*Avrupa Bilim ve Teknoloji Dergisi*
*Özel Sayı 40, S. 55-60, Eylül 2022*

***Araştırma Makalesi***

www.ejosat.com ISSN:2148-2683

*European Journal of Science and Technology*
*Special Issue 40, pp. 55-60, September 2022*

***Research Article***

# Türkçe Tweetlerden Makine Öğrenmesi ile Meslek Tahmini

İslam Mayda

Yıldız Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye, (ORCID: 0000-0001-5584-0259), islam.mayda@stu.khas.edu.tr

(1st International Conference on Innovative Academic Studies ICIAS 2022, September 10-13, 2022)

(**DOI:** 10.31590/ejosat.1168269)

**ATIF/REFERENCE:** Mayda, İ. (2022). Türkçe Tweetlerden Makine Öğrenmesi ile Meslek Tahmini. *Avrupa Bilim ve Teknoloji Dergisi*, (40), 55-60.

**Öz**

Sosyal medya platformlarının yaygınlaşması ve kullanıcı sayılarının hızla artmaya devam etmesiyle birlikte sosyal medyada üretilen veri miktarı da hızlı bir şekilde büyümektedir. Bu veriden bilgi çıkarmaya yönelik yapılan bilimsel çalışmaların hedeflerinden biri de meslek tahminidir. Sosyal medya kullanıcılarının meslek bilgisi, akıllı öneri sistemleri başta olmak üzere birçok farklı alanda kullanılabilmektedir. Bu çalışmada da Türkçe tweetler kullanılarak meslek tahmini yapılması amaçlanmıştır. Çalışma kapsamında öncelikle 25.000 Türkçe tweetten oluşan meslek veri kümesi oluşturulmuş ve kamuya açık olarak paylaşılmıştır. Bu veri kümesi üzerinde çeşitli önişleme adımları uygulanmış, hem kelimelerin kendileri hem de kelime kökleri kullanılarak özellik kümeleri çıkarılmıştır. Yapılan testlerde tweetler hem tekil olarak hem de 5'li ve 10'lu gruplar halinde birleştirilerek kullanılmıştır. Destek Vektör Makinesi ve Lojistik Regresyon yöntemlerinin uygulandığı deneylerde özellik seçimi yapılarak testler tekrar edilmiştir. Tekil tweetlerle yapılan deneylerde en iyi sonuç %74,90 doğruluk oranı olarak elde edilirken, 5'li gruplar halinde birleştirilmiş tweetlerle yapılan deneylerde %96,20 ve 10'lu gruplar halinde birleştirilmiş tweetlerle yapılan deneylerde %99,00 doğruluk oranları en iyi performanslar olarak raporlanmıştır. Testlerde kelime köklerinin kullanılmasının kelimelerin kendilerini kullanmaya göre daha yüksek başarı gösterdiği ve özellik seçiminin genel olarak başarıyı yükselttiği görülmüştür. Çalışmanın sonunda, alınan bu sonuçlar tartışılmış ve gelecek çalışmalara dair öneriler sunulmuştur.

**Anahtar Kelimeler:** Meslek tahmini, Meslek tespiti, Makine öğrenmesi, Doğal dil işleme, Twitter.

# Predicting Occupation with Machine Learning from Turkish Tweets

**Abstract**

With the spread of social media platforms and the rapid increase in the number of users, the amount of data produced in social media is growing rapidly. One of the goals of scientific studies to extract information from this data is occupation prediction. Social media users' occupation information can be used in many different areas, especially in smart suggestion systems. In this study, it is aimed to make occupation prediction using Turkish tweets. Within the scope of the study, an occupation dataset consisting of 25,000 Turkish tweets was created and shared publicly. Various preprocessing steps were applied on this dataset, and feature sets were extracted using both the words themselves and the word roots. In the tests, tweets were used both singularly and combined in groups of 5 and 10. In the experiments in which Support Vector Machine and Logistic Regression methods were applied, tests were repeated by feature selection. While the best result was obtained as 74.90% accuracy in the experiments with singular tweets, the best performances were reported as 96.20% accuracy in experiments with tweets combined in groups of 5, and 99.00% accuracy in experiments with tweets combined in groups of 10. It has been seen that the using of word roots in the tests has higher success than using the words themselves, and the feature selection generally increases the success. At the end of the study, these results were discussed and suggestions for future studies were presented.

**Keywords:** Occupation prediction, Profession identification, Machine learning, Natural language processing, Twitter.

# 1. Giriş

Günümüz dünyasında sosyal medya hayatımızın ayrılmaz bir parçası haline gelmiştir. İstatistiklere göre dünya çapında 2022 Temmuz itibariyle 4,7 milyar sosyal medya kullanıcısı mevcuttur ve bu sayı da küresel nüfusun %59'una karşılık gelmektedir (Kepios, 2022). Kullanıcılar sosyal medya platformlarında her türlü duygu, düşünce, bilgi ve görüşlerini paylaşmaktadır. Milyarlarca kullanıcının metin, fotoğraf, video ve ses gibi farklı şekillerde yaptığı paylaşımlarıyla büyük veri oluşmaktadır. Bu veriden çeşitli amaçlarla bilgi çıkarımı üzerine özellikle son yıllarda çok sayıda bilimsel çalışma yapılmıştır.

Yaygın olarak kullanılan sosyal medya sitelerinin biri de dünya genelinde 400 milyondan fazla aktif kullanıcıya sahip olan Twitter'dır (Statista, 2022). Bir mikroblog servisi olan Twitter, bir bilginin hızlı bir şekilde geniş kitlelere yayılmasına imkân sağlarken, bu özelliğiyle gündemi oluşturmada önemli bir konuma sahiptir. Twitter kullanıcılarının gerçekte kim olduğu, yaşı, cinsiyeti, lokasyonu, milliyeti, eğitim durumu, politik görüşü, kişilik özellikleri gibi birçok özelliğini tahmin etmeye yönelik yapılan araştırmalar literatürde yer almaktadır. Üzerinde çalışılan konulardan biri de kullanıcıların mesleğinin tahmin edilmesidir. Bir Twitter kullanıcısının, daha geniş bir ifadeyle, bir sosyal medya kullanıcısının mesleğinin tahmin edilebilmesi akıllı öneri sistemleri başta olmak üzere birçok amaç için kullanılabilir. Örneğin, kullanıcılara mesleklerine özel içerikler, etkinlikler, ürünler veya reklamlar sunulabilir. Kullanıcılarla aynı meslekteki diğer kullanıcı profilleri, takip etmeleri için kendilerine önerilebilir. Ayrıca, kullanıcıların meslek bilgisi sosyal medya üzerindeki kamuoyu araştırmalarında da kullanılabilir. Mesela, COVID-19 ile ilgili atılan tweetlerden doktorların paylaştığı tweetler filtrelenip, sadece ilgili meslek grubunun bu konu üzerindeki görüşleri analiz edilebilir. Ya da bir kurumsal şirket, sosyal medyada kendisi ile ilgili yapılan paylaşımlarda hangi meslek gruplarının olumlu, hangi meslek gruplarının olumsuz görüşlerde bulunduğunu bilmek isteyebilir.

Bu araştırmada, Twitter kullanıcılarının paylaşımları kullanılarak meslekleri tahmin edilmeye çalışılmıştır. Yapılan literatür taramasında, Türkçe metinler üzerinde meslek tahmini çalışmasına rastlanmamıştır. Bu çalışmanın Türkçe tweetler üzerinde yapılması açısından bu anlamda bir ilk olduğu düşünülmektedir. Makalenin bundan sonraki bölümlerinde öncelikle literatürde yer alan diğer dillerdeki metinler üzerinde yapılmış meslek tespiti çalışmaları incelenecek, daha sonra bu çalışmada kullanılan Türkçe veri kümesi ve uygulanan metodoloji açıklanacak, son olarak araştırma kapsamında yapılan deneylerin sonuçları sunulacak ve gelecek çalışmalara yönelik görüşler paylaşılacaktır.

# 2. Literatür Taraması

Literatürde yer alan sosyal medyada meslek tahmini üzerine yapılan çalışmaların sayısının yaş, cinsiyet, lokasyon, vb. özelliklerin tahminine yönelik çalışmalardan çok daha az olduğu görülmüştür. Bu bölümde, konuyla ilgili daha önce yayınlanmış olan az sayıdaki çalışma özetlenecektir.

Zhou ve diğerleri (2012) Çin'in en büyük mikroblog sitesi Sina Weibo[*]'dan 4 farklı sektörden 500'er kullanıcının verileri üzerinde bir sınıflandırma çalışması yapmıştır. Spor, eğlence, bilişim teknolojileri ve emlak sektörlerinden olan bu kullanıcıların kendi paylaştıkları, cevapladıkları, retweet ettikleri mesajlardan ve takipleşme ilişkilerinden çeşitli özellikler çıkarılmış ve Sıralı Minimal Optimizasyon (Sequential Minimal Optimization) sınıflandırıcı ile deneyler gerçekleştirilmiştir. Farklı özellik kümeleri ile yapılan deneylerde sınıf bazında en yüksek F-ölçüm değerlerinin 0,65-0,80 arasında değişkenlik gösterdiği raporlanmıştır.

Huang ve diğerleri (2015) yine Sina Weibo kullanıcıları üzerine bir meslek tahmini çalışması sunmuştur. Ulaşım, finans, devlet, eğitim, eğlence, elektronik, emlak, medya, sağlık, hizmet, kamu yararı ve diğer olmak üzere 12 sektörden farklı oranlarda toplanmış 65.828 kullanıcı hesabıyla geniş bir veri kümesi oluşturulmuştur. Özellik çıkarımında hem mesaj içeriklerinden hem de kullanıcılar arasındaki etkileşimlerden faydalanılmıştır. Naive Bayes, Karar Ağacı, Destek Vektör Makinesi, Lojistik Regresyon yöntemlerinin ve farklı özellik kümelerinin kullanıldığı deneyler sonucunda en iyi başarı 0,80 civarında F-ölçüm değeri olarak alınmıştır. Bu çalışmada, kullanıcıların kendilerine benzer kullanıcılarla etkileşim kurma eğiliminde olması durumunu ifade eden homofili (homophily) terimine özellikle vurgu yapılmış ve bunun meslek tahmini çalışmasında kullanılabileceği belirtilmiştir.

Preoţiuc-Pietro ve diğerleri (2015) tarafından yapılan araştırmada 9 farklı meslek grubundan 5.191 Twitter kullanıcısı üzerinde çalışma gerçekleştirilmiştir. Özellik olarak kullanıcının takipçi sayısı, takip ettiği hesap sayısı, toplam tweet sayısı, içinde etiket geçen tweetlerinin oranı, bir günde attığı ortalama tweet sayısı gibi kullanıcı seviyesindeki niteliklerin yanı sıra tweetlerin içerikleri de kullanılmıştır. Lojistik Regresyon, Destek Vektör Makinesi ve Gauss Süreci (Gaussian Process) yöntemlerinin uygulandığı testlerde en iyi sonuç, Gauss Süreci yöntemiyle %52,7 doğruluk değeri olarak elde edilmiştir. Pan ve diğerleri (2019) kendi çalışmaları için, aynı veri kümesindeki 4.557 kullanıcının takip ettiği kullanıcıları ve takipçilerini de toplamışlardır. Veri kümesindeki diğer kullanıcıların hesapları askıya alındığı veya özel (private) hesaba çevrildiği için profillerine erişilemediği belirtilmiştir. Ana kullanıcıların bu takipleşme ilişkilerinden çıkarılan komşuluk matrisi, Çizge Evrişimsel Ağ (Graph Convolutional Network) sınıflandırıcıya verilerek %61 doğruluk oranıyla Preoţiuc-Pietro ve diğerlerine göre daha yüksek performans alınmıştır.

Tu ve diğerleri (2015) de Sina Weibo sitesinden 62.415 kullanıcı hesabını toplayarak oluşturdukları veri kümesi üzerinde bir çalışma sunmuştur. PRISM (PRofession Identification in Social Media) adı verilen model ile 14 farklı meslek grubuna (medya, devlet, eğlence, emlak, finans, bilişim, spor, eğitim, moda, oyun, edebiyat, hizmet, sanat, sağlık) ait bu kullanıcıların meslekleri 0,82 F-ölçüm oranıyla tahmin edilebilmiştir. Kullanıcıların kendi yazdıkları biyografi açıklamaları, paylaştıkları mesajlar, mesajlardaki etiketler, mesajlardaki URL adresleri gibi bilgiler özellik çıkarımında kullanılmıştır. Eğitim verisini zenginleştirmek ve sınıflandırma performansını artırmak amacıyla, etiketlenmemiş 150.000 yeni kullanıcı hesabını otomatik etiketlemek için mevcut veri kümesi temel sınıflandırıcı

[*] https://weibo.com

olarak kullanılmıştır. Yarıdan fazla temel sınıflandırıcının, meslekleri konusunda hemfikir olduğu yeni kullanıcılar seçilerek ilgili meslek etiketiyle birlikte eğitim kümesine dâhil edilmiştir. Temel sınıflandırıcılar tekrar eğitilerek bu işlem yinelenmiştir.

Hu ve diğerleri (2016) yaptıkları çalışmada, 9.800 Twitter kullanıcısının paylaştıkları tweetler üzerinden 8 farklı mesleğin (pazarlama, yönetici, girişimci, editör/yazar, yazılım mühendisi, halkla ilişkiler, ofis memuru ve tasarımcı) farklı karakteristiklerini çıkarmışlardır. Daha sonra, özellik kümesi olarak tweetlerde geçen kelimeleri, ikili ve üçlü kelime gruplarını kullanarak sınıflandırma deneyleri gerçekleştirmişlerdir. Kullanıcıların %95'inden fazlasının kullandığı terimler ile %10'undan azının kullandığı terimler özellik kümesinden çıkarılmıştır. Yapılan deneylerde ortalama F-ölçüm değeri 0,78 olarak hesaplanmıştır.

Lv ve diğerleri (2017) bu konuda literatürde yer alan çalışmaların genelde meslek grupları ile ilgili olduklarını ve bu çalışmaların spesifik meslekler üzerinde yapılmadığı için gerçek hayattaki uygulamalar için yetersiz kaldıklarını belirtmiştir. Bu yüzden araştırmacılar kendi çalışmalarında spesifik mesleklerden kullanıcıların paylaşımlarını içeren bir veri kümesi oluşturmuşlardır. 8 farklı meslekten (yazar, muhabir, avukat, fotoğrafçı, aktör, şarkıcı, doktor ve diyetisyen) 8.000 Sina Weibo kullanıcısını rastgele seçmişler ve 100 ila 500'er paylaşımlarını toplamışlardır. Klasik özellik kümelerinin yanı sıra meslek odaklı sözlük tabanlı kelime temsillerini (word embedding) kullandıkları deneylerde en yüksek başarıyı %87,12 doğruluk oranıyla Destek Vektör Makinesi sınıflandırıcı ile elde etmişlerdir.

Meslek tespiti üzerine bugüne kadar yapılmış çalışmalara bakıldığında bunların çoğunun Sina Weibo sitesinin verilerini kullanıldığı görülmektedir. Bunun başlıca sebebi Sina Weibo'da onaylanmış hesaplar için meslek etiketinin zorunlu olmasıdır. Bu sayede onaylanmış olan tüm kullanıcıların meslekleri de otomatik olarak çekilebilmektedir. Öte yandan, Twitter'da hem normal kullanıcılar için hem de onaylanmış kullanıcılar için meslek bilgisi alanı yoktur. Kullanıcılar isterse mesleklerini biyografi açıklamalarında kendileri yazmaktadır. Dolayısıyla, Twitter kullanıcılarının meslek bilgilerine doğrudan erişmenin kolay bir yolu olmadığı için tweetlerle bir meslek veri kümesi oluşturmak daha meşakkatlidir.

Literatürde fotoğraflar üzerinden (Song vd., 2011; Shao vd., 2013; Chu & Chiu, 2014; Chu & Chiu, 2016) ve el yazısı görüntülerinden (Kumar vd., 2020) meslek tahmini çalışmaları da mevcuttur, ancak bu araştırmada metin verileri üzerinde çalışıldığı için bu bölümde söz konusu çalışmalara dair detay verilmesine gerek görülmemiştir.

# 3. Materyal ve Metot

Bu bölümde çalışmada kullanılan veri kümesi, veriler üzerinde uygulanan önişleme adımları, özellik çıkarımı ve deneylerde kullanılan sınıflandırma yöntemlerine dair bilgiler sunulmaktadır.

## 3.1. Veri Kümesi ve Önişleme

Literatürde Türkçe dili üzerine meslek tahmini çalışmasında kullanılabilecek kamuya açık olarak paylaşılmış bir veri kümesi bulunmamaktadır. Bu nedenle, bu araştırma için öncelikle bir veri kümesi oluşturulması gerekmiştir. Veri toplama kolaylığı nedeniyle, veri kaynağı olarak Türkiye'de en çok kullanılan sosyal medya ağlarından biri olan Twitter (T.C. Cumhurbaşkanlığı İletişim Başkanlığı, 2022) tercih edilmiştir.

Araştırmada kullanıcıların meslek grubunun değil, doğrudan mesleğinin tahmin edilmesi hedeflenmiştir. Bu nedenle, veri kümesi oluşturulurken öncelikle 10 spesifik meslek belirlenmiştir. Bu meslekler şunlardır: avukatlık, diyetisyenlik, doktorluk, ekonomistlik, öğretmenlik, psikologluk, spor yorumculuğu, tarihçilik, yazılımcılık, ziraat mühendisliği. Daha sonra Twitter'da her bir meslekten 5'er kullanıcı bulunmuştur. Bu aşamada, biyografi açıklamalarında açıkça mesleklerini belirten gerçek kullanıcılar seçilmiştir, parodi hesaplar veya sahte olabilecek kullanıcılara ait hesaplar tercih edilmemiştir. Son olarak, bu kullanıcıların paylaştıkları 500'er tweet toplanmıştır. Kullanıcıların tweetleri toplanırken güncel olandan eskiye doğru gidilmiştir. Sağlıklı bir veri kümesi oluşturmak amacıyla aşağıda belirtilen tweetler veri kümesine dâhil edilmemiştir:

- Türkçe yazılmamış olan tweetler
- İçeriği farklı bir kullanıcıya ait olan retweetler
- Karşılıklı bir sohbet akışının parçası olan ve tek başına bir anlam ifade etmeyen tweetler
- Sadece bağlantı (link) içeren tweetler
- Tekrar eden tweetler

Bu şekilde her bir meslekten 2.500 tweet olmak üzere toplam 25.000 Türkçe tweetten oluşan bir meslek veri kümesi ortaya çıkarılmıştır. Tweetlerdeki metinler küçük harfe çevrilmiş ve içerisinde geçen bağlantılar bu çalışma için anlamlı olmadığından temizlenmiştir. Çalışma kapsamında oluşturulan veri kümesi, bu alanda çalışma yapmayı düşünen diğer araştırmacıların da kullanabilmesi amacıyla kamuya açık olarak Github'da paylaşılmıştır[†].

Deneylerden önce tweetlerdeki tüm noktalama işaretleri silinmiştir. Kelimelerin köklerini bulmak için açık kaynak kodlu Türkçe doğal dil işleme kütüphanesi olan Zemberek (Akın & Akın, 2007) kullanılmıştır.

## 3.2. Özellik Çıkarımı ve Sınıflandırma

Sınıflandırma aşamasında özellik olarak hem kelimelerin kendileri hem de kökleriyle ayrı ayrı deneyler gerçekleştirilmiştir. Özellik kümesi oluşturulurken doğal dil işleme çalışmalarında yaygın olarak kullanılan TF-IDF (Term Frequency-Inverse Document Frequency/Terim Frekansı-Ters Doküman Frekansı) metin temsil yöntemi tercih edilmiştir. Sınıflandırma başarısını yükseltmek amacıyla özellik seçimi yapılarak, mevcut özelliklerin yarısı ile deneyler tekrar edilmiştir. Özellik seçimi için Ki-kare (Chi-squared) yöntemi kullanılmıştır.

Twitter'da paylaşılan her bir tweet için 280 karakter sınırı vardır. Söz konusu limitin tamamının kullanılması durumunda Türkçe yazılmış anlamlı bir tweet en fazla 40 civarında kelime içerebilmektedir. Çalışma kapsamında oluşturulan veri kümesinde, önişleme adımlarından sonra her bir tweette (kelimeler arasındaki boşluklar dâhil) ortalama 147,8 karakter ve 20,7 kelime mevcuttur. Öte yandan, bazı tweetler sadece birkaç kelimeden oluşmaktadır. Bu durum da tweetlerin metin sınıflandırma çalışmalarında kullanılmasının en önemli kısıtı olarak ifade edilmektedir. Çünkü kısa metinlerin sınıflandırılması uzun metinlere göre genellikle daha zordur. Meslek tahmini

[†] https://github.com/imayda/occupation-dataset-in-turkish

üzerinden düşünülecek olursa, kullanıcının paylaştığı tek bir tweete bakarak, onun hangi meslekten olduğunu anlamak oldukça güçtür. Bir tweet üzerinden meslek tahmini yapmaya çalışmak yerine kullanıcının birkaç tweetine birlikte bakmak bu işi çok daha kolaylaştırır. Bu nedenle, kullanıcılara ait tweetler hem tekil olarak hem de 5'li ve 10'lu gruplar halinde birleştirilerek deneylerde kullanılmıştır. Birleştirme yapılırken art arda gelen tweetler bir araya getirilerek, testlerde tek bir veri olarak değerlendirilmiştir.

Sınıflandırma aşamasındaki deneylerde, literatürdeki çalışmalarda da sıkça tercih edilen Destek Vektör Makinesi ve Lojistik Regresyon yöntemleri kullanılmıştır. Testler Google Colab ortamında Python dili ile yazılmış, kullanılan algoritmalar için açık kaynak kodlu makine öğrenmesi kütüphanesi olan Scikit-learn[‡]'den faydalanılmıştır. Destek Vektör Makinesi yönteminde parametre olarak lineer kernel kullanılırken, Lojistik Regresyon yönteminde maksimum iterasyon sayısı 200 olarak belirlenmiştir. Makine öğrenmesi yöntemlerinin diğer parametreleri değiştirilmeden kütüphanedeki varsayılan değerleriyle çalıştırılmıştır. Tüm deneyler 10-katlı çapraz doğrulama yapılarak gerçekleştirilmiştir. Bu iki yöntem haricinde K-En Yakın Komşu algoritması farklı komşuluk değerleriyle (10, 20, 30) ve Rastgele Orman algoritması farklı tahmin edici sayıları (10, 50, 100) ile test edilmiş ancak sınıflandırma başarıları daha düşük olduğu için bu testlerin sonuçları makalede sunulmamıştır.

## 4. Araştırma Sonuçları ve Tartışma

Kelimelerin kendilerinin kullanıldığı deneylerde 86.218, kelimelerin köklerinin kullanıldığı deneylerde ise 29.375 terimden oluşan özellik kümeleri makine öğrenmesi yöntemlerine verilmiştir. Öncelikle her bir tweet tek başına bir veri olarak testlerde kullanılmıştır. Bu testlerin sonuçları Tablo 1'de sunulmuştur. Bu sonuçlarda kelime köklerinin kullanıldığı testlerde daha yüksek başarı elde edildiği ve özellik seçiminin başarıya olumlu etki ettiği görülmüştür. Tekil tweetlerin kullanıldığı deneylerde en yüksek doğruluk oranı %74,90 olmuştur.

*Tablo 1. Tekil tweetler ile yapılan deneylerin sonuçları*

| Algoritma | Özellik Oranı | Kelimelerin kendileriyle | | Kelimelerin kökleriyle | |
|---|---|---|---|---|---|
| | | Doğruluk (%) | F-ölçüm | Doğruluk (%) | F-ölçüm |
| Destek Vektör Makinesi | %100 | 73,59 | 0,74 | 74,05 | 0,74 |
| | %50 | 74,34 | 0,74 | 74,36 | 0,74 |
| Lojistik Regresyon | %100 | 72,86 | 0,73 | 74,69 | 0,75 |
| | %50 | 73,43 | 0,74 | **74,90** | 0,75 |

Daha sonra, kullanıcıların paylaştıkları tweetler art arda 5'li ve 10'lu gruplar halinde birleştirilerek tek bir veri olarak değerlendirilmiş ve böylece daha uzun veriler elde edilmiştir. Bu şekilde tekrar edilen deneylerin sonuçları Tablo 2 ve 3'te verilmiştir. Verinin içeriğinin büyümesi başarıyı ciddi oranlarda artırmıştır. 5'li gruplar halinde birleştirilmiş tweetlerle en yüksek %96,20 doğruluk oranı elde edilirken, bu başarı 10'lu gruplar halinde birleştirilmiş tweetlerle yapılan deneylerde %99,00'a kadar yükselmiştir. Bu deneylerde de kelime köklerinin kullanılması ve özellik seçimi daha iyi sonuçlar alınmasını sağlamıştır. Ayrıca, bu testlerde Destek Vektör Makinesi yönteminin Lojistik Regresyon yöntemine göre daha başarılı olduğu görülmüştür.

*Tablo 2. 5'li gruplar halinde birleştirilmiş tweetlerle yapılan deneylerin sonuçları*

| Algoritma | Özellik Oranı | Kelimelerin kendileriyle | | Kelimelerin kökleriyle | |
|---|---|---|---|---|---|
| | | Doğruluk (%) | F-ölçüm | Doğruluk (%) | F-ölçüm |
| Destek Vektör Makinesi | %100 | 95,54 | 0,96 | 95,86 | 0,96 |
| | %50 | 95,96 | 0,96 | **96,20** | 0,96 |
| Lojistik Regresyon | %100 | 94,26 | 0,94 | 95,22 | 0,95 |
| | %50 | 94,88 | 0,95 | 95,30 | 0,95 |

*Tablo 3. 10'lu gruplar halinde birleştirilmiş tweetlerle yapılan deneylerin sonuçları*

| Algoritma | Özellik Oranı | Kelimelerin kendileriyle | | Kelimelerin kökleriyle | |
|---|---|---|---|---|---|
| | | Doğruluk (%) | F-ölçüm | Doğruluk (%) | F-ölçüm |
| Destek Vektör Makinesi | %100 | 98,80 | 0,99 | 98,92 | 0,99 |
| | %50 | 98,84 | 0,99 | **99,00** | 0,99 |
| Lojistik Regresyon | %100 | 98,16 | 0,98 | 98,44 | 0,98 |
| | %50 | 98,36 | 0,98 | 98,64 | 0,99 |

[‡] https://scikit-learn.org

Tablo 3'te görüldüğü gibi, veri kümesindeki 10 meslekten birine sahip bir Twitter kullanıcısına ait sadece 10 tweet ile o kullanıcının mesleğini yüksek bir başarıyla tahmin etmek mümkündür. Bu çalışmada %99 gibi yüksek bir sınıflandırma başarısının alınmasının birkaç sebebi vardır. Bunlardan birincisi meslek sayısının 10 ile sınırlı tutulmasıdır. Ancak, gerçek hayatta yüzlerce meslek olduğu bilinmektedir. Örneğin, Wikipedia'nın meslekler listesinde[§] 754 meslek bulunmaktadır. Veri kümesindeki meslek sayısı arttırıldıkça bu tahmin başarısı da düşecektir. Alınan yüksek başarının bir diğer sebebi de veri kümesi oluşturulurken seçilen 10 mesleğin birbirine yakın alanlardan olmamasıdır. Bir avukat ile bir doktorun çok fazla ortak ilgi alanı olması beklenmez. Öte yandan, bir avukat ile bir savcının veya bir doktor ile bir hemşirenin çok daha fazla ortak ilgi alanları vardır. Birbirine yakın mesleklerin verileri birbiriyle daha çok karıştırılacağından, aynı sektör içinde faaliyet gösteren farklı mesleklerden kullanıcıların ayırt edilmesi daha zordur. Bu nedenle, veri kümesinde aynı sektörden mesleklerin bulunması da sınıflandırma başarısını düşürecektir. Ayrıca, yapılan deneylerde eğitim ve test verileri rastgele seçildiği için, bir kullanıcıya ait tweetlerin bir kısmı eğitimde bir kısmı testte kullanılmıştır. Bu durumda, kullanıcının tweet yazma tarzındaki ayırt edici özellikler de sınıflandırma başarısına olumlu etkide bulunmuştur. Örneğin, veri kümesindeki diğer kullanıcılara oranla çok fazla emoji kullanan bir kullanıcının test için ayrılan tweetleri bu farklılıktan dolayı daha kolay tahmin edilebilir. Daha sağlıklı test sonuçları için tweetleri rastgele olarak eğitim ve test için ayırmak yerine doğrudan kullanıcıları rastgele olarak eğitim ve test için ayırmak daha doğru olabilir. Ancak bu çalışmada böyle yapılmamıştır, bu da çalışmanın zayıf yönlerinden biridir.

## 5. Sonuç

Bu çalışmada, sosyal medyada paylaşım yapan kullanıcıların mesleğinin makine öğrenmesi yöntemleri ile tahmin edilmesi amaçlanmıştır. Bunun için öncelikle 10 farklı meslekten Twitter kullanıcılarının tweetlerini içeren Türkçe meslek veri kümesi oluşturulmuş ve bu veri kümesi kamuya açık olarak paylaşılmıştır. Deneylerde Destek Vektör Makinesi ve Lojistik Regresyon yöntemleri uygulanmış, özellik kümesi olarak hem kelimelerin kendileri hem de kökleri kullanılmıştır. Başarıyı arttırmak için özellik seçimi yapılmış ve özelliklerin yarısı ile testler tekrar edilmiştir. Test sonuçlarında kelime köklerinin kullanılmasının kelimelerin kendilerini kullanmaya göre daha iyi performans gösterdiği ve özellik seçiminin başarıya olumlu etkisi olduğu görülmüştür. Bu çalışmada en iyi sonuç, 10'lu gruplar halinde birleştirilmiş tweetlerle yapılan deneylerde kelime köklerinin kullanıldığı ve özellik seçiminin yapıldığı Destek Vektör Makinesi testinde %99,00 doğruluk oranı olarak elde edilmiştir.

Gelecek çalışmalarda daha fazla meslekten daha çok sayıda kullanıcının olduğu daha kapsamlı bir veri kümesi oluşturulması hedeflenmektedir. Veri kümesindeki meslek ve kullanıcıların sayısı arttıkça sınıflandırma başarısının düşmesini önlemek için literatürdeki çalışmalarda yapıldığı gibi kullanıcıların sosyal medyadaki bağlantılarının ve etkileşimlerinin de veri kümesine dâhil edilmesi gerekmektedir. Zira homofili prensibine göre sosyal medya kullanıcıları kendisiyle benzer kullanıcıları takip etmeye ve onlarla etkileşim kurmaya daha meyillidir. Buradaki benzerlik tanımında mesleğin de önemli bir yeri vardır. Dolayısıyla kullanıcıların sosyal medyadaki bağlantıları ve etkileşimleri mesleklerini tahmin etmede önemli ipuçları vermektedir.

Veri kümesi boyutunun arttırılmasının yanı sıra deneylerdeki başarı oranını yükseltmeye yönelik olarak gelecek çalışmalarda farklı yöntemlerin uygulanması da planlanmaktadır. Sosyal medya paylaşımlarında yazım hatalarına sık rastlandığı için önişleme adımlarında bu hataların düzeltilmesi başarıyı olumlu yönde etkileyebilir. Özellik kümesini zenginleştirmek için karakter ve kelime n-gramları veya FastText, Glove gibi Türkçe için çıkarılmış kelime vektörleri kullanılabilir. Son olarak sınıflandırma aşamasında LSTM, CNN, GRU veya BERT gibi derin öğrenme modelleri test edilebilir.

## 6. Teşekkür

Veri toplama aşamasındaki desteklerinden dolayı Murat Karabulut'a teşekkür ediyorum.

## Kaynakça

Akın, M. D., & Akın, A. A. (2007). Türk Dilleri İçin Açık Kaynaklı Doğal Dil İşleme Kütüphanesi: Zemberek. *Elektrik Mühendisliği, 431*, 38-44.

Chu, W., & Chiu, C. (2014, Aralık). *Predicting Occupation from Single Facial Images*. IEEE International Symposium on Multimedia, Taichung, Tayvan. https://doi.org/10.1109/ISM.2014.13

Chu, W., & Chiu, C. (2016). Predicting Occupation from Images by Combining Face and Body Context Information. *ACM Transactions on Multimedia Computing, Communications, and Applications, 13*(1), 1-21. https://doi.org/10.1145/3009911

Hu, T., Xiao, H., Luo, J., & Nguyen, T. T. (2016, Mayıs). *What the Language You Tweet Says About Your Occupation*. The Tenth International AAAI Conference on Web and Social Media (ICWSM), Köln, Almanya. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13020

Huang, Y., Yu, L., Wang, X., & Cui, B. (2015). A multi-source integration framework for user occupation inference in social media systems. *World Wide Web, 18*, 1247-1267. https://doi.org/10.1007/s11280-014-0300-6

Kepios. (2022, Temmuz). Global Social Media Statistics. https://datareportal.com/social-media-users

Kumar, P., Gupta, M., Gupta, M., & Sharma, A. (2020). Profession Identification Using Handwritten Text Images. *Computer Vision and Image Processing (CVIP 2019), Communications in Computer and Information Science, 1148*, 25-35. https://doi.org/10.1007/978-981-15-4018-9_3

Lv, X., Jin, P., Mu, L., Wan, S., & Yue, L. (2017). Detecting User Occupations on Microblogging Platforms: An Experimental Study. *Web and Big Data, APWeb-WAIM 2017, Lecture Notes in Computer Science (LNCS), 10366*, 331-345. https://doi.org/10.1007/978-3-319-63579-8_26

Pan, J., Bhardwaj, R., Lu, W., Chieu, H. L., Pan, X., & Puay, N. Y. (2019, Temmuz). *Twitter Homophily: Network Based Prediction of User's Occupation*. The 57th Annual Meeting of the Association for Computational Linguistics, Floransa, İtalya. http://doi.org/10.18653/v1/P19-1252

[§] https://tr.wikipedia.org/wiki/Meslekler_listesi

Preoţiuc-Pietro, D., Lampos, V., & Aletras, N. (2015, Temmuz). *An analysis of the user occupational class through Twitter content*. The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Pekin, Çin. http://doi.org/10.3115/v1/P15-1169

Shao, M., Li, L., & Fu, Y. (2013, Aralık). *What Do You Do? Occupation Recognition in a Photo via Social Context*. IEEE International Conference on Computer Vision (ICCV), Sidney, Avustralya. https://doi.org/10.1109/ICCV.2013.451

Song, Z., Wang, M., Hua, X., & Yan, S. (2011, Kasım). *Predicting Occupation via Human Clothing and Contexts*. IEEE International Conference on Computer Vision (ICCV), Barselona, İspanya. https://doi.org/10.1109/ICCV.2011.6126355

Statista. (2022, Ocak). Most popular social networks worldwide as of January 2022, ranked by number of monthly active users. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

T.C. Cumhurbaşkanlığı İletişim Başkanlığı. (2022, Mayıs). Sosyal Ağ Haritası, Twitter Kullanım Raporu. http://sosyalagharitasi.gov.tr/report

Tu, C., Liu, Z., & Sun, M. (2015). PRISM: Profession Identification in Social Media with Personal Information and Community Structure. *Social Media Processing (SMP 2015), Communications in Computer and Information Science, 568*, 15-27. https://doi.org/10.1007/978-981-10-0080-5_2

Zhou, M., Xu, Y., & Zhao, X. (2012, Aralık). *Study of Feature Extract on Microblog User Occupation Classification*. Fourth International Symposium on Information Science and Engineering (ISISE), Şangay, Çin. https://doi.org/10.1109/ISISE.2012.14