DEU FMD 27(80) (2025) 267-271

Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering
Elektronik/Online ISSN: 2547-958X

RESEARCH ARTICLE / ARAŞTIRMA MAKALESI

Occupation Prediction from Twitter Data

Twitter Verilerinden Meslek Tahmini

Tolga İzdaş *, Hikmet İskifoğlu, Banu Diri

Yıldız Technical University, Faculty of Electrical and Electronics Engineering, Department of Computer Engineering, İstanbul, TÜRKİYE
*Corresponding Author / Sorumlu Yazar **: tolgaizdas@gmail.com

Abstract

Today, the use of social media has become quite widespread. Among social media platforms, Twitter, now known as X, stands out with its number of users and abundance of data. This data can be used in many studies. In this study, it is aimed to predict occupation based on Turkish tweets. In the study, 5 datasets of different sizes were used. The tweets are evaluated and compared as single and pairwise. In the pre-processing step, different machine learning and deep learning methods and pre-trained models were tested using 2 different natural language processing libraries. Among the machine learning methods, the highest accuracy of 88% was obtained from the Logistic Regression model with pairwise tweet data, while the highest accuracy of 88% was obtained with the Multi-layer Perceptron from deep learning models. The BERT and "ytu-ce-cosmos/turkish-base-bert-uncased" developed by Yıldız Technical University COSMOS AI Research Team were used as pre-trained models. Although these models gave different results on different datasets, both of them achieved the highest success with a ratio of 89% on pairwise tweet data.

Keywords: *Twitter, occupation prediction, machine learning, deep learning, BERT, COSMOS*

Öz

Günümüzde sosyal medya kullanımı oldukça yaygınlaşmıştır. Sosyal medya platformları arasında artık X olarak bilinen Twitter, kullanıcı sayısı ve veri fazlalığı özellikleriyle öne çıkıyor. Bu veriler pek çok çalışmada kullanılmaya müsaittir. Bu çalışmada Türkçe tivitler üzerinden meslek tahmini yapılması hedeflenmektedir. Çalışmada farklı boyutlarda 5 adet veri seti kullanılmıştır. Tivitler tekli ve ikili olarak değerlendirilerek karşılaştırması yapılmıştır. Ön işleme adımında 2 farklı doğal dil işleme kütüphanesi kullanılarak farklı makine öğrenmesi ve derin öğrenme metotları ve hazır modeller test edilmiştir. Makine öğrenmesi yöntemlerinden en yüksek başarı ikili tivit verileri ile %88 oranında Lojistik Regresyon modelinden alınırken derin öğrenme modellerinden Çok Katmanlı Algılayıcı ile yine %88'lik başarı elde edilmiştir. Hazır model olarak BERT ve Yıldız Teknik Üniversitesi COSMOS AI Araştırma Ekibi tarafından geliştirilen "ytu-ce-cosmos/turkish-base-bert-uncased" modeli kullanılmıştır. Bu modeller farklı veri setlerinde farklı sonuçlar vermelerine rağmen her ikisi de en yüksek başarısını ikili tivit verilerinde %89 oranı ile elde etmişlerdir.

Anahtar Kelimeler: *Twitter, meslek tahmini, makine öğrenmesi, derin öğrenme, BERT, COSMOS*

1. Introduction

The use of social media platforms has become part of everyday life through which information, communication, and expression are passed across. There are around 5.07 billion social media users around the world [1]. Billions of people share information about their lives on these platforms.

Particularly, Twitter, now known as X, stands out because of its reach, speed of information transmission, and aspects of real-time interaction. Over 368 million monthly active users worldwide make a data mine for carrying out many analytical functions [2].

Twitter is popularly known for its specialties, like the 280-character limit for tweets, which encourages people to be effective and impactful with their message in very few words. There are many people on Twitter from different locations, different age groups, different nationalities, and with different personality traits. Its membership represents a wide spectrum of demographics, from teenagers to senior citizens, and a broad array of professions. This kind of diversity makes Twitter an ideal source for occupational data collection. Leveraging Twitter's large and diverse user base allows for more robust and generalized conclusions about occupational trends.

The study is important in many uses. For example, its ability to predict occupation accurately could allow good targeting of advertisements, better job matching algorithms, and sociological studies. It can also help in user profiling for personalization and a good user experience of different services. For these reasons, Twitter is a very suitable source for collecting data.

The article analyzed the potential capacity of Twitter data to predict the occupation of a person. The language of the tweets used as data in the study is Turkish. The project deals with machine learning and deep learning techniques applied to the tweets. The general idea is to make educated guesses concerning their professional background. The literature review is done in the second section of the article. In the third section, the dataset, preprocessing part, and feature extraction parts are covered. The fourth section includes results and discussion. Finally, a conclusion summarizes and interprets the results.

DOI: https://doi.org/10.21205/deufmd.2025278013

Geliş Tarihi / Received: 21.08.2024
Kabul Tarihi / Accepted: 18.09.2024

2. Literature Review

Preoţiuc-Pietro et al. (2015) conducted a study on predicting the occupational class of Twitter users by analyzing the content of their tweets [3]. Their dataset included 9 different occupation groups. The research involved a new annotated corpus of Twitter users and considered their job titles, the text of their tweets, and platform-related attributes such as the number of followers, number of accounts followed, the total number of tweets, the ratio of tweets containing hashtags, and the average number of tweets per day. Word clusters and embeddings, two examples of latent feature representations, were used to frame the task as a classification problem. The study included a variety of techniques, including Support Vector Machines, Gaussian Processes, and Logistic Regression. The Gaussian Process method produced the best accuracy, 52.7%.

Hu et al. (2016) used 8 different occupation categories in their dataset contained from Twitter [4]. These categories are marketing, administrator, start-up, editor, software engineer, public relations, office clerk, and designer. Linguistic Inquiry and Word Count (LIWC) were used in the study. The average F1-score of all eight jobs was 0.78.

Aletras et al. (2018) used both network and linguistic data to estimate the wealth and occupational class of Twitter users [5]. They presented a new technique that learns low-dimensional vector representations of users from their social networks by using graph embeddings. Their research showed that when compared to models that only use textual features, the inclusion of social network data greatly improves prediction performance. They obtained better results by fusing language-based and network-based elements, demonstrating the complementary roles of textual and social network data in predicting intricate socioeconomic characteristics of Twitter users.

Pan et al. (2019) explored the significance of social network information in predicting the occupational class of Twitter users [6]. By contrasting content-based data, such as profiles and tweets, with network-based data, like the relationships between followers, they show that integrating social network homophily greatly improves prediction performance. Their results show that utilizing network features produces superior classification results than depending only on tweet content or conventional bag-of-words models, especially when combined with user profiles.

A study by Margaret L. Kern et al. (2019) used linguistic data gathered from social media to forecast people's preferred careers [7]. The linguistic content of tweets from 128,279 people in 3,513 different jobs was examined. According to their research, social networking may help people to match with their ideal jobs, supporting career guidance for new graduates, disengaged employees, career changers, and the unemployed.

Zainab et al. (2021) used word embedding and deep neural networks to analyze biographical content on Twitter to forecast the professions of medical users [8]. They used several cutting-edge neural network models, such as Gated Recurrent Unit (GRU), Bidirectional Encoder Representations from Transformers (BERT) [9], a lite version of the BERT (ALBERT), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM). According to the study, word embedding combined with these neural network models performed better than conventional machine learning methods in terms of F1-score, accuracy, precision, and recall. With an F1-score of 0.90, ALBERT performed the best. This technique made it unnecessary to manually generate some features, which improved accuracy and simplified the categorization process.

In the study by Mayda (2022), occupation prediction was done using Turkish tweets [10]. An occupation dataset that includes 25,000 Turkish tweets was created and publicly shared. The words themselves as well as their roots were used to extract feature sets after several pre-processing procedures. Support Vector Machine and Logistic Regression techniques were used in the research. Tests were conducted on tweets individually and in groups of five and ten. For individual tweets, the best accuracy was 74.90%; for groups of five tweets, it was 96.20%; and for groups of ten tweets, it was 99.00%. The findings demonstrated that feature selection often improved success and that employing word roots yielded higher accuracy than using the words themselves.

Shaojie Yan et al. (2022) gathered data from Sina Weibo, the biggest microblogging platform in China, to examine the connection between people's use of social media and their occupation [11]. 67 occupations are used in the dataset obtained from 20,452 active bloggers. Later, they were narrowed down to six fundamental professions. For occupation estimate, the Time Aware-Long Short-Term Memory (T-LSTM), SVM, and LR models were used. The T-LSTM + LIWC approach yielded the greatest results, with an accuracy of 59.13%.

Shayan et al. (2022) used word embedding and deep neural networks to analyze user bios and tweets to predict the job titles of Twitter users [12]. They used a dataset of 1,314 samples that included both bios and tweets of users, with job titles found using Wikipedia crawling. The study uses a deep neural network (DNN) for job prediction, using the TF-IDF word embedding technique. Emojis and hashtags in tweets were also considered as inputs to enhance prediction accuracy. The researchers tested three embedding models (TF-IDF, word2vec, and Glove) alongside three classifiers: logistic regression, DNN, and convolutional neural networks (CNN). The TF-IDF embedding with a DNN classifier produced the best results of all these combinations, predicting job titles in nine different categories with an accuracy of 54%. The authors noted that the accuracy was better than random classification and suggested that further research with larger datasets and more complex models could improve prediction performance.

By utilizing multimodal learning, Liu et al. (2024) presented an innovative method for predicting the occupation of Twitter users [13]. Their investigation captured the geographic and environmental context of users' residentials and workplaces using both textual data from tweets and visual data from Google Street View (GSV) photos. These high-dimensional multimodal features were integrated into a multilayer transfer learning model, which allowed them to categorize jobs like office workers, students, and others or jobless people with great accuracy. The findings showed that when textual and visual embeddings are combined, occupation classification models perform far better than when either modality is used alone.

Zeki Ciplak et al. (2024) investigated the application of machine learning techniques to the prediction of Turkish Twitter users' occupational groups from their tweets [14]. The study manually identified Twitter accounts associated with various occupations and collected over 500,000 tweets. The Zemberek library was utilized for pre-processing, and the Count and TF-IDF vectorizers were employed to turn the tweets into numerical data. 36 different occupational groups, some single and some multiple, were used in the study. Machine learning algorithms Multinomial Naive Bayes (MNB) and Multinomial Logistic Regression (MLR) were tested with 2 different vectorizers, Count and TF-IDF vectorizers. The effect of the Zemberek library was also tested by comparing results with and without its use. With all these

parameters, and by grouping the tweets by 1, 5, and 10, there was a total of 24 models. After testing all the models, the study concluded that the Multinomial Logistic Regression algorithm with Count vectorization, Zemberek pre-processing, and grouping by 10 produced the best results, with an accuracy of 97.3%.

Compared to previous studies, this study evaluates the results using more machine learning and deep learning algorithms, pre-trained models, and different NLP libraries for pre-processing. The study also examines the variation in the results of the models using datasets of different sizes. Moreover, it investigates whether the results can be improved by treating the data in different ways, singly and pairwise.

3. Dataset and Methods

In this section, the dataset used in the study, pre-processing steps, feature extraction, and classification methods are covered.

3.1. Dataset and Pre-processing

There are 5 datasets used in this research. The first one is the dataset [15] used in Islam Mayda's research, "Predicting Occupation with Machine Learning from Turkish Tweets". This dataset includes 25,000 tweets from 10 occupations, with each occupation having 2500 tweets. Each occupation's data comes from 5 users equally, which have their occupation stated in their biography. The occupations are lawyer, dietitian, doctor, economist, teacher, psychologist, sports commentator, historian, software developer, and agricultural engineer. The dataset does not include tweets that are not in Turkish, are part of a conversation, contain only links, or are repetitive.

The second dataset is the increased version of the first dataset, which includes 30,000 tweets. It includes 5000 additional tweets from each occupation added equally.

The other 3 datasets are a randomized reduction of the second dataset to 10,000, 15,000 and 20,000 tweets.

On the pre-processing part, firstly all punctuation in tweets is deleted. After that, all the tweets have been converted to lowercase and the links in them are deleted. Finally, Zeyrek [16] and Zemberek [17] libraries are used separately to find the root of words. Zeyrek and Zemberek are both NLP libraries used for the Turkish language. The reason for using two NLP libraries is to compare their performance and try to achieve the best accuracy. Although the differences between the results were very slight, overall, the Zemberek library was more successful than the Zeyrek library. Therefore, the Zemberek library is used in the results section.

3.2. Feature Extraction

In the feature extraction part, the TF-IDF (Term Frequency-Inverse Document Frequency) statistical value of the data was obtained for all machine learning and deep learning methods and Bag-of-Words vectors were created for each data. At this point, the vectors obtained from data of different lengths were padded to be equal to the longest vector in the dataset. In this way, the inputs of the models were reduced to a single dimension. For the pre-trained models, similar padding operations were performed while no vectorization was applied to the dataset.

3.3. Classification

Classification was performed using four different machine learning methods: Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine. Additionally, deep learning approaches, including Multilayer Perceptron, Convolutional Neural Networks, and various Recurrent Neural Networks (such as Long Short-Term Memory, Bidirectional Long Short-Term Memory, and Gated Recurrent Unit models), were applied for more complex analysis. Brief descriptions of these models are provided below.

The BERT model pre-trained by Google and a Turkish LLM model [18] fine-tuned from BERT developed by Yıldız Technical University COSMOS AI Research Group [19] were also used for classification.

Support Vector Machine (SVM): SVM finds the optimal hyperplane that maximizes the margin between different classes in a dataset.

Logistic Regression (LR): Logistic Regression models the probability of a binary outcome using a logistic function.

Random Forest (RF): Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions for more accurate and stable results.

Naive Bayes (NB): Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence between predictors.

Multilayer Perceptron (MLP): MLP is a type of neural network consisting of multiple layers of nodes, each fully connected to the next, used for various types of predictions.

Convolutional Neural Network (CNN): CNN is a deep learning model that uses convolutional layers to automatically and adaptively learn spatial hierarchies of features from input images.

Long Short-Term Memory (LSTM): LSTM is a type of recurrent neural network (RNN) that can learn long-term dependencies and retain information over extended sequences.

Bidirectional Long Short-Term Memory (bi-LSTM): bi-LSTM is an extension of LSTM that processes data in both forward and backward directions to capture context from both past and future states.

Gated Recurrent Unit (GRU): GRU is a simplified version of LSTM that uses gating units to regulate the flow of information without separate memory cells, making it computationally efficient.

In machine learning methods, the maximum number of iterations for Logistic Regression in the classification phase was set as 1000. In the Naive Bayes method, Multinomial Naive Bayes, which is widely used in text classification, was preferred. Finally, the "linear" kernel function was used in the Support Vector Machine method. In the Random Forest method, the default values were preserved.

3.4. Pairwise Tests

After the models were tested using the original texts in the dataset, they were also tested with new and longer texts obtained by combining 2 consecutive tweets (pairwise) from the same Twitter user into 1 individual tweet in each occupational class to avoid the fact that the dataset contains Twitter data, and this data may not contain enough information for classification in some cases. In this way, by grouping the tweets, texts with short and insufficient information were combined to obtain more comprehensive data [10].

4. Results and Discussion

All 5 datasets were given to the relevant models and tests were performed. At this point, different Turkish natural language processing libraries were used. The effect of increasing the dataset size on the results was analyzed. At the same time, the results of treating the data as singular and pairwise were also observed.

4.1. Accuracy

In this subsection, the accuracies of the models on data sets of different sizes are analyzed.

Table 1. Results for singular data.

Method Type	Method	10K	15K	20K	25K	30K
Machine Learning	SVM	**0.71**	0.70	0.69	**0.71**	0.69
	LR	0.71	0.71	0.72	**0.74**	0.73
	RF	0.59	0.58	0.59	**0.60**	0.59
	NB	0.70	0.72	0.71	**0.73**	0.72
Deep Learning	MLP	0.70	0.71	0.72	**0.74**	0.73
	CNN	0.68	0.69	0.70	**0.72**	**0.72**
	LSTM	0.62	0.63	0.65	0.66	**0.67**
	bi-LSTM	0.65	0.64	0.68	**0.71**	0.69
	GRU	0.62	0.64	0.64	**0.68**	0.65
Pre-Trained	BERT	0.75	0.75	0.76	0.76	**0.78**
	COSMOS	0.76	0.75	0.74	0.77	**0.78**

When Table 1 is examined, it is seen that the increase in the dataset size has a mostly positive effect on the results. In cases where the number of data is 30,000, it is seen that the accuracy of machine learning methods decreases. In deep learning methods, the accuracy of CNN is maintained while the accuracy of LSTM increases by 1%.

On the other hand, for the pre-trained models BERT and COSMOS, both models outperformed all other models by achieving an accuracy of 78%. At the same time, it is observed that the performance increases in parallel with the increase in the size of the dataset. This is expected since large language models require a large amount of data.

Table 2. Results for pairwise data.

Method Type	Method	10K	15K	20K	25K	30K
Machine Learning	SVM	0.85	0.85	**0.87**	0.85	0.83
	LR	0.85	0.85	**0.88**	0.85	0.85
	RF	0.75	**0.76**	**0.76**	0.74	0.73
	NB	0.82	0.85	**0.86**	0.84	0.82
Deep Learning	MLP	0.83	0.85	**0.88**	0.84	0.85
	CNN	0.79	0.81	**0.84**	0.83	0.82
	LSTM	0.71	0.74	**0.79**	0.77	0.75
	bi-LSTM	0.72	0.76	**0.79**	0.78	**0.79**
	GRU	0.71	0.76	**0.77**	0.74	0.74
Pre-Trained	BERT	0.86	0.88	**0.89**	0.88	0.87
	COSMOS	0.86	**0.89**	**0.89**	0.88	0.87

When Table 2 is examined, it is seen that the accuracy values are much higher than the singular data. This is due to the longer and more comprehensive treatment of the inputs. At this point, it is observed that the effect of increasing the dataset size on the accuracy value is not stable. In general, the highest accuracy was achieved with a dataset size of 20K. The pre-trained models achieved the highest accuracy of 89%, as was also the case for the singular data.

When both tables are analyzed, it is seen that the increase in dataset size changes the model performance. Although this was as small as 1% for machine learning models, it increased the success of deep learning and pre-trained models up to 8%.

One of the reasons why increasing the size of the dataset does not have a positive effect on some results may be the quality of the data added. The fact that posts on Twitter are from a certain occupational class but do not contain information relevant to that occupation may cause the models to misclassify such data. On the other hand, for example, the fact that the occupations of doctor and dietitian in the dataset contain similar data may have caused the models to misclassify these occupations. To prevent such situations, the dataset may need to be carefully examined, and problematic data may need to be removed from the dataset.

It is also seen that the pairwise data has a great impact on the model performance. At this point, BERT and COSMOS models, which have the highest accuracy in both cases, provide an increase of 11% in pairwise data. This improvement increases up to 16% in models such as RF and SVM.

4.2. Training Duration

In this subsection, the training durations of the models are analyzed for pairwise data on datasets of different sizes.

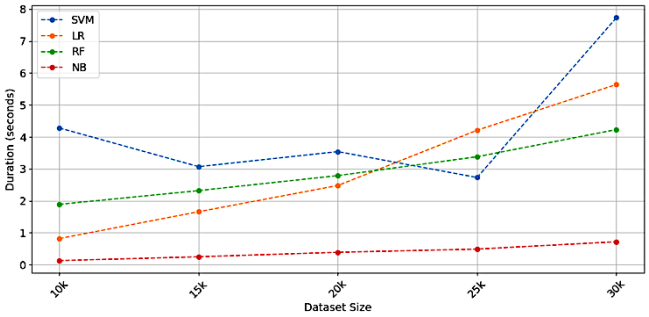

Figure 1. Training Duration on Pairwise Data by Dataset Size for Machine Learning Models.

Figure 1, which shows the training duration of machine learning models on pairwise data according to dataset size, shows that the Naive Bayes method gives much faster results than other methods due to its probabilistic approach.

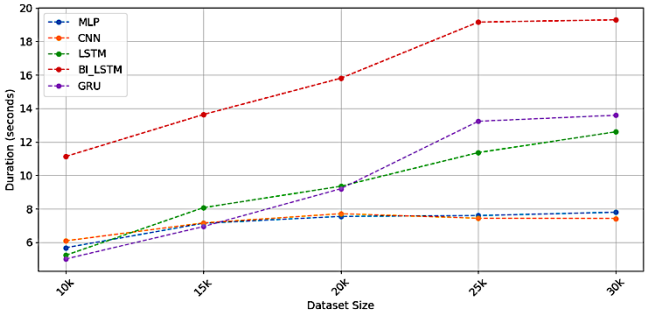

Figure 2. Training Duration on Pairwise Data by Dataset Size for Deep Learning Models.

Figure 2, which shows the training duration of deep learning models on pairwise data according to dataset size, shows that the bi-LSTM method takes much more time than the other methods. This may be because the data is analyzed bidirectionally, which causes this method to be slow.

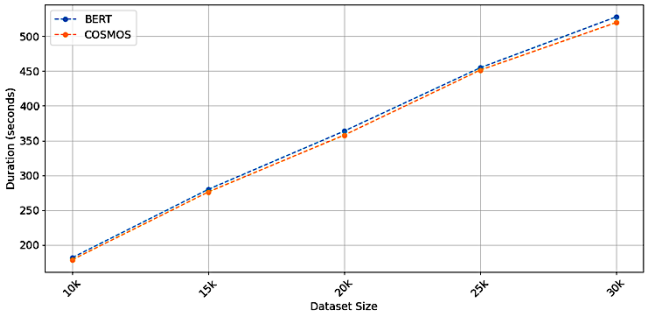

Figure 3. Training Duration on Pairwise Data by Dataset Size for Pre-Trained Models.

Figure 3, which shows the training duration of the pre-trained models on pairwise data according to the dataset size, shows that training is completed in similar durations for both models. However, it is observed that the COSMOS model is slightly faster than the BERT model.

It is also seen that the increase in the number of data increases the training duration for most models as expected. Even if this is not very evident in machine learning and deep learning methods, it shows significant differences in pre-trained models.

5. Conclusion

In this study, Turkish tweets belonging to 10 occupational classes collected from Twitter are classified with various machine learning and deep learning methods and pre-trained models for the model to predict the correct occupational class. For this purpose, firstly, the texts in the dataset were decomposed into their roots with Turkish natural language processing methods. Then, various feature extraction methods were applied, and the texts were given to the relevant models.

To evaluate the effect of increasing the dataset size on model performance, an additional 5,000 tweets were added to the original dataset of 25,000, forming a new dataset with 30,000 entries. Five datasets of varying sizes —10,000, 15,000, 20,000, 25,000, and 30,000— were subsequently created to systematically analyze the performance across different data scales.

The results revealed that the highest accuracy rates were obtained when the data were grouped in pairs, demonstrating the effectiveness of pairwise learning for this task. Notably, the highest accuracy in the study, 89%, was achieved by the BERT and COSMOS models when tested on pairwise data from the 20,000-entry dataset. COSMOS also achieved the same 89% accuracy on the 15,000-entry dataset.

It was observed that increasing the dataset size generally improved performance, particularly for deep learning and pre-trained models. This study builds on prior research by thoroughly investigating the impact of dataset size on model performance, a factor often overlooked in previous studies. Additionally, the use of pairwise data grouping is shown to significantly enhance model accuracy, which sets this study apart from earlier approaches.

6. Future Work

In future studies, the models can be tested by increasing the size of the dataset even more. At the same time, by adding different occupational classes or removing existing occupations, it can be investigated in which occupational classes the models show lower or higher performance. The tests can be repeated with different pre-trained models such as RoBERTa, XLNet, and GPT to measure their performance.

Ethics committee approval and conflict of interest statement

This article does not require ethics committee approval. This article has no conflicts of interest with any individual or institution.

Author Contribution Statement

T. İzdaş and H. İskifoğlu designed the study and implemented the models. T. İzdaş drafted the manuscript. H. İskifoğlu conducted the literature review and prepared the dataset. B. Diri supervised the project and provided overall guidance. All authors read and approved the final version of the manuscript.

References

[1] Smart Insights. 2024. Global social media statistics research summary. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/(Accessed: 2024-05-01).

[2] Backlinko. 2024. X (Twitter) Statistics: How Many People Use X? https://www.statista.com/statistics/303681/twitter-users-worldwide/ (Accessed: 2024-05-24).

[3] Preoţiuc-Pietro, D., Lampos, V., Aletras, N. 2015. An analysis of the user occupational class through Twitter content, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1754-1764.

[4] Hu, T., Xiao, H., Luo, J., Nguyen, T. 2016. What the language you tweet says about your occupation, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 10, No. 1, pp. 181-190.

[5] Aletras, N., Chamberlain, B. P. 2018. Predicting Twitter user socioeconomic attributes with network and language information, in: Proceedings of the 29th on Hypertext and Social Media, pp. 20-24.

[6] Pan, J., Bhardwaj, R., Lu, W., Chieu, H. L., Pan, X., Puay, N. Y. 2019. Twitter homophily: Network based prediction of user's occupation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2633-2638.

[7] Kern, M. L., McCarthy, P. X., Chakrabarty, D., Rizoiu, M. 2019. Social media-predicted personality traits and values can help match people to their ideal jobs, Proc. Natl. Acad. Sci. USA, Vol. 116, No. 52, pp. 26459-26464.

[8] Zainab, K., Srivastava, G., Mago, V. 2021. Identifying health related occupations of Twitter users through word embedding and deep neural networks, BMC Bioinformatics, Vol. 22, Suppl 10, p. 630.

[9] Devlin, J., Chang, M., Lee, K., Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, pp. 4171-4186. DOI: 10.18653/v1/N19-1423.

[10] Mayda, İ. 2022. Türkçe Tweetlerden Makine Öğrenmesi ile Meslek Tahmini, Avrupa Bilim ve Teknoloji Dergisi, Vol. 40, pp. 55-60.

[11] Yan, S., Zhao, T., Deng, J. 2022. Predicting social media user occupation with content-aware hierarchical neural networks, in: 2022 8th International Conference on Big Data and Information Analytics (BigDIA), pp. 388-395.

[12] Vassef, S., Toosi, R., Akhaee, M. A. 2022. Job title prediction from tweets using word embedding and deep neural networks, in: Proceedings of the 2022 30th International Conference on Electrical Engineering (ICEE), pp. 577-581.

[13] Liu, X., Peng, B., Wu, M., Wang, M., Cai, H., Huang, Q. 2024. Occupation prediction with multimodal learning from tweet messages and Google Street View images, AGILE: GIScience Ser., Vol. 5, p. 36.

[14] Ciplak, Z., Yildiz, K. 2024. Occupational groups prediction in Turkish Twitter data by using machine learning algorithms with multinomial approach, Expert Syst. Appl., p. 124175.

[15] Mayda, I. 2022. Occupation Dataset in Turkish. https://github.com/imayda/occupation-dataset-in-turkish(Accessed: 2024-05-10).

[16] Bulat, O. 2020. Zeyrek: Morphological Analyzer and Lemmatizer. https://github.com/obulat/zeyrek (Accessed: 2024-05-25).

[17] Akın, A. A. 2014. Zemberek-NLP. https://github.com/ahmetaa/zemberek-nlp (Accessed: 2024-05-25).

[18] Kesgin, H. T., Yuce, M. K., Amasyali, M. F. 2023. Developing and evaluating tiny to medium-sized Turkish BERT models, arXiv preprint arXiv:2307.14134.

[19] YTU COSMOS AI Research Group. 2024. https://ce.yildiz.edu.tr/genel-sayfa/tr/cosmosrg (Accessed: 2024-07-08).