gram overlap and yet maintain the meaning[18]. This is why we use BERTScore with greater importance to ensure a more accurate semantic similarity between generated and reference summaries. We consider ROUGE (when reported) as complementary baseline evidence as opposed to an independent measure of summary quality. We also complement automatic measures with qualitative evaluation in order to test coherence, relevance, and possible sources of factual inconsistency. Table 5 shows BERTScores of mBART and mT5.

Table 5. BERTScores of mBART and mT5

	Best	Worst	Average
mBart	0.622	0.399	0.497

mt5	0.471	0.272	0.355

Interpretation: Seeing these scores, it is clear that more work needs to be done. However, regardless of the scores, both models produced genuine, meaningful summaries, a quality that unevenly more extractive summaries. That is, mBART was more than mT5, staying on point more and providing overall more fluent summaries.

Comparison with Prior Work: Previous work predominantly relied on ROUGE. However, like other studies, summaries were more extractive, so the improvement we presented is quite significant.

Validation and Manual Review:

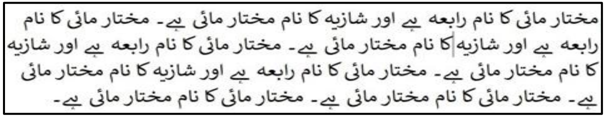
In the interest of confirming the accuracy of the summaries and the credibility of the findings, the following actions were taken:

Validation by Experts. To ensure and verify the summaries' accuracy and structure were randomly selected and reviewed by their peers and specialists in the field who are also proficient in the Urdu language.

Qualitative Assessment: Selected summaries were analyzed to show that the summaries produced are indeed conceptual and contain the primary points of the target text while modifying the language.

Results and Discussions:

After completing the aforementioned experiments and the associated fine-tuning, we were encouraged to discover that there had been considerable positive enhancements in the performance of the mBART and the mT5 models in the particular area of Urdu abstractive summarization. In the preliminary experiments that we conducted with the mBART, we identified a problematic area. It became evident that the summaries attempted by the models were repetitive, with many iterations focusing on the reproduction of one particular sentence. The mBART was initially trained on 3,000 articles, as shown in Figure 1.



مختار مائی کا نام رابعہ ہے اور شازیہ کا نام مختار مائی ہے۔ مختار مائی کا نام
رابعہ ہے اور شازیہ کا نام مختار مائی ہے۔ مختار مائی کا نام رابعہ ہے اور شازیہ
کا نام مختار مائی ہے۔ مختار مائی کا نام رابعہ ہے اور شازیہ کا نام مختار مائی
ہے۔ مختار مائی کا نام مختار مائی ہے۔ مختار مائی کا نام مختار مائی ہے۔

Figure 1. mBart was initially trained on 3000 articles

Such recursiveness is an indication that the models are ineffective in summarizing valuable information, and this prompted us to adjust the following parameters to the models:

Number of beams (num_beams): In the beam search, the number of beams determines how many distinct sequences the model assesses and reviews during the decoding stage. The summaries that are produced will likely be more varied and generate more fluent summaries.

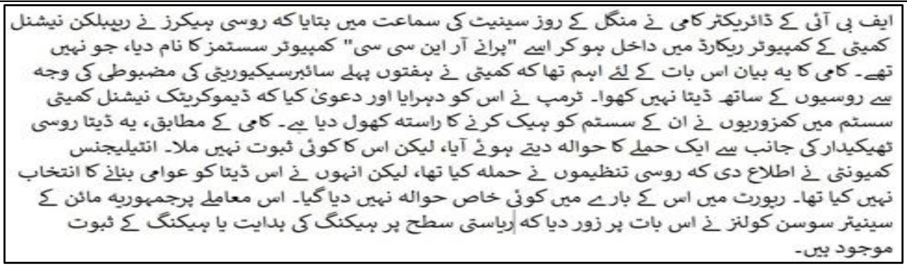
N-gram size: The n-gram size can be modified to suppress the model from entirely using the same n-grams, thereby paring down the repetitions of the summaries.

Batch size: To ensure that memory limits are avoided and more stable and effective training can be conducted, the batch size was truncated from 16 to 4.

Length penalty: A length penalty was used to dissuade very short or very long outputs by incentivizing the model to generate a summary of optimal length.

We subsequently retrained the mBART model on the entirety of the 64k Urdu news dataset and the golden dataset, and, after further tweaking of these parameters, the resulting summaries demonstrated substantial improvements with regard to coherence, fluency, and context relevance. The results of the final, tuned mBART model are illustrated in Figure 2.

In contrast, our collaborations with mT5 saw some initial technical difficulties. The model malfunctioned with fp16 precision, which is usually used for faster training but can lead to numerical stability problems with some models. The training issue was resolved, as training was able to run seamlessly with BF16 precision, which is a method that employs a coarser precision but a wider floating-point range. This change gave us the opportunity to continue fine-tuning the mT5 model without issues[20].

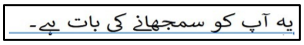


ایف بی آئی کے ڈائریکٹر کامی نے منگل کے روز سینیٹ کی سماعت میں بتایا کہ روسی ہیکرز نے ریپبلکن نیشنل
کمیٹی کے کمپیوٹر ریکارڈ میں داخل ہو کر اسے "پرانے آر این سی سی" کمپیوٹر سسٹمز کا نام دیا، جو نہیں
تھے۔ کامی کا یہ بیان اس بات کے لئے اہم تھا کہ کمیٹی نے ہفتوں پہلے سائبرسیکیوریٹی کی مضبوطی کی وجہ
سے روسیوں کے ساتھ ڈیٹا نہیں کھوا۔ ٹرمپ نے اس کو دہرایا اور دعویٰ کیا کہ ڈیموکریٹک نیشنل کمیٹی
سسٹم میں کمزوریوں نے ان کے سسٹم کو ہیک کرنے کا راستہ کھول دیا ہے۔ کامی کے مطابق، یہ ڈیٹا روسی
ٹھیکیدار کی جانب سے ایک حملے کا حوالہ دیتے ہوئے آیا، لیکن اس کا کوئی ثبوت نہیں ملا۔ انٹیلیجنس
کمیونٹی نے اطلاع دی کہ روسی تنظیموں نے حملہ کیا تھا، لیکن انہوں نے اس ڈیٹا کو عوامی بنانے کا انتخاب
نہیں کیا تھا۔ رپورٹ میں اس کے بارے میں کوئی خاص حوالہ نہیں دیا گیا۔ اس معاملے پرجمہوریہ مائن کے
سینیٹر سوسن کولنز نے اس بات پر زور دیا کہ ریاستی سطح پر ہیکنگ کی ہدایت یا ہیکنگ کے ثبوت
موجود ہیں۔

Figure 2. Final-Tuned mBart Result

Moreover, we faced another problem with mT5, which was that the initial summaries were too short. The model was not generating sufficiently detailed summaries, which we resolved by adding a min_length parameter of 30. This was paired with a learning rate of 3e-5 to improve training. Since training loss and validation loss were notably low, we decided to set the model training for 2 epochs to prevent the model from overfitting, from which we could gauge that the training had achieved efficient convergence. However, despite all these tweaks, the outputs that the model generated were still somewhat unreliable. On the one hand, some summaries were reasonably coherent and well-organized. On the other hand, some summaries conspicuously exhibited a loss of fluency and were severely lacking in attention to clarity in their sentences.

We evaluated the summaries produced from each model after training. Out of the summaries, the mT5 summaries were less fluent and coherent compared to the mBART summaries, although some of the mT5 summaries were of reasonable quality, demonstrating the potential to perform better; the performance gap was evident. Figure 3 shows early-stage mT5 training outcomes.



یہ آپ کو سمجھانے کی بات ہے۔

Figure 3. mT5 initially training Result

We used BERTScore to measure the quality of the automated summaries, considering the generated summaries and the reference summaries [25]. The BERTScore results indicated that mBART produced better quality summaries, as shown by the score of 0.497, compared to 0.355 from mT5. While these results were better for mBART, these results need to be supplemented by additional information, as the BERTScore value is limited in detailed measurement of the quality of the output summaries. The quality of any summaries produced is subjective, and in the case of an abstractive summary, there are additional factors, such as fluency, coherence, and relevance to the context, that also influence the quality of the summary produced. To strengthen the automated metrics, a qualitative evaluation of the outputs produced was also undertaken [17]. The review of the output summaries demonstrated that the summary reports from both models were semantically accurate, and the results were satisfactory, although there was a greater degree of fluency and coherence in the outputs of mBART.

Our results using mBART have improved upon previous research on Urdu abstractive summarization, which relied on either extractive summarization or smaller multilingual models. The mBART has not been widely researched or systematically tested in previous published literature, especially in large-scale studies involving semantic evaluation to abstractively summarize Urdu. This study is a valuable contribution in this regard. The distance we have gone in generating quality summaries is a testament to mBART's capability to summarize even in low-resourced languages like Urdu.

Future Work and Recommendations:

To enhance model performance, the focus in this case is on expanding the training dataset. In dealing with low-resourced languages like Urdu, the quantity and diversity of high-quality datasets need to be improved. The insufficient availability of large, diverse, high-quality datasets that pertain to the Urdu language is a significant limitation in the field and often requires a great deal of manual effort and curation to overcome. However, the effort of a more diverse dataset is worth incorporating, as it will improve multi-domain model handling and overall model robustness.

Hyperparameter optimization will be a secondary area of focus, in addition to data augmentation and expanding the dataset. If the model is to generate accurate summaries, it is important to balance learning rate and batch size. Because of the amount of power that is required to train the models, more high-performing GPUs will need to be purchased. The more powerful GPUs will allow for faster elaborations and the more efficient fine-tuning of the models, especially for larger data sets. The power of the more expensive and higher-quality GPUs will make a difference in both the level of the experiments and the model's performance. High demand and more powerful GPUs will improve the development of Urdu abstractive summarization models.

References:

[1] A. Raza, H. S. Raja, and U. Maratib, "Abstractive Summary Generation for the Urdu Language," May 2023, Accessed: Dec. 30, 2025. [Online]. Available: http://arxiv.org/abs/2305.16195

[2] M. H. S. Asif Raza, "Abstractive Text Summarization for Urdu Language | Journal of Computing & Biomedical Informatics," Journal of Computing & Biomedical Informatics. Accessed: Dec. 30, 2025. [Online]. Available: https://jcbi.org/index.php/Main/article/view/596

[3] N. Shafiq, I. Hamid, M. Asif, Q. Nawaz, H. Aljuaid, and H. Ali, "Abstractive text summarization of low- resourced languages using deep learning," *PeerJ Comput. Sci.*, vol. 9, p. e1176, Jan. 2023, doi: 10.7717/PEERJ-CS.1176/SUPP-2.

[4] M. Awais and R. Muhammad Adeel Nawab, "Abstractive Text Summarization for the Urdu Language: Data and Methods," *IEEE Access*, vol. 12, pp. 61198–61210, 2024, doi: 10.1109/ACCESS.2024.3378300.

[5] H. Raza and W. Shahzad, "End to End Urdu Abstractive Text Summarization With Dataset and Improvement in Evaluation Metric," *IEEE Access*, vol. 12, pp. 40311–40324, 2024, doi: 10.1109/ACCESS.2024.3377463.

[6] S. Khan, I., Khalil, M. I. K., Nawaz, A., Khan, I. A., Zafar, L., & Ahmed, "Urdu Language Text Summarization using Machine Learning | Journal of Computing & Biomedical Informatics," Journal of Computing & Biomedical Informatics.

[7] M. S. T. Muhammad Ayaz, "The Impact of Technology on Pakistan's Political Discourse: Integrating Islamic Values," *Tanazur Res. J.*, vol. 5, no. 3, 2024, [Online]. Available: https://tanazur.com.pk/index.php/tanazur/article/view/322

[8] Simm, J., & Potts, S., "Multiclass Classification of German News Articles Using Convolutional Neural Networks," *Learn. Deep Textwork*, 2021.

[9] S. Kasim, A. Amjad, D. A. Dewi, and S. Kasim, "Evaluating Classical and Transformer-Based Models for Urdu Abstractive Text Summarization: A Systematic Review," Jul. 2025, doi: 10.20944/PREPRINTS202507.1846.V1.

[10] R. Aharoni, S. Narayan, J. Maynez, J. Herzig, E. Clark, and M. Lapata, "Multilingual Summarization with Factual Consistency Evaluation," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 3562–3591, 2023, doi: 10.18653/V1/2023.FINDINGS-ACL.220.

[11] A. Bhattacharjee, T. Hasan, W. U. Ahmad, Y. F. Li, Y. Bin Kang, and R. Shahriyar, "CrossSum: Beyond English-Centric Cross-Lingual Summarization for 1,500+ Language Pairs," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, pp. 2541–2564, 2023, doi: 10.18653/V1/2023.ACL-LONG.143.

[12] A. Scirè, K. Ghonim, and R. Navigli, "FENICE: Factuality Evaluation of summarization based on Natural language Inference and Claim Extraction," *Proc.*

Annu. Meet. Assoc. Comput. Linguist., pp. 14148–14161, 2024, doi: 10.18653/V1/2024.FINDINGS-ACL.841.

[13] J. Z. Forde *et al.*, "Re-Evaluating Evaluation for Multilingual Summarization," *EMNLP 2024 - 2024 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 19476–19493, 2024, doi: 10.18653/V1/2024.EMNLP-MAIN.1085.

[14] Y. Ye *et al.*, "GlobeSumm: A Challenging Benchmark Towards Unifying Multi-lingual, Cross-lingual and Multi-document News Summarization," *EMNLP 2024 - 2024 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 10803–10821, 2024, doi: 10.18653/V1/2024.EMNLP-MAIN.603.

[15] R. Zhang, J. Ouni, and S. Eger, "Cross-lingual Cross-temporal Summarization: Dataset, Models, Evaluation," *Comput. Linguist.*, vol. 50, no. 3, pp. 1001–1047, Sep. 2024, doi: 10.1162/COLI_A_00519.

[16] S. Mille *et al.*, "The 2024 GEM Shared Task on Multilingual Data-to-Text Generation and Summarization: Overview and Preliminary Results," *INLG 2024 - 17th Int. Nat. Lang. Gener. Conf. Proc. Gener. Challenges*, pp. 17–38, 2024, doi: 10.18653/V1/2024.INLG-GENCHAL.2.

[17] D. Li *et al.*, "From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge," pp. 2757–2791, Nov. 2025, doi: 10.18653/V1/2025.EMNLP-MAIN.138.

[18] Z. Luo, Q. Xie, and S. Ananiadou, "Factual consistency evaluation of summarization in the Era of large language models," *Expert Syst. Appl.*, vol. 254, p. 124456, Nov. 2024, doi: 10.1016/J.ESWA.2024.124456.

[19] I. Mondshine, T. Paz-Argaman, and R. Tsarfaty, "Beyond N-Grams: Rethinking Evaluation Metrics and Strategies for Multilingual Abstractive Summarization," Jul. 2025, Accessed: Dec. 30, 2025. [Online]. Available: https://arxiv.org/pdf/2507.08342

[20] N. Dahan and G. Stanovsky, "The State and Fate of Summarization Datasets: A Survey," pp. 7259–7278, Jun. 2025, doi: 10.18653/V1/2025.NAACL-LONG.372.

[21] M. Azam *et al.*, "Current Trends and Advances in Extractive Text Summarization: A Comprehensive Review," *IEEE Access*, vol. 13, pp. 28150–28166, 2025, doi: 10.1109/ACCESS.2025.3538886.

[22] M. Rodríguez-Ortega, M., Rodríguez-Lopez, E., Lima-López, S., Escolano, C., Melero, M., Pratesi, L., ... & Krallinger, "Overview of MultiClinSum task at BioASQ 2025: evaluation of clinical case summarization strategies for multiple languages: data, evaluation, resources and results," *CLEF 2025 Work. Notes*, 2025.

[23] N. Hussain, A. Qasim, G. Mehak, O. Kolesnikova, A. Gelbukh, and G. Sidorov, "Hybrid Machine Learning and Deep Learning Approaches for Insult Detection in Roman Urdu Text," *AI 2025, Vol. 6, Page 33*, vol. 6, no. 2, p. 33, Feb. 2025, doi: 10.3390/AI6020033.

[24] S. Ali, U. Jamil, M. Younas, B. Zafar, and M. Kashif Hanif, "Optimized Identification of Sentence-Level Multiclass Events on Urdu-Language-Text Using Machine Learning Techniques," *IEEE Access*, vol. 13, pp. 1–25, 2025, doi: 10.1109/ACCESS.2024.3522992.

[25] S. Haque, Z. Eberhart, A. Bansal, and C. McMillan, "Semantic Similarity Metrics for Evaluating Source Code Summarization," *IEEE Int. Conf. Progr. Compr.*, vol. 2022-March, pp. 36–47, Oct. 2022, doi: 10.1145/3524610.3527909;CSUBTYPE:STRING:CONFERENCE.