SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
http://saucis.sakarya.edu.tr/
e-ISSN: 2636-8129
Publisher: Sakarya University
Vol. 7, No. 3, 2024, 346-360
DOI: 10.35377/saucis...1504388
SAKARYA UNIVERSITY
RESEARCH ARTICLE

Evaluation-Focused Multidimensional Score for Turkish Abstractive Text Summarization

Nihal Zuhal Kayalı[1,2*], Sevinç İlhan Omurca[2]

[1]Turkish-German University, Faculty of Engineering, Department of Computer Engineering, İstanbul, Türkiye
[2]Kocaeli University, Faculty of Engineering, Department of Computer Engineering, Kocaeli, Türkiye

Corresponding author:
Nihal Zuhal Kayalı, Turkish-German University, Faculty of Engineering, Department of Computer Engineering, İstanbul, Türkiye, nihal.kayali@tau.edu.tr

Article History:
Received: 25.06.2024
Accepted: 11.10.2024
Published Online: 30.10.2024

ABSTRACT
Despite the inherent complexity of Abstractive Text Summarization, which is widely acknowledged as one of the most challenging tasks in the field of natural language processing, transformer-based models have emerged as an effective solution capable of delivering highly accurate and coherent summaries. In this study, the effectiveness of transformer-based text summarization models for Turkish language is investigated. For this purpose, we utilize BERTurk, mT5 and mBART as transformer-based encoder-decoder models. Each of the models was trained separately with MLSUM, TR-News, WikiLingua and Fırat_DS datasets. While obtaining experimental results, various optimizations were made in the summary functions of the models. Our study makes an important contribution to the limited Turkish text summarization literature by comparing the performance of different language models on existing Turkish datasets. We first evaluate ROUGE, BERTScore, FastText-based Cosine Similarity and Novelty Rate metrics separately for each model and dataset, then normalize and combine the scores we obtain to obtain a multidimensional score. We validate our innovative approach by comparing the summaries produced with the human evaluation results.

Keywords: Natural language processing, Abstractive summarization, Transformers, Evaluation metrics, ROUGE

1. Introduction

Text summarization is the task of generating a short and illustrative representation of the source text. The primary motivation of this process is to emphasize not only the main ideas but also the significant details of the source text. Another important point is to filter out unnecessary information from a longer text. Due to the rapidly growing text data resources today, automatic text summarization is needed to save time, explore the content quickly, and draw useful information from huge text data sources. However, accurate text summarization is a challenging task because it requires not only fusing the primary information of the text but also understanding long dependencies, reasoning about the contents, and producing fluent and grammatically correct text [1]. From this perspective, several issues must be considered in a precise and efficient text summarization system. These points include the quality of the training data, the chosen neural language model, text feature specifications, the optimal summary length, the assessment of summarization performance, and the implementation of multilingual capabilities.

The summarization tasks have been extensively studied in the literature and broadly categorized into two groups: Extractive and Abstractive Text Summarization (ATS) [2]. In extractive summarization, sentences are selected from the source text based on their relevance scores. These sentences are chosen to preserve specific characteristics, such as keywords, main ideas, and critical concepts from the original text. Extractive text summarization is often used because it creates a quick and concise summary while preserving a significant portion of the text. On the other hand, new sentences not in the original text and expressing more related content are generated in the abstractive summarization. Therefore, ATS is one of the challenging tasks due to the need for a deeper understanding of text and language generation [3]. Although extractive methods are more commonly used in the literature, it has been observed that abstractive methods can produce higher-quality abstracts compared to extractive methods.

ATS involves linguistic difficulties and is done similarly to how people use cause-effect relationships when describing a text. This summarization method aims to generate a summary that looks like it was created by a human and uses semantic structure instead of structural elements. Achieving semantic understanding is difficult because the generated text needs to maintain proper grammar and fluency, which is currently a significant challenge for existing models [4,5]. Furthermore, the challenge is exacerbated by the variability in human language, where the same idea can be expressed in numerous ways, making it

Cite as: N. Z. Kayalı, S. İlhan Omurca, "Evaluation-Focused Multidimensional Score for Turkish Abstractive Text Summarization," *Sakarya University Journal of Computer and Information Sciences*, vol.7, no. 3, pp. 346-360, 2024. Doi: 10.35377/saucis…1504388

difficult for models to generalize effectively [6]. ATS has the potential to produce high-quality summaries that can generate an innovative summary utterly different from the statements contained in the original text and incorporate external knowledge bases [7]. From this point of view, the fields of Natural Language Processing (NLP), Natural Language Understanding (NLU), and generation tasks are critical. Deep learning methods have recently gained a significant interest in these three research areas. In particular, the attention mechanism [8] and the Transformer model [4] gained massive interest in deep neural networks, especially in sequence-to-sequence (Seq2Seq) tasks. Transformer, which is a network architecture constructed by Vaswani et al. [9], depends on feed-forward networks and a multi-head attention mechanism. Transformer-based models capture semantic connections by leveraging the self-attention mechanism, which allows them to consider the relationship between all words in the text simultaneously. Unlike traditional models like RNNs, Transformers do not rely on sequential processing, which makes them more efficient in capturing long-range dependencies across the entire text. Therefore, they are ideal for summarizing longer texts. In addition, Transformer models can more accurately summarize texts by considering language structure features where word order is important. Beyond that, another significant breakthrough in abstractive summarization tasks is pre-trained language models such as BERT [10] or GPT-3 [11], which are built on the Transformer model.

This study investigated the effectiveness of pre-trained Seq2Seq language models represented by BERTurk, mT5, and mBART in summarizing Turkish texts on four different datasets. We interpreted the experimental results separately for each model and dataset by looking at BERTScore, FastText-based Cosine Similarity, Novelty Rate, and the widely used ROUGE score. We normalized the obtained scores and transformed them into a holistic score called "Multidimensional Score" (MDS). MDS was formed with the dimensions of BERTscore for semantic accuracy, ROUGE for superficial word similarity, FastText Cosine Similarity for word-level semantic closeness, and Novelty Score for novelty and originality. Then, we compared the MDS results with human evaluation and interpreted the findings.

Our study is significant as it contributes to the limited research on Turkish abstractive text summarization using Transformer models, incorporating multiple evaluation metrics. Additionally, we provide a comprehensive comparative analysis by evaluating three different models across four datasets with five evaluation criteria. The study emphasizes evaluation metrics, an area with scarce research, all conducted within a consistent experimental setup. The remaining sections of the article are structured as follows: Section 2 gives the Turkish text summarization literature, Section 3 briefly explains the methods used in our study, Section 4 describes the datasets and evaluation criteria used in our experiments, and Section 5 discusses the quantitative and qualitative results. Lastly, conclusions are given in Section 6.

2. Related Works

Automatic text summarization systems have been a widely studied research area in NLP literature since early times. As a result of the in-depth examination of various ATS studies in the literature, some challenges have been highlighted by the researchers: selecting the most informative sentences from the source text, summarization of long single documents, evaluation of the computer-generated summary without human resources and generating a human-like abstractive summary [12]. While initial efforts in ATS were primarily focused on extractive summarization techniques, abstractive summarization has recently become the center of interest within the research community. This shift is due to its ability to address the challenges previously noted. However, there are few studies in the field of ATS in the Turkish language. The predominant approach in Turkish text summarization studies is extraction-based [13].

The morphological structure of the Turkish language is quite demanding and complex. Therefore, tasks such as automatic summarization and heading generation of Turkish texts become challenging. For example, in Turkish, word roots can be modified with various affixes to acquire different meanings. This makes it difficult to understand and summarize Turkish texts. Moreover, more complex summarization methods, such as abstractive summarization, have yet to be researched due to the difficulty of the Turkish language.

Ülker and Özer pointed out that there is not enough dataset for Turkish text summarization [14]. A TTSD-Turkish Text Summarization Dataset was presented for inferential and abstractive summarization tasks in a study. The results obtained from TextRank, Lexrank, LSA-based, Luhn, and Random methods are compared using the ROUGE evaluation metric. The presented dataset gave successful results in every method.

In their 2021 study, Beken Fikri, Oflazer and Yanıkoğlu stated that existing evaluation metrics for Turkish abstractive text summarization systems are insufficient and presented STSb-TR, the first semantic text similarity dataset developed for Turkish [15]. The presented dataset provided a high-quality translation with machine translation and showed that it can be used without the need for expensive human annotations. The study emphasizes that the ROUGE metric is insufficient in agglutinative languages such as Turkish and that semantic similarity models are more effective in this regard. In particular, it was shown with Pearson and Spearman correlation analyses that the proposed models provided higher correlation with human judgments compared to ROUGE. The quantitative and qualitative analysis results of the study revealed that the proposed models captured semantic equivalence more accurately and that these models can be used as evaluation metrics in Turkish abstractive summarization systems.

Baykara and Güngör investigated human evaluations summarization and heading generation tasks on a Turkish dataset using pre-trained Seq2Seq models [16]. They evaluated the performances of mBART, mT5, and BERT models on TR-News and MLSUM. Monolingual BERT models achieved better results than multilingual BERT models for both of the targeted tasks.

In 2022, a study stated that Turkish NLP studies have made significant progress in recent years, and Turkish NLP has become comparable to other languages [17]. Four new Turkish benchmark datasets have been introduced for Turkish NLP tasks, including language modeling, sentence segmentation, and spelling and correction. In addition, MUKAYESE, a comprehensive benchmark suite containing baselines for Turkish NLP tasks, has also been introduced, including language modeling, machine translation, named entity recognition, sentence segmentation, spelling and correction, summarization, and text classification.

Bech et al. [18] examined the current status of summarization studies in the Turkish language and the difficulties encountered in this field. They performed three different experiments: unsorted, token-based sorted, and novelty-based sorted on a dataset obtained by combining TR-News, WikiLingua, and MLSUM datasets at specific rates. ROUGE and score performances were measured using the mT5 model. It was determined that the token-based ordered model gave a better result than other models.

In another study conducted in 2023, an ATS study was conducted using the T5 model on a dataset containing Turkish news and summaries [19]. The researchers collected the dataset and published it for academic use. The summaries generated by the models were evaluated using the ROUGE score and BERTScore performance metrics. As a result of the evaluation, it was observed that more successful results were obtained compared to the studies in the literature.

Baykara and Güngör addressed the limitations of existing evaluation metrics for abstractive text summarization in morphologically rich languages like Turkish by proposing new evaluation metrics [20]. They pointed out that existing metrics, such as ROUGE and METEOR, are insufficient for assessing the performance of summarization systems in agglutinative languages, as these metrics rely on surface-level n-gram matching. This poses significant challenges, particularly for abstractive summarization, where words can be generated in various forms and enriched with affixes. In their study, they proposed using evaluation metrics that consider morphosyntactic properties and conducted correlation analyses with human judgments by training mT5 and BERTurk models on the TR-News dataset. The results demonstrated that using morphosyntactic tokenization during evaluation led to better alignment with human judgments compared to common metrics like ROUGE and METEOR. This study also emphasizes the importance of preprocessing and the morphosyntactic structure of the language in the evaluation process by presenting a new manually annotated dataset for Turkish.

Yüksel and Çebi published a dataset named "TR-News-Sum" which was created for Turkish summarization systems [21]. Attention Based, Pointer Generator, and Reinforcement Learning methods from Seq2Seq Neural Network models were studied on this dataset. ROUGE-1, ROUGE-2, and ROUGE-L were used as evaluation metrics.

3. Methods and Materials

3.1. Encoder-Decoder Architecture

Encoder-decoder networks are an essential and powerful tool in a wide range of Seq2Seq tasks, playing a crucial role in neural abstractive summarization. Encoder-Decoders consist of two main components: an Encoder that takes a word sequence as input and outputs a context vector, and a Decoder that takes the context vector and predicts the subsequent token in the target summary. While the responsibility of the encoder is to encode the entire sequence into a fixed-length vector called a context vector, the responsibility of the decoder is to decode the context vector into a desired summary.

3.2. Transformers

Recurrent models such as LSTM [22] have long been used for encoder-decoder models in various NLP tasks like text summarization. However, more recently, transformers [9] based on self-attention -which is the primary building block of the Transformer- have started to dominate the research field as state-of-the-art networks, especially for Seq2Seq models. One of the reasons behind this situation is that while transformers can parallelize text processing, recurrent models use sequential text processing over time. Another reason is that recurrent models couldn't handle long text sequences. There are some main qualities that make a transformer not suffer from long dependency issues as much as an LSTM network. Through the Attention mechanism [8], the information at the beginning of the sentence becomes equally well represented in the context vector, especially for long sentences. Beyond that, the attention mechanism can capture the words contributing more information from the whole input sequence. For text summarization systems, this means that some related words in the original text are considered more than nonrelated ones when creating the words in the summary. Another important contribution of transformers is the effectiveness of pre-trained language models such as BERT [10], BART [23], and T5 [24] with transformer structures, which has become evident. We use BERT2BERT, mBART, and mT5 models in our experiments. While BART and T5 use both encoder and decoder components, BERT uses encoders only.

3.3. BERTurk

Based on a masked language model, BERT is a contextualized text representation model that undergoes pre-training with a bidirectional transformer encoder architecture [9] BERT2BERT architecture uses a public BERT checkpoint to initialize the

encoder and again chooses the BERT model as the decoder for text generation. The encoder-decoder attention is randomly initialized [25]. When initiated with BERT decoder checkpoints, it autonomously generates summary text, much like Transformers [25], utilizing BERT's predictive capability for masked tokens with bidirectional text representations as its input. In our study, we use a publicly available checkpoint, BERTurk [26], which is a monolingual Turkish BERT model.

3.4. mT5

In their work, Raffel and colleagues [24] introduced the transformer-based T5 (Text-to-Text Transfer Transformer) framework with the intention of treating all text processing problems as 'text-to-text' challenges. This model uses a standard encoder-decoder Transformer as proposed by [9]. As in the BERT-BASE [10] configuration, there are 12 blocks in both the encoder and decoder, with each block composed of two basic components: a self-attention layer and a feed-forward network. A multilingual variation of the T5 model, which is trained on common crawl-based data and covers 101 different languages, is called mT5 [27]. Our studies used the mT5 model as the second language model after the BERTurk model.

3.5. mBART

BART, functioning as a denoising autoencoder in the pretraining of sequence-to-sequence models, excels notably when fine-tuned, especially in the context of text generation tasks. Its architecture is built with a bidirectional encoder and an auto-regressive decoder. Corrupting the original text with a noise function and then reconstructing the original text by a Seq2Seq model learning are two stages of Pretraining in this model [23]. In our study, we used mBART (Multilingual BART), which is a multilingual version of the BART, to fine-tune the datasets. While BART has been pre-trained only for English, mBART, utilizing the same BART architecture, has undergone large-scale monolingual pretraining on multiple languages [28].

3.6. Suggested Method

Our proposed summarization model encompasses the multi-dimensional approach in the performance evaluation process along with the improvement of the text summarization process using the advanced set of parameters. Each of the BERTurk, mT5, and mBART models mentioned in the previous sections is trained individually and separately for processing the dataset. After training the models, the "max_length" and "min_length" parameters are used to determine the maximum and minimum lengths of the summaries to be produced when summarizing the texts in the test dataset. This ensures that the summaries are short enough and strike a balance with the requirements of the content. The "num_beams" parameter specifies the number of beams used in the beam search algorithm, which contributes to a more comprehensive and accurate summarization. "no_repeat_ngram_size" prevents repetitive n-grams (groups of words) from being generated by the model, which increases the diversity and uniqueness of the text. "repetition_penalty" and "length_penalty" control how the model handles repetitions and length. early_stopping=True allows the model to stop summarizing when it reaches a good result. Combining these parameters improves the text summarization process's accuracy and efficiency while improving the quality of the model's output. We propose a comprehensive, multidimensional approach to evaluate these outputs using MDS: Rouge score, BERTScore, Novelty Rate, and FastText-based Cosine Similarity. We compare the MDS results with human evaluation and analyze its usability instead of human evaluation.

4. Implementations and Experiments

4.1. Datasets

In our experiments, four public datasets, MLSUM [29], TR-News [16], WikiLingua [30], and Fırat_DS [19], were used.

The MLSUM (Multi-Language Summarization) dataset is a multilingual large-scale summarization dataset containing more than 1.5 million articles/abstracts from online newspapers in five different languages: Turkish, French, German, Spanish, and Russian. The Turkish news set of MLSUM was taken from a news website. MLSUM dataset has 249277 news items and summary pairs in train, 11565 in validation, and 12775 in test.

The TR-News dataset is a monolingual dataset consisting of Turkish news taken from popular news websites between 2009-2020. The data set includes the news texts' URL, title, summary, content, subject, tags, date, author, and source information. TR-News dataset has 277573 news items and summary pairs in the train part, 14610 in the validation part, and 15379 in the test part.

WikiLingua is a large-scale dataset that can be used for NLP tasks such as summarizing in various languages and extracting semantics from text. The dataset includes summaries and full texts of articles from WikiHow, which is a high-quality data source that provides “how-to” guides covering different topics by various authors. WikiLingua has been prepared in 18 languages to increase language diversity and offer a rich resource for multilingual NLP models.

The dataset we refer to as Fırat DS, which we use in the experiments, is the dataset named "Text Summarization-Keyword Extraction Dataset" made available by the Fırat University Big Data and Artificial Intelligence Laboratory.

We applied the following preprocessing steps to normalize all documents before summarization. The duplicate lines and the lines with blank abstract or text content were deleted. The noise characters, such as unnecessary characters, numbers, and punctuation marks, were cleaned. The tokenization, which is the essential process of dividing text into smaller units, is

realized. All letters in the text are converted to lowercase. In the model training, the datasets are divided into 90% as a training subset, 5% as a validation subset, and 5% as a test subset.

4.2. Evaluation Metrics

Using more than one metric to measure the quantitative results may provide more information about summary quality. That's why we made a comprehensive assessment using different evaluation metrics commonly used in text summarization. In addition to the ROUGE scores, the BERTScore, the Novelty Rate, and FastText-based Cosine Similarity between the original text and summary have also been reported and discussed.

One of the most popular evaluation measures used in summarization systems is the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) performance measure [31]. ROUGE is an n-gram metric that measures the overlapping n-gram units between the reference and model-generated summaries. The F-scores of ROUGE-1 (unigram), ROUGE-2 (bi-gram), and ROUGE-L (the longest common sequence) are reported. ROUGE-L is a Longest Common Subsequence (LCS) based ROUGE metric. LCS automatically identifies the longest co-occurrences in sequence n-grams, naturally taking into account structure similarity at the sentence level. The formula used to calculate the ROUGE-N score is given in equation (1), and the formulas used to calculate the ROUGE-L score are given in equations (2), (3) and (4).

$$\mathbf{ROUGE} - \mathbf{N} = \frac{\sum_{\mathbf{S}\in\{\mathbf{ReferenceSums}\}}\sum_{\mathbf{gram_m}\in\mathbf{S}}\mathbf{Count_{match}(gram_m)}}{\sum_{\mathbf{S}\in\{\mathbf{ReferenceSums}\}}\sum_{\mathbf{gram_m}\in\mathbf{S}}\mathbf{Count(gram_m)}} \tag{1}$$

$$\boldsymbol{P_{LCS}} = \frac{\boldsymbol{LCS(ModelSum, ReferenceSum)}}{\boldsymbol{n}} \tag{2}$$

$$\boldsymbol{R_{LCS}} = \frac{\boldsymbol{LCS(ModelSum, ReferenceSum)}}{\boldsymbol{m}} \tag{3}$$

$$\boldsymbol{F_{LCS}} = \frac{(\mathbf{1}+\boldsymbol{\beta^2})\times \boldsymbol{R_{LCS}} \times \boldsymbol{P_{LCS}}}{\boldsymbol{R_{LCS}} + (\boldsymbol{\beta^2}\times \boldsymbol{P_{LCS}})} \tag{4}$$

"m" is the length of sequences of the model summary, and "n" is the length of sequences of the reference summary. When calculating the F score, the β parameter controls the P importance of recall R and sensitivity. The F score is the harmonic mean of recall and precision. By setting the β value, recall or precision can be prioritized in the evaluation. When β is set to 1, recall and precision are weighted equally, resulting in a balanced F score. If one wants to give more importance to recall, stating that it is more important that most of the n-grams in the references are found in the candidates, one can increase the value of β to a value greater than 1. In this study, β is chosen to be 1.

We also used BERTScore [32] to evaluate our experiments. BERTScore computes a semantic similarity score by interpreting the reference and model-generated summaries. Unlike ROUGE, BERTScore measures text similarity by considering semantic similarity in addition to word-level similarity. Therefore, reporting ROUGE and BERTScore together is essential for a more detailed analysis of a text summarization system. The BERTScore-based precision, recall, and F-score are given in equations (5), (6), and (7), respectively.

$$\boldsymbol{P_{BERTScore}} = \frac{\mathbf{1}}{|\hat{\boldsymbol{x}}|}\sum_{\hat{\boldsymbol{x}}_j\in\hat{\boldsymbol{x}}} \boldsymbol{max}(\boldsymbol{x_i}\in\boldsymbol{x})\, \boldsymbol{x_i^T}\hat{\boldsymbol{x}}_j \tag{5}$$

$$\boldsymbol{R_{BERTScore}} = \frac{\mathbf{1}}{|\boldsymbol{x}|}\sum_{\boldsymbol{x_i}\in\boldsymbol{x}} \boldsymbol{max}(\hat{\boldsymbol{x}}_j\in\hat{\boldsymbol{x}})\, \boldsymbol{x_i^T}\hat{\boldsymbol{x}}_j \tag{6}$$

$$\boldsymbol{F_{BERTScore}} = \mathbf{2}\times\frac{\boldsymbol{P_{BERTScore}}\times \boldsymbol{R_{BERTScore}}}{\boldsymbol{P_{BERTScore}} + \boldsymbol{R_{BERTScore}}} \tag{7}$$

BERTScore converts the words in two texts that are compared into high-dimensional vectors using a BERT model. Each word is transformed into a vector through the model: BERT(x_i) and BERT($\hat{x}_j$), which is the vector representation of each word. These vectors are representations that capture the meaning of each word in a specific context. In the BERTScore formulas, "x" represents the reference summary representations, and "$\hat{x}$" represents the candidate summary representations. Precision and recall metrics are calculated for BERTScore by comparing each token representation "x_i" of the reference summary with each token representation "$\hat{x}_j$" of the candidate summary.

To evaluate text summarization in a broader context, we also used FastText-based Cosine Similarity to compare summaries with each other. FastText is a library developed by Facebook AI Research (FAIR), and its primary purpose is to produce word and sentence representations quickly and effectively for NLP tasks [33]. This library provides word embeddings for each word in the summary, representing them as high-dimensional vectors based on their semantic meaning to capture word relationships and similarities. Each word in the reference and generated summaries is mapped to its corresponding vector using FastText. The word vectors are averaged to form a single vector representing the entire summary. The cosine similarity between the two summary vectors is calculated.

In abstractive summarization, evaluating the abstractness level (text novelty) of reference summaries in data sets and the summaries created is important. The calculation of the Novelty Rate typically begins after the summarization process is complete by examining each sentence or phrase within the summary. This involves checking whether the original text's sentences, phrases, or n-grams already exist. If they are not exactly found in the original text, they are considered as new content. The Novelty Rate is calculated by dividing the number of new n-grams, sentences, and phrases in the entire summary by the total number of n-grams, sentences, and phrases. This metric, which is used to calculate Novelty Rates in detail for each data set to evaluate whether the abstract is creative and original, is presented in Formula 8.

$$\boldsymbol{Novelty\ Rate} = \frac{\boldsymbol{Count\ of\ Novel\ n-grams\ in\ Reference\ Summaries}}{\boldsymbol{Count\ of\ n-grams\ in\ Reference\ Summaries}} * \mathbf{100} \quad (8)$$

Unfortunately, the text can be assessed using various metrics, including ROUGE and BERTScore, which may describe an overlap between the reference and the generated summaries. Nevertheless, human evaluation should always be part of the process. Human evaluation adds an additional touch in determining the coherence, style, and context of the summaries, which most automated systems cannot provide. Humans can judge the subtleties of language, including irony, humor, and emotional tone, which computerized systems might overlook. Moreover, human evaluators can assess summaries' factual accuracy and overall quality, ensuring they are not only statistically like reference summaries but also meaningful and informative to readers. Given these considerations, our study has also incorporated this metric, reinforcing our findings with a comprehensive view that blends algorithmic precision with human insight. In addition, in order to observe whether the MDS results can replace human evaluation, the results of both were compared with each other.

5. Results

Our study was conducted on a server with 16 Core AMD Ryzen Thread ripper 1950X 16-Core Processor CPU, 32 GB RAM, and Quadro GV100 GPU graphics card with Ubuntu Server operating system. The models were trained with the datasets for three epochs, and each model took an average of 70 hours to train using the datasets. The Adam optimizer was utilized with a learning rate of 1e-3. We used cross-entropy loss for training, which is calculated by comparing the generated output with the reference summary. To prevent overfitting, a dropout rate of 0,1 was applied in the embedding and attention layers of the encoder. Additionally, Layer Normalization was extensively used throughout the encoder to stabilize and enhance the training process. Furthermore, a specific strategy was employed to improve summarization performance: the length of the input text was analyzed, and 10% of this value was set as the "min_length" parameter for the model's summary generation. This approach aimed to ensure that the generated summaries were both concise and adequately detailed, adapting to the input text's length to enhance overall summarization quality.

In our experiments, a method has been employed to enhance the summarization performance of the models. The length of the input text to be summarized has been determined, and 10% of this value has been set as the “min_length” parameter for the model's summary generation function. With this approach, it is aimed to generate more concise and comprehensive summaries by adjusting the length of the generated summary according to the length of the input text.

In evaluating summarization models, it is important to determine the level of abstraction between the reference summary and the model-generated summary. Because of that, our study discussed the novelty dimension in the text summarization task. The Novelty Rates of summaries are reflected as percentages in Table 1, and these results are calculated according to the rates between the whole content of the news text, the reference summary, and the generated summary.

Table 1. The Obtained Novelty Rates for Datasets

		BERTurk			**mT5**			**mBART**		
		1-gram	**2-gram**	**3-gram**	**1-gram**	**2-gram**	**3-gram**	**1-gram**	**2-gram**	**3-gram**
TR-News	**Content / Reference Summary**	0,410	0,655	0,763	0,410	0,655	0,763	0,410	0,655	0,763
	Content / Generated Summary	0,118	0,267	0,393	0,088	0,185	0,279	0,121	0,240	0,358

MLSUM	**Content / Reference Summary**	0,386	0,617	0,718	0,386	0,617	0,718	0,386	0,617	0,718
	Content / Generated Summary	0,123	0,285	0,432	0,095	0,211	0,315	0,121	0,245	0,360
WikiLingua	**Content / Reference Summary**	0,510	0,872	0,965	0,510	0,872	0,965	0,510	0,872	0,965
	Content / Generated Summary	0,071	0,144	0,202	0,060	0,117	0,166	0,089	0,155	0,231
Firat_DS	**Content / Reference Summary**	0,390	0,611	0,714	0,390	0,611	0,714	0,390	0,611	0,714
	Content / Generated Summary	0,120	0,275	0,410	0,205	0,300	0,399	0,299	0,387	0,503

In a summarization task, the newer words and phrases the generated summary contains compared to the original text, the higher the Novelty Rate. In other words, a high Novelty Rate score means that the generated summary contains different words and sentence structures from the original text content and its reference summary. The results in Table 1 reflect the Novelty Rates of each transformer model on each dataset. In Table 1, the lines expressed as "content-reference summary" for each dataset reflect the Novelty Rate between the original text content and the reference summary. Similarly, the lines expressed as "content-generated summary" for each data set reflect the Novelty Rate between the original text content and the model-generated summary. If we examine these results obtained on a 1-gram scale, it is seen that the Novelty Rates obtained with BERTurk, mT5, and mBART models are close to the Novelty Rates in the reference summaries. It is observed that the mT5 model mainly provides a higher Novelty Rate compared to the mBART and BERTurk models.

In our experiments, we use FastText-based Cosine Similarity scores to evaluate the term similarities between the original text and summaries. Table 2 illustrates the Cosine Similarity scores.

Table 2. FastText-based Cosine Similarity scores

Datasets	**Models**	**Content - Reference Summary**	**Content-Generated Summary**	**Reference Summary – Generated Summary**
TR-News	**BERTurk**	0.838	0.886	0.706
	mT5	0.838	0.791	0.748
	mBART	0.838	0.792	0.750
MLSUM	**BERTurk**	0.833	0.770	0.733
	mT5	0.833	0.809	0.734
	mBART	0.833	0.825	0.718
WikiLingua	**BERTurk**	0.888	0.878	0.875
	mT5	0.888	0.905	0.841
	mBART	0.888	0.908	0.871
Firat_DS	**BERTurk**	0.907	0.916	0.889
	mT5	0.907	0.934	0.860
	mBART	0.907	0.949	0.908

The score between Content and Generated Summary indicates how similar the summary generated by the model is to the original text content. A high-performance score implies that the model has skillfully captured the core concepts present in the original text content. When Table 2 is examined, it is seen that these scores are higher than both the "Content-Reference summary" similarity scores and "Reference Summary-Generated Summary" similarity scores. When the obtained results are analyzed, it is concluded that BERTurk and mBART performed the best on both TRNEWS and MLSUM datasets, with the mBART model achieving the highest scores on the MLSUM dataset. The mT5 model, on the other hand, achieved lower scores than BERTurk and mBART.

Table 3 and Table 4, respectively, present the ROGUE and BERT scores from all models for a more comprehensive comparison.

Table 3. ROUGE Scores

		BERTurk			**mT5**			**mBART**		
		R1	R2	RL	R1	R2	RL	R1	R2	RL
TR-News	P	0.388	0.291	0.338	0.390	0.269	0.344	0.280	0,176	0,223
	R	0.725	0.569	0.652	0.527	0.369	0.472	0,494	0,309	0,399
	F1	**0.490**	**0.372**	**0.431**	0.438	0.303	0.391	0,343	0,217	0,275
MLSUM	P	0.362	0.223	0.288	0.410	0.246	0.359	0.371	0.230	0.300
	R	0.503	0.301	0.393	0.399	0.230	0.345	0.419	0.253	0.340
	F1	**0.410**	**0.252**	0.325	0.392	0.228	**0.344**	0.378	0.232	0.307
WikiLingua	P	0.433	0.146	0.272	0.421	0.145	0.253	0.356	0.114	0.203
	R	0.210	0.070	0.134	0.236	0.081	0.141	0.279	0.087	0.161
	F1	0.271	0.091	0.171	**0.293**	**0.101**	**0.177**	0.286	0.089	0.163
Firat_DS	P	0.327	0.222	0.278	0.354	0.208	0.278	0.311	0.212	0.260
	R	0.594	0.434	0.521	0.527	0.320	0.412	0.561	0.388	0.475
	F1	0.407	**0.282**	**0.350**	**0.412**	0.247	0.324	0.387	0.265	0.325

In Table 3, precision (P), recall (R), and F1-scores for ROGUE-1 (R1), ROUGE-2 (R2), and ROUGE-L (RL) are reflected, and the highest F1-scores are highlighted. Considering the ROGUE scores, it is realized that the BERTurk model gives more successful results for datasets other than WikiLingua. To provide an example of other striking results, the precision values of the mT5 model had the higher results.

Table 4. BERT Scores

		BERTScore		
		Precision	Recall	F1
TR-News	**BERTurk**	0.648	0.805	**0.716**
	mT5	0.648	0.719	0.680
	mBART	0.544	0.650	0.589
MLSUM	**BERTurk**	0.611	0.686	0.646
	mT5	0.663	0.661	**0.661**
	mBART	0.611	0.650	0.627
WikiLingua	**BERTurk**	0.598	0.589	**0.592**
	mT5	0.599	0.566	0.580
	mBART	0.565	0.598	0.579

Firat_DS	**BERTurk**	0.581	0.733	**0.645**
	mT5	0.589	0.700	0.638
	mBART	0.589	0.720	**0.645**

Table 4 concludes that BERTurk, mT5, and mBART transformers have similar summarization abilities. However, BERTurk has the highest BERTScore value in all datasets except MLSUM.

Proper evaluation of text summarization systems requires incorporating qualitative and quantitative analysis. For the qualitative analysis, in Table 5 and Table 6, we present some sample summaries generated by four models in terms of semantic similarity to the original text contents. While these results are reflected, new words synthesized by the models or derived by the suffixes are highlighted in bold.

Table 5. Sample Summaries From TR-News and MLSUM Datasets.

TR-News	**MLSUM**
Original Text Content	**Original Text Content**
"Fransa'nın girişimiyle düzenlenen konferansın amacı, kimyasal silah kullanımını engellemek. 30 ülkenin katılımıyla düzenlenen konferansta Türkiye'yi Dışişleri Bakanı Mevlüt Çavuşoğlu temsil ediyor. Dışişleri Bakanı Mevlüt Çavuşoğlu'nun toplantı kapsamında ikili görüşmelerde bulunması da bekleniyor. Konferans sonunda katılımcı ülkeler, kimyasal silah kullanan ya da geliştirenlerin cezasız kalmamalarına yönelik bir dizi taahhütte bulunacak. Ayrıca ülkeler, kimyasal saldırıların failleri hakkında elde ettikleri bilgileri uluslararası kuruluşlarla paylaşma konusunda da mutabakata varacak."	"KAYSERİ'nin Develi ilçesine bağlı Yukarı Künye Mahallesinde bulunan Homurlu Şelalesi'ne ailesiyle birlikte gelen 18 yaşındaki Serkan Turan Timurbuğa, iddiaya göre fotoğraf çektirmek için şelalenin üst tarafındaki kayalıklara çıktı. Fotoğraf çekinirken dengesini kaybeden Timurbuğa, dereye düşerek gözden kayboldu. Yakınlarının haber vermesi üzerine olay yerine AFAD, jandarma, 112 ve gönüllü Türkuaz Arama Kurtarma ekipleri sevk edildi. AFAD ekipleri, jandarma ile Turkuaz ekipleri tarafından sürdürülen arama kurtarma çalışmaları devam ediyor."
Reference Summary	**Reference Summary**
"Dışişleri Bakanı Mevlüt Çavuşoğlu, "Kimyasal silah kullanımının cezasız kalmaması için uluslararası ortaklık" konulu dışişleri bakanları toplantısına katılacak."	"KAYSERİ'de bulunan Homurlu Şelalesi'nde kayalıklarda fotoğraf çekinirken dengesini kaybederek dereye düşen genci arama çalışmaları sürüyor."
Summary generated with BERTurk	**Summary generated with BERTurk**
"Fransa'nın girişimiyle düzenlenen **konferansta** Türkiye'yi temsil **eden** 30 ülkenin katılımıyla konferansta kimyasal silah **kullanımı** konusunda mutabakata **varacağı bildirildi**."	"KAYSERİ'de **bir kişi** fotoğrafı çektirmek için **şelaleye** çıktı. Ancak dengesini **kaybetti** ve gözden kayboldu."
Summary generated with mT5	**Summary generated with mT5**
"Fransa'nın girişimiyle düzenlenen konferansta Türkiye'yi Dışişleri Bakanı Mevlüt Çavuşoğlu temsil ediyor."	"Fotoğraf çektirmek için şelalenin üst tarafındaki kayalıklara **çıkan** 18 yaşındaki **genç** gözden kayboldu."
Summary generated with mBART	**Summary generated with mBART**
"Fransa'nın girişimiyle gerçekleştirilen konferansta Türkiye'yi Dışişleri Bakanı Mevlüt Çavuşoğlu temsil ediyor."	"Kayseri'nin Develi ilçesine bağlı Yukarı Künye Mahallesinde bulunan Homurlu Şelalesi'ne ailesiyle birlikte çıkan genç adam gözden kayboldu."

Table 6. Sample Summaries From WikiLingua and Firat_DS Datasets

WikiLingua	**Firat_DS**
Original Text Content	**Original Text Content**
"Kötü şans getiren şeylere ilişkin çoğu batıl inancı herkes bilir ama belli başlılarını tekrarlamak iyi bir fikir. Bu şekilde kötü şanstan kaçmak için stratejik olarak davranışlarını değiştirmeye çalışabilirsin. Kötü şanstan kaçış olmasa da işaretleri tanıyabilecek, tersine çevirmek için hemen harekete geçebileceksin. Kötü şansın bazı yaygın belirtileri şunlardır: Ayna kırmak – bunun yedi yıl kötü şans getirdiği söylenir. Karga görmek – karşına karga çıkmasının kötü şans getirdiği söylenir. Ama karşına iki karga çıkarsa kötü şans tersine döner. Merdiven altından geçmek – bunun kötü şans getirdiğine inanılır çünkü duvara dayanan merdiven üçgen oluşturur – bu, Kutsal Üçlü'nün yani Baba, Oğul ve Kutsal Ruh'un simgesidir. Üçgenin içinden geçmekle kutsal zemini bozmuş olursun. Kendine "uğursuzluk getirmek" – bu, sana olacağını düşündüğün kötü bir şeyi yüksek sesle söylemek demektir. Bir nevi kadere meydan okumaktır. Bunu tersine çevirmek için masaya ya da herhangi bir zemine 3 kez vur ama vurma sesinin duyulduğundan emin ol. Opal taşı takmanın kötü şans getirdiğine inanılır – tabii eğer Ekim'de doğmadıysan. Kaldırımdaki çatlaklara basmak. Eski bir deyişin söylediği gibi: "Bir çatlağa basarsan kader, annenin belini kırar!" Karşına kara kedi çıkmasının kötü şans getirdiği söylenir – bu batıl inanç, kedilerin cadılarla ve büyüyle olan bağlantısından gelir. İçeride şemsiye açmanın kötü şans getirdiği düşünülür – bu, gölge için şemsiye kullanan Eski Mısırlılardan gelen bir batıl inançtır. O zamanlarda içeride şemsiye açmanın Güneş Tanrısı'na hakaret olduğuna inanılırdı. Bazı batıl inançlar daha az bilinir. Böyle batıl inançlar hakkında okumak ve belli eylemlerin yaratacağı risklerin farkında olmak iyi fikir. Yoksa geri dönüşü olmayan bir şekilde kendine kötü şans getirebilirsin…"	"Yangın akşam saatlerinde Efeler Mahallesi 2296 sokaktaki Sağlık Evleri sitesi G Blok 2. Katta meydana geldi. Edinilen bilgiye göre; Seyhan T.'ye ait evde kiracı olarak oturan aile mutfakta olduğu sırada salon bir anda yanmaya başladı. Kısa sürede yangının büyümesi üzerine dairede oturanlar evden çıkarak canlarını kurtardı. Elektrik kontağından çıktığı düşünülen yangına itfaiye ekipleri hemen müdahale etti. Binanın doğalgazlı olması nedeniyle bölgeye ilgili ekipler de çağrıldı. İtfaiye ekipleri bir yandan yangını söndürmeye çalışırken, diğer yandan da binanın doğalgaz ve elektrikleri kesildi. Aydın Büyükşehir Belediyesi İtfaiye Dairesi Başkanı Serdar Adanır'ın da söndürme çalışmalarına katıldığı yangında, polis ekipleri de güvenlik önlemi aldı, sağlık ekipleri de her ihtimale karşı hazır bulundu. Söndürme çalışmaları tamamlandıktan sonra tahliye edilen binanın sakinleri elektrik ve doğalgazın açılması ile evlerine yerleşti. Polis olayla ilgili soruşturma başlattı."
Reference Summary	**Reference Summary**
"Kötü şans getiren eylem ya da durumlardan kaçın. Daha az bilinen batıl inançlar hakkında bilgi edin. Ölüm ya da hastalık habercisi olan batıl inanışlara karşı dikkatli ol., Tuz kullan. Kırık ayna parçalarını atma. Tütsü yak. Koruyucu tılsım taşı. Adaçayı yak. Kristal ya da taş kullan. Sevap işle. Çakralarını taze çiçeklerle arındır. Dua et. Evini topla. Işığın gücünü kullan. Başka bir ülkeye seyahat et., Şansın dönmeye başladığında bunu fark et."	"Aydın Efeler ilçesinde akşam saatlerinde çıkan yangın paniğe neden oldu. Kısa sürede tüm daireyi saran yangın binanın diğer katlarına sıçramadan söndürüldü. Yangın çıkan binada oturanlar ekipler tarafından tahliye edilirken, yangında yaralanma ve can kaybı yaşanmaması teselli kaynağı oldu."
Summary generated with BERTurk	**Summary generated with BERTurk**
"Kötü şans getiren şeylere ilişkin kontrol et. İyi şansın kötü şans getirdiğine inanılmaya çalış., Bir batıl inanç hakkında okumak ya da yazmaktan kaçın. Olumsuzluklara karşı uyanık ol. Başkalarına yardım etmekten hoşlandığını söyle. Hoşlandığın birini seç."	"Aydın'ın Efeler **ilçesinde** bir apartmanın mutfağında çıkan yangın itfaiye ekipleri tarafından söndürüldü. Alevlerin büyümesi üzerine dairede oturanlar canlarını kurtarıp, evlerinden çıktı. Yangında ölen ya da yaralanan olmazken, olay yerine çok sayıda itfaiye ekibi sevk edildi."

Summary generated with mT5	**Summary generated with mT5**
"Kötü şanstan kaçmak için işaretleri tanı. Kendine "uğursuzluk getirme" sesini söyle. Karga görmek için masaya ya da herhangi bir zemine 3 kez vur., İçeride şemsiye aç. Bir kedi var mı diye düşün. İyi bir inançla bağlantıyı ortaya koy."	"**Aydın'ın** Efeler **ilçesindeki** Sağlık Evleri sitesi G Blok 2. Katta yangın meydana geldi. İtfaiye ekipleri yangını söndürmeye çalışırken, diğer yandan da binanın doğalgazlı olması nedeniyle bölgeye ilgili ekipler çağrıldı."
Summary generated with mBART	**Summary generated with mBART**
"Kötü şans getiren şeyleri tekrarlamaktan kaçın. Kendine meydan oku. Başka kötü niyetli davranışları görmezsen batıl inançlara karşı dikkatli ol., Güneş Tanrısı'na hakaret etme. Bir şemsiye açmanın kötü şans getirdiğini unutma. Kıskançlık ve öfkelilik arasındaki çizgiyi öğren. Dikkatin dağılırsa bir kart oyunu oynamayı dene. Herşeyden haberdar olup olmadığını bil. Hayatına devam et. Başarısız olduğun şeyin altında yatan nedenlerin neler olduğuna karar ver. Arada bir kendini suçlu hissettir. Kişisel hataların üstesinden gel."	"Bursa'nın Efeler **Mahallesinde** evde kiracı olarak oturan aile mutfakta çıkan yangına itfaiye ekipleri müdahale etti. Yangını **söndürme çalışmaları devam ederken**, diğer yandan da binanın doğalgaz ve bir anda yanmaya başladı. Kısa sürede yangının büyümesi üzerine dairede oturanlar evden çıkarak canlarını kurtardı."

When the results in Table 6 are examined for the TR-News dataset, it is observed that the summary generated by the BERTurk is more informative than the other models in terms of the news content and the subject of the meeting mentioned in the original summary. However, since it does not include the name of the person who will attend the meeting, the sentence is semantically incomplete. It has been observed that the summaries produced by mT5 and mBART are inferential; that is, they can combine the sentences in the original text without making many changes.

The sample news content of the MLSUM dataset was summarized by all models as meaningful and compatible sentence structures. The word "one person" is in the summary of the BERTurk model; the word "young" in the mT5 summary is generated by the model. In the summary produced by the mBART model, the "young man" is entirely accurate information produced by the model. The model produced the word "man" by inferring that the person in the news text was male from the name. Similarly, it produced the word "young" from the knowledge of his age.

As the WikiLingua dataset is translated from another language, the sentence structures are influenced by the language structure of the original text. This situation caused the translation effects to be reflected in the flow of the text by different grammar rules or word order between some languages. When the texts in the WikiLingua database were analyzed, it was observed that most of the sentences ended with words with imperative or infinitive endings. While the texts in other news content datasets use more objective and informative language regarding grammar and sentence structure, WikiLingua contains more colorful and narrative elements. These differences directly affected the summarization performance of the models. As a result, although all three models could capture some concepts from the original text accurately, each contained significant misunderstandings, inconsistencies, and information outside the original text. The quality of the summaries is relatively lower due to word errors and misunderstandings.

The BERTurk model for the Firat_DS dataset produced the closest result to the original summary but omitted some crucial details. The mT5 and mBART models produced summaries with some incorrect information. It should be remembered that summarization models often fail to convey the details in the text entirely and accurately and even create false information. The BERTurk summary has retained the essential elements of the original summary. The information that the fire broke out in an apartment, the firefighters intervened, and the residents were evacuated was accurately summarized. The mT5 summary accurately stated the location of the building where the fire occurred. In addition, there is information that firefighters are trying to extinguish the fire and that the relevant teams are called because it is natural gas. However, the mT5 summary does not explain why the fire broke out or the residents' condition. The mBART summary accurately summarized where the fire started, where firefighters intervened, and where extinguishing efforts are ongoing. However, this summary contains incorrect information. For example, the phrase "Efeler Mahallesi of Aydın" in the original text has been changed to "Efeler Mahallesi of Bursa" in the mBART summary. In addition, the phrase "natural gas and suddenly it started to burn" was added to the mBART summary, providing a detail that was not included in the original text.

After all evaluation metrics were calculated and analyzed separately, the scores required to calculate the MDS (Multi-Dimensional Score) were collected. The pre-normalization scores prepared using the F1 values calculated according to the 1 and 2-gram values of all metrics used are presented in Table 7. Normalization is necessary to ensure that metrics of different scales contribute fairly, and to maintain consistency in the results when combined in MDS calculations. The formula used in the normalization calculation of the values in our study is presented in equation (9). " x " represents the original metric value, "x_{min}" represents the lowest value observed for the relevant metric, "x_{max}" represents the highest value observed for the relevant metric, and "x_n" represents the normalized value between 0 and 1. The final versions of the values presented in Table

7, normalized between 0 and 1, are presented in Table 8. Additionally, in Table 7 and Table 8, "R" stands for Rouge score, "B" stands for BERTScore, "F" stands for FastText-based Cosine Similarity, and "N" stands for Novelty Rate.

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{9}$$

Table 7. Scores of Evaluation Metrics Before Normalization

Datasets	Models	1 GRAM				2 GRAM			
		R	B	F	N	R	B	F	N
TR-News	BERTurk	0.490	0.716	0.706	0,118	0.372	0.716	0.706	0,267
	mT5	0.438	0.680	0.748	0,088	0.303	0.680	0.748	0,185
	mBART	0,343	0.589	0.750	0,121	0,217	0.589	0.750	0,240
MLSUM	BERTurk	0.410	0.646	0.733	0,123	0.252	0.646	0.733	0,285
	mT5	0.392	0.661	0.734	0,095	0.228	0.661	0.734	0,211
	mBART	0.378	0.627	0.718	0,121	0.232	0.627	0.718	0,245
WikiLingua	BERTurk	0.271	0.592	0.875	0,071	0.091	0.592	0.875	0,144
	mT5	0.293	0.580	0.841	0,060	0.101	0.580	0.841	0,117
	mBART	0.286	0.579	0.871	0,089	0.089	0.579	0.871	0,155
Firat_DS	BERTurk	0.407	0.645	0.889	0,120	0.282	0.645	0.889	0,275
	mT5	0.412	0.638	0.860	0,205	0.247	0.638	0.860	0,300
	mBART	0.387	0.645	0.908	0,299	0.265	0.645	0.908	0,387

Table 8. Scores of Evaluation Metrics After Normalization

Datasets	Models	MDS	1 GRAM					2 GRAM				
			R	B	F	N	MDS1	R	B	F	N	MDS2
TR-News	BERTurk	0,600	1,000	1,000	0,000	0,243	0,561	1,000	1,000	0,000	0,556	0,639
	mT5	0,472	0,763	0,737	0,208	0,117	0,456	0,756	0,737	0,208	0,252	0,488
	mBART	0,259	0,329	0,073	0,218	0,255	0,219	0,452	0,073	0,218	0,456	0,300
MLSUM	BERTurk	0,418	0,635	0,489	0,134	0,264	0,380	0,576	0,489	0,134	0,622	0,455
	mT5	0,377	0,553	0,599	0,139	0,146	0,359	0,491	0,599	0,139	0,348	0,394
	mBART	0,318	0,489	0,350	0,059	0,255	0,288	0,505	0,350	0,059	0,474	0,347
WikiLingua	BERTurk	0,252	0,000	0,095	0,837	0,046	0,244	0,007	0,095	0,837	0,100	0,260
	mT5	0,187	0,100	0,007	0,668	0,000	0,194	0,042	0,007	0,668	0,000	0,180
	mBART	0,246	0,068	0,000	0,817	0,121	0,252	0,000	0,000	0,817	0,141	0,239
Firat_DS	BERTurk	0,614	0,621	0,482	0,906	0,251	0,565	0,682	0,482	0,906	0,585	0,664
	mT5	0,609	0,644	0,431	0,762	0,607	0,611	0,558	0,431	0,762	0,678	0,607
	mBART	0,764	0,530	0,482	1,000	1,000	0,753	0,622	0,482	1,000	1,000	0,776

As presented in Table 8, MDS scores for 1-gram and 2-gram were first calculated, and then these values were averaged. When the MDS results are examined, there are significant parallels between the human evaluation results. While these parallels and similarities indicate a positive outcome for our study, it should be noted that further validation, including human annotations and correlation measurements, is required to conclusively determine the reliability of MDS as an evaluation metric for text summarization. This alignment between MDS scores and human evaluation suggests that MDS may serve as an effective complementary metric for assessing model performance. For example, for the TR-News Dataset, the BERTurk model showed the highest performance according to both MDS and human evaluation results. MDS scores reflect that BERTurk is strong in superficial similarity and semantic accuracy, confirming the superiority of the model. Both mT5 and mBART models showed poor performance in terms of MDS and human evaluation. MDS evaluation reflected the performance difference between the models well. In terms of MLSUM Dataset, mBART was the best in both MDS and human evaluation. MDS scores provided a fair estimate of the model's ability to gather information and form ideas, as well as the overall human evaluation. The existing BERTurk and mT5 models again ranked in the average range in terms of MDS and human evaluation results. In summary, the results showed that it is reasonable to claim that MDS scores adequately represent the summary performance. In the WikiLingua Dataset, the mBART model had the highest MDS scores and also performed relatively better in human evaluation results. MDS supported the relative superiority of this model. Both mBART and mT5 showed low performance in both MDS and human evaluation. MDS was also observed to be useful in identifying poor performance. In the Firat_DS Dataset, mBART showed the best performance in both MDS and human evaluation results. This means that MDS is not only able to capture semantic information with high accuracy but also to identify information that is opposite to that high accuracy. BERTurk and mT5 performed well against both evaluation methods but were outperformed by mBART. This showed that MDS accurately depicted the distinction between the performances of the models.

6. Conclusions

In this study, we explored the potential of models BERTurk, mT5, and mBART in performing the task of abstractive text summarization in Turkish using TR-NEWS, MLSUM, WikiLingua, and Firat_DS as datasets. The summaries and scores produced were evaluated comprehensively, and the scores we obtained from ROUGE, BERTScore, Novelty Rate, and FastText Based Cosine Similarity metrics were normalized and compared with the manual evaluation results by creating a new score that we called MDS. This comprehensive approach allowed us to gain a multifaceted understanding of summarization quality.

In ATS studies, it is essential to evaluate Rouge values and Novelty Rates, especially to reveal the effectiveness of summarization models. In addition, instead of a summarization model that summarizes the text by copying the sentences in the original text one-to-one, a model that produces the summary with new sentences is preferred. According to the ROUGE values obtained, BERTurk gave more accurate results in summarizing the automatic abstraction text in Turkish compared to other models. mT5 produced summaries with higher Novelty Rates compared to BERTurk and mBART. Except for the MLSUM dataset, BERTurk obtained the highest BERTScore values. mBART and mT5 also have BERTScore performance values close to the BERTurk model. Considering the comprehensive evaluations, it was concluded that although the summarization performances of BERTurk, mBART, and mT5 are close, each model has shortcomings and strengths.

Assessing a model's summaries in conjunction with MDS proved helpful in understanding the subtleties of a model's performance, emphasizing the positives and negatives in detail. Through this extensive evaluation, we found that while the summarized performances of the BERTurk, mBART, and mT5 models were quite close, each had distinct benefits and drawbacks. This approach to evaluation proved that different metrics are indeed helpful and necessary for a thorough assessment of various models for the task of human evaluations summarization in Turkish. In addition, the correlation between the results of human evaluations and the results of MDS was a good reason to claim the efficiency of MDS in case of replacing human evaluation.

In future work, we plan to develop innovative approaches for Turkish abstractive text summarization. Performing the data summarization task in layers can enable more effective capture of different forms of information and data. In this process, the summaries produced by the model can be divided into specific layers of information and provide more prosperous and multidimensional content. Furthermore, developing personalized summarization models in the future can be an up-and-coming area. In particular such models will be especially useful in the age of Information Overload, where users will be able to indicate the total number of words they prefer to read from the summaries and the modalities. At this juncture, it is possible to construct adaptive summary models that correspond with individual user preferences and reading practices leveraging on learning devices. In particular, the multidimensional evaluation approach we propose in this study can provide a solid foundation for personalized and multidimensional summarization models in the future and more consistent results by evaluating the performance of the models from different perspectives.

References

[1] M. Zhang, G. Zhou, W. Yu, N. Huang, & W. Liu. (2022). A comprehensive survey of abstractive text summarization based on deep learning. Computational intelligence and neuroscience, 2022(1), 7132226.] [Akhmetov, I.,

Nurlybayeva, S., Ualiyeva, I., Pak, A., & Gelbukh, A. (2023). A Comprehensive Review on ATS. Computación y Sistemas, 27(4), 1203-1240.
[2] I. Mani, & M. T. Maybury (Eds.). (1999). Advances in ATS. MIT Press.
[3] D. Jain, M. D. Borah, & A. Biswas (2021). Summarization of legal documents: Where are we now and the way forward. Computer Science Review, 40, 100388.
[4] D. Suleiman and A. Awajan (2020). Deep learning based abstractive text summarization: Approaches, datasets, evaluation measures, and challenges. Mathematical Problems in Engineering, 2020, 1-29. https://doi.org/10.1155/2020/9365340.
[5] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. Trippe, J. Gutiérrez … & K. Kochut, (2017). Text summarization techniques: A brief survey. https://doi.org/10.48550/arxiv.1707.02268.
[6] S. Gehrmann, Z. Ziegler, & G. Rushton, (2019). Generating abstractive summaries with fine-tuned language models. https://doi.org/10.18653/v1/w19-8665.
[7] A. See, P. J. Liu & C. D. Manning (2017). Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.
[8] D. Bahdanau, K. Cho &Y. Bengio (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit,, L. Jones, A. N. Gomez, ... & I. Polosukhin. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[10] J. Devlin, M. W. Chang, K. Lee, & K. Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, ... & D. Amodei. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
[12] W. S. El-Kassas, C. R. Salama, A. A. Rafea, & H. K. Mohamed (2021). ATS: A comprehensive survey. Expert systems with applications, 165, 113679.
[13] B. Baykara & T. Güngör (2023). Turkish abstractive text summarization using pre-trained sequence-to-sequence models. Natural Language Engineering, 29(5), 1275-1304.
[14] M. Ülker, & A.B. Özer (2021, June). TTSD: A novel dataset for Turkish Text Summarization. In 2021 9th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-6). IEEE.
[15] F. B. Fikri, K. Oflazer, & B. Yanikoglu (2021, August). Semantic similarity based evaluation for abstractive news summarization. In *Proceedings of the 1st workshop on natural language generation, evaluation, and metrics (GEM 2021)* (pp. 24-33).
[16] B. Baykara & T. Güngör (2022). Abstractive text summarization and new large-scale datasets for agglutinative languages Turkish and Hungarian. Language Resources and Evaluation, 56(3), 973-1007.
[17] A. Safaya, E. Kurtuluş, A. Göktoğan,, & D. Yuret (2022). Mukayese: Turkish NLP strikes back. arXiv preprint arXiv:2203.01215.
[18] R. Bech, F. Sahin, & M. F. Amasyali (2022, September). Improving Abstractive Summarization for the Turkish Language. In 2022 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-6). IEEE.
[19] B. Ay, F. Ertam, G. Fidan, & G. Aydin (2023). Turkish abstractive text document summarization using text-to-text transfer transformer. Alexandria Engineering Journal, 68, 1-13.
[20] B. Baykara & T. Güngör (2023, June). Morphosyntactic Evaluation for Text Summarization in Morphologically Rich Languages: A Case Study for Turkish. In *International Conference on Applications of Natural Language to Information Systems* (pp. 201-214). Cham: Springer Nature Switzerland.
[21] Y. Yüksel., & Y. Çebi (2021, October). TR-SUM: An ATS Tool for Turkish. In the International Conference on Artificial Intelligence and Applied Mathematics in Engineering (pp. 271-284). Cham: Springer International Publishing.
[22] S. Hochreiter, & J. Schmidhuber. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.
[23] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.
[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, ... & P. J. Liu, (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1-67.
[25] S. Rothe, S. Narayan, & A. Severyn (2020). Leveraging pre-trained checkpoints for sequence generation tasks. Transactions of the Association for Computational Linguistics, 8, 264-280.
[26] S. Schweter, (2020). Berturk-bert models for Turkish, April 2020. URL https://doi. org/10.5281/zenodo, 3770924.
[27] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, ... & C. Raffel, (2020). mT5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934.
[28] Y. Liu. (2020). Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.
[29] T. Scialom, P. A. Dray, S. Lamprier, B. Piwowarski, & J. Staiano, (2020). MLSUM: The multilingual summarization corpus. arXiv preprint arXiv:2004.14900.
[30] F. Ladhak, E. Durmus, C. Cardie, & K. McKeown, (2020). WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. arXiv preprint arXiv:2010.03093.

[31] C. Y. Lin (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).
[32] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, & Y. Artzi, (2019). Bertscore: Evaluating text generation with Bert. arXiv preprint arXiv:1904.09675.
[33] P. Bojanowski, E. Grave, A. Joulin, & T. Mikolov, (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, 135-146.

Article Information Form

Authors Contributions: All authors contributed equally to the writing of this paper. All authors read and approved the final manuscript.

Conflict of Interest Notice
Authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Approval
It is declared that during the preparation process of this study, scientific and ethical principles were followed, and all the studies benefitted from are stated in the bibliography.

Plagiarism Statement: This article has been scanned by iThenticate ™.