Occupation Prediction Using Twitter Data

Berkan Eti, Ömer Diner
Department of Computer Engineering
Yildiz Technical University, 34220 Istanbul, Türkiye
{berkan.eti, omer.diner}@std.yildiz.edu.tr

***Özetçe* —Twitter, dünya genelinde milyonlarca kullanıcının kısa mesajlar aracılığıyla duygu ve düşüncelerini paylaştığı bir sosyal medya platformudur. Kullanıcıların paylaşımları, özellikle metinsel verilerle çalışmalar yapan araştırmacılar için önemli bir veri kaynağı sunmaktadır. Bu proje kapsamında 5 farklı makine öğrenmesi, 2 derin öğrenme algoritması ve 1 adet önceden eğitilmiş Türkçe BERT modeli [1], farklı doğal dil işleme teknikleri ve modellerin eğitimlerinde kullanılan tivitlerin tekli ve çoklu olarak kombinasyonlar şeklinde modellere verilmesi sonucunda ortaya çıkan sonuçlar gösterilmiş ve modellerin başarı oranları arasındaki farklar yorumlanmıştır. Bu çalışma için 43.000 veri içeren hazır veri setine ek olarak 11.000 adet yeni veri toplanmıştır. Bu veriler, veri temizleme ve köklerine ayırmayı içeren ön işleme adımlarına sokulduktan sonra modellere girdi olarak verilmiştir. Makine öğrenmesi modellerinde en yüksek performansı veren yöntem tekli verilerde %74.2 doğruluk oranı ile Destek Vektör Makineleri yöntemi olmuştur. Derin öğrenme modellerinde ise tekli verilerde en yüksek başarı oranına %73.5 ile Çok Katmanlı Algılayıcı ulaşmıştır. Eğitilen modeller arasında en yüksek performansı veren kombinasyon %95.1 başarı oranı ile beşli olarak gruplanan veriler kullanılarak Zemberek kütüphanesi [2] ile doğal dil işleme adımları uygulanan Destek Vektör Makineleri olmuştur.**

Anahtar Kelimeler—Twitter, Makine Öğrenmesi, Metin Sınıflandırma, Derin Öğrenme, Meslek Tahmini

***Abstract*—Twitter is a social media platform where millions of users around the world share their thoughts and feelings through short messages. Users' posts provide an important data source, especially for researchers working with textual data. Within the scope of this project, 5 different machine learning algorithms, 2 deep learning algorithms and 1 pre-trained Turkish BERT model [1], different natural language processing techniques and the results obtained by training the models with single and multiple combinations of tweets are shown and the differences between the success rates of the models are interpreted. For this study, 11,000 new data were collected in addition to the ready-made dataset containing 43,000 data. These data were given as input to the models after preprocessing steps including data cleaning and lemmatization. In machine learning models, the highest performing method was the Support Vector Machines method with an accuracy rate of 74.2% on single data. In deep learning models, Multilayer Perceptron achieved the highest success rate with 73.5% on single data. Among the trained models, the combination with the highest performance was Support Vector Machines with 95.1% success rate, which applied natural language processing steps with the Zemberek library [2] using data grouped into fives.**

Keywords—Twitter, Machine Learning, Text Classification, Deep Learning, Occupation Prediction

I. Introduction

The growing integration of social media platforms into daily life has transformed them into essential tools for communication, information exchange and self-expression. As of 2024, there are approximately 5.07 billion social media users worldwide [3], contributing to an immense and diverse pool of data. These platforms offer researchers valuable opportunities to analyze user behavior and societal trends. Among these platforms, Twitter stands out as a significant data source due to its real-time interaction capabilities, brevity in communication and global reach. With over 586 million active users as of 2024 [4], Twitter provides a unique platform for textual data analysis, including natural language processing (NLP). Twitter's diverse user base, spanning different age groups, nationalities and professions, makes it an ideal environment for collecting occupational data and understanding professional behaviors.

This article investigates the potential of Twitter data to predict a user's profession, focusing on tweets written in Turkish. The study applies machine learning and deep learning techniques to analyze the textual data, aiming to develop accurate predictive models. The second section presents a comprehensive literature review, examining prior research on social media data analysis and occupation prediction. The third section explains the dataset used in this study, detailing the data collection methods, preprocessing techniques and feature extraction processes. In the fourth section, the results are presented and discussed, emphasizing the performance of the developed models and the factors influencing their accuracy. Finally, the conclusion provides a summary of the findings, evaluates their implications and offers suggestions for future work.

II. Literature Review

Hu et al. (2016) examined the relationships between language use, interests and personality traits of different occupational groups. While creating the dataset, users with LinkedIn and Twitter profiles were preferred and IBM Watson Personality Insights API was used to extract personality traits. In the data cleaning phase, unnecessary expressions were removed by applying PMI (Pointwise Mutual Information) method and occupation prediction was made with 78% F-score [5].

Çıplak et al. (2024) reduced 65 different occupational groups into 36 categories and collected data from 304 Twitter users. The roots of words were found using the Zemberek library [2] and MNB (Multinomial Naive Bayes) and MLR (Multiple linear regression) algorithms were

applied for classification. Using different vectorization and preprocessing methods, 24 models were tested and the best result with 97.3% accuracy was obtained with MLR algorithm and CVec (CountVectorizer) method [6].

Shayan et al. (2022) studied occupation prediction using Twitter users' tweets. In the data collection process, data of 1,314 users were obtained using emoji and hashtags. K-means clustering algorithm was applied to create occupational groups of users. In the classification phase, text processing methods such as TF-IDF (Term Frequency-Inverse Document Frequency), BoW (Bag of Words), Word2Vec (Word to Vector) and GloVe (Global Vectors for Word Representation) were used to digitize textual expressions. Among the classifier algorithms, the best performance was achieved by the combination of DNN (Deep Neural Network) and TF-IDF with an accuracy of 54% [7].

Roobaea et al. (2020) aimed to predict age, gender and occupation by analyzing celebrity followers. In the data cleaning phase, HTML (HyperText Markup Language) tags and URLs were removed from the texts and the WordNet-Similarity tool was used to classify the words according to their meanings. LSTM (Long Short-Term Memory) algorithm was applied for prediction. As a result of the study, gender prediction was completed with 69.6% accuracy and occupation prediction was completed with 59.8% accuracy [8].

The study by Mayda (2022) is the first research conducted in Turkish in this field. The dataset was created by identifying occupational groups and selecting real users from Twitter profiles for each occupation. In the data cleaning steps, operations such as lowercase conversion and punctuation removal were applied and word roots were extracted with the Zemberek library. Support Vector Machine (SVM) and Logistic Regression (LR) methods were used in the classification phase and the experiments were conducted with 10-fold cross-validation. The study showed that clustered data, rather than individual tweets, yielded higher success rates. An accuracy rate of 96.2% was achieved with clusters of 5 and 99% with clusters of 10. It was also found that the SVM method gave better results than LR. The study suggests that the dataset can be improved as Twitter features change over time and the use of profile-based data can provide higher success [9].

III. Dataset and Preprocessing

The dataset consists of the ready-made dataset used in Islam Mayda's study [9] and new data collected recently for this project. In addition to the ready-made dataset of approximately 43,000 data, 11,000 new data were obtained through an API for various social media platforms. Thus, a new dataset of 54,000 data was created in total. This dataset includes tweets of social media users belonging to 10 predetermined occupational groups. During the new data collection process, the Twitter accounts of the target users were accessed by using keywords and tags specific to the relevant professions.

A. Data Analysis

The dataset was analyzed for each occupation based on basic statistical analysis such as the number of data, the presence of empty rows and columns and tweet lengths. As a result of the analysis, psychologists and historians were the most represented professions in the dataset, while lawyers and dietitians were underrepresented. When the lengths of tweets were analyzed, it was observed that the occupational group with the highest average character count was doctors (*180.93 characters*), while the occupational group with the shortest tweets was software developers (*114.11 characters*). This difference suggests that occupations may have a distinctive effect on social media posts. Tweets that were too short to be empty or meaningless were also analyzed and cleaned as part of the data preprocessing steps to be applied in the next stage.

Table 1 Number of Tweets by Occupation Based on Data Source

Occupation	Source	Number of Tweets
Lawyer	API	892
	Mayda	2809
Dietician	API	809
	Mayda	3526
Doctor	API	1328
	Mayda	4910
Economist	API	1375
	Mayda	3867
Psychologist	API	1304
	Mayda	5835
Sport Commentator	API	995
	Mayda	4180
Historician	API	1444
	Mayda	4878
Software Developer	API	987
	Mayda	3974
Agricultural Engineer	API	473
	Mayda	4695
Teacher	API	1204
	Mayda	4350

B. Data Preprocessing

In order to make the dataset suitable for the models, a data preprocessing process consisting of two main steps was performed: **Data Cleaning** and **Lemmatization**.

a) Data Cleaning: Emojis, links, usernames, hashtags and punctuation marks were removed during the cleaning of the tweet content. Texts have been completely converted to lower case, numbers and unnecessary spaces have been removed. A Turkish-specific *stop word* list was used to remove words such as meaningless conjunctions and question suffixes. In this process, the Python NLTK package [10] was used to identify the words in this list and remove them from the text.

b) Lemmatization: The cleaned texts were separated into words and the root of each word was found. The Python libraries Zeyrek [11] and Zemberek [2] were used

in this process. The root words were combined and tweets containing less than two words were removed from the dataset.

As a result of all these processes, approximately 50,000 cleaned and pre-processed data were saved in a JSON file, including profession tags and data source information and made ready for use.

IV. Methods and Results

Five different machine learning methods were used in this project: LR, Multinomial Naive Bayes (MNB), Random Forest (RF) and SVM, Gradient Boosting, while Multi-Layer Perceptron (MLP) and Convolutional Neural Network (CNN) were used for deep learning methods. The BERT model pre-trained by Turkish dataset was also used for classification.

During the training of the models, different NLP techniques and various combinations of tweets were used to evaluate the performance results.

Table 2 Accuracy Results for the Combined Dataset

Method Type	Method	NLP	Single	Binary	Triple	Quintuple
Machine Learning	Logistic Regression	Zeyrek	0.727	0.839	0.907	0.942
		Zemberek	0.721	0.844	0.905	0.946
	Random Forest	Zeyrek	0.664	0.764	0.828	0.879
		Zemberek	0.659	0.760	0.830	0.887
	Support Vector Machines	Zeyrek	0.742	0.846	0.912	0.946
		Zemberek	0.735	0.850	0.911	0.951
	Gradient Boosting	Zeyrek	0.649	0.762	0.827	0.864
		Zemberek	0.647	0.755	0.827	0.862
	Multinomial Naive Bayes	Zeyrek	0.721	0.829	0.895	0.927
		Zemberek	0.713	0.826	0.896	0.927
Deep Learning	CNN	Zeyrek	0.705	0.810	0.873	0.931
		Zemberek	0.698	0.812	0.867	0.924
	MLP	Zeyrek	0.735	0.830	0.898	0.945
		Zemberek	0.719	0.837	0.884	0.933
Pre-trained Model	BERT	Zeyrek	0.780	0.868	0.920	0.934
		Zemberek	0.760	0.864	0.906	0.943

When Table 2 is analyzed, it is seen that the best performance among machine learning models is achieved with LR and SVM algorithms, while MLP algorithm provides the highest success among deep learning models. The BERT model trained for Turkish provided a higher accuracy rate than all models.

In the LR model, the best results were obtained when $C = 1$ and L2 penalty function were used for single data. In multiple combinations, with the increase in data diversity, C was set to 10 and more complex structures were learned. Similarly, the SVM algorithm provided better accuracy rates with higher C values as the number of combinations increased.

In the MLP model, three hidden layers consisting of 256, 128 and 64 neurons were used respectively and ReLU activation function and Adam optimization algorithm were preferred. Dropout values were set as 0.4, 0.3 and 0.2 and the best performance was achieved with this structure.

A comparison with Islam Mayda's research shows that the results are close to each other due to the similar characteristics of the two datasets. In the training with single and multiple tweet combinations, increasing the number of combinations significantly increased the accuracy rates. For example, while the CNN algorithm achieved 71% accuracy with single combinations, this rate increased to 93% with five combinations. This increase can be explained by the effect of data diversity and context richness.

Table 3 Comparison with Results from [9]

Method	Combination Count	[9]	Recommended Technique
Logistic Regression	Single	0.747	0.721
	Quintuple	0.952	0.946
Support Vector Machines	Single	0.740	0.735
	Quintuple	0.958	0.951

A decrease in accuracy rates was observed in cases where training and test datasets came from different sources. Higher accuracy was achieved when both training and test datasets came from the same source, while this rate decreased when data from different sources was used. For example, while the average tweet length in the API-sourced dataset is 182.65 characters, the average length in Mayda's dataset is 142.57 characters. This difference caused accuracy rates to vary due to the formal diversity of data sources. Additionally, differences in content quality of tweets from different data sources and noise rates independent of occupational groups were also considered among factors affecting the results.

The pre-trained BERT model for Turkish achieved higher accuracy rates compared to other machine learning and deep learning models. The main reason for this is BERT's richer pre-training dataset and its ability to better contextualize words within context. The BERT model showed superior performance particularly in classification problems requiring context. This proves that context is an important factor in occupation prediction.

Finally, relationships between occupational groups with similar terminology were also evaluated. Due to terminological similarities between occupational groups such as doctors, dietitians and psychologists, there was confusion in model outputs in some cases. For example, some tweets that should have been labeled as doctor were classified as dietitian or psychologist. This situation was evaluated as a natural confusion between terminologically close occupations. No such relationship was observed between other occupational groups.

As a result of these analyses, it was observed that expanding the dataset and applying different NLP methods increased model success, but the heterogeneity of data sources had a negative effect on accuracy rates. Additionally, using multiple tweet combinations and ensuring context richness significantly improved classification accuracy.

V. Conclusion

In this study, models were developed to determine which of the 10 professions a given Turkish tweet was tweeted by a person belonging to which occupational group. In this process, data was collected in addition to Islam Mayda's dataset. The dataset was organized by applying some NLP methods on the obtained dataset. Two different methods, Zeyrek [11] and Zemberek [2], were used for data preprocessing and lemmatization. For training the models, 5 ML methods, 2 DL methods and BERT [1] were used. In addition, the tweets given as input to the models; 4 different combinations of single, double, triple and quintet were tried and the results were compared. When the results are observed, it is seen that SVM gives the best result among ML methods, while MLP has a higher accuracy rate than CNN among DL methods. BERT, on the other hand, achieved better results than both ML and DL methods. It was concluded that the models using Zeyrek, one of the NLP methods, gave better results compared to the models using Zemberek. In addition, it was understood that the results changed when the data sets used in the comparisons were divided into training and test without merging.

VI. Future Studies

In future studies, the size of the data set can be increased and the performance of the models can be tested. In addition, occupations with similar and different terminology can be added to the existing 10 occupational groups and the probability distribution can be observed. In addition, the results of the study can be enriched by vectorization by using different methods such as Word2Vec and GloVe in the preprocessing stage and including ready-made datasets for Turkish. New project-specific models can also be created by using large language models, which have become increasingly popular recently.

References

[1] B. Staatsbibliothek. (2024) dbmdz turkish bert model. [Online]. Available: https://huggingface.co/dbmdz/bert-base-turkish-cased

[2] Loodos. (2020) Zemberek. [Online]. Available: https://pypi.org/project/zemberek-python

[3] Statusbrew. (2024) 100+ social media statistics you need to know in 2024 [all networks]. [Online]. Available: https://statusbrew.com/insights/social-media-statistics

[4] TweetDelete. (2024) Twitter karakter sayısı: X'in tweet sınırlarını anlamak. [Online]. Available: https://tweetdelete.net/tr/kaynaklar/twitter-character-count

[5] T. Hu, T. N. Thuy-vy, H. Xiao, and J. Luo, "What the language you tweet says about your occupation," *ICWSM 2016*, pp. 181–190, 2016.

[6] Z. Ciplak and K. Yildiz, "Occupational groups prediction in turkish twitter data by using machine learning algorithms with multinomial approachwhat the language you tweet says about your occupation," *Expert Systems With Applications*, pp. 2–12, 2024.

[7] S. Vassef, R. Toosi, and M. A. Akhaee, "Job title prediction from tweets using word embedding and deep neural networks," *30th International Conference on Electrical Engineering (ICEE)*, pp. 577–581, 2022.

[8] R. Alroobaea, A. H. Almulihi, F. S. Alharithi, S. Mechti, M. Krichen, and L. H. Belguith, "A deep learning model to predict gender, age and occupation of the celebrities based on tweets followers," *CLEF 2020*, 2020.

[9] İslam MAYDA, "Türkçe tweetlerden makine Öğrenmesi ile meslek tahmini," *Avrupa Bilim ve Teknoloji Dergisi*, pp. 55–60, 2022.

[10] (2016) node-nltk-stopwords. [Online]. Available: https://github.com/xiamx/node-nltk-stopwords

[11] O. Bulat. (2020) Zeyrek. [Online]. Available: https://zeyrek.readthedocs.io