Journal of Information Systems and Telecommunication
Vol.13, No.4, October-December 2025, 278-288
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Automatic Concept Extraction from Persian News Text Based On Deep Learning

ZahraSadat Hosseini[1], Sayed Gholam Hassan Tabatabaei[1*]

[1].Department of Electrical and Computer Engineering, Malek-Ashtar University of Technology, Tehran, Iran

Received: 18 Dec 2024/ Revised: 04 Oct 2025/ Accepted: 02 Nov 2025

Abstract

One of the most critical issues in natural-language understanding is extracting concepts from the text. The concept expresses essential information from the text. Concept Extraction to the process of extracting and generating keyphrases that may exist or not in the text. Automatic concept extraction from the Persian news text is a challenging problem due to the complexity of the Persian language. In this paper, we first review traditional and deep learning-based models in keyphrase extraction and generation. Then, an automated Persian news concept extraction algorithm is presented, which exploits encoder-decoder models. Specifically, our proposed models use the output vector of BERT-Base and ParsBERT language models as a word embedding. The evaluation results have shown that changing the word embedding layer has improved recall, precision, and F1 measures about 3.15%. Since encoder-decoder models get inputs consecutively, the training time increases. Also, if the sentence is long, they cannot store much information from the sentences. Therefore, for the first time, we have used mT5-Base with Transformer architecture, which receives and processes data parallelly. Recall, precision, and F1 measures used for the concept extraction results of the mT5-Base model are 55.66%, 55.47%, and 55.48%, respectively. The F1 score has increased by 19.8% compared to the previous models. Therefore, this model is effective for extracting the concept of Persian news texts.

Keywords: Concept Extraction; Deep Learning; Keyphrase; BERT-BASE; ParsBERT; mT5.

1- Introduction

Many texts on different topics are published on social media every day. With the increasing volume of documents and texts, fast and reliable methods are needed to extract useful information from this vast amount of unstructured data. Concept extraction is a tool for generating and extracting keyphrases from an unstructured text that provides summary information about the text. Digital information management uses concepts for document clustering, information retrieval [1], and text summarization [2]. The concept consists of Keyphrases that may be directly present or not in the text [3]. Keyphrases can be single-word or multi-words expressions that summarize the main semantic meaning of unstructured text data and are divided into two categories [4]: absent and present Keyphrases. Unlike present Keyphrases, absent Keyphrases do not exist in the text and are implicitly mentioned in the text. In order to identify the present keyphrase in the text, keyphrase extraction algorithms are used. On the other hand, the keyphrase generation process performs the task of extracting explicit and implicit keyphrases from the text. The concept extraction is the task to extract and generate keyphrases at the same time [5].

The Internet provides people's information and contains a large amount of textual data. Therefore, it is difficult, expensive, and time-consuming for humans to extract concepts from huge documents. Hence, automatic concept extraction systems are needed [6]. So far, various automated systems have been designed for generating and extracting Keyphrases, but Persian concept extraction is still a challenge. This is for some reasons: First, most of the algorithms have been presented for the English language and little research has been done on the Persian language. Second, the structural complexity of the Persian language is higher than many languages such as English, and the other important reason is the existence of ambiguities in natural languages such as ambiguity in reference, lexical ambiguity due to polysemy, and ambiguity in distinguishing subject and object due to

✉ **Seyed Gholam Hassan Tabatabaei**
tabatabaei@mut.ac.ir

omission from the sentence. Therefore, other methods are needed to extract the concepts of Persian texts.

Given the mentioned challenges and the importance of extracting concept from Persian texts, this paper focuses on proposing an automated method for extracting concepts from Persian news texts. The proposed method is based on deep learning models, leveraging their ability to process large volumes of text data and capture complex patterns and semantic relationships. These models enable a more nuanced understanding of language, making them well-suited for tasks such as concept extraction and keyphrase generation, particularly in the context of Persian language processing.

2- Related Work

Concept extraction refers to the extraction and generation of Keyphrases. It is different from the text summarization process. Therefore, we review this literature in two parts: 1) Extracting and generating of keyphrases methods, and 2) summarization methods.

2-1- Extracting and Generating of Keyphrases

In early works, the text keyword extraction methods often included three steps: First, some keywords are extracted as text concept candidates. Second, the extracted concepts are refined using prior knowledge. As a result, the probability of reaching higher-level concepts increases, and finally, keywords are scored based on statistical information or prior knowledge [7]. Some automated keyword extraction systems are based on supervised approaches that attempt to map the sample space into two classes "key semantic units" and "non-key semantic units". Witten et al. [8] proposed a simple keyword extraction algorithm (KEA) that selects candidate keywords by calculating the TF-IDF (Term Frequency–Inverse Document Frequency) [9] and obtains the final keywords by the Naive Bayes algorithm. Zhang et al. [10] extracted Keyword from Chinese documents using a conditional random fields algorithm. The conditional random field model performed better than other machine learning methods such as linear regression and support vector machine model. Barla et al. [11] extracted key concepts instead of words for document classification using naïve Bayes model and obtained better results on news documents compared to TF-IDF keyword model.

Unlike supervised methods, unsupervised methods use unlabeled data to extract keywords. These can be divided into the graph-based, statistics-based, and language model-based methods. Khozani et al. [12] presented a statistical-based algorithm to extract keywords. In the first step, they removed the redundant words and weighed the remaining words with the TF-IDF criterion. Then using the n-gram method and based on the words' position, the weight of the words was updated and the key sentences were determined. Finally, keywords were extracted from the selected sentences. The experimental results have shown that the inference time and accuracy of this method for extracting keywords are high. Unsupervised statistical techniques such as KP-MINER [13], RAKE [14], and YAKE [15] use statistical features of texts to extract keywords. These methods are more complex due to a large number of operations. TextRank [16], SingleRank [17], and their extensions TopicRank [18] and ExpandRank [17] are graph-based methods that construct graphs to rank words based on their location in the graph. These techniques perform poorly in identifying cohesiveness between different words that constitute a keyword. Language model-based techniques utilize language model-derived statistics to extract keywords from the text [19-20]. Doostmohammadi et al. [21] conducted a comprehensive assessment to compare the performance of supervised and unsupervised methods in news keyphrase extraction and generation. Their research showed that 1) contrary to expectations, KP-Miner is better than the supervised method, 2) unsupervised approaches based on statistics are also better than graph-based methods, 3) The use of machine translation evaluation such as BLEU and ROUGE provides a more realistic evaluation for the task of keyphrase extraction and generation, 4) And all the keyphrase are not explicitly mentioned in the text. Therefore, generative models are needed to extract the non-expressed or absent keyphrases.

Recent research underscores the importance of discourse-level analysis for concept extraction, emphasizing the need to account for relations spanning sentences. For example, dependency graphs and discourse relations effectively capture linguistic structures. Techniques such as Clause Matching, as highlighted by I-Hung Hsu et al. [22], leverage dependency arc types to extract cohesive concepts from multi-sentence texts. This perspective aligns with the growing use of deep learning models, which excel in modeling complex semantic relationships and discourse structures [22].

Ontology-based concept extraction builds on these methods by integrating domain-specific knowledge to refine candidate keyphrases and associate terms with hierarchical structures. Gayathri and Kannan [23] developed a system for Ayurvedic texts that leverages domain ontologies, semantic weighting with TF-IDF, and k-Nearest Neighbors (kNN) classifiers for document classification, achieving superior results compared to traditional methods [24].

Deep learning-based methods have outperformed other machine learning methods in numerous natural language processing tasks, especially in keyphrases generation. The idea behind these methods is to learn complex features directly from data. Yuan et al. [25] adopted the RNN-based seq2seq architecture with a copy mechanism for keyphrase generation. This architecture predicts a group of keyphrases with variable length, which is considered its

advantage. Swaminathan et al. [26] proposed a CGAN generative architecture to generate keyphrases from research articles. Sun et al. [27] designed the DivGraphPointer architecture by combining traditional graph-based ranking methods and neural network-based approaches to generate keyphrases. The CopyRNN architecture was presented by Meng et al. [28] for keyphrase generation, consisting of an encoder for learning the representation of the text and a decoder for generating keyphrases based on that representation. Various modifications of the CopyRNN architecture have been proposed recently. Zhang et al. [29] proposed another architecture called CopvRNN to manage the repetition of keywords during generation based on the CopyRNN architecture. This architecture uses a bidirectional GRU for encoding and a forward GRU for decoding. CopyRNN-based architectures consistently predict N keyphrases for any input text, while in real-world examples, the number of keyphrases may vary among different texts and should be determined based on the document's content. Chen et al. [30] improved the performance of the generative model using an integrated model. The integrated model distinguishes the semantic features of present keyphrases from absent keyphrases. However, this model is not trained end-to-end and only uses a bottom shared encoder to implicitly capture the hidden semantic relationship between absent keyphrase generation and present keyphrase extraction. The first research in the field of extracting and generating keyphrases from Persian news articles was done by Doostmohammadi et al. [31]. They showed that sequence-to-sequence deep models not only perform well in keyphrase generation, but also significantly outperformed common methods such as Topic Rank, KPMiner, and KEA in keyphrase extraction.

Glazkova and Morozov [32] investigated fine-tuned generative models for keyphrase selection in Russian scientific texts, such as mT5, and mBART. Their experiments revealed that mBART achieved the best performance in in-domain evaluations, surpassing baseline methods across multiple domains such as mathematics, history, medicine, and linguistics. This study highlighted the efficacy of generative models for multilingual keyphrase extraction tasks, particularly in scientific domains. Recently, large language models (LLMs), such as GPT-4, have demonstrated impressive performance across various tasks without requiring fine-tuning. Glazkova et al. [5] also explored the use of LLMs for keyphrase generation, specifically for Russian scientific texts. Their result shows, mBART consistently outperforms LLMs and other baselines in in-domain evaluations, achieving up to higher F1 scores in fields like Mathematics and Medicine. Overall, LLMs can perform well on a variety of tasks without needing additional fine-tuning for each specific task. However, the performance of LLMs is highly reliant on the quality and design of the prompts. A poorly designed prompt may lead to inaccurate or irrelevant results, limiting the model's reliability and consistency [33].

Thomas and Vajjala [34] introduced an approach to separate present keyphrase extraction and absent keyphrase generation into distinct tasks, focusing on increasing diversity in absent keyphrase generation through specialized attention mechanisms. Their findings demonstrated improved performance across six English datasets, particularly for absent keyphrase generation tasks, emphasizing the role of distinct modeling strategies for present and absent keyphrases.

A recent study by Song et al. [35] explores the use of prompt-based unsupervised keyphrase extraction by leveraging large pre-trained language models like T5. The authors demonstrate that designing effective prompts significantly impacts performance, with complex prompts performing better for long documents. However, simple prompts often suffice for shorter texts. Their experiments on six benchmark datasets, including Inspec, SemEval2010, DUC2001, SemEval2017, Nus, and Krapivin, which are primarily English datasets, reveal that well-crafted prompts can significantly enhance keyphrase extraction performance, and automating prompt generation could further improve efficiency and scalability in real-world applications [35]. Similarly, Shen and Le [36] proposed the TAtrans model, which leverages title attention and sequence order embeddings to enhance keyphrase generation. The model showed superior performance across several datasets, including Chinese abstracts, showcasing the potential of Transformer-based methods for keyphrase tasks in diverse languages [36].

Most of the work done on keyphrase extraction and generation has focused on non-Persian texts. Therefore, in this paper, we focus on concept extraction (keyphrase generation and extraction) from Persian news texts. Given the success of language models such as T5 in various languages, we have, for the first time, employed the mT5 language model to extract and generate keyphrases from Persian news texts.

2-2- Summarization

In the present era, alongside progress in scientific and technological fields, there is a remarkable surge in the volume of accessible data. Consequently, is beneficial to have concise information that encapsulates the essence of the original document while occupying a reduced space. Although human-generated text summarization offers advantages such as precision, comprehensiveness, and coherence, it remains a laborious and costly undertaking [37]. Summarization is the process of compressing the source text into a brief version, which contains the key information of the source text. There are two types of summarization: abstractive and extractive [38]. Extractive methods choose

essential sentences, phrases, or paragraphs from the source text to form a summary while abstractive summarization methods use linguistic methods to generate a brief text [38-39]. The abstractive summarization might contain words that are not explicitly present in the source text [39]. Most of the research has been done on extractive summarization. Recently, researchers have turned to abstractive summarization. Abstractive summarization is a complex and challenging task due to the complexities of natural language text [40]. Abstractive summarization methods are broadly divided into three categories: 1) structure-based approaches, 2) semantic-based approaches, 3) deep learning-based approaches. The structure-based approach filters the most important information from the text using abstract or cognitive algorithms and includes template-based methods, tree-based methods, and ontology-based methods. Semantic-based approaches take text as input and construct a semantic representation of it. Information item-based methods, semantic graph-based methods, and multimodal semantic models use the semantic-based approach [41].

Some abstractive summarization algorithms give more scores to the summaries with more words in common with the source text and pay less attention to the semantic similarity between generated sentences and the source text. Therefore, Salemi et al. [42-43] presented a deep learning-based architecture to extract text summaries. This architecture is a pre-trained encoder-decoder model that has shown good performance in summarizing Persian text. Similarly, Shanthakumari et al. [44] used the PEGASUS model for abstractive summarization, which generates summaries by capturing key information from the original text, offering improved coherence and relevance. Their experiments demonstrated that PEGASUS, a transformer-based model, excels at generating human-like summaries by maintaining semantic integrity and reducing redundancy, addressing some of the common limitations of previous methods. Furthermore, research [45] into clinical text summarization highlights PEGASUS's capacity to distill large textual datasets into concise, coherent summaries, demonstrating comparable advantages in the medical domain where context, precision, and relevance are crucial. Liu et al. [46] proposed a hybrid summarization approach combining fine-tuned mT5 and large language models like ChatGPT, specifically evaluated on the LCSTS dataset—a large-scale Chinese short-text summarization corpus. Their approach involved using mT5 to generate initial summaries, which were refined by ChatGPT to enhance fluency and coherence, achieving high ROUGE scores and addressing key limitations in traditional models. Notably, T5's ability to treat all NLP tasks as a text-to-text problem allows it to achieve superior performance in both semantic accuracy and context preservation. This is due to its encoder-decoder transformer architecture, where the encoder captures context from the input text and the decoder generates corresponding outputs, making it particularly effective for tasks like summarization. Additionally, T5's self-attention mechanism, a core feature of the transformer architecture, enables it to focus on the most relevant parts of the input text, improving its ability to generate coherent and contextually accurate summaries. These strengths were leveraged by Liu et al. [46], where mT5 played a critical role in generating initial summaries before refinement by ChatGPT.

Encoder-decoder-based models, including the T5 model, have demonstrated good performance in both summarization and key phrase extraction tasks. Since encoder-decoder models are specifically designed and fine-tuned for tasks such as keyphrase extraction or generation, the aim of this paper is to propose a model based on the encoder-decoder architecture for extracting and generating keyphrases from Persian text, specifically news texts. The main contributions of our paper are: 1) we modify the base Encoder-Decoder [28] to extract the Persian text concept. 2) Then we change its word embedding layer by using the BERT-base [47] and ParsBERT [48] language model and present a modification of it 3) And finally, for the first time, we use the pre-trained Multilingual T5 (mT5-Base) model [3] to Persian text concept extraction. Our proposed models obtained significant results in extracting the concept of Persian news text.

The rest of this paper is organized as follows: Section 3 describes the proposed method, then the experimental setup is described in Section 4, and the experimental results are given in Section 5. Finally, it is concluded in section 6.

3- Proposed Method

The proposed method consists of two phases: pre-processing and the extension of deep learning-based language models for concept extraction. Pre-processing converts the data into a suitable format and making the process of calculations and extraction of information faster and simpler. The output of the pre-processing step is fed into the input of deep learning-based architectures. The details of each step are as follows:

3-1- Pre-processing

We use the Perkey dataset to evaluate our proposed model. This dataset has been preprocessed, as described in [21], and is publicly available. The preprocessing includes removing sentences containing specific keywords from Persian web pages and JavaScript code. Since some texts use different encodings and languages, it is necessary to unify the text to improve its analysis. For example, two Arabic letters, "ک" and "ی" are converted into their Persian equivalents. In addition to the preprocessing performed in [21], we applied further preprocessing to normalize the data using the Hazm library. The normalization process,

carried out with Hazm, consists of seven steps, which are detailed as follows:

- Analyzing the "ء ", which is a non-vowel letter and its different spellings, and correcting them.
- Removing the mentioned letter in the first step from the end of a word (such as modifying 'شهداء' to 'شهدا').
- Removing letters consisting of ' ', ' ', ' ', and ' ', from the words.
- Converting Arabic and English numbers into Persian equivalents.
- Correcting written half-spaces.
- Removing extra spaces and half-spaces used in the text.
- Correcting two-part words which are incorrect.

After normalization, the Tokenizer, a tool of BERT-Base [37], is used to split words into tokens. Tokenization recognizes the boundaries between words in texts and assigns a specific identifier to each semantic unit. As a result, a dictionary is created to convert the input text into a sequence of numbers and identifiers. Deep learning-based models are fed with the same dimensions. Since, in this paper, the BERT-BASE language model is used for word embedding, the maximum length for each token is considered 512.

3-2- The Proposed Concept Extraction Model

In this paper, the models used to extract the concept of news texts are pre-trained mT5-Base [3] and a modified encoder-decoder model. The encoder-decoder framework tries to extract the present keyphrases from the text and predict the absent keyphrases. On the other hand, with the transfer learning technique, the knowledge obtained from pre-trained models can be generalized to solve other tasks. In the following, the structure of the base models will be explained first. Then the proposed architectures are described:

3-2-1- Encoder-Decoder Structure

The encoder-decoder model was first introduced in 2014 by Chao. et al. [49] to solve translation problems. The encoder-decoder architecture [28] consists of two streams: an encoder path containing RNN blocks to learn hierarchical features from the input text. If $x = (x_1, x_2, \dots, x_T)$ is the input sequence, the hidden representation vector $h = (h_1, h_2, \dots, h_T)$ is obtained by applying the non-linear function 'f' on the x at the time step t and the previous hidden state. Then by applying the non-linear function q on the hidden representation vector, the concept vector c is obtained according to Eq. (2):

$$h_t = f(x_t, h_{t-1}) \quad (1)$$

$$c = q(h_1, h_2, \dots, h_T) \quad (2)$$

The second stream also includes RNN blocks, and its purpose is to convert the concept vector into keyphrases. Hence, this path is called the decoder. In each time step, the non-linear function f takes the concept vector, the output of the previous hidden state, and the predicted word at the time step t-1 as input and produces the hidden state s_t.

$$s_t = f(y_t, h_{t-1}, c) \quad (3)$$

Then, using the conditional language model, the predicted word y_t is obtained [21, 28].

In general, the encoder-decoder model has worked well in solving natural language processing problems, especially the generation of keyphrases, but it also has several drawbacks: 1) It is difficult to train the model for long sentences because the information containing the relationship between words is lost as the sentence length increases. Therefore, the model's accuracy in generating the main keyphrases decreases. 2) The vocabulary words of RNN models consist of a limited number of words (e.g. 30,000 words in [49]). Therefore, some keyphrases may not be included in these vocabulary words. 3) Most language models use common methods such as Word2Vec, Elmo, etc. for embedding words, but these methods do not accurately capture the relationships between words. On the other hand, the word embedding layer is the most critical part of the concept generation algorithm because its output is used as an input for the encoder-decoder model. Therefore, the design of a strong and appropriate word embedding layer is needed.

3-2-2- The Proposed Encoder-decoder Models

Inspired by the work of Meng et al. [28], we developed an encoder-decoder model utilizing bidirectional LSTM blocks for concept extraction. The structure of the proposed model, shown in Fig. 1, incorporates BERT-BASE or ParsBERT as the embedding layer, along with attention and copy mechanisms, to enhance performance. This architecture consists of two primary components: a contextual embedding layer and a modified encoder-decoder framework. The encoder processes the input text, leveraging the contextualized embeddings provided by BERT-BASE or ParsBERT, while the decoder generates output sequences. The details of each component are:

Embedding layer: According to Fig. 1, the encoder uses BERT-BASE or ParsBERT to generate contextual embeddings for the input text. In fact, we generate word embedding for textual data using the word embedding layer of ParsBERT [48] and BERT-BASE [47] models and propose BERT-BASE+Encoder-Decoder and ParsBERT+Encoder-Decoder models for the concept extraction. To use the BERT-BASE word embedding layer, the process of fine-tuning the model should be done. In this process, a list of 512 symbols is entered into the network and a 768-dimensional vector is generated. This vector is used as the input of the encoder-decoder model. Embedding from BERT-BASE would be different for the different occurrences of a word as it generates embedding's based on the context of the sentence. Other advantages of the BERT-BASE are: First, unlike other Encoders, the

BERT-BASE Encoder receives the entire sequence of words simultaneously. As a result, this is considered a bidirectional model that can learn the relationship of a semantic unit with all surrounding units [47]. Second, since BERT-BASE receives all the words of a text at once, the Masked Language Model (MLM) technique is used to train the model. In this technique, some words are randomly masked during training to increase the model's ability to learn the concept of the input sentence.

The BERT-BASE model is considered a multilingual model because it has been trained on 104 different languages. The extension of the BERT-BASE model for the Persian language under the name Pars Bert [48] was presented by Farahani et al. This model has been trained on Persian documents from various topics (such as science, novels, and news). Our experiments have shown that changing the word embedding layer using language models, especially BERT-Base, has led to improved evaluation criteria (see Section 5).

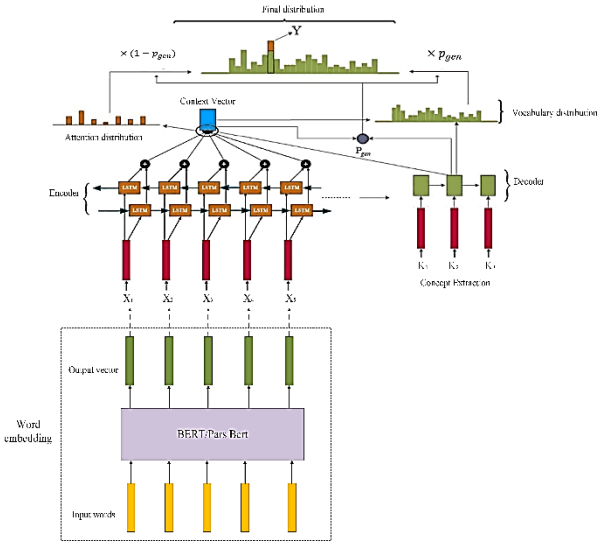

Fig. 1 The general scheme of the proposed Encoder-Decoder models with BERT-BASE /ParsBERT word embedding layer

Modified encoder-decoder: According to the explanation in Section 3.2.1 and similar to [28], the proposed architecture consists of an encoder-decoder for concept extraction, but we use bidirectional LSTMs instead of RNNs in the encoder. This is because bidirectional LSTMs, with their gating mechanisms, are better at capturing long-term dependencies and mitigating the vanishing gradient problem. These properties allow the model to effectively learn contextual relationships in both forward and backward directions, leading to more accurate concept extraction, especially in cases of complex or lengthy input texts.

Also, similar to [28], the decoder generates output by leveraging attention and copy mechanisms, addressing key challenges in sequence generation. Different words of a sentence have different importance for generating each output at each time step [50-52]. Therefore, the attention mechanism receives the output of the encoder's LSTM blocks and assigns a different weight to each of them to generate the final output. On the other hand, the copy mechanism copies certain parts of the source text exactly in the output. In this way, important key phrases that may not be present in the LSTM vocabulary are considered for concept generation.

By blending generative and extractive strategies, the final output is determined by a soft-switch parameter, p_{gen} , which dynamically adjusts the balance between generating tokens from the vocabulary and copying tokens from the input text [28].

3-3- MT5-Base Structure

The pre-trained mT5-Base model is an extension of the T5 (Text-To-Text Transfer Transformer) model which is considered an advanced version of BERT-based models. The T5 model is built using the transformers architecture, so its input and output can be text sequences. The transformer is a sequence-to-sequence model that consists of several blocks, which are connected as shown in Fig. 2: 1) an encoder block combined of a multi-head self-attention module, a position feed-forward network (FFN), residual connections to prevent gradient vanishing problem, and batch normalization layers, 2) and a decoder block, which has additional cross-attention modules between multi-head self-attention modules and position-based FFNs. The attention mechanism, as a core block of the transformer, is well-suited for long-range dependencies modeling, which is achieved by the adaptive weighting of the features according to the importance of the input. The main feature of this model is the use of relative positional embedding instead of sinusoidal positional embedding [53]. Relative positional embedding is a method for explicit and effective encoding of positional information, representing the relative position of a word in an input sentence as a vector or scalar.

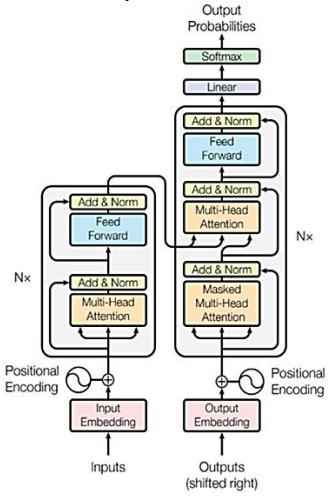

Fig. 2 The model architecture of The Transformer that was used in the mT5-Base model [47].

The use of Transformer blocks has enhanced the ability of the T5 model to perform multiple NLP tasks, including summarization, machine translation, question-answering, and classification [54]. The T5 model has been trained on a considerable amount of English texts, it cannot be generalized to other languages. The mT5-Base model is presented to solve this problem, which supports 101 different languages [55]. MT5-Base is capable of zero-shot learning and can be used in NLP tasks, including concept extraction. In this paper, we use the transfer learning technique and the pre-trained mT5-Base model to extract concepts from Persian news texts. For this purpose, we first convert the news data into text format. Then we load the mT5-Base model and use the simpleT5 class built on PyTorch-lightning and Transformers to train our model.

4- Experimental Setup

4-1- Dataset

All of the methods are validated on a subset of the Perkey dataset which includes 395,645 Persian news articles collected from 6 websites and news agencies. Each news article has at least 3 keyphrases and provides comprehensive information: {title, keyphrases, body, summary, category, URL} [21]. This dataset is divided into three subsets: training (345645 news articles), validation (2500 news articles), and test (2500 news articles) portions. The analysis carried out in [21] has shown that 31.44% of all keyphrases are not present in the text of news articles. Additionally, the number of keyphrases in news texts varies from 2 to more than 9. All this shows that the Perkey dataset can provide diverse examples with enough information to train deep learning models.

4-2- Training

The proposed method is implemented in PyTorch and evaluated on a computing server with a 3090 GPU. In the training process, the Negative Log Likelihood Loss function and Adam optimizer (initial learning rate = 10^{-4}, gradient clipping=0.1) are used.

4-3- Evaluation criteria

Various criteria were used to evaluate the performance of the proposed method in text concept extraction. First, three common criteria, namely Precision, Recall, and F1 score were used. These criteria are defined in formulas Eq. (4) to Eq. (6), respectively.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1 = 2 * \frac{precision*recall}{precision+recall} \quad (6)$$

Where 'TP' is the number of true keyphrases, 'FP' is the number of false keyphrases, 'TN' is the number of true non-keyphrases, and 'FN' is the number of false non-keyphrases. Furthermore, the results of traditional models were examined in terms of ROUGE-1 and ROUGE-2 metrics. The ROUGE-1 criterion refers to the overlap of unigrams (a subsequence of n words) between the candidate summary and the reference summary. While the ROUGE-2 criterion refers to the overlap of bigrams between the candidate summary and reference summary. According to the ROUGE definition, Precision and Recall criteria are described by formulas Eq. (7) and Eq. (8), respectively.

$$Precision = \frac{number\_of\_overlapping\_words}{total\_words\_in\_system\_summary} \quad (7)$$

$$Precision = \frac{number\_of\_overlapping\_words}{total\_words\_in\_system\_summary} \quad (8)$$

5- Experimental Results and Analysis

To confirm the performance of the proposed models in extracting the concept of Persian news texts, the test results were analyzed from different perspectives:

5-1- Keyphrase Extraction

Table 1 presents the results of keyword extraction on the Perkey dataset based on ROUGE-1 and ROUGE-2 criteria for traditional models. It is observed that the KEA model performs better than other methods in terms of precision and recall. In addition, Table 2 shows the quantitative evaluation results of all methods in extracting the keyphrases of the test set from the Perkey dataset. As can be seen, the supervised learning method performs better than the statistical models and graph-based methods due to the use of labeled data. Compared to traditional methods, deep learning-based methods have achieved better results in extracting keyphrases from Persian texts due to automatic feature extraction. Therefore, the good performance of the encoder-decoder model can be seen from an increase in the F1 score to 43.04%.

Table 1: The performance of traditional models for extracting keywords on the Perkey dataset

Method		ROUGE-1			ROUGE-2		
		Precision	Recall	F1	Precision	Recall	F1
Statistical Models	TF-IDF	36.34%	27.91%	29.83%	5.51%	4.78%	4.76%
	KP-Miner	39.89%	26.06%	29.12%	5.52%	4.37%	4.47%
	YAKE	18.99%	21.81%	18.69%	3.31%	4.35%	3.40%
Graph-based Model	Single Rank	20.40%	32.48%	23.59%	4.98%	9.96%	6.19%
Supervised Model	Kea	38.14%	29.39%	31.39%	6.46%	5.88%	5.72%

Table 2: Comparison of the different keyphrase extraction methods on the Perkey dataset

Method		Dataset	Precision	Recall	F1
Statistical Models	TF-IDF	Perkey	17.24%	20.60%	18.77%
	KP-Miner	Perkey	19.00%	19.48%	19.24%
	YAKE	Perkey	7.26%	8.20%	7.70%
Graph-based Model	Single Rank	Perkey	5.32%	6.71%	5.94%
Supervised Model	Kea	Perkey	18.37%	22.26%	20.13%
Deep learning-based Models	Encoder-Decoder model	Perkey	37.24%	62.87%	43.04%
	Proposed ParsBERT+Encoder-Decoder model	Perkey	38.70%	65.01%	44.83%
	Proposed BERT-BASE +Encoder-Decoder model	Perkey	39.41%	64.68%	45.32%
	mT5-Base	Perkey	56.79%	58.54%	59.63%

On the other hand, it has been observed that by using the 768-dimensional concept vectors obtained from ParsBERT's model as the input of the proposed Encoder-Decoder model, all the evaluation criteria were improved by about two percent. Alongside this, the performance of the BERT-BASE model is better than the former because it has been trained on a large corpus of multilingual data. A high prediction F1 of 45.32% for the proposed BERT-BASE+Encoder-Decoder model confirms this. Although the performance of the proposed encoder-decoder models is superior to other methods, the precision criterion obtained from these models is significantly lower than the recall criterion. This means that the number of extracted incorrect key phrases (FP) is more than the extracted incorrect non-key phrases (FN). We applied the pre-trained mT5-Base model to overcome this problem and achieved a significant improvement (59.63% F1-score) over the results of the previous models in the keyphrase extraction task. The mT5-Base model can generate word vectors more precisely due to the use of parallel processing and relative position embedding.

5-2- Keyphrase Generation

As mentioned earlier, some keyphrases do not appear in the input text. Hence, generating absent keyphrases is a challenging task. It should be noted that traditional methods cannot generate keyphrases. Therefore, Table 3 only provides the performances of deep learning-based models for the absent keyphrases prediction task. It can be seen from Table 3 that the proposed encoder-decoder models perform better than the base encoder-decoder architecture [28] in generating absent keyphrases. Also, the findings show that using the mT5-Base model has led to the improvement of all metrics. For example, the F1 score has increased by about 25%. This is because the mT5-Base model is a multilingual model and fine-tuning it on Persian news texts helps to improve the accuracy of keyphrase prediction results. It should be noted that deep learning models must be trained on large amounts of data. Hence, Fine-tuning the pre-trained model is very useful when a small training dataset is available.

Table 3: Comparison of the different keyphrase generation models methods in the Perkey dataset

Method		Dataset	Precision	Recall	F1
Deep learning-based Models	Encoder-Decoder model	Perkey	12.38%	34.60%	17.46%
	Proposed ParsBERT+Encoder-Decoder model	Perkey	14.84%	42.01%	21.04%
	Proposed BERT-BASE +Encoder-Decoder model	Perkey	15.40%	41.93%	21.52%
	mT5-Base	Perkey	44.58%	44.39%	46.86%

5-3- Concept Extraction

The concept of a text includes both absent and present keyphrases. Table 4 presents the results related to the overall performance of all deep learning-based methods, i.e. generating absent keyphrases and extracting present keyphrases. For the proposed BERT-BASE +Encoder-Decoder model the F1-score increased by approximately 3.15%. This shows that using BERT-BASE's language model for word embedding is effective. Also, as expected, after the proposed encoder-decoder models, the best performance belongs to the mT5-Base model. The overall performance of the proposed models, i.e. BERT+Encoder-Decoder and ParsBERT+Encoder-Decoder over the entire dataset are summarized in Tables 5 and 6. It can be seen from both tables that the proposed Encoder-decoder models have predicted fewer incorrect keyphrases compared to the base encoder-decoder. Also, the keyphrases generated by the mT5-Base model are more consistent with the true keyphrases.

Table 4: .Comparison of the different Concept Extraction methods in the Perkey dataset

Method		Dataset	Precision	Recall	F1
Deep learning-based Models	Encoder-Decoder model	Perkey	31.54%	46.75%	35.68%
	Proposed ParsBERT+Encoder-Decoder model	Perkey	33.24%	49.73%	37.99%
	Proposed BERT-Base +Encoder-Decoder model	Perkey	34.23%	50.91%	38.83%
	mT5-Base	Perkey	55.47%	55.66%	55.48%

Table 5: The output of the models - Example (1).

News text	نبض «پایتخت» در دست تنابنده است هومن حاجی عبداللهی مجری، صداپیشه و بازیگری است که در همه این عرصه‌ها فعالیت دارد، اما با حضور در سریال «پایتخت» بیشتر توانایی‌هایش در زمینه بازیگری به نمایش گذاشته شد. شیرینی‌های نقش رحمت شاسی در این سریال محبوب تلویزیون مدیون بازی هوشمندانه حاجی عبداللهی است که با توجه به استقبال مخاطبان باعث پررنگ‌ترشدن حضور این بازیگر در ادامه این سریال شد. به بهانه پخش سری پنجم این مجموعه پرطرفدار پای حرف‌های این بازیگر نشسته‌ایم.
True keyphrases	تلویزیون. سریال ایرانی. بازیگران سینما و تلویزیون ایران
Encoder-Decoder model	تلویزیون. سریال ایرانی. سینما. پایتخت. هومن حاجی عبداللهی
Proposed BERT-BASE +encoder-decoder model	بازیگران سینما و تلویزیون ایران. سازمان صدا و سیما، مجری رادیو و تلویزیون. تلویزیون. سریال ایرانی. سینمای تلویزیون. مجموعه تلویزیونی پایتخت. برنامه‌های تلویزیونی. هومن حاجی عبداللهی .
Proposed ParsBERT+encoder-decoder model	بازیگران سینما و تلویزیون ایران. مجری رادیو و تلویزیون. تلویزیون. سریال ایرانی. مجموعه تلویزیونی پایتخت. برنامه‌های تلویزیونی. هومن حاجی عبداللهی.
mT5-Base	تلویزیون. سریال ایرانی. شبکه نمایش خانگی. بازیگران سینما و تلویزیون ایران

Table 6: The output of the models - Example (2).

News text	رسمی؛ جواد نکونام با تیم ملی قرارداد یکساله بست به گزارش خبرگزاری خبرآنلاین؛ پس از ابراز علاقه کارلوس کی روش سرمربی تیم ملی به حضور نکونام در کادر فنی ایران، فدراسیون فوتبال با کاپیتان سابق تیم ملی وارد مذاکره شد. البته این مذکرات زیاد به طول نکشید و سرانجام روز گذشته قرارداد جواد نکونام با تیم ملی دیروز امضا شد. شماره 6 سابق ایران روز گذشته در فدراسیون فوتبال و در حضور مهدی تاج و محمدرضا ساکت مدیر تیم‌های ملی قراردادش را امضا کرد. قرارداد او یکساله است و در پایان هر سال تمدید خواهد شد.
True keyphrases	تیم ملی فوتبال ایران. جواد نکونام.
Encoder-Decoder model	تیم ملی فوتبال ایران. فدراسیون فوتبال. کارلوس کی روش. مهدی تاج. بازی‌های انتخابی. باشگاه پرسپولیس.
Proposed BERT-Base +encoder-decoder model	تیم ملی فوتبال ایران. جواد نکونام. فدراسیون فوتبال. کارلوس کی روش. محمدرضا نکونام. مهدی نکونام. بازی‌های انتخابی جام‌جهانی. باشگاه پرسپولیس. جام جهانی.
Proposed ParsBERT+encoder-decoder model	تیم ملی فوتبال ایران. جواد نکونام. فدراسیون فوتبال. کارلوس کی روش. محمدرضا نکونام. مهدی نکونام. باشگاه پرسپولیس. جام جهانی.
mT5-Base	تیم ملی فوتبال ایران. جواد نکونام. بازی‌های انتخابی جام‌جهانی

5-4- Title and Summary Extraction

The mT5-based models have a high potential to accurately learn the relationship between the words of a text due to their use of the transformer block. The performance of mT5-Base and mT5-Small models in extracting the title and summarizing the text is reported in Table 7. The mT5-Base model has more parameters than mT5-small and is considered a stronger model. Thus it has a better performance in summary extraction. In addition, the outputs of the mT5-Base model for two examples of news texts are shown in Tables 8 and 9, respectively. The results of Tables 9 and 10 show that the text produced by this model is conceptually very similar to the true labels (i.e. True Title and True Summarization). For example, in Table 8, although the summary text produced by mT5-Base has more words (i.e. "جواد نکونام" and "با تیم ملی") than the True summary, both sentences have the same meaning. The ROUGE score is low for two sentences that have different words but the same meaning because it only computes syntactical matches. Solving this challenge requires the definition of new criteria to consider the conceptual similarity of texts, which could be a fascinating issue for future studies.

Table 7: The output of the mT5-based models in extracting the title and summary in the Perkey dataset

Method		ROUGE-1			ROUGE-2		
		Precision	Recall	F1	Precision	Recall	F1
mT5-small [47]	Title	-	-	28.5 %	-	-	12.55%
	summary	-	-	33.88%	-	-	19.17%
mT5-Base	Title	40.69%	41.76%	41.99%	23.63%	24.32%	24.30%
	summary	38.16%	39.54%	39.75%	23.34%	24.07%	24.12%

Table 8: The output of the mT5-based models in extracting the title and summary from example 1

True Title	نبض «پایتخت» در دست تنابنده است.
True Summary	هومن حاجی‌عبداللهی که حضورش در فصل‌های جدید «پایتخت» پررنگ‌تر از قبل شده است، می‌گوید این اتفاق به دلیل استقبال مخاطبان از شخصیت رحمت شاسی افتاده است.
mT5-Base (Title)	سریال «پایتخت».
mT5-Base (summary)	هومن حاجی عبداللهی که حضورش در سریال «پایتخت» پررنگ تر از قبل شده است، می‌گوید به دلیل استقبال مخاطبان از شخصیت رحمت شاسی این اتفاق افتاده است.

Table 9: The output of the mT5-based models in extracting the title and summary from example 2

True Title	رسمی؛ جواد نکونام با تیم ملی قرارداد یکساله بست.
True Summary	کاپیتان سابق تیم ملی ایران سرانجام قراردادش را امضا کرد.
mT5-Base (Title)	جواد نکونام با تیم ملی قرارداد بست.
mT5-Base (summary)	جواد نکونام کاپیتان سابق تیم ملی سرانجام با تیم ملی قرارداد بست.

6- Conclusions

In this paper, deep-learning models are used for concept extraction from Persian news texts. First, a multi-stage pre-processing technique is applied to modify the Persian text and normalize the Persian text. Then, BERT-Base +Encoder-Decoder and ParsBERT+Encoder-Decoder models are proposed to extract the concept from news text. The proposed models utilize the output vector of BERT-BASE and ParsBERT language models for word embedding. The experimental results showed that the performance of the proposed models are significantly better than previous models. One of the disadvantages of encoder-decoder-based models is the generation of many incorrect keyphrases. The pre-trained mT5-Base model performs well in title extraction and abstractive text summarization tasks. Therefore, this model was also used to extract the concept. It was observed that this model has a significant ability to predict the concept of the text.
In general, compared to traditional methods, deep learning-based models not only extract the keyphrases of the text but also generate the missing keyphrases. Future work on the concept extraction task can also extend this study to other languages.

References

[1] S. Jones and M. S. Staveley, "Phrasier: A system for interactive document retrieval using keyphrases", in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 160-167.

[2] Y. Zhang, N. Zincir-Heywood, and E. Milios, "World wide web site summarization", Web intelligence and agent systems: an international journal, Vol. 2, No. 1, 2004, pp. 39-53.

[3] E. Papagiannopoulou and G. Tsoumakas, "A review of keyphrase extraction", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 10, No. 2, 2020, p. e1339.

[4] J. Chen, X. Zhang, Y. Wu, Z. Yan, and Z. Li ", Keyphrase generation with correlation constraints," arXiv preprint arXiv:1808.07185, 2018.

[5] F. Boudin, Y. Gallina, and A. Aizawa, "Keyphrase generation for scientific document retrieval", arXiv preprint arXiv:2106.14726, 2021.

[6] S. Mehrabi, S. A. Mirroshandel, and H. Ahmadifar, "DeepSumm: A Novel Deep Learning-Based Multi-Lingual Multi-Documents Summarization System", Journal of Information Systems and Telecommunication (JIST), 2019, p. 204.

[7] K. Barker and N. Cornacchia, "Using noun phrase heads to extract document keyphrases", in Advances in Artificial Intelligence: 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, 2000, pp. 40-52.

[8] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "KEA: Practical automatic keyphrase extraction", in Proceedings of the fourth ACM conference on Digital libraries, 1999, pp. 254-255.

[9] S. N. Kim and M.-Y. Kan, "Re-examining automatic keyphrase extraction approaches in scientific articles", in Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE), 2009, pp. 9-16.

[10] C. Zhang, "Automatic keyword extraction from documents using conditional random fields", Journal of Computational Information Systems, 2008, vol. 4, no. 3, pp. 1169-1180.

[11] M. Barla and M. Bieliková, "From ambiguous words to key-concept extraction", in 24th International Workshop on Database and Expert Systems Applications, 2013, pp. 63-67: IEEE.

[12] S. M. H. Khozani and H. Bayat, "Specialization of keyword extraction approach to persian texts", in International Conference of Soft Computing and Pattern Recognition (SoCPaR), 2011, pp. 112-116.

[13] S. R. El-Beltagy and A. Rafea, "KP-Miner: A keyphrase extraction system for English and Arabic documents", Information systems, 2009, Vol. 34, No. 1, pp. 132-144.

[14] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents", in Text Mining: Applications and Theory, 2010, pp. 1-20.

[15] R. Campos et al., "Yake! collection-independent automatic keyword extractor", in Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Vol. 40, 2018, pp. 806-810.

[16] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text", in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004, pp. 404-411.

[17] X. Wan, and J. Xiao, "Single document keyphrase extraction using neighborhood knowledge", in Proceedings of the 23rd AAAI Conference on Artificial Intelligence, Vol. 8, 2008, pp. 855-860.

[18] A. Bougouin, F. Boudin, and B. Daille, "Topicrank: Graph-based topic ranking for keyphrase extraction", in Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), 2013, pp. 543-551.

[19] T. Tomokiyo and M. Hurst, "A language model approach to keyphrase extraction", in Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment, 2003, pp. 33-40.

[20] Z. Liu, X. Chen, Y. Zheng, and M. Sun, "Automatic keyphrase extraction by bridging vocabulary gap", in Proceedings of the Fifteenth Conference on Computational Natural Language Learning, 2011, pp. 135-144.

[21] E. Doostmohammadi, M. H. Bokaei, and H. Sameti, "Perkey: A persian news corpus for keyphrase extraction and generation", in 2018 9th International Symposium on Telecommunications (IST), 2018, pp. 460-465.

[22] I. Hsu, G. Xiao, N. Premkumar, and P. Nanyun", Discourse-level relation extraction via graph pooling", arXiv preprint arXiv:2101.00124 ,2021.

[23] E. Oro, R. Massimo, and S. Domenico, "Ontology-based information extraction from pdf documents with xonto", International Journal on Artificial Intelligence Tools, Vol. 18, No. 05, 2009, pp. 673-695.

[24] M. Gayathri, and R. J. Kannan, "Ontology based concept extraction and classification of ayurvedic documents", Procedia Computer Science, Vol. 172, 2020, pp. 511-516.

[25] X. Yuan, T. Wang, R. Meng, K. Thaker, P. Brusilovsky, D. He, A. Trischler, "One size does not fit all: Generating and

evaluating variable number of keyphrases", arXiv preprint arXiv:1810.05241, 2018.
[26] A. Swaminathan, R. K. Gupta, H. Zhang, D. Mahata, R. Gosangi, and R. R. Shah, "Keyphrase generation for scientific articles using gans (student abstract) ", in Proceedings of the AAAI Conference on Artificial Intelligence, 2020, Vol. 34, No. 10, pp. 13931-13932.
[27] Z. Sun, J. Tang, P. Du, Z.-H. Deng, and J.-Y. Nie, "Divgraphpointer: A graph pointer network for extracting diverse keyphrases", in Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 755-764.
[28] R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky, and Y. Chi, "Deep keyphrase generation", arXiv preprint arXiv:1704.06879, 2017.
[29] Y. Zhang and W. Xiao, "Keyphrase generation based on deep seq2seq model", IEEE access, Vol. 6, 2018, pp. 46047-46057.
[30] W. Chen, H. P. Chan, P. Li, L. Bing, and I. King, "An integrated approach for keyphrase generation via exploring the power of retrieval and extraction", arXiv preprint arXiv:1904.03454, 2019.
[31] E. Doostmohammadi, M. H. Bokaei, and H. Sameti, "Persian keyphrase generation using sequence-to-sequence models", in 2019 27th Iranian Conference on Electrical Engineering (ICEE), 2019, pp. 2010-2015.
[32] A. Glazkova, and D. Morozov, "Exploring Fine-tuned Generative Models for Keyphrase Selection: A Case Study for Russian", arXiv preprint arXiv:2409.10640, 2024.
[33] A. Glazkova, D. Morozov, and T. Garipov, "Key Algorithms for Keyphrase Generation: Instruction-Based LLMs for Russian Scientific Keyphrases", arXiv preprint arXiv:2410.18040, 2024.
[34] E. Thomas, and S. Vajjala, "Improving Absent Keyphrase Generation with Diversity Heads", in Findings of the Association for Computational Linguistics: NAACL 2024, 2024, pp. 1568-1584.
[35] M. Song, Y. Feng, and L. Jing, "A Preliminary Empirical Study on Prompt-based Unsupervised Keyphrase Extraction", arXiv preprint arXiv:2405.16571, 2024.
[36] L. Shen, and X. Le, "An enhanced method on transformer-based model for one2seq keyphrase generation", Electronics, Vol. 12, No. 13, 2023, p. 2968.
[37] N. S. Shirwandkar and S. Kulkarni, "Extractive text summarization using deep learning", in 2018 fourth international conference on computing communication control and automation (ICCUBEA), 2018, pp. 1-5.
[38] M. E. Khademi, M. Fakhredanesh, and S. M. Hoseini, "Farsi conceptual text summarizer: a new model in continuous vector space", Journal of Information Systems and Telecommunication (JIST), Vol. 1, No. 25, 2019, p. 23.
[39] M. Afsharizadeh, H. Ebrahimpour-Komleh, A. Bagheri, and G. Chrupała, "A Survey on Multi-document Summarization and Domain-Oriented Approaches", Journal of Information Systems and Telecommunication (JIST), Vol. 1, No. 37, 2022, p. 68.
[40] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E.D. Trippe, J.B. Gutierrez, and K. Kochut, "Text summarization techniques: a brief survey", arXiv preprint arXiv:1707.02268, 2017.
[41] S. Gupta, and S. K. Gupta, "Abstractive summarization: An overview of the state of the art", Expert Systems with Applications, Vol. 121, 2019, pp. 49-65.
[42] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert", arXiv preprint arXiv:1904.09675, 2019.
[43] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization", in International Conference on Machine Learning, 2020, pp. 11328-11339: PMLR.
[44] L. Shen, and X. Le, "An enhanced method on transformer-based model for one2seq keyphrase generation", Electronics, Vol. 12, No. 13 , 2023, p. 2968
[45] N. Datta, "Extractive Text Summarization of Clinical Text Using Deep Learning Models", in 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE), 2024, pp. 1-6.
[46] F. Liu, C. Xiong, " A Generative Text Summarization Method Based on mT5 and Large Language Models", in 2023 Eleventh International Conference on Advanced Cloud and Big Data (CBD), 2023, pp. 174-179.
[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805, 2018.
[48] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, "Parsbert: Transformer-based model for persian language understanding", Neural Processing Letters, Vol. 53, 2021, pp. 3831-3847.
[49] K. Cho, B. v. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation", arXiv preprint arXiv:1406.1078, 2014.
[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need. Advances in neural information processing systems", Advances in neural information processing systems, Vol. 30, 2017.
[51] D. Wang, C. Hansen, L.C. Lima, C. Hansen, M. Maistro, J.G. Simonsen, and C. Lioma, "Multi-Head Self-Attention with Role-Guided Masks", in Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43,2021, pp. 432-439.
[52] T. Xiao, Y. Li, J. Zhu, Z. Yu, and T. Liu, "Sharing attention weights for fast transformer", arXiv preprint arXiv:1906.11024, 2019.
[53] S. Yildirim and M. Asgari-Chenaghlu, "Mastering Transformers: Build state-of-the-art models from scratch with advanced natural language processing techniques", Packt Publishing Ltd, 2021.
[54] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P.J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer", The Journal of Machine Learning Research, Vol. 21, No. 1, 2020, pp. 5485-5551.
[55] L. Xue, "mT5: A massively multilingual pre-trained text-to-text transformer", arXiv preprint arXiv:2010.11934, 2020.