PROCEEDINGS OF
INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS AND NEW APPLICATIONS
https://proceedings.icisna.org/
3rd International Conference on Intelligent Systems and New Applications (ICISNA'25), Antalya, December 12-14, 2025.

Embedding-Based Machine Learning Approach for Automatic Classification of Turkish News Articles

Ahmet Atasoglu[1], Yavuz Selim Taspinar[2]

[1]*Mechatronics Engineering Department, Selcuk University , Konya*
258265001007@ogr.selcuk.edu.tr, ORCID: 0009-0008-8178-2177

[2]*Mechatronics Engineering Department, Selcuk University , Konya*
ytaspinar@selcuk.edu.tr, ORCID: 0000-0002-7278-4241

***Abstract*— In this study, an automatic text classification approach for Turkish news articles is presented. The savasy/ttc4900 dataset from HuggingFace, consisting of seven news categories, was used. News texts were converted into 768-dimensional vector representations using the embeddinggemma model on the Ollama framework. These embeddings were then used to evaluate the performance of several machine learning algorithms. Seven models were tested: Support Vector Classifier (SVC), Logistic Regression, Multilayer Perceptron, K-Nearest Neighbors, Random Forest, Gaussian Naive Bayes, and Decision Tree. Model performance was assessed using accuracy, precision, recall, and F1-score metrics. Results showed that SVC and Logistic Regression achieved the highest accuracy in the high-dimensional embedding space. The findings demonstrate that embedding-based representations offer strong discriminative capability for Turkish news classification and that deep learning–derived vector embeddings can be effectively combined with traditional machine learning methods. These results emphasize the importance of vectorized text representations in natural language processing research.**

***Keywords*—** natural language processing, text embeddings, language model, text classification, Gemma language model

I. Introduction

The ability of computers to understand and generate human languages has long been one of the fascinating topics in computer science. The idea that machines could solve problems by exhibiting intelligent behavior like humans was strikingly presented by Alan Turing's question, “Can machines think?” [1], which made significant contributions to computer science. Today, it is possible to process human language not only at the symbolic or grammatical level, but also in terms of semantic relationships [2]. This has paved the way for transferring a level of understanding and interpretation similar to human intuition to computers. Developments in statistical modeling [3] and deep learning [4], in particular, have enabled the emergence of data-driven solutions based on data-driven training rather than rule-based solutions by teaching complex linguistic features to computers. These solutions are now important fundamental methods in the field of natural language processing. Today, with the development of deep learning-based approaches in the field of natural language processing and their increasing applicability, studies using deep learning methods have become quite widespread [5]. This study addresses the problem of text classification. Text classification is a problem that involves determining which category texts created in different contextual categories belong to. [6] Research in this field contributes to the development of natural language understanding systems. The complexity of the text classification problem is often increased by the diversity of large data sets, differences in language structure, and ambiguities in meaning. By addressing the text classification problem, this study aims to understand the existing challenges and obtain findings on overcoming these challenges using various methods. Furthermore, by presenting an analysis of the findings, the study aims to contribute to future work by highlighting the implications of these results. [7] introduced the long short-term memory (LSTM) architecture, a new deep learning method. This architecture aims to solve the long-term information storage problem caused by insufficient backpropagation and diminishing error feedback. The LSTM architecture consists of gates specialized for different tasks and is capable of handling complex tasks by directing the error flow of these gates. The evaluations shared in the study demonstrate that the model in the new architecture can solve complex and long-delay tasks that recursive neural network algorithms cannot solve.

In a study [8], the relationships between words and sentences were analyzed and an attempt was made to model important elements in texts using a PageRank (Brin & Page, 1998)-based ranking model. The similarity operation was applied to the calculated graph structure, identifying the most similar ranked nodes to reveal keywords and important sentences in the text. The study particularly addressed its use in tasks such as extracting important information from large texts or identifying

keywords related to specific topics. In their work [9], two new model architectures were introduced for the continuous representation of word vectors from large datasets. The quality of these representations was measured in a semantic similarity task, and the results were compared with previously best-performing techniques based on different types of neural networks. The evaluation results reported significant reductions in computational cost and substantial improvements in accuracy. The study reported that learning word vectors based on semantic similarities from a 1.6 billion-word dataset took less than a day. [10] analyzed the model features required to reveal patterns in word vectors. A new global regression model was developed that combines the advantages of two important model families: global matrix factorization and local context window methods. The newly developed model, GloVe, was trained on a large text corpus to effectively utilize statistical information. The model was then evaluated on semantic similarity and named entity recognition tasks and compared with other models in the literature. [11] introduced a new algorithm, Adam, for stochastic objective function optimization. The algorithm is based on adaptive estimates of low-order moments and is easy to implement, computationally efficient, has low memory requirements, is invariant to cross-scaling of gradients, and is suitable for problems with large data and parameters. Theoretical convergence properties of the algorithm are analyzed, and experimental results demonstrate that Adam can yield results comparable to those of other optimization methods. In their work [12], a new artificial neural network-based approach, distinct from statistical machine translation, was proposed. They argued that the use of fixed-length vectors was a bottleneck in improving the performance of the basic encoder-decoder architecture, and a new mechanism was developed that allows the algorithm to automatically search for important fragments in the source sentence to predict the target language word. The developed model was evaluated on an English-French translation task, and its performance was compared with existing results. The Transformer architecture presented in [13] is presented as a significant new step in the field of natural language processing. The Transformer architecture was based on [12], which introduced an attention mechanism that allows the model to pay more attention to the most important parts of the input array elements during training. The attention mechanism is central to the new Transformer architecture and, as presented in [14], consists of two blocks: an encoder and a decoder. The developed model has been evaluated on machine translation tasks and has been shown to yield the best results in comparisons in the literature.

[15] examined text summarization methods based on sentence extraction. Feature extraction and summary generation were carried out using genetic algorithms. A dataset consisting of Turkish summaries of news texts was used as the dataset. During training, the genetic algorithm determined the optimal weight values for the documents' features, and summaries were generated accordingly.

In their study [16], a dataset for evaluating different models across nine natural language comprehension tasks, called General Language Understanding Assessment (GLUE), and a test dataset for evaluating and comparing the models, were presented. The study also presented basic evaluation results using the ELMo language model [17] using a transfer learning technique. [18] introduced a new pretrained language model consisting solely of the encoder block of the Transformer architecture. The BERT language model presented in this study produces output vectors that represent input sequences from both perspectives. It has also been demonstrated that it can work on different tasks by retraining on the output layer. The model has been evaluated on various natural language processing tasks, demonstrating particularly high performance in natural language understanding tasks. In their study [19], a new variant of BERT [18] is proposed, named Sentence-BERT (SBERT), to address the network's inadequacy in semantic similarity search and clustering tasks. SBERT proposes a new ternary network structure, making it suitable for semantic similarity tasks. The developed model is evaluated on similarity tasks and transfer learning tasks, and comparisons with the literature are shared. In their work [20], a new language model called BART, equivalent to the original Transformer architecture [13], was introduced, with both encoder and decoder blocks. In this study, we used denoising approaches by randomly permuting the order of sentences in the learned texts to generalize the encoder- and decoder-based models. Specifically, we evaluated these approaches in natural language generation and question-answering tasks. [21] introduced the GPT-3 language model, which was developed to demonstrate that augmenting language models can significantly improve task-independent, small-sample performance, sometimes reaching a level that rivals the best previous transfer learning approaches. The developed model consists of 175 billion parameters. GPT-3's performance was evaluated solely through text interaction on tasks with small training samples, without model updating. GPT-3 demonstrated strong performance on several tasks, including translation, question-answering, gap-filling tasks, and word scrambling. However, methodological issues related to training on small-sample learning datasets and large text corpora, which still struggled for GPT-3, were identified. It was also reported that GPT-3 was able to write articles that were difficult to distinguish between human-written and human-written. [22] introduces a new variant of the T5 model [23], called mT5, trained on a new dataset covering 101 languages. mT5 is evaluated on tasks that currently evaluate existing language models and also introduces a simple technique to prevent the model from translating into the wrong language. The code and model used in the study are open source. [24] presented a new approach aimed at understanding and improving pairwise word vectors used in machine translation. This new approach proposes a simple merging method based on text vectors. Furthermore, a new norm-based method focuses on more efficient use of word vectors. The developed methods were also evaluated in machine translation and string-to-string tasks. [25] demonstrated the use of word vectors as a binary classification method. The problem identified was the detection of fake news. Furthermore, various preprocessing steps were introduced before classification. Another aspect of

the study was to expand binary classification to six categories of falsehood, ranging from true to completely false. However, it was determined that this method was not as effective as the results obtained with binary classification. In their study [26], a classification problem using convolutional neural networks (CNNs) was investigated for Turkish texts. The developed model was also compared with other machine learning methods on the same data. The selected datasets varied in terms of text and number of classes, and the effect of word vector size on classification success was investigated. Stemming and deletion of filler words were applied in the text preprocessing, and the TF-IDF method was applied for vector representations. The performance of different preprocessing and vector representations was evaluated against each other on the developed model. [27] used deep learning technologies to summarize Turkish texts and generate leading headlines. The study used a dataset consisting of over 50,000 collected news texts and their respective headlines, and applied the necessary preprocessing for training. High-performance results were achieved as a result of the use of the transformer model. The results showed that the transformer architecture performed better with less training content than other deep learning models and demonstrated a higher level of grammatical performance. [28] evaluate the performance of the GPT-3 large language model in classifying tweets containing and not containing cyberbullying on a dataset of Turkish tweets. The results demonstrate that GPT-3 achieves sufficient accuracy to be used in detecting cyberbullying in tweet content. The study examines ChatGPT, a prominent large language model. The study discusses ChatGPT's general benefits and usage scenarios, and discusses its potential and potential risks in various fields such as law, medicine, mathematics, finance, and academic writing.

II. Materials and Methods

A. Data Processing

The dataset used in this study is the Turkish news dataset, savasy/ttc4900, provided through the HuggingFace Datasets library. The dataset contains seven different categories: politics, world, economy, culture, health, sports, and technology. The aim of the study is to classify news texts belonging to one of these categories using machine learning models. The dataset was first retrieved using the load_dataset() function, and the number of samples for each category was examined using the dataset_samples() function. The data distribution was observed to be balanced across categories. The texts used in model training were taken from the text column, and the classes were drawn from the category column. The dataset was split into 70% training and 30% test sections using stratified sampling. This approach ensured that each class was balanced across the training and test sections.

B. Encoding Texts to Vector Embeddings

Text Data To enable the use of text data in machine learning models, it must be converted into vector representations. For this purpose, the study utilized the Ollama library to generate 768-dimensional feature vectors from news articles using the embeddinggemma model. The encode_text() function provides a single interface for embedding texts. To reduce processing costs for large datasets, the embedding process was applied using a 16-element mini-batch structure. After each batch was processed, the results were combined to create training and test data matrices.

C. Preparing Train and Test Sets

After creating the text vectors, X_train_final was defined as the training vector matrix, X_test_final as the test vector matrix, and y_train and y_test as the training and test labels. By converting the dataset into a format suitable for a vector-based model, different classification algorithms were enabled to work directly on high-dimensional vectors.

D. Classification Models

In this study, a total of seven different machine learning classifiers were used: Support Vector Classifier (SVC), Multi-Layer Perceptron Classifier (MLPClassifier), K-Nearest Neighbors (KNN), Random Forest Classifier (RF), Gaussian Naive Bayes (GNB), Logistic Regression (LR), and Decision Tree Classifier (DT). All models were trained with their default hyperparameters, thus comparing the fundamental effects of vector-based features across different algorithms. Throughout the training process, each model was trained with the training vector matrix and produced predictions on the test vector matrix.

E. Evaluating Models

Performance metrics were calculated separately for each classifier: Accuracy, Precision, Recall, and F1 Score. Additionally, to examine the class-based behavior of the models, confusion matrices (separate for each classifier), category-based F1 score comparisons, One-vs-Rest and ROC curves, and graphs containing AUC values were created. In the ROC analysis, decision function outputs were used for models without probability outputs, and if these were also unavailable, a warning message was placed on the relevant subgraph. All results were summarized in the final section by creating a comparison table.

III. Experimental Results

This section presents comparative performance analyses of seven different classification algorithms trained on embedding representations obtained from news texts. The models were evaluated on the test dataset using accuracy, precision, recall, and F1-score metrics, and the results are summarized in Figure 1. Figure 2 presents the confusion matrices of all the machine learning models evaluated in the study.

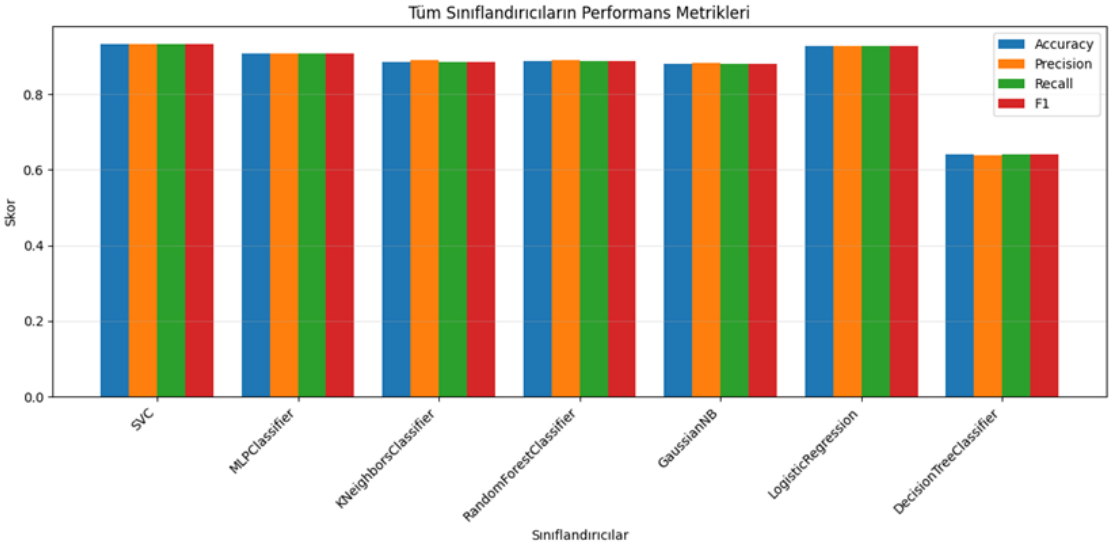

Fig 1. Performance metrics

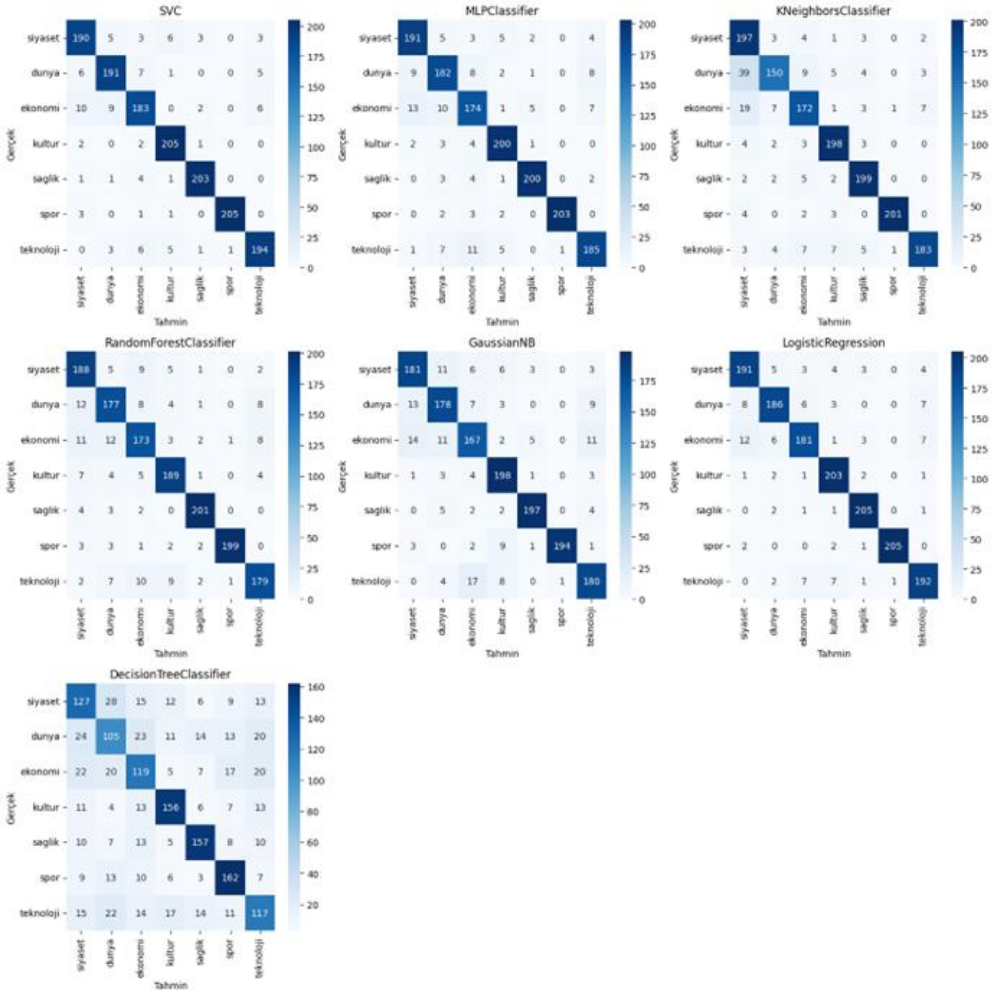

Fig. 2. Confusion matrix of all models

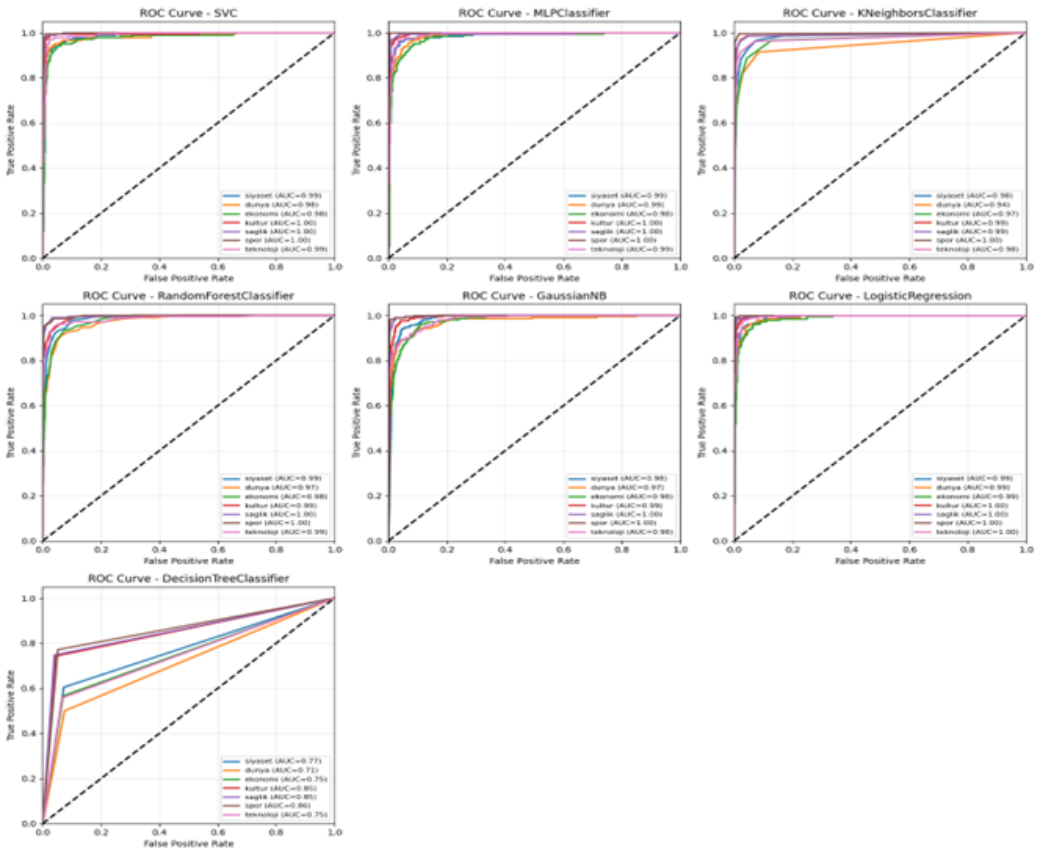

Fig 3. ROC of all modes

According to the results obtained, the Support Vector Classifier (SVC) model was the most successful classifier with 93.27% accuracy across all metrics. Logistic Regression (92.72%) followed SVC with a very close performance. The high performance of these two models demonstrates that linear/kernel-based methods are effective in high-dimensional embedding spaces.

When examining other models, the MLPClassifier achieved relatively high success (90.82%), while RandomForest (88.84%) and KNN (88.44%) performed at an intermediate level. Considering the continuous and dense structure of embedding-based representation, it was observed that tree and neighborhood-based methods could perform more limited discrimination in this space.

The GaussianNB (88.10%) model achieved reasonable accuracy despite its generative structure; however, it showed lower success compared to other discriminative models due to the complexity of the class distribution. The lowest performance was achieved with the DecisionTree classifier (64.15%), revealing that the single tree structure was insufficient to capture the complex data distribution based on embedding.

Overall, the results show that embedding-based vector representations provide high discriminative power for Turkish news classification; specifically, linear/kernel-based methods such as SVC and Logistic Regression perform best with these types of representations. Figure 3 presents the ROC curves of the machine learning models.

IV. Conclusions

In this study, a vector-based approach was used to classify Turkish news texts by category, and seven different machine learning algorithms were compared. Since the text representation vectors derived from the Gemma model can directly learn semantic information in natural language, high classification performance was achieved in all models. The results obtained showed that some models are naturally more compatible with vector-based data. In particular, SVC, Logistic Regression, and Random Forest achieved the highest accuracy on the test set. The Gaussian Naive Bayes and Decision Tree models performed less well than other models on high-dimensional text vectors. When examining the One-vs-Rest ROC curves, it was observed that the AUC values were high for all classes in many models. This indicates that vector-based representations provide strong distinctions between classes. Overall, the findings of the study reveal that large language model-based vector models are quite effective for Turkish news classification. For future work, model performance can be further improved by performing hyperparameter optimization, different vector models (BERT, RoBERTa, Mistral, etc.) can be added to the comparison, end-to-end learning can be applied with larger neural network models, and data augmentation techniques can be tested on the news content in the dataset. In

conclusion, this study demonstrates the high accuracy performance achieved by vector-based text classification approaches using different algorithms in Turkish news categories.

REFERENCES

[1] A. M. Turing, "I.—Computing Machinery And Intelligence", *Mind*, c. LIX, sy 236, ss. 433-460, Eki. 1950, doi: 10.1093/mind/LIX.236.433.

[2] Y. LeCun, Y. Bengio, ve G. Hinton, "Deep learning", *Nature*, c. 521, sy 7553, Art. sy 7553, May. 2015, doi: 10.1038/nature14539.

[3] Y. Bengio, R. Ducharme, ve P. Vincent, "A Neural Probabilistic Language Model", içinde *Advances in Neural Information Processing Systems*, MIT Press, 2000. Erişim: 14 Ocak 2024. [Çevrimiçi]. Erişim adresi: https://papers.nips.cc/paper_files/paper/2000/hash/728f206c2a01bf572b5940d7d9a8fa4c-Abstract.html

[4] F. Hu, "Survey on Neural Networks in Natural Language Processing", içinde *2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)*, Nis. 2023, ss. 591-594.doi: 10.1109/ICCECT57938.2023.10141113.

[5] D. Küçük ve N. Arici, "Doğal dil işlemede derin öğrenme uygulamalari üzerine bir literatür çalişmasi", *UYBİSBBD*, c. 2, sy 2, Art. sy 2, Ara. 2018.

[6] A. C. Tantuğ, "Metin Sınıflandırma", *TBV-BBMD*, c. 5, sy 2, Art. sy 2, Haz. 2016.

[7] S. Hochreiter ve J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, c. 9, sy 8, ss. 1735-1780, Kas. 1997, doi: 10.1162/neco.1997.9.8.1735.

[8] R. Mihalcea ve P. Tarau, "TextRank: Bringing Order into Text", içinde *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, D. Lin ve D. Wu, Ed., Barcelona, Spain: Association for Computational Linguistics, Tem. 2004, ss. 404-411. Erişim: 13 Ocak 2024. [Çevrimiçi]. Erişim adresi: https://aclanthology.org/W04-3252

[9] T. Mikolov, K. Chen, G. Corrado, ve J. Dean, "Efficient Estimation of Word Representations in Vector Space", 06 Eylül 2013, *arXiv*: arXiv:1301.3781. doi: 10.48550/arXiv.1301.3781.

[10] J. Pennington, R. Socher, ve C. Manning, "Glove: Global Vectors for Word Representation", içinde *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, ss. 1532-1543. doi: 10.3115/v1/D14-1162.

[11] D. P. Kingma ve J. Ba, "Adam: A Method for Stochastic Optimization", *CoRR*, Ara. 2014, Erişim: 13 Ocak 2024. [Çevrimiçi]. Erişim adresi: https://www.semanticscholar.org/paper/Adam%3A-A-Method-for-Stochastic-Optimization-Kingma-Ba/a6cb366736791bcccc5c8639de5a8f9636bf87e8

[12] D. Bahdanau, K. Cho, ve Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", *CoRR*, Eyl. 2014, Erişim: 13 Ocak 2024. [Çevrimiçi]. Erişim adresi: https://www.semanticscholar.org/paper/Neural-Machine-Translation-by-Jointly-Learning-to-Bahdanau-Cho/fa72afa9b2cbc8f0d7b05d52548906610ffbb9c5

[13] A. Vaswani *vd.*, "Attention Is All You Need", 01 Ağustos 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

[14] I. Sutskever, O. Vinyals, ve Q. V. Le, "Sequence to Sequence Learning with Neural Networks", *ArXiv*, Eyl. 2014, Erişim: 13 Ocak 2024. [Çevrimiçi]. Erişim adresi: https://www.semanticscholar.org/paper/Sequence-to-Sequence-Learning-with-Neural-Networks-Sutskever-Vinyals/cea967b59209c6be22829699f05b8b1ac4dc092d

[15] O. Kaynar, Y. E. Işik, Y. Görmez, ve F. Demirkoparan, "Otomatik Metin Özetleme İçin Genetik Algoritma Tabanli Cümle Çikarimi", *Yönetim Bilişim Sistemleri Dergisi*, c. 3, sy 2, Art. sy 2, Ara. 2017.

[16] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, ve S. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding", içinde *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium: Association for Computational Linguistics, 2018, ss. 353-355. doi: 10.18653/v1/W18-5446.

[17] M. E. Peters vd., "Deep Contextualized Word Representations", içinde Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), M. Walker, H. Ji, ve A. Stent, Ed., New Orleans, Louisiana: Association for Computational Linguistics, Haz. 2018, ss. 2227-2237. doi: 10.18653/v1/N18-1202.

[18] J. Devlin, M.-W. Chang, K. Lee, ve K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 24 Mayıs 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[19] N. Reimers ve I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", içinde Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), K. Inui, J. Jiang, V. Ng, ve X. Wan, Ed., Hong Kong, China: Association for Computational Linguistics, Kas. 2019, ss. 3982-3992. doi: 10.18653/v1/D19-1410.

[20] M. Lewis *vd.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension", içinde *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, ve J. Tetreault, Ed., Online: Association for Computational Linguistics, Tem. 2020, ss. 7871-7880. doi: 10.18653/v1/2020.acl-main.703.

[21] T. B. Brown *vd.*, "Language Models are Few-Shot Learners", *ArXiv*, May. 2020, Erişim: 13 Ocak 2024. [Çevrimiçi]. Erişim adresi: https://www.semanticscholar.org/paper/Language-Models-are-Few-Shot-Learners-Brown-Mann/6b85b63579a916f705a8e10a49bd8d849d91b1fc

[22] L. Xue vd., "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer", içinde Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online: Association for Computational Linguistics, 2021, ss. 483-498. doi: 10.18653/v1/2021.naacl-main.41.

[23] C. Raffel *vd.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", 19 Eylül 2023, *arXiv*: arXiv:1910.10683. doi: 10.48550/arXiv.1910.10683.

[24] X. Liu, "Exploring Word Embeddings to Enhance Neural Machine Translation", Ph.D., Ann Arbor, United States, 2021. Erişim: 13 Ocak 2024. [Çevrimiçi]. Erişim adresi: https://www.proquest.com/docview/2601500412/abstract/881E8D8897F74E1EPQ/12

[25] J. L. Hauschild, "Examining the Effect of Word Embeddings and Preprocessing Methods on Fake News Detection", Ph.D., Ann Arbor, United States, 2023. Erişim: 13 Ocak 2024. [Çevrimiçi]. Erişim adresi: https://www.proquest.com/docview/2809436764/abstract/881E8D8897F74E1EPQ/3

[26] G. Alparslan ve M. Dursun, "Konvolüsyonel Sinir Ağları Tabanlı Türkçe Metin Sınıflandırma", *Bilişim Teknolojileri Dergisi*, c. 16, sy 1, Art. sy 1, Oca. 2023, doi: 10.17671/gazibtd.1165291.

[27] A. Karaca ve Ö. Aydin, "Transformatör mimarisi tabanlı derin öğrenme yöntemi ile Türkçe haber metinlerine başlık üretme", *GUMMFD*, c. 39, sy 1, Art. sy 1, Ağu. 2023, doi: 10.17341/gazimmfd.963240.

[28] Ç. Koçak ve T. Yiğit, "Gpt-3 Sınıflandırma Modeli İle Türkçe Twitlerin Siber Zorbalık Durumlarının Belirlenmesi", *GMBD*, c. 9, sy 4-ICAIAME 2023, Art. sy 4-ICAIAME 2023, Ara. 2023.fficient Estimation of Word Representations in Vector Space", 06 Eylül 2013, *arXiv*: arXiv:1301.3781. doi: 10.48550/arXiv.1301.3781.