Topic Classification of Text-Based Lesson Questions in Turkish with BERTurk

Ayşegül Albayrak Doğan[1(✉)], Ahmet Sayar[1], and İlker Çetiner[2]

[1] Kocaeli University, 41380 Izmit, Kocaeli, Turkey
aysegulalbayrak1@gmail.com, ahmet.sayar@kocaeli.edu.tr
[2] ArgeLabs Information Technologies, 41275 Yeniköy, Başiskele, Kocaeli, Turkey
ilker.cetiner@argelabs.com.tr

Abstract. With the corona virus pandemic, social contact has been kept to a minimum, and education in schools can be carried out remotely. As a result of this, the concept of distance education has gained importance. In this study, natural language processing (NLP) and its effects on distance education are discussed, and by using NLP, a topic classification system is proposed. Classification is applied to text-based lesson questions in Turkish language. In this way, the questions asked by the students can be quickly directed to the teachers in the relevant specialty through a system to be designed, and the processes can be accelerated. In the data preparation phase, the real-world lesson questions were collected and converted from image to text using the EasyOCR library, and topic classification was performed on the data set using the Berturk model. Since the image-to-text method was used in the data set preparation phase, we encountered some noise in the data. To clean the data, different data preprocessing and cleaning techniques are applied. Finally, the training has been performed, and accuracy rates are presented.

Keywords: Natural Language Processing · Intention Classification · Data Preprocessing · Topic Classification · Bert Model · NLP

1 Introduction

In changing and developing world conditions; especially due to effects of the Covid-19 pandemic and the effects of the earthquake disaster centered in Kahramanmaraş, which occurred in Turkey on February 6, 2023, a decision was made to provide distance education until the major impact of the earthquake was overcome in order to continue educational activities. It has been experienced that the effects of distance education, which gained importance with the pandemic conditions and the effects of natural disasters experienced afterwards, continue. In such situations of necessity, it is inevitable to ensure sustainability and adapt quickly to changing conditions [3]. As a result of the aforementioned conditions, the world of distance education, which is defined as time and space independent learning, continues to occupy the agendas of teachers, students and families [12].

S. Kadry and R. Prasath (Eds.): MIKE 2023, LNAI 13924, pp. 87–94, 2023.
https://doi.org/10.1007/978-3-031-44084-7_9

NLP, which is becoming more and more important in the age of the digitalizing world and has an impact in almost every field, can be used in many areas such as social media, banking transactions, appointment systems, chat applications, question and answer systems. In the world of education, it is widely integrated in many areas such as research, science, linguistics, e-learning, assessment systems and contributes to achieving positive results in other educational environments such as schools, higher education system and universities [1]. NLP develops methods that produce solutions by understanding human needs in different fields. One of these methods, BERT that was developed by Google in November 2018 to better understand human language, has remained popular in the field of natural language processing in recent years.

In this study, the factors observed in the process of automatically labeling which course class the questions asked in dialog systems, which will provide fast interaction in the field of education, are evaluated. The accuracy rates of the classification process using the BERTurk language model were observed on 4027 data belonging to History, Turkish, Philosophy, Geography, Biology, Mathematics, Physics, Chemistry, History, Turkish, Philosophy, Geography, Biology, Mathematics, Physics, Chemistry courses among the high school courses in Turkish language and in the National Education curriculum of the Republic of Turkey [15]. The data used in the training consists of a pool of data converted from image to text via EasyOCR library and course labeled during the conversion. It was observed some noisy data such as underscores (_) and hyphens (-) occurred in the image-to-text data depending on the quality. In this context, in order to examine the effect of the detected noisy data and the preprocessing work to be done on the questions on the classification of the data, different combinations were made and the training were repeated with the same hyper-parameters and the accuracy rates were examined. The same number and the same training set were used in each training. In the training, 20% test data was used. After the training, it was observed that the performance rates did not differ significantly even if cleaning operations were performed. Despite noisy data, successful classification was achieved at a rate of 0.97.

Related studies are presented in the second part to the general one in our study. The work done in the third section is summarized. Tests and evaluation studies were carried out in the fourth section. The fifth section contains the results and comments.

2 Relatedwork

Classification techniques have been used in many different scientific and application domains. Those classification studies can be grouped into three categories. The first group is applied to image and video data; the second group is a hybrid classification applied to text and image data together; and the third group's classification techniques are applied merely to text-based data.

The samples of image and video data classification are given as follows: [7] study the performance of the vehicle classification algorithms by using deep

learning algorithms on video streams. Topçu et al. [17] have applied the classification to remote satellite image data by using capsule networks. Some of them are based on image data, and others are based on text data. Şentaş et al. [13] studied the performance of Support Vector Machine (SVM) and Convolutional Neural Network Algorithms (CNN) for real-time vehicle type classification. Tasiev et al. [16] presented a real-time vehicle type classification using CNN. The hybrid classifications are applied to text and image data together. Omurca et al. [11] proposed a document image classification system by fusing deep and machine learning models. Sevim et al. [14] studied multi-class document image classification by using deep visual and textual features. Yurtsever et al. [18] worked on a search technique enabling figure search by text in large scale digital document collections.

In this study, we focus on the third group of classification techniques, which are applied merely to text-based data. The topic classification problem is a Natural Language Processing (NLP) problem that is often studied in the literature. BERT, who has created a solution to this problem, has given strong results in most of the studies conducted in the field of subject classification in many languages. In this section, studies that use BERT modeling in the field of subject classification and evaluate the effect of preprocessing stages on performance rates are examined. The sources examined in the literature are usually in foreign languages and the number of studies on Turkish is in the minority. In our study, the studies that can guide the effect of the data consisting of a set of Turkish questions on the performance rate of the data preprocessing stage when using the BERT model are shared below in chronological order.

Hazrati et al. [5], in their study to detect irony in texts published on social media, stated that non-standard expressions are generally included in the texts used by social media users. The BERT model was trained before and after these detected expressions were cleaned and the success rates were shared. They stated that data cleaning has a great impact on the success rate in sentiment analysis. After the data cleaning process we performed on the Turkish dataset, it was observed that there were not very big differences in the success rates. In this context, it is seen that the data set in different fields has an effect on the success rate on BERT.

Kurniasih and Manik [8] investigated the effects of training with and without data preprocessing on deep learning. In this study, the effect of the processes such as correction of abbreviated words, removal of repeated syllables, removal of hashtags on the accuracy performance rate of the Bert model was examined.

Another study examining the effect of systematic practical perturbations on the performance of deep learning-based text classification models such as CNN, LSTM and BERT was conducted by Miyajiwala et al. [10]. In the study where punctuation marks, stop words, ineffective words, and unwanted words are expressed as perturbations, it is stated that BERT is a more sensitive model than other models when perturbations are added or removed.

Bayrak and Issifu [4] worked on two main tasks in their study. The first one is dialect recognition and the second one is sentiment analysis with BERT

model on tweet data and they shared their classification results. In the data preprocessing stages, they applied cleaning steps such as Html tags, URLs, leaving spaces after Arabic numbers, removing Arabic-specific accent marks. They showed that preprocessing has a positive effect on Dialect Recognition unlike Sentiment Analysis.

In their study, Zhu et al. [19] showed that for text classification tasks with modern NLP models such as BERT, methods applied over various types of noise do not always improve performance and may even degrade it. For different types of noise in the dataset, they show that BERT is robust to injected noise, but not necessarily under weak supervision noise. In our study, it was observed that the removal of ineffective words relatively decreased the performance rate.

In this study for author profile detection, Alzahrani and Jololian [2] discussed the effects of data preprocessing techniques before Bert model training. In the preprocessing stages, training was carried out after the removal of stop words, reteweet tags, hashtags, mentions, and urls, and the success rates obtained were shared. It was observed that the highest performance rate was obtained on data trained without any data preprocessing.

In this study by Jiang et al. [6], a pre-trained and fine-tuned BERT model was used to classify OCR translated texts into book excerpts according to subject areas. As a result of the classification, the effect of OCR noise on the performance rate was analyzed.

Maharani [9] stated that data quality greatly affects the classification performance in classification using the Bert model. In this study, as a result of the study conducted to classify tweets related to emergencies, it was emphasized that data quality should be improved in order to avoid misclassification and it was stated that this issue will be studied in future studies.

3 Architecture

The general flow we followed during our study to conduct the training and evaluate the results is shown in Fig. 1. During the creation of the labeled data, the question set was converted from image to text using EasyOcr and a total of 4027 data were obtained. The dataset was split into training and test data by 20% before training. The same number of questions was used in all experiments. In the data preprocessing stages, the preprocessing method was presented with different combinations depending on the data set content and training was performed. The training was named according to the order in which the training was performed. For the training, library management was provided by installing Anaconda on a computer with GeForce RTX 3080 Nvidia graphics card. Hyperparameters used for each training are available in Table 1. Training was performed on a preprocessed dataset and took 4 min on average. Six different training trials were conducted, are listed below. The data sample used in the trainings can be seen in Table 2.

In the first training, "A), B), C), D), E), A) Yalnız/Yalniz, B) Yalnız/Yalniz, C) Yalnız/Yalniz, D) Yalnız/Yalniz, E) Yalnız/Yalniz" answer choices were removed from the sentence and the training was realized.

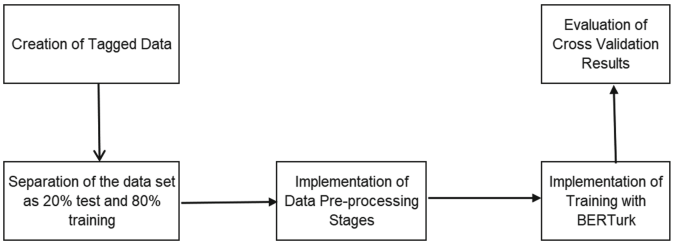

Fig. 1. Training Model Flowchart

Table 1. Model Parameters

Parameter	Value
Hugging Face Model	dbmdz/bert-base-turkish-uncased
Use Early Stopping	True
Early Stopping Delta	0.01
Early Stopping Metric	mcc
Early Stopping Metric Minimize	False
Early Stopping Patience	5
Evaluate During Training Steps	6000
fp16	False

In the second training, the Natural Language Toolkit (NLTK) library was used to remove stopwords from the dataset and training was performed.

In the third training, the training was performed by removing only the options starting with "Yalnız/Yalniz" from the sentence in the data set.

In the fourth training, the Stopwords belonging to the NLTK library on the dataset and the choices starting with "Yalnız/ Yalniz" were removed and the training was performed.

In the fifth training, no cleaning was performed on the data set.

The sixth training, words were merged by removing hyphens (-) and underscores (_) from the data set. The method followed during word merging is as follows:

The hyphen represents subtraction in mathematical formulas, so there are rules to be considered during word merging. For example, since the hyphen in the formula "x = a − b" should not be removed and proceeded as an expression in the form of "ab", the merging method here is based on the rule of Turkish grammar that the word at the end of the line is divided while the word at the end of the line is divided, and no single letter is left at the end of the line and at the beginning of the line. In addition, it was checked that the word following the line does not correspond to a variable in the formula in Mathematics. For formula variables, a fixed list was prepared according to the content of the data set: ab, ba, abc, acb, bac, bca, cab, cba, a, b, c. If the word checked before and

Table 2. Data Sample

Lesson	Question	Question After Pre-Processing
Matematik	a < b < 0 < C < d, olduguna göre, asagidakilerden hangisi sifir olabi- lir?, A) a + c = 3d, B) a − b = c, C) a − b + c, D) d − c = b, E) a + 2b − d	a < b < 0 < C < d, olduguna göre, asagidakilerden hangisi sifir olabilir?, A) a + c = 3d, B) a − b = c, C) a − b + c, D) d − c = b, E) a + 2b − d
Fizik	Elektronun karsit parçacigi pozitron için; I Kütlesi elektronunki ile aynidir. II. Yük miktari elektronunki ile aynidir. III. Yükünün isareti elektroninki ile aynidir. yargilarindan hangileri doğrudur? 1, A) Yalniz I, B) Yalniz II D) ve III E) II ve III, C) I ve II	Elektronun karsit parçacigi pozitron için; I Kütlesi elektronunki ile aynidir. II. Yük miktari elektronunki ile aynidir. III. Yükünün isareti elektroninki ile aynidir. yargilarindan hangileri doğrudur?

after the hyphen in the word to be merged was present in the above list, the word was not merged.

4 Tests and Evaluation

The success rates of the experiments conducted within the scope of the study are available in Table 3. According to the data preprocessing stages performed on the data, the steps are indicated in the columns according to the trainings. The number of test data used for success rate calculation and the number of incorrectly predicted questions are shared. In the validation process performed on the data model obtained after the training, the course distributions of the incorrectly predicted questions are in Table 4.

Table 3. Training Cross Validation Results

Training	Accuracy	Test Question Count	False Prediction Count	Remove Stopwords	Remove "-, _"	Remove Start With "Yalnız"	Remove All Answers Choices
First	0.985472	826	12				✓
Second	0.984261	826	13	✓			
Third	0.979418	826	17			✓	
Fourth	0.978208	826	18	✓		✓	
Fifth	0.978208	826	18				
Sixth	0.970944	826	24		✓		

Table 4. Number of False Predicted Questions by Lessons

Training	Biyoloji	Coğrafya	Felsefe	Fizik	Kimya	Matematik	Tarih	Türkçe	TOTAL
First	1	3	2	3	0	0	1	2	12
Second	3	3	1	3	0	0	1	2	13
Third	3	2	1	5	1	0	1	4	17
Fourth	3	3	3	3	2	0	1	3	18
Fifth	7	2	0	6	2	0	1	0	18
Sixth	7	4	2	5	0	0	1	5	24

5 Conclusions

We investigated the performance rates of the pre-trained and fine-tuned BERTurk model on data converted from image to text using EasyOCR, and the effect of data preprocessing on classification by course categories on a dataset of high school questions. Our analysis shows that the pre-trained and fine-tuned BERTurk language model achieves good performance rates despite the OCR noise. The most successful result was obtained in the first training as 0.98. When the results are analyzed, it can be seen from Table 4 that the highest number of incorrect predictions was in the Physics course and there was no misclassification in the Mathematics course.

In the scope of our study the training accuracy performance rates did not show large differences in the data preprocessing steps. In particular, it was observed that the incorrect predictions were obtained based on similar question contents. Therefore, the number of incorrect predictions can be improved in future studies by providing diversity over the training data set. Different combinations can be added to the training steps by diversifying the data preprocessing stages. For example, although the word “ka- lem” was corrected as “kalem”, no processing was performed on the word “ka - lem”. Therefore, these words were included separately in the training. The training can be performed again by performing the cleaning process for these words.

References

1. Alhawiti, K.M.: Natural language processing and its use in education. Int. J. Adv. Comput. Sci. Appl. **5**(12) (2014)
2. Alzahrani, E., Jololian, L.: How different text-preprocessing techniques using the BERT model affect the gender profiling of authors. arXiv preprint arXiv:2109.13890 (2021)
3. Aras, K.S., Kocasaraç, H.: Eğitimin dijital boyutunda öğrenme-öğretme araçları. Uluslararası Karamanoğlu Mehmetbey Eğitim Araştırmaları Dergisi **4**(2), 117–134 (2022)
4. Bayrak, G., Issifu, A.M.: Domain-adapted BERT-based models for nuanced Arabic dialect identification and tweet sentiment analysis. In: Proceedings of the the Seventh Arabic Natural Language Processing Workshop (WANLP), pp. 425–430 (2022)

5. Hazrati, L., Sokhandan, A., Farzinvash, L.: Profiling irony speech spreaders on social networks using deep cleaning and BERT. In: CLEF, pp. 1613–0073 (2022)
6. Jiang, M., Hu, Y., Worthey, G., Dubnicek, R.C., Underwood, T., Downie, J.S.: Impact of OCR quality on BERT embeddings in the domain classification of book excerpts. In: CHR, pp. 266–279 (2021)
7. Kul, S., Eken, S., Sayar, A.: Trafik gözetim videolarında araç sınıflandırma algoritmalarının etkinliğinin Ölçülmesi (2016)
8. Kurniasih, A., Manik, L.P.: On the role of text preprocessing in BERT embedding-based DNNs for classifying informal texts. Neuron **1024**(512), 256 (2022)
9. Maharani, W.: Sentiment analysis during Jakarta flood for emergency responses and situational awareness in disaster management using BERT. In: 2020 8th International Conference on Information and Communication Technology (ICoICT), pp. 1–5. IEEE (2020)
10. Miyajiwala, A., Ladkat, A., Jagadale, S., Joshi, R.: On sensitivity of deep learning based text classification algorithms to practical input perturbations. In: Arai, K. (ed.) SAI 2022. LNNS, vol. 507, pp. 613–626. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-10464-0_42
11. Omurca, S.İ, Ekinci, E., Sevim, S., Edinç, E.B., Eken, S., Sayar, A.: A document image classification system fusing deep and machine learning models. Appl. Intell. **53**, 1–16 (2022)
12. Sayılır, K., Sarı, Y.P., Pepele, H.R., Yetkin, S.G.: Uzaktan eğitim faaliyetlerine ilişkin ortaokul öğrencilerinin görüşlerinin değerlendirilmesi. Ulusal Eğitim Dergisi **3**(2), 417–435 (2023)
13. Şentaş, A., et al.: Performance evaluation of support vector machine and convolutional neural network algorithms in real-time vehicle type and color classification. Evol. Intell. **13**, 83–91 (2020)
14. Sevim, S., et al.: Multi-class document image classification using deep visual and textual features. Int. J. Comput. Intell. Appl. **21**(02), 2250013 (2022)
15. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune BERT for text classification? In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) CCL 2019. LNCS (LNAI), vol. 11856, pp. 194–206. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32381-3_16
16. Tashiev, İ., et al.: Konvolüsyonel sinir ağı kullanarak gerçek zamanlı araç tipi sınıflandırması real-time vehicle type classification using convolutional neural network
17. Topçu, M., Dede, A., Eken, S., Sayar, A.: Multilabel remote sensing image classification with capsule networks. In: 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pp. 1–3. IEEE (2020)
18. Yurtsever, M.M.E., Özcan, M., Taruz, Z., Eken, S., Sayar, A.: Figure search by text in large scale digital document collections. Concurr. Comput.: Pract. Exp. **34**(1), e6529 (2022)
19. Zhu, D., Hedderich, M.A., Zhai, F., Adelani, D.I., Klakow, D.: Is BERT robust to label noise? A study on learning with noisy labels in text classification. arXiv preprint arXiv:2204.09371 (2022)