[1]Sumayya Afreen*

Dr. P V Sudha[2]

Evaluating Multilingual Models: A Comparative Evaluation of MT0 and MT5 for Urdu Text Articles Summarisation

Abstract: - Abstractive text summarization is a critical task in natural language processing aimed at generating concise summaries while preserving the core meaning of the original text. This paper focuses on abstractive summarization of Urdu articles using the MT5 pre-trained model, a variant of T5 that supports multilingual capabilities. Specifically fine-tuned for Urdu, MT5 harnesses cross-linguistic knowledge to enhance summarization quality. We evaluate its performance against the MT0 model, a baseline, using ROUGE metrics. Our experiments, conducted on a novel Urdu dataset, demonstrate significant improvements in summary quality with MT5 compared to MT0. This research underscores the utility of multilingual models in advancing abstractive summarization and highlights avenues for future development in natural language processing for underrepresented languages.

Keywords: Abstractive summarization, Multilingual models, MT0 model, MT5 model, ROUGE metrics, Urdu dataset

INTRODUCTION

Urdu, one of the predominant languages of South Asia, boasts a rich literary tradition and is spoken by millions of people worldwide. Despite widespread use, a significant gap remains in advanced natural language processing (NLP) tools tailored specifically to the Urdu language. This gap poses a challenge in managing and summarizing the vast amounts of digital content available in Urdu, particularly in the realm of news. To address this, research focuses on performing abstractive text summarization on Urdu news articles, using a cutting-edge language model known as MT5 (Multilingual Text-to-Text Transfer Transformer).

The increasing digitalization of news outlets has led to an overwhelming influx of information, necessitating efficient methods to digest and comprehend news content. News articles, due to their complexity and depth, require sophisticated summarization techniques that can capture the essence while maintaining coherence. Abstractive text summarization, unlike extractive approaches that compile snippets from the original text, generates new sentences that succinctly convey the core ideas. This method is particularly valuable for news content as it provides a more readable and meaningful summary.

This research leverages the extensive repository of Urdu news articles from the BBC Urdu News website. BBC Urdu is a reputable source, offering a rich and diverse range of news covering various domains such as politics, economy, culture, and technology. This source provides a robust dataset ideal for training and evaluating the summarization model.

The objective of this paper is to evaluate two large language models, MT0 and MT5, for the task of summarizing Urdu news articles. Following the evaluation, a comparative analysis of the results is conducted to determine the performance and effectiveness of each model.

RELATED WORK

The paper "Abstractive Text Summarization for the Urdu Language: Data and Methods" by Muhammad Awais and Rao Muhammad Adeel Nawab introduces a large benchmark corpus of 2,067,784 Urdu news articles for abstractive text summarization, addressing a gap in Urdu language processing. The authors evaluate several deep learning models and transformers on this corpus, including LSTM-based encoder-decoder architecture, Bi-LSTM-based encoder-decoder architecture, GRU-based encoder-decoder architecture, Bi-GRU-based encoder-decoder architecture, LSTM-based encoder-decoder architecture with attention, GRU-based encoder-decoder architecture with attention, BART (Bidirectional Auto-Regressive Transformers), and GPT-3.5 (Generative Pre-

[1] [1*]Research Scholar, Department of Computer Science and Engineering, University College of Engineering, Osmania University

[1*]Assistant Professor, Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women

[2]Professor, Department of Computer Science and Engineering, Unversity College of Engineering, Osmania Unversity

trained Transformer). They find that the GRU with attention model performs best, achieving notable ROUGE scores. This work provides a significant resource for future research in Urdu text summarization and is made publicly available under a Creative Commons License [1].

The paper "End to End Urdu Abstractive Text Summarization" by H. Raza and W. Shahzad addresses the growing need for automatic text summarization due to the increasing volume of textual data available on the internet. The authors discuss the challenges of manual summarization and the necessity for automatic techniques, particularly for the Urdu language, which lacks ready resources like corpora, tag sets, and embedding vectors. The paper distinguishes between Extractive Text Summarization (ETS) and Abstractive Text Summarization (ATS), noting that ATS involves paraphrasing and natural language generation to create summaries. The authors highlight the lack of work in Urdu abstractive summarization and review various methodologies, including machine learning-based methods, graph-based methods, and evolutionary algorithms. They also discuss the challenges of translating and preprocessing Urdu text, emphasizing the need for accurate translations and the importance of diacritic marks in understanding the language.

The paper concludes by suggesting future work to enhance datasets, train better embeddings, and develop summarization models with optional summary lengths. The authors make a significant contribution by providing a structured approach to Urdu text summarization and highlighting the need for further research in this area.

A new evaluation metric named the "disconnection rate" is introduced to provide a contextually informed assessment of a summary. This metric, referred to as the Context Aware RoBERTa Score, aims to improve the evaluation process by considering contextual factors [2].

Large Language Models (LLMs) have demonstrated extraordinary capabilities in various natural language processing tasks, including language translation, text generation, and question answering. This review provides a comprehensive overview of LLMs, covering their evolution, architectures, transformers, resources, training methods, applications, societal impacts, and open issues and challenges. LLMs have advanced the field of natural language processing through the development of models like BERT, GPT, and T5, which leverage deep learning and large datasets to achieve state-of-the-art performance. The review examines the influence of machine learning models on LLMs and the datasets used in their training. It also explores the hardware implementation and the diverse applications of LLMs in healthcare, education, social media, business, and agriculture. Finally, the paper discusses the open issues and challenges facing LLMs, including ethical concerns, security, privacy, and the need for improved generalization and few-shot learning capabilities.

LLMs have evolved from early language models and neural networks, with the introduction of the transformer architecture as a significant milestone.

LLMs utilize diverse datasets, including web pages, books, news, scientific data, and code, to train models with varying capabilities and configurations.

LLMs have been successfully applied in various domains, such as healthcare, education, social media, business, and agriculture, demonstrating their versatility and potential.

LLMs have had a significant impact on society, enabling advancements in natural language processing, automation, and content generation, but also raising ethical concerns and challenges.

Open issues and challenges faced by LLMs include ethical and responsible AI, multimodal integration, energy efficiency, security and adversarial attacks, privacy and data protection, generalization and few-shot learning, and cross-lingual and low-resource settings.

Future research directions focus on enhancing bias mitigation, optimizing efficiency, improving context handling, enabling continuous learning, increasing interpretability, incorporating multimodal capabilities, and developing ethical and legal frameworks [3].

The paper "A Comprehensive Review of Arabic Text Summarization" discusses the challenges of summarizing large volumes of text data, emphasizing the three main techniques: extractive, abstractive, and hybrid. Despite advancements, automated summarization, especially for Arabic, lags behind human quality due to the language's morphological complexity, dialect variety, and data scarcity. The review highlights issues like the lack of golden tokens, out-of-vocabulary words, and repetitive summaries, calling for improved evaluation metrics and deep learning models. Key challenges include Arabic's orthographic ambiguity, lack of capitalization, and absence of standardized resources. The paper concludes that while recent deep learning advancements show promise, more research and refined methodologies are essential to bridge the gap between human and automated summarization for Arabic texts [4].

Attention is all you need is one of the finest paper which has brought a revolutionary change in the field of natural language processing. A transformer which includes encoder and decoder is constructed uisng this attention mechanism which is the base for many deep learning and large language model like BERT, BART, T5 etc [5]

MT5 (Multilingual Text-to-Text Transfer Transformer) is a variant of the T5 (Text-to-Text Transfer Transformer) model specifically designed to handle multiple languages. It is part of the family of transformer-based models developed by Google Research. The large and diverse training dataset allows MT5 to perform well across a wide range of languages and tasks. The extensive multilingual training helps the model learn patterns and representations that are useful for various NLP tasks in different languages [6].

DATASET DESCRIPTION:

Data set used is BBC-news data set provided by xl-sum (https://github.com/csebuetnlp/xl-sum) which has around 84.6 k news articles published on https://www.bbc.com/urdu. The data set consist of 67.7k articles for training, 8.46k for training, and 8.46k for validating.

Number of words in longest summary (train): 250

Number of words in shortest summary (train): 55

Number of words in longest article (train): 11230

Number of words in shortest article (train): 24

Number of words in longest summary (test): 69

Number of words in shortest summary (test): 11

Number of words in longest article (test): 1492

Number of words in shortest article (test): 93

Average number of words in summaries (train): 35.89

Average number of words in articles (train): 541.14

Average number of words in summaries (test): 33.06

Average number of words in articles (test): 476.87

ABSTRACTIVE TEXT SUMMARIZATION FOR URDU LANGUAGE USING LLM

Large Language Models which are based on transformer model as shown in Fig. 1 can be used for abstractive Text Summarization of Urdu language as shown in and the result can be compared with the human generated summary to check accuracy of the model. Not all models are trained with languages like Urdu. Where as few models like MT0, MT5, mBert are trained for multiple languages like Urdu, Hindi etc.

A transformer is a deep learning model architecture introduced in the paper *"Attention is All You Need"* by Vaswani et al. in 2017. The transformer architecture is designed to handle sequential data and is particularly effective for tasks involving natural language.

Key Components of Transformers

Self-Attention Mechanism: This allows the model to weigh the importance of different words in a sentence relative to each other. For example, in the sentence "The cat sat on the mat," the model can learn to focus on the relationship between "cat" and "mat."

Multi-Head Attention: This extends the self-attention mechanism by running multiple attention processes in parallel. It helps the model capture different aspects of the relationships between words.

Positional Encoding: Since transformers do not inherently process sequences in order, positional encodings are added to input embeddings to give the model information about the position of each word in the sequence.

Feedforward Neural Networks: After the attention layers, the model uses fully connected feedforward networks to transform the representations of the words.

Layer Normalization and Residual Connections: These techniques help stabilize training and allow for deeper models by normalizing the outputs of each layer and adding the input of the layer back to its output.

Transformers in Large Language Models

In large language models, transformers are used as the foundational architecture. Here's how they are applied:

1. **Pretraining**: Transformers are used to train models on large corpora of text to understand and generate human-like text. During this phase, models learn to predict missing words, understand context, and grasp various linguistic patterns.
2. **Fine-Tuning**: After pretraining, models can be fine-tuned on specific tasks such as translation, summarization, or question answering. This involves adjusting the model parameters to improve performance on these tasks.
3. **Scalability**: Transformers scale effectively with larger datasets and more computational resources. This scalability is a key reason why transformer-based models like GPT-3, GPT-4, and BERT have achieved impressive results across a wide range of NLP tasks.
4. **Transfer Learning**: Transformers enable transfer learning, where a model trained on one task (like general language understanding) can be adapted to perform well on a different but related task (like sentiment analysis).

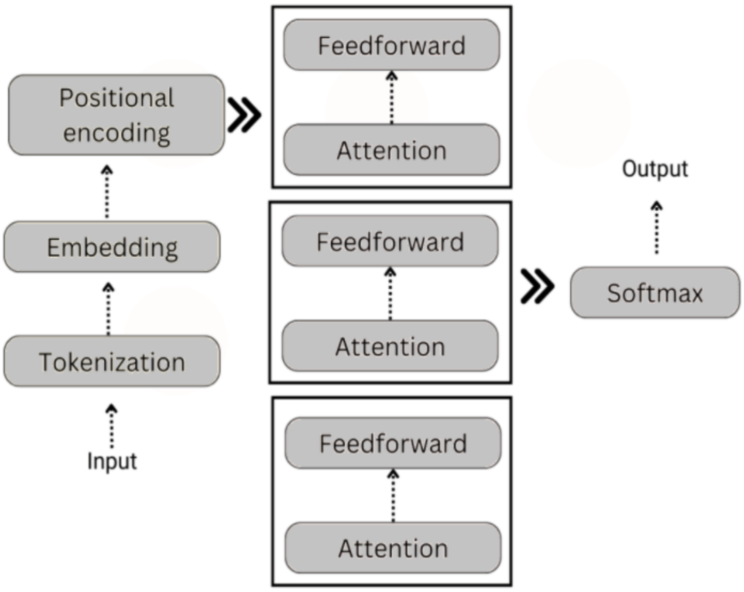

Fig 1: Working of Transformer

MT5

MT5 (Multilingual T5) is a variant of the T5 (Text-To-Text Transfer Transformer) model, specifically designed to handle multiple languages. Like T5, MT5 treats every NLP task as a text-to-text problem, meaning that both input and output are always in the form of text.

MT5 is the extended version of the C4 pre-training dataset that covers 101 languages and introduce changes to T5 to better suit this multilinguality. The C4 dataset was designed to be strictly English, using langdetect2 to ensure each page had at least a 99% probability of being in English. In contrast, mC4, which covers over 100 languages, uses cld3 for language identification and includes data from 71 monthly web scrapes by Common Crawl, a much larger source compared to the single April 2019 web scrape used for C4.

C4 filtered out lines that did not end with English punctuation, but since many languages don't use the same punctuation marks, mC4 applies a "line length filter." This filter requires pages to have at least three lines of text, each with 200 or more characters. Both C4 and mC4 remove duplicate lines and pages containing inappropriate content. Additionally, mC4 filters out pages where the primary language has less than 70% confidence according to cld3.

After these filters, mC4 groups the remaining pages by language, including only those languages with at least 10,000 pages. This results in text from 107 languages, though six are script variants of the same spoken language (like Russian in Cyrillic and Latin scripts). The appendix provides detailed statistics on the dataset, including the number of tokens per language, and a histogram shows the page counts for each language. Table 1. Demonstrate variants of MT5. The variant used in this paper is MT5-Small.

MT5 follows the same architecture as the original T5 model, which is based on the Transformer model. It consists of:

Encoder: Processes the input text and creates a sequence of context-aware representations.

Decoder: Generates the output text based on the encoded input representations and previous generated tokens.

Training

MT5 is trained using a large multilingual corpus and employs the following key techniques:

Pre-training:

- Self-Supervised Learning: MT5 is pre-trained using a denoising autoencoding objective. This involves corrupting the input text by randomly masking tokens and training the model to predict the original text.
- Causal Language Modeling: Like autoregressive models, where the model predicts the next token in a sequence, given the previous tokens.

Fine-Tuning:

- After pre-training, MT5 can be fine-tuned on specific tasks, such as translation, summarization, or question answering, by providing task-specific examples.

Variants of MT5	**Layers**	**Hidden Units**	**Parameters**
MT5-Small	6 encoder, 6 decoder	512	~60M
MT5-Base	12 encoder, 12 decoder	768	~220M
MT5-Large	24 encoder, 24 decoder	1025	~600M
MT5-XL	24 encoder, 24 decoder	2048	~3B
MT5-XXL	48 encoder, 48 decoder	4096	~13B

Table 1: MT5 Variants

The results calculated using **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) metrics by comparing model generated summary with human generated summary are shown in table 2.

	Rouge 1			Rouge 2			Rouge L		
1	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
2	34.50	44.20	38.77	40.625	50.0	46.9	29.87	38.3	41.11

Table 2: MT5 Results for BBC news data set

MT0

mT0 is a variant of the MT5 model, specifically fine-tuned on the xP3 dataset. This dataset includes multilingual data with prompts in English, enhancing the model's capability to handle multiple languages while maintaining the structure and functionality of the MT5 architecture.

MT5 Foundation:

mT0 builds on the MT5 model, which is designed for multilingual natural language processing tasks. MT5 is part of the T5 family, known for its text-to-text framework, where all tasks are converted into a text generation problem.

xP3 Dataset:

The xP3 dataset consists of multilingual datasets with English prompts. This dataset is specifically designed to improve the model's performance in multilingual settings by providing diverse linguistic data with a common prompt language (English).

Fine-Tuning:

mT0 is fine-tuned on the xP3 dataset, which means it has been further trained on this specific dataset to optimize its performance for the tasks and languages included in xP3.

Training and Fine-Tuning:

- **Pre-training:**

  The base MT5 model undergoes extensive pre-training on a large multilingual corpus. This step helps the model learn linguistic structures and patterns across various languages.

- **Fine-Tuning on xP3:**

  After pre-training, mT0 is fine-tuned on the xP3 dataset. This involves adjusting the model's parameters to better handle the specific tasks and languages included in xP3, with English prompts guiding the tasks.

	Rouge 1			Rouge 2			Rouge L		
1	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
2	22.69745	30.85896	26.15634	18.75	26.08695	28.27301	18.255129	25.059125	27.65826

Table 3: MT0 Results for BBC news data set

COMPARATIVE EVALUATION OF MT0 AND MT5

In our research, we executed both the MT5 and MT0 models on the same dataset to perform a comparative analysis. This analysis was carried out using ROUGE metrics, which measure the quality of the generated summaries by comparing them to reference summaries. Specifically, we utilized ROUGE-1, ROUGE-2, and

ROUGE-L metrics to evaluate the performance of each model in terms of n-gram overlap and longest common subsequence respectively. The results of this rigorous evaluation indicate that the MT5 model consistently outperforms the MT0 model across all key metrics. This superior performance of MT5 is evidenced by higher ROUGE scores, demonstrating its greater efficacy in producing accurate and concise summaries for the given dataset.

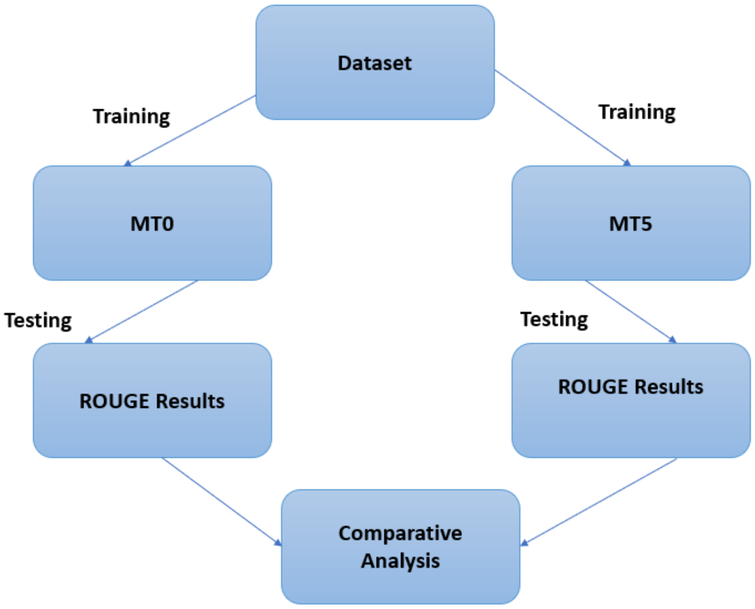

Fig 2: Comparative evaluation of MT0 and MT5

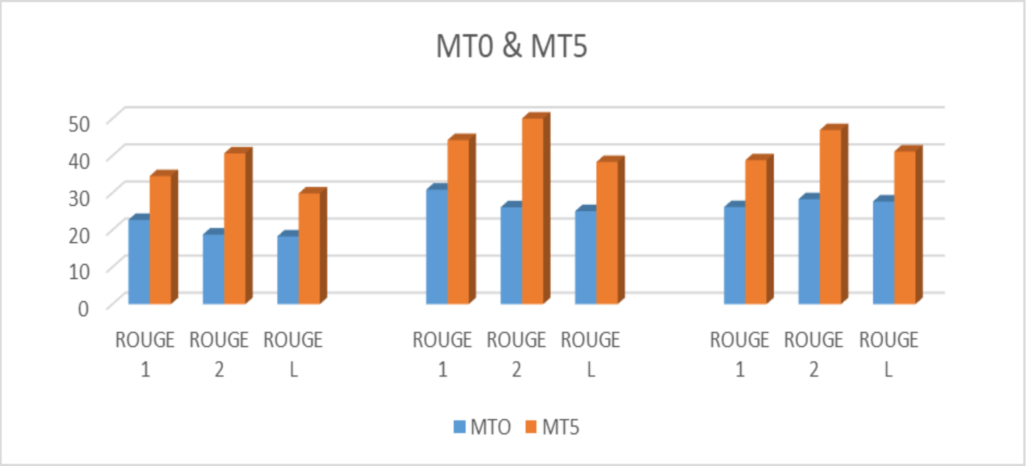

Fig 3: Performance evaluation of MT0 and MT5

CONCLUSION AND FUTURE ENHANCEMENT

In this paper we have performed summarization of the news artcles uidng two large language models. MT0 and MT5, these are variants of T5 (Text-to-text- Transfer Transformer). There is significance difference between the results of both the models. Though MT0 IS a fine-tuned model on XP3 dataset, MT5 which is only pretrained gives the better results for urdu language. Future enhancements could focus on fine-tuning MT5 specifically for the Urdu language using a larger, more diverse dataset, and developing hybrid models that combine the strengths of both MT0 and MT5. Incorporating advanced pre-processing techniques, such as better tokenization and contextual data augmentation, and exploring additional fine-tuning strategies like transfer learning and task-specific training could also enhance performance. Additionally, expanding evaluation metrics to include qualitative measures and integrating user feedback for continuous refinement would provide a more comprehensive assessment. Deploying these improved models in real-world applications, such as news aggregators, and developing interactive summarization tools would further enhance their practical utility and user relevance.

REFERENCES

1. Awais, M., & Nawab, R. M. A. (2024). Abstractive Text Summarization for the Urdu Language: Data and Methods. *IEEE Access*.
2. Raza, H., & Shahzad, W. (2024). End to End Urdu Abstractive Text Summarization with Dataset and Improvement in Evaluation Metric. *IEEE Access*.
3. Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., ... & Azam, S. (2024). A review on large Language Models: Architectures, applications, taxonomies, open issues, and challenges. *IEEE Access*.
4. Elsaid, A., Mohammed, A., Ibrahim, L. F., & Sakre, M. M. (2022). A comprehensive review of arabic text summarization. *IEEE Access*, *10*, 38012-38030.
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
6. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.
7. Asif, Muhammad, Syed Ali Raza, Javed Iqbal, Nousheen Perwaiz, Tauqeer Faiz, and Shan Khan. "Bidirectional Encoder Approach for Abstractive Text Summarization of Urdu Language." In 2022 International Conference on Business Analytics for Technology and Security (ICBATS), pp. 1-8. IEEE, 2022.
8. Shafiq, Nida, Isma Hamid, Muhammad Asif, Qamar Nawaz, Hanan Aljuaid, and Hamid Ali. "Abstractive text summarization of low-resourced languages using deep learning." PeerJ Computer Science 9 (2023): e1176.
9. Muhammad, Aslam, Noman Jazeb, Ana Maria Martinez-Enriquez, and Ali Sikander. "EUTS: extractive Urdu text summarizer." In 2018 seventeenth mexican international conference on artificial intelligence (MICAI), pp. 39-44. IEEE, 2018.
10. Nawaz, Ali, Maheen Bakhtyar, Junaid Baber, Ihsan Ullah, Waheed Noor, and Abdul Basit. "Extractive text summarization models for Urdu language." Information Processing & Management 57, no. 6 (2020): 102383.
11. Humayoun, Muhammad, Rao Muhammad Adeel Nawab, Muhammad Uzair, Saba Aslam, and Omer Farzand. "Urdu summary corpus." In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 796-800. 2016.
12. Basit, Rida Hijab, Muhammad Aslam, A. M. Martinez-Enriquez, and Afraz Z. Syed. "Semantic similarity analysis of urdu documents." In Pattern Recognition: 9th Mexican Conference, MCPR 2017, Huatulco, Mexico, June 21-24, 2017, Proceedings 9, pp. 234-243. Springer International Publishing, 2017.
13. Agarwala, Saurabh, Aniketh Anagawadi, and Ram Mohana Reddy Guddeti. "Detecting semantic similarity of documents using natural language processing." Procedia Computer Science 189 (2021): 128-135.
14. Himaja, Prathi, Guduru Anvitha, Syed Fasih Ahsan, Manikya Chowdary, and T. Vignesh. "A Survey on Text Summarization in Urdu Language using Machine Learning Techniques." In 2023 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-6. IEEE, 2023.
15. Naseer, Asma, Tanzeela Shakeel, Kinza Arshad, and Zeenia Ather. "Analysis of Corpus Development for Urdu Language." In 2021 International Conference on Innovative Computing (ICIC), pp. 1-5. IEEE, 2021.
16. Iqbal, Muntaha, Bilal Tahir, and Muhammad Amir Mehmood. "CURE: Collection for urdu information retrieval evaluation and ranking." 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2). IEEE, 2021.
17. Nazir, Shahzad, Muhammad Asif, Shahbaz Ahmad Sahi, Shahbaz Ahmad, Yazeed Yasin Ghadi, and Muhammad Haris Aziz. "Toward the development of large-scale word embedding for low-resourced language." IEEE Access 10 (2022): 54091-54097.
18. Farooq, Aman, Safiyah Batool, and Zain Noreen. "Comparing Different Techniques of Urdu Text Summarization." In 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), pp. 1-6. IEEE, 2021.
19. Wazery, Yaser Maher, Marwa E. Saleh, Abdullah Alharbi, and Abdelmgeid A. Ali. "Abstractive Arabic text summarization based on deep learning." Computational Intelligence and Neuroscience 2022 (2022).
20. Ghafouri, Arash, Mohammad Amin Abbasi, and Hassan Naderi. "AriaBERT: A Pre-trained Persian BERT Model for Natural Language Understanding." (2023).
21. Bhatti, Muhammad Wasif, and Muhammad Aslam. "ISUTD: Intelligent System for Urdu Text De-Summarization." In 2019 International Conference on Engineering and Emerging Technologies (ICEET), pp. 1-5. IEEE, 2019.

22. Tahir, Bilal, and Muhammad Amir Mehmood. "UBERT22: Unsupervised Pre-training of BERT for Low Resource Urdu Language." In 2022 16th International Conference on Open-Source Systems and Technologies (ICOSST), pp. 1-6. IEEE, 2022.
23. El-Kassas, Wafaa S., Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. "Automatic text summarization: A comprehensive survey." Expert systems with applications 165 (2021): 113679.
24. Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen et al. "A survey on evaluation of large language models." ACM Transactions on Intelligent Systems and Technology 15, no. 3 (2024): 1-45.
25. Luo, Yun, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. "An empirical study of catastrophic forgetting in large language models during continual fine-tuning." arXiv preprint arXiv:2308.08747 (2023).
26. 26. Brașoveanu, Adrian MP, and Răzvan Andonie. "Visualizing transformers for nlp: a brief survey." In 2020 24th International Conference Information Visualisation (IV), pp. 270-279. IEEE, 2020.
27. Sumayya Afreen, S. Sameen Fatima. “A survey on Text Summarization for Urdu Language using deep Learning Techniques”, IJREAM, Vol 9, Issue 11,2024.
28. Burney, Aqil, Badar Sami, Nadeem Mahmood, Zain Abbas, and Kashif Rizwan. "Urdu text summarizer using sentence weight algorithm for word processors." International Journal of Computer Applications 46, no. 19 (2012): 38-43.
29. Tahir, Bilal, and Muhammad Amir Mehmood. "Corpulyzer: A novel framework for building low resource language corpora." IEEE Access 9 (2021): 8546-8563.
30. Abdul-Mageed, Muhammad, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. "ARBERT & MARBERT: deep bidirectional transformers for Arabic." arXiv preprint arXiv:2101.01785 (2020).
31. Hamza, Syed Ali, Bilal Tahir, and Muhammad Amir Mehmood. "Domain identification of urdu news text." 2019 22nd International Multitopic Conference (INMIC). IEEE, 2019.
32. Nikhila, P., G. Jyothi, K. Mounika, M. C. Kishor, K. Reddy, and R. Murthy. "Chatbots using artificial intelligence." J. Appl. Sci. Comput 6, no. 2 (2019): 103-115.
33. Anisha, P. R., Nhu Gia Nguyen, and G. Sreelatha. "A text mining using web scraping for meaningful insights." In Journal of Physics: Conference Series, vol. 2089, no. 1, p. 012048. IOP Publishing, 2021.
34. Mahalakshmi, C., T. Sharmila, S. Priyanka, M. Sastry, D. B. V. R. M. Reddy, and M. K. K. Reddy. "A SURVEY ON VARIOUS CHATBOT IMPLEMENTENTION TECHNIQUES." JASC: Journal of Applied Science and Computations.