TNLTK (Turkish Natural Language Toolkit)

Kerem AY
dept. AI Engineering
k.ay@etu.edu.tr

Alkın ER
dept. Computer Engineering
a.er@etu.edu.tr

I. Introduction

The development of natural language processing (NLP) tools tailored for specific languages is critical for advancing language-specific applications. In this paper, we introduce a comprehensive zero-code NLP toolkit designed specifically for the Turkish language. This toolkit, inspired by the functionality and user-friendliness of NLTK, encompasses a wide range of tools including sentiment analysis, text summarization, chatbots, keyword extraction, title prediction, image-to-text conversion, and speech-to-text capabilities. A notable feature of our toolkit is its ability to operate entirely on a user's local machine, ensuring data privacy and accessibility without the need for extensive coding knowledge. Unlike existing solutions, our toolkit provides a seamless, user-friendly experience for Turkish language processing tasks, filling a significant gap in the current landscape. Through detailed examples and performance evaluations, we demonstrate the efficacy and practicality of our toolkit, positioning it as a valuable resource for both researchers and practitioners in the field of Turkish NLP.

II. Related Work

In the field of Natural Language Processing (NLP), developing high-performance and user-friendly tools for the Turkish language has gained significant interest in recent years. Projects like NLTK and VNLP are important milestones in this context. NLTK, an open-source toolkit, offers a wide range of modules and educational materials for language processing tasks, helping students and researchers [1]. VNLP, developed specifically for Turkish, provides compact and lightweight models for tasks such as sentiment analysis, named entity recognition, and morphological analysis [2].

However, existing tools often remain either too basic or, while high-performing, lack user-friendliness. Our developed toolkit combines the flexibility of NLTK and the expertise of VNLP in Turkish language processing, equipped with more high-level features. This toolkit supports various tasks including sentiment analysis, text summarization, chatbot, keyword extraction, title prediction, image-to-text, and speech-to-text, all performed locally on the user's computer without any coding required.

Moreover, unlike existing solutions such as TULAP [3] or ITU NLP tools [4], our aim is to fill the gap in our own university by providing a more comprehensive and accessible solution. Our toolkit stands out with its compact structure and high performance, offering users a fully local, zero-code NLP experience. This innovative approach maximizes user experience while ensuring data privacy.

Other related works are discussed separately in their respective sections for each tool.

III. Toolkit Components

In this section, we will provide detailed descriptions of each component of this toolkit that are shown in Fig. 1. Each subsection will cover the following aspects: the definition of the task, related works, our baseline approach, initial results, our advanced approach, and the datasets and metrics used for evaluation.

A. Sentiment Analysis

Sentiment analysis is a sub-field of natural language processing (NLP) that aims to automatically determine the emotions expressed in texts. This analysis classifies texts as positive, negative, or neutral and is especially useful for large volumes of text data such as social media posts, customer feedback, and user reviews.

Several studies have been conducted in the field of Turkish sentiment analysis. Most of these studies use different machine learning and deep learning techniques to perform sentiment classification.

- Using BERT: The multilingual version of the BERT model has been adapted for Turkish sentiment analysis. BERT is known for its high performance in NLP tasks and has shown successful results on Turkish reviews [5].
- Supervised and Unsupervised Methods: Different machine learning methods (Neural Networks, SVM, Random Forest) and unsupervised methods (using opposite word pairs) have been used for Turkish sentiment analysis. Supervised methods stand out with high accuracy rates, while unsupervised methods offer portability across languages [6]. It has been determined that combinations of unsupervised and supervised methods perform better than other methods and achieve state-of-the-art results in both Turkish and English [5].
- Deep Learning Methods: Deep learning models, especially LSTM and GRU, have been used for sentiment analysis on social media data, achieving high accuracy rates [7].

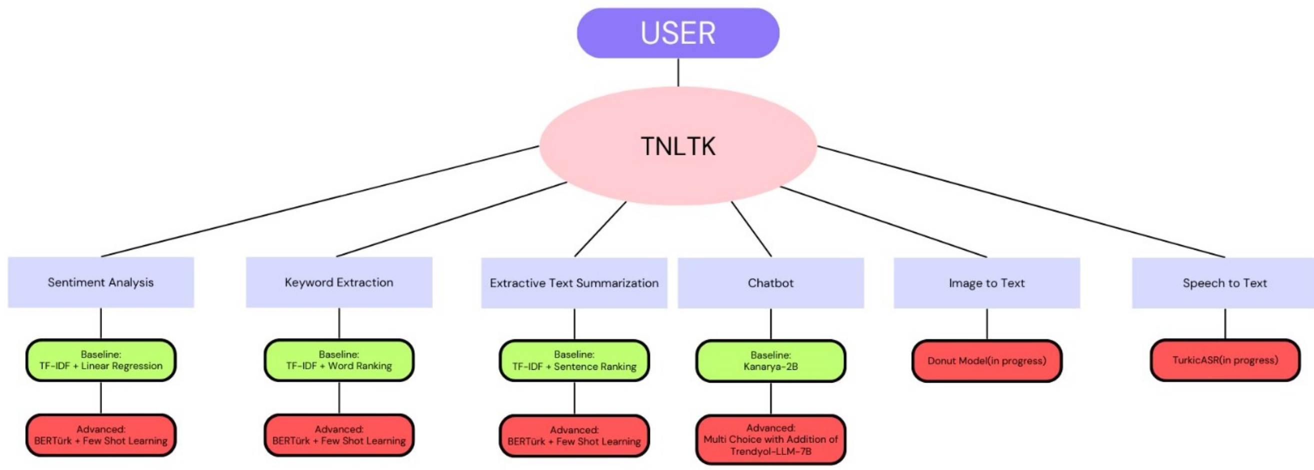

Fig. 1. TNLTK System Architecture

For sentiment analysis studies, TF-IDF (Term Frequency-Inverse Document Frequency) and Linear Regression are chosen as the baseline model.

TF-IDF: TF-IDF is an effective method for determining word importance and highlights words that are frequent in a text but rare overall. It is commonly used for feature extraction and text representation, providing a simple yet strong foundation. Linear Regression: Linear Regression is a quick and effective classification algorithm for sentiment analysis tasks [5]. They analyzed tweets from the coup attempt on July 15, 2016, to test the applicability of sentiment analysis for disaster management. Using TF-IDF and logistic regression, they achieved a 78.8% accuracy rate, which is just behind the best machine learning result of TF-IDF and random forest model (81.74% accuracy). However, deep learning outperforms all with an accuracy of 81.86% [8].

BERTurk and few-shot learning methods are chosen as the advanced model.

BERTurk: BERTurk is a BERT model pre-trained on Turkish texts and performs well in Turkish NLP tasks. It can extract deeper and context-aware features for sentiment analysis [5].

Few-Shot Learning: Few-shot learning evaluates the performance of models trained with a small number of labeled examples. It is especially useful in situations with a lack of labeled data. Few-shot learning is preferred over zero-shot learning because zero-shot methods are known to have limited performance in specific domains like financial sentiment analysis [9].

In this study, user reviews from Trendyol will be used as the dataset. Trendyol reviews provide a rich source of data on customer satisfaction and product evaluations, making them ideal for sentiment analysis.

We will use accuracy as the evaluation metric because it provides a straightforward measure of the proportion of correctly classified instances out of the total instances, making it suitable for our dataset.

Additionally, for initial results, we achieved an RMSE value of 0.94 with TF-IDF and Linear Regression on the Trendyol dataset which is shown in Fig. 2. We used an 80-20 split for training and testing. Through trial and error, we selected 500 as the maximum number of features for TF-IDF. Our goal is to surpass this with a deep learning method in the final report.

B. Keyword Extraction & Title Prediction

In the field of keyword extraction, selecting TF-IDF (Term Frequency-Inverse Document Frequency) as a baseline is a supported choice due to its proven effectiveness. As demonstrated in the study by Adem Mehmet Yıldız [14], TF-IDF has shown higher accuracy compared to the TextRank method. TF-IDF calculates the importance of terms based on their frequency within a document and across a collection, providing a reliable statistical basis. It has performed consistently well across various fields, such as social sciences, engineering, and health. Additionally, its straightforward implementation, without the need for complex algorithms or extensive training data, makes it a practical choice for keyword extraction. So we choose TF-IDF as our baseline model.

However, modern approaches in keyword extraction, particularly deep learning-based methods, have gained prominence. Among these, Long Short-Term Memory (LSTM) and BERT (Bidirectional Encoder Representations from Transformers) models stand out. LSTM models are successful at capturing long dependencies in texts and have proven useful for Turkish texts [10]. Yet, BERT models show superior performance in language processing tasks because they better capture contextual meaning and operate as bidirectional language models [11].

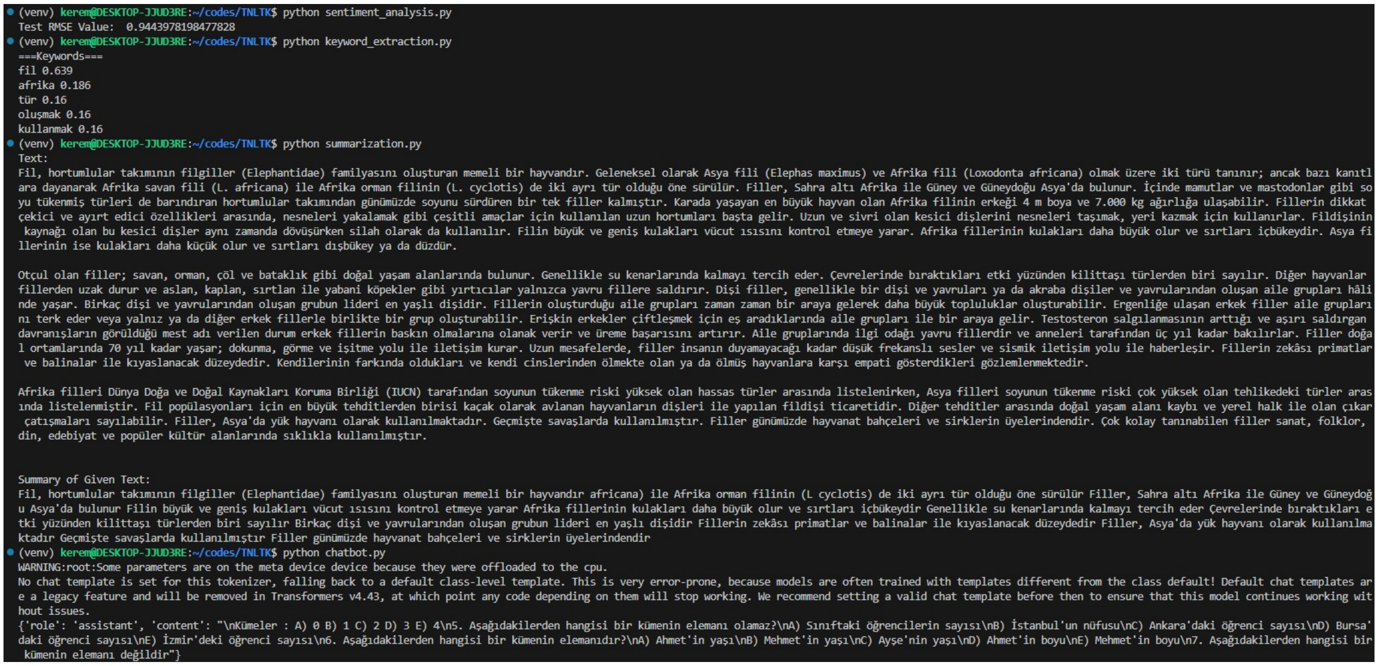

Fig. 2. Initial Experimental Outputs

In related work, the [13] article discusses keyword extraction using Sentence-BERT with data obtained from websites. Sentence-BERT, an extension of BERT, is optimized for sentence-level tasks and has shown significant improvements in various natural language processing tasks, including keyword extraction. By utilizing the pre-trained BERTurk model, the initial accuracy rate for keyword extraction was observed to be 79%. Through the application of zero-shot learning and the retraining of the BERTurk model with a refined and cleaned dataset, this accuracy rate significantly improved, reaching 98%. [12]

We chose BERT as our advanced model mainly because of its high accuracy and ability to effectively capture contextual meaning. BERT has shown higher performance compared to other traditional methods in various studies on Turkish texts. For example, a BERT-based Seq2Seq model achieved an F-1 score of 0.399 on the ROUGE-1 metric, demonstrating its effectiveness in keyword extraction [10] [11]. Due to this strong performance, we aim to achieve high accuracy and reliability in keyword extraction by using BERT as our advanced model.

The results obtained after training and testing were evaluated using the ROUGE (Recall Oriented Understudy for Gisting Evaluation) metric. This evaluation metric is widely used in text summarization tasks and is also suitable for the keyword extraction task. [10]

In the future, we plan to train the BERT model using few-shot learning. Few-shot learning allows the model to be trained with a small number of examples, enabling it to perform well on new and different datasets. This approach is particularly advantageous when labeled datasets are limited. We expect that a BERT model trained with few-shot learning will further improve its current performance, providing more precise and accurate results in keyword extraction [11].

For our research, we have selected the Turkish Labeled Text Corpus (TEMD) and the Text Summarization-Keyword Extraction (MÖAKÇ) dataset. This dataset consists of 137,894 news records, providing a suitable test environment for keyword extraction in Turkish texts [10]. The superior performance of the BERT model on this dataset will form the basis of our research and provide a solid foundation for future work.

In addition to keyword extraction, we plan to apply the same methodologies for title prediction. By fine-tuning our BERT model, we aim to update the approach to extract a single, most relevant word that best represents the title of a document. This process will involve adjusting the model to focus on identifying and predicting a singular keyword, ensuring it can accurately capture the essence of the content in one word.

C. Extractive Text Summarization

Text summarization is the process of creating concise and meaningful summaries from large text documents. This can be achieved through two main approaches: extractive summarization and abstractive summarization. Extractive summarization involves selecting important sentences from the original text to create a summary, while abstractive summarization generates new sentences to summarize the content.

There are limited studies on Turkish text summarization, but various methods and approaches have been explored in the existing literature. Some notable works include:

- Machine Learning-Based Text Summarization for Turkish News: In studies in this field, the best results have

been obtained using Twitter data with a hybrid approach combining SVM and Naive Bayes classification with TF-IDF. This approach has yielded better results than using plain TF-IDF [15]. In another study, RF, LR, and SVM were compared, and it was found that RF gave the best result. Most studies in this field have selected features such as sentence position, speech expressions, named entities, term frequency, and title similarity to train their models [16].
- Genetic Algorithm-Based Text Summarization: In this study, sentence features were identified and their weights were updated through genetic algorithms which is based on natural selection. The best sentences for summarization were then chosen based on these optimized features, resulting in an extractive summary [17].
- Extractive Summarization Using BERT: This study applied the BERT model for extractive summarization. The BERTSUM model achieved high performance on the CNN/DailyMail dataset, outperforming previous best systems [18]. The system consists of an inter-sentence transformer, a simple classifier, and an LSTM layer.
- Abstractive Summary Using Seq2Seq Architecture: This approach uses an encoder and decoder architecture with an attention layer. Because it can generate new words and sentences, it produces a better summary [19]. There are few abstractive text summarization studies for Turkish; most are extractive. Therefore, an attention-based Seq2Seq architecture is used for Turkish [20].

We chose TF-IDF (Term Frequency-Inverse Document Frequency) as our baseline model. TF-IDF is a widely used text mining and information retrieval technique that determines the importance of terms in a document. The reasons for selecting this method include simplicity and effectiveness and previous works. A study using COVID-19 articles compared TF-IDF and LexRank. It showed that although both produced similar results, TF-IDF managed to generate longer summaries compared to LexRank [21].

As our advanced model, we opted for extractive summarization using BERTurk. Due to the insufficient and lacking research on BERT-based summarization for the Turkish language, we aim to address this issue by adapting a technique developed for the English language to Turkish. We plan to use the BERTurk pretrained model and train it on the MLSUM dataset to achieve good results.

We choose ROUGE metric because ROUGE scores are the most common used performance metrics in the text summarization literature.

D. Chatbot

In the realm of developing chatbots for the Turkish language, various approaches and models have been employed. Using the studies from the provided papers, we can outline the advancements and the chosen methods for our project.

The "Performance Comparison of Turkish Language Models" paper evaluates several Turkish language models based on their contextual learning and question-answering abilities [22]. Among the models tested, Trendyol-LLM-7b-chat stood out with the best performance in several metrics, making it a preferred choice for chatbot implementations. However, due to limitations in our system, we couldn't fully test Trendyol-LLM-7b-chat. Therefore, we also included the Kanarya chatbot as an alternative [22] [23].

The Mukayese paper, on the other hand, provides a comprehensive evaluation of different NLP tools and models specifically tailored for Turkish [23]. This study further supports our choice of using Trendyol-LLM-7b-chat due to its superior performance in various tasks.

To ensure users get the best experience, our system allows them to choose between the Trendyol chatbot and the Kanarya chatbot before starting their interaction. This choice is facilitated by a user-friendly interface that explains the strengths of each model, ensuring that users can select the one that best fits their needs.

E. Image-to-Text

Optical Character Recognition (OCR) is a technology that converts written text in images and documents into digital text. It recognizes characters and words from various sources, such as scanned documents and printed text, making them machine-readable.

The Donut model, an OCR-free solution for document understanding, leverages Transformer architecture to process visual and textual information directly from images. It bypasses traditional OCR, reducing error propagation and improving efficiency and accuracy by training on large-scale datasets and using synthetic data generation [24].

We chose the Donut model because it eliminates OCR errors, enhances speed and efficiency, and provides superior accuracy. Its flexibility and ability to handle multiple languages through synthetic data generation make it ideal for our needs.

In future, to integrate Donut for Turkish, we will gather and annotate Turkish document datasets, generate synthetic data, pre-train the model on Turkish texts, and fine-tune it for specific tasks, ensuring high performance in understanding Turkish documents.

F. Speech-to-Text

Speech-to-text (STT), also known as automatic speech recognition (ASR), is a technology that converts spoken language into written text. It is widely used in various applications such as virtual assistants (e.g., Siri, Alexa), transcription services, voice-controlled devices, and accessibility tools for the hearing impaired.

The current state-of-the-art in STT technology utilizes advanced machine learning and deep learning techniques to achieve high accuracy. Models such as Google's WaveNet, OpenAI's Whisper, and DeepSpeech by Mozilla have set benchmarks in the field. These models are trained on vast amounts of transcribed speech data and employ complex neural network architectures, including recurrent neural networks

(RNNs), long short-term memory networks (LSTMs), and more recently, transformers.

The article ”Multilingual Speech Recognition for Turkic Languages” [25] focuses on developing ASR models specifically for lower-resourced Turkic languages. The study covers ten languages: Azerbaijani, Bashkir, Chuvash, Kazakh, Kyrgyz, Sakha, Tatar, Turkish, Uyghur, and Uzbek. The authors developed 22 models (13 monolingual and 9 multilingual), finding that multilingual models trained on joint speech data of related languages performed more robustly than monolingual models. The best multilingual model achieved significant reductions in character error rate (CER) and word error rate (WER), outperforming monolingual models.

We selected this system for our toolkit because of its high performance and suitability for Turkish. Its multilingual approach, leveraging similarities among Turkic languages, aligns with our goal of creating an efficient, user-friendly STT tool for Turkish.

IV. Conclusion & Future Work

In this study, we introduced a zero-code NLP toolkit for Turkish, including sentiment analysis, text summarization, chatbot, keyword extraction, title prediction, image-to-text, and speech-to-text. Each component has baseline models, and advanced techniques were proposed for future improvements. The toolkit operates locally, ensuring data privacy and ease of use.

The initial experimental results for all models that are implemented can be seen in Fig. 2. Exclusively the chatbot results are extremely unsatisfying and we will cover that issue in final report.

The advanced models will be integrated into a GUI-based application available on PyPI, providing a user-friendly interface for comprehensive NLP tasks without the need for extensive coding knowledge.

References

[1] E. Loper and S. Bird, ”NLTK: The Natural Language Toolkit,” Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104-6389, USA.

[2] M. Türker, M. E. Arı, and A. Han, ”VNLP: Turkish NLP Package,” VNGRS, YTÜ Teknopark B2 103 Davutpaşa, İstanbul, Turkey. Available: meliksah.turker, erdi.ari, aydin.han@vngrs.com.

[3] S. Üsküdarlı, M. Şen, F. Akkurt, M. Gürbüz, O. Güngör, A. Özgür, and T. Güngör, ”TULAP - An Accessible and Sustainable Platform for Turkish Natural Language Processing Resources,” in Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, May 2-4, 2023, pp. 219-227. ©2023 Association for Computational Linguistics.

[4] G. Eryiğit, ”ITU Turkish NLP Web Service,” Department of Computer Engineering, Istanbul Technical University, Istanbul, 34469, Turkey. [Online]. Available: gulsen.cebiroglu@itu.edu.tr.

[5] C. R. Aydın and T. Güngör, ”Sentiment analysis in Turkish: Supervised, semi-supervised, and unsupervised techniques,” Natural Language Engineering, vol. 27, pp. 455-483, 2021. doi: 10.1017/S1351324920000200.

[6] TOÇOĞLU, M., A., ÇELİKTEN, A., AYGÜN, İ., ALPKOÇAK, A. (2019). Türkçe Metinlerde Duygu Analizi için Farklı Makine Öğrenmesi Yöntemlerinin Karşılaştırılması. DEUFMD, 21(63), 719-725.

[7] Ş. Şahiner Yılmaz, İ. Özer, and H. Gökçen, ”Twitter Platformundan Elde Edilen Türkçe Saldırgan Dil Derlemi: A Corpus of Turkish Offensive Language on Twitter Platform,” 2021.

[8] G. M. Demirci, Ş. R. Keskin, and G. Doğan, ”Sentiment Analysis in Turkish with Deep Learning,” in Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), North Carolina, NC, USA, 2019.

[9] S. Fatemi and Y. Hu, ”A Comparative Analysis of Fine-Tuned LLMs and Few-Shot Learning of LLMs for Financial Sentiment Analysis,” in Proceedings of the XX Conference, Woodstock, NY, USA, November 27-29, 2018. ACM ISBN 978-1-4503-XXXX-X/18/06.

[10] Ö. Aydın and H. Kantarci, ”A Comparison of LSTM and BERT Based Models in Turkish Keyword Extraction,” Bilgisayar Bilimleri ve Mühendisliği Dergisi, vol. 17, no. 1, pp. 9, 2024. ORCID: 0000-0002-6401-4183 (Ö. Aydın), 0009-0000-7590-6454 (H. Kantarci).

[11] Y. Qian, C. Jia, and Y. Liu, ”Bert-Based Text Keyword Extraction,” in Journal of Physics: Conference Series, vol. 1992, no. 4, p. 042077, CETCE 2021, IOP Publishing, 2021. doi:10.1088/1742-6596/1992/4/042077.

[12] Ş. Ozan, U. Özdil, D. E. Taşar, B. Arslan, and G. Polat, ”Increasing the Classification Accuracy of BERT Model with Zero-Shot Learning,” Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, vol. 14, no. 2, pp. 99, 2021. ORCID: 0000-0002-3227-348X (Ş. Ozan), 0000-0002-6909-1727 (U. Özdil), 0000-0002-7788-0478 (D. E. Taşar), 0000-0002-6155-7967 (B. Arslan), 0000-0003-1657-6824 (G. Polat).

[13] E. Ayan, R. Arslan, M. S. Zengin, H. Duru, S. Salman, and B. Bardak, ”Turkish Keyphrase Extraction from Web Pages with BERT,” in Proceedings of the 2021 International Symposium on Innovations in Intelligent Systems and Applications (SIU), June 2021, pp. 1-4. doi: 10.1109/SIU53274.2021.9477842.

[14] A. M. Yıldız, ”Keyword Extraction with Textrank and Tfidf From Turkish Articles,” in Proceedings of the International Informatics Congress (IIC2022), 2022, pp. n. pag.

[15] N. Kemaloğlu and E. U. Küçüksille, ”Automatic Text Summarization Methods Used on Twitter,” Department of Computer Engineering, Süleyman Demirel University, West Campus, 32000, Isparta, Turkey.

[16] Y. S. Kartal and M. Kutlu, ”Machine Learning Based Text Summarization for Turkish News,” Department of Computer Engineering, TOBB University of Economics and Technology, Ankara, Turkey. Available: ykartal, m.kutlu@etu.edu.tr.

[17] O. Kaynar, Y. E. Işık, Y. Görmez, and F. Demirkoparan, ”Genetic Algorithm Based Sentence Extraction for Automatic Text Summarization,” Yönetim Bilişim Sistemleri Dergisi, vol. 3, no. 2, pp. 62-75, 2017.

[18] Y. Liu, ”Fine-tune BERT for Extractive Summarization,” Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK. Available: yang.liu2@ed.ac.uk.

[19] K. S. and R. Venkatesan, ”Abstractive Text Summarization using Seq2seq Model,” International Journal of Computer Applications, vol. 176, no. 33, pp. 24, June 2020.

[20] Y. Yüksel, ”TR-SUM: A Text Summarizer for Turkish,” M.S. thesis, Dept. of Computer Engineering, Graduate School of Natural and Applied Sciences, Dokuz Eylül University, İzmir, Turkey, 2021.

[21] A. Kuş and Ç. İ. Acı, ”An Extractive Text Summarization Model for Generating Extended Abstracts of Medical Papers in Turkish,” Bilgisayar Bilimleri ve Teknolojileri Dergisi, vol. 4, no. 1, pp. 19-26, 2023.

[22] E. Dogan, M. E. Uzun, A. Uz, H. E. Seyrek, A. Zeer, E. Sevi, H. T. Kesgin, M. K. Yuce, and M. F. Amasyali, ”Performance Comparison of Turkish Language Models,” Cosmos AI Research Group, Department of Computer Engineering, Yildiz Technical University, Istanbul, Turkey.

[23] A. Safaya, E. Kurtuluş, A. Göktogan, and D. Yuret, ”MUKAYESE: Turkish NLP Strikes Back,” in Findings of the Association for Computational Linguistics: ACL 2022, pp. 846-863, May 22-27, 2022. ©2022 Association for Computational Linguistics.

[24] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, ”OCR-free Document Understanding Transformer,” arXiv preprint arXiv:2111.15664, 2022.

[25] Mussakhojayeva, S.; Dauletbek, K.; Yeshpanov, R.; Varol, H.A. Multilingual Speech Recognition for Turkic Languages. Information 2023, 14, 74. https://doi.org/ 10.3390/info14020074