Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 187-194, Kasım 2020

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 187-194, November 2020

Research Article

Twitter Platformunda Makine Öğrenmesi Algoritmalarıyla Cinsiyet ve İlgi Analizi

Enes Günçe[1*], Aydın Carus[2]

[1] Trakya Üniversitesi, Fen Bilimleri Enstitüsü, Edirne, Türkiye (ORCID: 0000-0001-8546-2324)
[2] Trakya Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Edirne, Türkiye (ORCID: 0000-0003-3370-5974)

(International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) 2020 – 22-24 Ekim 2020)

(**DOI:** 10.31590/ejosat.819722)

ATIF/REFERENCE: Günçe, E., & Carus, A. (2020). Twitter Platformunda Makine Öğrenmesi Algoritmalarıyla Cinsiyet ve İlgi Analizi. *Avrupa Bilim ve Teknoloji Dergisi*, (Özel Sayı), 187-194.

Öz

Twitter gibi sosyal ağlar, insanların iletişim kurması için popüler bir platform haline gelmiştir. Bireysel kullanıcıların yanı sıra kurumlar ve şirketler de ürün tanıtımı, pazarlama ya da herhangi bir konu hakkında geri bildirim alma gibi daha birçok nedenden dolayı bu sahaya ilgi duymaktadır. Kurumların ve şirketlerin hedefi, kişilerin ilgilendikleri ürün ve alanlar dışında gereksiz bilgiler ile rahatsız edilmemesini sağlamaktır. Bunun için de kurum ve şirketler, paylaşım yapanın kadın veya erkek oluşu, tweetin ilgili olduğu alan gibi bilgilere ihtiyaç duymakta ve bu bilgilere bağlı olarak, kendi hedef kitlelerine ulaşmak için çeşitli çalışmalar yapmaktadır. Bu çalışmada Twitter'da üretilen içeriklerden yola çıkılarak, paylaşım yapanın cinsiyeti ve paylaşılan tweetin ilgi alanı için tahmin yapılmıştır. Bu amaçla, Twitter Uygulama Programlama Arayüzü (API- Application Programming Interface) kullanan bir uygulama geliştirilmiştir. Bu uygulama kullanılarak, iki farklı eğitim seti oluşturmaya yönelik veriler toplanmıştır. Cinsiyet tespitine yönelik eğitim seti için, tweetler filtreleme yapılmadan toplanmıştır. İlgi alanı tespitine yönelik eğitim seti için, tweetler farklı ilgi alanları için belirlenmiş anahtar kelime kümeleri yardımıyla, filtreleme yapılarak toplanmıştır. Daha sonra, bu tweetler, etiketleme çalışmasına kolaylık sağlaması amacıyla uygulama kullanılarak el ile etiketlenmiştir. Çeşitli denemeler yapılarak, özniteliklerin belirlenmesinin ardından, gözetimli makine öğrenmesinde kullanılacak iki farklı eğitim seti oluşturulmuştur. Oluşturulan bu eğitim setleri kullanılarak; Naive Bayes, K-En Yakın Komşu Algoritması (KNN- K-Nearest Neighbors), C4.5, Destek Vektör Makineleri (SVM- Support Vector Machine) ve Ardışık Minimal Optimizasyon algoritmaları (SMO- Sequential Minimal Optimization) için modeller oluşturulmuştur. Modellerin başarımı, kappa istatistik ve doğruluk ölçütleri dikkate alınarak değerlendirilmiştir. Elde edilen modellerin başarımları değerlendirildiğinde; cinsiyet tahmini için oluşturulan modeller içinde, en düşük başarıma %44,6 doğruluk ve 0.17 kappa değeri ile SVM algoritması sahipken en yüksek başarımı %99,9 doğruluk ve 0.99 kappa değeri ile SMO algoritması sağlamıştır. Aynı şekilde ilgi alanı için oluşturulan modeller içinde en düşük başarımı %47,9 doğruluk ve 0.37 kappa değeri ile SVM algoritması vermişken en yüksek başarım %93,18 doğruluk ve 0.91 kappa değeri ile KNN algoritması tarafından sağlanmıştır. Doğruluk değerleri ve kappa değerlerinin birbiri ile uyumlu olduğu görülmüştür.

Anahtar Kelimeler: Twitter, Makine Öğrenmesi, Cinsiyet Analizi, İlgi Alanı Analizi

Gender and Interest Area Analysis on Twitter using Machine Learning Algorithms

Abstract

Social networks have become popular platforms that help people to connect with each other. In addition to individuals, companies and institutions are also interested in social networks for several reasons such as promoting and marketing their products or getting feedback on a specific topic. The goal of companies and institutions is to ensure that people are not targeted by unnecessary information except for products and areas they are interested in. To achieve their business goals, companies and institutions would like to determine the gender of a person who shares the post and the interest area a social media post is related to. Using this information, they carry out various studies to reach their target audiences. In this study, we analyze tweets to identify the genders of Twitter users and interest areas tweets are related to. We develop an application that uses the Twitter Application Programming Interface (API). We collect data using this application to create two different training sets: the gender determination training set and the interest area determination training set. For the gender determination training set, we collect tweets without filtering them. For the interest area determination training set, we collect the tweets by filtering them with the help of the sets of keywords that are created separately for each interest area. After

[*] Sorumlu Yazar: Trakya Üniversitesi, Fen Bilimleri Enstitüsü, Edirne, Türkiye, ORCID: 0000-0001-8546-2324, enesgunce@trakya.edu.tr

collecting the Tweets, we tag them manually with the help of the application in order to facilitate the tagging process. By performing various experiments, after the determination of the attributes, two different training sets were created which are then used in supervised machine learning. Models were created using these training sets for Naive Bayes, K-Nearest Neighbor Algorithm (KNN-K-Nearest Neighbors), C4.5, Support Vector Machines (SVM-Support Vector Machine) and Sequential Minimal Optimization algorithms (SMO-Sequential Minimal Optimization). The performances of the models were evaluated taking into account kappa statistics and accuracy criteria. When the performances of the obtained models were evaluated, among the models created for gender prediction, the lowest success rate was 44.6% with an accuracy of 44.6% and a kappa value of 0.17. While SVM algorithm had the highest performance, SMO algorithm provided 99.9% accuracy and 0.99 kappa value. Likewise, SVM algorithm gave the lowest performance with 47.9% accuracy and 0.37 kappa value among the models created for the area of interest, while the highest performance was achieved by the KNN algorithm with 93.18% accuracy and 0.91 kappa value. It is observed that the accuracy values and kappa values are compatible with each other.

Keywords: Twitter, Machine Learning, Gender Analysis, Interest Field Analysis.

1. Giriş

Günümüzde teknolojik olarak gelişmiş özellikler içeren taşınabilir cihazların ucuzlaması ve bu sayede yaygınlaşması, çok fazla yeni kullanıcının ortaya çıkmasına da neden olmuştur. Kullanıcı sayısının artması ve beraberinde, bu taşınabilir cihazlar üzerinden veri paylaşmaya yönelik platform ve uygulamaların geliştirilmiş olması, veri paylaşımını bugüne kadar olmadığı bir hızda arttırmıştır. Kullanıcıların görüşlerini paylaşma amacıyla en çok tercih edilen platformlardan birisi olan Twitter, kullanım kolaylığı ve paylaşımları anlık bir şekilde birçok kişiye ulaştıran tasarımı sayesinde, insanların kendini ifade etmek için kullandığı platformlardan biri haline gelmiştir. Aynı zamanda birçok kurum ve kuruluşun da kullanıcı olarak yer aldığı bu platform, kullanıcıların fikirlerini, şikâyetlerini ve sorularını istedikleri kullanıcıya anlık ulaştırıp, cevap alabildiği bir iletişim aracı olarak da kullanılabilmektedir. Twitter platformundaki tüm bu paylaşımlar değerlendirilerek, kişilerin belirli özelliklerini ortaya çıkarabilmek mümkündür. Twitter, kullanıcıları hakkında birçok bilgi sağlamasına rağmen kullanıcının yaşı, cinsiyeti, eğitim durumu, ilgi alanları gibi bilgileri ise sağlamamaktadır. Bu çalışmada, Twitter' dan toplanan tweetler üzerinden, tweeti paylaşanın, makine öğrenmesi algoritmaları ile cinsiyet ve paylaşılan tweetin ilgi alanı tahmininin yapılması hedeflenmiştir.

Matematiksel teoriye dayanan algoritmaları kullanan makine öğrenmesi, geçmiş deneyimlerden elde edilmiş veriyi işleyerek gelecek durumlar için kestirimler de gerçekleştirebilmektedir. Bu kestirimlere dayanarak, veriler işlendikten sonra değerli bilgiler ortaya çıkarılabilmektedir. Bu bilgiler kurum ve şirketlere önemli yararlar sağlayabilir. Bu sayede, kişilerin sosyal medyada yaptığı paylaşımlardan elde edilen bilgiler kullanılarak, kişiye özel ürün önerisi ya da ilgili olduğu alanlara özel mesajlar gönderilebilmektir. Buradaki amaç, kişilerin ilgilenmedikleri ürün ve alanların dışında, gereksiz bilgiler ile rahatsız edilmemesini sağlamaktır.

Twitter'da gönderilen her paylaşım kullanıcıların kendi özgün fikri olduğu için kurum ve şirketler, paylaşılan bu veriye önem vermektedir. Kurum ve şirketlerin amacı, olabildiğince kendi hedef kitlesi olan kullanıcıya erişmektir. Dolayısıyla kullanıcıların paylaşımlarından çıkarılabilecek bilgiler, bu kurum ve şirketler için daha cazip hale gelmiştir. Twitter, kişilerin cinsiyeti ve paylaşılan tweetlerin ilgi alanları hakkında bilgi sunmadığı için bu bilgilerin ortaya çıkarılması, bu bilgileri kullanacak kurum ve şirketler için önem arz etmektedir. Tweeti paylaşan kişilerin cinsiyetinin ve tweetlerin ilgi alanlarının bilinmesi, kurumlar ve şirketler için hedef kitleye düşük maliyet ve kısa sürede ulaşmayı sağlayacaktır. Bu çalışmanın önemli bir özelliği ise Türkçe veriler üzerinden gerçekleştirilmiş olmasıdır. Türkçe için yapılan bu çalışma, benzer çalışmaların azlığı nedeniyle önem arz etmektedir. Bu çalışmanın 2. bölümünde, konuyla ilgili yapılmış literatürdeki çalışmalara yer verilmiştir. 3. bölümde, gerçekleştirilen çalışmanın detayları sunulmuştur. 4. bölümde, deneysel sonuçlar açıklanmaktadır. Son bölümde, çalışmadan elde edilen sonuçlar değerlendirilmiştir.

2. Literatür Taraması

Sosyal medya verileri üzerinden sınıflandırma ve bu verileri kullanarak çıkarımlar yapmak, günümüzde önem arz eden bir çalışma alanı haline gelmiştir. Hızla büyüyen sosyal medya platformlarından biri olan Twitter, kolay ulaşılabilen verileri sayesinde, birçok çalışmada tercih edilmiştir. Literatürde, kullanıcıların paylaştıkları tweetlerin üzerinden içeriklerinin belirlemesine yönelik çalışma yapılmıştır [1]. Fakat bunlar ya İngilizce ya da farklı diller için yapılmıştır. Birden fazla dil için paylaşılan tweetleri makine öğrenmesi algoritmaları ile analiz ederek kullanıcıların yaşı, cinsiyeti ve demografik özelliklerini ortaya çıkarmak için yapılan çalışma da vardır [2]. Sadece isimler üzerinden kişilerin demografik özelliklerinin çıkartılması, yapılan çalışmalar arasındadır [3]. Portekizce dili için tweet içeriği ve tweet sahibi kişilerin cinsiyetlerini tahmin eden çalışma da yapılmıştır [4]. Yapılan bir başka çalışmada, Sayyadiharikandeh vd. öznitelik olarak sadece; ekran ismi, profil resmi ve tweet metnini kullanmışlar ve sonuç olarak 89.9% gibi bir doğruluk değerine ulaşabilmişlerdir [5]. Bhattacharya vd. kullanıcıların ilgi alanları üzerine yaptıkları çalışma ile "Who Likes What" isminde, milyonlarca Twitter kullanıcısının verilerini toplayan bir sistem tasarlayıp, insanların ilgi alanlarını tahmin eden çalışma yapmışlardır [9].

"Ontology-based sentiment analysis of twitter posts" isimli çalışmada [10], kişilerin tweetleri üzerinden duygu analizi yapılmıştır. Bu çalışma için duygu analizinin yapılma amaçlarından biri, kurum ve kuruluşlar için müşteriler açısından sorunların ne olduğuna dair cevaplar aramaktır. "Predicting the Future With Social Media" [11] isimli çalışmada, Twitter verileri, sinema filmlerinin, gişede elde edeceği gelirleri, filmler henüz vizyona girmeden tahmin etmek için kullanılmıştır. Twitter'dan veri alınırken, anahtar kelime olarak film isimleri kullanılmış, 3 aylık bir zaman aralığı boyunca 24 farklı filmle ilgili 2.890.000 tweet toplanmıştır. Kullanıcıların paylaştığı tweetler kullanılarak yapılan tahminler, Hollywood Stock Exchange adlı siteden alınan veriler ile karşılaştırılmıştır. Culotta yaptığı çalışmada [12], Twitter üzerinden alınan, 10 haftalık bir zaman aralığını kapsayan 500.000'den fazla tweet kullanarak, grip salgınlarını

tahmin etmeye çalışmıştır. Basit doğrusal regresyon ve çoklu doğrusal regresyon kullanılan bu çalışmada, elde edilen veriler, ABD Hastalık Kontrol ve Korunma Merkezleri'nden alınan veriler ile %78 oranında benzerlik göstermiştir.

Soler vd. yaptıkları çalışmada [13], geliştirilen Taratweet isimli aracı kullanılarak 2011 ve 2012 yıllarında İspanya'da yapılan 3 farklı seçim sonuçlarını tahmin etmeye çalışmışlardır. Yerel seçimler için yapılan bu çalışma 105.282; genel seçimler için yapılan çalışma ise 259.016; Endülüs' teki seçim için yapılan çalışma 176.000 adet tweet kullanılarak gerçekleştirilmiştir. Çalışmada siyasi partilere verilen oylar ile bu partilerin Twitter' da bahsedilme oranları karşılaştırılmış ve uyumlu oldukları görülmüştür. Hussein vd. tarafından yapılan çalışmada [14], arapça dilinde paylaşılan tweetlerden kullanıcıların cinsiyetleri tahmin edilmeye çalışılmıştır. Sonuç olarak %87,6 bir doğruluk oranı ile kullanıcıların cinsiyetleri tahmin edilebilmiştir.

3. Gerçekleştirilen Çalışma

3.1. Eğitim Setleri

Yapılan bu çalışma için tweet toplama, veri ön işleme ve veri temizleme işlemlerini yerine getirmek için bir uygulama geliştirilmiştir. Eğitim seti oluşturma amacıyla bu uygulama ile veriler toplanırken, Twitter'ın sunduğu Uygulama Programlama Arayüzü (API- Application Programming Interface) kullanılmıştır. Visual Studio ortamında, C# programlama dili ile verilerin toplanması için Tweetinvi[2] kütüphanesi kullanılmıştır. Toplanan verilerin içeriğinin temizlenmemiş, ham veri olmasından dolayı, paylaşılan veriler uygulama aracılığı ile bir taraftan toplanırken diğer taraftan gerçek zamanlı olarak, Düzenli İfadeler (Regex- Regular Expression) kütüphanesi yardımıyla, içindeki özel karakterler ve etkisiz kelimeler temizlenmiştir. Tweet toplama sürecinde cinsiyet ve ilgi alanı için veri toplama süreci ayrı ayrı yürütülmüştür. Cinsiyet Eğitim setine yönelik tweet toplama sürecinde, filtreleme yapılmamıştır. Oluşturulacak ikinci Eğitim seti olan ilgi alanı için ise önceden belirlenmiş anahtar kelime kümeleri ile filtreleme yapılarak paylaşılan tweetler toplanmıştır. Her biri için sınıf sayısını da gösterir şekilde, cinsiyet ve ilgi alanı Eğitim setleri oluşturmaya yönelik, toplanan tweet sayıları sırasıyla Tablo 1 ve Tablo 2 de verilmiştir.

Tablo 1. Cinsiyet Eğitim Seti için Toplanan Tweet Adetleri

Sınıf	**Tweet Adedi**
KADIN	*54.370*
ERKEK	*48.985*
DİĞER	*67.845*

Tablo 2. İlgi Alanı Eğitim Seti için Toplanan Tweet Adetleri

Sınıf	**Tweet Adedi**
SİYASET	*15.000*
SPOR	*15.000*
EĞİTİM	*15.000*
EKONOMİ	*15.000*
BİLİM VE TEKNOLOJİ	*15.000*
DİĞER	*15.000*

İlgi alanı Eğitim seti için öznitelik olarak sadece Paylaşılan tweet metni kullanılmıştır.

Cinsiyet tespiti için toplanan 171.200 adet tweetin ekran isimleri kontrol edilerek, farklı kullanıcılar belirlenmiş ve 171.200 adet tweetin, 59.750 adet eşsiz kullanıcı tarafından atıldığı görülerek, eşsiz kullanıcı eğitim seti oluşturmak için her bir kullanıcının rastgele bir tweeti alınmıştır. Tablo 3' de eğitim seti içindeki tweetlerin eşsiz kullanıcı olarak, sınıf bazlı sayıları gösterilmiştir.

Tablo 3. Eşsiz Kullanıcı Eğitim Seti Tweet Adedi

Sınıf	**Tweet Adedi**
KADIN	*24.520*
ERKEK	*26.740*
DİĞER	*8.490*

Cinsiyet Eğitim seti oluşturmak için toplanan tweetlerden öznitelik olarak belirlenen veriler Tablo 4' te açıklamalı olarak verilmiştir.

[2] https://github.com/linvi/tweetinvi, 2014

Tablo 4. Cinsiyet Eğitim Seti İçin Belirlenen Öznitelikler

Belirlenen Öznitelikler	
1- İsim alanı (erkek, kadın, diğer, erkek-kadın, kadın-erkek, erkek-sayı, kadın-sayı, diğer-sayı, erkek-emoji, kadın-emoji, diğer-emoji, erkek-noktalama, kadın-noktalama)	*26- Ekran ismindeki sayı adedi*
2- İsim	*27- Ekran ismindeki sayı adedi aralığı*
3- Ekran ismi	*28- Ekran ismindeki büyük harf adedi*
4- Tweet	*29- Ekran ismindeki büyük harf adedi aralığı*
5- Tweetteki kelime adedi	*30- Ekran ismindeki küçük harf adedi*
6- Tweetteki kelime adedi aralığı	*31- Ekran ismindeki küçük harf adedi aralığı*
7- Tweette tekrarlı kelime var mı?	*32- Tweet uzunluğu*
8- Tweetteki farklı kelime adedi	*33- Tweet uzunluğu aralığı*
9- Tweette ki sayı adedi	*34- Kaynak*
10- Tweetteki sayı adedi aralığı	*35- Arkadaş sayısı*
11- Tweetteki büyük harf adedi	*36- Arkadaş sayısı aralığı*
12- Tweetteki büyük harf adedi aralığı	*37- İsim' deki emoji sayısı*
13- Tweetteki küçük harf adedi	*38- İsim' deki emoji sayısı aralığı*
14- Tweetteki küçük harf adedi aralığı	*39- Tweetteki emoji sayısı*
15- Tweetteki hashtag adedi	*40- Tweetteki emoji sayısı aralığı*
16- Tweetteki hashtag adedi aralığı	*41- Tweet hangi karakter ile başlamış? (büyük/küçük/rakam)*
17- Tweetteki mention adedi	*42- Noktalama adedi*
18- Tweetteki mention adedi aralığı	*43- Noktalama adedi aralığı*
19- Tweet media (resim, video) içeriyor mu (var/yok)?	*44- Büyük kelime adedi*
20- İsim' deki sayı adedi	*45- Büyük kelime adedi aralığı*
21- İsim' deki sayı adedi aralığı	*46- Tweette büyük harfle başlayan kelime sayısı*
22- İsim' deki büyük harf adedi	*47- Tweette büyük harfle başlayan kelime sayısı aralığı*
23- İsim' deki büyük harf adedi aralığı	*48- Tweette tekrarlı harf sayısı*
24- İsim' deki küçük harf adedi	*49- Tweette etkisiz kelime sayısı*
25- İsim' deki küçük harf adedi aralığı	*50- Tweette etkisiz kelime sayısı aralığı*

Çalışmada öznitelik olarak kullanılan verilerden bazıları aşağıda açıklanmıştır:

İsim alanı: Kullanıcı isminin içinde hem erkem hemde kadın ismi varsa o halde "erkek-kadın" olarak belirlenmektedir. Herhangi bir erkek kullanıcının ismi "AHMET YILDIZ" ise "YILDIZ", bayan ismi olduğu için bu alan "erkek-kadın" olarak belirlenmiştir. Aynı şekilde isim alanında; sayı, noktalama veya emoji bulunuyor ise isimlendirme benzer şekilde "kadın-sayı", "kadın-noktalama", "erkek-emoji" şeklinde alınmaktadır.

Tweet hangi karakter ile başlamış?: Bu alanda kullanıcının attığı tweet eğer büyük harf, küçük harf veya rakam ile başlıyor ise sırasıyla "büyük harf", "küçük harf" , "rakam" olarak belirtilmiştir.

Tweette tekrarlı harf sayısı: Bu alan, tweette kelime içerisinde bir harfin, ardışık birden fazla geçmesi anlamına gelmektedir. Herhangi bir tweetin içerisinde "yaşaaaaa" gibi bir ifade geçerse, içindeki tekrarlanan harf adeti alınmaktadır.

Tweette tekrarlı harf sayısı: Bu alan, tweette yer alan fakat tweetin anlamına herhangi bir etkisi kelimelerin sayısı olarak alınmaktadır.

Cinsiyet eğitim seti için sınıf olarak Kadın, Erkek ve Diğer olarak 3 sınıf belirlenmiştir. Diğer ile ifade edilen sınıf, kurum ve kuruluşları kapsamaktadır.

İlgi alanı eğitim seti ise sınıf olarak Bilim ve Teknoloji, Eğitim, Spor, Siyaset ve Diğer olarak belirlenmiştir. Burada Diğer olarak ifade edilen sınıf ise ilk dört sınıf dışında olanları kapsamaktadır.

İlgi alanı eğitim seti için tweet toplarken filtre olarak nispeten ilgili tweetlerin elde edilmesi amacıyla, her bir sınıf için belirli anahtar kelimeler kullanılmıştır. Bu anahtar kelimeler sınıf bazlı olarak Tablo 5' te verilmiştir.

Toplanan tüm bu tweetlerden oluşturulan Eğitim setlerinin etiketleme işlemi, el ile yapılmıştır.

Tablo 5. İlgi Alanı Eğitim Setine Özgü Tweet Toplamak İçin Kullanılan Anahtar Kelimeler

Sınıf	Anahtar Kelimeler
EKONOMİ	*hisse senedi, ekonomi, sermaye, mali gelir, vadeli fiyat, vadeli hesap, dolar, vergi, cari kur, cari işlemler hesabı, enflasyon, dalgalı kur, deflasyon, döviz kuru, ekonomik güven endeksi, faiz oranı, finans, finansal, gayri safi milli hasıla, gsmh, hazine bonosu, ihracat, iskonto, ithalat, konut fiyat, konut kira, kredi riski, likidite, merkez bankası, para kurulu, para piyasası, para politikası, parite, resesyon, revalüasyon, sabit kur, stagflasyon, vadeli işlem, sterilizasyon, yatırım fonları*
EĞİTİM	*Yüksek öğretim kurumu, ösym, ösys, dgs, lisans eğitimi, önlisans eğitimi, ortaöğretim, ilköğretim, milli eğitim bakanı, milli eğitim, öğretmen ataması, öğretmen atamaları, atama, yüksek lisans, fakülte, yüksekokul, meslek yüksekokulu, örgün öğretim, ikinci öğretim, açıköğretim, devlet üniversitesi, kampüs, transkript, öğretim görevlisi, eğitim fakülteleri, akademi, seviye belirleme sınavı, dershane, hazırlık sınıfı, lisansüstü, formasyon, KPSS, KPDS, Öğretim üyesi, tez danışmanı, üniversite, Yükseköğretim*
SİYASET	*anayasa, kararname, milletvekili, tutanak, siyaset, bürokrasi, cumhurbaşkanı, politika, akp, chp, mhp, sayıştay, danıştay, parlamenter, demokrat, sosyalizm, radikalizm, laik, parlamento, emperyalist, referandum, sosyalist, radikalist, tbmm*
BİLİM VE TEKNOLOJİ	*anakart, işlemci, ram, bilişim, fiber, siber, veri tabanı, HDMI, algoritma, BIOS, IOS, ios, linux, root, yazılım, endüstri 4.0, inovasyon, sanal gerçeklik, yapay zeka, çözelti, elektrot, elektroliz, elektrolit, elektron, atom, fotoelektrik, kara delik, radyoaktif, kuantum, proton, nötron, nükleon, izotop, foton, quark, anti madde, hadron, fizyon, füzyon, nükleer, büyük patlama, kozmik ışınlar, karanlık madde, karanlık enerji, izafiyet, görelilik, pozitron, termodinamik, DNA, RNA, ışık hızı, süpernova, solucan deliği, katalizör, santrifüj, kimyasal reaksiyon, elektromanyetik, elektromanyetizma, inorganik kimya, izomer, anatomi, antibiyotik, bakteri, biyonik, kolestrol, genom, moleküler biyoloji, nörobiyoloji, nörobilim, patoloji, farmakoloji, prokaryot, ribozom, crispr, krispır, crispır, antibiyotik, bilim, nobel, AKN, adil kullanım kotası, bilim insanı, uzay istasyonu, insansız uzay aracı, iha, gökbilimciler, gökbilimci, diyabet, galaksi, uzay teleskobu, büyük veri, yüksek teknoloji*
SPOR	*asist, penaltı, defans, smaç, faul, şut, hakem, gol, kaleci, korner, rövaşata, UEFA, lig, ofsayt, şut, steps, ribaund, deplasman, turnuva, fair play, antrenör, hazırlık maçı, hat trick, oyun kurucu, galatasaray, fenerbahçe, beşiktaş*
DİĞER	*pasaport, akrostiş, alafranga, argo, asalet, betimleme, biyografi, fıkra, gazel, kafiye, kaside, şiir, mahlas, masal, roman, satranç, şarkı, akustik, magazin, acun, altın kelebek, demet akalın, ahmet kural, sıla, Seyahat Acentası, Konaklama, Rezervasyon, resepsiyon, suit oda , tek kişilik oda, film, dizi, cilt bakımı, hava durumu, tik tok, müzik, şarkı, yılbaşı, necip fazıl, kim milyoner olmak ister, namık kemal, flört, fragman*

3.2. Sınıflandırma Algoritmaları

Eğitim setleri oluşturulduktan sonra, bu eğitim setlerinden model oluşturmak için; Naive Bayes, K-En Yakın Komşu Algoritması (KNN- K-Nearest Neighbors), C4.5, Destek Vektör Makineleri (SVM- Support Vector Machine) ve Ardışık Minimal Optimizasyon algoritmaları (SMO- Sequential Minimal Optimization) olmak üzere beş adet makine öğrenmesi algoritması kullanılmıştır.

Basit ve kullanımı kolay bir sınıflandırıcı olarak tercih edilen Naive Bayes, temel istatistiksel sınıflandırıcılardan birisi olup ismini ünlü matematikçi Thomas Bayes'ten almıştır. Naive Bayes teoreminde her bir öznitelik, verilen sınıf içerisinde diğerlerinden bağımsız olarak kabul edilmektedir [6].

Destek Vektör Makineleri, sınıflandırma ve regresyon analizi için kullanılan bir makine öğrenmesi algoritmasıdır. Bu algoritma, iki sınıfı ayıran ve sınıflandırmayı yapmak için her sınıftan destek vektörleri adı verilen, az sayıda örneği seçip en yüksek kenar boşluğunu bulmaya çalışmaktadır [7].

K-En Yakın Komşu Algoritması (KNN- K-Nearest Neighbors), sınıflandırma için kullanılan bir diğer makine öğrenmesi algoritmalarından biridir. Veriler arasındaki uzaklığa bağlı olarak sınıflandırma işlemine dayanır. Sınıflandırılacak olan verinin en yakın "k" adet verisine bakılır. Bu mesafede, en çok olan hangi veri varsa tahmini yapılacak olan sınıfta bu verilere göre belirlenir [8].

C4.5, Ross Quinlan tarafından geliştirilen bir makine öğrenmesi algoritmasıdır. Bilgi entropisi yardımıyla eğitim setinden, ikili karar ağacı oluşturan C4.5 algoritması, eğitim süreci nispeten uzun olan, test süreci ise oldukça hızlı olan bir sınıflandırıcıdır [15].

Ardışık Minimal Optimizasyon algoritması [16], 1998 yılında makine öğrenmesi algoritmalarından, Destek Vektör Makinesi algoritmasına alternatif olarak ortaya çıkmıştır. Destek Vektör Makineleri algoritmasını eğitmek, büyük ve karmaşık sorunların çözülmesini gerektirir. Bu büyük karmaşıklık kuadratik programlama (KP) olarak isimlendirilmektedir. Ardışık Minimal Optimizasyon algoritması, bu büyük KP problemini, küçük parçalara böler. Bu parçalar analitik olarak çözülür ve optimizasyonun daha hızlı olmasını sağlar.

4. Deneysel Sonuçlar

Hazırlanmış olan eğitim setlerinden; Naive Bayes, K-En Yakın Komşu, C4.5, Destek Vektör Makineleri ve Ardışık Minimal Optimizasyon algoritmaları ile modeller oluşturularak testler yapılmıştır. Eğitim setinden oluşturulan modellerin başarımlarının ölçülebilmesi için, doğruluk ve kappa ölçütleri baz alınmıştır. Eğitim setlerindeki örnek sayılarına bağlı olarak elde edilen başarımlar, cinsiyet eğitim seti için Tablo 6' da ve benzer şekilde, ilgi alanı eğitim seti için Tablo 7' de verilmiştir.

Tablo 6. Cinsiyet Eğitim Seti Doğruluk ve Kappa Değerleri

	Eğitim Seti	**NB**		**SMO**		**KNN**		**C4.5**		**SVM**	
		Doğruluk (%)	*Kappa*	*Doğruluk (%)*	*Kappa*	*Doğruluk (%)*	*Kappa*	*Doğruluk (%)*	*Kappa*	*Doğruluk (%)*	*Kappa*
1	*300*	*86.0*	*0.79*	*98.3*	*0.97*	*82.3*	*0.73*	*97.6*	*0.96*	*44.6*	*0.17*
2	*600*	*87.6*	*0.81*	*98.8*	*0.98*	*83.0*	*0.74*	*98.1*	*0.97*	*44.0*	*0.16*
3	*900*	*88.5*	*0.82*	*99.4*	*0.99*	*86.4*	*0.79*	*98.4*	*0.97*	*47.1*	*0.20*
4	*1.200*	*89.9*	*0.84*	*99.4*	*0.99*	*88.0*	*0.82*	*98.4*	*0.97*	*48.4*	*0.22*
5	*1.500*	*90.2*	*0.85*	*99.8*	*0.99*	*87.3*	*0.81*	*98.4*	*0.97*	*48.6*	*0.23*
6	*1.800*	*90.6*	*0.85*	*99.9*	*0.99*	*86.8*	*0.80*	*98.7*	*0.98*	*49.2*	*0.23*
7	*2.100*	*90.5*	*0.85*	*99.9*	*0.99*	*86.7*	*0.80*	*98.9*	*0.98*	*49.9*	*0.24*
8	*2.400*	*90.7*	*0.86*	*99.9*	*0.99*	*87.3*	*0.81*	*99.0*	*0.98*	*50.9*	*0.26*
9	*2.700*	*91.1*	*0.86*	*99.9*	*0.99*	*87.6*	*0.81*	*99.0*	*0.98*	*51.5*	*0.27*
10	*3.000*	*91.2*	*0.86*	*99.9*	*0.99*	*88.0*	*0.82*	*99.4*	*0.99*	*51.1*	*0.26*
11	*3.300*	*91.0*	*0.86*	*99.9*	*0.99*	*88.3*	*0.82*	*99.4*	*0.99*	*52.1*	*0.28*
12	*3.600*	*91.0*	*0.86*	*99.9*	*0.99*	*88.4*	*0.82*	*99.3*	*0.99*	*52.4*	*0.28*
13	*3.900*	*91.3*	*0.87*	*99.9*	*0.99*	*88.5*	*0.82*	*99.4*	*0.99*	*51.7*	*0.27*
14	*6.000*	*91.3*	*0.86*	*99.9*	*0.99*	*89.8*	*0.84*	*99.6*	*0.99*	*54.2*	*0.31*
15	*9.000*	*91.0*	*0.86*	*99.9*	*0.99*	*90.3*	*0.85*	*99.4*	*0.99*	*56.6*	*0.34*
16	*15.000*	*91.2*	*0.86*	*99.9*	*0.99*	*91.1*	*0.86*	*99.6*	*0.99*	*60.2*	*0.40*
17	*45.000*	*91.1*	*0.86*	*99.9*	*0.99*	*93.5*	*0.90*	*99.6*	*0.99*	*69.6*	*0.54*
18	*171.200*	*91.4*	*0.87*	*99.9*	*0.99*	*95.7*	*0.93*	*99.4*	*0.99*	*81.9*	*0.72*

Eğitim setindeki örnek sayılarının, sınıflara göre dengeli olması durumu önem arz etmektedir. Eğer eğitim seti dengeli değil ise örnek sayısı fazla olan sınıf lehine bir yanlılık oluşacak ve bu da başarım ölçütlerine yansıyacaktır. Bu sebeple yapılan çalışmada buna önem verilmiş ve bu durumun önüne geçmek için mümkün olduğunca tüm eğitim setleri dengeli olarak hazırlanmıştır. Verilerin 300, 600, 900 gibi örnek sayısı içerecek şekilde hazırlanmıştır. Örneğin, örnek sayısı 300 adet olan eğitim seti ele alınırsa, eğitim setinin dengeli olması için bunun 100 adedi *erkek*, 100 adedi *kadın* ve 100 adet *diğer* sınıftan alınmıştır.

Tablo 6' da artan şekilde örnek sayısına sahip olan cinsiyet eğitim setlerinin, tüm algoritmalar için başarım sonuçları verilmiştir. Destek Vektör Makineleri algoritmasının, eğitim seti örnek sayısı 45.000 olduğu halde bile düşük sınıflandırma başarımına olduğu görülmektedir ancak eğitim seti örnek sayısı yaklaşık 4 katına çıkarıldığı durumda kabul edilebilir sınıflandırma başarımı seviyesine ulaşabilmiştir. Ancak geri kalan diğer algoritmaların, eğitim seti örnek sayısının 3.000' den başlayarak yüksek sınıflandırma başarımına ulaştığı görülmektedir. 45.000 örnek sayısında ise başarımlarının doğruluk ölçütü ele alındığında yaklaşık %90 ile %99 arasında olduğu tespit edilmiştir. Tüm bu sonuçlarda, kappa değeri ve doğruluk ölçütünün uyumlu olduğu söylenebilir.

Tablo 7. İlgi Alanı Eğitim Seti Doğruluk ve Kappa Değerleri

	Eğitim Seti	**NB**		**SMO**		**KNN**		**C4.5**		**SVM**	
		Doğruluk (%)	*Kappa*	*Doğruluk (%)*	*Kappa*	*Doğruluk (%)*	*Kappa*	*Doğruluk (%)*	*Kappa*	*Doğruluk (%)*	*Kappa*
1	*3.000*	*83.50*	*0.80*	*67.00*	*0.60*	*88.50*	*0.86*	*88.36*	*0.86*	*47.90*	*0.37*
2	*6.000*	*86.31*	*0.83*	*65.23*	*0.58*	*89.63*	*0.87*	*89.18*	*0.87*	*65.38*	*0.58*
3	*12.000*	*83.60*	*0.80*	*74.55*	*0.69*	*91.21*	*0.89*	*89.29*	*0.87*	*74.42*	*0.69*
4	*18.000*	*81.68*	*0.78*	*77.80*	*0.73*	*91.51*	*0.89*	*89.81*	*0.81*	*81.03*	*0.77*
5	*24.000*	*83.39*	*0.80*	*78.04*	*0.73*	*92.06*	*0.90*	*90.32*	*0.88*	*83.42*	*0.80*
6	*30.000*	*85.09*	*0.82*	*78.52*	*0.74*	*92.16*	*0.90*	*90.69*	*0.88*	*84.76*	*0.81*
7	*36.000*	*85.04*	*0.82*	*80.10*	*0.76*	*92.53*	*0.91*	*91.18*	*0.89*	*84.84*	*0.81*
8	*42.000*	*85.29*	*0.82*	*80.60*	*0.76*	*92.65*	*0.91*	*91.43*	*0.89*	*86.04*	*0.83*
9	*48.000*	*85.28*	*0.82*	*81.28*	*0.77*	*92.69*	*0.91*	*91.72*	*0.90*	*86.39*	*0.83*
10	*54.000*	*85.12*	*0.82*	*80.75*	*0.76*	*92.54*	*0.91*	*91.71*	*0.90*	*86.93*	*0.84*
11	*60.000*	*85.00*	*0.82*	*80.86*	*0.77*	*92.49*	*0.90*	*92.06*	*0.90*	*87.64*	*0.85*
12	*66.000*	*85.14*	*0.82*	*81.39*	*0.77*	*92.69*	*0.91*	*92.20*	*0.90*	*88.73*	*0.86*
13	*72.000*	*85.06*	*0.82*	*81.56*	*0.77*	*92.74*	*0.91*	*92.17*	*0.90*	*89.08*	*0.86*
14	*78.000*	*84.62*	*0.81*	*81.72*	*0.78*	*92.94*	*0.91*	*91.92*	*0.90*	*89.26*	*0.87*
15	*84.000*	*84.47*	*0.81*	*81.66*	*0.77*	*93.03*	*0.91*	*92.11*	*0.90*	*89.69*	*0.87*
16	*90.000*	*84.79*	*0.81*	*81.90*	*0.78*	*93.18*	*0.91*	*92.14*	*0.90*	*90.07*	*0.88*

Tablo 7' de ilgi alanı eğitim setinin doğruluk ve kappa değerleri görülmektedir. Tabloda 12.000 eğitim seti örnek sayısına kadar sınıflandırma başarımı en düşük olan algoritmanın Destek Vektör Makineleri olduğu görülmektedir. Bunun nedeni Destek Vektör Makineleri algoritmasının, nispeten daha yüksek eğitim seti örnek sayısına ihtiyaç duymasıdır. Destek Vektör Makineleri algortimasının, eğitim seti örnek sayısı 90.000' e çıkarıldığında Naive Bayes ve Ardışık Minimal Optimizasyon algoritmalarına göre daha yüksek sınıflandırma başarımına sahip olduğu görülmektedir. Doğruluk ölçütü olarak %93,18 değeri ile en yüksek sınıflandırma başarımı sağlamış olan algoritma ise K-En Yakın Komşu algoritması olmuştur. Tüm bu ilgi alanı eğitim setleri için algoritmaların 30.000 eğitim seti örnek sayısında kabul edilebilir sınıflandırma başarımına ulaştığı görülmektedir. Tablo 7' de görüldüğü üzere eğitim seti örnek sayısının 30.000 daha fazla olması sınıflandırma başarımlarına çok fazla katkı sağlamamaktadır.

Çalışmada; ayrıca cinsiyet sınıflandırma için 59.750 adet eşsiz kullanıcı eğitim seti oluşturulmuştur. Oluşturulan bu eğitim setinin de aynı algoritmalar ile testleri yapılmıştır. Bu eğitim seti için elde edilen sonuçlar Tablo 8' de verilmiştir. Bu sonuçlar değerlendirildiğinde aynı kullanıcının farklı tweetlerinin de bulunduğu eğitim seti sonuçları ile bir kullanıcının tek bir tweetini içeren eğitim seti sonuçları arasında anlamlı bir değişiklik olmadığı görülmektedir.

Tablo 8. Eşsiz Kullanıcı Eğitim Seti Doğruluk ve Kappa Değerleri

	Doğruluk (%)	**Kappa**
NB	*90.9*	*0.85*
SMO	*99.9*	*0.99*
KNN	*92.6*	*0.88*
C4.5	*99.8*	*0.99*

5. Sonuç

Bu çalışmada, Twitter'dan paylaşılan tweetler ile kullanıcıların cinsiyetleri ve atılan tweetlerin ilgi alanlarının tahmininin yapılması amaçlanmıştır. Farklı makine öğrenmesi algoritmaları ile ilgi alanı ve cinsiyet tahmini için oluşturulmuş eğitim setlerinden elde edilen modellerin başarımları test edilip, belirlenen algoritmaların bu eğitim setleri için sınıflandırma başarımının ortaya çıkarılması amaçlanmıştır. Eğitim setlerinin birden fazla algoritma ile test edilmesinin sebebi, her bir algoritmanın çalışma prensibinin farklı

olmasındandır. Deneysel sonuçlarda görüldüğü üzere; Naive Bayes, K-En Yakın Komşu Algoritması, C4.5, Destek Vektör Makineleri ve Ardışık Minimal Optimizasyon algoritmaları kullanılmıştır. Cinsiyet eğitim setinde %99,9 doğruluk değeri ile Ardışık Minimal Optimizasyon algoritması ve ilgi alanı eğitim setinde ise %93,18 doğruluk değeri ile K-En Yakın Komşu Algoritması en yüksek başarımı sağlayan algoritmalar olmuştur. Ardışık Minimal Optimizasyon algoritmasının sınıflandırma başarımının doğruluk değerinin yüksek olmasının sebebi, eğitim seti için az örnek sayısı ile yüksek başarım sağlamasıdır. Ardışık Minimal Optimizasyon algoritmasında, matris algoritması kavramı olmadığı için problemi daha az duyarlılıkla çözebilmektedir. Aynı şekilde cinsiyet eğitim setinde, 45.000 örnek sayısına sahip eğitim seti için doğruluk değeri yüksek olmayan algoritma olan Destek Vektör Makineleri, derin öğrenme mantığı ile çalışmasından dolayıdır. Destek Vektör Makine algoritmasının çalışma mantığı, eğitim seti için çok miktarda örnek sayısına ihtiyaç duymaktadır.

Yapılan bu çalışmayı daha ileri seviyeye taşımak için, cinsiyet ve ilgi alanı eğitim setlerini birleştirip kullanıcıların hem cinsiyetini hem de kullanıcıların ilgi alanlarını doğru tahmin edebilecek modeller geliştirilebilir.

Kaynakça

[1] Parantapa Bhattacharya, Muhammad Bilal Zafar, Niloy Ganguly, Saptarshi Ghos, Krishna P. Gummadi, “Inferring User interests in the Twitter Social Network”, In: Proceedings of the 8th ACM conference on recommender systems. ACM, pp 357–360, 2019.

[2] Mounica Arroju, Aftab Hassan, Golnoosh Farnadi, “Age, Gender and Personality Recognation using Tweets in a Multilingual Settings”, in CLEF 2015 working notes, Toulouse, France, 2015.

[3] Zach Wood-Doughty, Nicholas Andrews, Rebecca Marvin, Mark Dredze, “Predicting Twitter User Demographics from Names Alone”, Association for Computational Linguistics, USA, pp. 105-111, 2018.

[4] J.V.P.S Avinash and Rakshith Muniraju and Shreyas Shaligraman, “Gender Classification using Twitter Feeds”, CS-552-Advanced Data Mining – Final Project, Illinois Institute of Technology, Chicago, 2017.

[5] Mohsen Sayyadiharikandeh, Giovanni Luca Ciampaglia, Alessandro Flammini, “Cross-domain gender detection in Twitter”, Indiana University, School of Informatics and Computing, USA, pp. 39-54, 2019.

[6] A. McCallum, and K. Nigam, “A comparison of event models for Naïve Bayes text classification,” AAAI-98 Workshop on Learning for Text Categorization, pp. 41–48, 1998.

[7] C. Cortes, and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, pp. 273-297, 1995.

[8] S. Zhang1, M. Zong1, X. Zhu1, X. Li2, R. Wang3, “Efficient KNN Classification With Different Numbers of Nearest Neighbors”, IEEE Transactions on Neural Networks and Learning Systems, vol. 29, Issue: 5, May 2018.

[9] Bhattacharya P, Bilal Zafar M, Ganguly N, Ghosh S, Gummad P.K, “Inferring User Interest in Twitter Social Network”, IIT Kharagpur MPI-SWS, Germany, pp. 6-10, 2014.

[10] Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades Nick Bassiliades, “Ontology-based sentiment analysis of twitter posts”, vol. 40, Issue 10, pp. 4065-4074, August 2013.

[11] Sitaram Asur, Bernardo A. Huberman, “Predicting the Future With Social Media”, 11636305, IEEE, Toronto, ON, Canada, 01 November 2010.

[12] Aron Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages", Department of Computer Science, Southeastem Louisiana University, Hammond, LA 70402, 2010

[13] Juan M. Soler, Fernando Cuartero, Manuel Roblizo, "Twitter as a Tool for Predicting Elections Results", IEEE, Istanbul, Turkey, 04 February 2013.

[14] Shereen Hussein, Mona Farouk, ElSayed Hemayed,"Gender identification of egyptian dialect in twitter", vol. 20, Issue 2, pp. 109-116, July 2019.

[15] Edy Budiman, Haviluddin, Nataniel Dengan, Awang Harsa Kridalaksana1, Masna Wati, Purnawansyah, “Performance of Decision Tree C4.5 Algorithmin Student Academic Evaluation”, Faculty of Computer Science and Information Technology, Mulawarman University, Samarinda, Indonesia, Computational Science and Technology, pp. 380-389, 2018.

[16] S.S. Keerthi, E.G. Gilbert, “Convergence of Generalized SMO Algorithm for SVM Classsifier Design”, Dept. Of Mechanical and Production Engineering University of Singapore, pp. 351–360, 2002.