NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., 2023; 12(2), 325-333

Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences

Araştırma makalesi / Research article

www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

Türkçe sosyal medya mesajlarından kullanıcıların yaş ve cinsiyetini tahmin etme

Predicting users age and gender using Turkish social media messages

Mustafa Kaan Görgün[1], Gökçe Başak Demirok[2], Mücahid Kutlu[3,*]

[1,3] *TOBB Ekonomi ve Teknoloji Üniversitesi, Bilgisayar Mühendisliği Departmanı, 06510, Ankara Türkiye*
[2] *TOBB Ekonomi ve Teknoloji Üniversitesi, Yapay Zekâ Mühendisliği Departmanı, 06510, Ankara Türkiye*

Öz

Sosyal medya platformları insanların herhangi bir konu hakkındaki fikirlerine dair çok yüksek miktarda veri sunmaktadır. Bu yüzden, bu tip platformlar market analizi ve toplumsal görüş tahmini gibi birçok çalışma için çok önemli veri kaynaklarıdır. Ancak, sosyal medya kullanıcıları bir toplumu tam anlamıyla yansıtmadığından ötürü sosyal medya verisindeki yanlılığı azaltmak için kullanıcıların yaşı ve cinsiyeti gibi çeşitli bilgileri de göz önünde bulundurarak sayma işlemi gibi ek adımların atılması gerekmektedir. Bu çalışmada verilen bir Türkçe Twitter hesabının paylaştığı mesajları kullanarak hesap sahibinin yaş aralığını ve cinsiyetini tahmin etme problemi konusunu ele aldık. Çalışma kapsamında 1040 Twitter kullanıcısının yaş ve cinsiyet bilgilerinden oluşan etiketli bir veri kümesi hazırlanmıştır. Ardından kelime, karakter, retweet, fastText ve BERT tabanlı beş farklı yöntem geliştirilmiştir. Yaptığımız kapsamlı deneylerden kullanıcıların paylaştıkları mesajların insanların yaş ve cinsiyet bilgisine dair önemli ipuçları sunduğunu göstermektedir.

Keywords: Doğal dil işleme, Yaş tahmini, Cinsiyet tahmini, Yazar profili tahmini

Abstract

Social media platforms provide a huge amount of data on people's opinions on any topic. Therefore, such platforms are very important data sources for many studies such as market analysis and social opinion prediction. However, since social media users do not fully reflect a society, it is necessary to take additional steps to reduce bias such as weighted counting based on users' age and gender. In this study, we focus on the problem of predicting the age range and gender of the owner of a given Twitter account using the shared messages in Turkish. Within the scope of the study, we constructed a labeled dataset consisting of age and gender information of 1040 Twitter users. In addition, we developed five different methods based on words, characters, retweets, fastText, and BERT. Our extensive experiments show that the messages shared by users offer important clues about people's age and gender information..

Anahtar kelimeler: Natural language processing, Age prediction, Gender prediction, Author profiling

1 Giriş

Sosyal medya platformları kullanıcılarına fiziki uzaklıklara bakmaksızın dünyanın herhangi bir yerinde yaşayan diğer insanlarla kolaylıkla etkileşime geçme ve diledikleri konuda fikirlerini beyan etme imkânı sunmaktadır. Sağladıkları bu müthiş iletişim olanağından ötürü birçok insan aktif bir şekilde bu platformları kullanmaktadır. Örneğin, Twitter 2022 yılının ilk çeyreğinde yaklaşık olarak günlük 229 milyon aktif kullanıcısı olduğunu bildirmiştir [1]. Türkiye'de yaşayan insanlar da bu popüler platforma çok büyük bir ilgi göstermiştir. Türkiye 16.1 milyon kullanıcısı ile dünyada en çok kullanıcısı olan 6. ülke konumundadır [2].

Sosyal medya platformlarının sağladığı bu etkin iletişim imkânı ayrıca insanların çeşitli konularda ne düşündüğüne dair çok zengin bir veri kaynağı da oluşturmaktadır. İnsanların sosyal medya mesajları market analizi, spesifik bir konuda toplumsal görüş hakkında anket çalışması [3] veya toplumsal polarizasyonu analiz etmek [4] gibi birçok alanda kullanılmaktadır.

Sosyal medya platformları her ne kadar birçok araştırma çalışması için heyecan verici bir veri kaynağı olsa da, sağladığı verileri işlerken ve analiz ederken çok dikkatli olmak gerekmektedir. Öncelikle, sosyal medya platformları programlar tarafından yönetilen bot hesaplar içerebilir [5] ve bu bot hesaplar belli bir konuda birçok mesaj üreterek yapay gündem oluşturmada kullanılabilirler. Ayrıca sosyal medya kullanıcıları toplumun bir biçimli ("uniform") bir örneklemi olmayıp, toplumu yansıtma açısından çok ciddi bir oranda yanlılık ("bias") göstermektedir [6]. Bu yüzden sosyal medya verisinden anlamlı çıkarımlar yapabilmek için verideki yanlılığın azaltılması büyük bir önem arz etmektedir.

Sosyal medya verilerindeki yanlılığı azaltmak için, sağladığı içeriği olduğu gibi analiz etmek yerine, kullanıcılar üzerinde belli analizler yapıp, kullanıcı profillerinin toplumdaki ağırlığına göre bir çıkarım yapmak daha sağlıklı olacaktır [7]. Örneğin, Dwi Prasetyo ve Hauff [3] kullanıcıların cinsiyet bilgisini Endonezya seçim sonuçlarını sosyal medya üzerinden tahmin ederken kullanmıştır. Ancak kullanıcılar yaş ve cinsiyet bilgilerini direkt olarak

* Sorumlu yazar / Corresponding author, e-posta / e-mail: m.kutlu@etu.edu.tr (M. Kutlu)
Geliş / Recieved: 19.10.2022 Kabul / Accepted: 27.02.2023 Yayımlanma / Published: 15.04.2023
doi: 10.28948/ngmuh.1191719

profillerinde yazmadıklarından, bu bilgileri kullanıcıların paylaştıkları mesajlardan veya sosyal medya hesaplarındaki diğer bilgilerinden tahmin etmek gerekmektedir.

Metinlerin yazarlarının yaş ve cinsiyetini tahmin etme konusu birçok araştırmacının ilgisini çekmiştir. Özellikle PAN ortak-görevinde ("shared-task") birçok kez yaş ve cinsiyet tahmini konusu ele alınmıştır [8]. Ancak maalesef bu konuda Türkçe üzerine yapılan çalışmalar oldukça kısıtlıdır.

Bu çalışmada Türkçe Twitter hesaplarının paylaştığı sosyal medya mesajları kullanılarak yaş ve cinsiyet bilgisini tahmin etme problemi ele alınmıştır. Bu konudaki veri kümesi eksikliğinden ötürü, öncelikle 1040 kullanıcıdan oluşan etiketli bir veri kümesi oluşturulmuştur. Ardından her iki problem için de beş farklı model geliştirilmiştir. Bu modeller 1) kelime bazlı n-gram modeli, 2) karakter bazlı n-gram modeli, 3) retweet temelli sınıflandırma, 4) fastText [9] kelime vektörlerini ("word embedding") kullanarak sınıflandırma ve 5) hassas ayar ("fine-tune") yapılmış BERT [10] modelidir. Ayrıca BERT modelinin işleyebileceği kelime sayısı kısıtlı olduğundan, profillerin analizlerinde hangi tweet'lerinin kullanılması gerektiğine dair bir tweet seçim yöntemi önerdik.

Yaptığımız kapsamlı deneylerde, yaş tahmini probleminde son mesajlar kullanılarak hassas-ayar işleminden geçirilen BerTurk modeli ile en yüksek başarım elde edilmiştir. Cinsiyet tahmininde ise kadın ve erkeklerin sık kullandığı kelimelere göre seçilen mesajları kullanan BerTurk modeli diğer modellerden daha yüksek başarım elde etmiştir.

Bu çalışmanın temel olarak iki tane önemli katkısı bulunmaktadır. İlk olarak Türkçe'de yaş ve cinsiyet tahmini problemi için etiketli bir veri kümesi oluşturulmuştur. Bu veri kümesi araştırmacılar ile paylaşılacak olup, bu konudaki gelecek çalışmalar için önemli bir kaynak oluşturulmuştur. İkinci olarak, yaş ve cinsiyet tahmini için beş farklı yöntem geliştirilmiş olup, başarımları ölçümlenmiştir. Geliştirilen modellerin kodları paylaşılacak olup, deney sonuçlarımızın tekrarlanabilirliği sağlanacaktır.

Çalışmamızın geri kalan kısımlarında şu konular anlatılacaktır. Öncelikle, yaş ve cinsiyet tahmini üzerine literatürdeki ilgili çalışmalar sunulacaktır. Sonrasında oluşturduğumuz etiketli veri kümesini oluşturma yöntemi ve veri kümesine ait istatistiki bilgiler paylaşılacaktır. Ardından her iki problem için geliştirdiğimiz modeller sunulacaktır. En sonunda deneysel değerlendirmeler ile yöntemler karşılaştırılacak ve sonuçlar tartışılacaktır.

2 Literatür taraması

Yaş ve cinsiyet tahmini genel olarak yazar profili çıkarma problemlerindendir. O yüzden öncelikle yazar profili çıkarma konusundaki çalışmalara değineceğiz. Yazar profili çıkarma konusundaki çalışmalar birçok farklı metin türü ve kaynağını kullanmışlardır. Örneğin, Facebook mesajları [11], blog yazıları [12], e-postalar [13] ve tweet'ler [14] kullanılan metinler arasındadır. Biz çalışmamızda tweet'leri kullandık çünkü Twitter akademik araştırma için verisini kullanıma açmıştır ve diğer platformlar ile karşılaştırıldığında verisini paylaşma konusunda çok daha fazla kolaylıklar sağlamaktadır. Muhtemelen bu sebepten ötürü, literatürdeki çalışmaların büyük bir kısmı da Twitter'ı bir veri kaynağı olarak kullanmıştır.

Yazar profili çıkarma konusunda araştırmacılar yaş, cinsiyet, politik görüş, konum ve karakter tipi gibi birçok farklı konuyu ele almıştır. Örneğin, Rao vd. [15] profillerin cinsiyet, yaş, politik eğilim ve nereli olduğunu tahmin etmeye çalışmıştır. Flekova vd. [16] ise insanların kullandıkları dilden gelirlerini ve yaşlarını tahmin etmeye çalışmıştır. Schwartz vd. [11] ise yaş, cinsiyet ve kişilik tahmini konusuna eğilmiştir.

Yazar profilini çıkarma konusundaki çalışmalar ayrıca kullandıkları doğal dillere göre de ayrıştırılabilir. Literatürde Almanca [17], Felemenkçe [18], Yunanca [19,20], İngilizce [12], İspanyolca [12] ve Arapça [21] gibi birçok dilde çalışmalar mevcuttur. Türkçe'de de çeşitli konularda çalışmalar olmakla birlikte, maalesef diğer dillere göre oldukça az sayıda çalışma olduğunu söyleyebiliriz.

2.1 Yaş ve cinsiyet tahmini için etiketli veri kümeleri

Modelleri eğitmek ve test etmek için muhakkak etiketli veri kümelerine ihtiyacımız olduğundan birçok araştırmacı çeşitli yazar profili veri kümeleri oluşturmuştur. Bu veri kümelerinde Facebook mesajları [11], bloglar [12], elektronik postalar [13] ve tweet'ler [14] gibi farklı metin türlerini kullanmışlardır. Bu konuda ne yazık ki Türkçe veri kümeleri de oldukça sınırlı sayıdadır.

Wiegmann vd. [22] 37 farklı dilde toplam 71,706 ünlü kişinin Twitter hesabından oluşan bir veri kümesi oluşturmuştur. Ancak bu veri kümesinin çok küçük bir kısmı Türkçe hesaplardan oluşmaktadır. Cinsiyet tahmini için Sezerer vd. [23] Türk Dil Kurumu'ndan aldıkları cinsiyetler için kullanılan isimler listesindeki isimleri içeren profilleri toplayarak toplamda 5,292 kullanıcıyı içeren bir veri kümesi oluşturmuşlardır. Bu çalışmada ise, geneli ünlülerden olmak üzere 1,040 Twitter kullanıcısının yaş, cinsiyet ve meslek bilgilerini içeren bir veri kümesi oluşturduk. Bu açıdan literatürde bu üç konuyu da içeren ve bu büyüklükte olan başka bir veri kümesi bildiğimiz kadarı ile yoktur.

2.2 Yaş ve cinsiyet tahmini çalışmaları

Bu çalışmada olduğu gibi, geçmiş çalışmalar da genelde yaş ve cinsiyet tahmini konularının ikisini de kapsamaktadır. Örneğin, PAN 2013'te [24] ve sonraki bazı PAN ortak-görevlerinde de yaş ve cinsiyet tahmini ele alınmıştır. Biz de bu yüzden bu konulardaki çalışmaları birlikte anlatacağız.

Erkek ve kadınların yazım stillerinde ve bazı ifadelerin kullanım sıklığında farklılık olduğuna dair literatürde çeşitli bulgular bulunmaktadır. Örneğin Schwartz vd. [11] kadınların erkeklere nazaran başkalarının eşlerinden daha çok bahsederken, erkeklerin ise kendi eşlerinden daha sık bahsettiğini gözlemlemiştir. Ayrıca kadınların daha çok duygusal ifadeler ("seni seviyorum" vb.) ve birinci çoğul kişi çekimi kullanırken erkeklerin ise daha fazla argo içerikli kelime kullandığını belirtmiştir. Ayrıyeten, Rao vd. [15] kadınların "my husband" (Tr: kocam) ifadesini kullanma sıklığının erkeklerin "my wife" (Tr: karım) ifadesini kullanma sıklığından daha fazla olduğunu belirtmiştir. Park vd. [25] içerik özelliklerine göre cinsiyet tahmini yapmış ve

Tablo 1. Veri kümemizden örnek hesaplar ve tweet'leri

Kullanıcı ID	Yaş	Cinsiyet	Meslek	Örnek Tweet'ler
1361659642896654336	61+	Erkek	Siyasetçi	Tweet-1: Sizlerle benim hayat felsefemin bir parçası olmuş bir şiir paylaşmak istedim. #adamolmak #rudyardkipling Tweet-2: Oğullarım Emrah, Emre & Fatih'le çektirdiğimiz bu fotoğraf #TürkiyeninDönüşümYılları adlı kitabımın son bölümünde yer alıyor. Bu karede olmayan Selin'le birlikte her birinin varlığı, hayatta aldıkları yola şahitlik & eşlik etmek beni bir baba olarak çok mutlu ediyor. #BabalarGunu
569849244	61+	Kadın	Sanatçı	Tweet-1: Bütün kızlar toplandık geliyoruzzzzz...😍👧 kanalturk @bbturanli @bulbulmustafa @eeceyilmazz Tweet-2: İyi haftalar... #istanbulboğazı #toplantızamanı
466053456	18-30	Kadın	Sporcu	Tweet-1: Biz Ne Baharlar gördük , senle Ne kışlar. Hiçbir şeyi sevmedik inan senin kadar.. □ #besiktas Hakkari deplasmanı için yoldayız, Tüm taraftarlarımızı bekliyoruz Tweet-2: Kadın A Milli Takım Hazırlıklar Devam Ediyor Türkiye-Hollanda 08.11.2019 https://t.co/JntRiMXFaS
83000123	31-40	Erkek	Oyuncu	Tweet-1: Mühim olan artık buranın gediklisi olmak. Her sene burda olmalıyız !!!! 🖤❤ #FinalFour #FenerbahçeÜlker Tweet-2: Mükemmel galibiyet tebrikler beyler 🖤❤🖤❤🖤❤🖤 #Fenerbahce

kadınların mesajlarında daha çok pozitif duygu ve sosyal ilişki belirten kelimelerin (örn: arkadaşlar, aile, kız kardeş) ve emojilerin olduğunu; erkeklerin mesajlarında ise daha çok politika (örn: hükümet, vergi), spor ve yarışma (örn: futbol, sezon, kazanmak) konularında kelimelerin olduğunu gözlemlemiştir. Newman vd. [26] erkeklerde kelime uzunluğu, sayı ve edat kullanımının kadınlara göre fazla olduğunu, kadınların ise erkeklere göre daha fazla zamir ve sosyal kelime kullandıklarını belirtmiştir.

Yaşın insanların yazdıkları metinler üzerindeki etkisi hakkında da çeşitli çalışmalar vardır. Örneğin, Schwartz vd. [11] gençlerin "idk" ("I do not know") gibi sosyal medyada yaygın ifadeleri daha sık kullandığını ve beklenildiği üzere 19-22 yaş aralığında üniversite konularının daha çok rastlandığını belirtmişlerdir. Nguyen vd. [18] yaş arttıkça genellikle tweet uzunluğu, kelime uzunluğu, verilen link ve hashtag sayısının arttığını, "ben" zamirinin kullanımının ise azaldığını bildirmişlerdir. Pennebaker ve Stone [27] insanlar yaşlandıkça kendine referansın daha az kullanıldığını, pozitif kelime kullanımının ve gelecek zaman kullanımının arttığını bildirmektedir. Ancak Brandt ve Herzberg [28] yaşlı katılımcıların uygulamalarında daha fazla olumlu duygu kelimesi kullanılmadığını belirtmiş olup, Pennebaker ve Stone'nun [27] bulgusunu desteklememektedir.

Yaş ve cinsiyet tahmini için literatürde çeşitli yöntemler geliştirilmiştir. Nguyen vd. [29] yaş tahminini regresyon problemi olarak tanımlarken, Nguyen vd. [18], kategorik ve yaşam basamakları şeklinde ele almıştır. Özellikle Nguyen vd. [29] yaş tahmini yaparken cinsiyet bilgisini de öznitelik olarak kullanmıştır. Schwartz vd. [11] cinsiyet, yaş ve karakter tespiti için açık-sözlük ("open vocabulary") yaklaşımını kullanmıştır. Geçmiş bilgiye dayanmayan, sözlük kullanılmayan, sadece o anki veriden çıkarılan bu yaklaşımın, klasik kapalı-sözlük ("closed-vocabulary)" yöntemine göre daha çok bilgi içeren sonuçlar verdiğini belirtmişlerdir. Rao vd. [15] Twitter kullanıcılarının cinsiyet, yaş, bölgesel köken ve politik yönelimlerini tahmin etmek için sosyo-dilsel öznitelikleri ve n-gram modeli kullanan iki ayrı destek vektör makinesi ("support vector machine" - SVM) modelinin tahminlerini birleştiren bir model geliştirmişlerdir. Hirt vd. (2019) ise diğer çalışmalardan ayrı olarak kullanıcı ismi ve profil fotoğrafını da cinsiyet tahmininde kullanmıştır.

Literatürde Türkçe metinler üzerinden cinsiyet tahmini yapan çalışmalar da mevcuttur. Örneğin, Sezerer vd. [30] cinsiyet tahminini sadece yazım stili inceleyerek yapmışlardır. Bunun için standart makine öğrenimi yaklaşımları yerine stil özelliklerini çıkarmak için evrişimli sinir ağları ("convolutional neural network") kullanmışlardır. Sezerer vd. [23] de cinsiyet tahmini için oluşturduğu veri kümesinde SVM ile kelime-torbası yönteminin sonuçlarını paylaşmıştır. İlhami ve Hanbay [31] Sezerer vd.'nin [23] veri kümesi üzerinde BERT, DistilBert ve Electra gibi dönüştürücü modelleri, LSTM ve CNN gibi derin öğrenme modelleri ile SVM modellerini karşılaştırmıştır. Cinsiyet tespitinde yazılan içerikten çok yazım tarzının daha belirleyici bir özellik olduğunu düşündüklerinden, çok sık ve az bilgi içeren kelimeleri ("stop words"), noktalama işaretlerini ve emojileri veriden çıkarmamışlardır. En yüksek başarımı BERT modeli ile elde edildiğini bildirmişlerdir. Bu çalışmada Türkçe dili üzerine yapılan çalışmalardan farklı olarak BERT modelinde uygun tweet'lerin seçimi konusu, karakter bazlı n-gramlar ve retweet bazlı özniteliklerin kullanımı incelenmiştir.

3 Veri kümesi

Herhangi bir Twitter kullanıcısının yaş ve cinsiyet bilgisini el ile etiketlemek çok zor olacağı için, veri kümemiz genellikle internet üzerinden bilgilerine kolaylıkla ulaşılabileceğimiz ünlü kişilerden oluşturulmuştur. Aynı meslek grubundaki insanların kullandıkları üslubun benzer olması sebebiyle tek bir meslek grubundaki tanınmış kişiler yerine siyasetçi, sanatçı, sporcu, gazeteci, modacı, iş adamı, doktor, veteriner, avukat, ekonomist ve mühendis gibi çeşitli meslek gruplarından kişiler belirlenmiştir. Kullanılan dilin belirli bir konuya yönelik olmaması için mümkün olduğunca meslek gruplarında dengesiz dağılım olmamasına özen gösterilmiştir. Bu yüzden az sayıda kullanıcı içeren meslek gruplarındaki kullanıcıların takip ettiği kişilere bakılarak

aynı meslek grubundaki kullanıcıların sayısı artırılmıştır. Belirlenen kullanıcıların cinsiyetleri fotoğraf ve ismi üzerinden etiketlenmiştir. Yaşlarının tespiti için ise kişisel internet sayfaları ya da LinkedIn hesapları kullanılmıştır.

Literatürde de yaşlar genelde gruplandırılarak modeller eğitildiği için (örneğin, [18]), kullanıcıların yaşları belirlendikten sonra 18-30, 31-40, 41-50, 51-60 ve 61+ olmak üzere beş farklı yaş grubu kullanılmıştır. Genellikle genç yaştaki tanınmış kullanıcıların sporcu veya oyuncu olması sebebiyle meslek grupları ile yaş bilgisinin özdeşleşmesi, hazırlanan veri kümesinde yanlılığa yol açabilir. Bu nedenle genç yaştaki kullanıcıların veri kümesindeki oranını artırmak amacıyla #tercih2021 etiketi altına ÖSYM yerleştirme sonucunu paylaşan ve kullanıcı isminden veya profil resminden cinsiyeti belirlenebilen 100 adet kullanıcı seçilmiştir. Bu kullanıcıların 18 ila 30 yaş aralığında olduğu varsayılmıştır.

Veri kümemiz 553 erkek ve 487 kadın (%53.2 erkek, %46.8 kadın) kullanıcı olmak üzere toplamda 1040 adet Twitter kullanıcısını içermektedir. Tablo 1'de veri kümemizden örnekler sunulmuştur. Kullanıcıların yaş ve meslek gruplarının grafiği Şekil 1'de gösterilmiştir. Yaş açısından bakıldığında yüksek yaş gruplarının azınlıkta olduğu görülmektedir.

4 Geliştirilen yöntemler

4.1 Problem tanımı

Bu çalışmada Türkçe tweet'lere sahip herhangi bir Twitter kullanıcısının yaş ve cinsiyetinin tahmin edilmesi amaçlanmaktadır. Literatürde de genel olarak uygulandığı üzere (örneğin, [18]), kullanıcıların direkt olarak yaşını tahmin etmek yerine yaş aralığı tahmin edileceğinden dolayı, her iki problem de birer sınıflandırma problemi olarak tasarlanmıştır. Yaş aralığı probleminde kullanıcılar yaşlarına göre 18-30, 31-40, 41-50, 51-60, 61 ve üzeri olmak üzere 5 gruba ayrılmıştır. Cinsiyet probleminde ise ikili sınıflandırma problemi olarak "kadın" ve "erkek" sınıflarının belirlenmesi amaçlanmaktadır

4.2 Veri önişleme

Yaş ve cinsiyet tahminleri için aynı metinler kullanılmıştır. Uygulayacağımız her yöntem için şu önişlemler uygulanmıştır: Kullanıcıların yalnızca dahil olduğu yaş grubunda iken paylaştığı tweet'leri içerecek şekilde filtrelenmiştir. Örneğin, bir kullanıcı 32 yaşında ise 31-40 yaş aralığına girmesi sebebiyle kullanıcının son 2 yılda paylaştığı tweet'ler alınıp diğerleri filtrelenmiştir. Tweet'lerde "@", "\#", "https://" ile başlayan kelimeler, emojiler ve retweeti temsil eden "RT:" ifadeleri çıkartılmıştır. BERT ve fastText bazlı yöntemler hariç diğer yöntemler için kelimelere köke indirgeme ("stemming") işlemi uygulanmıştır.

4.3 Yöntemler

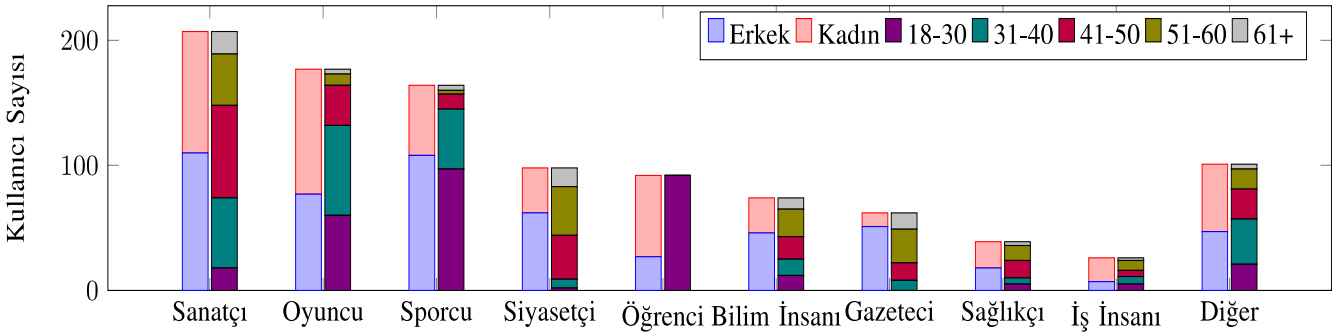
Önişlemden geçirilen veriyi girdi olarak alıp hesapların yaş grubu ve cinsiyetini tahmin eden beş farklı yöntem geliştirilmiştir. Bu yöntemler sırasıyla Kelime Bazlı N-Gram, Karaker Bazlı N-Gram, Retweet Bazlı Sınıflandırma, FastText Kelime Vektörleri ve BERT Dönüştürücü Modelleridir. Bu bölümde belirtilen beş yöntem anlatılmaktadır.

4.3.1 Kelime bazlı N-Gram

Yaş grubu ve cinsiyetin tespit edilebilmesinde kullanıcıların kullandıkları kelimeler ve bunların kullanılma sıklıkları belirleyici olabilir. Daha önce belirtildiği üzere, önceki çalışmalar insanların hayatlarının farklı dönemlerinde kullandıkları kelimelerin zamanla değişim gösterebildiğini ve benzer şekilde kadın ve erkeklerin sık kullandıkları kelimelerin de belirgin bir şekilde farklı olabildiğini belirtmişlerdir. O yüzden kelime torbası yöntemi etkin bir çözüm olabilir. Spesifik olarak, bu yöntemde, kelimelerin dokümanlarda ne kadar sık geçtiğini belirten *terim frekansı* ("term frequency", TF) ve tüm koleksiyonda ne kadar nadir geçtiğini belirten *ters doküman frekansı* ("inverse document frequency", IDF) skorlarını birleştiren TF-IDF skorları hesaplanıp, bu skorlara göre öznitelik çıkarımı yapılmıştır. Ardından destek vektör makineleri gibi çeşitli yapay öğrenme algoritmaları uygulanmıştır.

4.3.2 Karakter bazlı N-Gram

Tweet metinlerinde gündelik dil kullanımı ve birçok yazım hatası kelime bazlı öznitelik kullanan modellerin performansını düşürebilir. Bu nedenle, bu yöntemde kelimelerin frekansı yerine, 3 karakter uzunluğundaki kelime alt gruplarının TF-IDF skorlarını hesaplayıp öznitelikleri çıkardık. Ardından benzer şekilde çeşitli yapay öğrenme algoritmaları ile modelimizi eğittik. Hangi algoritmanın kullanılacağı bir sonraki bölümde deneylerde belirledik.



Şekil 1. Oluşturulan veri kümesindeki etiket dağılımı

4.3.3 *Retweet bazlı sınıflandırma*

Aynı yaş veya cinsiyet grubundaki kullanıcılar benzer ilgi alanlarına sahip olup, aynı kişileri takip edebilir ve onların tweet'lerini "retweet" edebilirler. Bu nedenle, bu yöntemde önce en çok retweet edilen hesaplar tespit edilip, bu hesaplar birer öznitelik olarak kullanılmıştır. Spesifik olarak, her hesaptan kaç kez retweet yapıldığı öznitelik değeri olarak kullanılmıştır.

4.3.4 *FastText kelime vektörleri*

Bojanowski vd.'nin [9] geliştirmiş olduğu fastText kelime vektörleri ("word embedding") kelimelerin semantik ve sentaktik özelliklerini yakalayabilen popüler bir yöntemdir. Ayrıca fastText karakter n-gramlarına göre bir öğrenme gerçekleştirdiği için diğer kelime vektörleri yöntemlerine göre yazım hatalarından çok daha az etkilenmektedir. Biz de bu modelimizde fastText'in sunduğu metin sınıflandırma modelini yaş ve cinsiyet tahmini problemimizde kullandık.

4.3.5 *BERT dönüştürücü modelleri*

BERT [10] modelleri ile birçok doğal dil işleme görevinde yüksek başarımlar elde edilmiştir. Bu sebeple biz de Türkçe metinler ile öneğitilmiş BerTurk [32] modeline hassas ayar işlemini uyguladık. BERT modeli sadece belirli bir uzunluktaki (512 token) metinleri girdi olarak alabilmektedir. Ancak kullanıcıların toplam mesaj uzunluğu BERT'in işleyebileceğinden daha uzun olabilmektedir. Bu yüzden bu yöntemde iki farklı tweet seçimi yöntemi uygulanarak iki farklı BerTurk modeli geliştirilmiştir. 1) $BerTurk_{Son}$: Tweet'ler paylaşıldığı zamana göre sıralanıp en son atılan tweet'ler kullanılmıştır. 2) $BerTurk_{Filtre}$: Önce sınıflara ait en sık geçen kelimeler belirlenmiştir. Ardından birden çok sınıfa ait sık kelimeler elenmiştir. En son olarak bu kelimeleri içeren tweet'ler BERT modelinde kullanılmıştır.

5 Deneyler

Bu bölümde öncelikle deney düzeneği anlatılmaktadır. Ardından makine öğrenmesi algoritmalarının karşılaştırılması ve özniteliklerin etkileri incelenmiştir.

5.1 *Deney düzeneği*

Makine öğrenmesi modelleri eğitilirken ve tweet'lerden kelime frekans vektörleri oluşturulurken Scikit kütüphanesi [33] kullanılmıştır. Önişleme aşamasında metinleri türcelere ("token") ayırma ve köke indirgeme işlemleri için sırasıyla NLTK [34] ve TurkishStemmer [35] kütüphaneleri kullanılmıştır. Ayrıca BERT uygulaması için ktrain kütüphanesinden [36] faydalanılmıştır.

Kelime bazlı n-gram modelinde tekli ve çiftli (unigram ve bigram) kelime grupları seçilmiştir. BERT modelinde her bir problem için 5 epoch, 2e-5 öğrenme oranı ve 6 batch büyüklüğü parametreleri kullanılmıştır. FastText modeli için ise Türkçe için hazırlanan kelime vektörleri [37] kullanılarak varsayılan model parametreleri kullanılmıştır.

Oluşturduğumuz veri kümesinin yüzde 60'ı eğitim, yüzde 20'si değerlendirme ("validation") ve yüzde 20'si test için kullanılmıştır. Modellerin başarımlarını ölçebilmek için makro-ortalama F1 skoru hesaplanmıştır. Rassal orman ("random forest") ve BERT modellerinin rastgele atanan kök ("seed") parametrelerinden ötürü 3 defa çalıştırılarak ortalama değerleri raporlanmıştır.

Filtreleme yönteminde en sık geçen kelimeler olarak cinsiyet tahmini probleminde 1000'den fazla geçen kelimeler arasından iki sınıfta ortak olmayan kelimeler seçilmiştir. Yaş aralığı probleminde ise 100'den fazla geçen kelimelerden yaş gruplarında ortak olmayan kelimeler seçilmiştir. Her bir kullanıcı için, belirlenen bu kelimelerin en az birinin geçtiği tweet'ler uç uca eklenip birleştirilmiş ve BERT modelinde kullanılmıştır.

5.2 *Deney sonuçları*

İlk deneyimizde kelime torbası, karakter bazlı n-gram ve retweet bazlı modellerde vektör uzunluğunu belirlemek için çeşitli uzunluklardaki vektörler kullanılarak lojistik regresyon modeli eğitilmiş ve eğitilen modellerin değerlendirme kümesindeki performansı ölçülmüştür. Yapılan deney sonuçları Tablo 2'de gösterilmiştir.

Tablo 2. Farklı boyutlardaki retweet, kelime ve karakter bazlı vektörler ile eğitilmiş lojistik regresyonu modelinin yaş ve cinsiyet tahminindeki F_1 skorları. Her durum için en yüksek skor kalın gösterilmiştir.

Yöntem	Öznitelik Sayısı	Yaş Tahmini	Cinsiyet Tahmini
Retweet Bazlı Vektörler	50	**0.368**	0.565
	100	0.353	0.565
	200	0.348	**0.599**
Kelime Bazlı N-Gram	5,000	0.361	**0.757**
	10,000	0.363	0.747
	20,000	**0.367**	0.737
Karakter Bazlı N-Gram	5,000	0.277	**0.755**
	10,000	0.279	0.726
	20,000	**0.286**	0.726

Öznitelik sayısı arttıkça kelime torbası ve karakter bazlı n-gram yöntemlerinin yaş tahmini problemindeki performansları artmış; cinsiyet tahmini problemindeki performansları ise azalmıştır. Retweet bazlı modelin performansı ise öznitelik sayısı arttıkça yaş tahmininde düşmüştür. Cinsiyet tahmininde ise öznitelik sayısı 200'e arttırılınca modelin performansında gözle görülür bir artış olmuştur. Yöntemleri karşılaştırdığımızda ise, yaş tahmininde retweet bazlı yöntem, cinsiyet tahmininde ise kelime bazlı n-gram yöntemi en yüksek başarımı elde etmiştir.

Bir sonraki deneyimizde retweet, kelime ve karakter bazlı öznitelikler kullanılarak destek vektör makineleri (support vector machines -SVM), Rassal Ormanlar ("Random Forest" -RF), K-en-yakın-komşu ("K-Nearest Neighbor" -KNN) ve lojistik regresyon ("logistic regression" -LR) algoritmaları eğitilmiş ve her modelin performansı değerlen dirme kümesinde ölçülmüştür. Sonuçlar Tablo 3'te gösterilmiştir.

Her algoritmanın diğerlerine göre daha yüksek başarım elde ettiği bir durum olduğu gözlemlenmektedir. Kelime

bazlı n-gram yönteminde her iki tahmin problemi için de en iyi sonuç lojistik regresyon kullanıldığında alınmıştır. SVM modeli de retweet bazlı öznitelikler ile yaş tahmininde ve karakter bazlı n-gram öznitelikleri ile cinsiyet tahmininde en yüksek sonucu elde etmiştir. RF ve KNN ise sırasıyla retweet bazlı cinsiyet tahmininde ve karakter n-gram bazlı yaş tahmininde en yüksek sonucu elde etmiştir.

En son deneyimizde ise, tüm yöntemlerimizin test kümesindeki F1, doğruluk ("accuracy"), hassasiyet ("precision") ve duyarlılık ("recall") değerleri hesaplanmıştır. Sonuçlar Tablo 4'te gösterilmiştir. Kelime, karakter ve retweet bazlı modeller için değerlendirme kümesinde en az bir problemde en yüksek başarımı gösteren modellerin sonuçları gösterilmiştir.

Yaş aralığı tahmini problemi için kelime vektörleri ve dönüştürücü modeller, yani genel olarak daha modern yöntemlerin diğer klasik yöntemlerden daha başarılı olduğu gözlemlenmektedir. Spesifik olarak, en son tweet'leri kullanarak geliştirilen BertTurk modeli en yüksek F1 değerini elde etmiştir, fastText modeli ise ikinci olmuştur.

Cinsiyet tahmini probleminde ise önerdiğimiz kelime bazlı filtre uygulanarak geliştirilen BertTurk modeli en yüksek başarımlı model olmuştur. Ancak cinsiyet modelinde klasik yöntemlerin de yüksek başarım elde ettiği gözlemlenmektedir. Özellikle, karakter n-gram öznitelikleri kullanılarak eğitilen SVM modeli BerTurkFiltre dışındaki tüm modellerden daha yüksek sonuç elde etmiştir. Daha önce belirtildiği gibi BERT bazlı modellerimizi üç kere çalıştırıp ortalama değerlerini sunduk. Ancak BERT modellerinin bu veri kümesindeki performansının standart sapmasının yüksek olduğu gözlemlenmiştir. Ayrıca farklı epoch değerlerinde de performansının düştüğü gözlemlenmiştir. Bu sebeple bu veri kümesindeki BERT sonuçlarına dikkatli yaklaşmak gerekmektedir.

Tablo 4'te kullandığımız yöntemlerin performansını Sezerer vd. [23]'nin cinsiyet tahmini için oluşturduğu veri kümesinde de ölçtük. Sezerer vd. veri kümesinde 3,368 kullanıcı eğitim için, 1,924 kullanıcı ise test için ayrılmıştır. Her kullanıcı için 100'er tweet paylaşılmıştır. Bu veri kümesinde kullanıcıların retweet ettikleri mesajlar olmadığı için retweet temelli yöntemimizi çalıştıramadık. Ayrıca hafıza yetersizliği probleminden ötürü 'batch size' parametresini BerTurkSon ve BerTurkFiltre için sırasıyla 4'e ve 2'ye düşürdük. Tablo 5'te cinsiyet tahmininde kullandığımız yöntemlerin Sezerer vd. [23] veri kümesindeki sınıflandırma performansları gösterilmektedir. Sezerer vd. veri kümelerinde temel sistem olarak SVM ile kelime torbası yönteminin performansını 0.72 doğruluk olarak belirtmişlerdir. Bizim yöntemlerimizden KNN Karakter N-Gram hariç hepsi Sezerer vd.'nin temel yönteminden daha yüksek başarım elde etmiştir. İlhami ve Hanbay [31] aynı veri kümesinde çeşitli derin öğrenme yöntemleri geliştirmiştir. Elde ettikleri en yüksek başarım BERT modeli ile 0.8012 olmuştur. Hesapların en son tweet'lerini kullanan BERT modelimiz de benzer bir sonuç almıştır. Ancak filtreleme yöntemi ile tweet'leri seçtiğimizde çok daha yüksek bir başarım elde etmekteyiz. Bu da birçok tweet'in aslında veride kirliliğe sebep olduğunu göstermektedir. Ek olarak, bazı spesifik ifadelerin kişilerin cinsiyetini tahmin etmede çok önemli olduğunu göstermektedir.

Her algoritmanın diğerlerine göre daha yüksek başarım elde ettiği bir durum olduğu gözlemlenmektedir. Kelime bazlı n-gram yönteminde her iki tahmin problemi için de en iyi sonuç lojistik regresyon kullanıldığında alınmıştır. SVM modeli de retweet bazlı öznitelikler ile yaş tahmininde ve karakter bazlı n-gram öznitelikleri ile cinsiyet tahmininde en yüksek sonucu elde etmiştir. RF ve KNN ise sırasıyla retweet bazlı cinsiyet tahmininde ve karakter n-gram bazlı yaş tahmininde en yüksek sonucu elde etmiştir.

En son deneyimizde ise, tüm yöntemlerimizin test kümesindeki F1, doğruluk ("accuracy"), hassasiyet ("precision") ve duyarlılık ("recall") değerleri hesaplanmıştır. Sonuçlar Tablo 4'te gösterilmiştir. Kelime, karakter ve retweet bazlı modeller için değerlendirme kümesinde en az bir problemde en yüksek başarımı gösteren modellerin sonuçları gösterilmiştir.

Yaş aralığı tahmini problemi için kelime vektörleri ve dönüştürücü modeller, yani genel olarak daha modern yöntemlerin diğer klasik yöntemlerden daha başarılı olduğu gözlemlenmektedir. Spesifik olarak, en son tweet'leri kullanarak geliştirilen BertTurk modeli en yüksek F1 değerini elde etmiştir, fastText modeli ise ikinci olmuştur.

Cinsiyet tahmini probleminde ise önerdiğimiz kelime bazlı filtre uygulanarak geliştirilen BertTurk modeli en yüksek başarımlı model olmuştur. Ancak cinsiyet modelinde klasik yöntemlerin de yüksek başarım elde ettiği gözlemlenmektedir. Özellikle, karakter n-gram öznitelikleri kullanılarak eğitilen SVM modeli BerTurkFiltre dışındaki tüm modellerden daha yüksek sonuç elde etmiştir. Daha önce belirtildiği gibi BERT bazlı modellerimizi üç kere çalıştırıp ortalama değerlerini sunduk. Ancak BERT modellerinin bu veri kümesindeki performansının standart sapmasının yüksek olduğu gözlemlenmiştir. Ayrıca farklı epoch değerlerinde de performansının düştüğü gözlemlenmiştir. Bu sebeple bu veri kümesindeki BERT sonuçlarına dikkatli yaklaşmak gerekmektedir.

Tablo 4'te kullandığımız yöntemlerin performansını Sezerer vd. [23]'nin cinsiyet tahmini için oluşturduğu veri kümesinde de ölçtük. Sezerer vd. veri kümesinde 3,368 kullanıcı eğitim için, 1,924 kullanıcı ise test için ayrılmıştır.

Her kullanıcı için 100'er tweet paylaşılmıştır. Bu veri kümesinde kullanıcıların retweet ettikleri mesajlar olmadığı için retweet temelli yöntemimizi çalıştıramadık. Ayrıca hafıza yetersizliği probleminden ötürü 'batch size' parametresini $\text{BerTurk}_{\text{Son}}$ ve $\text{BerTurk}_{\text{Filtre}}$ için sırasıyla 4'e ve 2'ye düşürdük. Tablo 5'te cinsiyet tahmininde kullandığımız yöntemlerin Sezerer vd. [23] veri kümesindeki sınıflandırma performansları gösterilmektedir. Sezerer vd. veri kümelerinde temel sistem olarak SVM ile kelime torbası yönteminin performansını 0.72 doğruluk olarak belirtmişlerdir. Bizim yöntemlerimizden KNN Karakter N-Gram hariç hepsi Sezerer vd.'nin temel yönteminden daha yüksek başarım elde etmiştir. İlhami ve Hanbay [31] aynı veri kümesinde çeşitli derin öğrenme yöntemleri geliştirmiştir. Elde ettikleri en yüksek başarım BERT modeli ile 0.8012 olmuştur. Hesapların en son tweet'lerini kullanan BERT modelimiz de benzer bir sonuç almıştır.

Tablo 3. Farklı makine öğrenmesi algoritmaları ve öznitelikler kullanılarak eğitilen modellerin değerlendirme kümesindeki F1 skorları. Her durum için en yüksek skor kalın gösterilmiştir.

Yöntem	Yaş			Cinsiyet		
	Kelime T.	Kr. N-Gram	RT	Kelime T.	Kr. N-Gram	RT
SVM	0.306	0.285	**0.377**	0.755	**0.779**	0.609
RF	0.309	0.316	0.313	0.741	0.763	**0.633**
LR	**0.367**	0.286	0.368	**0.757**	0.755	0.599
KNN	0.317	**0.357**	0.323	0.670	0.624	0.605

Tablo 4. Tüm yöntemlerin test kümesinde elde ettikleri F_1, doğruluk, hassasiyet ve duyarlılık değerlerinin karşılaştırılması.

Yöntem	Yaş				Cinsiyet			
	F_1	Doğruluk	Hassasiyet	Duyarlılık	F_1	Doğruluk	Hassasiyet	Duyarlılık
LR – Kelime N-Gram	0.388	0.502	0.388	0.405	0.722	0.729	0.723	0.722
SVM Karakter N-Gram	0.366	0.498	0.374	0.390	0.745	0.749	0.744	0.746
KNN Karakter N-Gram	0.397	0.483	0.432	0.400	0.664	0.633	0.670	0.659
SVM - Retweet	0.372	0.449	0.435	0.370	0.641	0.633	0.640	0.642
RF - Retweet	0.328	0.420	0.355	0.334	0.659	0.655	0.657	0.661
FastText	0.413	**0.527**	0.416	0.424	0.673	0.681	0.674	0.673
$\text{BerTurk}_{\text{Son}}$	**0.432**	0.512	**0.613**	**0.426**	0.742	0.744	0.740	0.743
$\text{BerTurk}_{\text{Filtre}}$	0.406	0.498	0.470	0.410	**0.762**	**0.764**	**0.760**	**0.763**

Ancak filtreleme yöntemi ile tweet'leri seçtiğimizde çok daha yüksek bir başarım elde etmekteyiz. Bu da birçok tweet'in aslında veride kirliliğe sebep olduğunu göstermektedir. Ek olarak, bazı spesifik ifadelerin kişilerin cinsiyetini tahmin etmede çok önemli olduğunu göstermektedir.

Modellerin başarımının ne derece iyi olduğunu anlamak için insanların yaş ve cinsiyet tahmin etmedeki başarımını ölçtük. Bunun için veri kümemizden her yaş grubundan rastgele 5 tane kadın ve 5 tane erkek hesap seçtik. Böylece toplamda 50 hesap seçilmiş oldu. Ardından kullanıcıların hesap isimlerini maskeleyip, en son 20 tweet'ini seçtik ve 3 kişiye bu hesapların sadece tweet'lerine bakarak yaş grubunu ve cinsiyetini tahmin etmelerini istedik. 3 kişinin yaş ve cinsiyet tahminindeki başarımları Tablo 6'da gösterilmiştir.

Tablo 5. Yöntemlerim Sezerer vd. [23] veri kümesinde elde edilen F_1, doğruluk, hassasiyet ve duyarlılık değerlerinin karşılaştırılması. Her durumdaki en iyi sonuç kalın fontla gösterilmiştir. Doğ: Doğruluk, Hassasiyet: H, Duyarlılık: D

Yöntem	F_1	Doğ	H	D
LR – Kelime N-Gram	0.821	0.804	0.823	0.819
SVM Karakter N-Gram	0.828	0.813	0.823	0.833
KNN Karakter N-Gram	0.713	0.680	0.730	0.697
FastText	0.827	0.809	0.833	0.820
$\text{BerTurk}_{\text{Son}}$	0.800	0.803	0.800	0.801
$\text{BerTurk}_{\text{Filtre}}$	**0.986**	**0.987**	**0.987**	**0.986**

Tablo 6. Rastgele seçilen 50 hesabın sadece tweet'lerine bakarak yapılan insan tahminlerinin doğruluk oranları.

	Yaş Tahmini	Cinsiyet Tahmini
Kişi - 1	0.34	0.66
Kişi - 2	0.34	0.78
Kişi - 3	0.34	0.64

Bizim en iyi modellerimiz yaş tahmininde 0.527 doğruluk, cinsiyet tahmininde ise 0.764 doğruluk elde etmiştir. Farklı kümeler olduğu için skorları direkt karşılaştırmak doğru değildir. Ancak modellerin performanslarına dair önemli bir referans değer olarak kabul edilebilir. Bu sebeple geliştirdiğimiz modeller yaş tahmininde insanlardan çok daha yüksek bir doğrulukta çalışmaktadır. Ayrıca insanların 0.34 doğruluğunda tahmin etmeleri de sadece metne bakarak insanların yaşlarını tahmin etmenin çok zor olduğunu göstermektedir. Geliştirdiğimiz modelin daha yüksek bir başarım elde etmesi metinler arasında görmesi zor olan benzerlikleri yakaladığını göstermektedir.

Önerdiğimiz yöntemlerin performansını daha iyi değerlendirebilmek için literatürdeki diğer diller için yapılan çalışmaların performans değerleri ile karşılaştırdık. Burada not edilmesi gerekir ki, farklı veri kümelerinde elde edilen sonuçlar olduğu için birebir karşılaştırma yapmak doğru değildir. PAN 2018 [38] ortak-görevinde metin bazlı cinsiyet tahmininde İngilizce, Arapça ve İspanyolca dilleri için elde edilen en yüksek skorlar sırasıyla 0.822, 0.817 ve 0.820'dir. PAN 2016 ortak-görevinde hem yaş hem cinsiyet tahmini

İngilizce ve İspanyolca dilleri için ele alınmıştır. Cinsiyet tahmininde en yüksek başarı İngilizce ve İspanyolca için sırasıyla 0.5575 ve 0.7031'dir. Yaş tahmininde elde edilen en yüksek başarım ise İngilizce ve İspanyolca için sırasıyla 0.3879 ve 0.3594'tür. Sonuç olarak, literatürde veri kümesine bağlı olarak çok farklı sonuçlar alınmaktadır. Farklı kümelerindeki skorlar direkt karşılaştırılamamakla birlikte, elde ettiğimiz sonuçların düşük olmadığını göstermektedir. Ancak diğer dillerde de olduğu üzere, bu konuda hala daha yüksek performanslı sistemlere ihtiyaç vardır. Paylaştığımız veri kümesinin diğer araştırmacıların daha yüksek performanslı sistemler geliştirmelerine katkı sağlayacağını umuyoruz.

6 Sonuçlar

Bu çalışmada Türkçe sosyal medya hesaplarının yaş ve cinsiyetini tahmin etmek için önce etiketli bir veri kümesi oluşturulmuştur. Ardından da beş farklı yöntem geliştirilmiş ve oluşturulan veri kümesinde performansları karşılaştırılmıştır. Geliştirilen modellerde dört farklı öznitelik kullanılmıştır: kelime bazlı n-gram, karakter bazlı n-gram, kullanıcıların retweet ettikleri hesaplar ve fastText kelime vektörleri. Bunlara ek olarak, BERT dönüştürücü modelleri incelenmiş ve kullanıcıların tweet'lerinin hangilerinin BERT'te kullanılması gerektiğine dair bir yöntem önerilmiştir.

Yaptığımız deneylerde son tweet'leri kullanarak geliştirilen BerTurk modeli yaş aralığı tahmininde en yüksek başarımı elde ederken, kelime bazlı tweet filtreleme yöntemi kullanılarak hassas ayarlanmış BerTurk modeli cinsiyet probleminde en yüksek başarımı elde etmiştir.

Oluşturduğumuz veri kümesini araştırmacılar ile paylaşıp bu konudaki önemli bir açığı kapatmayı planlamaktayız. İleriki çalışmalarda farklı öznitelikler kullanılarak daha güçlü modeller geliştirilebilir. Ayrıca ileriki çalışmalarda, diğer dillerdeki veri kümeleri kullanılarak diller-arası ("cross-lingual") modeller üzerine çalışmayı planlamaktayız. Buna ek olarak, deneylerimizde BERT modelleri için seçilen tweet'lerin performansı etkilediği gözlemlenmiştir. Bu sebeple, ileriki çalışmalarımızda bu konuya da eğilip, uzun metinlerin sınıflandırılması için metinlerin hangi kısmının BERT modellerinde kullanılması gerektiği konusunda da çalışmayı düşünmekteyiz. Son olarak, oluşturduğumuz veri kümesinde kullanıcıların meslek bilgileri de yer almaktadır. O yüzden kullanıcıların mesleklerini tahmin eden bir model geliştirmek de gelecek çalışma planlarımız arasında bulunmaktadır.

Teşekkür

Bu çalışma TÜBİTAK 3501 programının 120E514 nolu projesi tarafından desteklenmiştir.

Çıkar Çatışması

Yazarlar çıkar çatışması olmadığını beyan etmektedir.

Benzerlik Oranı (iThenticate): % 6

Referanslar

[1] Aljazeera, Twitter daily user growth rises as Musk readies to take control, https://www.aljazeera.com/economy/2022/4/28/twitter-daily-user-growth-rises-as-musk-readies-to-take-control, Erişim Tarihi: 29 Mart 2023.

[2] Statista, Leading countries based on number of Twitter users as of January 2022, https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/, Erişim Tarihi: 29 Mart 2023.

[3] N. Dwi Prasetyo, and C. Hauff, Twitter-based election prediction in the developing world. Proceedings of the 26th ACM Conference on Hypertext & Social Media, pp. 149-158, Guzelyurt, TRNC, Cyprus, 2015.

[4] A. Rashed, M. Kutlu, K. Darwish, T. Elsayed, and C. Bayrak, Embeddings-Based Clustering for Target Specific Stances: The Case of a Polarized Turkey. Proceedings of the International AAAI Conference on Web and Social Media, pp. 537-548, 2021.

[5] P. Suárez-Serrato, M. E. Roberts, C. Davis, and F. Menczer, On the influence of social bots in online protests. International Conference on Social Informatics, pp. 269-278, Bellevue, USA, 2016

[6] A. Mislove, S. Lehmann, Y. Y. Ahn, J. P. Onnela, and J. Rosenquist, Understanding the demographics of Twitter users. Proceedings of the International AAAI Conference on Web and Social Media, Vol. 5, No. 1, pp. 554-557, Barcelona, Spain, 2011

[7] C. Bayrak M. Kutlu, Predicting Election Results via Social Media: A Case Study for 2018 Turkish Presidential Election. IEEE Transactions on Computational Social Systems. 2022. https://doi.org/10.1109/TCSS.2022.3178052.

[8] PAN, Shared Tasks, https://pan.webis.de/shared-tasks.html, Erişim Tarihi: 29 Mart 2023.

[9] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, 135-146, 2017. https://doi.org/10.1162/tacl_a_00051.

[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171-4186, Minneapolis, MN, USA, 2019.

[11] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, and L. H. Ungar, Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8 (9), e73791, 2013, https://doi.org/10.1371/journal.pone.0073791.

[12] K. Santosh, R. Bansal, M. Shekhar, and V. Varma, Author profiling: Predicting age and gender from blogs. Notebook for PAN at CLEF. 2, 2013.
[13] W. Deitrick, Z. Miller, B. Valyou, B. Dickinson, T. Munson, and W. Hu, Author Gender Prediction in an Email Stream Using Neural Networks. Journal of Intelligent Learning Systems and Applications, 4, 169-175, 2012, https://doi.org/10.4236/jilsa.2012.43017.
[14] 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 (Working Notes), Thessaloniki, Greece, 2020.
[15] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, Classifying latent user attributes in twitter. Proceedings of the 2nd international workshop on Search and mining user-generated contents, pp. 37-44, Toronto, Canada, 2010.
[16] L. Flekova, D. Preoţiuc-Pietro, and L. Ungar, Exploring stylistic variation with age and income on twitter. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 313-319, Berlin, Germany, 2016.
[17] R. Hirt, N. Kühl, and G. Satzger, Cognitive computing for customer profiling: meta classification for gender prediction. Electronic Markets, 29(1), 93-106, 2019, https://doi.org/10.1007/s12525-019-00336-z.
[18] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, "How old do you think I am?" A study of language and age in Twitter. Proceedings of the International AAAI Conference on Web and Social Media, Vol. 7, No. 1, pp. 439-448, Ann Arbor, MI, USA, 2013.
[19] G. K. Mikros and K. Perifanos, Authorship attribution in greek tweets using author's multilevel n-gram profiles. AAAI Spring Symposium: Analyzing Microtext. pp. 17-23, 2013.
[20] S. Baxevanakis, S. Gavras, D. Mouratidis, and K. L. Kermanidis, A machine learning approach for gender identification of Greek tweet authors. Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, pp. 1-4, Corfu, Greece, 2020.
[21] K. Alrifai, G. Rebdawi, and N. Ghneim, Arabic Tweeps Gender and Dialect Prediction. CLEF (Working notes). Dublin, Ireland, 2017.
[22] M. Wiegmann, B. Stein, and M. Potthast, Celebrity profiling. Proceedings of the 57th annual meeting of the Association for Computational Linguistics, pp. 2611-2618, Florence, Italy, 2019.
[23] E. Sezerer, O. Polatbilek, and S. Tekir, A Turkish Dataset for Gender Identification of Twitter Users. Proceedings of the 13th Linguistic Annotation Workshop, pp. 203-207, Florence, Italy, 2019.
[24] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches, Overview of the author profiling task at PAN 2013. CLEF Conference on Multilingual and Multimodal Information Access Evaluation, pp. 352-365, Valencia, Spain, 2013.
[25] G. Park, D. B. Yaden, H. A. Schwartz, M. L. Kern, J. C. Eichstaedt, M. Kosinski, D. Stillwell, L.H. Ungar, and M. E. Seligman, Women are warmer but no less assertive than men: Gender and language on Facebook. PloS one, 11(5), e0155885. 2016, https://doi.org/10.1371/journal.pone.0155885.
[26] M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker, Gender differences in language use: An analysis of 14,000 text samples. Discourse processes, 45(3), 211-236, 2008, https://doi.org/10.1080/01638530802073712.
[27] J. W. Pennebaker and L. D. Stone, Words of wisdom: language use over the life span. Journal of personality and social psychology, 85(2), 291, 2003, https://doi.org/10.1037/0022-3514.85.2.291.
[28] P. M. Brandt and P. Y. Herzberg, Wisdom of words? Age differences in language and social media use in job applications. Current Psychology, 1-11, 2022, https://doi.org/10.1007/s12144-021-02646-y.
[29] D. Nguyen, N. A. Smith, and C. Rose, Author age prediction from text using linear regression. Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities, pp. 115-123, Portland, OR, USA, 2011.
[30] E. Sezerer, O. Polatbilek, Ö. Sevgili, and S. Tekir, Gender prediction from Tweets with convolutional neural networks: Notebook for PAN at CLEF 2018. 19th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF 2018. CEUR Workshop Proceedings. Avignon, France, 2018
[31] S. E. L. İlhami and D. Hanbay, Ön Eğitimli Dil Modelleri Kullanarak Türkçe Tweetlerden Cinsiyet Tespiti. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 33(2), 675-684, 2021. https://doi.org/10.35234/fumbd.929133.
[32] Stefan Schweter, BERTurk - BERT models for Turkish, https://zenodo.org/record/3770924, Erişim Tarihi: 29 Mart 2023.
[33] Scikit-Learn, Scikit-Learn Machine Learning in Python, https://scikit-learn.org/stable/index.html, Erişim Tarihi: 29 Mart 2023.
[34] NLTK, Natural Language Toolkit, https://www.nltk.org, Erişim Tarihi: 29 Mart 2023.
[35] O. Tunçelli, Turkish Stemmer Python, https://github.com/otuncelli/turkish-stemmer-python, Erişim Tarihi: 29 Mart 2023.
[36] A. S. Maiya, ktrain, https://github.com/amaiya/ktrain, Erişim Tarihi: 29 Mart 2023.
[37] É. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, Learning Word Vectors for 157 Languages. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 3483-3487, Miyazaki, Japan, 2018.
[38] F. Rangel, P. Rosso, M. Montes-y-Gómez, M. Potthast, and B. Stein, Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working notes papers of the CLEF*, pp. 1-38, Avignon, France, 2018.