Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi
International Journal of Engineering Research and Development
UMAGD, (2022) 14(3), s204-s210.
10.29137/umagd.1214018
Cilt/Volume:14 Sayı/Issue:3 Aralık/December 2022 Özel Sayı / Special Issue
Seçilmiş Konferans Makalesi / Selected Conference Paper

Nitelik Seçimi Kullanarak Twitter Kullanıcısının Cinsiyet Sınıflandırması üzerine Bir Çalışma

A Study on Twitter User Gender Classification using Feature Selection

Tuba Parlar*[1]

[1]*Mustafa Kemal Universitesi, Antakya, Hatay, TÜRKİYE*

***Başvuru/Received**: 03/12/2022* ***Kabul / Accepted**: 21/12/2022* ***Çevrimiçi Basım / Published Online:** 31/12/2022*
***Son Versiyon/Final Version**: 31/12/2022*

Öz
Günümüz iş modellerinde kurum veya kuruluşlar, karar alma süreçlerini iyileştirmek için kullanıcıların görüşlerini bilmek istemektedirler. Dünyanın dört bir yanındaki milyonlarca insan, sosyal ağ uygulamaları aracılığıyla metin mesajları, videolar veya fotoğraflar kullanarak günlük yorumlarını ve düşüncelerini ifade etmektedir. Facebook, Instagram, Twitter ve YouTube gibi sosyal ağ uygulamalarının hızla büyümesi, burada paylaşılan büyük veri içeriğini araştırmak ve kullanıcı davranışlarını analiz etmek için araştırmacılara çekici bir alan sunmaktadır. Sosyal ağlardan gelen bu muazzam miktardaki veri, etkili pazarlama, kişiselleştirilmiş öneri sistemleri, fikir liderleri bulma, ilaç endüstrisi veya politik analizler için kullanılmaktadır. Sosyal ağ uygulamaları aracılığıyla elde edilen büyük miktarda veri, makine öğrenme yöntemleriyle analiz edilmektedir. Bu çalışmada Twitter kullanıcılarının otomatik cinsiyet sınıflandırması performansını artırmak için nitelik seçim yöntemi kullanılmıştır. Twitter kullanıcı tanımları, twit metinleri ve her ikisinin bir arada kullanıldığı üç veri kümesi üzerinde uygulanan nitelik seçim yönteminin performansı naive bayes ve lojistik regresyon sınıflayıcıları ile değerlendirilmiştir. Deney sonuçları ki-kare nitelik seçim yöntemi ile seçilen niteliklerin lojistik regresyon ile sınıflandırma başarısının çok daha üstün olduğunu göstermektedir.

Anahtar Kelimeler
"Cinsiyet sınıflandırması, makine öğrenme, nitelik seçimi, sosyal ağlar, Twitter"

Abstract
In today's business models, institutions or organizations want to know users' opinions to improve their decision-making processes. Millions of people all around the world express their daily comments and thoughts using text messages, videos, or photos via social network applications. The rapid growth of social networking applications such as Facebook, Instagram, Twitter, and YouTube provides an attractive field for researchers to investigate the content of big data shared here and analyze user behavior. This enormous amount of data from social networks is used for effective marketing, personalized recommendation systems, finding opinion leaders, the pharmaceutical industry, or political policy making. A big amount of data obtained through social network applications is analyzed by machine learning methods. In this study, feature selection method is used to improve the automatic gender classification performance of Twitter users. The performance of the feature selection method that is applied on three datasets: user descriptions, tweets and where both are used together is evaluated with naive bayes and logistic regression classifiers. The results of the experiments show that the classification success of the selected features using chi-square feature selection method is much better with logistic regression classifier.

Key Words
"Feature selection, gender classification, machine learning, social networks, Twitter"

***Responsible Author:** tparlar@mku.edu.tr*

1. Giriş

Twitter, kullanıcılarının kısa cümlelerle görüşlerini paylaştığı en yaygın sosyal medya platformudur. Web 2.0 ile artan sosyal medya kullanımı ile bu platformlarda ürünler, hizmetler, kuruluşlar, bireyler, konular ve olaylar hakkında paylaşılan duygu ve düşünceler, başkalarının karar verme süreçlerinde çok daha etkili hale gelmiştir. Twitter, Instagram, Facebook gibi sosyal ağlar, e-ticaret siteleri, forumlar, bloglar aracılığıyla büyük miktarda veri elde edilmektedir. Elde edilen veriler makine öğrenme yöntemleri kullanılarak yüksek doğrulukta sınıflandırılabilmektedir. Bilgisayar bilimleri, yönetim bilimleri, pazarlama, finans, siyaset bilimi, iletişim, sağlık gibi çeşitli alanlarda sosyal medya uygulamalarının kullanıcılarının bir ürün veya hizmet hakkındaki görüşlerinden bilgi çıkarımı ve otomatik sınıflandırma her geçen gün daha da önem kazanmaktadır. Kullanıcısının duygularını maksimum 280 karakterle ifade etmesi gerektiğinden Twitter verileri çok daha öz ve anlamlı bilgi içermektedir (Brzustewicz & Singh, 2021; Dahal, Kumar, & Li, 2019; Kim, Ganesan, Dickens, & Panda, 2021). Ancak diğer sosyal ağlardan farklı olarak Twitter, kullanıcısı hakkında çok daha sınırlı bilgi taşımaktadır. Twitter kullanıcı profili, kullanıcının ekran adı, kullanıcı adı, yer, tanımlama gibi bazı temel nitelikleri içermektedir. Twitter kullanıcısının profilini analiz etmek, sosyal ağ uygulamaları alanında çalışan mühendislik, sosyoloji, psikoloji, pazarlama gibi farklı disiplinlerden araştırmacılar için ilgi çekici olmaya başlamıştır. Kullanıcıların profil analizi duygu analizi, doğal dil işleme gibi konu başlıkları altında yer almaktadır. Bu alandaki çalışmalarda makine öğrenme algoritmaları, metin madenciliği, veri madenciliği yöntemleri kullanılmaktadır.

Twitter kullanıcılarının profilini belirleme ile ilgili çalışmalar PAN konferanslarının temel konu başlıklarındandır. PAN CLEF 2017 ve 2018 konferanslarında farklı dillerde Twitter veri kümeleri üzerinde kullanıcı profili, özellikle cinsiyet belirlemek için yöntemler geliştirmek üzere çeşitli çalışmalar gerçekleştirilmiştir (Rangel, Rosso, Montes-y-Gómez, Potthast, ve Stein, 2018; Rangel, Rosso, Potthast, ve Stein, 2017). PAN CLEF 2018 konferansında İngilizce Twitter verisinden cinsiyet tespitinde en iyi sonucu, Daneshvar ve Inkpen (Daneshvar ve Inkpen, 2018), saklı anlamsal analiz (latent semantic analysis) yöntemi ile seçtikleri nitelikleri destek vektör makineleri algoritmasını kullanarak %82.21 doğrulukla sınıflandırmışlardır. PAN CLEF 2019 konferansında (Rangel ve Rosso, 2019) n-gram yöntemi kullanılarak, emoji ve özel karakterleri metin içerisine eklenerek lojistik regresyon sınıflayıcı ile %84'ün üzerinde doğruluk değeri elde edilmiştir (Valencia, Adorno, Rhodes, ve Pineda, 2019).

Khandelwal vd. (Khandelwal, Swami, Akhtar, ve Shrivastava, 2018) İngilizce-Hintçe karışık atılan twitlerden oluşan bir veri kümesini etiketleyerek cinsiyet sınıflandırması için farklı nitelik çıkarım yöntemleriyle sınıflandırma performans analizlerini yapmışlardır. Çalışmada en iyi sonuç, karakter n-gram yöntemiyle yapılan nitelik çıkarımı ve destek vektör makineleri sınıflayıcısı kombinasyonu ile %89.7 doğruluk değeri ile elde edilmiştir. Yang ve diğerleri (Yang, Al-Garadi, Love, Perrone, ve Sarker, 2021) yaptıkları çalışma ile biyomedikal araştırmalarda kullanılmak üzere Twitter verilerinden otomatik olarak cinsiyet analizi yapan bir model sunmuşlardır. Kullanıcı adları, kullanıcıların ekran adları, kullanıcı tanımları, twit metinleri ve profil renklerinden oluşan nitelikler ile destek vektör makineleri, rastgele orman, uzun-kısa vadeli bellek algoritmaları ile sınıflandırarak cinsiyete bağlı trendleri analiz etmeyi hedeflemişlerdir. Vashist ve Meehan (Vashisth ve Meehan, 2020) Twitter kullanıcı cinsiyetini otomatik olarak sınıflandıran çalışmalarında tf-idf, word2vec, glove nitelik çıkarım yöntemleriyle lojistik regresyon, destek vektör makineleri, naive bayes makine öğrenme algoritmalarını bir arada kullanmışlardır. Kaggle Twitter veri seti (Kaggle, 2016) üzerinde yaptıkları çalışmalarında en iyi sonucu word2vec yöntemiyle lojistik regresyon sınıflayıcı kullanarak %57.14 doğruluk değeri ile elde etmişlerdir.

Vicente vd. (Vicente, Batista, ve Carvalho, 2015) yalnızca kullanıcı adı ve ekran adı gibi profil bilgilerine dayalı olarak çıkarttıkları niteliklere dayanarak cinsiyet tahmini yaptıkları çalışmalarında naive bayes, lojistik regresyon, destek vektör makineleri, bulanık ortalamalar, ve k-ortalamalar gibi denetimli ve denetimsiz makine öğrenme yöntemleri ile performans analizleri yapmışlardır. Sonuç olarak, bulanık ortalamalara dayalı denetimsiz makine öğrenme yöntemini kullanarak %96 doğrulukla sınıflandırma performansı ile başarılı bir sonuç elde etmişlerdir. Vicente vd. (Vicente, Batista, ve Carvalho, 2019) İngilizce ve Portekizce dillerinde yaptıkları sınıflandırma çalışmasında ilave olarak profil fotoğrafı bilgisini de kullanmışlardır. Kullanıcı adı, ekran adı, kullanıcı tanımı, twit metni ve profil fotoğrafı bilgilerini bir arada kullanarak destek vektör makineleri sınıflayıcısı ile İngilizce için %93.2 Portekizce için %96.9 sınıflandırma başarısı elde etmişlerdir.

Bu çalışma ile kullanıcı tanımları ve kullanıcı tweetlerinden elde edilen niteliklere dayanarak kullanıcının cinsiyetini otomatik olarak sınıflandırmak için yöntem önerilmektedir. Kullanıcının profilinden (adı, ekran adı, tanımlaması) ve yazdığı twit metinlerinden elde edilen nitelikler hem ayrı ayrı hem de bir arada kullanılarak oluşturulan veri kümeleri üzerinde ki-kare nitelik seçim yöntemi uygulanmış ve elde nitelik alt kümeleri naive bayes ve lojistik regresyon makine öğrenme algoritmaları kullanılarak performans analizleri yapılmıştır. Literatürde yapılan çalışmalar incelendiğinde farklı veri ön işleme adımlarında, nitelik çıkarımlarında ve nitelik seçim algoritmaları farklı yöntemler izlendiği görülmüştür. Çalışmanın ikinci bölümünde kullanılan yöntemler ve değerlendirme ölçütleri açıklanmaktadır. Üçüncü bölümde veri kümesi, önişleme adımları, nitelik çıkarımı süreci ve deneyler sonuçları ile sunulmaktadır. Dördüncü bölümde ise genel sonuçlar ve öneriler sunulmaktadır.

2. Kullanılan Yöntemler

Bu bölümde Twitter kullanıcılarının tanımları ve yazdıkları mesajlardan yararlanılarak cinsiyet tahmininde kullanılan nitelik seçim yöntemi, makine öğrenme algoritmaları ve performans değerlendirme ölçütleri hakkında genel bir bilgi verilmektedir. Bu çalışmada sınıflandırma performansını artırmak için ki-kare nitelik seçim yönteminden yararlanılmaktadır. Performans analizi için olasılık tabanlı Naive Bayes ve Lojistik regresyon makine öğrenme algoritmaları kullanılmıştır.

2.1. Ki-kare (χ^2) nitelik seçme yöntemi

Ki-kare nitelik seçim yöntemi, her bir niteliğin, sınıf değişkeni ile arasındaki ilişkiyi ölçmektedir. Yöntem, sınıf frekansının sınıfın beklenen frekansına bölünmesi ile elde edilen değerlerin karşılaştırılmasına dayanır. Düşük puanlı bir *f* niteliği, ilgili *c* sınıfı için daha az bilgi vericidir ve bu nedenle kaldırılabilir. Denklem (1)'de; *A, c* sınıfında *f* niteliğini içeren belge sayısını; *B,* diğer sınıftaki *f* niteliğini içeren belge sayısını; C, *c* sınıfında *f* niteliğini içermeyen belge sayısını, *D* diğer sınıftaki *f* niteliğini içermeyen belge sayısını ifade ederken, *N* de toplam belge sayısını ifade etmektedir (Jin vd., 2015).

$$\chi^2(f,c) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{1}$$

Ki-kare nitelik seçim yöntemi veri kümesindeki her nitelik için sınıf ile arasındaki bağlantıya göre bir değer hesaplar ve sonra da nitelikleri bu değerlere göre sıralar. Büyük değere sahip olan nitelikler kullanıcının cinsiyet bilgisini belirlemede sınıflandırma başarısı için önemli niteliklerdir.

2.2. Naive Bayes makine öğrenme algoritması

Naive Bayes makine öğrenme algoritması, Bayes teoremine dayalı geliştirilen olasılık tabanlı bir sınıflandırma yaklaşımıdır. Koşullu olasılıktan yararlanılan Bayes teoremi bir niteliğin hangi olasılık ile bir sınıfa ait olduğu bilgisini hesaplar. Naive Bayes (NB) yaklaşımında ise, birçok nitelikten ($f_1,\ldots f_n$) oluşan veri kümelerinde $P(c \mid f_1,\ldots f_n)$ hesaplama maliyeti oldukça yüksek olduğundan bu sorunu çözmek için naive varsayımı ile sınıf bilgisi koşullu bağımsız olarak kabul edilmektedir. $P(c \mid f_1,\ldots f_n)$ denklem (2)'ye göre hesaplanmaktadır (Han ve Kamber, 2006).

$$P(c_i|f_1,\ldots f_n) = \frac{P(c_i)P(f_1,\ldots f_n|c_i)}{P(f_1,\ldots f_n)} \tag{2}$$

Burada $P(f_1,\ldots f_n)$ tüm sınıflar için sabit olduğundan amaç $P(c_i)P(f_1,\ldots f_n|c_i)$ değerinin maksimize edilmesidir.

2.3. Lojistik Regresyon makine öğrenme algoritması

Lojistik regresyon (LR) makine öğrenme algoritması ile hedeflenen bağımlı ve bağımsız değişkenler arasındaki ilişkinin doğrusal olarak modellenmesidir. Lojistik regresyon yaklaşımı ile veri kümesinden çıkarılan niteliklerin olasılıkları denklem (3)'e göre hesaplanmaktadır. Denklem (3)'te de görüldüğü gibi üstel bir fonksiyondan yararlanarak olasılık değerleri hesaplanarak nitelikler için bir sınıf belirlenmektedir.

$$P(c_i|f_1,\ldots f_n) = \frac{\exp\left(\sum_{i=1}^{N} w_i f_i\right)}{\sum_c \exp\left(\sum_{i=1}^{N} w_i f_i\right)} \tag{3}$$

2.4. Performans değerlendirme ölçütleri

Makine öğrenme algoritmalarının sınıflandırma performansları karmaşıklık matrisine (Tablo 1) dayalı olarak hesaplanan temel ölçütlere göre değerlendirilmektedir. Çalışmamızda Twitter kullanıcı verilerinden elde edilen niteliklerle kullanıcı cinsiyet analizi yapılması hedeflenmiştir. Tablo 1'i incelediğimizde, çalışmamızdaki niteliklerin sınıflandırma sonucu tahmin kadın ve gerçek sınıf da kadın ise doğru negatif (true negative-*TN*); gerçek sınıf erkek ise yanlış negatif (false negative-*FN*); sınıflandırma sonucu tahmin erkek iken gerçek sınıf kadın ise yanlış pozitif (false-positive-*FP*); gerçek sınıf erkek ise doğru pozitif (true positive-*TP*) olarak ifade edilmektedir. Doğruluk (accuracy), Kesinlik (Precision), Duyarlılık (Recall), F1 formüllerini gösteren Denklem (4-7), karmaşıklık matrisinden yararlanılarak geliştirilen performans ölçütleridir (Han ve Kamber, 2006; Sokolova ve Lapalme, 2009).

Tablo 1. Karmaşıklık Matrisi

		Tahmin Edilen Sınıf	
		Kadın (0)	**Erkek (1)**
Gerçek Sınıf	**Kadın (0)**	*TN*	*FP*
	Erkek (1)	*FN*	*TP*

Doğruluk (accuracy) ölçütü, her iki sınıf (kadın ve erkek) için doğru tahmin ile yapılan sınıflandırma toplamının, toplam sayıya bölümü ile elde edilir:

$$Doğruluk = \frac{TN + TP}{TN + TP + FP + FN} \tag{4}$$

Kesinlik (precision-*P*) ölçütü, başarılı tahmin oranı yani erkek (1) sınıfı için yapılan tahminlerin sınıflandırmadaki doğru pozitif oranını gösterir:

$$P = \frac{TP}{TP + FP} \tag{5}$$

Duyarlılık (recall-*R*) ölçütü, erkek (1) sınıfı için yapılan tahminlerin sınıflandırmada gerçek doğru olma oranını gösterir:

$$R = \frac{TP}{TP + FN} \tag{6}$$

*F*1 ölçütü kesinlik ve duyarlılık ölçütlerinin harmonik ortalaması alınarak elde edilir:

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{7}$$

3. Kullanılan Veri Kümesi ve Deney Sonuçları

3.1. Veri kümesi

Önerilen yöntemleri kullanarak deneyleri yapmak için Kaggle Twitter veri kümesi kullanılmıştır (Kaggle, 2016). Kaggle Twitter veri kümesi 20 bin kayıttan oluşmaktadır. Her bir kayıt, kullanıcı adı, twit, kullanıcı hesap bilgileri, yer, bir bağlantı bilgisi içermektedir. Veri kümesinden yararlanarak kullanıcı cinsiyet sınıflandırması denetimli öğrenme algoritmalarıyla otomatik olarak yapılacağından cinsiyet bilgisi olmayan kayıtlar elenmiştir. Deneyler, 5725 kadın (sınıf:0) ve 5469 erkek (sınıf: 1) olmak üzere toplam 11194 kayıt ile gerçekleştirilmiştir.

3.2. Deneysel sonuçlar

Çalışmada kullanılan masaüstü bilgisayar Intel core i5-10400F CPU işlemci 16 GB RAM bellek, NVIDIA GeForce RTX 3070 ekran kartı özelliklerine sahiptir. Deneyler için kodlar Python NLTK[1] ve scikit-learn (Pedregosa vd., 2011) kütüphanelerinden yararlanılarak geliştirilmiştir. Nitelik çıkarımı için öncelikle Twitter veri kümesi üzerinde veri önişleme yapılmıştır. Tüm kelimeler küçük harfe çevrilmiş, @, RT, bağlantılar, etiketler, emojiler, noktalama işaretleri, rakamlar, tek karakterler, etkisiz kelimeler (sıklıkla tekrar eden and, a, an, the vs…) çıkarılmıştır. Daha sonra kök çözümleme (lemmatization) gerçekleştirilmiştir. Böylece farklı ek alan aynı kelime köklerine inildiği için nitelik sayıları indirgenmiştir. Örneğin "studies, studied, studying" kelimeleri yerine "study" kelimesi kullanılmıştır. Nitelik çıkarımı, kelime torbası (bag-of-words) yöntemi ile terim frekanslarına göre gerçekleştirilmiştir. Veri temizleme, indirgeme ve bütünleştirme aşamalarından sonra elde edilen veri kümesi, kullanıcı tanımları ve kullanıcıların attığı twitler olmak üzere iki bölümden oluşmaktadır. Tablo 2'de görüldüğü gibi kullanıcı tanımları ve twit metni bir arada kullanıldığında 33464 nitelik, yalnızca kullanıcı tanımı 21138 nitelik, kullanıcı twit metinleri 19673 nitelikten oluşmaktadır. Temel bir performans değerlendirmesi yapmak amacıyla nitelik seçim yöntemi uygulamadan önce, her bir deneyde, kayıtların %80'i eğitim verisi %20'si test verisi olarak bölündükten sonra tüm nitelikler için naive bayes ve lojistik regresyon makine öğrenme algoritmaları ile sınıflandırma yapılmıştır. Deney sonuçları her bir veri kümesi için Tablo 2'de sunulmuştur. Tablo 2'de görüldüğü gibi en iyi sonuç kullanıcı tanımları veri kümesi ile elde edilen %67 F1 ölçütü ile lojistik regresyon sınıflayıcı ile elde edilmiştir.

Tablo 2. Nitelik seçimi öncesi Sınıflandırma Sonuçları

		NB		**LR**	
	Tüm Nitelikler	**Doğruluk**	**F1**	**Doğruluk**	**F1**
Kullanıcı Tanımı+Metin	33464	**59.89**	**59**	65.61	66
Kullanıcı Tanımı	21138	57.84	55	**66.67**	**67**
Kullanıcı Twit Metni	19673	57.17	54	58.6	58

Tüm nitelikler için temel değerlendirme sonuçları elde edildikten sonra, ki-kare nitelik seçim algoritması ile kullanıcı tanımı ve twit metinleri, kullanıcı tanımları ve kullanıcı twit metinleri olmak üzere 3 ayrı veri kümesi için performans değerlendirme yapılmıştır. Ki-kare nitelik seçim yöntemiyle skorları hesaplanan nitelikler önem sırasına göre sıralanmıştır. En değerli ilk 250, 750, 1250, 2250 ve 2750 nitelik sayıları için sınıflandırma deneyleri gerçekleştirilmiştir. Tablo 3'de görüldüğü gibi, kullanıcı tanımlarının ve twit metinlerinin bir arada kullanıldığı verilerin sınıflandırılmasında en iyi sonuç %74 F1 ölçütü ile 2250 nitelik ile elde edilmiştir. 33464 nitelik ile elde edilen %66 F1 değeri sınıflandırma başarısı 2250 adet ki-kare yöntemiyle seçilen nitelikle çok daha yüksek performans ile sınıflandırmıştır. Bu sonuç hesaplama zamanı ve maliyeti açısından da önemlidir.

[1] https://nltk.org

Tablo 3. Ki-kare Nitelik Seçim Yöntemi ile Performans Değerlendirme Sonuçları

Kullanıcı Tanımı+Metni	**NB**		**LR**	
Nitelik Sayıları	**Doğruluk**	**F1**	**Doğruluk**	**F1**
250	60.12	54	68.20	67
750	62.97	58	70.30	70
1250	65.74	62	72.18	72
1750	66.99	64	72.67	73
2250	70.70	69	**73.60**	**74**
2750	**71.95**	**71**	73.51	73
33464 (Tüm nitelikler)	59.89	59	65.61	66

Tablo 4'te gözlemlendiği gibi sadece kullanıcı tanımlarına göre oluşturulan niteliklere göre sınıflandırma yapıldığında en iyi sonuç 2750 nitelik kullanılarak %72.58 doğruluk ve %72.02 F1 ölçütü ile lojistik regresyon sınıflayıcı ile elde edilmiştir. Tablo 2'deki sonuçların bir miktar daha iyi sınıflandırma başarıları sunduğu gözlenmektedir.

Tablo 4. Ki-kare Nitelik Seçim Yöntemi ile Performans Değerlendirme Sonuçları

Kullanıcı Tanımı	**NB**		**LR**	
Nitelik Sayıları	**Doğruluk**	**F1**	**Doğruluk**	**F1**
250	62.04	57.47	66.23	64.66
750	63.06	57.92	70.03	69.28
1250	63.82	59	71.37	70.86
1750	**69.36**	**67.21**	71.91	71.56
2250	65.52	61.47	72.18	71.80
2750	68.56	65.55	**72.58**	**72.02**
21138 (Tüm nitelikler)	57.84	55.0	66.67	67.0

Tablo 5'te sadece kullanıcı twit metinlerinden yararlanılarak cinsiyet bilgisi sınıflandırılması deney sonuçları verilmiştir. Buradaki sonuçların Tablo 3 ve 4'teki değerlerden çok daha düşük olduğu görülmektedir. Yine de nitelik seçiminin sınıflandırma başarısını artırdığı görülmektedir. Tüm nitelikler için %58 F1 sınıflandırma başarısı, seçilen 2750 nitelik ile %67.6 F1 değerine yükselmiştir. Deney sonuçlarından da anlaşıldığı gibi en iyi sınıflandırma sonucu twitter kullanıcı tanımları ve twit metinlerinden bir arada elde edilen niteliklerin ki-kare nitelik seçim yöntemiyle en değerli niteliklerin seçilmesiyle 33464 nitelikten seçilen 2250 değerli niteliğin sınıflandırılmasıyla %74 F1 değeri ile LR sınıflayıcısıyla elde edilmiştir (Tablo 3).

Tablo 5. Ki-kare Nitelik Seçim Yöntemi ile Performans Değerlendirme Sonuçları

Twitter Metni	**NB**		**LR**	
Nitelik Sayıları	**Doğruluk**	**F1**	**Doğruluk**	**F1**
250	57.17	48.88	59.85	55.68
750	60.47	54.23	64.63	63.37
1250	63.47	59.28	65.88	65.19
1750	63.15	58.17	66.64	66.18
2250	66.19	62.54	67.44	67.05
2750	**66.64**	**63.11**	**67.89**	**67.61**
19673 (Tüm nitelikler)	57.17	54.0	58.6	58.0

4. Sonuç ve Öneriler

Sosyal ağ verilerinden yararlanarak kullanıcı profilini çıkarmak alışveriş alışkanlıkları, siyasi yönelimleri, sağlık ihtiyaçları gibi farklı alanlarda kullanılabilecek yararlı bilgiler edinmek açısından giderek daha önemli hale gelmektedir. Kullanıcılarının günlük hayatlarında karar alma mekanizmalarında, iş ve alışveriş alışkanlıklarında etkilendiklerinde bu mecralar ilişkiler, demografi ve sosyal davranışlar açısından psikolojik, sosyolojik, pazarlama gibi alanlarda ilgi çekici olmayı hak etmektedir. Makine öğrenme yöntemleriyle

sınıflandırma yapmak için kullanıcıların verilerinden yararlı niteliklerin çıkarılması giderek zorlaşmaktadır. Çünkü paylaşımlar artmakta ve veri çoğaldıkça gürültülü veri dediğimiz anlamsız veri de artmaktadır. Nitelik seçim yöntemlerinin kullanımının önemi giderek artmaktadır.

Bu çalışma ile Twitter veri kümesinden elde edilen kullanıcı tanımları ve twit metinlerinden yararlanılarak otomatik olarak kullanıcının cinsiyet tahmini yapılması için bir model önerilmiştir. Önerilen model ile Twitter verileri önişlemden geçirildikten sonra nitelik çıkarımı yapılmış, nitelikler terim frekansı ile ağırlıklandırılıp ki-kare nitelik seçim yöntemi ile önemine göre sıralanmıştır. Yapılan deneylerden sonra en başarılı sınıflandırma performansı kullanıcı tanımlarının ve twit metinlerinin bir arada kullanıldığı veri kümesi üzerinde yapılan nitelik seçimi ile gerçekleştirildiği gözlenmiştir. Tüm nitelikler (33464 adet) üzerinde yapılan sınıflandırma performansı %66 F1 değerinde iken ki-kare yöntemi ile seçilen en değerli 2250 nitelik ile %74 F1 değeri ile sınıflandırma başarısı %8 artırılmıştır. Ki-kare nitelik seçim yöntemi hem veri kümesinde gürültü yaratan ve sınıflandırmaya katkısı olmayan nitelikleri elenmiş hem de çok daha az nitelikle çalışan sınıflayıcının hesaplama zamanı maliyeti azaltılmıştır. Sınıflayıcılar karşılaştırıldığında lojistik regresyon sınıflayıcısının çok daha başarılı olduğu gözlenmiştir. Lojistik regresyon sınıflayıcısı korelasyonu yüksek niteliklerle daha başarılı sonuç verdiği bilinmektedir. Veri kümesinde çok fazla korelasyonu yüksek nitelik olması nedeniyle naive bayes sınıflayıcısından daha yüksek sınıflandırma başarıları elde edilmiştir. Gelecek çalışmalarda, sınıflandırma başarısını artırmak amacıyla farklı nitelik çıkarım yöntemleriyle hibrit nitelik seçim yöntemleri kullanarak deneyleri farklı veri kümeleri üzerinde geliştirmeyi ve daha başarılı sınıflandırma sonuçları elde etmeyi planlamaktayız.

Bilgilendirme

Bu çalışma ICSAR 2022 (1st International Conference on Scientific and Academic Research) konferansında sunulmuştur.

Referanslar

Daneshvar, S., ve Inkpen, D. (2018). *Gender identification in twitter using n-grams and lsa.* Paper presented at the proceedings of the ninth international conference of the CLEF association (CLEF 2018).

Han, J., ve Kamber, M. (2006). *Data Mining: Concepts and Techniques* (Second ed.): The Morgan Kaufmann Series in Data Management Systems.

Jin, C., Ma, T., Hou, R., Tang, M., Tian, Y., Al-Dhelaan, A., ve Al-Rodhaan, M. (2015). Chi-square statistics feature selection based on term frequency and distribution for text categorization. *IETE journal of research, 61*(4), 351-362.

Kaggle. (2016). Twitter User Gender Classification. Retrieved from https://www.kaggle.com/datasets/crowdflower/twitter-user-gender-classification?select=gender-classifier-DFE-791531.csv

Khandelwal, A., Swami, S., Akhtar, S. S., ve Shrivastava, M. (2018). Gender Prediction in English-Hindi Code-Mixed Social Media Content: Corpus and Baseline System. *Computacion Y Sistemas, 22*(4), 1241-1247. doi:10.13053/CyS-22-4-3061

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

Rangel, F., ve Rosso, P. (2019). *Overview of the 7th author profiling task at PAN 2019: bots and gender profiling in twitter.* Paper presented at the Proceedings of the CEUR Workshop, Lugano, Switzerland.

Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., ve Stein, B. (2018). Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working notes papers of the CLEF*, 1-38.

Rangel, F., Rosso, P., Potthast, M., ve Stein, B. (2017). Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working notes papers of the CLEF*, 1613-0073.

Sokolova, M., ve Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437. doi:10.1016/j.ipm.2009.03.002

Valencia, A. I. V., Adorno, H. G., Rhodes, C. S., ve Pineda, G. F. (2019). *Bots and gender identification based on stylometry of tweet minimal structure and n-grams model.* Paper presented at the Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum, Lugano, Switzerland.

Vashisth, P., ve Meehan, K. (2020). *Gender classification using twitter text data.* Paper presented at the 2020 31st Irish Signals and Systems Conference (ISSC).

Vicente, M., Batista, F., ve Carvalho, J. P. (2015). *Twitter gender classification using user unstructured information.* Paper presented at the 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE).

Vicente, M., Batista, F., ve Carvalho, J. P. (2019). Gender detection of Twitter users based on multiple information sources. In *Interactions between computational intelligence and mathematics part 2* (pp. 39-54): Springer.

Yang, Y. C., Al-Garadi, M. A., Love, J. S., Perrone, J., ve Sarker, A. (2021). Automatic gender detection in Twitter profiles for health-related cohort studies. *Jamia Open, 4*(2). doi:10.1093/jamiaopen/ooab042