Journal of Artificial Intelligence and Data Science (JAIDA)
e-ISSN: 2791-8335
https://dergipark.org.tr/pub/jaida

Vol. 3, No. 1, pp. 27-35, 2023
Research Article
Received: 04/05/2023 Accepted: 23/06/2023

Performance Evaluation of a Pretrained BERT Model for Automatic Text Classification

Sercan ÇEPNİ, Amine Gonca TOPRAK, Aslı YATKINOĞLU *, Öykü Berfin MERCAN, Şükrü OZAN

Abstract

This study presents a pre-trained BERT model application on texts that are extracted from website URLs automatically to classify texts according to the industry. With the aim of doing so, the related dataset is first obtained from different kinds of websites by web scraping. Then, the dataset is cleaned and labeled with the relevant industries among 42 different categories. The pre-trained BERT model which was trained on 101.000 advertisement texts in one of our previous ad text classification studies is used to classify texts. Classification performance metrics are then used to evaluate the pre-trained BERT model on the test set and 0.98 overall average accuracy and 0.67 average F1 score for 12 among 42 categories are obtained. The method can be used to test the compatibility of texts to be used in online advertising networks with the advertiser's industry. In this way, the suitability of the texts, which is an important component in determining the quality of online advertising, within the industry will be tested automatically.

Keywords: Ad text; bidirectional encoder representations from transformers; digital marketing; natural language processing; text classification.

1. Introduction

Digital marketing is the marketing of products or services using digital technologies, mainly on the Internet, but also including mobile phones, display advertising, and any other digital medium [1]. The term digital marketing has evolved over time from a specific term that describes the marketing of products and services using digital channels to a term that describes the process of using digital technologies to acquire customers and create customer preferences, promote brands, retain customers and increase sales. Accordingly, many types of research have begun to be conducted on effective advertising studies examining the relationship between investment and conversion in online advertising. Within the scope of these researches, machine learning algorithms and natural language processing (NLP) techniques are used to draw meaningful results from advertisement data such as written and even visual or audio content of advertisements, user movements, device information, location information, or demographic data [2]. Search network ads, which have been used for a long time, especially in search engines such as Google, are mostly displayed to users as text-based. The most important feature of a quality search ad work is that the content presented to the consumer is relevant and effective. Ad quality on online advertising platforms depends on many different factors, including how relevant the texts are to searches, the likelihood of consumers clicking the ad, and the experience on the landing page after the ad is clicked. Higher ad quality leads to better ad position and lower cost [3]. At this point, the preparation of advertising texts that respond to consumer needs and calls plays an important role in effectively transferring the product/service offered by the advertiser to potential customers. In this study, a text classification method BERT architecture [4] is proposed to evaluate which text context may be more relevant to a relevant industry. In the studies [4]-[5] studies were examined and it was observed that BERT architecture gave better results than models such as Word2vec and LSTM in similar text classification problems. For this reason, it was decided to use the BERT model in the study. It was also observed that the average accuracy rate of the classification process we performed with the BERT model which was trained on advertisement texts in our previous study for 42 categories [6], can successfully classify a large number of texts automatically generated from the URLs of the websites according to the sectors.

*Corresponding author

SERCAN ÇEPNİ; AdresGezgini Inc., R&D Department, Türkiye; e-mail: sercancepni@outlook.com; 0000-0002-3405-6059

AMİNE GONCA TOPRAK; AdresGezgini Inc., R&D Department, Türkiye; e-mail: goncatoprak@adresgezgini.com; 0000-0003-2425-5342

ASLI YATKINOĞLU; AdresGezgini Inc., R&D Department, Türkiye; e-mail: aslicankut@adresgezgini.com; 0009-0000-5702-1281

ÖYKÜ BERFIN MERCAN; AdresGezgini Inc., R&D Department, Türkiye; e-mail: boykumercan@gmail.com; 0000-0001-7727-0197

ŞÜKRÜ OZAN; AdresGezgini Inc., R&D Department, Türkiye; e-mail: sukruozan@gmail.com; 0000-0002-3227-348X

In the following sections of this article, the details of the studies we have carried out and the results obtained as a result of these studies are presented.

2. Related Work

Deep learning is a type of artificial neural network that uses sequential processing unit layers for feature extraction and each layer output feeds the next layer's input [7]. The theoretical foundations of deep learning, which has created an exciting new trend in machine learning in recent years, are well rooted in the classical neural network (NN) literature [8]. Thanks to the high capacity of neural networks to learn attributes from data, deep learning has been the subject of research in many fields besides NLP. One of the subjects of natural language processing, which has been studied for a long time and has important application areas, is text classification. Deep learning methods are also used for text classification and there are various current studies on this subject.

The use of deep learning representations for document classification has become increasingly popular in natural language processing (NLP) tasks. Statistical methods have been quickly replaced by deep learning as the state-of-the-art for many text classification tasks. Over the years, the focus has shifted from word representation generation (e.g. Word2Vec [9] or GloVe [10]) to generating embeddings of sentences or texts. Many architectures have been explored for this task, but the most commonly used are those based on the Transformers architecture [11]. This architecture provides self-supervised training and variable-length input.

BERT (Bidirectional Encoder Representations from Transformers) is a sentence encoder model based on the Transformers architecture introduced in 2018 [12]. It can be considered a state-of-the-art embedding model [13]. BERT models typically consist of interconnected encoder and decoder layers, and include a self-attention layer, a feed-forward pass-through layer, and a redundant hopping link. Many variants of BERT have been introduced since 2018, such as ALBERT [14], a lightweight version of BERT for smaller memory consumption and faster training, and DistilBERT [15], also smaller and faster, but pre-trained with information distillation.

An interesting extension is SBERT (SentenceBERT) [16], where a BERT model is fine-tuned with a Siamese or triplet architecture. This model is computationally efficient and produces embeddings that can be compared using cosine similarity, which reflects semantic meaning. An important extension is RoBERTa [17], which is also available because it is a multilingual model that can perform well in low-resource languages.

A sample study proposes a new approach for sequential short-text classification that combines recurrent neural networks (RNNs) and convolutional neural networks (CNNs) [18]. The proposed model in the study first uses an RNN to learn long-range dependencies between words in a short text. The output of the RNN is then fed into a CNN, which learns local patterns in the text. The CNN's output is then used to classify the short text. Another study proposes a new approach for sentence classification with CNNs by learning local patterns in the text [19]. They achieved state-of-the-art results on all four different datasets. In this study, they aimed to classify customer reviews as negative or positive. In another study, they used the BERT model for conflict event annotation [20]. The authors collected a dataset of news articles from the Global Terrorism Database (GTD) and extracted features from the text of the news articles. The features included the presence of certain keywords, the sentiment of the text, and the topic of the text. These texts then classified into different categories, such as conflict event, non-conflict event, and non-event. They achieved an accuracy of 88.7%. In another study, comparative analyzes of BERT model and traditional natural language processing approaches were carried out on data sets such as messages shared on social media, movie-series reviews, news content, and the results were presented. It has been proven by empirical values that BERT gives better results in different areas than traditional NLP approaches [21].

Many papers have been published and models exist for English texts, but the topic of NLP in other languages is limited but has been gaining momentum in recent years [22]. There were such studies that compared the BERT model with other models and in one such, the authors evaluated the BERT and DistilBERT model performances on collected two datasets of text from social media in the Filipino language [23]. According to the study, authors evaluated the two models on a held-out test set. The evaluation results showed that the BERT model achieved an accuracy of 80.2%, and the DistilBERT model achieved an accuracy of 79.5%.

In the study that this study was inspired from, the authors collected a dataset of ad text from Google Ads. The dataset consists of over 1 million ad texts [24]. The ad texts then preprocessed by removing stop words and punctuation marks. They also converted the ad text to lower case. They used the fine-tuned BERT model to classify ad text into different categories, such as product, service, and event. Then they fine-tuned the BERT model on the preprocessed ad text. The fine-tuning process was done using the Adam optimizer and the cross-entropy loss function and the fine-tuned BERT model achieved an accuracy of 94.5%. In other similar study, the authors was develop a new approach for ad text classification using the BERT model and achieved achieve an accuracy of 90.2% on a relatively small dataset [25].

Thanks to existing studies in the literature, it is proven that the BERT model achieves a significant accuracy for both text and ad text classification. This study aims to contribute to the state-of-the-art by developing an automated text classification system for the digital marketing field.

Table 1. *Data of Randomly Selected Categories Examples*

ID	Categories	URL
0	Art Education	https://www.opus-muzik.net/ https://www.akademipendik.com/
1	Cosmetics	https://www.espadekozmetik.com/ https://cosming.com/
2	Flower Order	https://yasmincicek.com/ https://antalyacicekcilik.com.tr/
3	Foreign Language Education	https://www.akademikdilkursu.com/ https://globaldilakademisi.com.tr/
4	Machine	https://adilmakina.com.tr/ https://mstendustriyelmakina.com/
5	Occupational Retraining	https://www.onlinedersmerkezi.com/ https://akademizeka.com/
6	Office Furnitures	https://theoffice.com.tr/ https://bahcesehirmobilyacarsisi.com/
7	Packaging	https://oggetti.com.tr/ https://patpat.com.tr/
8	Promotional Items	https://kcpromosyon.com/ https://www.mercanpromosyon.com/
9	Psychological Counseling	https://www.cocukpsikiyatrisi.com/ https://izmirpsikolojikdanismanlik.com/
10	Rent a Car	https://www.dervishotokiralama.com/ https://www.modernotokiralama.com/
11	Software Services	https://www.surrealyazilim.com/ https://www.akinsoftistanbul.org/
12	Technical Service	https://kombitamirci.net/ https://elektrotamir.com/
. .	. .	. .
42	Investment	https://www.dijitalpara.com.tr/ https://mbsyatirim.com/

3. Materials and Method

3.1 Dataset

For the study, the dataset is obtained from various websites by web scraping and contains approximately 101.000 data that belong to 42 different types of industries. In order to perform web scraping, AdresGezgini's database is used to select various URLs and manually labeled them according to the addressed industry itself to

create ground truth labels and some of the URLs with corresponding industries are given in Table 1 in order to understand the dataset better. After gathering different URLs from the database, texts are obtained by web scraping. Since we have the obtained texts and their ground truth labels in the dataset, the next step would be data preprocessing texts for the pre-trained BERT model. To do so, several data preprocessing methods are performed (Figure 1). In the data preprocessing stage, the dataset is first cleared of repetitive data. Punctuation and symbols are omitted, both the leading and the trailing characters are removed and all texts are lowercased. Because the texts are obtained from random URLs, there is of course an imbalance of data for each category namely industry. Since each category has different numbers of texts and those texts have different numbers of words, the outputs of the categorical probabilistic classification method taken from the BERT model trained with this imbalanced dataset were analyzed and it was observed that it showed a superior success.

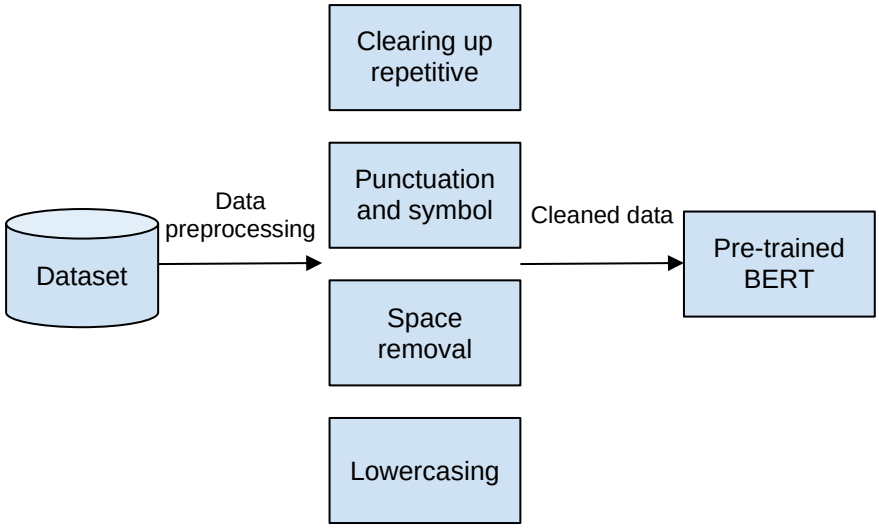

Figure 1. Data Preprocessing Steps

3.2 Method

3.2.1 Classification with BERT Model

The BERT model [26] is a transformer model introduced by Google in 2018, designed to pre-train bidirectional representations from unlabeled text and then fine-tune it using labeled text for different NLP tasks. With BERT, the translation success of the Google Translate application has also increased significantly, and now this application can translate even very long sentences between different languages with great accuracy without loss of meaning. Transducers have a structure that works with the self-attention mechanism formed by the tail structure (Figure 2).

The BERT model maps a query and a set of key-value pairs to an output. Here, vectors are formed that can express the correlation between the query, keys, values and the output itself. The output is calculated by the weighted sum of the values and the weight assigned to a value, with the rate of agreement with the key corresponding to the query [28]. The BERT model processes a text both from right to left and from left to right, so it can learn the relationships between elements in the text. In the training phase, MLM (Masked Language Modeling) and NSP (Next Sentence Prediction) techniques are used. In MLM technique, masked words are tried to be predicted using open (unmasked) words. With this technique, analysis and estimation are made on the words in the sentence. In the NSP technique, the relationship of the sentences with each other is examined. The relationship of a sentence with the sentence that follows it is examined. Structures built with the BERT model require a pre-trained model. For this reason, in our study, the BERT-BASE-TURKISH-UNCASED [29] model, which is a pre-trained BERT model for Turkish language by the Loodos team, using the 200GB dataset obtained from sources such as e-books, news articles, online blog posts and wikipedia has been preferred. In this study, the pretrained BERT model is fine-tuned and is used as a language model that can be used to classify text into different categories. In this study, the BERT model was used to classify text that was extracted from website URLs. The

text was first extracted from the URLs using a web scraping tool. The text was then cleaned and labeled with the relevant industry among 42 different categories. During the fine-tuning, early stopping criteria is employed to mitigate overfitting and promote generalization. Therefore, during fine-tuning, no over-fitting situation is observed. The BERT model was then used to classify the text. The classification performance of the BERT model was evaluated on a test set.

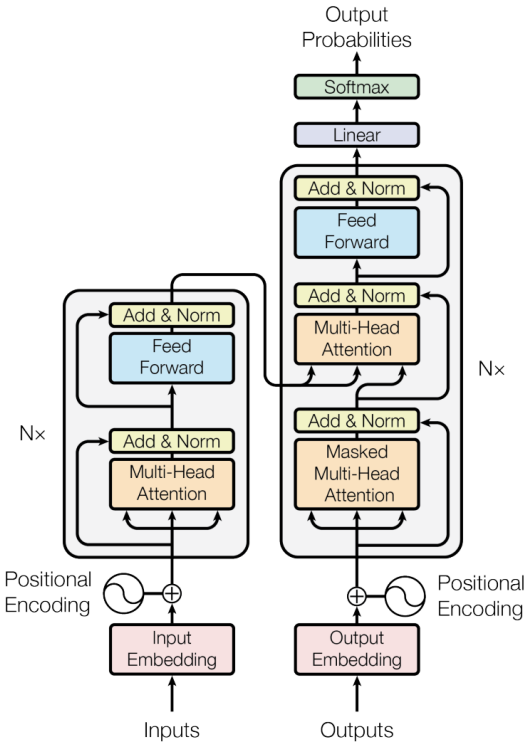

Figure 2. BERT Model Architecture [27]

The BERT model is applied to the problem of industry classification in the following steps:

- The text is extracted from website URLs using a web scraping tool.
- The text is cleaned and labeled with the relevant industry among 42 different categories.
- The BERT model is used to classify the text.
- The classification performance of the BERT model is evaluated on a test set.

The BERT model is able to achieve high accuracy for industry classification because it is pre-trained on a large dataset of text and code. The BERT model is also able to understand the context of a text, which is important for industry classification.

The method presented in this study can be used to test the compatibility of texts to be used in online advertising networks with the advertiser's industry. In this way, the suitability of the texts, which is an important component in determining the quality of online advertising, within the industry will be tested automatically.

3.2.2 Model Performance Evaluation

In order to accurately evaluate the performance of the model, the F1 score was followed along with the accuracy value. True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) values are used to calculate the F1 score. TP is the positive result of both the model's estimated and the true value, TN the negative results of both the model's prediction and the true value, FP the negative result when the model prediction is positive, and the FN the positive result when the model's prediction is negative, and the true value. In this case, TP and TN are considered correct results, FP and FN are incorrect results.

While calculating the accuracy value, it is calculated by the ratio of the TP and TN values that the model predicts correctly to all the predicted TP, TN, FP, FN values (1). The precision value is the ratio of the number of TP values predicted by the model to the number of TP and FP values, which are all positive results produced by

the model (2). The recall value can be found by the ratio of the number of TP values predicted by the number of TP and FN, which are all positive results that the model should produce (3). F1 score can be defined as the harmonic mean of precision and recall values (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\ Score = 2\ x\ (\frac{Precision\ x\ Recall}{Precision + Recall}) \quad (4)$$

4. Results

The aim of this study was to evaluate the performance of a pretrained BERT model for the task of classifying texts extracted from various types of URLs. A total of 101.000 samples were used to train and evaluate the model, with 840 samples of the data set aside for testing.

The pre-training parameters of the model used are given in Table 2. This model has been fine-tuned and made ready for use. The pretrained BERT model was trained for 20 iterations (epochs) using a batch size of 128.For the training of the pre-trained BERT model in Turkish language we used in our study, the dataset consisting of advertisement texts was divided into predetermined categories, 70% of the dataset was used as the training dataset and 30% was used as the test dataset to test the model. Since the training-test separation was made categorically, training and testing were carried out with equal amounts of data from each category.

Table 2. *Parameters of the Pretrained BERT Model*

Model	Hidden Size	Max Sequence Length	Attention Heads	# of Hidden Layers	Architecture	Vocabulary Size
(Loodos 2020)	768	512	12	12	BERT For Masked LM	32000

Then, after the model was trained in 10 iterations with 14,349 training data and 3,588 test data, the analyzes were performed on the results of the 3rd iteration, since there was no increase in accuracy, precision, sensitivity and F-1 values after the 3rd iteration of the model. Here, different training-test set partitions were compared and it was aimed to produce results in accordance with the 70%-30% ratio used in similar studies in the literature.

It was seen that the lowest value among the precision values of the categories in the test data set belonged to the 5th category, which contains 25 test data. In the sensitivity analysis, the lowest value was found in the class containing 97 test data belonging to the 1st category. For the F-1 score, the 5th category and the minimum value were compared. It is seen that the lowest successful categories of the model consist of the data with the lowest rate in the training and test data set. Based on this situation, it is seen that an accuracy of over 90% can be obtained in cases where the training and test data, which is ideal for similar studies, are more than 1000 on a category basis. Table 3 shows the classification report of the 3rd iteration. As can be seen here, the model reaches a high training accuracy from the first iteration, but can reach a high test accuracy after the third iteration.

The results showed that the pretrained BERT model achieved a high average accuracy of 0.98, indicating that it was effective in accurately classifying the majority of the samples. The accuracy was computed as the ratio of correctly classified samples to the total number of samples. The high accuracy achieved by the model suggests that it can be useful for identifying texts from different types of URLs.

However, the overall mean precision and recall values were relatively low at 0.33 and 0.28, respectively. On the other hand, it is observed that F1-score is calculated relatively higher for 12 categories among 42 categories and these categories are given in Table 4. The average F1-score is calculated as 0.67 for these 12 categories. The average F1-score is calculated as 0.67 for those categories. The reason behind this situation is in the previous study, the BERT model is trained with advertisement texts rather than randomly scraped texts from websites.

Precision measures the proportion of correctly classified positive samples out of all samples classified as positive, while recall measures the proportion of correctly classified positive samples out of all actual positive samples. The low precision and recall metrics suggest that the model had limitations in correctly identifying all positive samples. Specifically, the model had a tendency to incorrectly classify negative samples as positive, resulting in a low precision value. Additionally, the model was not able to correctly identify all actual positive samples, resulting in a low recall value.

Table 3. *Iteration Test Report for Pretrained BERT Model*

ID	Precision	Sensitivity	F1 Score	Ad Texts
0	0.90	0.95	0.92	353
1	0.99	0.77	0.87	97
2	0.92	0.92	0.92	414
3	0.93	0.87	0.90	378
4	0.88	0.87	0.87	187
5	0.78	0.84	0.81	25
6	0.95	0.94	0.90	474
7	0.87	0.93	0.91	560
8	0.92	0.90	0.94	444
9	0.91	0.98	0.94	168
10	0.94	0.93	0.94	363
11	0.87	0.82	0.84	125
Accuracy			**0.91**	**3588**
Average	**0.91**	**0.91**	**0.91**	**3588**

Table 4. *Test Results Higher Than %45 F1 Score*

Category Names	F1 Score
Art Education	0.77
Energy	0.48
Flower Order	0.90
Foreign Language Education	0.81
Occupational Retraining	0.51
Office Furnitures	0.59
Packaging	0.82
Promotional Items	0.71
Psychological Counseling	0.65
Rent a Car	0.78
Technical Service	0.50
Investment	0.47
Average	**0.67**

The average F1-score, which is the weighted harmonic mean of precision and recall, was also relatively low at 0.28. This further highlights the limitations of the model's performance in terms of correctly identifying positive

samples. However, the average F1-score for certain 12 categories that are mentioned in Table 1 is 0.67 which indicates that the pre-trained BERT model is still successful to classify texts.

In summary, the pretrained BERT model demonstrated a high accuracy in classifying texts extracted from different types of URLs. However, the model's overall performance in terms of precision and recall was limited except for specific categories, indicating the potential for further improvement. Future work could focus on fine-tuning the model on specific data or incorporating additional features to enhance its performance. Future work also might be focused on using a transfer learning method rather than fine-tuning an existing model. The results demonstrate the potential of using pretrained BERT models for classification tasks while highlighting the importance of evaluating the model's precision and recall metrics in addition to its accuracy.

5. Conclusions

The aim of this study was to evaluate the BERT model performance that was trained on advertisement texts. Based on the results obtained from the pre-trained BERT model, the model shows a high average accuracy of 0.98, which suggests that it is effective in classifying texts extracted from different types of URLs. However, the relatively low average precision and recall values of 0.33 and 0.28, respectively, indicate that the model has limitations in correctly identifying all positive samples. However, for the 12 categories given in Table 1, the pre-trained BERT model gives an average 0.67 F1-score. The reason why the pre-trained BERT model was not able to classify texts with higher F1-score is the insufficiency of the dataset. With a higher number of data, the model's performance would be better. And, the model was also trained on advertisement texts but tested on scraped texts from different websites. These limitations suggest that further improvements could be made to enhance the model's precision and recall metrics, potentially by incorporating additional features or fine-tuning the model on specific data. Overall, these findings demonstrate the potential of using pre-trained BERT models for classification tasks, while highlighting the importance of evaluating the model's precision and recall in addition to its accuracy. Future work will be focused on to trained the BERT model with data that are obtained from websites and evaluating its performance with more samples.

Declaration of Interest

The authors declare that there is no conflict of interest.

Author Contributions

Aslı Yatkınoğlu: Writing - review & editing, Collecting the data, Results and discussions. Amine Gonca Toprak: Writing - review & editing, Validation, Training, Results and discussions. Sercan Çepni: Software, Methodology, Formal analysis, Collecting the data. Öykü Berfin Mercan: Methodology, Validation, Training, Results and discussions. Şükrü Ozan: Supervision, Conceptualization, Methodology.

References

[1] Desai, Vaibhava, and B. Vidyapeeth. "Digital marketing: A review." International Journal of Trend in Scientific Research and Development 5.5 (2019): 196-200.

[2] Z, A. and Adali, E., "Opinion mining and sentiment analysis for contex-tual online-advertisement," in 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT). IEEE, 2016, pp. 1–3.

[3] "Reklam kalitesi hakkında - Google Ads Yardım," https://support.google. com/google-ads/answer/156066?hl=tr&ref topic=10549746, May 2021.

[4] Ş. Ozan and D. E. Taşar, "Auto-tagging of Short Conversational Sentences using Natural Language Processing Methods," 2021 29th Signal Processing and Communications Applications Conference (SIU), 2021, pp. 1-4, doi: 10.1109/SIU53274.2021.9477994

[5] Rønningstad, E., "Targeted sentiment analysis for norwegian text," 2020.

[6] Özdil, U., Arslan, B., Taşar, D. E., Polat, G., & Ozan, Ş. (2021, September). Ad Text Classification with Bidirectional Encoder Representations. In 2021 6th International Conference on Computer Science and Engineering (UBMK) (pp. 169-173). IEEE.

[7] Deng, L., & Yu, D. (2014). Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing (Cilt 7, s. 197-387).

[8] Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2017). "Deep learning for health informatics". IEEE journal of Biomedical and Health Informatics 21(1), 4-21.

[9] Mikolov, C. (2013). Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space, CoRR.

[10] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[13] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning--based text classification: a comprehensive review. ACM computing surveys (CSUR), 54(3), 1-40.

[14] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

[15] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[16] Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

[17] Liu, Y., Ott, M., & Goyal, N. (2019). Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. Computing Research Repository.

[18] Lee, J. Y., & Dernoncourt, F. (2016). "Sequential short-text classification with recurrent and convolutional neural networks". arXiv preprint arXiv:1603.03827.

[19] Kim, Y. (2014). "Convolutional neural networks for sentence classification". arXiv preprint arXiv:1408.5882.

[20] Olsson, F., Sahlgren, M., Abdesslem, F. B., Ekgren, A., & Eck, K. (2020, May). Text categorization for conflict event annotation. In Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020 (pp. 19-25).

[21] Gonzalez-Carvajal, S. and Garrido-Merchan, E. C., "Comparing bert against traditional machine learning text classification," arXiv preprint arXiv:2005.13012, 2020

[22] Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. arXiv preprint arXiv:2004.09813.

[23] Cruz, J. C. B., & Cheng, C. (2020). Establishing baselines for text classification in low-resource languages. arXiv preprint arXiv:2005.02068.

[24] Şükrü, O. Z. A. N., et al. "BERT Modeli'nin Sınıflandırma Doğruluğunun Sıfır-Atış Öğrenmesi ile Artırılması." Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi 14.2 (2021): 99-108.

[25] Özdil, U., Arslan, B., Taşar, D. E., Polat, G., & Ozan, Ş. (2021). Ad Text Classification with transformer-based natural language processing methods. arXiv preprint arXiv:2106.10899.

[26] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., "Bert: Pre-trainingof deep bidirectional transformers for language understanding," arXivpreprint arXiv:1810.04805, 2018.

[27] Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021, June). Attention is all you need in speech separation. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 21-25). IEEE.

[28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez,A. N., Kaiser, L., and Polosukhin, I., "Attention is all you need," arXivpreprint arXiv:1706.03762, 2017.

[29] Loodos., "loodos/bert-base-turkish-uncased · hugging face," https://github.com/Loodos/turkish-languagemodels, Aug. 2020.