IUST
Iran University of
Science and Technology

Iranian Journal of Electrical and Electronic Engineering

Journal Homepage: ijeee.iust.ac.ir

KurdSM: Transformer-Based Model for Kurdish Abstractive Text Summarization with an Annotated Corpus

Pedram Yamini*, Fatemeh Daneshfar*(C.A.) and Abouzar Ghorbani**

Abstract: With the exponential growth of unstructured data on the Web and social networks, extracting relevant information from multiple sources; has become increasingly challenging, necessitating the need for automated summarization systems. However, developing machine learning-based summarization systems largely depends on datasets, which must be evaluated to determine their usefulness in retrieving data. In most cases, these datasets are summarized with humans' involvement. Nevertheless, this approach is inadequate for some low-resource languages, making summarization a daunting task. To address this, this paper proposes a method for developing the first abstractive text summarization corpus with human evaluation and automated summarization model for the Sorani Kurdish language. The researchers compiled various documents from information available on the Web (rudaw), and the resulting corpus was released publicly. A customized and simplified version of the mT5-base transformer was then developed to evaluate the corpus. The model's performance was assessed using criteria such as Rouge-1, Rouge-2, Rouge-L, N-gram novelty, manual evaluation and the results are close to reference summaries in terms of all the criteria. This unique Sorani Kurdish corpus and automated summarization model have the potential to pave the way for future studies, facilitating the development of improved summarization systems in low-resource languages.

Keywords: Kurdish Language; Abstractive Summarization; Text Processing; Annotated Corpus.

1 Introduction

THE texts available on the internet are widely recognized as a significant and extensive source of information in today's world. News websites, in particular, draw the attention of diverse groups of people. However, with the vast amount of text available on these sites, it is essential to use summarization tools to condense the information. This is where natural language processing (NLP) comes in. Text summarization is valuable to researchers because it aids in information retrieval [1], content abstraction [2], time efficiency [3], data synthesis [4], and content generation [5], thereby facilitating various aspects of research in the field of NLP. The two primary methods for text summarization are extractive and abstractive [6]. In extractive summarization, selected sentences are extracted from the original text and included in the summary text. On the other hand, abstractive summarization generates a new, concise summary that captures the essential ideas of the source text [7, 8]. This method creates new phrases through a combination of words. However, abstractive summarization faces a significant challenge due to the lack of available data (corpus).

Typically, there are common methods for abstractive text summarization including language models such as PEGASUS [9], T5 [10], and BART [11]. Although these transformer-based models perform well in many NLP tasks, they require a rich corpus to train. Their outputs depend greatly on the amount of training data.

Iranian Journal of Electrical & Electronic Engineering, 2024.
Paper first received 26 May 2024 and accepted 06 December 2024.

* The authors are with the Department of Computer Engineering and Information Technology, University of Kurdistan, Sanandaj, Iran.
E-mail: f.daneshfar@uok.ac.ir.
** The author is with the Department of Computer Engineering, University of Isfahan, Isfahan, Iran.
Corresponding Author: Fatemeh Daneshfar.

Table 1 Available Kurdish Language Sources

Resources	Ref	Description	Kurdish dialect
Corpora	[12]	Open Super-large Crawled ALMAnaCH Corpus	Sorani and Kurmanji
	[13]	Kurdish folkloric lyrics corpus	Sorani
	[14]	AsoSoft corpus	Sorani
	[15]	Zaza-Gorani corpus	Zazaki and Gorani
	[16]	Kurdish Textbooks Corpus	Sorani
	[17]	Pewan	Sorani and Kurmanji
Parallel Corpora	[18]	FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation	Sorani
	[19]	Ahmadi et al's corpus	English-Kurmanji-Sorani aligned texts
	[20]	A parallel corpus of Sorani-English text	Sorani-English text
	[21]	Tanzil: one Quran translation alignable with many other translations in other languages	Sorani-English
	[22]	Ataman's Bianet corpus	Turkish-English-Kurmanji aligned texts
Dictionaries, terminologies and ontologies	[23]	Kurdish lexicographical resources in Ontolex-Lemon	Sorani, Kurmanji, Gorani and Southern Kurdish
	[24]	KurdNet-the Kurdish wordNet	Sorani
Datasets	[25]	Datasets for text to Kurdish Sign Language	Sorani
	[26]	A dataset for speech recognition	Sorani
	[27]	Universal dependency	Kurmanji
	[28]	Dataset for sentiment analysis	Sorani

Moreover, rich corpora are not available for some languages; therefore, researchers have resorted to developing specific corpora tailored to their needs.

The Kurdish language boasts a significant population of 30-45 million speakers worldwide [29, 30]. However, despite its wide usage, there is a shortage of resources available for in-depth research, resulting in a scarcity of adequate references. As shown in Table 1, there are limited sources of Kurdish language data available for different Kurdish dialects [31]. Moreover, there is currently no established corpus for Sorani Kurdish text summarization, highlighting a crucial area for further development in the field.

This paper presents the development and evaluation of the first annotated corpus with human evaluation for Sorani Kurdish text summarization, which includes a dataset of numerous news articles covering various topics. At the time of writing this paper, another related work emerged, featuring human-written summaries for 40 languages, including low-resource ones (The paper is available at aclanthology.org/2023.findings-acl.427). The corpus contains abstractive news summaries, and it was designed to be a representative and comprehensive corpus for evaluating Sorani Kurdish text summarization techniques. The paper also introduces KurdSM, the first Sorani Kurdish text summarization model, which is based on a customized and simplified version of the mT5-base transformer model by [32]. The mT5-base transformer is a multilingual model trained on a large training set of 101 languages, making it an efficient option for non-English language training. The performance of KurdSM was compared to that of a simple model using ROUGE, N-gram novelty measures, and manual evaluation.

This study marks a significant milestone in the field of Kurdish language processing. The proposed corpus, which is the first of its kind for Sorani Kurdish language text summarization, has the potential to be utilized in the future for summarizing news articles, academic papers, and books by evaluating the models trained on this corpus using a cross-domain dataset. Moreover, this study opens up opportunities for other researchers in Sorani Kurdish language processing to leverage the proposed corpus and build models to further advance the field.

This paper presents two significant contributions to the field of Sorani Kurdish text summarization:

- The first annotated dataset with human evaluation for Sorani Kurdish text summarization, which is made freely available for development purposes (github.com/pedramyamini/KurdSM)
- A customized and simplified implementation of the mT5 model for Sorani Kurdish abstractive text summarization, trained on the proposed dataset and also released open-sourced (huggingface.co/pedramyamini/ku_t5_base)

The paper is structured into several sections. Section 2 provides a comprehensive review of the relevant literature, while Section 3 offers a detailed account of the development process of the KurdSM corpus. Section 4 describes the transformer-based abstractive text summarization model utilized in this study. Section 5 presents the experimental results, and finally, Section 6 summarizes the research conclusions.

2 Related Work

In this section, we explore several studies that have focused on creating text summarization datasets, with a particular emphasis on languages that are low-resource.

Farhani *et al.* (2021) proposed a corpus of summaries for Persian texts. They fine-tuned Parsbert and mT5 models on this corpus and then assessed the outputs. A remarkable point of this paper is the summarization of Persian texts [33]. At the time of publication, this study was one of the limited research projects on Persian. Landro *et al.* (2022) developed two corpora in abstractive text summarization for the Italian language by using two Italian and Spanish news websites [34]. The Spanish website was translated into Italian through a translator. They adopted mBART and T5 methods for evaluating their corpora and claimed that the proposed corpora were the only abstract summarization corpora in Italian. According to their findings, if we want to obtain good results in a language with few sources, a dataset should be created in that language [34]. In [35], Chowdhury *et al.* proposed a graph-based abstractive text summarization method for Bengali texts. This method needs to employ a part-of-speech label and a pre-trained model of Bengali language. The proposed method was also evaluated through the dataset presented in the paper; however, it faces the limited generation of new words. Parida and Motlicek (2019) adopted the advanced transformer model to summarize German abstractive texts [36]. They used the iterative data augmentation method for resolving the lack of German language data. This method uses both synthetic and actual summarization data to generate a new corpus in German. These data were collected and produced in different domains by using crawler tools. The proposed method was evaluated through ROUGE and BLEU criteria. According to the outputs, the resultant data had a positive effect on the performance of text summarization in German. Since collecting data for text summarization is costly and time-consuming, the collected documents sometimes have a limited number of long documents, something which has a negative effect on the quality of text summarization in pre-trained models. The paper reviewed by Bajaj *et al.* (2021) evaluated the compression of a long document into a coherent concise document by preserving the important information [37]. This paper proposed a method for integrating GPT-2 and BERT classifications to solve the issue. The integrated framework generates coherent and fluent summaries through a pre-trained BART model. The paper reviewed by Baykara and Güngör (2022) employed a pre-trained Seq2Seq model for text summarization and title generation. It was evaluated on two Turkish datasets named TR-News and MLSum [38]. According to the experimental results, the BERTurk-case monolingual model was able to obtain satisfactory results compared with the mT5 model in summarizing Turkish texts. Shilpa and Shashi Kumar (2019) proposed a method for abstractive summarization of an Indian dialect. The proposed method performed text summarization by integrating information extraction rules and template-based models TF/IDF [39]. The lexical analysis was also conducted to reduce the complexity of the rules. In [40], a cross-lingual method was proposed for abstractive text summarization. For this purpose, a pre-trained model of the English language was set to the Slovenian language for text summarization. According to the experimental results, zero-shot models were outperformed by few-shot models, and data quality had positive effects on text summarization. The automated evaluation of text summarization in this study showed that the quality of summarization was higher than that of the language model set on the target language. Table 2 shows a summarizes of the mention paper with their evaluation.

3 Sorani Kurdish Abstractive Text Summarization Corpus

The corpus used for fine-tuning is sourced from Sorani Kurdish news articles across various categories available at www.rudaw.net, a multimedia and multilingual website that publishes news in Kurdish, English, Arabic, and Turkish. This source is particularly suitable for collecting a corpus for text summarization as each article has a corresponding summary and is well-written and accurately revised. This designated corpus is the first Sorani Kurdish abstractive summarization corpus with human evaluation ever collected, making it a valuable reference for evaluating Kurdish abstractive summarization models against other similar models. Notably, Rudaw's news articles are almost free of errors, and each article has its corresponding summary highlighted in bold type, which served as reference abstractive summaries. The website covers various categories, including business, world, Iraq, Iran, Syria, Turkey, Kurdistan, health, sports, and culture-style, as depicted in Figure 1. It's worth noting that most of the news articles are in the Kurdistan category, followed by Iraq, the world, Turkey, Iran, and Syria, respectively. This distribution was taken into account when dividing the corpus into training (90%), test (5%), and validation (5%) sets to prevent domain gaps, as shown in Figure 2.

Table 2 Literature Review of the Text Summarization Papers in Terms of Methods, Datasets, and Evaluation Results

Ref	Key Contributions	Dataset, Language	Evaluation Results
[33]	• Introduced a corpus of Persian text summaries. • Fine-tuning Parsbert and mT5 models	pn-summary, Persian	R-1: 44.01 R-2: 25.07 R-L: 37.76
[34]	• Introduced two corpus of Italian text summaries. • Fine-tuning mBART and T5 models	Two datasets, Italian	R-1: 38.91 R-2: 17.44 R-L: 26.17
[35]	• Introduced an abstractive dataset with document-summary pairs. • Introduced an unsupervised abstractive sentence generation model.	NCTB and BNLPC, Bengali	R-1: 61.62 R-2: 55.97 R-L: 61.09
[36]	Introduced an iterative data augmentation approach which uses synthetic data for low resource dataset.	Swiss text, German	R-1: 55.7 R-2: 41.8 R-L:57.6
[37]	• Introduced new approach long summaries task. • Fine-tuning BART model	Amicus dataset	R-1: 47.07 R-2: 17.64 R-L:24.40
[38]	• Fine-tuned BART, T5, GPT, BERT and XLM models. • Used monolingual BERTurk models outperform the multilingual BERT models. • Showed how pre-trained sequence-to-sequence models can reach state-of-the-art summary and title generation tasks.	TRNews and MLSum datasets, Turkish.	R-1: 42.26 R-2: 27.81 R-L: 37.96
[39]	• Created abstractive summaries of individual Kannada documents that are content-aware. • Introduced a method to generate diverse sentences that effectively present extracted information.	Indian	-
[40]	The proposed approach does not require any resources in the target language, apart from a monolingual corpus.	Two datasets, Slovene	R-1: 24.97 R-2: 7.43 R-L:21.50

Figure 3 exhibits distribution plots illustrating the tokenized lengths of articles and summaries using the mT5-base tokenizer, displayed in token counts. The histograms in the plot demonstrate the frequency of instances with varying lengths, while the scatter plots provide better visualization for any outliers, such as extremely lengthy documents. This information is crucial in establishing two critical hyper-parameters, namely the maximum length of source and target texts. These hyper-parameters play a vital role in enhancing the efficiency of training and inference processes while ensuring that the summary lengths remain similar to reference summaries. Moreover, it is necessary to truncate very lengthy articles (outliers) to maintain a certain degree of quality in the summaries. Figure 3a and Figure 3b illustrates the token counts of articles, which are usually less than 1024 tokens. The tokenization was performed using the mT5-base tokenizer splitter.

To address the resource limitations, particularly in GPU graphic memory, the maximum length (tokens) of input articles was set to 512 using the mT5-base tokenizer. Truncating longer inputs may result in missing crucial information from news articles, which can affect the coverage of important information in the generated summaries. The majority of the summary token counts are less than 128 tokens. Therefore, we set the maximum input and target lengths to 512 and 128 tokens, respectively, after performing tokenization using the mT5-base tokenizer.

Table 3 presents the statistics of article news token count for the prepared corpus consisting of 38,611 articles. The input max length of 512 tokens is reasonably close to the mode, mean, and median of the article news token count, taking into account the resource limitations.

Table 4 provides the statistics of the summary token count, and the target max length of 128 tokens is reasonably close to the mode, mean, and median of the summaries token count. However, it is worth noting that the generated summaries have a minimum length of 96 tokens, which is close to the mean of the summaries token count (92.59 tokens). Figure 4 illustrates the step-by-step process of developing the corpus. The Sorani or central Kurdish news from www.rudaw.net were crawled using the Selenium python library. To accelerate the crawling process, we used the Microsoft Edge headless driver with disabled image loading. Each news entry had a summary highlighted in bold, followed by the article. We collected these articles, summaries, and news hyperlinks to create a raw dataset. Next, we used the free normalizer library introduced by [14] to normalize the articles and corresponding summaries. Finally, to comply with the recommended dataset format by Huggingface [41], we obtained a JSON dataset from the normalized raw dataset. JSON is a widely used standard object notation for storing data in various datasets.

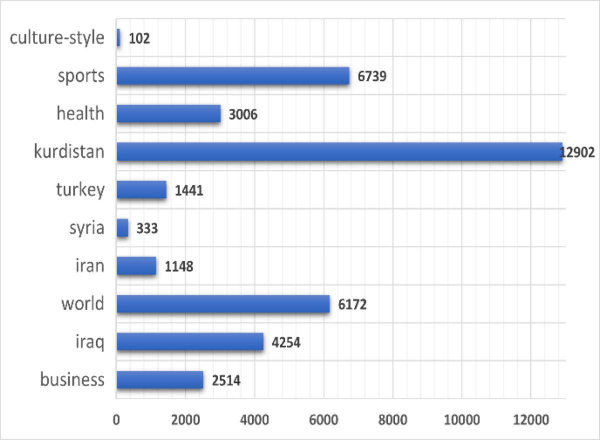


Fig. 1 Distribution of categories

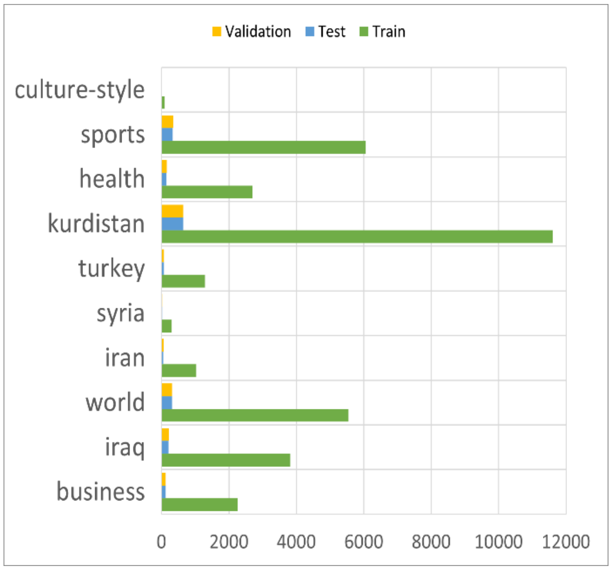
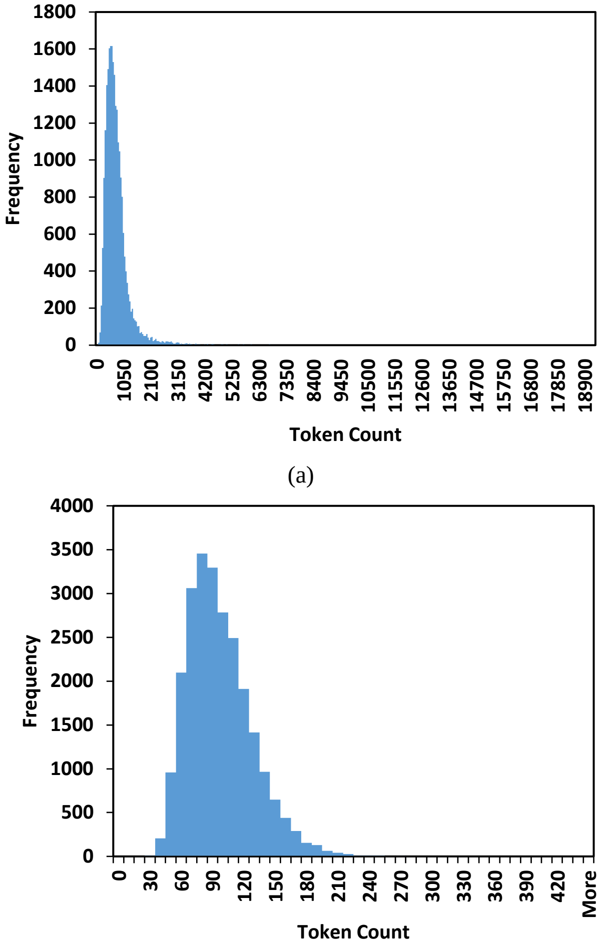


Fig. 2 Distribution of categories in training/test/validation splits.

To prevent domain gaps, we grouped the dataset into categories and then created training, testing, and validation splits. Specifically, we split 90% of instances from each category for the training split, 5% for the test split, and the remaining 5% for the validation split. While Iraq, Iran, Syria, and Turkey are subcategories of the main Middle East category, we treated them as four main categories due to their significance in simplifying the mT5-base and providing insights into the importance of each language.



(a)

(b)

Fig. 3 Distribution of article and summary token counts a) article token counts histogram b) summary token counts histogram.

Table 3 Article news token counts statistics

Maximum	Minimum	Mode	Median	Mean
19227	59	569	654	778.62

Table 4 Summary news token counts statistics

Maximum	Minimum	Mode	Median	Mean
439	25	77	88	92.59

3.1 Normalization

Text normalization is a crucial step in reducing randomness and variety in incorrect data. It involves addressing issues such as the use of different forms of characters, Unicode types, and digits. In our study, we utilized the Asosoft Normalizer library (The documentation of Asosoft normalization library is at github.com/AsoSoft/AsoSoft-Library) [14] for Sorani Kurdish text normalization to ensure accurate and consistent data. In [14], the most common abnormalities in Sorani Kurdish are discussed in detail. Some of the key corrections include replacing "ر" at the beginning of

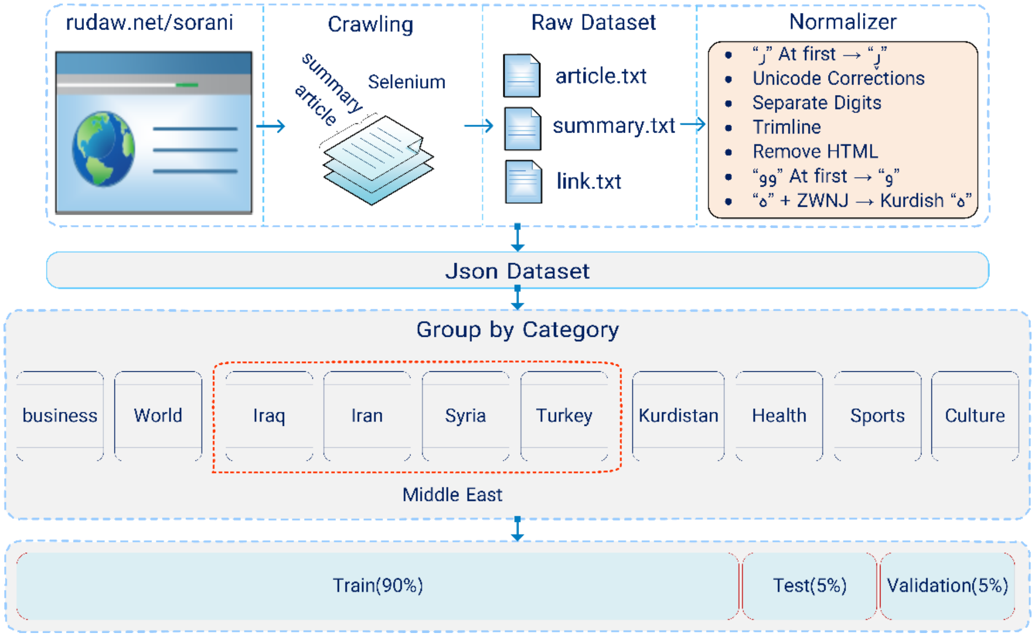


Fig 4. The process of corpus development

words with "ڕ", replacing "ه" + ZWNG (zero-width non-joiner) with Kurdish "ە", replacing "وو" at the beginning of words with "و", separating concatenated digits of words with spaces, and removing HTML mark-ups. Additionally, Arabic "ك" and "ي" must be replaced with Kurdish "ک" and "ی" (Figure 4). Once the corpus is normalized, the text can be tokenized and embedded to capture relevant information using the attention mechanism in the transformer architecture for encoding and decoding words.

4 Models

4.1 Definition 1 – mT5-base

T5 is a powerful transformer model that can be used for a variety of NLP tasks. It follows a text-to-text format, where the same model, loss function, and parameters can be used for multiple NLP tasks, including machine translation, document summarization, question answering, and classification [10]. While T5 is a remarkable model, one of its drawbacks is the lack of support for non-English languages. However, a new version called multilingual T5 (mT5) addresses this issue. This version has been trained on 101 different languages and comes in small, base, large, XL, and XX sizes (as shown in Figure 5). It is similar to the original T5 model but with the added capability to support multilingual inputs and outputs. This makes mT5 a valuable tool for natural language processing tasks involving multiple languages.

The success of multilingual models heavily depends on their data sampling method. In the mT5 transformer, the zero-sum game method [32] is used for data sampling. This method involves two players competing with each other, where the winner is the player with more points than the rival. In mT5, the two players represent the amounts of data from rich languages and low-resource languages. However, multilingual models can face two major problems. The first problem is over-fitting, which can occur if a large amount of data is sampled from low-resource languages. On the other hand, under-fitting can occur if rich languages are not trained enough. Therefore, achieving a balance of data usage between rich and low-resource languages is crucial for multilingual models. Additionally, it is important to ensure that the distribution of linguistic data is not biased towards a specific language to prevent model bias. Overall, the distribution of linguistic data is a critical aspect of multilingual models. Proper data sampling methods, such as the zero-sum game method, can help address these issues and ensure that the model is effective for all supported languages.

Figure 5 illustrates the workflow of the mT5 transformer model, which begins by selecting a multilingual corpus for training. The data sampling method is then used to determine the probability of selecting languages and the amount of data to be selected from rich and low-resource languages. This is crucial to prevent over-fitting and under-fitting problems in the model. Next, the desired NLP task, such as machine translation or text summarization, is selected by tagging the input text in a specific language. The input text is then fed into the 12-layer transformer model along with the selected task and training samples. The model is trained sequentially on each language, and the

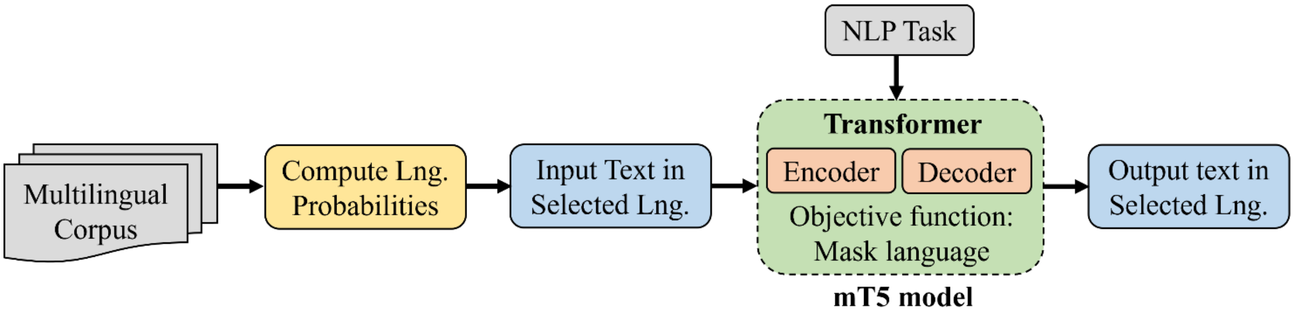

NLP Task
Multilingual Corpus
Compute Lng. Probabilities
Input Text in Selected Lng.
Transformer
Encoder
Decoder
Objective function: Mask language
mT5 model
Output text in Selected Lng.

Fig. 5 The overall structure of the mT5 framework

languages can affect each other during the training process. Finally, the output text is generated according to the selected task.

4.2 Definition 2 – KurdSM

Fine-tuning is a process that involves adding a task-specific head to a pre-trained model's body, which has general knowledge of languages from primary tasks, such as mask language modelling in mT5. This process has been made easier thanks to the Huggingface transformer library. In this study, the mT5-base checkpoint was fine-tuned using the Rudaw news corpus. To reduce the model's size and increase its efficiency, unnecessary word embedding samples from the mT5-base were removed (More detail is available at towardsdatascience.com. “how to adapt a multilingual t5 model for a single language”), leaving only five languages out of 101 (because their linguistic structures are closely related to Kurdish). The mT5-base tokenizer was employed to process five separate datasets, each comprising 100,000 sentences from the five remaining languages. This resulted in a total of 500,000 sentences being processed, and the most frequent word embedding samples of these five remaining languages were retained to prune the unnecessary mT5-base word embedding samples. As a result, the model size was reduced from 2.2 gigabytes to 1.06 gigabytes, and the model was pushed to the Huggingface models hub as *ku_t5_base*.

Table 5 shows the remaining languages and their corresponding number of word embedding samples. As discussed in Section 3, most of the news articles in the Rudaw corpus fall under the Kurdistan category, followed by Iraq, World, Turkey, Iran, and Syria categories, respectively. Therefore, 20k word embedding samples for each of the Kurdish and English languages (as a global language) and 10k word embedding samples for each of the Persian, Arabic, and Turkish languages were kept. Overall, this fine-tuning process helped to reduce the model's size while retaining its effectiveness in processing the Rudaw news corpus. By removing unnecessary word embedding samples, the model's efficiency was improved, and it was able to perform better on the specific task at hand.

Table 6 presents the hyper-parameters utilized for fine-tuning the mT5-base model. In order to fine-tune the model on a single GPU, the batch size was set to four, allowing for more parallel processing by passing further training instances in one iteration. The input and target max lengths were set to 512 and 128 tokens, respectively. Any sequences longer than 512 tokens were truncated, while shorter sequences were padded to the maximum length with a special padding token defined in mT5-base. The model was trained for 12 epochs, yielding the best possible results within the resource and time constraints. The learning rate and weight decay were based on the default configurations of the Adam optimizer. For generation, the minimum and maximum lengths were set to 96 and 128 tokens, respectively, as most of the reference summaries fell within this length range, as calculated by the mT5 tokenizer prior to generation.

Figure 6 provides a comprehensive visual representation of our fine-tuning process, illustrating each step involved. The dataset partitions, including training, validation, and test sets, are stored in JSON files, and we streamlined data handling and preprocessing using the Huggingface datasets library. This library offers numerous advantages, such as a unified API for diverse datasets, automatic caching, and support for multiprocessing, optimizing our workflow. Given the absence of prior research on abstractive summarization in Sorani Kurdish, we employed a straightforward baseline method known as Lead-3. This approach selects the initial three sentences from the source text as the summary. To implement this baseline, we utilized the Kurdish Language Processing Toolbox (KLPT) sentence tokenizer [31] to segment the input text into sentences and then selected the first three as the baseline summary.

In the subsequent experiments section, we present a detailed performance comparison between our proposed model and the Lead-3 baseline, followed by a comprehensive analysis of the evaluation results.

5 Experiments and Comparison Results

5.1 ROUGE

For evaluating our results, we employed ROUGE metrics, a widely accepted evaluation method in the field. These metrics rely on calculating the n-gram overlap between the reference and candidate summaries, along with assessing the longest common substring (LCS). Specifically, ROUGE-1 gauges the overlap of unigrams between the reference and candidate summaries, providing a measure of basic overlap. In contrast, ROUGE-2 assesses the overlap of bigrams, offering a more comprehensive view of summary quality, particularly in terms of fluency. This dual approach ensures a more thorough evaluation, as higher ROUGE-2 scores can indicate enhanced fluency and coherence in the generated summaries. Additionally, ROUGE-L calculates the longest common substring, which is valuable for assessing fluency and overlap between the reference and candidate summaries, even when the matching words are not in consecutive order. Furthermore, it's important to note that precision and recall measures play a crucial role in the ROUGE evaluation framework. Precision calculates the overlap concerning the candidate summary, while recall calculates the overlap concerning the reference summary. The F-measure, computed as the geometric mean of precision and recall, is a commonly used metric for comparing the overall performance of the ROUGE metrics.

In our analysis, we not only considered F-measure but also scrutinized precision and recall values to provide a comprehensive assessment of the quality of our generated summaries.

Table 7 presents an example of calculating the ROUGE-1, ROUGE-2, and ROUGE-L metrics, using a candidate summary for the Sorani Kurdish language as a reference. The LCS consists of words in the correct order, but they may not be consecutive. The overlapping n-grams are highlighted in red, and the ROUGE-1, ROUGE-2, and ROUGE-L equations are represented and calculated according to [42]. In Table 8, we compare the results of the baseline model with those of the proposed model by computing the ROUGE-1, ROUGE-2, and ROUGE-L metrics. The proposed model outperforms the baseline model, as demonstrated by the metrics' values and results. These findings are significant and meaningful.

It is apparent that the recall measure is higher in the baseline model than in the proposed model because the baseline assumed the first three sentences of each input text as summaries and the first sentences and conclusions are the most informative sentences of texts. Moreover, the baseline summary exhibits a greater degree of overlap with the reference summary. This can be attributed to the fact that the baseline summary (lead-3) is authored by the same individual who created the reference summary. Additionally, the baseline summary is presented after the reference summary, offering a more detailed exposition of the reference summary's content. In contrast, the model-generated summary often faces limitations due to GPU constraints, leading to truncation and loss of some input information. Nevertheless, it's worth noting that the proposed model demonstrates higher precision. In essence, the summary produced by our model is more concise and manages to encompass a broader spectrum of the input text within a reasonable token count. This results in a denser summary that provides improved coverage of the entire input text, despite the inherent limitations.

5.2 N-gram Novelty

While ROUGE metrics are useful for measuring the overlaps between reference and candidate summaries, they do not provide a measure of how abstractive the summaries are. To address this limitation, the N-gram novelty metric [43] can be used to measure the abstractness of a summary. This metric analyzes how many novel N-grams appear in the summary but do not appear in the source document. Equation (1) shows the formula for computing the N-gram novelty, where the summary N-grams and source N-grams are both sets. Subtracting the source N-grams from the summary N-grams results in a set that includes novel N-grams. Dividing the number of novel N-grams by the number of summary N-grams provides the N-gram novelty.

$$N - gram\ Novelty = \frac{|\{\text{summary n}-\text{grams}\}-\{\text{source n}-\text{grams}\}|}{|\{\text{summary}|} \quad (1)$$

In addition to the N-gram novelty metric, we also computed the unigram novelty and bigram novelty for the reference, model, and baseline summaries, and the results are presented in Figure 7A. Higher N-gram novelty score indicates that the summary is more abstractive. As shown in Figure 7, the reference summaries are highly abstractive, while the model summaries are more abstractive than the baseline summaries. It is worth noting that the Lead-3 baseline is an extractive summary, which explains why its N-gram novelty score is the lowest. The summaries generated by our proposed model are moderately abstractive.

5.3 Manual Evaluation

While automated evaluation metrics such as ROUGE can provide insights into the quality of summaries, they may not always accurately reflect human judgments of summary quality. To ensure that the generate

Table 5 The simplified mT5-base kept languages (all the sentences No. is 100k)

Language	English	Kurdish	Persian	Arabic	Turkish
Word Embeddings No.	20k	20k	10k	10k	10k

Table 6 The hyper-parameters for fine-tuning the mT5-base

Hyper-parameter	Value
Batch size	4
Max input length (tokens)	512
Max target length (tokens)	128
Training epochs No.	12
Learning rate	5.6e-5
Weight decay	0.01
Generation minimum length	96
Generation maximum length	128

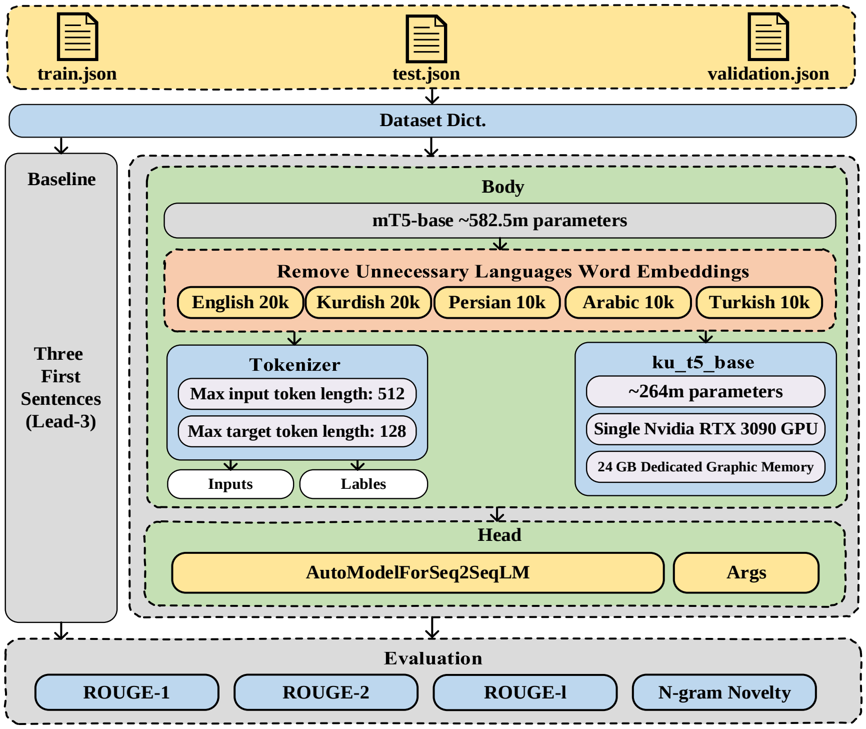


Fig 6. The process of fine-tuning ku_t5_base

summaries are of appropriate quality, a manual evaluation was also conducted. Specifically, two Sorani Kurdish literature students were asked to read 50 randomly selected articles along with their corresponding reference, model-generated, and baseline summaries without knowing which was which. The students were then asked to assign three scores ranging from 1 to 10 to each article based on readability (i.e., coherence and fluency), relevance (i.e., coverage of important information from the source text), and length (i.e., whether the summary was too short or too long). The scores were then compared to the average scores across all samples and students. The results of the manual evaluation are presented in Table 9. It is worth noting that the baseline summaries, which consist of three lead sentences, were approximately twice as long as the reference and model-generated summaries. Although the baseline summaries may contain more information, they may also be too long. The model-generated summaries were found to be comparable to the reference and baseline summaries in terms of readability and relevance.

Additional information, including the distributions of manual evaluation results, can be found in Figure 8 and Figure 9.

Table 7 Reference and candidate summaries unigrams and bigrams. The common unigrams and bigrams between reference and candidate summaries are highlighted in red. P, R and F are Precision, Recall and F-measure respectively.

<table>
<tr><th colspan="9">Reference Summary</th></tr>
<tr><td colspan="9">"وەزارەتی تەندروستیی هەرێمی کوردستان ئاماری ٢٤ کاژێری رابردووی کۆرۆنای بڵاوکردەوە و ئاماژەی بەوە کردووە، ٦٧١ تووشبووی نوێ دەستنیشانکراون و ٣٢ تووشبووش گیانیان لەدەستداوە".</td></tr>
<tr><th colspan="9">Reference Summary in English</th></tr>
<tr><td colspan="9">The Kurdistan Regional Government (KRG) has reported 671 new cases of coronavirus in the past 24 hours, while 32 people have died.</td></tr>
<tr><th colspan="9">Candidate Summary</th></tr>
<tr><td colspan="9">"وەزارەتی تەندروستیی هەرێمی کوردستان ئاماری ٢٤ کاژێری رابردووی تایبەت بە ڤایرۆسی کۆرۆنای بڵاوکردەوە و رایگەیاند، ٦٧١ تووشبووی نوێی کۆرۆنا تۆمارکراون و ٢٥ تووشبووش گیانیان لەدەستداوە. هاوکات ٦٧١ تووشبووی پێشووی ڤایرۆسەکەش چاکبوونەتەوە."</td></tr>
<tr><th colspan="9">Candidate Summary in English</th></tr>
<tr><td colspan="9">The Kurdistan Regional Government (KRG) has reported 671 new cases of coronavirus in the past 24 hours, while 25 people have died. Meanwhile, 671 previous cases have recovered.</td></tr>
<tr><th colspan="9">Reference unigrams</th></tr>
<tr><td colspan="9">R1-grams = {وەزارەتی , تەندروستیی, هەرێمی, کوردستان, ئاماری, ٢٤, کاژێری, رابردووی, کۆرۆنای, بڵاوکردەوە, و, ئاماژەی, بەوە, کردووە, ٦٧١, تووشبووی, نوێ, دەستنیشانکراون, و, ٣٢, تووشبووش, گیانیان, لەدەستداوە}</td></tr>
<tr><th colspan="9">Reference unigrams in English</th></tr>
<tr><td colspan="9">R1-grams = <b>{</b>The, Ministry, of, Health, of, the, Kurdistan, Region, has, released, the, statistics, of, the, past, 24, hours, and, indicated, 671, new, cases, and, 32, deaths<b>}</b></td></tr>
<tr><th colspan="9">Candidate unigrams</th></tr>
<tr><td colspan="9">C1-grams = {وەزارەتی, تەندروستیی, هەرێمی, کوردستان, ئاماری, ٢٤, کاژێری, رابردووی, تایبەت, بە, ڤایرۆسی, کۆرۆنای, بڵاوکردەوە, و, رایگەیاند, ٦٧١, تووشبووی, نوێی, کۆرۆنا, تۆمارکراون, و, ٢٥ , تووشبووش, گیانیان, لەدەستداوە, هاوکات, ٦٧١, تووشبووی, پێشووی, ڤایرۆسەکەش, چاکبوونەتەوە}</td></tr>
<tr><th colspan="9">Candidate unigrams in English</th></tr>
<tr><td colspan="9">C1-grams = {The, Ministry, of, Health, of, the, Kurdistan, Region, announced, 671, new, cases, of, coronavirus, in, the, past, 24, hours, and, 25, deaths , 671, previous, cases, recovered}</td></tr>
<tr><th colspan="9">Reference bigrams</th></tr>
<tr><td colspan="9">R2-grams = {وەزارەتی تەندروستیی, تەندروستیی هەرێمی, هەرێمی کوردستان, کوردستان ئاماری, ئاماری ٢٤, ٢٤ کاژێری, کاژێری رابردووی, رابردووی کۆرۆنای, کۆرۆنای بڵاوکردەوە, بڵاوکردەوە و, و ئاماژەی, ئاماژەی بەوە, بەوە کردووە, کردووە ٦٧١, ٦٧١ تووشبووی, تووشبووی نوێ, نوێ دەستنیشانکراون, دەستنیشانکراون و, و ٣٢, ٣٢ تووشبووش, تووشبووش گیانیان, گیانیان لەدەستداوە}</td></tr>
<tr><th colspan="9">Reference bigrams in English</th></tr>
<tr><td colspan="9">R2-grams = {The Ministry of Health, Regional Health, Kurdistan Region, Kurdistan statistics, statistics, statistics, 24, 24 hours, past, past hours, coronavirus, published, and, and indicated, indicated, indicated, that, made 671, 671 cases, new cases, new identified, identified and, and 32, 32 cases, cases died, died}</td></tr>
<tr><th colspan="9">Candidate bigrams</th></tr>
<tr><td colspan="9">C2-grams = {وەزارەتی تەندروستیی, تەندروستیی هەرێمی, هەرێمی کوردستان, کوردستان ئاماری, ئاماری ٢٤, ٢٤ کاژێری, کاژێری رابردووی, رابردووی تایبەت, تایبەت بە, بە ڤایرۆسی, ڤایرۆسی کۆرۆنای, کۆرۆنای بڵاوکردەوە, بڵاوکردەوە و, و رایگەیاند, رایگەیاند ٦٧١, ٦٧١ تووشبووی, تووشبووی نوێی, نوێی کۆرۆنا, کۆرۆنا تۆمارکراون, تۆمارکراون و, و ٢٥, ٢٥ تووشبووش, تووشبووش گیانیان, گیانیان لەدەستداوە, لەدەستداوە هاوکات, هاوکات ٦٧١, ٦٧١ تووشبووی, تووشبووی پێشووی, پێشووی ڤایرۆسەکەش, ڤایرۆسەکەش چاکبوونەتەوە}</td></tr>
<tr><th colspan="9">Candidate bigrams in English</th></tr>
<tr><td colspan="9">C2-grams = {Ministry of Health, Regional Health, Kurdistan Region, Kurdistan statistics, statistics, 24, 24 hours, past hour, past special, related, to, virus, coronavirus, coronavirus published, published and, and announced, announced 671, 671 cases, infected New, new coronavirus, coronavirus registered, registered and, and 25, 25 cases, cases died, died, died while, while 671, 671 cases, previous cases, previous cases of the virus, the virus recovered}</td></tr>
<tr><th colspan="9">LCS</th></tr>
<tr><td colspan="9">"وەزارەتی تەندروستیی هەرێمی کوردستان ئاماری ٢٤ کاژێری رابردووی کۆرۆنای بڵاوکردەوە و ٦٧١ تووشبووی و تووشبووش گیانیان لەدەستداوە"</td></tr>
<tr><th colspan="9">LCS in English</th></tr>
<tr><td colspan="9">The Kurdistan Regional Government (KRG) has reported 671 new cases of coronavirus in the past 24 hours</td></tr>
<tr><th colspan="3">ROUGE-1</th><th colspan="3">ROUGE-2</th><th colspan="3">ROUGE-L</th></tr>
<tr><th>P</th><th>R</th><th>F</th><th>P</th><th>R</th><th>F</th><th>P</th><th>R</th><th>F</th></tr>
<tr><td>0.54</td><td>0.73</td><td>0.62</td><td>0.4</td><td>0.54</td><td>0.46</td><td>0.54</td><td>0.73</td><td>0.62</td></tr>
</table>

Table 8 The ROUGE metrics results

	ROUGE-1			**ROUGE-2**			**ROUGE-L**		
Model	P	R	F	P	R	F	P	R	F
Baseline (three first sentences)	15.24	**40.70**	22.18	5.47	**15.78**	8.12	12.70	**34.39**	18.55
Fine-tuned ku_t5_base	**25.74**	33.47	**29.10**	**11.36**	15.02	**12.94**	**22.26**	29.05	**25.21**

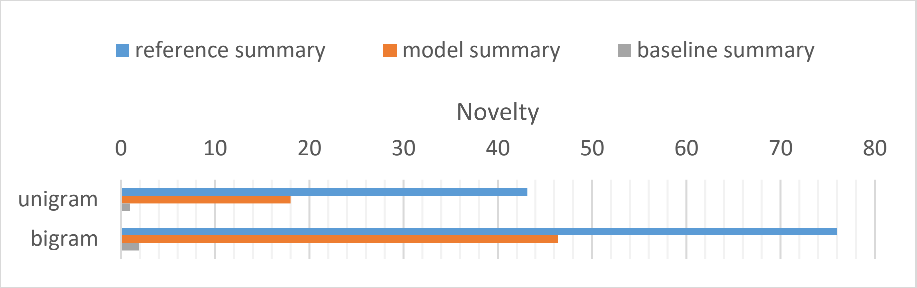


Fig.7 The unigram novelty and the bigram novelty. The horizontal axis indicates novelty.

Table 9 Manual evaluation

	Model-generated (1st summary)			Baseline (2nd summary)			Reference (3rd summary)		
	Len.	**Rel.**	**Read.**	**Len.**	**Rel.**	**Read.**	**Len.**	**Rel.**	**Read.**
Student 1	3.26	6	7.92	4.78	5.84	7.32	2.24	6.54	8.12
Student 2	2.32	4.14	6.6	4.6	5.98	8.26	2.22	6.02	8.2
Average	2.79	5.07	7.26	4.69	5.91	7.79	2.23	6.28	8.16

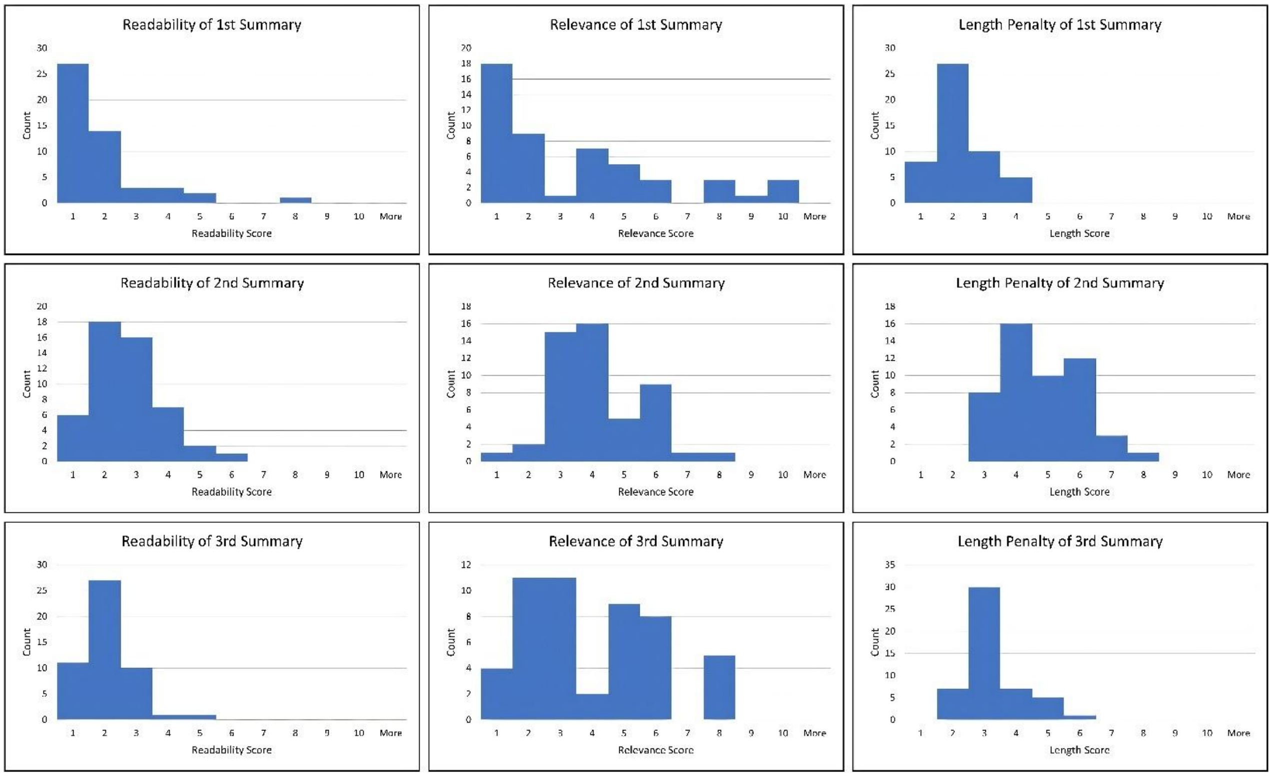


Fig. 8 Manual evaluation results generated by Student 1. The horizontal axis represents the manual evaluation scores, while the vertical axis indicates the frequency count corresponding to each manual score.

The horizontal axis represents the manual evaluation scores, while the vertical axis indicates the frequency count corresponding to each manual score. In these figures, we have presented three types of summaries evaluation: the first being the summary generated by our model, the second being the baseline (lead-3) summary, and the third being the reference summary. According to the evaluation results, it is observed that the lead-3 summaries evaluated to be longer in terms of text length compared to both the reference and generated summaries by both of students. Conversely, the generated summaries closely resemble the reference summaries in terms of length. In terms of readability metrics, the generated and reference summaries exhibit similar levels of readability, approximately. However, there is a discrepancy in perception among the students. Student 1

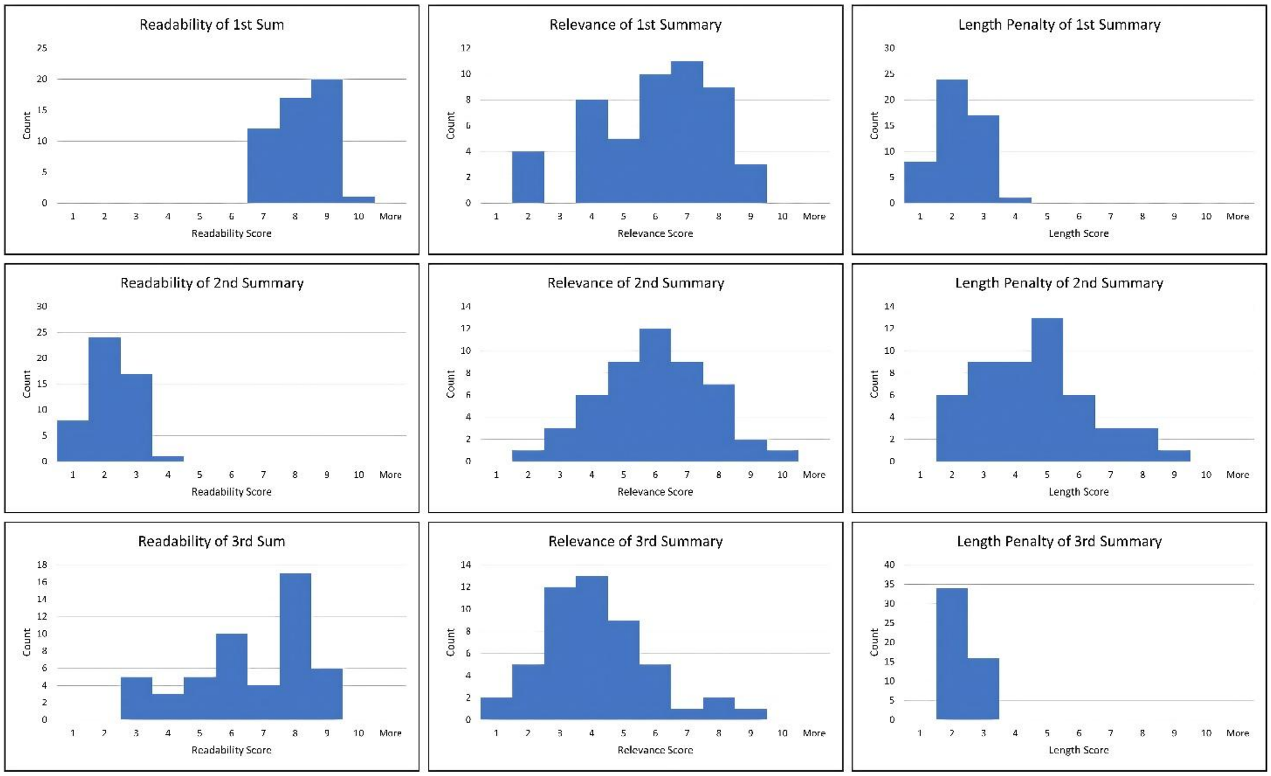


Fig. 9 Manual evaluation results generated by Student 2. The horizontal axis represents the manual evaluation scores, while the vertical axis indicates the frequency count corresponding to each manual score.

considers the lead-3 summary to be more readable, while Student 2 finds it less so. Interestingly, the generated summary demonstrates a readability level that is comparable to both the lead-3 and reference summaries.

Regarding the relevance of the generated summaries, they are generally considered highly relevant. Nevertheless, some of the generated summaries are perceived as less relevant. This discrepancy may be attributed to potential truncation of the input news article text, where the model might not capture the entire context necessary to generate a fully relevant summary. In conclusion, the generated summaries are on par with the reference summaries, suggesting that the model has the capability to produce high-quality summaries for Kurdish news articles and similar texts.

Due to the limited availability of Kurdish students, our manual evaluation was conducted with two students. A two-way ANOVA to analyze the results of manual evaluation was conducted.

The results showed significant differences between the metrics and summaries (p-value = 0.0004), indicating variability in the evaluation criteria. However, the differences between the subjects' ratings were not significant (p-value = 0.178), suggesting consistency in their evaluations. While the small sample size limits the statistical reliability of these results, manual evaluation results are included to provide qualitative insights into the performance of our summarization method. This approach, despite its limitations, highlights the achievements of our work and offers a foundation for future research.

6 Conclusion

The explosive growth of data available on the internet has led researchers to focus on summarizing articles, news reports, books, and other sources of information. In this study, we present a novel approach for creating the first corpus of abstractive text summarization with human evaluation in the Sorani Kurdish language. Additionally, we trained a customized simplified version of the mT5-base transformer model on the created corpus to develop the first abstractive text summarization model for the Kurdish language. The model was evaluated on several important criteria and showed promising results. Furthermore, a baseline model was used to compare the results with those of the proposed model, and potential areas for further improvement were identified. This study provides a foundation for future research on abstractive text summarization in the Kurdish language.

Conflict of Interest

The authors declare no conflict of interest.

Informed Consent Statement

Any research article describing a study involving informed consent was obtained from all subjects involved in the study.

Acknowledgment

This paper is derived from the research project of “Kurdistan Studies Institute” of University of Kurdistan. We express our gratitude and appreciation for providing financial resources.

References

[1] Perea-Ortega, J.M., et al., Application of text summarization techniques to the geographical information retrieval task. Expert systems with applications, 2013. 40(8): p. 2966-2974.

[2] Mohan, G.B. and R.P. Kumar, Lattice abstraction-based content summarization using baseline abstractive lexical chaining progress. International Journal of Information Technology, 2023. 15(1): p. 369-378.

[3] Yadav, D., et al., Qualitative analysis of text summarization techniques and its applications in health domain. Computational Intelligence and Neuroscience, 2022. 2022.

[4] Magooda, A. and D. Litman, Abstractive summarization for low resource data using domain transfer and data synthesis. arXiv preprint arXiv:2002.03407, 2020.

[5] El-Kassas, W.S., et al., Automatic text summarization: A comprehensive survey. Expert systems with applications, 2021. 165: p. 113679.

[6] Widyassari, A.P., et al., Review of automatic text summarization techniques & methods. Journal of King Saud University-Computer and Information Sciences, 2020.

[7] Daneshfar F, Saifee BS, Soleymanbaigi S, Aeini M. Elastic deep multi-view autoencoder with diversity embedding. Information Sciences. 2025 1;689:121482.

[8] Berahmand, K., et al., An Improved Deep Text Clustering via Local Manifold of an Autoencoder Embedding. 2022.

[9] Zhang, J., et al. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. in International Conference on Machine Learning. 2020. PMLR.

[10] 1Raffel, C., et al., Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 2020. 21(140): p. 1-67.

[11] Lewis, M., et al., Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461, 2019.

[12] Daneshfar, Fatemeh. "Enhancing Low-Resource Sentiment Analysis: A Transfer Learning Approach." Passer Journal of Basic and Applied Sciences 6.2 (2024): 265-274..

[13] Muhamad, S.S., et al., Kurdish end-to-end speech synthesis using deep neural networks. Natural Language Processing Journal, 2024. 8: p. 100096.

[14] Ahmadi, S. KLPT–Kurdish Language Processing Toolkit. in Proceedings of Second Workshop for NLP Open-Source Software (NLP-OSS). 2020.

[15] Xue, L., et al., mT5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934, 2020.

[16] Abadji, J., et al., Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. arXiv preprint arXiv:2201.06642, 2022.

[17] Ahmadi, S., H. Hassani, and K. Abedi. A corpus of the Sorani Kurdish folkloric lyrics. in Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL). 2020.

[18] Veisi, H., M. MohammadAmini, and H. Hosseini, Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus. Digital Scholarship in the Humanities, 2020. 35(1): p. 176-193.

[19] Ahmadi, S. Building a corpus for the Zaza–gorani language family. in Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects. 2020.

[20] Abdulrahman, R.O., H. Hassani, and S. Ahmadi, Developing a fine-grained corpus for a less-resourced language: the case of Kurdish. arXiv preprint arXiv:1909.11467, 2019.

[21] Esmaili, K.S., et al. Building a test collection for Sorani Kurdish. in 2013 ACS International Conference on Computer Systems and Applications (AICCSA). 2013. IEEE.

[22] Costa-jussà, M.R., et al., No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672, 2022.

[23] Ahmadi, S., H. Hassani, and D.Q. Jaff, Leveraging Multilingual News Websites for Building a Kurdish Parallel Corpus. Transactions on Asian and Low-Resource Language Information Processing, 2022. 21(5): p. 1-11.

[24] Amini, Z., et al., Central Kurdish machine translation: First large scale parallel corpus and experiments. arXiv preprint arXiv:2106.09325, 2021.

[25] Ahmadi, S. and M. Masoud, Towards machine translation for the Kurdish language. arXiv preprint arXiv:2010.06041, 2020.

[26] Ataman, D., Bianet: A parallel news corpus in turkish, kurdish and english. arXiv preprint arXiv:1805.05095, 2018.

[27] Azin, Z. and S. Ahmadi, Creating an Electronic Lexicon for the Under-resourced Southern Varieties of Kurdish Language. Electronic lexicography in the 21st century (eLex 2021) post-editing lexicography: p. 83.

[28] Aliabadi, P., et al. Towards building kurdnet, the kurdish wordnet. in Proceedings of the Seventh Global Wordnet Conference. 2014.

[29] Kamal, Z. and H. Hassani. Towards Kurdish text to sign translation. in Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives. 2020.

[30] Qader, A. and H. Hassani, Kurdish (sorani) speech to text: Presenting an experimental dataset. arXiv preprint arXiv:1911.13087, 2019.

[31] Gökırmak, M. and F. Tyers. A dependency treebank for Kurmanji Kurdish. in Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017). 2017.

[32] Hameed, R., S. Ahmadi, and F. Daneshfar, Transfer learning for low-resource sentiment analysis. arXiv preprint arXiv:2304.04703, 2023.

[33] Farahani, M., M. Gharachorloo, and M. Manthouri. Leveraging ParsBERT and Pretrained mT5 for Persian Abstractive Text Summarization. in 2021 26th International Computer Conference, Computer Society of Iran (CSICC). 2021. IEEE.

[34] Landro, N., et al., Two New Datasets for Italian-Language Abstractive Text Summarization. Information, 2022. 13(5): p. 228.

[35] Chowdhury, R.R., et al., Unsupervised abstractive summarization of bengali text documents. arXiv preprint arXiv:2102.04490, 2021.

[36] Parida, S. and P. Motlicek. 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). 2019.

[37] Bajaj, A., et al., Long Document Summarization in a Low Resource Setting using Pretrained Language Models. arXiv preprint arXiv:2103.00751, 2021.

[38] Baykara, B. and T. Güngör, Turkish abstractive text summarization using pretrained sequence-to-sequence models. Natural Language Engineering, 2022: p. 1-30.

[39] Shilpa, G. and D. Shashi Kumar, Abs-Sum-Kan an abstractive text summarization technique for an India regional language by induction of tagging rules. Int J Recent Technol Eng (IJRTE), ISSN, 2019: p. 2277-3878.

[40] Žagar, A. and M. Robnik-Šikonja, Cross-lingual transfer of abstractive summarizer to less-resource language. Journal of Intelligent Information Systems, 2022. 58(1): p. 153-173.

[41] Wolf, T., et al., Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771, 2019.

[42] Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. in Text summarization branches out. 2004.

[43] Gehrmann, S., Z. Ziegler, and A.M. Rush. Generating abstractive summaries with finetuned language models. in Proceedings of the 12th International Conference on Natural Language Generation. 2019.

Pedram Yamini received the B.E. degree in Computer Engineering from University of Kurdistan (UoK) IRAN and is currently pursuing the M.Sc. degree in Artificial Intelligence with the Department of Computer Engineering, University of Kurdistan (UoK), IRAN.

Fatemeh Daneshfar is currently an assistant professor with the Department of Computer Engineering, University of Kurdistan (UoK), IRAN. Her current research interests include Machine learning, text mining, natural language processing, and speech processing.

Abouzar Qorbani received the B.Sc. degree in computer engineering from the Mehrastan University, Gilan, Iran, in 2011, and the M.Sc. degree in Information Technology from the Foulad Institute of Technology University, Isfahan, Iran, in 2013. He is currently pursuing the Ph.D. degree with the Department of Computer Engineering, Isfahan University, Isfahan, Iran.