Information Processing and Management 61 (2024) 103734

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Meeting the challenge: A benchmark corpus for automated Urdu meeting summarization

Bareera Sadia [a], Farah Adeeba [b,*], Sana Shams [a], Kashif Javed [c]

[a] Center for Language Engineering (CLE), KICS, University of Engineering and Technology (UET), Lahore, 54890, Pakistan
[b] Department of Computer Science, University of Engineering and Technology (UET), Lahore, 54890, Pakistan
[c] Department of Electrical Engineering, University of Engineering and Technology (UET), Lahore, 54890, Pakistan

ARTICLE INFO

Keywords:
Natural language processing
Abstractive text summarization
Deep learning
Meeting corpus
Urdu
Fine-tuning

ABSTRACT

Meeting summarization has become crucial as the world is gradually shifting towards remote work. Nowadays, automation of meeting summary generation is really needed in order to minimize the time and effort. The surge in online meetings has made summarization an indispensable requirement, yet summarizing Urdu meetings poses a formidable challenge due to the scarcity of pertinent corpora. Abstractively summarizing Urdu meetings compounds this challenge. This research addresses these gaps by introducing the Center for Language Engineering (CLE) Meeting Corpus, a benchmark resource tailored for meeting summarization in administrative and technical domains where Urdu is the primary language. Comprising 240 recorded meetings, encompassing both scenario-based and natural discussions, the corpus spans approximately 7900 min (~132 h) of meeting duration. Beyond corpus creation, the study delves into the performance analysis of various deep learning models in Urdu abstractive meeting summarization. Models, including ur_mT5-small, ur_mT5-base, ur_mBART-large, ur_RoBERTa-urduhack-small, and GPT-3.5 with prompting, undergo comprehensive evaluation using both automated metrics and manual assessments based on five specific criteria. This research not only addresses the immediate challenges of Urdu meeting summarization but also contributes to advancing the capabilities of meeting summarization systems in diverse organizational contexts where Urdu is the language of communication during meetings.

1. Introduction

Today, an extensive percentage of the workforce conducts their regular meetings and interactions virtually. It takes a lot of work to make notes in meetings and obtain the desired information most of the times due to internet connectivity and bandwidth issues. This leads to the unexpected loss of information for the participants. For this purpose, the minutes of the meeting are typically recorded by a particular participant who is assigned to keep record of discussions, decisions, and actions taken during a meeting. These minutes serve as a summary of a meeting for planning purposes and to keep track of conversations. Due to several factors, generation of meeting minutes becomes a cumbersome task when it comes to virtual meetings. Technical difficulties in a meeting makes it more difficult to understand the details and also interrupt the process of taking notes. Furthermore, it can be time-consuming to accurately write down the important discussion and manage digital documents. Therefore, it is absolutely necessary to automate the entire procedure of generating the meeting minutes.

* Corresponding author.
E-mail addresses: bareera.sadia@kics.edu.pk (B. Sadia), farah.adeeba@uet.edu.pk (F. Adeeba), sana.shams@kics.edu.pk (S. Shams), kashif.javed@uet.edu.pk (K. Javed).

https://doi.org/10.1016/j.ipm.2024.103734
Received 22 November 2023; Received in revised form 23 February 2024; Accepted 30 March 2024
Available online 13 April 2024

In order to generate meeting minutes, a meeting needs to be summarized first which can be done in two ways i.e. abstractive or extractive. Abstractive summarization (Song, Huang, & Ruan, 2019) is a technique used in Natural Language Processing (NLP) (Qiu et al., 2020) that involves generating a concise and coherent summary of a given text document. Unlike extractive summarization (Rahimi, Mozhdehi, & Abdolahi, 2017), which selects and combines existing sentences or phrases from the original text, abstractive summarization involves understanding the meaning of the text and generating new sentences that convey the essential information.

Abstractive summarization algorithms use advanced NLP techniques, such as machine learning and deep learning, to comprehend the input text and generate a summary that captures the key points. These algorithms often rely on neural networks (Jadhav, Jain, Fernandes, & Shaikh, 2019), such as Recurrent Neural Networks (RNNs) (Dam et al., 2023) or transformers (Ranganathan & Abuka, 2022), to model the language and generate human-like summaries. Abstractive summarization has several advantages over extractive methods, as it can produce more concise and coherent summaries that are not limited to the sentences present in the original text. However, it is also a more challenging task due to the need for deeper language understanding.

1.1. Research objectives and contributions

In the context of Pakistan, where Urdu serves as the lingua franca, meetings are predominantly conducted in a dynamic linguistic environment characterized by the seamless integration of Urdu and English, known as code-mixing. Despite the prevalence of this linguistic phenomenon in everyday communication, the existing research landscape predominantly centers around languages with ample linguistic resources, notably English, as mentioned in a recent survey on abstractive meeting summarization (Rennard, Shang, Hunter, & Vazirgiannis, 2023). This linguistic bias poses a significant gap in the availability of resources and research insights for Urdu, particularly in the domain of meeting summarization. The dire need to bridge this gap is underscored by the unique challenges posed by code-mixed language usage in meetings, a distinctive linguistic practice that demands dedicated attention. Hence, the main objectives of this research study are as follows:

- To develop a novel dataset for automatic summarization of meetings in Urdu (a low-resource language).
- To develop guidelines for the manual transcription of recorded meetings in Urdu on Xtrans software.
- To analyze and compare various summarization models for abstractive summarization of meetings in Urdu.

In order to fulfill these objectives, the key contributions of our research work can be summarized as:

- **Development of the CLE Meeting Corpus:** A comprehensive corpus specifically for Urdu meeting summarization has been created,[1] which includes annotated data for summaries. This corpus is made available online for researchers to do further analysis. It consists of 240 recorded meetings covering four main administrative domains i.e. hiring, procurement, admin affairs and finance as well as technical domain related to computer science. Currently the work is still in progress to increase dataset further from 240 onwards. This recorded speech corpus also serves as a valuable resource for training of Automatic Speech Recognition (ASR) system (Khan, Rauf, Adeeba, & Hussain, 2021) specifically for the purpose of meeting transcription.
- **Annotation Guidelines:** For annotation recorded meeting corpus, comprehensive guidelines are developed. As the guidelines included in this paper are language-independent, therefore they can be used to enhance our developed corpus and develop new corpora for Urdu, as well as other languages.
- **Comparative Analysis of Models on Abstractive Summarization:** A comparative analysis of various deep learning models has been conducted for abstractive meeting summarization in order to evaluate their performance and identify the most effective approach.

The rest of the paper is structured as follows: Section 2 provides a detailed study on related work for abstractive meeting summarization in different languages as well as the available corpora for meeting summarization. Section 3 presents the complete development of CLE Meeting Corpus. Section 4 is focused on the experimentation and results are available in Section 5. Finally, Sections 6, 7, 8 provide the conclusion of the research, limitations of this research and the future work, respectively.

2. Related work

Within the constantly evolving field of NLP, researchers are concentrating on the challenge of distilling large amounts of textual data into concise, yet contextually rich, summaries. Abstractive summarization in NLP is at the forefront of this effort; its goal is to produce human-like, logical summaries that encapsulate the essence of the source material, rather than just extracting the most important sentences. The study of abstractive summarization techniques is becoming more and more relevant in the age of information overload, as the need for efficient information retrieval and content condensation grows. A plethora of studies have been conducted on text summarization for different languages. The goal of this literature review is to provide light on the important contributions made by researchers and the state of the art in this rapidly developing field by examining the various approaches, difficulties, and developments in the field of abstractive summarization. Table 1 presents a summary of the key studies in abstractive text summarization.

[1] https://github.com/farahadeeba/CLEMeetingCorpus.

Table 1
Summary of the research studies in the domain of abstractive text summarization in different languages.

Researcher and reference	Year	Architecture	Language
Mohammad Masum, Abujar, Islam Talukder, Azad Rabby, and Hossain (2019)	2019	Sequence to sequence RNNs	English
Parida and Motlicek (2019)	2019	State-of-the-art transformer model	German
Zaman, Shardlow, Hassan, Aljohani, and Nawaz (2020)	2020	Hybrid text summarization and simplification architecture	English
Li and Zhuge (2021)	2021	Semantic link networks	Chinese
Alahmadi, Wali, and Alzahrani (2022)	2022	RNN-based abstractive summarization	Arabic
Nagoudi, Elmadany, and Abdul-Mageed (2022)	2022	Multilingual T5 model (mT5) vs. Pre-trained Arabic T5-style models	Arabic
Phan, Tran, Nguyen, and Trinh (2022)	2022	Pre-trained Vietnamese T5-style model	Vietnamese
Ay, Ertam, Fidan, and Aydin (2023)	2023	Text to text transfer transformer (T5) technique	Turkish
Bani-Almarjeh and Kurdy (2023)	2023	RNN-based and transformer-based architectures	Arabic
La Quatra and Cagliero (2023)	2023	BART architecture	Italian
Shafiq et al. (2023)	2023	Multi-layer encoder and single layer decoder model	Urdu
Raza, Raja, and Maratib (2023)	2023	Transformer-based encoder–decoder approach	Urdu

Table 2
Summary of the research studies in the domain of abstractive meeting summarization in different languages.

Researcher and reference	Year	Architecture	Language
Li et al. (2019)	2019	Multi-modal hierarchical attention mechanism	English
Singhal et al. (2020)	2020	Transformer model based abstractive summarization	English
Motilal Lodhi et al. (2022)	2022	Machine learning model	Hindi
Nedoluzhko et al. (2022)	2022	Evaluation of different extractive and abstractive summarization methods	English and Czech
Hu et al. (2023)	2023	Comparative analysis of various extractive and abstractive summarization systems	English

After the arrival of COVID-19, the need to implement abstractive text summarization in the context of meeting summarization had increased. There had been observed a major shift of many organizations towards remote jobs, due to which virtual meetings had become a crucial mode of communication. However, with meetings happening every single day, there arised a need to extract useful information from that bulk amount of meeting data produced. In order to fulfill the today's need of abstractive meeting summarization, a significant advancement in this domain had been observed as the researchers' interest in this area had started to rise. Meeting summarization is different from text summarization because text summarization often deals with well-structured documents where information is presented in paragraphs, sections, or chapters. The structure is relatively predictable, facilitating the identification of key sentences and passages. Whereas, meetings are structurally complex, with interactions among multiple participants, interruptions, and varied discourse patterns. Summarizing meetings requires handling the dynamic nature of spoken language and the nuances of conversational exchanges. In the early years of 21st century, prominent meeting summarization corpora were the AMI Meeting Corpus and the ICSI Meeting Corpus. The AMI Meeting Corpus (Mccowan et al., 2005) contained a compilation of 100 h of meeting discussions, including both natural and scenario-based meetings. This valuable open-source corpus incorporated audio and video recordings, transcripts, and an extensive array of annotations including topic segmentation, dialogue acts, named entities, as well as extractive and abstractive summaries. Contrasting this, the ICSI corpus (Janin et al., 2003) consisted of 70 h of English language meetings in computer science domain. With durations spanning 17 to 103 min, these meetings contained participants ranging from 3 to 10 individuals, including non-native English speakers having diverse fluency levels.

An abstractive meeting summarizer (Li, Zhang, Ji, & Radke, 2019) was introduced in which a multi-modal hierarchical attention mechanism was proposed by jointly modeling topic segmentation and summarization. They incorporated both audios and videos of meeting recordings from AMI Meeting Corpus (Mccowan et al., 2005). In another research (Singhal, Khatter, Tejaswini, & Jayashree, 2020) abstractive summarization method was used to train model for dialog systems. In 2022, another study made a business meeting summarizer (Motilal Lodhi, Kharche, Kambri, & Saleem Khan, 2022) using machine learning in order to summarize business meetings held in their regional or professional languages. One of the first meeting minuting corpus, ELITR (Nedoluzhko, Singh, Hledíková, Ghosal, & Bojar, 2022) was introduced in 2021 which provided a comprehensive collection of English and Czech technical project meetings, featuring ASR-generated transcripts, manually corrected versions, and independently generated minutes. This corpus included 120 English and 59 Czech meetings, contributing around 180 h of meeting content. In research, performance analysis of various summarization is also conducted on the English meetings. Recently, MeetingBank (Hu et al., 2023), a new benchmark corpus created from the city council meetings collected from 50 major U.S. cities over the past decade, was presented. They used an innovative divide-and-conquer approach and provided a thorough analysis of data on various extractive and abstractive baselines. A summary of the key studies in abstractive meeting summarization is depicted in Table 2. Table 3 shows a comparison of all the existing meeting corpora with our CLE Meeting Corpus.

Majority research work available in literature is on languages other than Urdu. Lack of research in the domain of summarization for Urdu language had been observed. In the last five years, one of the research (Bhatti & Aslam, 2019) discussed the challenges and complexities faced while working with the Urdu language. The paper highlighted that Urdu inherited a lot of vocabulary from Arabic, Persian, and other native languages of South Asia. Additionally, the language was under-resourced in terms of available computational resources. The paper primarily focused on the task of de-summarization, which involved increasing the length of the document and explaining the substantial points of the text. Another research (Ali, 2021) proposed a method for producing high-quality summaries for Urdu roman text. The approach used the fuzzy logic model to generate fuzzy rules with uncertain property

Table 3
Comparison of all the existing meeting corpora with our CLE Meeting Corpus.

Corpus	Year	No. of meetings	Average duration of meetings	Average No. of speakers	Average No. of summaries	Meeting format	Language
ICSI	2003	59	1 h	6.2	Single	In-person	English
AMI	2005	137	30–40 min	4.0	Single	In-person	English
QMSum	2021	232	1 h	9.2	Single	In-person	English
ELITR (English)	2022	120	1 h	5.9	Multiple	Virtual	English
ELITR (Czech)	2022	59	1 h	7.6	Multiple	Virtual	Czech
AMC	2023	654	15–30 min	2.5	Single	In-person	Mandarin
MeetingBank	2023	1366	2.6 h	8.9	Single	In-person	English
CLE (ours)	2024	**240 (Work in progress)**	30–40 min	3.8	**Multiple**	**Virtual**	Urdu

Table 4
Summary of the corpuses available for Urdu summarization.

Corpus	Year	Domain	Summarization technique	Summary length	Open	Summaries per article
Urdu Summary Corpus (USC) (Humayoun, Nawab, Uzair, Aslam, & Farzand, 2016)	2016	News	Abstractive	Multi-liner	✓	Single
Xlsum (Hasan et al., 2021)	2021	News	Abstractive	One-liner	✓	Single
Urdu news dataset (Hussain, Mughal, Ali, Hassan, & Daudpota, 2021)	2021	News	Abstractive	One-liner	✓	Single
Urdu roman language dataset (Ali, 2021)	2021	News	Extractive	Multi-liner	X	Single
Corpus of Urdu extractive summaries (Humayoun & Akhtar, 2022)	2022	News	Extractive	Multi-liner	✓	Multiple

weight to produce an acceptable summary. One of the study (Nawaz et al., 2020) used local weights and global weights based approach to generate extractive summaries in Urdu. Their study concluded that local weights based approach gives better results for extractive summary generation. In 2022, a benchmark corpus for Urdu extractive summaries of news articles (Humayoun & Akhtar, 2022) was introduced. It contained 161 documents having hand-written extractive summaries. Despite this, they also built a supervised learning framework using machine learning models for Urdu extractive summarization. A recent study (Raza et al., 2023) explored the use of encoder–decoder approach for Urdu abstractive summarization by employing a transformer-based model using self-attention mechanism. A summary of the corpora available online for Urdu summarization are shown in Table 4.

From the literature review, it is evident that majority of the research in the field of Urdu summarization focused on extractive methods and that limited work was done in the context of Urdu abstractive meeting summarization, demonstrating that Urdu is far behind than other languages in terms of this research area. Despite all the improvements in text summarization (Widyassari et al., 2022), it is still necessary to focus on abstractive summarization when it comes to meetings in Urdu. This gap demands a system that can generate human-like, effective, and contextual summaries from Urdu recorded meetings. Our research fills this gap by not only introducing the first Urdu meeting corpus with abstractive summaries but also provides a comprehensive comparative analysis of various deep learning models in the area of Urdu abstractive meeting summarization.

2.1. Disposition from the existing work

This study significantly differs from the existing studies, Hu et al. (2023) and Nedoluzhko et al. (2022), on meeting summarization that appear to be related to this work. The primary difference is that existing studies have been conducted for the languages other than Urdu, whereas this study is conducted for the Urdu. Furthermore, unlike previous efforts, our study not only addresses the unique challenges posed by the Urdu language but also provides detailed guidelines for corpus transcription. These comprehensive annotation guidelines have been made publicly available.

It is noteworthy that the existing summarization endeavors in the Urdu language predominantly concentrate on news articles, with no prior exploration in the context of meetings. This study fills a crucial gap in the literature by delving into the distinctive requirements and challenges of meeting summarization within the Urdu linguistic landscape. Meeting summarization differs fundamentally from text summarization, which typically deals with well-structured documents featuring predictable information presentation in paragraphs, sections, or chapters. The structural complexity of meetings, characterized by interactions among multiple participants, interruptions, and varied discourse patterns, necessitates a unique approach. Summarizing meetings requires adeptly handling the dynamic nature of spoken language, capturing the nuances of conversational exchanges to distill key insights effectively.

3. CLE Meeting Corpus

In this section, process for design, collection and preparation of CLE Meeting Corpus is discussed. An overview of the methodology for generating meeting corpus is shown in Fig. 1. There are four key steps involved in the corpus development i.e. corpus design, corpus recording, corpus transcription and summary generation.

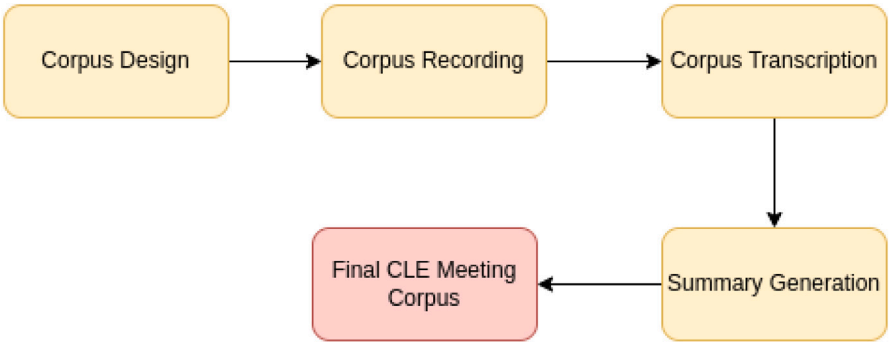


Fig. 1. An illustration of the overall protocol for the generation of CLE Meeting Corpus.

3.1. Corpus design

The design of CLE Meeting Corpus involves both corpus planning and script design. Corpus planning phase is focused on selection of meeting domains, scenarios and agendas in each domain, speaker demographics and online meeting platforms. Script design phase focused on the planning and guidelines for writing meeting script.

3.1.1. Corpus planning

In order to construct a comprehensive corpus, different domains, scenarios, and agendas were planned to be included. Specifically, five key domains were selected: "hiring", "procurement", "admin affairs", "finance", and "technical". The purpose of selecting these domains was to represent different aspects of the modern organizational activities and to cover an extensive variety of meeting topics. These five domains were specifically chosen because they were the most frequent and common domains seen in almost all sorts of organizations. Irrespective of the industry or sector, each organization is mostly involved in hiring human resources, deals with the purchase of goods and services, maintains administrative matters, controls financial operations and implements different technologies.

There were further scenarios for each domain within which there were three to four agendas. These scenarios were selected such that they mimicked the topics discussed in our own organization. Some of the scenarios were also taken from the internet which were widely observed in other organizations as well. For example, the domain "Hiring" was further categorized as "Intern's hiring" that may further be sub-categorized as "Internship-related requirements for linguists, Internship-related requirements for technical resources, and Summer internships". In addition to scenario-based meetings, it was also decided to include some of the natural meetings which were conducted online by the organizational peers regarding software development and NLP tasks covering the domain of computer science.

In order to keep the speaker demographics under consideration, an age range of 18 to 50 years was decided to be considered. Each of these participants were native Urdu speakers, either speaking Urdu as their primary or secondary language. It was also made sure to include participants from different educational backgrounds such as undergraduate, graduate, masters and PhD holders, the details of which would be discussed in a later section.

Zoom, Google Meet, and Microsoft Teams are the three most commonly used online meeting platforms in Pakistan. Therefore, it was planned to hold online sessions for corpus recording using these well-known platforms. This allowed the participants to participate in the meetings from the comfort of their homes. Although several participants joined each meeting remotely from their homes, some of them opted to join the meetings from the office premises. In all of these cases, participants connected to the meetings using their own laptops, mobile phones or PCs with their own headsets or hands-free.

3.1.2. Script design

Before meeting recording, script of each meeting was designed and it was ensured to write the script in dialogue format in Urdu. The script was aimed to be made as natural as possible by gathering relevant information about the specific agenda of the meeting. For this purpose, a proper discussion was carried out with the team members, managers, Human Resource (HR), and other project team members who were involved in real-time meetings related to a specific topic. Meeting script was mainly focused on the formal discussions but some informal greetings and chit chat was also decided to be included in the script to make it more natural. It was also instructed to make sure that the content of each meeting was different from the other meetings every single time when the meeting script was written.

During the writing of the scripts, it was essential to include speaker names to help participants during the meeting recording process. It was decided because this could help participants to quickly recognize their own lines when reading the meeting script because each line was explicitly attributed to the appropriate speaker. This ensured an easy recording process, enabling participants to effortlessly recognize and deliver their designated lines.

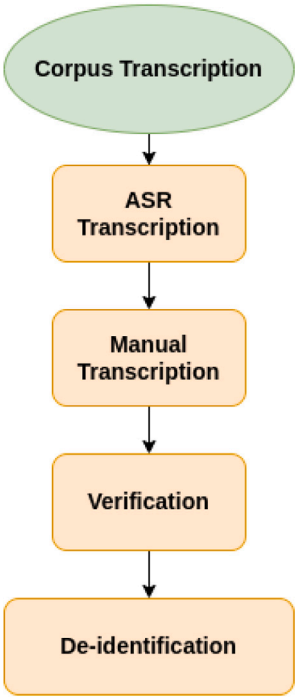


Fig. 2. An overview of the complete process of corpus transcription.

3.2. Corpus recording

Prior to the recording, all participants agreed to take part in meetings for the corpus generation purpose by signing a consent form. As it is obvious that corpus recording requires minimal training for participants, so for that before every meeting which was recorded, the script of the meeting was shared with all the participants. Sharing this script just served to give participants an idea of how the conversation would proceed. Participants were advised that they could add their own points if they wanted; the script was only meant to be a guide. It was decided to have at least 3 to 4 participants in each meeting. The recordings were saved in the wav and mp4 formats for later processing after being downloaded from the platform which was used for recording.

3.3. Corpus transcription

The four key steps of the corpus transcription process are depicted in Fig. 2. First, an initial transcript was generated by using ASR system (Khan et al., 2021) to convert audio recordings into text. The second stage was manual correction of transcript which was done by the transcribers due to probable errors in ASR output. Subsequently, a verification process was then carried out to further filter out any mistakes in the manual transcriptions. Last but not the least, a de-identification method was used to ensure confidentiality by removing any information that could be used to identify participants or projects.

3.3.1. ASR transcription

During the ASR transcription stage, the available ASR system (Khan et al., 2021) was used to generate initial transcript of the meeting recordings. It is important to note that the generated transcript needed further refinement because meetings frequently included domain-specific terminologies and jargon that was not in the ASR system's lexicon. Additionally, the ASR did not do speaker diarization. Due to these limitations, manual correction of transcript was required in order to do the post-processing necessary to ensure the correctness of the final transcript. The ASR-generated transcription therefore served as a starting point for the subsequent manual transcription step, where skilled transcribers reviewed and corrected the text, incorporating domain-specific terms to produce a high-quality transcript ready to undergo further analysis.

3.3.2. Manual transcription

The manual transcription phase involved the use of XTrans software[2] for transcribing audio data. Skilled transcribers employed XTrans to capture spoken content, including speech, speaker identification, speaker gender, and the assignment of specific speaker

[2] https://api.semanticscholar.org/CorpusID:37869734.

Table 5
Transcription guidelines for doing segmentation, handling overlapping and performing word level transcription of recorded meetings on XTrans software.

Segmentation:	1. Make sure that the length of each segment is not more than 15 s. 2. Distinguish speech segments in a syntactic way i.e. mark segments with the same syntactic structure as one. 3. Mark the segment boundary when the speaker takes a pause during his utterance i.e. Distinguish the sentence boundaries in spontaneous speech in a prosodic way. 4. Create new segments whenever a new speaker is identified in order to incorporate speaker diarization, ensuring that each speaker's utterances are transcribed separately.
Overlapping:	1. Mark the segment as overlapping if person A is making sounds (hmmm, ahem, etc.) while person B is speaking. Mark them as separate speakers if these sounds can be separated.
Word level transcription:	1. Use transliterated word dictionary whenever an English word needs to be annotated. 2. Use "+" as a mispronunciation marker whenever a word is mispronounced. For example 'کرائیٹیریے+', here the word 'criteria' is mispronounced. 3. Use "–" as a partial speech marker whenever a speaker wants to utter a word but could not speak completely i.e. when the word is pronounced incompletely by the speaker. For example 'لیپ-', here the word "laptop" is pronounced partially. 4. Transcribe what is precisely heard in the audio without attempting to make corrections or assuming correctness. 5. Do not add words which are not present in the audio, and do not delete spoken words which are grammatically incorrect. 6. Do not normalize dialectal words or attempt to transcribe accent features; instead, use standard orthography. 7. Use (()) for clarification in cases where words are challenging to understand. 8. Enclose person names in "{ }" and project names in "[ ]".

Table 6
Additional transcription guidelines for using vocal tags and handling year and dates while transcribing recorded meetings data on XTrans software.

Vocal tags:	1. Mark the segment as overlapping in cases where person A is laughing while person B is talking. If both person A and B are laughing during a segment, mark it as either overlapping or leave it unmarked. 2. Tag the laughter only when the same person is speaking and laughing simultaneously. 3. Mark a segment as a cough if person A coughs while speaking. On the other hand, if person A is speaking and person B coughs, mark the segment as overlapping. 4. Use special vocal tags to describe sounds made using the mouth or nose, which lack standard lexical representations. These tags, such as <laugh/>, <cough/>, <breath/>, <background> </background>, serve to annotate and identify specific non-verbal sounds encountered in the meeting scripts. 5. Use hesitation marks to denote hesitation in the speaker's speech.
Year and dates:	1. Write all the digits in Urdu, including years pronounced in Urdu. 2. Write standard Urdu counting shared with all of the transcribers. 3. Write the statements completely in Urdu which contain phrases such as 1960s etc.

IDs. To facilitate Urdu lettering, transcribers used the Center for Research in Urdu Language Processing (CRULP) Urdu Phonetic[3] Keyboard.

Given the common practice of Urdu-English code mixing in Pakistani meetings, it was decided to transcribe English words in transliterated form. To ensure consistency and prevent multiple orthographies, a transliterated word dictionary was meticulously maintained (see Tables 5 and 6).

3.3.3. Verification

During the verification step, qualified linguists meticulously reviewed and verified the manually transcribed meetings. A comprehensive 100 percent review, encompassing all meetings, was conducted to minimize human error. The guidelines provided to the reviewers mirrored those given to the transcribers initially.

Reviewers diligently corrected any errors that may have been inadvertently overlooked by the transcribers, contributing significantly to the overall enhancement of corpus quality. This meticulous verification process ensured the accuracy and reliability of the transcribed content.

3.3.4. De-identification

Upon the completion of the comprehensive verification of meeting transcriptions, a pivotal de-identification step was implemented. Following specific guidelines to protect the privacy and confidentiality of individuals featured in the meeting recordings, a systematic approach was employed.

In this process, names of individuals, whether actively participating speakers or not, enclosed within curly brackets "{}", were replaced with generic identifiers such as "person01". This measure ensured anonymity while maintaining the integrity of the transcript. Similarly, project names, encapsulated within square brackets "[]", were substituted with generic labels like "project01". This meticulous approach guaranteed the preservation of contextual discussions while safeguarding sensitive information related to the projects discussed.

[3] https://www.cle.org.pk/software/localization/keyboards/CRULPphonetickbv1.1.html.

Table 7
Corpus Statistics showing the count of meetings for both the scenario-based and the natural meetings. Additionally, the distribution of each domain in scenario-based meetings is also mentioned. In the end, the total number of meetings in CLE Meeting Corpus is shown.

Category	Domain	No. of meetings
Scenario-based	Hiring	30
	Procurement	30
	Admin affairs	07
	Finance	07
	Technical	136
Natural	Computer science	30
Total		240

3.4. Summary generation

In order to capture the key points and the main content of the meetings, manual summary writing of the meeting transcript was done. Following were the guidelines for summary generation:

- The main points should be covered.
- There was no particular word or sentence limit for a summary. It was dependant upon the discussion and the major points that were necessary to be added in a summary.
- Usually, it should cover around ten percent of the whole meeting. For example, if there were 80 utterances in a meeting, the summary should be of 8 to 10 utterances depending upon the important content to be covered.

The summarization process was made diverse by obtaining three distinct summaries for each meeting, written by three different individuals. A total of 9 annotators were involved in writing summaries of all meetings. These annotators were native Urdu speakers. The instructions given to each annotator were uniform, but it was ensured that the generated summaries should not be identical and should capture different viewpoints and aspects of the discussions since each individual had their own distinct understanding and interpretation of the topic.

3.5. Final CLE meeting corpus

In this study, a comprehensive corpus of Urdu meetings has been compiled providing an important resource for analysis of summarization models on meetings. The corpus covered diverse domains such as hiring, procurement, admin affairs, finance, and technical domain ensuring a complete representation of organizational discussions. A snippet of a sample meeting is shown in Fig. 3.

Specifically, a total of 30 meetings had been conducted on hiring, 30 meetings on procurement, 7 meetings on admin affairs, 7 meetings on finance and 136 meetings on technical domain. This thorough methodology ensured that a wide range of actual meeting scenarios were included, which assisted in the design and evaluation of effective techniques to generate abstractive summaries for utilization in various organizational contexts specifically for generating meeting minutes. The distribution of overall corpus is shown in Table 7. The detailed distribution of corpus is illustrated in Table 8.

In a series of meetings conducted, a total of 98 participants were involved including 55 males and 43 females, with each meeting having a maximum number of 8 participants and a minimum of 2 participants. The distribution of the number of participants across these meetings is depicted in Table 9 which shows a breakdown of the number of participants and the corresponding count of the meetings conducted. This table shows that most meetings had a reasonable number of participants, with a significant proportion falling under the category of three and four participants. Specifically, 106 meetings had three participants, making it the most frequent occurrence among all the recorded meetings. Similar to the meetings with three participants, 87 meetings had four participants. On the other hand, a relatively smaller number of meetings accommodated the two, six and eight number of participants.

Another Table 10 provides valuable insights into the speaker variation within the corpus, showcasing the distribution of participants across different age ranges, along with their respective education levels. The youngest age range, 18 to 22, had the highest number of participants, with a count of fifty eight individuals. These participants were studying at the undergraduate level of education. The next age range, 22 to 24, had twenty five participants. This group consisted of individuals pursuing their graduate studies. Moving further in age, the 24 to 26 range had ten participants, and they possessed a master's level of education. Lastly, there were five participants in the "26 and more" age range who holded a phd degree. This data of Table 10 highlights the diversity of participants in terms of age and education, who were contributing to the corpus.

Table 11 provides insights into the distribution of meeting durations and their corresponding frequency. Analyzing this table reveals that the shortest meeting durations (06 to 15 min) were observed in 5 meetings. Next, 23 meetings were having duration 16 to 25 min. The most common meeting duration category was between 26 and 35 min, encompassing a total of 146 meetings. Additionally, 54 meetings had durations falling within the range of 36 to 45 min, representing another substantial portion of the recorded meetings. Notably, longer meetings were relatively less common, with only 12 meetings having durations in the range from 46 to 100 min.

Speaker1	السلام علیکم
Speaker2	وعلیکم السلام :م: جی ماورا انٹرنز ہائرنگ کا کیا سٹیٹس ہے کتنی ایپلیکیشنز کنسیڈر کی ہیں ایس این جی کے لیے آپ نے
Speaker1	:م: چار سی ویز ہیں ہمارے پاس جو آپ نے کہا تھا ایس این جی کے لیے ناں تو ان میں سے دو کو تو میں نے کال کی ہے آج اور باقی جو دو تین لوگ ہیں ان کو بھی ایک دو دن تک کر دوں گی
	ان میں سے فاطمہ اور عائشہ کو میں بلا لیتی ہوں آج تو آپ انٹرویو کر لیجیے گا
	جو باقی دو گرلز ہیں ان میں سے ایک کہہ رہی تھی کہ وہ پارٹ ٹائم کے لیے اویلیبل ہوں گی
Speaker2	نہیں پارٹ ٹائم کے لیے تو ہم ہائر ہی نہیں کر رہے ناں پارٹ ٹائم کا کیا فائدہ فل ٹائم چاہیے ہمیں تو
Speaker1	ہاں جی میں نے ان کو بتایا تھا ویسے کہ ہم اس پروجیکٹ کے لیے پارٹ ٹائم نہیں لاؤ کریں گے
Speaker2	پھر
Speaker1	تو کہتی کہ اچھا چلیں میں پھر آپ کو انفارم کروں گی کیونکہ وہ کہہ رہی تھی کہ وہ شاید اکیڈمی ٹیچنگ کرتی ہے تو وہ مارننگ میں اویلیبل ہوا کرے گی
Speaker2	نہیں کہنا تھا کہ فل ٹائم اویلیبلٹی ہے تو بتائیں
Speaker1	صحیح اچھا مس اب ان کی جو سیلری رینج ہے وہ کتنی ہوگی مطلب ایس این جی اور براڈ کاسٹنگ کی سیم سیلری ہے
Speaker2	اچھا یہ دیکھو ایسا ہے کہ جو ایس این جی لینوٹیشن ہے ٹونٹی فائیو منٹس کا ٹارگٹ ہے جن کا ان کو ٹونٹی تھاؤزنڈ آفر کرتے ہیں اور ٹونٹی منٹس پہ سکیئن

Reference Summary:

میٹنگ میں نیو انٹرنز کی ہائرنگ کے حوالے سے ڈسکشن کی گئی ہے۔ انٹرنشپ کے لیے جو جو سکلز ریکوائرڈ ہیں ان پر ڈسکشن کی گئی ہے۔ اس کے علاوہ انٹرنز کو جو اویلوئیشن ٹاسکس اسائن کیے جائیں گے ان پر بات کی گئی ہے اور سیلری ڈسکس کی گئی ہے۔ جاب ریکوائرمنٹس میں اردو ٹائپنگ سے واقفیت، ورڈ ریکگنیشن، ڈیٹا ویریفکیشن، لینوٹیشن، ڈیٹا انٹری، ڈیٹا اپ لوڈنگ اور مینجمنٹ، ایم ایس ایکسل اور فائل ہینڈلنگ سکلز کا ہونا ضروری ہے۔ اور آخر میں نیکسٹ میٹنگ میں پراجیکٹ ٹو یعنی براڈ کاسٹنگ کے لیے انٹرنز ہائرنگ، سمر انٹرنشپس اور کینڈیڈیٹ شارٹ لسٹنگ کے حوالے سے بات کی جائے گی۔

Fig. 3. A sample of a transcript snippet of our CLE Meeting Corpus for a starting segment of a meeting. This is a conversation between a manager and HR regarding Intern's hiring status. A reference summary generated by the annotator is shown in the bottom. (Translation of reference summary : In the meeting, a discussion has been held regarding the hiring of new interns. The skills required for internship are discussed. Apart from this, the evaluation tasks to be assigned to the interns are discussed and the salary discussed. Job requirements include familiarity with Urdu typing, word recognition, data verification, annotation, data entry, data uploading and management, MS Excel and file handling skills. And finally, in the next meeting, project two, i.e. hiring of interns for broadcasting, summer internships and candidate shortlisting will be discussed.)

While analyzing important aspects of the meeting durations, it was discovered that the average meeting lasted approximately 33 min. The total sum of all the meeting durations amounted to 7906 min, reflecting the collective time. The maximum meeting duration was 95 min, while the minimum was only 6 min, illustrating the considerable variation in meeting lengths.

Employing a quantitative analysis of the whole corpus, Fig. 4 represents the relationship between the number of utterances in the meeting and the corresponding number of meetings falling within specific line count ranges. The x-axis of Fig. 4 shows the range of utterances in the meetings, such as less than 200, 200 to 300, 300 to 400, 400 to 500 and so on, while the y-axis represents the count of meetings falling within each utterance count range. From Fig. 4, it can be observed that the maximum number of meetings,i.e., 55 in total had utterances from 400 to 500 and the minimum number of meetings had utterances greater than 1000 with a count of only 4 meetings. Another type of analysis is shown in the column chart Fig. 5 which provides a comprehensive relationship between number of summary utterances and the corresponding number of meetings within the specific line count ranges.

3.6. Transcriber details

A team of 5 qualified transcribers, possessing expertise in linguistics and native Urdu, was engaged in the manual transcription process for our research. Furthermore, during the verification step, 4 reviewers, who were also expert linguists, were involved. In total, the entire corpus transcription process was carried out by 8 linguists. Prior training had been undergone by each

Table 8
Detailed distribution of CLE Meeting Corpus across all the five domains. It also illustrates the duration of meetings in minutes in each scenario of a specific domain.

Domain	Scenario	Duration in Minutes
Hiring	Job Description (JD)	081 min
	Interns' Hiring (IH)	091 min
	Cross Project Hiring (CPH)	101 min
	Full Time Project Staff Hiring (FTPS)	113 min
	Hiring ad Development (HAD)	104 min
	Pre Hiring Project Team Planning (PTP)	101 min
	Post Evaluation Performance Discussion (EPD)	105 min
	Project Discussion (PD)	111 min
	Remote Hiring (RH)	099 min
	Interviews' Planning (IP)	098 min
Procurement	Call for Proposal (CP)	103 min
	Request for Purchase (RP)	109 min
	Tender (T)	096 min
	Budget Discussion (BD)	102 min
	Purchase Order (PO)	094 min
	Equipment Delivery (ED)	096 min
	Invoice (Inv)/Billing Payment (BP)	096 min
	Complaints (C)	099 min
	Equipments' Warranty (EW)	092 min
	Demand for New Equipments (DNE)	096 min
Admin affairs	HR Hiring (HH)	072 min
	Budget Officer Hiring (BOH)	067 min
	Salary Discussion for Admins (SDA)	069 min
	Job Appraisal (JA)	036 min
Finance	Bank Accounts (BA)	099 min
	Contingency Bills (CB)	074 min
	Budget Review and Analysis (BR&A)	100 min
Technical	Data Processing (DP)	1041 min
	Requirements (R)	313 min
	System Design (SD)	835 min
	Development (D)	390 min
	Testing (T)	760 min
	Deployment (D)	487 min

Table 9
Count of the meetings having a specific number of participants.

No. of participants	No. of meetings
2	04
3	106
4	87
5	31
6	11
8	01

Table 10
Variation of the participants in terms of age and their education.

Age range	Count of participants	Education
18–22	58	Undergrad
22–24	25	Grad
24–26	10	Masters
26 and more	05	PhD

transcriber, ensuring a high level of expertise in script annotation. This collaborative effort of skilled professionals contributed to the strengthened accuracy and reliability of the transcription process.

3.7. Ethical consideration

The recognition of the imperative to uphold ethical norms and preserve the privacy and confidentiality of individuals engaged in meetings led to the implementation of stringent ethical practices. Prior to their participation in the recordings, speaker agreement was obtained, and all participants completed speaker consent forms. A thorough explanation of the study's goals, the confidentiality

Table 11
Number of meetings lying in a specific range of meeting duration. The range for meeting duration is in minutes.

Meeting duration	No. of meetings
06–15 min	05
16–25 min	23
26–35 min	146
36–45 min	54
46–100 min	12

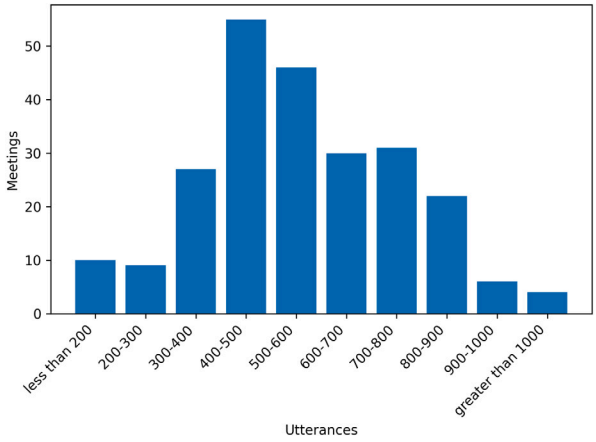


Fig. 4. A distribution of the number of meetings lying in each of the specific range of meeting utterances.

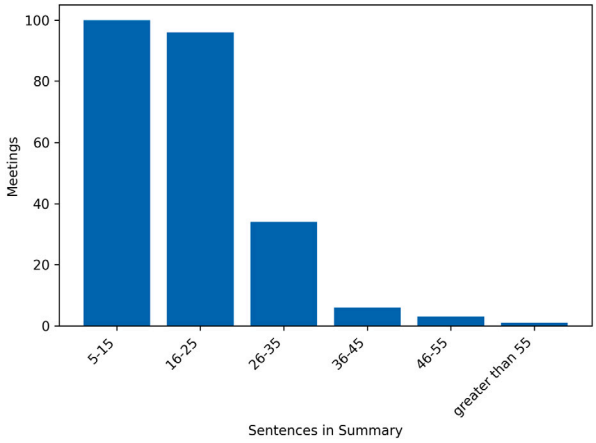


Fig. 5. A plot showing the distribution of number of meetings containing specific range of number of sentences in reference summaries of our meeting corpus.

mechanisms employed, and the potential benefits and impacts of the research was provided to participants in advance. Assurances were given that their contributions would be strictly utilized for research purposes, and their identities would be kept confidential.

In addition to obtaining speaker consent, a comprehensive de-identification process was employed to safeguard the secrecy of participants. This approach not only upheld the privacy rights of the participants but also minimized the risks associated with the disclosure of sensitive or confidential data.

4. Experimentation

This section serves as the starting point for the exploration of experiments conducted using a range of techniques, including various deep learning models, graph-based approaches, and keyword extraction algorithms. Within this section, the employed models, utilized corpora, followed methodologies, and obtained results are comprehensively discussed. The outcomes of these experiments undergo thorough evaluation and analysis, providing a comprehensive understanding of the effectiveness and implications of the applied methodologies.

Table 12
Comprehensive overview of the base model specifications.

Model	Base model	Transformer	No. of layers	No. of heads	Hidden size
mT5	T5	Text-to-Text Transfer Transformer	24	16	1024
mBART	BART	Bidirectional and auto-regressive transformer	12	16	1024
RoBERTa	BERT	Bidirectional encoder representations from transformer	12	12	768

4.1. Models

In this study, three state-of-the-art deep learning models for abstractive text summarization are employed, namely mT5 (multilingual T5) (Xue et al., 2020), mBART (multilingual BART) (Liu et al., 2020) and RoBERTa (Liu et al., 2019). Each model is fine-tuned on a summary corpus, the details of which are given in the next subsection.

For automatic summarization systems, the Urdu CLE Meeting Corpus's inherent code-mixing issue presents a serious challenge. This phenomenon, where speakers alternate between English and Urdu throughout a conversation adds complexity to standard language processing tasks. To effectively address this complexity, models with built-in multilingual capabilities are given priority during the model selection process. Although not specifically created for Urdu, these selected models show different levels of support for multilingual data, which is used to handle the unique characteristics of code-mixed Urdu conversation observed in the CLE Meeting Corpus.

The recent "Text-to-Text Transfer Transformer" (T5) (Mastropaolo et al., 2021) employed a unified text-to-text format and scale to attain state-of-the-art results on a wide variety of English language NLP tasks. mT5 (Xue et al., 2020), a multilingual variant of T5, is pre-trained on a new common crawl-based corpus covering 101 languages and emerged as a powerful tool for various natural language processing tasks, including abstractive text summarization. Although there is no official pre-trained mT5 model specifically for Urdu at the time, the framework can be adapted for Urdu summarization through fine-tuning.

Moreover, mBART (Liu et al., 2020), another transformer-based model designed for multilingual tasks, is employed. After fine-tuning mBART on the Urdu summarization corpus, the model could be exhibited to produce high-quality abstractive summaries. Another model, RoBERTa (Liu et al., 2019), is a pre-trained transformer model developed by Facebook AI, similar in many ways to BERT (Devlin, Chang, Lee, & Toutanova, 2019), a pre-trained transformer model developed by Google. These models are designed to perform well on a variety of NLP applications. However, RoBERTa is designed to be a more powerful and computationally efficient version of BERT, with significantly more parameters and training data. It is pre-trained on a massive amount of text data and fine-tuned for specific tasks to achieve good results on a variety of NLP benchmarks. Table 12 shows a comprehensive overview of the base model specifications and related details for each model.

In addition to these three deep learning models, three of the extractive techniques have also been employed which includes TextRank (Zieve et al., 2023), LexRank (Dalal, Singhal, & Lall, 2023) and RAKE (Rapid Automatic Keyword Extraction) (Huang, Wang, & Wang, 2020). TextRank and LexRank are both unsupervised graph-based ranking models for automatic text summarization. RAKE, on the other hand, is a domain-independent algorithm of keyword extraction which uses word frequency and co-occurrence in order to identify the keywords.

Additionally, GPT-3.5 (Zhang, Dong, Xiao, & Oyamada, 2023) is also utilized with prompting. By giving a prompt of "write abstractive summary of this text in Urdu", GPT-3.5 generates a summary on the basis of its own understanding of the text (Goyal, Li, & Durrett, 2023).

4.2. Dataset

This experiment utilized two distinct corpora: the news corpus (XL-Sum) and the meetings corpus (CLE Meeting Corpus). XL-Sum (Hasan et al., 2021) is an extensive corpus featuring over 1 million article-summary pairs sourced from the BBC in 44 different languages, spanning from low to high-resource languages. Notably, this corpus stands as the largest abstractive summarization corpus, both in terms of the number of samples collected from a single source and the number of languages covered. The second category of the news corpus includes 50 news articles (Humayoun & Akhtar, 2022), released for the study of supervised learning in Urdu extractive summarization (Muhammad, Jazeb, Martinez-Enriquez, & Sikander, 2018).

The second type of data, the CLE Meeting Corpus, is categorized into two sections. The first category encompasses 147 scripted meetings, while the second category comprises 230 recorded meetings.

Three distinct types of experiments were conducted: zero-shot, news fine-tuned, and meetings fine-tuned. In the zero-shot experiment, models were evaluated without any fine-tuning, eliminating the need for training on a specific corpus. For the news fine-tuned experiment, models were trained on a combination of 67k instances from the XL-Sum corpus and an additional 50 news articles. In the meetings fine-tuned experiment, models were trained on a broader corpus, which included not only news data but also 147 scripted meetings and 230 recorded meetings. In each case, 10% of the training set was allocated for validation. The performance of all models was evaluated on a test dataset comprising 10 recorded meetings.

4.3. Implementation

The experimentation begins with a zero-shot approach, employing pre-trained models including GPT-3.5 without additional training or fine-tuning. Following this, the pre-trained mT5-small model undergoes fine-tuning on the XL-Sum corpus for Urdu.

Table 13
Parameters setting for fine-tuning of mT5, mBART and RoBERTa.

Parameter	mT5	mBART	RoBERTa
Per_device_train_batch_size	1	1	1
Per_device_eval_batch_size	1	1	1
Num_train_epochs	10	10	10
Logging_steps	100	1000	1000
Save_steps	500	500	1000
Evaluation_strategy	Epoch	Steps	No
Eval_steps	100	300	1000
Weight_decay	0.01	0.01	0.00
Fp16	True	True	True
Learning_rate	5e−04	2e−05	5e−05
Lr_scheduler_type	Linear	Linear	Linear
Label_smoothing_factor	0.1	0.0	0.0
Optimizer	Adamw_torch	Adafactor	Adamw_torch

However, it is observed that, irrespective of the input text length, this fine-tuned model generates only one-line summaries due to the nature of the training corpus.

To address this limitation, a concatenation approach is adopted. The input text is divided into equal parts, and independent summaries are generated for each part. Subsequently, the results from each part are concatenated to form a comprehensive summary.

Further fine-tuning is conducted on a small corpus comprising 50 news articles with summaries longer than one line. This fine-tuning process is applied to the previously fine-tuned mT5-small model, resulting in more extended summaries compared to the initial case. To enhance the performance of the mT5 model, the mT5-base version undergoes fine-tuning on the same corpus using the same procedure. The same fine-tuning procedure is repeated for models mBART-large and RoBERTa-urduhack-small. The reason why mBART-large is used is that for mBART only the mBART-large variant is available online with Urdu support, while smaller variants such as mBART-small or base are not available online with Urdu support for comparison. Therefore, our evaluation reflects the performance of the available mBART variant against mT5-small, mT5-base and RoBERTa-urduhack-small, considering the limitations in the accessibility of different mBART model sizes.

Finally, all the news fine-tuned models undergo an additional round of fine-tuning on the CLE Meeting Corpus to produce the meetings fine-tuned models. This sequential fine-tuning process aims to optimize the models for generating summaries in the context of meetings. The parameters setting for fine-tuning of each of the above models is shown in Table 13. The complete code of implementation for setting up the experiment on mT5,[4] mBART and RoBERTa[5] are available.

5. Results and discussion

To compare the models' performances, the generated summaries are evaluated using standard evaluation metrics such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Ganesan, 2018) including ROUGE-1, ROUGE-2 and ROUGE-L which, when comparing the generated and reference summaries, correspond to the unigram, bigram, and longest common sub-sequence overlaps, respectively. The formula to compute ROUGE metrics is shown as follows:

$$ROUGE\_N = \frac{\sum_{i=1}^{N} \sum_{n\in\text{n-grams}} \text{CountMatch}(n, \text{Reference})}{\sum_{i=1}^{N} \sum_{n\in\text{n-grams}} \text{Count}(n)} \tag{1}$$

Here CountMatch$(n, \text{Reference})$ counts the number of n-grams that match between the generated summary and the reference summary, and Count(n) counts the total number of n-grams in the reference summary.

For example, consider the following reference summary and generated summary.

Reference summary: 'The team discussed the project progress and upcoming milestones'.

Generated summary: 'Project progress and future goals were the main topics of discussion in the meeting'.

For ROUGE-1 calculation, number of unigrams matched = 5 (the, project, progress, and, the)

Total number of unigrams in reference summary = 9

Let us assume that $N = 1$ for simplicity then by putting values in Eq. (1),

ROUGE-1 = 5/9 = 00.5555

[4] https://github.com/csebuetnlp/xl-sum/tree/master/seq2seq.

[5] https://github.com/BareeraSadia/CLE-Abstractive-Meeting-Summarization.

Table 14
Automatic evaluation of extractive methods using ROUGE metric and BLEU_score. Models are evaluated on test data of 10 recorded meetings.

Model name	ROUGE-1	ROUGE-2	ROUGE-L	BLEU_Score
TextRank	22.1191	03.2872	11.8578	00.6652
LexRank	**26.6292**	**06.2595**	**14.3953**	**01.9327**
RAKE	25.4952	04.9746	13.6122	01.8657

Table 15
Automatic evaluation of deep learning models and GPT-3.5 using ROUGE metric and BLEU_score. Models are evaluated on three types of experiments: zero-shot, news fine-tuned and meetings fine-tuned. These results have been computed on test data of 10 recorded meetings.

Model name	Base model	Version	ROUGE-1	ROUGE-2	ROUGE-L	BLEU_Score
Zero-shot	mT5	Small	01.5491	00.0012	01.5693	00.0000
	mT5	Base	03.6174	00.3306	00.3306	00.1081
	mBART	Large	19.6400	03.7000	11.3581	00.8780
	RoBERTa-urduhack	Small	01.5600	00.0668	01.1244	00.0237
	GPT (with prompting)	3.5	22.3501	06.0309	11.1774	01.4171
News Fine-tuned	mT5	Small	06.2191	01.3333	03.1985	00.0114
	mT5	Base	20.4054	04.1058	12.0013	00.7528
	mBART	Large	21.1469	03.6000	09.8100	01.3504
	RoBERTa-urduhack	Small	05.0914	00.0138	03.2561	00.0251
Meetings Fine-tuned	ur_mT5	Small	28.9883	10.8670	**17.6608**	03.4163
	ur_mT5	Base	**31.7262**	**11.1186**	17.2454	**04.1586**
	ur_mBART	Large	31.1564	10.5306	16.6504	03.9430
	ur_RoBERTa-urduhack	Small	10.1123	00.0401	03.8421	00.0312

Bilingual Evaluation Understudy (BLEU) measure (Saadany & Orăsan, 2021) is also utilized which measures how closely the generated text aligns with the reference summary. The BLEU metric can be computed using this formula:

$$BLEU = \mathrm{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2}$$

Here BP is the brevity penalty, p_n is the precision for n-grams, and w_n are the weights assigned to each n-gram precision. For better understanding, consider the above example of reference and generated summary again.

Number of unigrams matched = 5 (the, project, progress, and, the)

Total number of unigrams in generated summary = 14

Precision for unigrams = p_1 = 5/14 = 00.3571

Now, let us assume that BP = 1, $w_1 = 0.5$ and $N = 1$ for simplicity. By putting values in Eq. (2),

BLEU = 00.7996

Besides using these evaluation metrics, a thorough human evaluation is also conducted to identify the strengths and weaknesses of each model in summarizing Urdu text.

5.1. Automatic evaluation

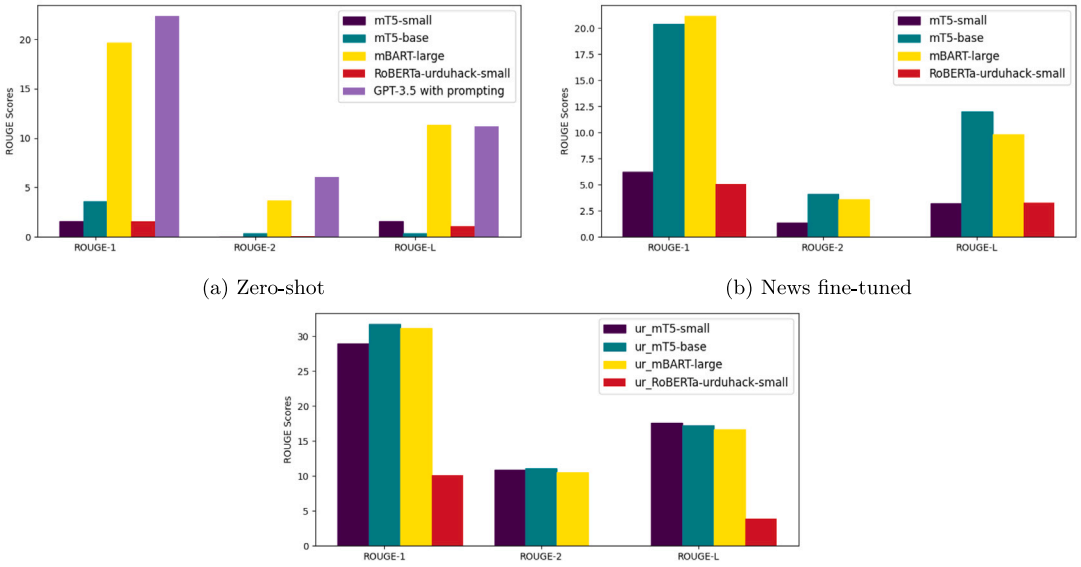
The evaluation of three extractive methods is presented in Table 14. A notable observation from Table 14 is that LexRank consistently outperforms the other two algorithms across all metrics, including ROUGE scores and BLEU_score. According to the results, the most effective technique for meeting extractive summarization is LexRank, with ROUGE-1, ROUGE-2, ROUGE-L, and BLEU_score values of 26.6292, 6.2595, 14.3953, and 1.9327, respectively.

5.1.1. Discussion on zero-shot experiment

The zero-shot experiment involved the evaluation of various models i.e. mT5-small, mT5-base, mBART-large, RoBERTa-urduhack-small, and GPT-3.5 with prompting—for meeting transcript summarization. Remarkably, the mBART-large model outperformed other abstractive models, achieving notable ROUGE scores of 19.64, 3.7, and 11.3581 for ROUGE-1, ROUGE-2, and ROUGE-L, respectively. Although RoBERTa-urduhack-small yielded poor results, GPT-3.5 with prompting excelled, garnering ROUGE-1 at 22.3501, ROUGE-2 at 6.0309, and ROUGE-L at 11.1774. These results are visually represented in Fig. 6(a).

5.1.2. Discussion on news fine-tuned experiment

Building upon the pre-trained models, the news fine-tuned models demonstrated performance improvements compared to their zero-shot counterparts. Among them, mBART-large outperformed with high ROUGE scores of 21.1469, 3.6, and 9.81 for ROUGE-1,



(a) Zero-shot

(b) News fine-tuned

(c) Meetings fine-tuned

Fig. 6. A demonstration of the results in the form of ROUGE-1, ROUGE-2 and ROUGE-L measure of all the models for three distinct type of experiments.

ROUGE-2, and ROUGE-L. The mT5-base model exhibited significant enhancement, particularly excelling in ROUGE-L with a score of 12.0013. Minor improvements were observed for mT5-small and RoBERTa-urduhack-small. The results underscore the benefits of fine-tuning for news summarization tasks, as depicted in Fig. 6(b).

5.1.3. Discussion on meetings fine-tuned experiment

Expanding on the news fine-tuned models, the meetings fine-tuned models showed substantial enhancements for summarizing recorded meetings. Both ur_mT5-base and ur_mBART-large models exhibited significant improvements, with ur_mT5-base outperforming having ROUGE-1, ROUGE-2, and ROUGE-L scores of 31.7262, 11.1186, and 17.2454 respectively. Similarly, ur_mBART-large showcased improved performance with a ROUGE-1 score of 31.1564 and a ROUGE-2 score of 10.5306. While ur_mT5-small demonstrated enhanced performance, ur_RoBERTa-urduhack-small displayed limited improvement, with a ROUGE-1 score of 10.1123, indicating challenges in learning for the summarization task. Visual representation of these results is provided in Fig. 6(c).

Distinct model performances can be attributed to various factors. The effectiveness of ur_mT5-base and ur_mBART-large is attributed to their larger architecture and robust fine-tuning. Moreso, the bidirectional architecture of ur_mBART-large enhances its capabilities in abstractive summarization tasks. The ur_mT5-small model's improved performance in the meetings fine-tuned experiment is attributed to the nature of the test data being meetings. However, ur_RoBERTa-urduhack-small, being a smaller model primarily used for masked language modeling, struggles in the broader context of summarization tasks.

In addition to the comprehensive discussion above, the distinction between extractive and abstractive summarization models is evident in the results presented in Tables 14 and 15. The ROUGE and BLEU measures consistently yield lower results for extractive models, such as TextRank, LexRank, and RAKE. These lower scores suggest that extractive models are comparatively less successful in capturing nuanced details and generating summaries that closely resemble human-like outputs, as opposed to abstractive models, including ur_mT5-small, ur_mT5-base, and ur_mBART-large. This observation underscores the inherent differences in the summarization approaches and the unique strengths of abstractive models in generating more contextually rich and human-like summaries.

5.2. Human evaluation

Recognizing the limitations of ROUGE metrics in providing a comprehensive assessment of summary quality and readability, a manual human evaluation is undertaken. The performance of each summarization model, encompassing fine-tuned abstractive models (ur_mT5-small, ur_mT5-base, ur_mBART-large, and ur_RoBERTa-urduhack-small) and GPT-3.5 with prompting, is assessed to examine the efficacy of the system-generated meeting summaries. Given the limited test data consisting of 10 recorded meetings, a single evaluator is assigned the task of conducting the human evaluation. The evaluator who has done human evaluation is totally not involved in the corpus preparation part. The evaluator reads the transcript and evaluates the quality of each model's generated summary based on five criteria (Fabbri et al., 2021), as detailed in Table 16. The evaluation criteria include informativeness, factuality, fluency, coherence, and redundancy.

5.2.1. Discussion on results of human evaluation

In order to evaluate each criterion, a 5-point Likert scale is utilized. The mapping of each scale value is shown in Table 17. In Table 18, the performance of each summarization model is presented in the form of average scores on test data of 10 recorded

Table 16
Definition of each of the five human evaluation criteria taken from Fabbri et al. (2021).

Criterion	Definition
Informativeness	Whether the generated summary properly covers all and only the significant aspects and main points discussed within the meeting?
Factuality	Whether the data in the generated summary is correct and true to the original content?
Fluency	What is the quality of individual sentence within the generated summary? How well a sentence is constructed, taking into account elements like writing style and grammatical correctness?
Coherence	How well the sentences flow from one sentence to the next sentence? is there any logical connection or transition between sentences?
Redundancy	Whether the summary contains the unnecessary repetitions or duplicate details?

Table 17
Representation of 5-point Likert scale in order to evaluate all the models on the five criteria mentioned in Table 16.

Scale	Representation
1	Totally disagree
2	Moderately disagree
3	Neutral
4	Moderately agree
5	Totally agree

Table 18
Results of human evaluation on each of the five criterion. The scores are averaged on all of the 10 recorded meetings. Last row represents the overall average score on all the five criteria.

Criterion	ur_mT5-small	ur_mT5-base	ur_mBART-large	GPT-3.5	TextRank	LexRank	RAKE
Informativeness	1.2	2.4	**2.9**	2.3	1.1	1.9	1.7
Factuality	2.0	2.6	3.0	3.3	**5.0**	**5.0**	**5.0**
Fluency	**3.5**	3.2	3.2	1.0	1.8	1.9	1.4
Coherence	3.1	3.1	2.6	**3.6**	1.0	1.4	1.0
Redundancy	4.1	4.2	4.3	**4.9**	4.6	**4.9**	4.7

meetings. Out of the five criteria, redundancy is generally seen to be low for all models with the exception of ur_RoBERTa-urduhack-small. GPT-3.5 outperformed other abstractive models with scores of 3.6 and 3.3 for coherence and factuality, respectively. However, GPT-3.5 faces challenges in terms of fluency; it received the lowest score of 1, indicating that it produces text that is not fluent.

ur_mBART-large also exhibits notable performance, excelling particularly in informativeness, factuality and fluency having scores of 2.9, 3.0 and 3.2 respectively. With a fluency score of 3.5, ur_mT5-small demonstrates strong fluency and coherence as compared to all the other models indicating its ability to generate a fluent summary. In case of informativeness and factuality, it indicates area for improvement. ur_mT5-base, on the other hand, performs well in terms of factuality, fluency and coherence with scores of 2.6, 3.2 and 3.1 respectively but needs improvement in informativeness. Overall, ur_RoBERTa-urduhack-small performs poor across all criteria having score 1 in all of them indicating its shortcoming in generating informative, accurate, and cohesive summary in addition to struggling with fluency and redundancy.

In contrast to abstractive models and GPT-3.5, TextRank, LexRank, and RAKE exhibit challenges in terms of informativeness, fluency, and coherence. While these extractive models demonstrate proficiency in preserving factual information and avoiding redundancy, they tend to struggle with the cohesion and natural flow of the generated summaries. The inherent limitation of extractive techniques lies in their extraction of summary lines directly from the original input, resulting in factual but disjointed sentences and phrases within those sentences. This limitation becomes apparent in their relative deficiency in fluency and coherence compared to abstractive techniques. The trade-off between factual content and the coherence of the summary is a notable drawback of extractive methods in meeting summarization tasks.

6. Conclusion

In this paper, we have introduced a benchmark corpus, the CLE Meeting Corpus, specifically tailored for Urdu meeting summarization. Our focus extended to fine-tuning various deep learning summarization models, namely mT5-small, mT5-base, mBART-large, and RoBERTa-urduhack-small, followed by comprehensive testing for abstractive meeting summarization in Urdu. The evaluation of these models involved both automated metrics, including ROUGE score and BLEU_score, and manual assessments using a 5-point Likert scale.

Our research findings highlight that, following fine-tuning on meetings data, both the ur_mT5-base and ur_mBART-large models demonstrated superior performance compared to other models, including GPT-3.5, in automated evaluations. Despite GPT-3.5 performing well in human evaluations, it exhibited limitations in terms of fluency. The nuanced differences observed in various evaluation aspects underscore the importance of considering both automated and human-centric assessments.

Table 19
An extension of the Table 8 representing more detailed distribution of CLE Meeting Corpus across the first domain i.e. hiring. It also illustrates the duration of meetings in minutes for each agenda of a specific scenario.

Domain	Scenario	Agenda	Duration in Minutes
Hiring	Job Description (JD)	Job description for RA technicals	00:28:25 min
		Job description for linguists	00:29:15 min
		Job description for social scientist	00:23:42 min
	Interns' Hiring (IH)	Internship related requirements for linguists	00:26:34 min
		Internship related requirements for technicals	00:32:13 min
		Summer internships	00:32:20 min
	Cross Project Hiring (CPH)	Interns' hiring	00:29:37 min
		Research officer's hiring	00:33:56 min
		Tasks' discussion	00:38:03 min
	Full Time Project Staff Hiring (FTPS)	Pre-hiring interview-selection /Shortlisting of candidate	00:39:33 min
		Post-hiring interview-selection of candidate-salary negotiation	00:36:21 min
		Project's beginning	00:37:54 min
	Hiring ad Development (HAD)	Gain project requirements to post a hiring Ad	00:35:24 min
		Inclusion/Exclusion of job requirements	00:35:17 min
		Remote internships	00:33:32 min
	Pre Hiring Project Team Planning (PTP)	Technicals' team planning	00:37:25 min
		Linguistics' team planning	00:31:12 min
		Interns' team planning	00:32:32 min
	Post Evaluation Performance Discussion (EPD)	Technicals' performance	00:31:37 min
		Linguists' performance	00:37:32 min
		Interns' performance	00:36:04 min

(*continued on next page*)

For researchers and practitioners, our curated CLE Meeting Corpus stands as a valuable resource, offering a foundation for the development of efficient Urdu meeting summarizers. The insights gained from our experiments contribute to the broader understanding of the strengths and limitations of different summarization models, aiding future advancements in this field.

7. Limitations

In introducing the novel meeting corpus for meeting summarization, it is crucial to acknowledge certain limitations. The corpus presented in this paper is constrained to specific administrative domains, namely hiring, procurement, admin affairs, finance and technical domain. The selection of above mentioned domains, was intentionally made to cover a wide range of typical office discourse. Specifically, this research's focus is to cover virtual meetings with a particular emphasis on the Pakistani industry. While it is true that the design rules may not perfectly align with every possible variation in input meeting across different domains, it is believed that the selected approach captures a representative sample of office meeting conversations. Furthermore, it is recognized that pre-processing may be necessary when applying this research's approach to new corpora with distinct characteristics. However, the evaluation framework is designed with flexibility in mind, anticipating the need for adaptation to different contexts. Additionally,this corpus also lacks in annotation of action items, which could provide valuable insights into various meeting aspects related to these administrative factors.

Despite these identified shortcomings, the CLE Meeting Corpus remains a valuable and useful resource for the development of meeting summarization systems and meeting minutes generation. While the current focus is on specific domains, the corpus still offers a solid foundation for research and development in the field of meeting summarization, with potential for future expansions and refinements. Researchers should be mindful of these limitations when considering the applicability of the corpus to their specific use cases.

8. Future work

Our future trajectory is centered on the expansion and enrichment of our existing corpus, with a multifaceted approach. First and foremost, there is a concerted effort to significantly increase the size of the dataset. This expansion aims to provide a more extensive and diverse set of meeting transcripts, contributing to the robustness of our models and their efficacy across various scenarios.

In addition to the quantitative augmentation, our future work involves comprehensive annotation of the corpus against action items. This annotation process will enhance the corpus by explicitly identifying and categorizing action items within the meeting content. Furthermore, our expansion efforts extend to covering a broader array of domains. By incorporating meetings from additional domains, the corpus becomes more representative of diverse organizational contexts, enabling our models to generalize

Table 19 (*continued*).

Domain	Scenario	Agenda	Duration in Minutes
	Project Discussion (PD)	Project details/Budget discussion	00:46:15 min
		Vacant posts/Post Requirement discussion	00:31:51 min
		Documentation of the project	00:33:13 min
	Remote Hiring (RH)	Employee's shortlisting/Salary negotiation	00:37:41 min
		Remote employee's evaluation Performance discussion	00:29:24 min
		Employee's joining discussion	00:32:13 min
	Interviews' Planning (IP)	Call for interview discussion/ Interview schedule	00:31:33 min
		Post interview discussion/ Performance discussion	00:33:09 min
		Appointment planning	00:33:42 min

Table 20
An extension of the Table 8 representing more detailed distribution of CLE Meeting Corpus across the remaining three domains i.e. procurement, admin affairs and finance. It also illustrates that the duration of meetings is in minutes for each agenda of a specific scenario.

Domain	Scenario	Agenda	Duration in minutes
Procurement	Call for Proposal (CP)	Vendors' shortlisting	00:34:45 min
		Vendor's selection	00:36:51 min
		Vendors' replacement	00:31:56 min
	Request for Purchase (RP)	Purchase refusal	00:36:40 min
		Purchase approval	00:37:36 min
		Urgent purchase	00:35:16 min
	Tender (T)	Newspaper Ad	00:33:08 min
		Tender acceptance	00:33:59 min
		Tender rejection	00:29:24 min
	Budget Discussion (BD)	Budget distribution	00:32:55 min
		Request for project funding	00:32:20 min
		Budget request for lab equipments	00:37:08 min
	Purchase Order (PO)	General purchase for electronic items	00:32:54 min
		General purchase for technical items	00:30:29 min
		General purchase for kitchen items	00:31:06 min
	Equipment Delivery (ED)	Date extention	00:31:03 min
		Damaged delivery of electronic items	00:31:32 min
		Damaged delivery of lab furniture	00:33:37 min
	Invoice (Inv)/Billing Payment (BP)	Amount installments	00:31:24 min
		Project payment delay	00:34:09 min
		Budget breakup details	00:31:15 min
	Complaints	Laptop batteries	00:31:16 min
		Electronic items	00:32:08 min
		Computer systems	00:36:20 min
	Equipments' Warranty (EW)	Demand for warranty	00:27:44 min
		Fake warranty	00:33:36 min
		Request for warranty extension	00:30:56 min
	Demand for New Equipments (DNE)	Request for extended RAMs	00:31:40 min
		Request for lab equipments /Microphones	00:33:06 min
		Request for lab furniture	00:31:31 min

(*continued on next page*)

and adapt to a wider range of topics and discussions. Simultaneously, we plan to annotate the corpus against specific topics discussed in the meetings. This annotation will provide a nuanced understanding of the thematic content, enabling the development of summarizers that can discern and emphasize key topics within a meeting.

CRediT authorship contribution statement

Bareera Sadia: Validation, Software. **Farah Adeeba:** Methodology, Formal analysis, Data curation, Conceptualization. **Sana Shams:** Supervision, Project administration. **Kashif Javed:** Investigation.

Data availability

Data will be made available on request.

Table 20 (*continued*).

Domain	Scenario	Agenda	Duration in minutes
Admin affairs ⌒	HR Hiring (HH)	HR shortlisting	00:33:07 min
		HR recruitement	00:39:51 min
	Budget officer hiring	Budget officers' shortlisting	00:34:38 min
		Budget officers' recruitment	00:32:35 min
	Salary discussion for admins	Salary discussion for HR members	00:37:45 min
		Salary discussion for budget officer	00:31:28 min
	Job appraisal	Salary increments	00:36:02 min
Finance	Bank Accounts (BA)	Project accounts	00:33:22 min
		Employees' accounts	00:36:23 min
		Project funding	00:30:11 min
	Contingency Bills (CB)	Employees' accommodation and traveling bills	00:30:10 min
		Lab equipments' bills	00:34:15 min
	Budget Review and Analysis (BR&A)	Current budget overview and revenue analysis	00:34:07 min
		Budget adjustments and approval	00:32:47 min
		Expense analysis and budget variances	00:34:04 min

Table 21
An extension of the Table 8 representing more detailed distribution of CLE Meeting Corpus across technical domain. It also illustrates that the duration of meetings is in hours for each agenda of a specific scenario.

Domain	Scenario	Agenda	Duration in minutes
Technical	Data Processing (DP)	Data design	03:41:06 h
		Data collection	04:55:25 h
		Data annotation	02:54:21 h
		Data cleaning	02:53:06 h
		Data verification	02:58:08 h
	Requirements (R)	Requirement inception	01:13:34 h
		Requirement documentation	01:00:27 h
		FS document finalization	03:00:10 h
	System Design (SD)	High level design	04:55:46 h
		Detailed design	02:52:47 h
		Database design	03:58:12 h
		Design document finalization	02:10:13 h
	Development (D)	Framework	03:11:17 h
		Status related	03:19:06 h
	Testing (T)	Web app testing	03:49:41 h
		Unit testing	03:40:02 h
		Model testing	02:26:36 h
		Error analysis	02:45:02 h
	Deployment (D)	System deployment	02:44:17 h
		Production server	03:14:05 h
		Deployment server	02:09:51 h

Appendix

See Tables 19–21.

References

Alahmadi, Dimah, Wali, Arwa, & Alzahrani, Sarah (2022). TAAM: Topic-aware abstractive arabic text summarisation using deep recurrent neural networks. *Journal of King Saud University - Computer and Information Sciences*, *34*(6, Part A), 2651–2665.

Ali, Zeshan (2021). Automatic text summarization for urdu roman language by using fuzzy logic. *Journal of Autonomous Intelligence*, *3*, 23.

Ay, Betul, Ertam, Fatih, Fidan, Guven, & Aydin, Galip (2023). Turkish abstractive text document summarization using text to text transfer transformer. *Alexandria Engineering Journal*, *68*, 1–13.

Bani-Almarjeh, Mohammad, & Kurdy, Mohamad-Bassam (2023). Arabic abstractive text summarization using RNN-based and transformer-based architectures. *Information Processing & Management*, *60*(2), Article 103227.

Bhatti, Muhammad Wasif, & Aslam, Muhammad (2019). ISUTD: Intelligent system for urdu text de-summarization. In *2019 international conference on engineering and emerging technologies* ICEET, (pp. 1–5).

Dalal, Sarthak, Singhal, Amit, & Lall, Brejesh (2023). LexRank and PEGASUS transformer for summarization of legal documents. In Dilip Singh Sisodia, Lalit Garg, Ram Bilas Pachori, & M. Tanveer (Eds.), *Machine intelligence techniques for data analysis and signal processing* (pp. 569–577). Singapore: Springer Nature Singapore.

Dam, Sumit Kumar, Shirajum Munir, Md., Raha, Avi Deb, Adhikary, Apurba, Park, Seong-Bae, & Hong, Choong Seon (2023). RNN-based text summarization for communication cost reduction: Toward a semantic communication. In *2023 international conference on information networking* ICOIN, (pp. 423–426).

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, & Toutanova, Kristina (2019). BERT: Pre-training of deep bidirectional transformers for language understanding.

Fabbri, Alexander R., Kryściński, Wojciech, McCann, Bryan, Xiong, Caiming, Socher, Richard, & Radev, Dragomir (2021). SummEval: Re-evaluating summarization evaluation.

Ganesan, Kavita (2018). ROUGE 2.0: Updated and improved measures for evaluation of summarization tasks.

Goyal, Tanya, Li, Junyi Jessy, & Durrett, Greg (2023). News summarization and evaluation in the era of GPT-3.

Hasan, Tahmid, Bhattacharjee, Abhik, Islam, Md Saiful, Samin, Kazi, Li, Yuan-Fang, Kang, Yong-Bin, et al. (2021). XL-Sum: Large-scale multilingual abstractive summarization for 44 languages.

Hu, Yebowen, Ganter, Tim, Deilamsalehy, Hanieh, Dernoncourt, Franck, Foroosh, Hassan, & Liu, Fei (2023). MeetingBank: A benchmark dataset for meeting summarization.

Huang, Han, Wang, Xiaoguang, & Wang, Hongyu (2020). NER-RAKE: An improved rapid automatic keyword extraction method for scientific literatures based on named entity recognition. *Proceedings of the Association for Information Science and Technology*, *57*(1), Article e374.

Humayoun, Muhammad, & Akhtar, Naheed (2022). CORPURES: Benchmark corpus for urdu extractive summaries and experiments using supervised learning. *Intelligent Systems with Applications*, *16*, Article 200129.

Humayoun, Muhammad, Nawab, Rao Muhammad Adeel, Uzair, Muhammad, Aslam, Saba, & Farzand, Omer (2016). Urdu summary corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, & Stelios Piperidis (Eds.), *Proceedings of the tenth international conference on language resources and evaluation* LREC'16, (pp. 796–800). Portorož, Slovenia: European Language Resources Association (ELRA).

Hussain, Khalid M., Mughal, Nimra, Ali, Irfan Z., Hassan, Saif, & Daudpota, Sher Muhammad (2021). Urdu news dataset 1M.

Jadhav, Anish, Jain, Rajat, Fernandes, Steve, & Shaikh, Sana (2019). Text summarization using neural networks. In *2019 international conference on advances in computing, communication and control (ICAC3)* (pp. 1–6).

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., et al. (2003). The ICSI meeting corpus. *Vol. 1*, In *2003 IEEE international conference on acoustics, speech, and signal processing, 2003. proceedings. (ICASSP '03)* (p. I).

Khan, Erbaz, Rauf, Sahar, Adeeba, Farah, & Hussain, Sarmad (2021). A multi-genre urdu broadcast speech recognition system. In *2021 24th conference of the oriental COCOSDa international committee for the co-ordination and standardisation of speech databases and assessment techniques (o-COCOSDa)* (pp. 25–30).

La Quatra, Moreno, & Cagliero, Luca (2023). BART-IT: An efficient sequence-to-sequence model for Italian text summarization. *Future Internet*, *15*(1).

Li, Manling, Zhang, Lingyu, Ji, Heng, & Radke, Richard J. (2019). Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2190–2196). Florence, Italy: Association for Computational Linguistics.

Li, Wei, & Zhuge, Hai (2021). Abstractive multi-document summarization based on semantic link network. *IEEE Transactions on Knowledge and Data Engineering*, *33*(1), 43–54.

Liu, Yinhan, Gu, Jiatao, Goyal, Naman, Li, Xian, Edunov, Sergey, Ghazvininejad, Marjan, et al. (2020). Multilingual denoising pre-training for neural machine translation.

Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach.

Mastropaolo, Antonio, Scalabrino, Simone, Cooper, Nathan, Nader Palacio, David, Poshyvanyk, Denys, Oliveto, Rocco, et al. (2021). Studying the usage of text-to-text transfer transformer to support code-related tasks. In *2021 IEEE/ACM 43rd international conference on software engineering* ICSE, (pp. 336–347).

Mccowan, Iain, Carletta, J, Kraaij, Wessel, Ashby, Simone, Bourban, S, Flynn, M, et al. (2005). The AMI meeting corpus. In *Int'l. conf. on methods and techniques in behavioral research*.

Mohammad Masum, Abu Kaisar, Abujar, Sheikh, Islam Talukder, Md Ashraful, Azad Rabby, A. K. M. Shahariar, & Hossain, Syed Akhter (2019). Abstractive method of text summarization with sequence to sequence RNNs. In *2019 10th international conference on computing, communication and networking technologies* ICCCNT, (pp. 1–5).

Motilal Lodhi, Pallavi, Kharche, Shubhangi, Kambri, Dikshita, & Saleem Khan, Sumaiya (2022). Business meeting summarization system. In *2022 2nd Asian conference on innovation in technology* ASIANCON, (pp. 1–6).

Muhammad, Aslam, Jazeb, Noman, Martinez-Enriquez, Ana Maria, & Sikander, Ali (2018). EUTS: Extractive urdu text summarizer. In *2018 seventeenth mexican international conference on artificial intelligence* MICAI, (pp. 39–44).

Nagoudi, El Moatez Billah, Elmadany, AbdelRahim, & Abdul-Mageed, Muhammad (2022). AraT5: Text-to-text transformers for arabic language generation.

Nawaz, Ali, Bakhtyar, Maheen, Baber, Junaid, Ullah, Ihsan, Noor, Waheed, & Basit, Abdul (2020). Extractive text summarization models for urdu language. *Information Processing & Management*, *57*(6), Article 102383.

Nedoluzhko, Anna, Singh, Muskaan, Hledíková, Marie, Ghosal, Tirthankar, & Bojar, Ondřej (2022). ELITR minuting corpus: A novel dataset for automatic minuting from multi-party meetings in english and czech. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 3174–3182). Marseille, France: European Language Resources Association.

Parida, Shantipriya, & Motlicek, Petr (2019). Abstract text summarization: A low resource challenge. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 5994–5998). Hong Kong, China: Association for Computational Linguistics.

Phan, Long, Tran, Hieu, Nguyen, Hieu, & Trinh, Trieu H. (2022). ViT5: Pretrained text-to-text transformer for Vietnamese language generation.

Qiu, XiPeng, Sun, TianXiang, Xu, YiGe, Shao, YunFan, Dai, Ning, & Huang, XuanJing (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872–1897.

Rahimi, Shohreh Rad, Mozhdehi, Ali Toofanzadeh, & Abdolahi, Mohamad (2017). An overview on extractive text summarization. In *2017 IEEE 4th international conference on knowledge-based engineering and innovation* KBEI, (pp. 0054–0062). IEEE.

Ranganathan, Jaishree, & Abuka, Gloria (2022). Text summarization using transformer model. In *2022 ninth international conference on social networks analysis, management and security* SNAMS, (pp. 1–5).

Raza, Ali, Raja, Hadia Sultan, & Maratib, Usman (2023). Abstractive summary generation for the urdu language.

Rennard, Virgile, Shang, Guokan, Hunter, Julie, & Vazirgiannis, Michalis (2023). Abstractive Meeting Summarization: A Survey. *Transactions of the Association for Computational Linguistics*, *11*, 861–884.

Saadany, Hadeel, & Orăsan, Constantin (2021). BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. In *TRITON 2021, Proceedings of the translation and interpreting technology online conference TRITON 2021*. BULGARIA: INCOMA Ltd. Shoumen.

Shafiq, Nida, Hamid, Isma, Asif, Muhammad, Nawaz, Qamar, Aljuaid, Hanan, & Ali, Hamid (2023). Abstractive text summarization of low-resourced languages using deep learning. *PeerJ Computer Science*, *9*, Article e1176.

Singhal, Daksha, Khatter, Kavya, Tejaswini, A., & Jayashree, R. (2020). Abstractive summarization of meeting conversations. In *2020 IEEE international conference for innovation in technology* INOCON, (pp. 1–4).

Song, Shengli, Huang, Haitao, & Ruan, Tongxiao (2019). Abstractive text summarization using LSTM-CNN based deep learning. *Multimedia Tools and Applications*, *78*, 857–875.

Widyassari, Adhika Pramita, Rustad, Supriadi, Shidik, Guruh Fajar, Noersasongko, Edi, Syukur, Abdul, Affandy, Affandy, et al. (2022). Review of automatic text summarization techniques & methods. *Journal of King Saud University - Computer and Information Sciences*, *34*(4), 1029–1046.

Xue, Linting, Constant, Noah, Roberts, Adam, Kale, Mihir, Al-Rfou, Rami, Siddhant, Aditya, et al. (2020). mT5: A massively multilingual pre-trained text-to-text transformer.

Zaman, Farooq, Shardlow, Matthew, Hassan, Saeed-Ul, Aljohani, Naif Radi, & Nawaz, Raheel (2020). HTSS: A novel hybrid text summarisation and simplification architecture. *Information Processing & Management*, *57*(6), Article 102351.

Zhang, Haochen, Dong, Yuyang, Xiao, Chuan, & Oyamada, Masafumi (2023). Large language models as data preprocessors.

Zieve, Morris, Gregor, Anthony, Stokbaek, Frederik Juul, Lewis, Hunter, Mendoza, Ellis Marie, & Ahmadnia, Benyamin (2023). Systematic TextRank optimization in extractive summarization. In Ruslan Mitkov, & Galia Angelova (Eds.), *Proceedings of the 14th international conference on recent advances in natural language processing* (pp. 1274–1281). Shoumen, Bulgaria: INCOMA Ltd..