T.C.
PAMUKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DERİN ÖĞRENMEYE DAYALI SOSYAL MEDYA PROFİLLEMESİ

YÜKSEK LİSANS TEZİ

VASFİ TATAROĞLU

DENİZLİ, AĞUSTOS - 2019

T.C.
PAMUKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DERİN ÖĞRENMEYE DAYALI SOSYAL MEDYA PROFİLLEMESİ

YÜKSEK LİSANS TEZİ

VASFİ TATAROĞLU

DENİZLİ, AĞUSTOS - 2019

KABUL VE ONAY SAYFASI

Vasfi TATAROĞLU tarafından hazırlanan "**Derin Öğrenmeye Dayalı Sosyal Medya Profillemesi**" adlı tez çalışmasının savunma sınavı 26.08.2019 tarihinde yapılmış olup aşağıda verilen jüri tarafından oy birliği ile Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans Tezi olarak kabul edilmiştir.

Jüri Üyeleri	İmza
Danışman Prof. Dr. Sezai TOKAT	.....
Üye Doç. Dr. Emre ÇOMAK	.....
Üye Dr. Öğr. Üyesi Meriç ÇETİN	.....

Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun ...04.09.2019... tarih ve ...35/11... sayılı kararıyla onaylanmıştır.

.....

Prof. Dr. Uğur YÜCEL

Fen Bilimleri Enstitüsü Müdürü

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırmalarının yapılması ve bulgularının analizlerinde bilimsel etiğe ve akademik kurallara özenle riayet edildiğini; bu çalışmanın doğrudan birincil ürünü olmayan bulguların, verilerin ve materyallerin bilimsel etiğe uygun olarak kaynak gösterildiğini ve alıntı yapılan çalışmalara atfedildiğine beyan ederim.

VASFİ TATAROĞLU

ÖZET

DERİN ÖĞRENMEYE DAYALI SOSYAL MEDYA PROFİLLEMESİ
YÜKSEK LİSANS TEZİ
VASFİ TATAROĞLU
PAMUKKALE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

(TEZ DANIŞMANI:PROF. DR. SEZAİ TOKAT)

DENİZLİ, AĞUSTOS - 2019

İnsanoğlu yüzyıllardır edindiği bilgi ve tecrübelere göre devamlı kendisini geliştirmekte ve bu tecrübelerle bazı kararlar vermektedir. İnsanın kendisine has özelliklerini, düşünce yapısını, kararlarını tahminlemek tüm dünyada siyasetçiler, siyasi partiler ve ürün pazarlaması yapan reklamcılık sektörünün önem verdiği konulardan biridir. Sosyal medyanın kullanım oranının artması ve neredeyse herkesin bir çevrim-içi sosyal ağa bağlı olması ile birlikte kişiler izledikleri faaliyetler, okudukları eserler, takip ettikleri kurumlar veya kişilerle ilgili tercihlerini, duygularını, özel bilgilerini açık bir şekilde bu ortamlarda paylaşmaya başlamıştır. Her yeni gelen nesil ile birlikte giderek sosyal hayatın parçası haline gelen bu durum, büyük veri ve sosyal medya profillemesine verilen önemin de artmasına ve bu konuyla ilgili birçok çalışmanın yapılmasına yol açmaktadır. Bu sebeple bilgisayar biliminin ürettiği güncel teknik, yöntem, araç ve gereçlerin bu alanda uygulamaları geliştirilmektedir. Derin öğrenme, makine öğrenmesinin özel bir şeklidir. Derin öğrenme ağlarının olumlu yönlerinden biri, verilerin boyutu arttıkça gelişmeye devam etmeleridir.

Bu tez çalışmasında da Türkiye'deki siyasetçilerin, siyasi liderlerin ve siyasetle uğraşan yazarların, gazetecilerin Twitter sosyal medya hesapları kullanılarak oluşturulan büyük boyutlu bir ilişki matrisi yardımı ile sosyal medya profilleme yapılması ve buradan elde edilen bilgilerle kullanıcıların siyasi eğilimlerinin tahmin edilmeye çalışılması amaçlanmıştır. Siyasi görüşü bilinen örnek eğitim verisi üzerinde literatürdeki k-NN, naive bayes, rassal orman ve derin öğrenme gibi farklı makine öğrenmesi algoritmaları çalıştırılarak uygun parametre ve modellerin seçilmesi sağlanmış, test verileri ile de bu algoritmaların başarımları karşılaştırılmıştır. Siyasi eğilimlerin tahmini için algoritmalar karşılaştırıldığında %87.77 doğruluk, %87.93 kesinlik değeri ile derin öğrenme yönteminin karşılaştırılan diğer yöntemlere göre daha başarılı sonuçlar verdiği gözlemlenmiştir.

ANAHTAR KELİMELER: Sosyal Medya, Profilleme, Twitter, Veri Madenciliği, Derin Öğrenme

ABSTRACT

DEEP LEARNING BASED SOCIAL MEDIA PROFILING
MSC THESIS
VASFİ TATAROĞLU
PAMUKKALE UNIVERSITY INSTITUTE OF SCIENCE
COMPUTER ENGİNEERİNG

(SUPERVISOR:PROF. DR. SEZAİ TOKAT)

DENİZLİ, AUGUST 2019

Mankind constantly develops itself according to the knowledge and experience gained for centuries and makes some decisions with these experiences. All over the world, it is one of the issues that politicians, political parties and the advertising sector that make marketing of products give importance to estimating the characteristics, thinking and decisions of human being. With the increase in the usage of social media and the fact that almost everyone is connected to an online social network, people have started to share their preferences, feelings, private information about these activities, the works they read, the institutions or the people they follow in these environments. As each generation becomes increasingly a part of social life, this situation leads to an increase in the importance of social media profiling and many studies on this subject arises. For this reason, the current techniques, methods, tools and materials produced by computer science are developed in this field.

In this thesis, it is aimed to make social media profiling with the help of a large-scale relationship matrix created using the social media accounts of the politicians, writers and leaders who are engaged in politics in Turkey and to try to predict the political tendencies of the users with the information obtained from it. Using the sample training data with labeled political views, training was obtained using different machine learning algorithms in the literature such as k-NN, naive Bayes, random forest and deep learning and the performance of these algorithms were compared. When the algorithms were compared for the prediction of political tendencies, it was observed that %87.77 accuracy, %87.93 precision values and deep learning method gave more successful results compared to other methods compared.

KEYWORDS: Social Media, Profiling, Twitter, Data Mining, Deep Learning

İÇİNDEKİLER

Sayfa

ŞEKİL LİSTESİ

Sayfa

TABLO LİSTESİ

Sayfa

SEMBOL LİSTESİ

P(A) **:** A hipotezinin olma olasılığı.
P(A|B) **:** B verisinde A hipotezinin olma olasılığı. Buna posterior olasılık denir.

KISALTMALAR LİSTESİ

API	**:**	Application Programming Interface, Uygulama Programlama Arayüzü
GBDT	**:**	Gradient Boosted Decision Tree, Gradyan Güçlendirmeli Karar Ağacı
k-NN	**:**	K-Nearest Neighbourhood, k-En Yakın Komşuluk
SVM	**:**	Support Vector Machine, Destek Vektör Makinesi
WWW	**:**	Word Wide Web, Dünya Çapında Ağ

ÖNSÖZ

Dijital dünyada ve bilgi işlem dünyasında, sınırları her defasında aşan bir hızla bilgi üretilmekte ve toplanmaktadır. 2020 yılında tüm dünyada 50 milyarın üzerinde cihazın Internet'e bağlı olması beklenmektedir. Büyük verinin bu hız ve oranda üretilir olmasında 5 milyarın üzerinde mobil cihaz kullanıcısının çevrim içi sosyal ağlarda toplanan verileri önemli bir yer kaplamaktadır. Teknolojik gelişmeler ve sanallaşma ile birlikte sosyal medya kullanımının artacağı belirgindir. Tüm bu veriler, sosyal medya ortamlarındaki bu verilerin değerlendirilmesinin işletmelerin gelecek ile ilgili planlarında önem kazanacağını göstermektedir. Derin öğrenme, verilerin gösterimini öğrenmek için çoklu işlem katmanlarından oluşan hesaplama modelleri kullanır ve konuşma tanıma, görsel nesne tanıma, nesne algılama, genetik bilimi gibi birçok alanda son teknolojiyi önemli ölçüde geliştirmiştir. Tez çalışmamı seçmemde; sosyal medya ortamında üretilen büyük verinin önem kazanacağı, derin öğrenme algoritmalarının bu süreçte giderek daha çok kullanılacağı düşüncesi ve bu konularda kendimi geliştirmek istemem etkili olmuştur.

Bu çalışmayı yapmayı sağlayacak bilgi birikimimi kazandıran lisans ve yüksek lisans eğitimim sırasında üzerimde emeği bulunan tüm hocalarıma, tüm yakın arkadaşlarıma teşekkür ederim. Tez çalışmamın her adımında bana yardımcı olan, desteğini esirgemeyen ve adeta beni bilgi bombardımanına tutan Sayın Prof. Dr. Sezai TOKAT'a ve hayatımın vazgeçilmez unsurları olan Geniş Aileme ve yol arkadaşım eşime teşekkürü bir borç bilirim.

1. GİRİŞ

1.1 İletişim ve Sosyal Ağ

Haberleşme ve bildirişim olarak da adlandırılan iletişim, kısaca tanımlamak gerekirse insanlar arasındaki bilgi akışıdır. Latince ortak görüş anlamını içeren "communis" ifadesinden türetilen ve toplumsallaşma ve birliktelik anlamına gelen "communication" İngilizce sözcüğünün Türkçe karşılığı olarak kullanılan iletişim; hedef ile kaynak arasındaki beraberce kurulan anlam aktarma sürecidir (Demirel vd. 2011). İletişim tek bir kişinin sadece bir başka kişiye mesaj aktarmasıyla biten bir süreç değildir; bir bilgi paylaşımı söz konusudur ve alıcı da kendisine iletilen mesajı aldığına dair bir geri bildirimde bulunur. Bu ise iletişimin gönderici ve alıcı arasında bir mesaj alışverişi olduğunu gösterir (Kıraç, 2012). İletişim süreci, insanın ortaya çıkışından itibaren, bireysel ve toplumsal yaşamı ve gelişimi belirlemiş ve yönlendirmiştir (Yılmaz, 2003). İletişimdeki tüm taraflar düşüncelerini bir diğerine ifade ediyorsa buna açık iletişim, sadece bir taraf aktif olarak iletişimde bulunuyorsa buna kapalı iletişim denilmektedir.

Bireyler iletişim yetenekleri sayesinde birbirlerine görünmeyen karmaşık bağlarla bağlıdırlar ve bu bir sosyal ağ oluşturur (Scott, 1988). Genel olarak, sosyal ağ bir etkileşim eylemidir ve düğümlerin aktörlerden oluştuğu ve kenarların bu aktörler arasındaki ilişkilerden veya etkileşimlerden oluştuğu bir etkileşimler veya ilişkiler çizgesi olarak tanımlanabilir (Aggarwal, 2011). Sosyal ağ, fikir ve bilgileri paylaşmak, insanlarla bağlantı ve iletişim kurmak, bir topluluk duygusu oluşturmak için sanal bir alandır (Clemons vd, 2007). Bir sosyal ağ, bir sosyal sistemin üyeleri arasında var olan dostluk, tavsiye, iletişim veya desteği modellemektedir. Sosyal ağ analizi çalışmaları sosyolojide sınıf yapıları, uluslararası ticaret, bilimsel atıf, göç, salgın konularında uygulanmıştır (Scott, 1988). Sosyal ağı oluşturan aktörler sadece insanlar veya işletmeler olarak düşünülmemelidir. Web sayfaları, gazete makaleleri, ülkeler, bir işletme içerisindeki birimler de sosyal ağ olarak ele alınabilir (Scott and Carrington, 2011).

İnternet dünya genelinde birçok bilgisayar sisteminin birbirine bağlı olduğu, dünya çapında yaygın olan, sürekli büyüyen ve ağların ağı olarak da bilinen bir iletişim ağıdır (Vural, 2006). İnternet sunduğu olanaklarla geleneksel kitle iletişim araçlarından ayrılmaktadır (Bektaş Şeker, 2005). Televizyon ve basılı medya organları tek yönlü bir iletişim kurdukları için amaca ulaşma konusunda yetersiz olabilmekte veya tek yönlü bilgi akışı manipülatif amaçla kullanıma neden olabilmektedir (Solmaz vd., 2013). İnternet ortamında ise iletişim eş zamanlı ve iki yönlü sağlanmaktadır (Sayımer; 2008). Sosyal ağ siteleri web sitelerinden farklı özelliklere sahiptir. Rice Üniversitesi tarafından 2007 yılında sosyal ağ siteleri üzerine yapılan bir çalışmada sosyal ağ sitelerini normal bir web sitesinden ayırt eden beş temel özellik bulunduğu belirtilmiştir. Bu beş özelliği kullanıcı tabanlı olma, etkileşimli olma, topluluk odaklı olma, bireyler arası ilişkiler üzerinden büyüme ve içerikteki yalın bilgiye değil duyguya odaklanma olarak incelemişlerdir (Dube, 2011).

Kişilerarası iletişim, içinde yaşadığımız teknolojik olarak biçimlenmiş dünya tarafından şekillendirilmektedir (Qing, 2007). Bilgisayar teknolojilerinin ve İnternet'in gelişmesi ile sosyal ağ kavramı da bu yeniliklerden etkilenmiştir. Artık günümüzde çevrim-içi sosyal ağlar yeni bir araştırma konusu haline gelmiştir. İnsanlar kafeler, alışveriş merkezleri gibi fiziksel ortak alanlarda buluşup tanışmak yerine, ortak ilgi alanlarına ya da benzer fikirlere sahip gruplarla sosyal ağ sitelerinde bir araya gelmekte ve sanal topluluklar aracılığıyla ilişkiler kurmaktadır (America, 2013). Çevrim-içi sosyal ağ sitelerindeki örüntülerin incelenmesi, psikoloji, sosyoloji ve pazarlama gibi alanlar için büyük önem taşımaktadır. Picard (2000), bilgisayar bilimleri alanında uygun yanıtları elde etmek için insan duygularını modellemekle ilgili olarak duygusal bilgi-işlem (affective computing) veya sosyal bilgi-işlem terimlerini ortaya atarak bu konunun önemini vurgulamıştır.

1.2 Sosyal Medya

Sosyal medya bireylerin video, fotoğraf, görüntü, yazı, karikatür, fikir, dedikodu, haberler gibi içerikleri paylaşmak için kullandığı yaygın erişimli Internet-tabanlı ve mobil-tabanlı çevrim-içi kaynaklardır ve bu kaynaklar blogları,

vlogları, sosyal ağları, mesaj panolarını, podcastleri, içerik topluluklarını, sanal oyunları ve sanal sosyal dünyaları içermektedir.

Britannica Çevrimiçi Ansiklopedisi, kişisel web sayfaları, sayfa veya reklam başına maliyet, sabit banner reklamlar Web 1.0 dönemine ait iken, Web 2.0 ile bunların yerini etkileşimli bilgi kaynakları, bloglar, wikiler, ortak projeler, tıklama başına maliyet, içerik ile uyumlu reklamlar almıştır (Kaplan ve Hainlein, 2010; Genç, 2010). Örneğin bir sosyal medya platformu olan Wikipedia'nın, Web 2.0'ın kullanıcı hizmetine sunulmasıyla birlikte, tek yönlü bilgi paylaşımından, çift yönlü ve eş zamanlı bilgi paylaşımına ulaşılmasını sağlayan bir medya sistemi olarak tanımlanmaktadır. Kaplan ve Haenlien'ın (2010) tanımına göre sosyal medya, Web 2.0'ın ideolojik ve teknolojik temelleri üzerine kurulu ve kullanıcı içeriğinin oluşturulmasına ve değiştirilmesine izin veren İnternet'e dayalı bir grup uygulamadır (Kaplan ve Hainlein, 2010).

Web 2.0 üzerine kurulu yapısı ile bir İnternet bağlantısı olan tüm bireylerin ve tüzel kişiliklerin istediği iletişim mesajlarını üretebilme ve bunları dağıtma imkânına sahip olduğu bir ortam oluşturarak açık iletişim biçimlerini kolaylaştıran sosyal medya, geniş kitlelerle bilgi paylaşmanın veya onlara bilgi iletmenin etkin ve etkili bir yoludur. Yüksek derecede paylaşımın gerçekleştiği, çevrim-içi medyanın yeni bir türü olarak fırsatlar sunan sosyal medya, kamuya açık Web siteleri ile kullanıcılara düşünce, ilgi, deneyim ve bilgi paylaşım imkânı tanıyarak karşılıklı etkileşim yaratan çevrim-içi araçlar ve web siteleri için ortak kullanılan bir terim (Sayımer, 2008) ve İnternet dünyasını hızla hayatımıza yerleştiren bir uygulama alanıdır (Weinberg, 2009). Blackshaw ve Nazzaro (2004) ise sosyal medyayı diğer bireyleri ürünler, markalar, kişiler ve konular hakkında bilgilendirmek amacıyla tüketiciler tarafından yaratılan, başlatılan, dağıtılan ve kullanılan yeni çevrim-içi bilgi kaynakları olarak tanımlamışlardır. Buna göre sosyal medya; çeşitli çevrim-içi bloglarını, tüketici forumlarını, işletme sponsorlu tartışma panellerini ve sohbet odalarını, tüketiciden-tüketiciye e-postaları, tüketici ürün veya hizmet puanlama sitelerini, tartışma panellerini ve forumlarını, mobloglarını (dijital ses, görüntü, film veya fotoğraflar) ve sosyal ağ sitelerini kapsayan bir iletişim aracı olarak konumlandırılmaktadır (Mangold ve Faulds, 2009).

Bilgiye ve iletişimin taraflarına ulaşmaya yönelik engellerin olabildiğince az, geri-bildirimlere ve katılımcılara olabildiğince açık olan sosyal medya; oylama, yorum ve bilgi paylaşımı gibi konularda kullanıcıları cesaretlendirir ve ilgili olan her bir kullanıcıdan geri bildirim alır (Mayfield, 2010). Geleneksel kitle iletişim araçları yayına ilişkin iken (içerik aktarımı ya da dinleyiciye bilgi ulaştırma), sosyal medya iki-yönlü iletişime yönelik olması bakımından farklılık gösterir (Mayfield, 2010). Sosyal medya topluluklara çabuk ve etkili bir oluşum için izin verir. Topluluklar da böylece sevdikleri fotoğraf, politik tercihler, favori TV şovları gibi ilgili oldukları şeyleri paylaşırlar (Mayfield, 2010). Sosyal medyanın çoğu türü, bağlantılı işler gerçekleştirir; diğer siteler, araştırmalar ve insanların ilgili oldukları herhangi bir konuda bağlantı (link) verilmesine olanak tanır (Mayfield, 2010). Sosyal medya asenkron iletişime izin verir; tarafların iletişim için aynı anda karşılıklı iletişim halinde olmaları gerekmez. Sosyal medyada katılımcıların ortama istedikleri anda 24/7 erişimi vardır (Chang vd, 2013). Sosyal medya, ulaşım ve mesafe gibi engelleri aşarak, hareketlilik, konuşma veya işitme problemleri olan bireylerin de kolaylıkla çevrim-içi etkileşimde bulunmalarını sağlar (Chang vd, 2013). Ayrıca toplum için hassas konuları tartışmak için göreceli bir anonimlik verildiğinden, toplumsal olarak küçük düşürülme, damgalanma korkusu olmadan bireylerin kendilerini ifade etmesi sağlanmış olur (Chang vd, 2013).

Basılı kitle iletişim araçlarında bilgi kalıcıdır. Dergi, gazete basımı ve dağıtımı yapmak hükümetler veya güçlü özel sektör sermayesi ile bu konuda uzmanlaşmış kişiler aracılığı ile sağlanabilir. Sosyal medyada ise az bir maliyetle bir site açarak veya hazır servisler üzerinden herhangi bir maliyet olmadan herkes bilgi paylaşımı yapabilir. Gazete, dergi gibi kitle iletişim araçlarında bilginin taraflara dağıtımı yapıldıktan sonra bir değişiklik yapılması, erişimin engellenmesi zordur. Sosyal medyada ise yayınlama, değişiklik veya engelleme hızlı bir şekilde yapılabilmektedir. Günümüzde artık kitle iletişim araçları da sosyal medyayı ve İnternet'i etkili olarak kullanmaktadır.

Eski iletişim araçlarının aksine kullanıcıların etkileşim içinde olmasına olanak veren çevrim-içi araçlar olarak da ifade edilen sosyal medya araçlarına, forumlar, bloglar, wikiler, paylaşım siteleri, sosyal ağ siteleri, mikro-blog siteleri ve çevrimiçi sanal dünyalar örnek olarak gösterilebilir (Nash, 2009).

Akıllı telefonların yaygınlaşması, bilgisayar teknolojilerindeki gelişmeler ve Web 2.0 ile birlikte sağlanan kullanım etkinliği İnternet kullanımının artmasına bu ise Twitter, Instagram, Facebook, YouTube, Tumblr, Flickr, MySpace gibi sosyal medya uygulamalarının hızla benimsenmesine yol açmıştır. Bu olgunun bir sonucu olarak, sosyal medya, çağdaş öğretim yöntemlerinin, reklamcılık ve halkla ilişkilerin, politik kampanyaların ve çok sayıda başka unsurun ayrılmaz bir parçası haline gelmiştir. Sosyal medyanın gelişmesi ve genişlemesinin bilimsel çalışmalar ile analiz edilmesi güncel çalışma konularından biridir. (Al-Deen and Hendricks, 2012).

Sosyal medya uygulamaları olarak da bilinen sosyal ağ siteleri (social network sites), sosyal medya şemsiyesi altında insanların birbirleriyle etkileşim kurdukları ortamlardır. Sosyal ağ sitelerinin yaygın örnekleri olarak çevrimiçi fotoğraf paylaşım siteleri olan Instagram ve Flickr, bilgi, referans servisi olan Wikipedia, sosyal ağ servisi Facebook ve Myspace, mikro-blog sitesi Twitter, işaretleme ve etiketleme servisi del.icio.us ve çevrim-içi oyun olan World of War Craft gösterilebilir (Drury, 2008). Bir sosyal ağ sitesi, bireylerin sınırlı bir sistem içinde kamuya açık ya da yarı-açık profil oluşturmasına, bir bağlantı paylaştığı diğer kullanıcıların listesini eklemelerine ve kendi bağlantı listelerini görüntülemesine ve bu listelere ulaşmasına izin veren web tabanlı bir hizmettir (Boyd and Ellison, 2007). Sosyal ağ sitelerinin zaman çizelgesi üzerinde gösterilimi Şekil 1.1'de verilmiştir (Boyd and Elison, 2007).

Şekil 1.1'de verilen her bir sosyal ağ sitesi, profil ve uygulanabilirlik özellikleri açısından diğerlerinden ayırt edilebilir özelliklere sahiptir. Sitelerin çoğunluğu kullanıcıları bir profil fotoğrafı yüklemeye teşvik eder (Boyd ve Ellison, 2007). Bir sosyal ağ sitesinde başkalarında olmayan birçok özellik vardır. Sosyal ağ sitesi profillerindeki ilk fark, görünürlük derecesidir. MySpace ve Facebook gibi bazı siteler, yalnızca bir arkadaş ağındaki kişilerin bir kullanıcının profil sayfasını görüntülemesine izin verir; Friendster gibi diğer siteler ise arama motorları tarafından taranır ve izleyicinin bir hesabına sahip olup olmadığına bakılmaksızın onları herkese görünür kılar (Boyd ve Ellison, 2007). Bireylerin bir sosyal ağ sitesini seçerken en önem verdikleri konulardan biri bu görünürlük derecesidir ve kişinin kullanım amacına göre çeşitlilik gösterir. Profil gizliliğine ek olarak, sosyal

ağ siteleri, anlık mesajlaşma özellikleri, kişinin ulaşabildiği ağ listesi, diğer kullanıcıların profillerine doğrudan yorum yapma yeteneği ve medya paylaşım yetenekleri gibi diğer profil özellikleriyle birbirlerinden ayırt edilebilirler (Boyd ve Ellison, 2007).

Şekil 1.1: Birçok büyük sosyal ağ sitesinin lansman tarihleri ile topluluk sitelerinin SNS özellikleriyle yeniden başlatıldığı tarihlerin zaman çizelgesi (Boyd and Elison, 2007)

1.3 Sosyal Medya Analitiği

Çevrim-içi sosyal ağ sitelerinin yayılması ile birlikte İnternet'in karmaşık veri eko-sisteminde beğenme (like), paylaşma (share), yorum (comment), tweet, dürtme (poking) vb. büyük veri kaynakları ortaya çıkmaktadır (Farook and Abeysekara, 2016). Bu bilgiler sadece sosyal etkileşimleri içermekle kalmamakta aynı zamanda bir bütün olarak incelendiğinde toplumsal yönelişler ile ilgili bilgileri

de içerebilmektedir. Sosyal medya analitiği, bilişim araçları ve altyapısının sosyal medya verilerini toplamak, izlemek, analiz etmek, özetlemek ve görselleştirmek, konuşmaları ve etkileşimleri kolaylaştırmak, faydalı örüntüleri ortaya çıkarmak için geliştirilmesi ve değerlendirilmesiyle ilgilenir (Zeng vd, 2014).

Sosyal medya analitiği sınıflandırma, profilleme ve dinleme gibi farklı amaçlarla gerçekleştirilebilir. Sınıflandırma ve kümeleme algoritmaları, veri madenciliği sürecindeki temel algoritmalardır. Sınıflandırma ve kümeleme, görevin bazı veri nesneleri için önceden tanımlanmış sınıf etiketi bilgilerini kullanıp kullanmamasına göre farklılık gösterir. Denetimli öğrenmenin bir örneği olan sınıflandırma, eğitim verilerinde sınıf etiketleri kullanarak test verileri için sınıf etiketlerini öngörmeyi hedefler. Sınıflandırma ve kümelemenin sosyal medyadaki uygulamalarına örnek olarak duyarlılık analizi, spam algılama, çizge ve düğüm sınıflandırma sayılabilir. Sınıflandırmadan farklı olarak denetimsiz öğrenmenin bir örneği olan kümeleme ise veri nesneleri öznitelik uzayındaki benzerlik veya farklarına gruplandırılması çalışmalarını içerir. Sosyal medyadaki kümeleme uygulamaları arasında topluluk algılama, aykırı değer tespiti sayılabilir. Ayrıca, yarı denetimli öğrenme, çağrışımlı kural madenciliği ve öznitelik / örnek seçimi de veri analizi için faydalıdır.

Sosyal dinleme (social listening) veya sosyal medya dinleme; insanların elektronik ve sosyal kanallar aracılığıyla gerçek zamanlı olarak bir etkinliğe tepki verdiği sırada çeşitli uyaranlara katılmalarını, uyaranları gözlemlemelerini, yorumlamalarını ve uyaranlara yanıt vermelerini büyük miktarda veri toplayarak ve analiz ederek toplumun düşüncelerini kavramanın etkin bir süreci olarak tanımlanmaktadır (Stewart ve Arnold, 2018, Ituski vd., 2013). Sosyal dinleme sadece ticari alanda değil, politika oluşturma, seçim kampanyaları vb. gibi politik alanlarda da kullanılmaktadır (Ituski vd., 2013). Bu aşamada sosyal medyanın etkinliğini değerlendirmek için çeşitli ölçütler önerilmiştir (Hofman ve Fodor, 2010). Twitter dahil mikroblog platformları için basit ölçütler olarak tweet ve takipçilerin sayısı (marka bilinirliği için); takipçilerin ve cevapların sayısı (marka katılımı için); ve retweet (ağızdan ağıza) sayısı verilmiştir. Bu ölçütler önemli bilgiler sunsalar da sosyal medya döneminde önemini artıran daha güçlü tekniklerin yerini tutamazlar.

Sosyal medya profilleme sosyal medyanın ve büyük veri kavramlarının yaygınlaşması sonucunda ortaya çıkmıştır. Profilleme çalışmalarında daha iyi müşteri bölütlemesi oluşturmak amacıyla farklı kullanıcıların geçmişlerini, zevklerini ve satın alma davranışlarını derinlemesine anlamak için sosyal medya giderek daha yaygın şekilde kullanılmaktadır. Yapılan bu segmentasyon, her bir marka için bilinirliğini ve kullanımını artırmada farklı stratejiler oluşturmak amacıyla işletmelere çeşitli gruplara daha etkili bir şekilde ulaşmalarında yardımcı olur. Profili oluşturma hem ürün geliştirmede hem de tüketicinin desteklediği müşteri hizmetlerinde görüşleri oldukça değerli olan sosyal topluluk liderlerinin veya uzmanlarının belirlenmesinde de yardımcı olabilir. Sosyal ağ analizi, konu modellemesi ve görsel analiz dahil olmak üzere çeşitli teknikler sosyal profilleme çalışmalarını desteklemektedir.

Sosyal medya profillemede temel amaç sosyal medya kullanan kişilerin sosyal medyada belirli bir süre boyunca paylaştıkları bilgilerden yola çıkarak, öngörülen birtakım kurallar ve önerilere yönelik analizler yapılmasını sağlamaktır. Büyük veri araçlarının yaygınlaşması, Microsoft, Google, Amazon gibi büyük firmaların sunduğu makine öğrenmesi, büyük veri ve yapay zekâ sistemlerinin kullanımını kolaylaştıran yaygın uygulamaların herkese sunulması sayesinde sosyal medya profilleme veri bilimcilerin en önemli unsurlarından biri haline gelmiştir.

Sosyal medya profilleme sistemleri, günümüzde, karar alma mekanizmalarına olumlu katkıda bulunmak amacı ile reklamcılar, halkla ilişkiler uzmanları, işverenler gibi farklı meslekler, kurum veya kuruluş tarafından farklı amaçlarla kullanılmaktadır. Örneğin; işveren kurumlar yüzlerce başvuru arasından kısa zamanda kendilerine en uygun olan kişileri işe almak istemektedir. Bunun için işverenler, iş başvurusu yapan kişilerin sosyal medya profillerine bakarak bugüne kadar yaptığı paylaşımları, takip ettiği kişileri ve sayfaları, beğendiği herhangi bir yazıyı, kişinin cinsiyetini, yaşını ve daha birçok veriyi kullanarak, bu verilere göre kişi hakkında bir profil çıkarmaktadır. Bu çıkarımlar genellikle iş başvuru sürecinin bir parçası olmadan, bilimsel tutarlılık içermeden ve gizlice yapılıyor olmasına karşın, işe alımlarda sosyal medya değerlendirmesi ile ilgili bilimsel çalışmalar da yapılmaktadır (Ross ve Slovensky, 2012). Bu sosyal medya değerlendirmesi

neticesinde işverenler hızlı ve etkili bir şekilde bu kişinin uygunluğu için her işletmenin belirlediği algoritma ve kurallara göre karar verilebilmektedir (Hartwell, 2015). Benzer şekilde kredi kurumları, kredi kullanmadan önce kredi başvuru sahiplerinin kredi notu için sosyal medya profilleme sisteminden faydalanabilmektedir.

Sosyal medya profilleme, en çok kullanıldığı ve kazanç sağlanan unsur olan reklamlarda da çok etkili bir şekilde fayda sağlamaktadır. Teknoloji devleri Google ve Facebook dijital reklam gelirlerinden büyük bir pay elde etmekte ve kazançları her geçen gün artmaktadır. Örneğin; Google "statista" verilerine göre (Clement, 2019) 2001 yılından 2018 yılına kadar kazançları katlanarak artmaktadır. Bu sebeple reklam konusu sosyal medya profilleme üzerinde büyük öneme sahiptir. Reklam veren firmaların amacı verdikleri reklamı gerçekten ilgi duyan kişilere gösterebilmektedir. Bunun için de reklamı gösterecek olan firmalar kişileri çok iyi profilleyecek sistemlere ihtiyaç duymaktadır.

Makine öğrenmesi sürecindeki gözetimli öğrenme ile yapılan sınıflandırma işlemi için verinin etiketlenmiş olması gerekmektedir. Sosyal medya profillemede en büyük problem gereksinimlere göre etiketlenmiş eğitim kümesinin belirlenmesidir. Bu etiketler çekilen verilerden otomatik olarak oluşturulabilir veya anket aracılığı ile doğrudan kişilere ulaşılarak etiketleme yapılabilir. Örneğin; bireyleri pizza sevme eğilimlerine göre profillemek istediğimizde, kişi eğer "pizza yemeyi çok seviyorum" gibi bir yazı paylaştıysa, bir pizza resmi paylaştıysa veya pizza sevenler isimli bir grubu takip ediyorsa bu kişinin pizza sevdiği öngörüsünde bulunulabilir. Fakat bu durum, ilgili kişinin gerçekten pizza sevip sevmediği konusunda, kişilere anketle yapılarak elde edildiği doğrulukta bir bilgi sunmayabilir. Bu tür analiz çalışmalarının yapılabilmesi için önceden kişilere pizza sevip sevmediğinin sorulması ve pizza seven kişiler için bu kişilerin ortak özelliklerinin çıkarılması ve etiketlerin oluşturulması gerekmektedir. X sayfasını takip ediyor, Y kişinin paylaşımlarını beğeniyor, Z sayfası pizza seven kişimizi takip ediyor ve bu tüm pizza seven kişilerde ortak ise x sayfasını takip eden, y kişinin paylaşımlarını beğenen, z pizza sayfası takip ediyorsa bu kişi pizza seviyor grubundadır diyebiliriz. Etiketleme işlemi için bu iki yöntem karma bir şekilde de kullanılabilir.

Sosyal medya profilleme çalışmalarında en çok kullanılan algoritma türleri sınıflandırma algoritmalarıdır. Sınıflandırma algoritmaları var olan verileri başlangıçta belirli olan farklı gruplara ayırma işlemini gerçekleştirir. Kullanılan bir sınıflandırma algoritması resimleri tanımlamada kullanılabileceği gibi yazılımda herhangi bir değişiklik yapmadan aynı sınıflandırma algoritması farklı bir amaç için de kullanılabilir. Örneğin hayvan resimlerini ayırt etmek için kullanılan bir sınıflandırma algoritması e-posta servislerinde kullanılan spam (istenmeyen e-posta) ayırt etme mekanizması için de kullanılabilmektedir.

Günümüz teknolojisinin gelişmesi, veri analizi için özel bilgisayarların ve yazılımların ücretli veya ücretsiz bir şekilde sunulmasından dolayı veri analizleri çok elverişli hale gelmiştir. Tüm bu durumlar firmaların kazanç sağlama istemleri ile birleşince sosyal medya profilleme en önemli unsur ve yükselen bir değer haline gelmiştir.

1.4 Tezin İlgi Alanı

Pear Analytics (2009) çalışmasında Twitter'daki tweetlerin %50,9'unun İnternet ortamında ve gerçek dünyada değişik düşünceleri harekete geçirebilecek bazı yararlı bilgilere sahip olduğu kestirilmiştir. Bu nedenle, sosyal medya kullanıcılarının görüşleri farklı organizasyonlar için büyük stratejik değere sahiptir. Toplumun her kesiminden bireylerin ve tüzel kişiliklerin görüş ve düşüncelerini paylaştığı, bu görüş ve düşüncelerin diğer insanlar tarafından hızla ve rahatlıkla görülebildiği sosyal medyanın iletişimdeki gücünü siyasi kuruluşların liderleri de görmüş ve sosyal medya ortamlarında yer almaya ve kitlelerine buradan mesajlar göndermeye, iletişime geçmeye başlamışlardır (Özay, 2018). Sosyal medya günümüzde hem dünyada hem de ülkemizde siyasi kuruluşlar ve onların üye, gönüllü, yönetici ve lider bireyleri tarafından etkin olarak kullanılmaktadır. Derin öğrenme, yapay sinir ağlarına göre daha fazla katmandan oluşan, daha yüksek soyutlama seviyelerine izin veren ve verilerden gelişmiş tahminler yapılmasını sağlayan yapısı ile yapay sinir ağlarının gelişmiş bir biçimidir. Bu tez çalışmasında, derin öğrenme ve sıkça kullanılan k-NN, Naive Bayes ve Rassal Orman

algoritmalarının sosyal medya profilleme çalışmalarında sıradan bireylerin siyasi eğilimlerini tahmin etmek için kullanılması üzerinde durulmaktadır.

1.5 Tezin Amacı

Bu tez çalışmasında güncel bir mikro-blog hizmeti olan Twitter'dan elde edilecek veriler üzerinde derin öğrenmeye dayalı sosyal medya profilleme çalışması yapılması amaçlanmıştır. Tezde sosyal medya profilleme ile ilgili yapılmış akademik çalışmalar açıklanmış, derin öğrenmeden ve tezde ele alınan makine öğrenmesi algoritmalarından bahsedilmiş, kullanılan algoritmalarla birlikte elde edilen sonuçlar analiz edilmiştir.

1.6 Tezin Akışı

Tezin ikinci bölümünde Twitter ve sosyal medya profilleme ile ilgili çalışmalar incelenmiş, siyasi görüşün tahmin edilmesi ile ilgili yapılan çalışmalar ayrı bir başlık altında verilmiştir. Üçüncü bölümde tez çalışmasında büyük veri, yapay zekâ, Twitter'dan veri çekme amacı ile kullanılan teknolojiler kısaca tanıtılmıştır. Dördüncü bölümde verinin elde edilmesi, ön işlemler, yapay zekâ ile tahmin sonuçları verilmiş ve sonuçlar analiz edilmiştir. Sonuç bölümünde yapılan çalışma genel olarak değerlendirilmiş ve ileriye dönük çalışma konuları üzerinde durulmuştur.

2. TWITTER İLE SOSYAL MEDYA PROFİLLEME

2.1 Sosyal Medya Profilleme

Gerçek-zamanlı olarak yapılan sosyal dinleme (social listening) eyleminde geleneksel sorgulamadan farklı olarak veriler zaman serisi analizleri için de uygundur. Sosyal dinleme elbette her problemin çözümü olamaz. İlk olarak, sosyal medyadan seçilen bireylerden oluşan popülasyon, telefon ya da şahsi anket yoluyla gerçekleştirilmiş bir kamuoyu anketinde olduğu kadar net değildir. İkincisi, seçilen popülasyon ağırlıklı olarak İnternet ve sosyal medyayı kullanabilen belirli özelliklere sahip insanları temsil etmektedir. Bu sorunlara rağmen, siyasi ve seçim faaliyetlerini analiz etmek için sosyal dinleme yaygın olarak kullanılmaktadır (Ituski vd., 2013). Bir sosyal dinleme problemi olan sosyal medya profilleme konusu akademik olarak güncel bir çalışma konusudur.

Profilleme (profilling), kendisi hakkında önemli veya ilginç bilgileri içeren bir nesnenin betimlenmesini otomatik olarak oluşturmayı amaçlar (Schiaffino ve Amandi, 2009). Profili oluşturulacak nesne hakkındaki bilginin toplanması, temizlenmesi ve organize edilmesi ile ilgili süreçler otomatik veya yarı-otomatik olarak ele alınır. Profillemenin amacı uygulamaya göre farklılık gösterebilir. Bu amaçlardan bazıları kullanıcı profilleme, grup profilleme veya ilişki profilleme olarak adlandırılabilir (Hu ve Liu, 2015). Kullanıcı profillemede kullanıcı bilgileri toplanarak kullanıcı profiline ait belirli değerler oluşturulmaya çalışılır. Kullanıcı profillemede kişinin yaş, cinsiyet, eğitim, gelir, meslek, medeni durum, din, ırk, etnisite, dil, bölge/konum, şehir, milliyet vb. gibi demografik bilgilerine göre, kişilik, davranış, ruh hali, duygu, alışkanlık, sosyal etki, öncelik, sosyal bağlantı, etkinlik, ilgi, görüş, değer, tutum gibi psikografik bilgilerine göre vücut kitle indeksi, hastalık eğilimi gibi sağlık bilgilerine göre yapılan çalışmalar bulunmaktadır (Bilal, 2019). Reklamcılık, pazarlama ve tavsiye sistemlerinde önemli bir yeri vardır. Örneğin Twitter kullancılarının hesaplarında kullandıkları metin etiketlerinden kullanıcı ile ilgili bilgilere ulaşılabilmektedir (Hu ve Liu, 2015). Bir kullanıcının bu tür bilgilerinden elde edilecek ilgilendiği şehir ve konum bilgilerinden seyahat ile ilgili tavsiye sistemleri ile bu şehir ve konumlar ile ilgili

turizm bilgilerine ulaşması sağlanabilir. Twitter üzerinde makine öğrenmesi, veri madenciliği ve veri bilimi teknikleri kullanılarak yapılan bir çalışmada insanların attıkları tweetler ve Myers-Briggs kişilik tipi göstergesi (Briggs ve Myers, 1988) kullanılarak kişilik analizleri yapılmaya çalışılmıştır. Çalışma için ilk olarak kişilik tipi etiketlenmiş olan 64 kişiye dair 16 MBTI bilgisini de içeren 63384 tweeti toplanmıştır. Toplanan veriler üzerinden word-gram yöntemi ile özellik çıkarımı yapılmıştır (Şeker, 2015). Toplanan bu tweetler ile birçok algoritma kullanılarak denemeler yapılmış ve Naive Bayes, Random Tree, ve Gradient Boosted Tree algoritmalarından sonuç alınabilmiştir. En başarılı sonuç ise %54 oranı ile Naive-Bayes yöntemi olmuştur (Bastem ve Şeker, 2017).

Sosyal medya kaynaklarından yararlanarak yapılan kişilik profilleme çalışmasında; herhangi bir kaynaktan elde edilen (Twitter, blogger) verilere göre paylaşılan metnin uzunluğu, kullanılan kısaltmalar, söz dizimi kuralları, imla kuralları, dil bilgisi hataları gibi bilgiler kullanılarak kullanıcıların kişilik tiplerine ulaşılması hedeflenmiştir. (Chin ve Wright, 2014; Argamon vd., 2005). Ikeda vd. (2013) tarafından, Twitter kullanıcılarının demografik tahmini için metin tabanlı ve tweet geçmişi ve takipçi/takip edilen kümelerinden yararlanan topluluk temelli karma bir yöntem önerilmiştir. 100.000 Twitter kullanıcısından elde edilen deneysel sonuçlar, önerilen karma yöntemin sadece metin tabanlı yöntem kullanmaya göre doğruluğu artırdığını göstermektedir. Rao et al. (2010) Twitter'da düz metin tweet bilgisinden kullanıcı profil bilgisi, kullanıcı tweet davranışları (retweet frekansı), sosyal arkadaş ağ yapısı ve dilsel içerikten yararlanarak cinsiyet ve siyasi görüşün tahmini üzerine çalışmıştır. Pennacchiotti ve Popescu (2011) bu çalışmayı duygu analizi ve makine öğrenmesi yöntemleri kullanarak iyileştirmeye çalışmıştır.

Chen vd. (2010) istatistiksel modeller kullanılarak kullanıcı ilgi alanları üzerine bir profilleme sayesinde Twitter kullanıcısına URL tavsiyeleri yapmaya çalışmıştır. Bir başka çalışmada ise bilginin ve örüntülerin çoklu-etmen, çoklu bakış açısı veya çoklu veri kaynağı içeren teknikler kullanılarak edinildiği işbirlikçi filtreleme (coloborative filtering) ile Twitter kullanıcılarına takipçi tavsiyelerinde bulunulmaya çalışılmıştır (Hannon vd., 2010).

Kullanıcı profilleme çalışmalarının sınıflandırıldığı ve kullanılan öznitelikler ve veri kaynaklarının kategorilendirildiği bir çalışmada daha sonra

kullanıcı sosyal ağ davranışı, ağ trafiği vb. bilgiler kullanılarak siber-güvenlik amacı ile kullanıcı profilleme yapılmaya çalışılmıştır (Lashkari vd, 2019).

Grup profillemede karmaşık ilişkilerin bulunduğu topluluk ağlarında topluluğun belirli özelliklerinden tanımlayıcı profiller elde edilmeye çalışılır (Gomes vd, 2016). Grup profillemede kullanıcının değil ilgili grubun belirleyici nitelikleri ortaya çıkarılmaya çalışılır (Tang vd., 2011). Örneğin dil öğrenmeye yardımcı olacak şekilde kullanıcıları öğrenme biçimleri ve yeteneklerine göre kümeleyen bir yapı grup profilleme çalışması kapsamında incelenebilir (Troussas vd, 2013). Rhim vd. (2016) de cep telefonu kullanıcılarının grup profillerini oluşturarak cep telefonu kullanıcısı bir kişiye tavsiye sistemi oluşturmaya çalışmıştır.

İlişki profillemede kullanıcılar arasındaki ilişki çeşidi tanımlanmaya çalışılır. Kullanıcılar arası ilişki tipi (arkadaşlık, akrabalık, iş), ilişki derecesi (yakın, uzak vb), ilişki etkisi (etkilenen, etkileyen, lider, takipçi, vb.) gibi konularda profil çalışması yapılmaya çalışılır.

Sosyal medya, sosyal medya profilleme, sosyal medya analitiği konuları bilimsel literatürde üzerinde yoğun olarak çalışılan konulardır (Fan ve Gordon, 2014). 2012-2019 yılları arasında “social media analytics”, “social media profilling”, “social media” gibi bazı anahtar kelimeler için terimler tırnak içinde yazılarak Google Akademik üzerinde yapılan arama sonuçları Tablo 2.1’de verilmiştir. Tablo 2.1’den görüldüğü gibi sosyal medya konusu ve analitiği ile ilgili yapılan çalışmalar artarak devam etmektedir.

Sosyal medya profillemede derin öğrenme kullanılması üzerinde de literatür çalışmaları bulunmaktadır. Xue vd. (2018) onbir milyon Facebook kullanıcısnın kişilik özelliklerini metinsel veriler yoluyla anlamak için derin öğrenme yaklaşımı ortaya koymuştur. Segalin vd. (2017), Flickr kullanıcılarına ait görüntü bilgisinden yararlanarak yine kişilik özelliklerini belirlemeye yönelik bir sosyal medya profillemesi yapmıştır. Tang vd. (2014) Twitter’da duygu analizi için 10 milton tweet kullanılarak elde edilen mesaj metinleri kullanmış ve karma bir kayıp fonksiyonu kullanan derin öğrenme algoritması tasarlanmıştır. İslam ve Zhang

(2016), Twitter'dan çekilen 1269 görüntü üzerinde evrişimsel sinir ağları kullanarak görsel içerikten duygu analizi üzerinde durmuştur.

Tablo 2.1: Google akademik literatür tarama sayıları

Yıllar	Sosyal medya(social media)	Sosyal medya analizi(social media analytics)	Sosyal medya profillemesi(social media profilling)
2012	187.000	657	3
2013	215.000	1010	13
2014	284.000	1500	36
2015	306.000	1870	62
2016	202.000	2220	74
2017	147.000	2580	73
2018	110.000	2710	73
2019 (ilk 4 ay)	47.100	891	24

2.2 Twitter

Dube (2011) tarafından listelenen Rice Üniversitesinin sosyal ağ sitelerinin ayırt edici beş özelliğini, Twitter açısından tekrar ele alan America, (2013), Twitter'ın dört temel özelliğini kullanıcı tabanlı olma, etkileşimli olma, topluluk odaklı olma, bireyler arası ilişkiler üzerinden büyüme olarak listelemiştir.

Profilleme çalışması sırasında kullanıcıya ait bilgiler toplanır. Örneğin kullanıcının ilişki (relationship) bilgileri sıklıkla hem grup profilleme hem de kullanıcı profillemede kullanılır. Twitter için ilişki bilgisi takipçi (follower), takip edilen (following) kullanıcılar ile elde edilir. Takipçilerin takipçileri, takipçilerin takip ettikleri gibi ikinci seviye veya daha derin bilgilere de ulaşılarak daha ayrıntılı bilgi tabanları da oluşturulabilir. Bir kullanıcının kendisini diğer kullanıcıların Twitter'da takip etmesini sağlayabilmek amacı ile diğer kişilere istek göndermesi vb. gibi bir yöntem Twitter'da tanımlanmamıştır. Bunu sağlayan yardımcı yazılımlar ise Twitter tarafından hak ihlali olarak görülmekte ve kullanıcının hesabı askıya alınabilmektedir (Twitter Yardım Merkezi, 2019a).

Twitter'da kullanıcı gönderilerine tweet adı verilmektedir. Tweet özünde kısa biçimli bir mesajlaşma şeklidir. Mesaj uzunluğu sınırlı bir alanda karşılıklı iletişim sağlayan tüm sosyal ağ sitelerinde olduğu gibi bir mikro-blog sitesi olan

Twitter için de mesajlaşma gönderi uzunluklarının ne olacağı önemli bir stratejik konudur.

2017 yılına kadar Twitter gönderi sınırı 140 karakter iken Korece gibi bazı alfabeler dışında bu sınır tüm tweetler için 280 karaktere çıkarılmıştır. Görüntüler ve videolar bu sınırı etkilememektedir. Twitter'da doğrudan mesajlar (direct messages) 10.000 karakter uzunluğa kadar olabilmektedir. Twitter gönderi sınırı 280 karakter olmasına rağmen Buddy Media tarafından yapılan analizler sonucunda 71-100 karakter arası tweetlerin bir şekilde yerini bulması veya retweet edilmesinin diğerlerine göre %17 daha yüksek olduğu sonucuna ulaşılmıştır. Gönderi uzunluğunun uygun değeri ile ilgili çalışmalar farklı sosyal ağ siteleri için de incelenmiştir. Örneğin durum (statue), olay gönderisi (event post), reklamlar (ads) gibi farklı gönderi tipleri bulunan Facebook için birçok gönderi tipi için maksimum karakter sınırı 63.206 karakter uzunluğundadır ve görüntü ve video kullanımı için bir sınırlama da yoktur. Fakat Facebook uzmanları ve HubSpot gibi pazarlama şirketlerinin yaptıkları analizler sonucunda Facebook gönderilerinin 1-40 karakter arasında olduğunda daha uzun gönderilere göre %86 daha fazla bağlantı sağladığı görülmüştür (Social Report, 2019).

Twitter'da birisini etiketlemek istendiğinde veya yönlendirilmiş iletişim söz konusu ise genellikle "@kullanıcı" sözdizimini kullanılır. Bu ifade ilgili mesajda diğer kullanıcının da adreslenmesini sağlar (Honeycutt and Herring, 2009). Bunlara mention denilmektedir. Eğer "@kullanıcı" sözdiziminden oluşan mention tweet'in en başında yazılırsa tweet'i yazan kullanıcıyı takip eden tüm kullanıcıların zaman akışında (timeline) bu tweet görüntülenmez. Sadece hem tweet atan kişiyi hem de mention'da geçen kullanıcıyı takip edenlerin zaman akışında görülür. Mention ortaya veya sona yazıldığında ise tüm takipçier tarafından görülür. Örnek vermek gerekirse vasfi_tataroglu kullanıcı tarafından atılan iki tweet aşağıdaki gibi olsun:

Tweet-1: @yunus_sarica akşam maç için buluşuyoruz değil mi kanka?

Tweet-2: bence bu maçın en iyisi @ronaldo ve @mehmet_topal

Yukarıda örnek olarak verilen ilk tweet sadece yunus_sarica ve vasfi_tataroglu ile ikisini de takip eden kullanıcıların zaman akışında görülür. İkinci tweet ise tüm takipçiler tarafından görülebilecektir.

Hashtag (#) işaretleri ile de konusuna göre tweetler işaretlenerek belirli bir konu üzerindeki konuşmaların diğer kullanıcılar tarafından da kolayca takip edilmesi sağlanmış olur.

Twitter'da takipçilerinizle herkese açık olarak paylaştığınız bir tweet, retweet olarak adlandırılır ve kullanıcının ilginç bulduğu haberleri ve yeni bilgileri takipçilerine iletmesi için kullanılır. Bu kullanım amacı e-posta kavramındaki yönlendir (forward) işlemine benzetilebilir (Boyd vd, 2010). Retweet yapmadan önce kullanıcı kendi yorum veya medya verilerini de ekleyebilmektedir. Twitter'ın Retweet simgesi kullanıldığında, Retweet veya yorum eklenen Retweet, paylaşılan Tweeti referans alır. Yorum eklenen Retweet'e biri yanıt verdiği zaman, orijinal tweetin yazarı otomatik olarak sohbete eklenmez. Orijinal tweet'in yazarını eklemek için kullanıcı adının retweet içerisinde geçmesi gerekir. Twitter'da başkalarının tweetlerini paylaşmanın yanında, kullanıcılar kendi Tweeti ile birlikte Retweetleyebilmekte veya yorum ekleyerek Retweetleyebilmektedir. Bu işlev özellikle gündemle yeniden ilgili hale gelen eski Tweetlerinizden birini paylaşmak ya da tüm takipçilerinizin görmesini sağlamak amacıyla diğer kişilere verdiğiniz yanıtları Retweetlemek için faydalı bir özelliktir (Twitter Yardım Merkezi, 2019b).

Twitter'da Twitter API'leri yardımı ile şu bilgiler çekilebilir (Vergeer, 2015):

- Tweet zamanı
- Tweet tarihi
- Tweet yollayan kullanıcı
- Konum (kullanıcı tarafından izin verildiyse)
- Kullanılan uygulama
- Tweet'in retweet yapılma sayısı
- Tweet'in favori yapılma sayısı

Twitter'da tweet dışında kullanıcının kendisini tanıtmak için kullanacağı bölümler profil fotoğrafı ve maksimum 160 karakter uzunluğuna sahip kişisel

bilgiler bölümüdür (bio). Takip etme kararı için genellikle ilk bakılan alan olduğu için önemlidir. Örneğin Alashammari (2019) kişilerin favori tweet, zaman akışı tweet ve arkadaş listesi bilgilerinden yararlanarak etkisiz, az etkili ve çok etkili kullanıcı olarak profillemek için bir etki ölçme metriği geliştirmiştir. Kullanıcının oluşturduğu tweet'te geçen bağlantıların içeriğinden de yararlanarak dış kaynak kullanımıyla zenginleştirilen kullanıcı profillerinin, yalnızca Twitter'ın etkinliklerine dayandırılan profilleri geride bıraktığı görülmüştür (Esparza vd., 2013).

2.3 Twitter'da Politik Görüş Üzerine Yapılan Çalışmalar

Sosyal medya; sıradan vatandaşların siyasi figürler ve seçkinlerle iletişim kurmasını ve onları desteklemelerini sağladığı için, çevrim-içi sosyal medya araçlarının bu süreçte ürettikleri etkileşim kalıpları özellikle bireylerin benzer fikirli insanlarla etkileşime girme eğiliminde olduğu varsayımıyla ele alınırsa, sosyal medya kullanıcılarının ideolojik tercihleri hakkında zengin bilgiler içerebilmektedir (Briatte ve Gallic, 2015).

Çeşitli konularda halkın görüşünün ne olduğunu izleyebilmek için sosyal medyadan yararlanılması önemli bir konudur. Bu konuda üzerinde en çok çalışma yapılan çevrim-içi sosyal ağ sitelerinin Twitter ve Facebook olduğu dikkat çekmektedir. Twitter'ı, stratejiye dayalı ilişkiler sunan ve onu Facebook gibi karşılıklı (reciprocal) ağa dayanan diğer klasik sosyal ağ platformlarından farklı kılan bir sosyal medya web sitesi biçimi olarak görmek mümkündür (Alshammari, 2019). Twitter kullanıcıları arasında var olan ilişkiler, sadece bilgilendirme amaçlı veya sadece sosyal amaçlı olabileceği gibi her iki amaçla da olabilir. Bunun nedeni, bilgi edinme temel amacı ile kullanıcıların her zaman hem etkileşimler hem de ilişkiler ağında aktif rol alan diğer kullanıcıları takip etmeleridir (Abel vd. 2011; Vosoughi, 2015).

Halk arasındaki genel popülaritesinin yanında Twitter; kısa mesajlar ve medya eklentileri yoluyla birbirleriyle ve parti destekçileri gibi daha geniş kitlelerle iletişim kurmak için Twitter kullanmakta olan birçok siyasi partiyi, parti liderlerini ve adaylarını cezbetmektedir. Siyasetçiler basın yayın organlarına ve gazetecilere

erişimlerini sınırlayan kurumsal kısıtlamaları atlamak için de Twitter'a yönelmektedirler (Briatte ve Gallic, 2015). Politik reklam harcamalarının tüm dünyada her geçen gün arttığı da bir gerçektir. Bu durum da siyasi kampanyalar sırasında seçmenlerle bağlantı kurmak ve kullanıcılar arasındaki politik temelde katılımları teşvik etmek için düşük maliyetli bir platform olarak, sosyal medya araçlarına daha fazla önem verilmesinin bir başka nedenidir (Conover vd., 2011).

Twitter kullanıcılarının politik eğilimlerini tahmin etmek için kullanılan farklı yaklaşımlar incelendiğinde temel olarak tweet metin içeriklerinin, kullanıcı davranışlarının (tweet ve retweet hakkında nicel bilgiler) ve Twitter yapısını (kullanıcının takipçileri ve takip ettikleri hakkında nicel bilgiler) kapsayan özelliklerin araştırma konusu olduğu görülmektedir (Pla ve Hurtado, 2014).

Günümüzde seçimlerin yapıldığı demokratik tüm ülkelerde politikacılar ve vatandaşlar arasındaki uçurumun giderek arttığı, özellikle Avrupa ülkeleri olmak üzere birçok ülkede seçmen katılımının (Blais ve Rubenson 2013) ve siyasi kurumlara ve politikacılara duyulan güvenin (Dalton 2004) zamanla azaldığı görülmektedir. Bu demokratik zaafın bir nedeni, artık kampanya strateji uzmanları ve reklam ajansları aracılığı ile yürütülen iletişimin siyasetçilerin temsil ettikleri kişilerle temaslarını kaybetmesine neden olmasıdır (Anderson ve McLeod 2004). Genç nesil ve yeni kuşak gençliğin iletişim şekli de değişmektedir. Önceden kafelerde, spor sahalarında oluşturulan arkadaşlıklar artık mobil cihazlarla çevrim-içi sosyal medyada gerçeklenmektedir. Yeni nesil ile uzun vadeli kalıcı bir iletişim kurmak ve oylarını arttırmak için siyasetçilerin de bu ortamı en iyi şekilde kullanmaları gerekmektedir. Gazete, dergi gibi basılı kitle iletişim araçları ve siyaset mitinglerinde yapılan siyasi propaganda genellikle ilgili siyasi gruba yakın olan kişilere hitap etmektedir. Televizyon gibi pahalı kitle iletişim araçları kararsız seçmeni de hedef almaktadır. Internet'ten yararlanan günümüz kampanyaları ise daha bireysel kampanyalarla kişiye özel hitap edilmesini sağlayan yeni bir siyasi kampanya devrini başlatmıştır (Dennis, 2019; Wei ve Xu, 2019). Wegrzyn-Wolska ve Bougueroua (2012), farklı eğilimlerin sosyal medyadaki kitleleri farklı anket yöntemleri kullanarak nasıl etkilediğini analiz etmek amacı ile 2012'deki Fransa cumhurbaşkanlığı seçimlerinden önce bir çalışma yapmıştır.

Twitter kullanıcılarının siyasi özelliklerini tahmin etme konusu üzerine yapılan çalışmalar incelendiğinde duyarlılık analizine, ideolojiyi tahmin etmeye, belirli bir siyasi olayla ilgili siyasi duruşun tahmini veya Twitter'ın etkilerini analiz etmeye, otomatik anketlere ve Twitter'ı kullanarak uzak mesafedeki denetimlerin kullanımına ilişkin politik tahminlere odaklanıldığı görülmektedir.

Twitter'da duygu analizi kullanılarak siyasi görüşlerin tahmin edilmesi üzerine çalışmalar da yapılmıştır. Örneğin Pla ve Hurtado (2014), dünya siyasetinde, ekonomisinde, medya veya kültür dünyasında tanınmış 158 kişiye ait İspanyolca 68.000 Twitter mesajı üzerinde sözlük tabanlı bir duygu analizi yaparak öncelikle öznitelik çıkarımı yapmış ve bireyleri sağ görüşlü, sol görüşlü, merkez görüşlü ve görüşü tanımsız olmak üzere dört farklı kategoride sınıflandırmıştır. Bakliwal vd., (2013) Şubat 2011'de İrlanda genel seçimleri öncesinde üretilen 2.624 tweet üzerinde duygu analizi yöntemleri ile pozitif, negatif ve nötr duyarlılık sınıflandırması gerçekleştirmiştir. Alaycı tweetler bu setten çıkarılmış olmasına rağmen zorlu bir test kümesini temsil eden veri seti ile %61.6 doğruluk elde edilmiştir (Bakliwal vd., 2013).

Siyasi özellikleri tahmin etme konusunda Fernandes de Mello Araújo ve Ebbelaar (2018) yaptıkları çalışmada, tweetleri politik ve politik olmayan olarak sınıflandırmak için makine öğrenmesine dayalı bir yöntem önermişlerdir. Bu amaçla, etiketli eğitim verileri ile denetimli öğrenme yaklaşımı kullanılmış ve Twitter'daki politik içeriğin sınıflandırılmasının kural temelli bir yöntemden daha iyi performans gösterip göstermediği incelenmiştir. Sınıflandırıcının oluşturulması için, ilk olarak iki aylık bir süre zarfında 2.881 Felemenkçe tweet toplanmıştır. Korpus, bu proje için oluşturulmuş bir web uygulaması kullanılarak elle etiketlenmiştir. Daha sonra tweetler ön-işlemden geçirilmiş ve sınıflandırmayı iyileştirmek için meta verilerden ek özellikler çıkarılmıştır. Etiketli veri seti kullanılarak çeşitli makine öğrenmesi algoritmaları eğitilmiş ve doğru modelleri bulmak için sonuçlar karşılaştırılmıştır. Sonra da en iyi performans gösteren beş model oylama sistemi kullanan bir sınıflandırıcı oluşturmak için birleştirilmiştir (Fernandes de Mello Araújo and Ebbelaar, 2018).

Twitter kullanıcılarının tümünün oy kullanma hakkına sahip bireyleri temsil etmediği durumu göz önüne alarak, Dwi Prasetyo ve Hauff (2015) ile Sanders vd.

(2016) demografik dağılımları düzeltmek için bir yöntem geliştirmişlerdir. Dwi Prasetyo ve Hauff (2015) Twitter'daki erkek nüfusu egemenliği kaynaklı sapmayı azaltmak için kadınlardan gelen tweetlere daha yüksek ağırlık vermişler ve bu ayarlamanın ortalama mutlak hatayı %3.3'ten %1.99'a düşürerek tahmin doğruluğunu artırdığını tespit etmişlerdir.

Twitter kullanıcılarının profillenmesinde makine öğrenmesi yaklaşımlarının kullanılmasına yönelik yapılmış bir çalışmada; kullanıcıların durumları, ağ yapıları ve dil içeriklerinden çıkarımlar yapılarak değerler oluşturulmuş ve politik yönelimleri, etnik yapıları hakkında sonuca ulaşmak amaçlanmıştır. Makine öğrenmesi yöntemleri kullanılarak umut verici deneysel sonuçlar rapor edilmiştir. Makine öğrenmesi yöntemlerinden Gradient Boosted Decision Trees (Friedman, 2001) kullanılmıştır (Pennacchiotti ve Popescu, 2011).

Twitter ile seçim sonuçlarının öngörülmesine yönelik yapılan bir çalışmada; Almanya Federal bölge seçim içeriklerinden yararlanılarak bir sonuç elde edilmek istenmiştir. Metin analizi için LIWC (Linguistic Inquiry and Word Count) yöntemi kullanılmıştır (Pennebaker vd., 2007). LIWC, psikometrik olarak doğrulanmış bir iç sözlük kullanarak metin örneklerinin duygusal, bilişsel ve yapısal bileşenlerini değerlendirmek için geliştirilmiş bir metin analiz yazılımıdır. Tumasjan vd. (2010) yaptıkları çalışmada yüz binden fazla mesajı incelemiş ve çalışma sonucunda Twitter'ın kullanıcıların politik düşüncelerinin belirlenmesinde kullanılabileceği ve bir partinin tweet / mention sayısının seçimleri kazanma olasılığı ile doğru orantılı olduğu sonucuna varılmıştır (Tumasjan vd., 2010). Bhola (2014), Hindistan'da 2014 yılında yapılan genel seçimlerde, seçim başladıktan sonra çekilen Twitter verisi kullanarak dikkat çekici örüntüler elde etmeye çalışmıştır. İkisi kullanıcı tarafından yapılan tweetlerin içeriğine, birisi kullanıcı tabanlı özelliklere, diğeri ise retweet ve kullanıcı tarafından bahsedilen ağlarda topluluk algılama algoritmasına dayalı dört farklı algoritma kullanmışlardır. Topluluk algılama algoritmasının % 80'den fazla bir verimle en iyi şekilde çalıştığını tespit etmişlerdir. İçerik temelli yöntemlerin sınıflandırma sonuçlarında başarılı sonuçlar vermediği görülmüştür (Bhola, 2014). Oikonomou ve Tjortjis (2018), 8 Kasım 2016'da yapılan ABD başkanlık seçimlerine odaklanmış, seçimde özel olarak kazanma şansı en yüksek iki ana aday hakkında tweet toplamıştır. Veriler toplandıktan sonra, önerilen

yöntem bir sınıflandırma algoritması seçimi ve bunun uygulanmasından oluşmaktadır. Metin üzerinde sınıflandırma elde etmek için duyarlılık analizi de yapılmıştır. Önerilen yöntemle üç eyalet için yapılan çalışmada seçim sonuçlarının doğru şekilde tahmin edildiği gösterilmiştir.

Belirli bir siyaset olayı konusunda siyasetçilerin politik duruşunu tahmin etmek için yapılan çalışmalar da vardır. Johnson ve Goldwasser (2016), başkan adaylarının ve diğer önde gelen politikacıların mikroblog faaliyetlerini modellemek için politikacılar arasında belirli konudaki uzlaşma ve anlaşmazlık kalıplarının yanı sıra, geniş bir yelpazedeki meseleler üzerinde tahmin öngörüsü konusunda çalışmıştır.

Yapılan çalışmalardan görüldüğü gibi sosyal ağ, sosyal medya analizi ve siyasi amaçlar için metin madenciliği gelecekte hem siyasi hem de ekonomik eğilimleri tahmin etmenin kullanışlı ve doğru bir yöntemi haline gelebilecek elverişli bir yöntemdir.

Bu tez çalışmasında; Türkiye'de Twitter ortamında toplanan veriler ile İnternet üzerindeki veri kaynakları olan bloglar, sosyal ağlar veya herhangi bir mecradan elde edilen bilgiler üzerinde politik görüş belirleme amacı ile kullanıcı profilleme çalışmaları yapılmıştır.

3. KULLANILAN TEKNOLOJİLER VE PROGRAMLAMA DİLLERİ

Bu bölümde veri setinin elde edilmesinde ve sosyal profillemenin gerçeklenmesinde kullanılan teknoloji, platform ve programlama dilleri tezin işleyişinin anlaşılması açısından kısaca tanıtılmıştır.

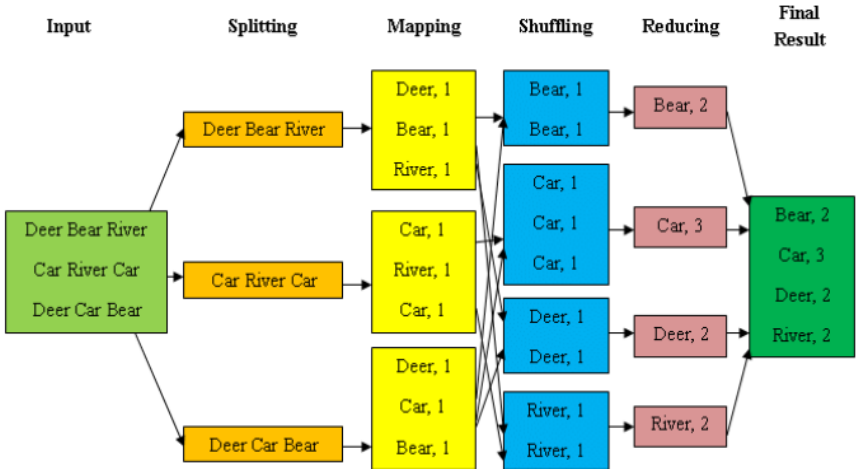
3.1 Kullanılan Teknolojiler

3.1.1 Apache Hadoop

Verinin yönetimi, işlenmesi ve depolanması üretilen verinin her geçen gün artması ile daha da zorlaşmakta ve ilişkisel veri tabanı yönetim sistemleri ile ölçeklenemeyecek boyutlara ulaşmaktadır. Doug Cutting ve Mike Cafarella adlı iki Yahoo çalışanı tarafından Nutch arama motoru projesi için tasarlanan Hadoop; açık kaynak bir çatı kuruluş olan Apache Yazılım Vakfı'nın kayıtlı ticari markası olarak geliştirilmeye devam etmektedir ve büyük veri kümeleri ile birden fazla makinede dağıtık olarak işlem yapılabilmesine olanak sağlayan Java tabanlı açık kaynak kodlu bir yazılım çatısıdır (Uzunkaya vd, 2015). Hadoop büyük verilerin birden fazla makinede saklanmasını ve yönetilmesini sağlar. Hadoop içerisinde büyük verileri sakladığımız bileşene Hadoop Dağıtık Dosya Sistemi (HDFS - Hadoop Distributed File System) adı verilir. Hadoop içerisinde büyük verilerin dağıtık ve paralel olarak işlenmesini sağlayan programlama modeline MapReduce denir (Ghazi ve Gangodkar, 2015).

Veri kümeleri HDFS üzerinden yüklendikten sonra Map ve Reduce fazları işletilir. Örnek olarak basit bir metin dosyasındaki kelime sayısını bulan MapReduce işlemi Şekil 3.1'deki adımlardan oluşur (Seethalakshmi, 2018). Şekil 3.1'deki ayırma (splitting) adımında veriler 64 MB'lık bloklara ayrılır. Bu değer değiştirilebilir. Eşleşme (mapping) adımında her bir kelime key(word) ve value(1) şeklinde bölümlere ayrılır. Shuffling adımında eşleşme (mapping) işleminden çıkan sonuçlar Reducer'a yönlendirilir. Şekil 3.1'deki örnekte amaç kelime-sayma

uygulaması oldugu için aynı kelime grubu aynı Reducer'a yönlendirilir. İndirgeme (reducing) adımında gelen sonuçlar üzerinden toplama işlemi yapılır ve sonuçlar istenilen kaynaklara (HDFS, SQL, NoSQL) yazılır (Seethalakshmi, 2018).



Şekil 3.1: Örnek bir Hadoop MapReduce kelime sayım süreci (Seethalakshmi, 2018).

Apache Hadoop duygu analizi ve verilerin dağıtık olarak işlenmesi için altyapı oluşturulmasında ve aynı zamanda kelime-sayımı benzeri süreçlerle veriler üzerinde birtakım çalışmalar yaparak doğruluk ve öznitelik çıkarımlarında bulunulmasında fayda sağlamaktadır.

3.1.2 Selenium

İlk olarak ThoughtWorks firmasında çalışan Jason Huggins tarafından Java Script Test Runner adıyla geliştirilen Selenyum; web sitesindeki tanımı ile bir web tarayıcı otomasyonudur (Selenium Home Page, 2019). Web tarayıcı otomasyonları, web sitelerinde ekranların ve bazı süreçlerin test edilmesinde kullanılan araçlardır. Selenium sayesinde, bir kişi adına kayıt formuna otomatik kayıt olunması, kişinin yerine birtakım butonlara tıklanması, bazı yerlere klavyeden veri gönderimi yapılabilmesi mümkündür. Selenyum araç seti, birlikte veya ayrı ayrı kullanılabilen ve farklı yazılım geliştiricileri tarafından Selenium yazılım paketine eklenen aşağıdaki bileşenleri içerir: Selenyum IDE, Selenyum Core, Selenyum 1 (Selenyum Remote Control), Selenyum 2 (Selenyum Web sürücüsü), Selenyum Grid (Srinivas ve Prakash, 2017).

Selenyum2, projenin gelecekteki yönünü belirleyen ve Selenyum araç setine en yeni eklenen bileşendir. Bu yepyeni otomasyon aracı, daha uyumlu, nesnelere yönelik bir API ve eski uygulamanın sınırlamalarına çözüm sunan özellikler içermektedir. Selenyum 2 testlerde maksimum esneklik için WebDriver API'yi, ilgili teknolojiyi ve WebDriver API altındaki Selenium 1 teknolojisini destekler. Ek olarak, Selenyum 2, geriye dönük uyumluluk için Selenium 1'in Selenium RC arayüzünü kullanmaktadır. Selenyum 1 çoğunlukla bakım modunda olmak üzere halen aktif olarak desteklenmektedir. Selenyum IDE (Tümleşik Geliştirme Ortamı), test komut dosyaları oluşturmak için bir prototipleme aracıdır. Bir Firefox eklentisidir ve otomatikleştirilmiş testler geliştirmek için kullanımı kolay bir arayüz sağlar. Selenyum IDE, kullanıcı eylemlerini gerçekleştirilirken kaydeden ve daha sonra çalıştırılabilecek birçok programlama dilinden birinde bunları yeniden kullanılabilir bir komut dosyası olarak dışa aktaran bir kayıt özelliğine sahiptir. Selenyum Grid ise Selenyum RC çözümünün büyük test grupları ve birden fazla ortamda çalıştırılması gereken test paketleri için ölçeklendirilmesine olanak sağlar (Razak ve Fahrurazi, 2011).

Bu tez çalışmasında kullanılacak olan Twitter verilerine ait niteliklerin tümü Twitter API (Application Programming Interface) aracılığı ile elde edilmemektedir. Bu yüzden Selenium yardımı ile istenilen verilerin çekilmesi sağlanmıştır. Uygulanan yöntemin ve analiz sonuçlarının açıklandığı Bölüm 4'te verinin elde edilişi aşamasında Selenium yardımı ile hangi veri niteliklerinin çekildiği ayrıntılı bir şekilde açıklanmıştır.

3.2 Programlama Dilleri ve Platformlar

Çalışmanın gerçekleştirilmesi sırasında iki farklı programlama dilinden yararlanılmıştır. Verilerin toplanması ve bazı ön işlemler için Java platformundan yararlanılmıştır. Analiz çalışmaları için Python ve Rapidminer tercih edilmiştir.

3.2.1 Python

Phyton, açık kaynak kodlu olan gelişmiş bir yazım şekli ile geniş bir içeriğe sahip olan, yorumlanan bir dildir. Yazılım paradigmalarından fonksiyonel programlama, yordamsal programlama, cephe yönelimli programlama ve nesneye yönelik programlama paradigmalarını destekler (Sarkar, 2019). Yazım şeklinin rahatlığı yazılımcı için konforlu bir alan sunmaktadır.

Python'da sınıflar, nesneler, veriler ve yöntemler ile soyutlama, kapsülleme, kalıtım ve polimorfizm gibi ilkeler de dahil olmak üzere birçok nesne yönelimli programlama kavramı bulunmaktadır (Sarkar, 2019). Python'da kod yazma yeteneklerini arttıran koleksiyonlar, itertools ve functools gibi modüller de dahil olmak üzere birçok gelişmiş özellik bulunmaktadır. Python standart kütüphaneleri, düşük seviyeli donanım arayüzlerinden dosya işlemeye ve metin verileriyle çalışmaya kadar çok çeşitli yetenek ve özellikler içermektedir (Sarkar, 2019). Yazılım geliştirilirken kolay genişletilebilirlik ve entegrasyon göz önünde bulundurulur, böylece mevcut uygulamalarla kolayca entegre edilebilir ve diğer uygulamalara ve araçlara arayüzler sağlamak için zengin uygulama programlama arayüzleri (API'ler) oluşturulabilir. Python ayrıca, İnternette faydalı kaynak ve belge üreten faydalı bir geliştirici topluluğuna sahiptir. Bu topluluk ayrıca dünya çapında çeşitli atölye çalışmaları ve konferanslar düzenlemektedir (Sarkar, 2019).

Python; Java, C++ ve C gibi diğer dillere kıyasla büyük boyutlu yazılım projelerinde geliştirme, çalıştırma, hata ayıklama, dağıtma ve bakım aşamalarında süreyi kısaltarak maliyeti düşürür ve üretkenliği artırır (Sarkar, 2019). Çok amaçlı bir dil olmasının yanı sıra, Python kullanılarak geliştirilen ve Python ile birlikte kullanılan veri tabanlarını kullanma kütüphaneleri, metin verileri, makine öğrenmesi, sinyal işleme, görüntü işleme, derin öğrenme, yapay zekâ çalışmaları için oluşturulan çeşitli çerçeveler, kütüphaneler ve platformlar ile bir ekosistem oluşturur (Sarkar, 2019). Python bu güçlü özellikleri, güçlü dil yapısı, çoklu-dil dökümantasyon desteği ve kolayca öğrenilebilmesi sayesinde analiz çalışmalarında kullanılan en önemli yazılım dili olmakta (Pine, 2019) ve birçok tez ve projede tercih edilmektedir.

3.2.2 RapidMiner

Java ile oluşturulmuş açık kaynak bir yazılım olan RapidMiner, makine öğrenmesi, veri madenciliği, öngörü analizi çalışmalarında sıkça kullanılmaktadır. RapidMiner ile veri madenciliği akışı XML'de tanımlanır ve bir grafik kullanıcı arayüzü (GUI) ile görüntülenir. RapidMiner, Weka ve R ile bütünleşik çalışır. Rapidminer'in işlevleri, çeşitli operatörler de dahil olmak üzere süreçlerin birbiri ile bağlantısı ile uygulanır. Tüm akış, orijinal veri girişi ve model sonuçları çıktısı ile sanki bir fabrikanın üretim hattı gibi kabul edilebilir. Operatörler, farklı giriş ve çıkış özelliklerine sahip bazı özel fonksiyonlar olarak kabul edilebilir (Chen vd., 2014).

Genel anlamda araştırma ve eğitim alanlarında kullanılan RapidMiner gelişmiş bir içeriğe sahiptir. Bu kapsamda toplanan Twitter verileri üzerinde makine öğrenmesi yöntemlerini uygulamak için RapidMiner platformu da kullanılmıştır.

3.2.3 Java

Java ilk olarak Sun firması tarafından 1995 yılında piyasaya sürülen bir programlama dilidir. Java programlama dili gelişmiş altyapısı ve gönüllü topluluğu sayesinde yıllar içinde hızla gelişmiş ve kullanıcı kitlesini arttırmıştır. Java günümüzde tüm dünyada masaüstü bilgisayarların %97'sinde, ABD'deki masaüstü bilgisayarların %89'unda bulunmaktadır. Dünya genelinde 9 milyon Java Geliştiricisi ile yazılım geliştiricilerinin ilk seçimlerinden ve en önemli yazılım geliştirme platformlarındandır. 3 Milyar Cep Telefonunda, 125 milyon TV cihazında Java bulunmaktadır. Blu-ray disk oynatıcıların tümünde, beş orijinal parça üreticisinin ürünlerinde Java ME kullanılmaktadır.

Java'nın yaygınlığı da göz önüne alınarak bu tez çalışmasında, Selenium IDE'si üzerindeki işlemler ve bir takım programlama işlemleri Java dili ile yapılmıştır.

4. TWITTER VERİLERİ İLE SİYASİ PROFİL ÇIKARIMI

Bölüm 2.2'de ayrıntılı olarak anlatılan Twitter mikro-blog ve sosyal ağ sitesi üzerinden verilerin elde edilmesi, ön işlemler ile verinin ayıklanması, başarım ölçümünde kullanılan ölçütler, kullanılan yöntemler ve elde edilen başarım sonuçları bu bölümde ayrıntılı olarak anlatılmıştır.

4.1 Verinin Elde Edilişi

Twitter, Facebook gibi büyük veri üreten ve depolayan kurumlar, veri çekim işlemleri için API (Application Programming Interface) denilen sistemler sağlamaktadır. Fakat bu API'lere verilen izinler kısıtlı olduğu için özel kişi izni veya kurum izinlerinin el verdiği ölçüde veri çekilebilmektedir. Tezde analiz çalışmalarımızda kullanılacak veri setindeki ilgili nitelikleri çekmek Twitter API çalışma yapısı ile bazı nitelikler için uzun süreler alacağı ve bazı niteliklerin çekilmesine de Twitter veri kullanım politikası gereği izin verilmediği için veri çekim işlemi, Twitter erişimi yapan bir bireyi taklit ederek giriş yapıp verileri tek tek kopyala yapıştır yapan bir bot yazılımı ile yapılmıştır. Buradaki amaç API kısıtlamalarından tamamen kurtulmak ve düzenli verileri hatta API'nın sunamadığı birçok veriyi kısa sürede tutmaktır. Bu amaçla Selenium Web Tarayıcı Otomasyonu adı verilen bir yazılım tasarlanmıştır. Selenium Web Tarayıcı Otomasyonu sayesinde web tarayıcıyı sanki birileri açıyormuş gibi Twitter sayfaları açılmakta, bireysel oluşturulmuş bir hesap üzerinden veri setindeki ilgili nitelikte verileri çekilmesi planlanan kişilerin bilgilerine ulaşılmıştır. Selenium Web Tarayıcı Otomasyonu ile Twitter'dan çekilen veriler SQL yardımı ile saklanmıştır. Bu veriler veri analiz programlarında görselleştirme ve karşılaştırılmalı analizler yapılabilmesi için anlamlı, veri analizine uygun, verinin anonimleştirilmesi sağlanarak uygun Excel tablosu biçimine dönüştürülmüştür.

Bu tez çalışmasında bireylerin siyasi eğilimlerinin belirlenmesi amaçlanmıştır. Bu amaçla siyasetle yakından ilgili kişilerin, siyasi liderlerin, yazarların, gazetecilerin sosyal medya kanallarında ve sosyal ağ sitelerinde en aktif oldukları dönem olan seçim dönemi öncesinde veri çekiminin gerçekleştirilmesine

karar verilmiştir. Bu amaçla Türkiye'de 24 Haziran 2018 tarihinde gerçekleştirilen Cumhurbaşkanlığı ve 27. Dönem Milletvekili Genel Seçimi öncesinde, en çok oy alan beş siyasi partinin verileri ile bu beş siyasi parti listesinde görev alan 3085 kişi ilgili analiz için kullanılarak verilerin çekim işlemi ve kayıt işlemi yapılmıştır.

Veri setinde ele alınan kişilerin siyasi parti organlarına ait Twitter hesaplarında takip ettiği 3085 birey için bazı özellikler dikkate alınmıştır. Buradaki amaç popüler kişilerin belli olması, ortak özellik çıkarımının sağlanması ve aynı zamanda kendi partisinde olmayan kişi veya kurumları takip edip etmediği bilgilerinden ortak bir çıkarım yapılması sağlanmıştır. Siyasi olarak aktif 3085 bireye ait aşağıdaki nitelikler üzerinde durulmuştur.

1. Kişilerin en son attığı iki tweet arasındaki gün sayısı;
2. Kişilerin web sitelerinin olup olmaması
3. Attığı toplam tweet sayısı
4. Takip ettiği kişi sayısı
5. Takip eden kişi sayısı
6. Beğendiği paylaşım sayısı
7. Açıklamasındaki karakter sayısı
8. Açıklamasında kullandığı # sayısı
9. Paylaştığı fotoğraf ve video sayısı
10. Günlük Ortalama tweet sayısı
11. Arkadaşlarının parti takip bilgisi
12. Kendisinin takip ettiği partiler
13. Tweet içerikleri

Yukarıda bahsi geçen 13 niteliğin tercih sebebi bir kişinin Twitter üzerinden elde edilebilecek bilgilerin tüm Twitter kullanıcıları için elde edilebilecek olmasıdır. Kişinin son attığı 2 tweet arasındaki gün sayısı kişinin aktifliğinin göstergesi olarak gözükmektedir. Örneğin bir kişi 100 gündür tweet atmadı ise bu kişinin ortalama günlük tweet sayısı ile kurulacak bir korelasyon ile aktif olarak Twitter kullanıp kullanılmadığı yorumu yapılması sağlanmıştır. Böylelikle Bölüm 4.2.2'de pasif içerikli bireylerin elenmesinde bu bilgiden yararlanarak eleme işlemleri yapılmıştır. Bir kişi ortalama ayda bir tweet atıyor ise bu kişinin Twitter kullanma sıklık karakterini belli etmektedir. Bunu da bizim ölçebilmemiz için

toplam ne kadar tweet attığını ve Twitter'a katılım sağladığı ilk tweet arasındaki gün farkına böldüğümüzde ortalama tweet sayısı çıkmıştır. Bazı siyasi parti üyeleri Twitter üzerinde genelde beğenme işlemini çok tweet atmayı az tercih edebilirken, kiminde ise çok tweet az beğenme karakteri gözükebilmektedir. Bu sebeple kişilerin beğendiği paylaşım sayıları, ortalama beğendiği paylaşım sayıları önem arz etmektedir. Bu tez çalışmasında ayrıca bir veri ilişki matrisinin elde edilmesi üzerinde yoğunlaşılmıştır. Çalışmanın önemli bir aşamasını içeren bu süreçte, siyasetle uğraşan uzman bir kişi yardımı ile Twitter'da seçilen kişiler tek tek önemle ayırt edilmiş, parti başkanları, yazar, siyasetle uğraşmış olan bu tür kişiler tespit edilmiştir. Buradaki amaç rastgele seçilen Twitter verilerinden öte aktif siyasette görev almış ve paylaşımları ile kendi savundukları değerleri anlatan kişilerin tercih edilmesidir.

Veri seti oluşturma çalışmamızın başarımı için önemli bir aşamasıdır. Özellikle 24 Haziran Cumhurbaşkanlığı Seçimi öncesindeki bir aylık dönemde çekilmiş olduğu için yoğun ve hatasız bir şekilde veri toplamanın yapılması gerekmiştir. Bu açıdan Selenium Web Otomasyon Aracı uygun bir seçim olduğunu göstermiş ve bu aşamayı başarı ile geçmemizi ve 3085 bireye ait tüm nitelikleri planlanan sürede çekebilmemizi sağlamıştır.

Veri setinde Türkiye'de son seçimlerde en çok oy alan 5 siyasi partinin verilerinden yararlanılmıştır. Bu tez çalışmasında, ele alınan siyasi partilerin isimleri P_1, P_2, …, P_5 şeklinde ifade edilmiştir. Analiz için veri çekim işlemi yapılan toplam 3085 Twitter kullanıcısı bireyin partilere göre dağılımı Tablo 4.1'de verilmiştir. Tablo 4.1'deki veriler, ön işlemden aşamasında pasif içerikli bireylerin elendiği durumdaki kişi dağılımıdır.

Tablo 4.1: Partilere göre kişi sayılarının dağılımı

P_1	P_2	P_3	P_4	P_5
122	1192	75	712	556

Veri setimiz 24 Haziran seçimlerinden önce oluşturulmuştur. Bu nedenle güncel seçim ile tweetler birebir veri analizinde gözlemlenmiştir. Örneğin Apache Hadoop üzerinde kurduğumuz büyük veri analizi aracı ile kelime-sayım (word-

count) sonucu 24 kelimesinin en çok kullanılan kelimeler arasında olup tüm partiler tarafından ortak olarak kullanıldığı gözlenmiştir.

4.2 Ön İşlemler

4.2.1 Bilgi Doğrulama

İlk olarak Selenium ile çekilen verilerde herhangi bir hata olup olmadığı, çapraz matris mantığı ile test edilmiştir. Çapraz matris ile kast edilen, bir kişinin kendisini takip edemeyeceği böylelikle bir matris oluşturulduğunda kendi kendine takip edilmemesi üzerinden ilk test işlemi gerçekleştirilmiştir.

Tablo 4.2: Çapraz matris kontrolü

Kişiler	**A**	**B**	**C**	**D**	**E**
A	**0**				
B		**0**			
C			**0**		
D				**0**	
E					**0**

Tablo 4.2'de Twitter üzerinden takip edilen kişi bilgisinde kendisini takip eden kişi olmaması gerekmektedir. Bu sebeple elde edilen verileri kişilerin birbirinin takiplerini gösteren bir matris haline getirerek veri kontrolünün ilk adımı sağlanmıştır.

Çapraz matris kontrolü sağlandıktan sonra veri setindeki bilgilerin tutarlılığı ve kontrolü için rastgele seçilen 50 kişi üzerinde veri kontrolleri sağlanmıştır. Bu veri kontrolleri doğru çıktıktan sonra aynı çekim işleminin 200 kişi için tekrarlanmasına ve bu kişilerin çekilen verileri arasında uyumsuzluk olup olmadığının kontrol edilmesine geçilmiştir. Farklılık çıkan işlemlerin neler olduğu hangi hesaplar ve hangi bilgilerde olduğu gözlemlenmiştir ve gerekli güncellemeler sağlanmıştır.

4.2.2 Pasif İçerikli Bireylerin Elenmesi

Tezin başarı kriterlerinde etkili bir rol alan aktif kişilerin ve seçimden 1 ay önce muhakkak belli bir seviyede tweet atmış kişiler olması önemli bir etkendir. Eğer bir kişi 24 Haziran seçimlerinden önce hiçbir tweet atmadı ise bu kişilerin bir düşünce savunduğu ve bunu yaymak için çaba harcamadıklarını düşündüğümüzden eleme işlemleri yapılmıştır. Bu eleme işlemi veri setimizde eğitim için önemli bir adım oluşturmaktadır. Çünkü eğitim verisinin kalitesi ve çeşitliliği eğitim sonuçlarının etkin bir şekilde kullanılmasına olanak sağlamaktadır. Pasif içerikli bireylerin elenmesiyle birlikte elimizdeki 3085 kişilik veri seti 2657 kişiye düşmüştür.

4.2.3 Kelime Analizleri

Twitter üzerinden belirlenen 5 siyasi parti için ilgili siyasi partide yöneticilik yapan veya yönetim kurulunda aktif görevi olan kişiler tespit edilmiştir. Bu kişilerin 24 Haziran seçimlerinden hemen 1 ay önceki bazı zaman aralıklarında tweet veri çekim işlemi gerçekleştirilmiştir. Bu veriler yaklaşık olarak 56693 adet tweet verisini içermektedir. Bu veriler eleme ve tüm siyasi partilerden ortak sayıda alarak oluşturulan bir tweet deposudur. Bu veriler sadece siyasi parti etiketi ile txt dosyalarına yazma işlemi yapıldı. Bu txt'ler üzerinden Bölüm 3.2.1'de anlatılan Apache Hadoop kuruldu ve Hadoop üzerinde word_count uygulaması ile her partide en çok kullanılan kelime çıkarımları yapıldı. Şekil 4.1'de görüleceği üzere ittifak yapan bazı partilerin kullanılan kelimelerle arasında güçlü bir bağ olduğu özellikle bazı partilerin bazı kelimeler üzerinde çok durduğu gözükmektedir. Partiler arası karşılaştırma tablosu yine Şekil 4.1'de verilmiştir. 31 adet kelime üzerinde her parti tarafından kaç kez söylendiği tespit edilmiştir. Şekil 4.1'de görüleceği üzere Başkanımız kelimesi p1 ve p3 partileri tarafından çok sık kullanılmıştır. Bu ortak kelime aynı zamanda partiler arasında ittifak olduğunu, "Başkanımız" diyerek başkanlık sistemi konusundaki ortak görüş ve sahiplenmeyi, bu olguya destek verildiğini göstermektedir.

Bu verinin doğruluğunu test edebilmek için bu partilerin desteklediği adaylar ve sistem incelenmiştir. Bu inceleme sonucunda çok güçlü bir ilişki olduğu

gözlenmiştir. Aynı şekilde diğer 3 parti "Başkanımız" kelimesini çok az kullanmış ve bu kavrama karşı bir duruş olduğunu göstermişlerdir. Partiler incelendiğinde elde ettiğimiz sonuçlar ile güçlü bir bağ olduğu gözlenmektedir.

Şekil 4.1'de çekilen tweet'lerde sık kullanılan kelimelerde 24 Haziran seçimi ile ilişkilendirilebilecek çok sayıda ifade bulunması çektiğimiz veri setlerinin ve bu veri setlerinde ele aldığımız kişilerin kalitesi ve doğruluğunu teyit etmek için kullandığımız diğer bir yöntemdir. Bu kelime sıklık analizi sistemi ile veri setimizin doğruluğu ve seçim ile ilişkisi test edilmiş ve aynı zamanda gelecekte yapılacak olan çalışmalarda bir takım duygu analizi çalışmaları için alt yapı çalışması hazırlanmıştır.

KELİMELER/PARTİLER	P1	P2	P3	P4	P5
çok	7351	8601	9369	11881	7130
bir	52512	46529	72590	50232	38356
her	8386	6080	12232	7203	7596
daha	11973	6843	7139	6490	5509
birlikte	17535	6162	9165	4092	6048
...	33110	45483	26623	47043	35180
Haziran	8156	3302	8128	5083	4293
Türkiye	17859	7501	11646	5088	2661
24	12481	6443	9386	2530	7760
Başkanımız	14560	4867	13211	2512	3079
bu	14966	19306	17692	24372	17306
tüm	11442	6217	6422	5636	5801
Cumhurbaşkanı	4532	7733	6430	4724	3285
kadar	5751	6007	8899	8424	4586
için	31928	25046	21527	21757	21327
Bu	8706	14416	15239	12146	10035
ziyaret	20065	3007	8509	4311	4260
da	11563	12779	19424	15258	14441
de	14840	15453	18269	20190	16470
gibi	5193	4775	6368	7039	3768
Bugün	6602	4157	5095	3734	3464
ile	37648	16877	14527	16269	12066
en	11166	7322	10523	7986	6617
ve	130658	74491	179425	64750	60033
devam	11280	3952	4419	2935	4864
Genel	11259	7471	19521	4972	5892
büyük	10651	5267	6114	4412	4001
olarak	7470	5632	10554	5280	4762
bugün	5287	2687	4386	2757	4503
olan	14049	6987	9808	6366	4737

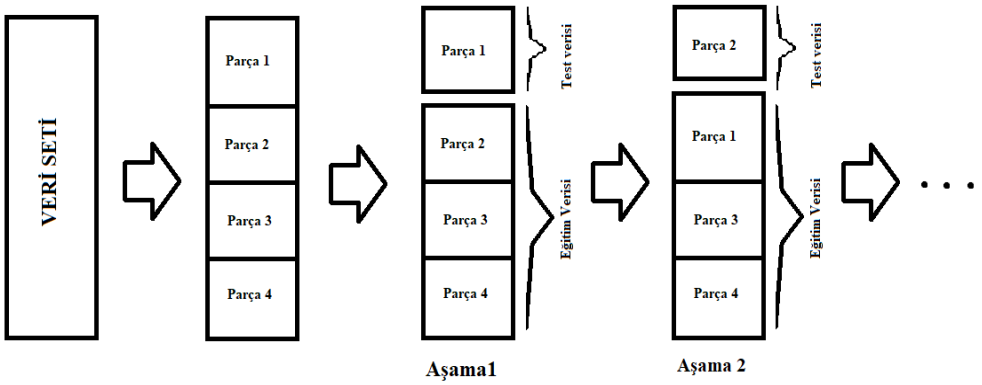
Şekil 4.1: Sık kullanılan kelimelerin partilere göre dağılımı

4.2.4 k-Katlamalı Çapraz-Doğrulama

Bölüm 4'te yapılacak analiz çalışmaları için geleneksel olarak kullanılan veri setinin eğitim ve test verisi şeklinde doğrudan ikiye ayrılması yerine k-katlamalı çapraz doğrulama (k-fold cross-validation) yöntemi tercih edilmiştir. k-katlamalı çarpraz doğrulama yöntemi doğrudan veri setini eğitim ve test kümesi olarak ayırmak yerine veri setindeki her verinin eğitim ve test verisi olarak belirli kereler kulanılmasını hedefler. Böylece eğitim ve test kümesinin seçimi ile oluşacak analiz sonuçlarının olumlu veya olumsuz yönde etkilenmesini ve dürüst olmayan analiz sonuçları elde edilmesini engellemiş olur.

k-katlamalı çapraz doğrulama için ilk olarak veri seti istenilen sayıda parçalara ayrılır (3, 4 , 5 ,10 vb). Bu parametre *k* ile gösterilir. Ardından ayrılan bu parçalardan 1. aşama olarak bir tanesi test verisi diğer kalanları eğitim verisi olarak kullanılır. 2. aşamada kullanılmayan bir parça test verisi diğer kalan veriler eğitim verisi olarak seçilir. Bu işlem kullanılmayan test verisi kalmayana kadar devam eder.

Sonuç olarak her veri hem test verisi hem de eğitim verisi olarak kullanılmaktadır. Bunun sonucunda verilerin test ve eğitim kümesine ayrılması sırasında oluşacak veri bağımlılığı ortadan kaldırılmış olacaktır.



Şekil 4.2: k-fold çapraz doğrulama görsel anlatımı (Web-Sadi-Seker)

Şekil 4.2'de olduğu gibi parça sayısı kadar aşama olacaktır. İlk parça test verisi olarak seçildiğinde kalan parçalar eğitim verisi olarak seçilmiştir. Aşama 2 de test verisi olarak parça 2 seçilmiştir, bu parçadaki veriler bu aşamada eğitime

katılmayacak test verisi olarak değerlendirilecektir. Tüm aşamaların tamamlanmasından sonra her aşamanın doğruluk verisinin (accuracy) sonucunun ortalaması alınır. ∓ yanılma payıyla hesaplanır.

4.3 Başarımın Ölçülmesi

Veri madenciliği, sosyal ağ analizi, finansal analiz, pazarlama araştırmaları vb. alanlarda yapılan çalışmalar sonucunda bazı tahminlerde bulunulmaktadır. Bu tahminlerin başarımının ve doğruluğunun ölçülebilmesi için bazı başarım ölçütleri kullanılmaktadır. Bunlar arasında en çok bilinenleri doğruluk, hata oranı, kesinlik, duyarlılık, fl ölçütü, hata karelerinin ortalaması, mutlak hata ve özgünlük ölçütleridir. Bu çalışmada, veri madenciliği uygulamalarında en çok kullanılan ölçütler olan doğruluk, kesinlik, duyarlılık ve fl ölçütü kullanılmıştır. Ayrıca bu tür uygulamalarda sınıflandırma sonuçlarında gerçek değer ile tahmin edilen değerin karşılaştırılması için karışıklık matrisinden yararlanılır. Tablo 4.3'te S_1, S_2 ve S_3 olmak üzere 3 sınıf için bir karışıklık matrisi örneği bulunmaktadır. Bu tabloda X_{ij} gerçekte ait olduğu sınıf i.sınıfken, tahminleme sonucu elde edilen sınıf etiketinin j.sınıf olduğu örnek sayısıdır.

4.3.1 Karışıklık Matrisi

Karışıklık matrisi, tahminde bulunulmuş n adet sınıfın gerçekte hangi sınıfta olduğunu ve her sınıf için gerçekte olması gereken değer ile tahmin edilen değerin bu tabloya bakılarak incelenmesini sağlayan bir yapıdır. (Özkan 2016). Tablo 4.3'te n sınıf için verilen karışıklık matrisi yapısında satırlar, ilgili örneğin gerçekte hangi sınıflara ait olduğunu sütunlar ise tahmin sonucunun ne olduğunu göstermektedir. Bu durumda oluşan bu tablodaki köşegen değerleri başarılı sınıf etiketlemelerine karşı düşer (Şahin, 2018).

Tablo 4.3: *n* sınıf için karışıklık matrisi (Şahin, 2018)

	Tahmin edilen Sınıf					
Gerçek Sınıf		$\boldsymbol{S_1}$	$\boldsymbol{S_2}$	$\boldsymbol{S_3}$	…	$\boldsymbol{S_n}$
	$\boldsymbol{S_1}$	X_{11}	X_{12}	X_{13}	…	X_{1n}
	$\boldsymbol{S_2}$	X_{21}	X_{22}	X_{23}	…	X_{2n}
	$\boldsymbol{S_3}$	X_{31}	X_{32}	X_{33}	…	X_{3n}
	⋮	⋮	⋮	⋮	⋮	⋮
	$\boldsymbol{S_n}$	X_{n1}	X_{n2}	X_{n3}	…	X_{nn}

4.3.2 Doğruluk

Doğruluk (Accuracy), tahmin edilen örneğin gerçek sınıf değerlerinin arasındaki oranın ne olduğu ile ilgilidir. (Özkan 2016). Tablo 4.3'te verilen karışıklık matrisi dikkate alındığında doğruluk ölçütü şu şekilde hesaplanabilir:

$$doğruluk = \frac{\sum_{i=1}^{n} X_{ii}}{\sum_{i=1}^{n} \sum_{j=1}^{n} X_{ij}} \quad (1)$$

Özet olarak buradaki doğruluk yukarıdaki karşıklık marisinde gösterdiğimiz köşegenlerindeki sayının toplamının örneklem kümemizin tamamına bölünmesiyle bulunur. Optimum değeri 1, minumum değeri yani kötü olduğunda alacağı değer 0'dır. Hata oranı ise basit mantık kuramı ile 1'den çıkarılması ile bulunur (Şahin, 2018).

Yani kısaca doğruluk ölçütü karışıklık matrisinin köşegenlerinde yer alan örnek sayısının toplam örnek sayısına bölünmesiyle bulunmaktadır. En iyi durumda alacağı değer 1, en kötü durumda alacağı değer 0'dır.

4.3.3 Kesinlik

Kesinlik (Precision), tahminde bulunduğumuz verilerden doğru bildiklerimizin doğru ve yanlış bildiklerimize oranı diyebiliriz. Karışıklık matrisi üzerindeki ele alınan bir sınıfın doğru bilindiği satırdaki değerinin bir sütundaki

değerlerin toplamına bölünmesi ile ifade edebiliriz. Literatürdeki farklı bir aktarımla, bir sınıfla etiketlenmiş örnek verilerin gerçekten o verilerin o sınıfa ait olma ihtimalidir (Precision_Score, 2017).

$$kesinlik_{S_k} = \frac{X_{kk}}{\sum_{i=1}^{n} X_{ki}} \tag{2}$$

Benzer şekilde toplam ortalama kesinlik ölçütü şu şekilde hesaplanabilir:

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} X_{ij} * kesinlik_{S_i}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}} \tag{3}$$

Yukarıdaki anlatımdan anlaşılacağı üzere kesinlik ölçütü her bir grup için ayrı olarak hesaplanabilmekle birlikte tüm sınıfların kesinlik ölçütünün ağırlıklı ortalaması alınarak ortalama kesinlik ölçütü de hesaplanabilmektedir. Optimum durumdaki değeri 1, en kötü durumda alacağı değer 0'dır (Şahin, 2018).

4.3.4 Duyarlılık

Duyarlılık sınıflandırmanın tüm doğru sınıflandırmaları bulma becerisi diyebiliriz.

$$duyarlılık_{S_k} = \frac{X_{kk}}{\sum_{i=1}^{n} X_{ki}} \tag{4}$$

Benzer şekilde toplam ortalama kesinlik ölçütü şu şekilde hesaplanabilir:

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} X_{ij} * duyarlılık_{S_i}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}} \tag{5}$$

Duyarlılık ölçütü de aynı kesinlik gibi her sınıf için ayrı ayrı hesaplanabileceği gibi tüm sınıfların duyarlılık ölçütünün ağırlıklı ortalaması alınarak ortalama duyarlılık ölçütü de hesaplanabilmektedir. En kötü durumda 0, en iyi durumda 1 değerini almaktadır (Şahin, 2018).

4.3.5 Ölçütlerin Önemi

Bir veri tahminleme işleminde sadece doğruluk oranına bakılarak yorumlama yapılırsa o yorumlama her zaman eksik bir yorumlama olacaktır. Diğer bir deyişle algoritmaların başarı kriteri sadece doğruluk (accuracy) olursa dengesiz veri setlerinde (imbalanced data sets) bize pek bilgi vermez. Dengesiz veri setleri sınıflar arasındaki dağılımın yakın olmadığı veri setlerini tanımlarken kullanılır. Bunu biraz detaylı anlatabilmek için bir örnek üzerinden gidelim. Bazı niteliklerden oluşmuş verilere bakarak bir insana sağlıklı ve hasta diyebilmek için bir tahminde bulunduğumuzu varsayalım. 100.000 Kişi içerisinden sadece 50 kişinin hasta olduğunu varsayalım. Herkese sağlıklı olarak tahminde bulunursak başarı oranımız bölüm 4.2.1'de hesaplama şekline göre

$$Doğruluk = \left(\frac{99.950}{100.000}\right) * 100 = \%\ 99.95 \tag{6}$$

Görüldüğü gibi herhangi bir yapay zeka algoritması kullanmadan sadece herkese sağlıklı diyerek yüksek bir başarı kaydettiğimizi söyleyebiliriz. Fakat bu başarı oranının bizim hiçbir işimize yaramayacağı aşikardır. Bu sebeple bölüm 4.2.2 ve 4.2.3'deki kesinlik ve duyarlılık ölçütlerine de ihtiyaç duyarız. Bu örnek üzerinde gidilerek bazı durumlarda bazı değerlerin daha çok önem arz ettiği görülebilir.

Örneğin anormal durumları doğru tespit edebilmek için yanlış alarmlar üretmekten daha önemli bir hale getirebilir. Yani hasta birisini tespit edemeyip onun ölümüne neden olmaktansa olmayan bir kişiye hasta olduğu söylenip onu çağırarak gerekli kontrollerin yapılmasını sağlamak daha çok kabul edilebilir bir durumdur. Fakat bu durumun aynı zamanda bir eksisi de vardır. Böylelikle herkesi hastaneye çağırıp hasta demiş oluruz hasta olanları yüzde 100 tespit edip fakat birçok kişiyede yanlış yönlendirme bilgilendirme yapmış oluruz. Birine hasta demeden önce muhakkak çok iyi düşünüp taşınmamız gerekiyor. Herkese hasta dersek kesinlik üzerindeki eksisi ise bir kişiye hasta dedik ve o gerçekten hasta ama geriye kalan kişilere hasta diyip tespit edemedik.

Bu anlatılan örnekte açık bir şekilde görülmektedir ki, Kesinlik ve Duyarlılık arasında önemli bir ilişki vardır. Buna iktisattaki adı ile "ödünleşme(trade-off)" diyebiliriz. Kesinliğin veya Duyarlılığın tercihinin objektif doğrusu yoktur. İkisini de aynı anda istemek, ikisinden de mahrum bırakabilir.

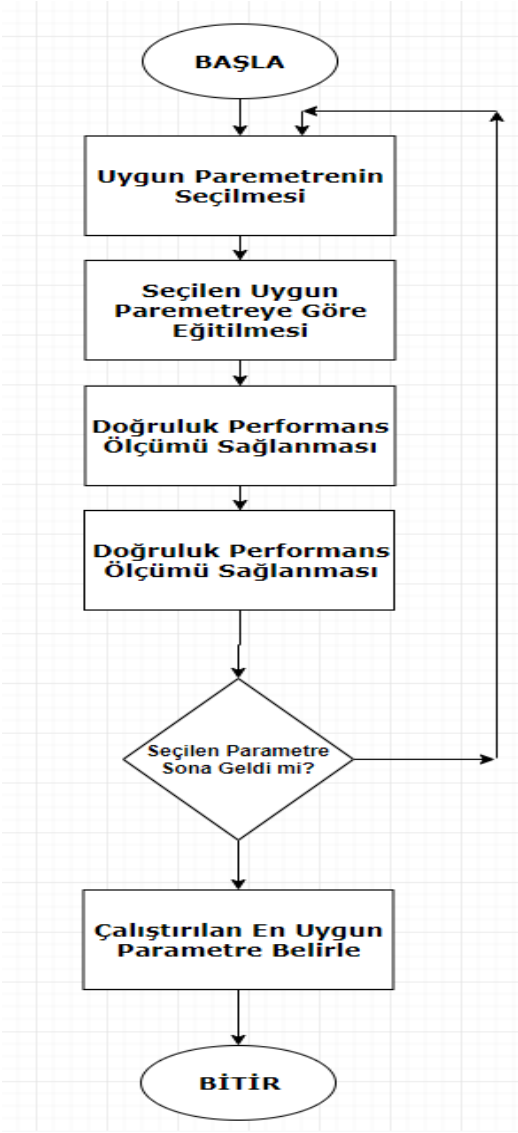
Bu sorunu çözmek ve farklı bir değer ile tahminleme kontrolü sağlamak adına literatürdeki adı F Skoru olan değere bakılmaktadır.

F Skoru aslında bu 2 değişkenin harmonik ortalamasıdır.

$$F1\ Skor = 2 * \left(\frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık}\right) \quad (7)$$

4.3.6 Parametrelerin Optimizasyonu

Şekil 4.3'te görüldüğü gibi algoritmamız örneğin bir k değeri ele alacak olursak k değeri 1'den N'e kadar tüm değerler için denenip en çok başarılı bulunan k değeri için için doğruluk oranları paylaşılacaktır. Bu tezdeki parametrelerin değerlerine göre en optimum sonucun olacağı gösterilmektedir.



Şekil 4.3: Parametre optimizasyonu işleminin algoritmik gösterimi

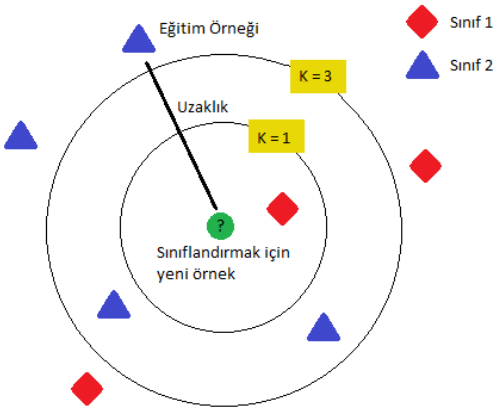
4.4 Yöntemler ve Sonuçları

Bu alt bölümde, tez çalışmasında kullanılan yöntemler ele alınarak bu yöntemlerin Bölüm 1 ve 2 de araştırma yapılan ve referans verilen çalışmaların bazılarının benzerlerinin Türkiye'deki seçim sonuçları için uygulanmasına bazı yöntemlerin ise değiştirilerek tekrardan farklı bir şekilde uygulamasının sağlanması hedeflenmiştir. Doğrulama işlemleri için K Katmanlı Çarpraz Doğrulama (K-fold Cross Validation) kullanılmıştır. Kullanılan bazı yöntemler ise şu şekildedir; k-NN, Naive Bayes, Rassal orman, Deep Learning.

Kullanılan tüm yöntemlerde veri setinin eğitim ve test olarak ayrılma işlemindeki k-fold yönteminin *k* değerinin 10 olarak seçilmesi uygun görülmüştür.

4.4.1 k-NN

Yaygın bir sınıflandırma tekniği olarak kullanılan k-NN algoritması Denetimli öğrenme sınıfındadır. k-NN algoritması tembel bir öğrenme algoritmasıdır, aynı zamanda parametrik olmayan ve örnek tabanlı bir algoritmadır. k-NN algoritmasının amacı bir çıktıyı tahmin etmek için farklı veri noktaları içeren bir ortam kullanmaktır. Parametrik olmayan araçlar, veriler hakkında varsayımda bulunmaz. Genelde teorik varsayımlara uygulamaz. Örnek tabanlı, algoritma veri setini öğrenmek yerine ezberlemeyi çalışıyor ve bunu tahmin aşamasında bilgi olarak kullanıyor. (Şeker, 2010) k-NN algoritması düşük hesaplama zamanı ve kolay yorumlanabildiği için diğer algoritmalara göre daha çok tercih edilir.



Şekil 4.4: k-NN algoritması görsel anlatımı

Şekil 4.4'teki yeni örnek olarak gelen yuvarlak noktanın çıktı sınıfını tahmin etmek için ilk başta *k*=1 aldığımızdan dolayı kendisine en yakın *k* komşusu baklava dilimi olduğundan baklava dilimi nesnenin bulunduğu Sınıf-1 olarak sınıflandırmıştır.

Ancak *k*=3 alındığında yuvarlak noktanın en yakın 3 noktasını seçmemiz gerekmektedir. Şekil 4.4'teki örneğe göre öklid mesafesi (euclidean distance) hesaplandıktan sonra 2 üçgen ve 1 baklava dilimi noktaya yakındır. Buradan da

yeni gelen yuvarlak noktanın üçgen nesnenin sınıfı olan Sınıf-2'ye ait olacağı sonucu çıkar.

k-NN ile ilgili tahminleme algoritmamızı veri setimizde k-fold ile veri setini test eğitim olarak devamlı ayrılması ile ilgili karışıklık matrislerini Tablo 4.4'teki gibi görebilirsiniz.

k-NN algoritmamız öncelikle veriler üzerinde herhangi bir değişiklik yapmadan veri setimiz üzerinde çalıştırma yapılmıştır. Daha sonra ise verilerimiz Normalizasyon işleminden geçirildikten sonra tekrar eğitim yapılmış aynı *k* değerler ve k-fold *k* değeri için karşılaştırmaları yapılmıştır.

k-NN algoritmasındaki *k* parametresinin optimizasyonunun yapılması için şekil 4.3'te belirtilen parametre optimizasyon işlemi yapılmıştır. En başarılı *k* değeri elde edilinceye kadar çalışma devam ettirilmiştir. Tüm tabloları paylaşmak yerine *k*=1,5,13,19 sonuçları paylaşılmıştır.

Tablo 4.4'te k-NN algoritmasındaki *k* komşu sayısı 1 seçilmiş olup k-fold veri setimiz *k*=10 seçilmiştir. Elde edilen değerlere göre doğruluk oranımız Tablo 4.5'te görüldüğü üzere ortalama %41.89 başarı elde edildiği gözlenmiştir. Fakat bu başarı oranın çok düşük olduğu çok net gözlenmektedir. Veri setimizdeki oluşturduğumuz değerlerin doğrudan kullanımda k-NN algoritması için uygun olmadığı görülmektedir. Bu durumu araştırmak ve k-NN algoritmasındaki komşu sayımızı 1'den 19'a kadar tek komşuluklarını deneyerek devam edilmiştir ve sonuçlar aşağıdaki tablolarda paylaşılmıştır.

Tablo 4.4: k-NN *k*=1 ve k-fold *k*=10 için karışıklık matrisi

k-fold=> *k*=10 için | k-NN => *k*=1 için						
	DOĞRU P1	**DOĞRU P2**	**DOĞRU P5**	**DOĞRU P4**	**DOĞRU P3**	**SINIF KESİNLİK**
TAH. P1	15	66	29	14	1	12.00%
TAH. P2	64	625	263	234	37	51.10%
TAH. P5	27	244	139	119	18	25.41%
TAH. P4	15	226	110	331	16	47.42%
TAH. P3	1	31	15	14	3	4.69%
SINIF HASSSASİYET	12.30%	52.43%	25.00%	46.49%	4.00%	

Tablo 4.5: k-NN *k*=1 ve k-fold *k*=10 için doğruluk, kesinlik ve hassasiyet değerleri

	DOĞRULUK	KESİNLİK	HASSASİYET
DEĞERLER	41.89% (+/-) 2.71%	28.36% (+/-) 2.73%	28.06% (+/-) 2.11%

Tablo 4.6'da k-NN algoritmasındaki *k* komşu sayısı 5 seçilmiş olup k-fold veri setimiz *k*=10 seçilmiştir. Elde edilen değerlere göre doğruluk oranımız Tablo 4.7'de görüldüğü üzere ortalama %44.03 başarı elde edildiği gözlenmiştir.

Tablo 4.6: k-NN *k*=5 ve k-fold *k*=10 için karışıklık matrisi

k-fold=> *k*=10 için | k-NN => *k*=5 için						
	DOĞRU P_1	**DOĞRU P_2**	**DOĞRU P_5**	**DOĞRU P_4**	**DOĞRU P_3**	**SINIF KESİNLİK**
TAH. P_1	19	48	26	7	2	18.63%
TAH. P_2	72	666	288	218	45	51.67%
TAH. P_5	20	245	130	126	17	24.16%
TAH. P_4	10	217	103	355	11	51.01%
TAH. P_3	1	16	9	6	0	0.00%
SINIF HASSSASİYET	15.57%	55.87%	23.38%	49.86%	0.00%	

Tablo 4.7: k-NN *k*=5 ve k-fold *k*=10 için doğruluk,kesinlik ve hassasiyet değerleri

	DOĞRULUK	KESİNLİK	HASSASİYET
DEĞERLER	44.03% (+/-) 2.%	29.22% (+/-) 2.55%	28.98% (+/-) 2.67%

Tablo 4.8'de k-NN algoritmasındaki *k* komşu sayısı 13 seçilmiş olup k-katlamalı çapraz doğrulama için *k*=10 seçilmiştir. Elde edilen değerlere göre doğruluk oranı Tablo 4.9'da ortalama %49.60 başarı elde edildiği gözlenmiştir.

Tablo 4.8: k-NN *k*=13 ve k-fold *k*=10 için karışıklık matrisi

k-fold=> *k*=10 için | k-NN => *k*=13 için						
	DOĞRU P1	**DOĞRU P2**	**DOĞRU P5**	**DOĞRU P4**	**DOĞRU P3**	**SINIF KESİNLİK**
TAH. P1	4	22	8	2	3	10.26%
TAH. P2	102	852	356	262	56	52.33%
TAH. P5	12	135	94	80	8	28.57%
TAH. P4	4	183	98	368	8	55.67%
TAH. P3	0	0	0	0	0	0.00%
SINIF HASSSASİYET	3.28%	71.78%	16.91%	51.69%	0.00%	

Tablo 4.9: k-NN *k*=13 ve k-fold *k*=10 için doğruluk,kesinlik ve hassasiyet değerleri

	DOĞRULUK	KESİNLİK	HASSASİYET
DEĞERLER	%49.60 (+/-) 2.06%	% 29.67 (+/-) 4.49%	28.67% (+/-) 1.92%

Tablo 4.10'da k-NN algoritmasındaki *k* komşu sayısı 19 seçilmiş, k-katlamalı çapraz doğrulama için *k*=10 seçilmiştir. Elde edilen değerlere göre doğruluk oranı Tablo 4.11'de ortalama %50.84 başarı elde edildiği gözlenmiştir.

Tablo 4.10: k-NN *k*=19 ve k-fold *k*=10 için karışıklık matrisi

k-fold=> *k*=10 için | k-NN => *k*=19 için						
	DOĞRU P1	**DOĞRU P2**	**DOĞRU P5**	**DOĞRU P4**	**DOĞRU P3**	**SINIF KESİNLİK**
TAH. P1	0	5	1	0	0	0.00%
TAH. P2	111	890	372	268	57	52.41%
TAH. P5	6	111	79	62	8	29.70%
TAH. P4	5	186	104	382	10	55.60%
TAH. P3	0	0	0	0	0	0.00%
SINIF HASSSASİYET	0.00%	74.66%	14.21%	53.65%	0.00%	

Tablo 4.11: k-NN *k*=19 ve k-fold *k*=10 için doğruluk, kesinlik ve hassasiyet değerleri

	DOĞRULUK	KESİNLİK	HASSASİYET
DEĞERLER	%50.84 (+/-) 2.06%	% 27.57 (+/-) 2.35%	28.50% (+/-) 1.54%

Yukarıdaki k-NN karışıklık matrisi tabloları K-NN Algoritmasının farklı komşu sayıları için uygulanmıştır. *k* parametresi 1,3,5,7,9,11…19'a kadar çalıştırılmıştır. Her komşu sayısı farklı farklı matrisler çıkmaktadır, ancak ortaya çıkan bu matrislerin sonuçları birbirine yakındır.

k-NN algoritmasında genel bir kanı olarak *k* komşu sayısı ne kadar büyük olursa o oranda doğru sonuç elde edilecek düşüncesi mevcuttur. Fakat bu durum veri setlerine göre değişim göstermektedir. Farklı veri setlerine göre farklı komşuluk sayısı vardır. Genel doğruluk (accuracy) yüzdesi bu veri setine göre farklı *k* değerine göre farklı yüzde değerleri almıştır.

Bu değerler arasından en büyük doğruluk oranına sahip *k* değerini seçmemizin nedeni, yapılan tahminin doğru olma olasılığını arttırmak ve elimizdeki veri setinin *k* parametrelerine karşı nasıl bir sonuç verdiğinin izlenmesidir. Bu sonuçlara göre *k* değeri arttıkça başarı oranı artmıştır. Tablo 4.13'te k-NN algoritmasındaki *k* komşu sayısı 25 seçişmiş olup k-fold veri setimiz *k*=10 seçilmiştir.

Tablo 4.12: k-NN *k*=25 ve k-fold *k*=10 için karışıklık matrisi

k-fold=> *k*=10 için | k-NN => *k*=25 için						
	DOĞRU P1	**DOĞRU P2**	**DOĞRU P5**	**DOĞRU P4**	**DOĞRU P3**	**SINIF KESİNLİK**
TAH. P1	0	1	1	0	1	0.00%
TAH. P2	110	933	391	262	58	53.19%
TAH. P5	8	85	61	62	5	27.60%
TAH. P4	4	173	103	388	11	57.14%
TAH. P3	0	0	0	0	0	0.00%
SINIF HASSSASİYET	0.00%	78.27%	10.97%	54.59%	0.00%	

Tablo 4.13'te k-NN algoritmasındaki *k* komşu sayısı 35 seçilmiş olup k-fold veri setimiz k=10 seçilmiştir. Görüldüğü üzere *k* değerinin belli bir noktaya kadar artması doğruluğun artmasını sağlamıştır, fakat bir süre sonra *k* değerinin büyüklüğünün sonuç üzerinde değiştirici bir etkisi olmadığı gözlemlenmiş, hatta ağırlıklı kesinlik (precision) değerleri düşmekte sadece 3 parti için tahminlemede bulunulmaktadır.

Tablo 4.13: k-NN *k*=35 ve k-fold *k*=10 için karışıklık matrisi

k-fold=> *k*=10 için | k-NN => *k*=35 için						
	DOĞRU P1	**DOĞRU P2**	**DOĞRU P5**	**DOĞRU P4**	**DOĞRU P3**	**SINIF KESİNLİK**
TAH. P1	0	0	0	0	0	0.00%
TAH. P2	115	958	408	277	60	52.70%
TAH. P5	4	62	50	56	5	28.25%
TAH. P4	3	172	98	379	10	57.25%
TAH. P3	0	0	0	0	0	0.00%
SINIF HASSSASİYET	0.00%	80.37%	8.99%	53.23%	0.00%	

k-NN ile ilgili çıkaracağımız sonuç, veri setimizin çoklu girdi yapısı olması, arkadaşlık matrisinin olması gibi durumlar k-NN ile siyasi parti tahminlemesi yapılması veri setimiz için uygun olmadığı gözlemlenmiştir.

k-NN ile ilgili diğer elde edebileceğimiz sonuç ise optimum parametrenin *k* değerinin arttıkça veri setimize göre arttığında başarı oranın arttığı gözlenmiştir. k değeri bir süre sonra ne kadar atarsa artsın sonuçlar üzerinde çok büyük etken gözlenmemiştir, *k* için 19 değerinden sonra sonucu değiştirecek herhangi bir doğruluk oranı olmadığı gözlenmiştir. Ayrıca *k* değerinin artması ile p1 partisinin kesinliğinin yüzde 0 'a düştüğü gözlenmiştir.

k-NN algoritması yukarıda verilen veri setinin normalizasyon işlemi sonrasında ise uygulanan z-transformation ile %20 'lik bir doğruluğun artmasına sebep olmuştur. Bu da veri setlerinin türleri ve yapısına göre algoritmaların nasıl bir başarı gösterceğinin bir kanıtıdır. Bu normalizasyon işleminden sonra ise *k* değerinin arttıkça başarı oranı artmamış aksine düşüş göstermiştir.

Tablo 4.14'te k-NN algoritmasındaki *k* komşu sayısı 3 seçilmiş olup k-fold veri setimiz k=10 seçilmiştir. Elde edilen değerlere göre doğruluk oranımız Tablo 4.15'te görüldüğü üzere ortalama %79.31 başarı elde edildiği gözlenmiştir.

Tablo 4.14: Normalizasyon sonrası k-NN *k*=3 ve k-fold *k*=10 için karışıklık matrisi

k-fold=> *k*=10 için | k-NN => *k*=3 için						
	DOĞRU P1	**DOĞRU P2**	**DOĞRU P5**	**DOĞRU P4**	**DOĞRU P3**	**SINIF KESİNLİK**
TAH. P1	103	23	9	16	1	67.76%
TAH. P2	12	1056	143	121	6	78.92%
TAH. P5	6	65	391	45	0	77.12%
TAH. P4	1	41	6	484	1	90.81%
TAH. P3	0	5	1	46	67	56.30%
SINIF HASSSASİYET	84.43%	88.74%	71.09%	67.98%	89.33%	

Tablo 4.15: Normalizasyon sonrası k-NN *k*=3 ve k-fold *k*=10 için doğruluk,kesinlik ve hassasiyet değerleri

	DOĞRULUK	KESİNLİK	HASSASİYET
DEĞERLER	%79.31 (+/-) 2.13%	% 74.76 (+/-) 3.55%	80.36% (+/-) 3.56%

Tablo 4.16'da k-NN algoritmasındaki *k* komşu sayısı 9 seçilmiş olup k-fold veri setimiz *k*=10 seçilmiştir. Elde edilen değerlere göre doğruluk oranımız Tablo 4.17'de görüldüğü gibi ortalama %77.05 başarı elde edildiği gözlenmiştir.

Tablo 4.16: Normalizasyon sonrası k-NN *k*=9 ve k-fold *k*=10 için karışıklık matrisi

k-fold=> *k*=10 için | k-NN => *k*=9 için						
	DOĞRU P1	**DOĞRU P2**	**DOĞRU P5**	**DOĞRU P4**	**DOĞRU P3**	**SINIF KESİNLİK**
TAH. P1	93	11	7	19	1	70.99%
TAH. P2	28	1130	220	209	8	70.85%
TAH. P5	1	29	322	21	0	86.33%
TAH. P4	0	13	1	430	0	96.85%
TAH. P3	0	7	0	33	66	66.26%
SINIF HASSSASİYET	76.23%	94.96%	58.55%	60.39%	88.00%	

Tablo 4.17: Normalizasyon sonrası k-NN *k*=9 ve k-fold *k*=10 için doğruluk,kesinlik ve hassasiyet değerleri

	DOĞRULUK	KESİNLİK	HASSASİYET
DEĞERLER	%77.05 (+/-) 2.18%	% 78.43 (+/-) 3.55%	75.66% (+/-) 2.75%

Tablo 4.18'de k-NN algoritmasındaki *k* komşu sayısı 19 seçilmiş olup k-fold veri setimiz k=10 seçilmiştir. Elde edilen değerlere göre doğruluk oranımız Tablo 4.19'da ortalama %71.12 başarı elde edildiği gözlenmiştir.

Tablo 4.18: Normalizasyon sonrası k-NN *k*=19 ve k-fold *k*=10 için karışıklık matrisi

k-fold=> *k* = 10 için | k-NN => k=19 için						
	DOĞRU P1	**DOĞRU P2**	**DOĞRU P5**	**DOĞRU P4**	**DOĞRU P3**	**SINIF KESİNLİK**
TAH. P1	82	11	5	14	1	72.57%
TAH. P2	40	1158	305	333	14	62.59%
TAH. P5	0	9	240	4	0	94.86%
TAH. P4	0	7	0	344	0	98.01%
TAH. P3	0	5	0	17	60	73.17%
SINIF HASSSASİYET	67.21%	97.31%	43.64%	48.31%	80.00%	

Tablo 4.19: Normalizasyon sonrası k-NN *k*=19 ve k-fold *k*=10 için doğruluk, kesinlik ve hassasiyet değerleri

	DOĞRULUK	KESİNLİK	HASSASİYET
DEĞERLER	%71.12 (+/-) 3.23%	% 81.05 (+/-) 3.85%	67.32% (+/-) 4.26%

Yukarıdaki sonuçlardan anlaşılacağı üzere k-NN algoritması veri setine bağımlılığı ve aynı veri setindeki birimlerin türleri ve değerlerinden dolayı normalizasyon gibi bir ön işlemden geçirildiğinde başarı oranının %19 arttığı gözlenmiştir. Ayrca bu artışla birlikte *k* değerinin ise 3, 9,19 değerleri paylaşılmış fakat optimum *k* parametresi için *k* değeri en küçüğü en güçlü doğruluk oranı elde edilmiştir.

4.4.2 Naive Bayes

İsmini matematikçi Thomas Bayes'ten alan Naive Bayes algoritması olasılıkları ve olasılık dağılımlarının girdilerini belirleme yöntemidir. Bayes sınıflandırmaları yakın zamanda popülerlik kazanmış ve bu sınıflama işleminin gayet başarılı bir biçimde iyi performans sergilediği görülmüştür. Bayes teoremi önceden bilgimiz olan bir hipotezin olasılığını bulmamıza yardımcı olur. Olasılık teorisi ve istatistiklerinde Bayes teoremi olayla ilgili olabilecek önceden bilinen bilgilere dayanarak olayın olasılığını açıklar.

Olasılık ifadesinin birçok kullanım şekli mecvuttur. Rassal bir K olayının herhangi başka bir olaydan bağımsız olarak gerçekleşme ifadesini kullanmak için p(K) notasyonunu kullanırız. Bayes teoremi ise rassal(stokastik) bir duruma bağlı olarak ortaya çıkan rastgele bir A olayı ile diğer rassal B olayı için aralarındaki ikişkiyi tanımlamaktadır. (Web-Bayes)

$$P(B \mid A) = \left(\frac{P(A|B)\ P(B)}{P(A)}\right) \tag{8}$$

Yukarıda bahsedilen formül genel yapısını B örnek uzayının parçalanmış şekli olarak şu şekle getirebiliriz.

$$B_1, B_2, \dots, B_k\ lar\ bir\ B\ örnek\ uzayının\ bir\ parçalanışı\ diyelim \tag{9}$$

$$B_i \cap\ B_j = \emptyset\ tüm\ i\ \neq j\ ler\ için \tag{10}$$

$$\bigcup_{i=1}^{k} B_i\ =\ X \tag{11}$$

$$P(B_i) > 0\ tüm\ i\ ler\ için \tag{12}$$

$$P(B_r|A) = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)} \tag{13}$$

Şekline getirebiliriz. Naive Bayes algoritmasının anlaşılması kolay olması, son yıllarda popüler olması bununla ilgili birçok akademik çalışmanın yapılmış olması tercih sebebidir. Metin sınıflandırma, Eposta spam filtreleme, Ağ Analizi vb. birçok konuda sınıflandırma işlemi için kullanılmaktadır. Naive Bayes yönteminin çalıştırılması ile elde edilen sonuçlar ise şu şekildedir.

Tablo 4.20 ve Tablo 4.21'de yukarıdaki verilerden anlaşılacağı üzere Naive Bayes yöntemi k-NN göre daha başarılı sonuçlar ortaya koymuştur. Bayes teoremini temel alan olasılıklı bir sınıflayıcıdır. Basitleştirilmiş varsayımlara rağmen Naive Bayes sınıflandırıcı gerçek dünya durumlarında beklenenden çok daha iyi sonuçlar vermiştir. Bir durumun olma ihtimalinin en yüksek olma koşuluna göre hareket eden bir tür sınıflandırma algoritmasıdır. Büyük veri kümeleri için

oldukça kullanışlı bir algoritma olduğunu, siyasi tahminleme eğilimlerinde veri setimizde Naive Bayes bir beklenti olasılığı üzerinden gitmesinden dolayı k-NN'e göre daha başarılı bir sonuç çıkarımında sağlanmıştır.

Tablo 4.20: Naive bayes algoritması için karışıklık matrisi

	DOĞRU P1	**DOĞRU P2**	**DOĞRU P5**	**DOĞRU P4**	**DOĞRU P3**	**SINIF KESİNLİK**
TAH. P1	87	68	18	7	0	48.33%
TAH. P2	13	887	38	67	3	88.00%
TAH. P5	7	181	496	29	0	69.57%
TAH. P4	15	55	4	599	32	84.96%
TAH. P3	0	1	0	10	40	78.43%
SINIF HASSSASİYET	71.31%	74.41%	89.21%	84.13%	53.33%	

Tablo 4.21: Naive bayes algoritması için doğruluk,kesinlik ve hassasiyet değerleri

	DOĞRULUK	KESİNLİK	HASSASİYET
DEĞERLER	%79.37 (+/-) 2.74%	% 74.70 (+/-) 3.51%	74.40% (+/-) 4.81%

4.4.3 Rassal Orman

Rassal Orman (random forest) diğer bir söylemle rastgele orman yöntemini anlayabilmek için öncelikle karar ağaçlarının öğrenilmesi ve bilgi sahibi olunması gerekmektedir.

Karar ağaçları, isminden anlaşılacağı üzere bir karar verilebilmesi için ağaç yapısına benzer bir yapı oluşturmaktadır. (Sayad,2015; Zhang 2006) Karar ağaçları yaprak düğümleri, karar düğümleri gibi terimleri barındırır. Karar düğümleri elde edilen veri seti içerisinde bir karara ulaşabilmek için, sınıflandırma veya tahminleme için kullanılır. Yapraklar ise sonuç olarak kararları tutmaktadır.

Karar ağaçları da denetimli bir öğrenme algoritmasıdır. Aynı k-NN algoritmasındaki gibi örnek verilerle bir sistem oluşturulur ve bu sisteme göre

sınıflandırma yapılmaktadır. Karar ağaçlarında ise en üstte kök düğümden en aşağıdaki yaprak uçlara kadar ilerlendiğinde elde edilen bütünlüğe yol adı verilir.

Rassal orman (Rassal orman) aynı karar ağaçları gibi denetimli bir öğrenme algoritmasıdır. Adından da anlaşılacağı üzere bir orman oluşturur ve bunu rastgele bir şekilde oluşturmaktadır. Oluşturulan orman genelde torbalama (bagging) yöntemiyle eğitilmiş karar ağaçları topluluğudur diyebiliriz. Torbalama yönteminin genel fikri öğrenme modellerinin kombinasyonunun tüm sonucu arttırmasıdır. Kısaca özetlemek gerekirse Rassal Orman algoritması oluşturduğu karar ağaçlarını doğruya en yakın ve istikrarlı bir tahmin etmek için bir araya toplamaktadır.

Rassal Orman algoritması geleneksel karar ağaçlarındaki veri ezberleme gibi dezavantajları ortadan kaldırmak adına ağaçları büyütürken algoritma tarafından rastgesellik katılmaktadır. Rassal orman algoritmasında kullanılan hiper parametresi modeli oluşturken veri setinin eğilimine ve problemin durumunda göre değişiklik parametresine hiper parametresi denilir.

Örnek olarak Bölüm 4.4.1'de k-NN algoritmasındaki *k* değerinin bizim tarafımızdan belirlenmesi diyebiliriz. Ağaç sayısı 100, maksimum derinlik 10 alınarak elde edilen sonuçlar Tablo 4.22'de Rassal orman algoritması karşıklık matrisinde gösterilmiştir. Tablo 4.23'te doğruluk, kesinlik ve hassasiyet değerleri verilmiştir.

Tablo 4.22: Rassal orman algoritması için karışıklık matrisi

	DOĞRU P1	**DOĞRU P2**	**DOĞRU P5**	**DOĞRU P4**	**DOĞRU P3**	**SINIF KESİNLİK**
TAH. P1	23	0	0	1	0	95.83%
TAH. P2	99	1176	308	280	18	62.52%
TAH. P5	0	9	0	425	2	97.48%
TAH. P4	0	1	0	4	55	91.67%
TAH. P3	0	5	1	46	67	56.30%
SINIF HASSSASİYET	18.85%	98.82%	44.00%	59.69%	73.33%	

Tablo 4.23: Rassal orman algoritması için doğruluk,kesinlik ve hassasiyet değerleri

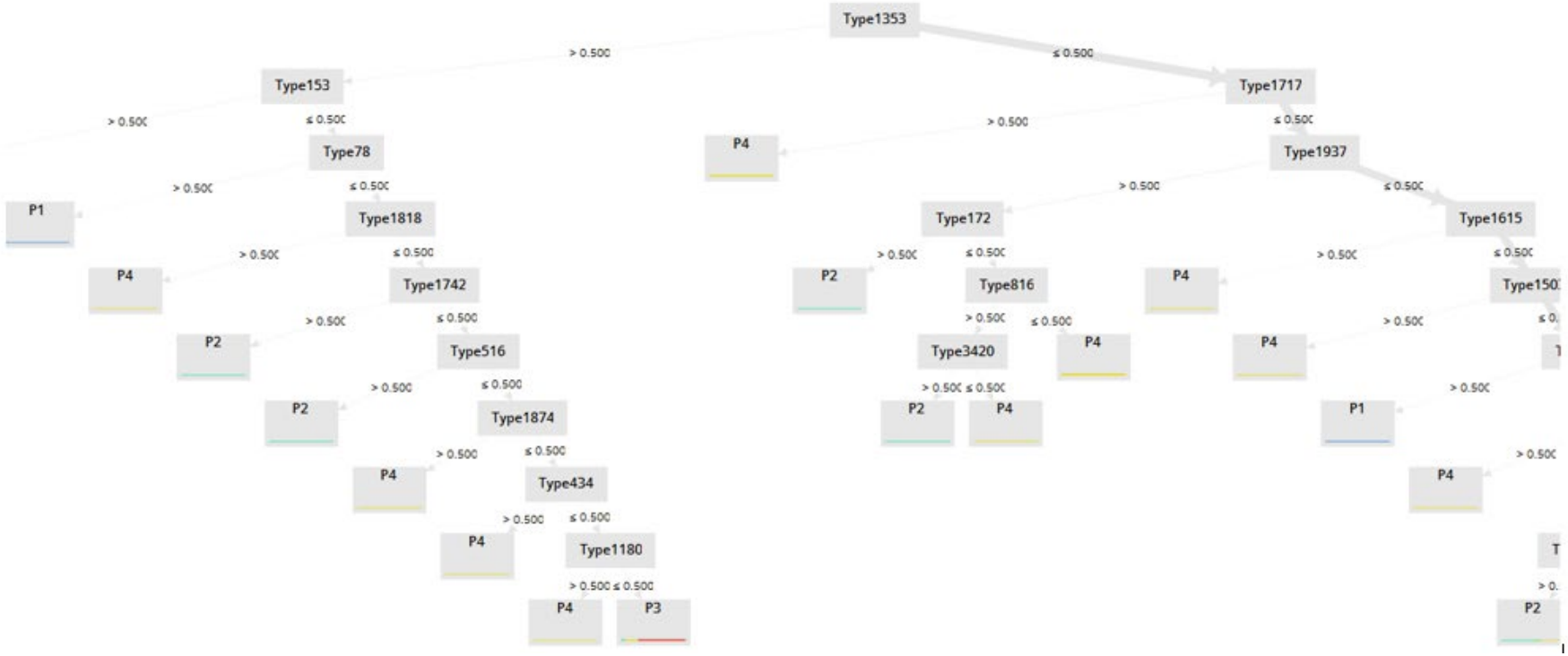
	DOĞRULUK	KESİNLİK	HASSASİYET
DEĞERLER	%72.52 (+/-) 2.29%	% 89.65 (+/-) 2.70%	59.01% (+/-) 2.59%

Rassal orman yöntemi k-NN gibi yakın bir başarı oranına sahip olmasına rağmen Bayes kadar başarılı bir doğruluk oranı yakalayamamıştır. Ağaç yapısı ve şekli 3000'den fazla kriter sütunu olmasından dolayı elde edilen her bir türe göre yeni bir kök ve dallar belirlenmiş rassallıktan dolayı birçok ağaç oluşması gerekmektedir.

Şekil 4.5'te gösterilen Type1234, Type234, Type546 gibi bilgiler aslında bir kişiyi siyasal anlamda kimliği belli olan kişilerle bağları belli etmektedir. Verinin anonimleştirme işlemi ile kişilerin isimleri Type123 gibi bir ifadeye çevrilmiştir. Şekil 4.5'te Rassal ormanda oluşturulan karar ağaçlarından bir tanesi örnek olarak gösterime sunulmuştur.

Çok fazla sütun olduğu ve çok fazla tür olduğu için ağaç yapısı Şekil 4.5'te resimde net olarak seçilememektedir. Ağaç yapısındaki gösterim olması için bir kesit sunulmuştur. Fakat bazı türlerin sütunların parti ayrımında kesin bir çizgi sağladığı gözlenmektedir. Bu tür sütunların incelemesi yapıldığında siyasi parti temsilcilerinden bazı kişilerin önemli bir siyasi eğiliminin olduğu gözükmektedir.

Yukarıdaki örneklerden de anlaşılacağı üzere ağaç sayısının (number of tree) degeri 100 olarak belirlenirse sınıflandırma sonucu doğruya en yakın olmaktadır.



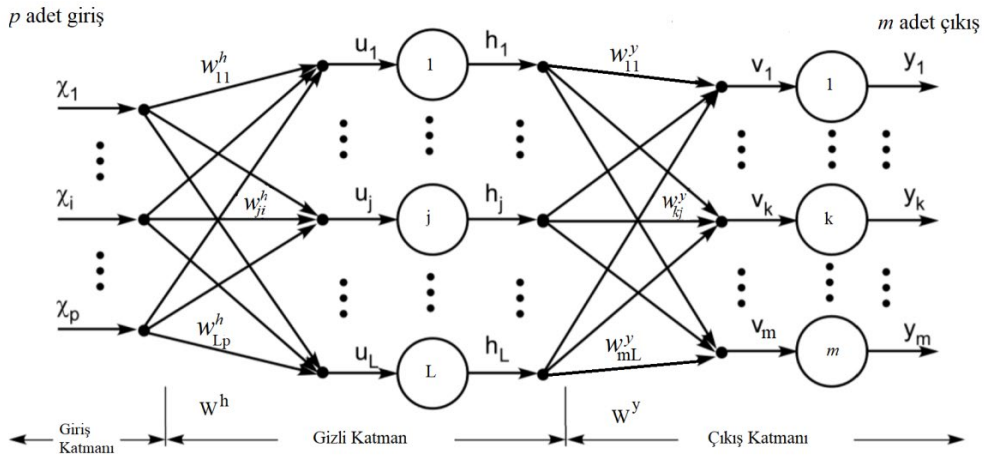
Şekil 4.5: Rassal orman yöntemi ile oluşturulan ağaç yapısı

4.4.4 Derin Öğrenme

RapidMiner derin öğrenme (deep learning) yönteminde kullanılan temel yapı çok katmanlı ileri beslemeli yapay sinir ağları geri yayılımlı bir yapı önerilmektedir. İleri beslemeli yapay sinir ağlarında temel olarak 3 çeşit katman (layer) bulunur. Giriş, gizli ve çıkış olmak üzere 3 katman sırası ile yapay sinir ağlarına giren verileri tutan giriş katmanı, işlemlerin yapıldığı ve çıkış değerinin gözlemlendiği katmanıdır.

İleri beslemeli sinir ağlarında işaret akışına tek yönlü olarak izin verilir. Örnek gösterilmesi açışısından tek gizli katmanı olan üç katmanlı bir ileri beslemeli sinir ağı örneği Şekil 4.6'da gösterilmiştir.

Şekil 4.6'da görüldüğü gibi çok katmanlı ileri beslemeli yapay sinir ağlarında, hücreler katmanlar şeklinde düzenlenir ve bir katmandaki hücrelerin çıkışları bir sonraki katmanlardaki ağırlıklar üzerinden giriş yapılır. En çok bilinen ve kullanılan geriye yaylım öğrenme algoritması ise bu yapar sinir ağları ile etkin bir şekilde kullanılmaktadır.



Şekil 4.6: Çok katmanlı ileri beslemeli yapay sinir ağı modeli

Geriye yayılım algoritması verilerin sınıflandırılmasında kullanılmış ve ilk olarak Webos daha sonra Paerker, Tummelhart ve McCleland tarafından geliştirilen bir geri yayılım ağıdır. Yayılma ve uyum gösterme olmak üzere temel 2 aşaması bulunmaktadır. İsminin bu şekilde olmasının sebebi ise var olan hataları geriye

doğru yani çıkış katmanından giriş katmanına doğru ilerletmesidir (Çetin vd., 2006). Yapay Sinir Ağları (YSA) birçok farklı yöntemi ve algoritmayı içerisinde barındıran gelişmiş bir sistemdir.

YSA'ların bir evrimi olan derin öğrenme sinir ağları, bilgisayar teknolojilerindeki yeni gelişmelerle birlikte, neo-kortekste bulunan nöronların insan beyni aktivitesini daha iyi taklit etmeye çalışır. Derin öğrenme için geliştirilen modeller, ses, görüntü ve diğer verilerin dijital gösterimlerindeki karmaşık desenleri tanımak için herhangi bir öznitelik çıkarım işlemine gerek duymaksınız eğitilebilir (Krizhevsky, 2012). Bilgisayar teknolojilerindeki gelişmeler nedeniyle, artık insanlar şimdiye kadar olduğundan çok daha fazla yapay sinir ağı hücre tabakasını modelleyebilmektedir (LeCun vd., 2015; Nguyen ve Halem, 2019). Derin katman mimarileri ile bu yeni eğitim teknikleri konuşma ve görüntü tanımada kayda değer gelişmeler sağlamaktadır. Bu gelişmeler insan DNA dizilimlerinin incelenmesinden, tıpta yeni ilaçlara yol açan molekülleri tanımlamaya kadar farklı alanlarda kullanılmaktadır. Uzun-kısa süreli bellek (long-short term memory) modeline sahip tekrarlayan sinir ağları (recursive neural networks) uzun süreli geçici bağımlılıklar veya dil modellemesi için zaman serileri verilerini başarıyla öğrenebilen derin öğrenme modellerinden biridir (Hochreiter and Schmidhuber, 1997). Dünya ve uzay bilimleri ve diğer bilim disiplinleri gibi uygulamalar için derin öğrenme modellerinin genişletilmesi hem kavramsal açıdan hem de bilgi işlem gücünde daha ileri gelişmeler gerektirecektir (Nguyen ve Halem, 2019).

Bu tez çalışmasında, tahminleme aşaması için kullanılan Derin Öğrenme modelindeki temel amaç siyasi parti eğilimlerinin Twitter üzerinden elde ettiğimiz veri setlerinde kullanılmasıdır. Tahminleme üzerinde kullanılan birçok algoritma ile derin öğrenme arasındaki farklılıkların olması ve YSA'nın herhangi bir parametresi ile oynayarak başarı oranında düşüş veya yükseliş gözüktüğü için tercih edilmiştir. Bu gelişmiş yapıyı kullanmak için RapidMiner'ın bize sunduğu deep learning modeli üzerinden ilerleme kaydedilmiştir.

Şekil 4.7'da RapidMiner derin öğrenme ekran görüntüsü gösterilmiştir. Bu şekilde gözüktüğü gibi bu model bize gelişmiş bir altyapıyı analizler kullanmamız için sunmuş durumdadır. Burada kullanılan parametrelerin değerlerine göre elde

edilen sonuçlar farklı olacağı için kullanılan parametre değer açıklamalarının incelemesi aşağıdaki gibi yapılmıştır.

Bazı aktivasyon fonksiyonlarının (Activation Function), herhangi bir çıktı üretimi sağlanmadan önce erişilmesi gereken bir eşik değeri vardır. RapidMiner 8 çeşit aktivasyon yöntemini kullanıma sunmaktadır. Ayrıca gizli katmanların sayısının ne olacağını, kaybetme fonksiyonun(loss function) ne olacağını (quadric mi crossentropy mi) ve dağıtma fonksiyonun(distributian function) türünün ne olacağını (bernoulli mi multinomial mi) gibi birden fazla fonksiyon seçimine kadar gelişmiş bir alt yapı sistemi sunmaktadır.

Şekil 4.7'de görüldüğü üzere derin öğrenme yöntemimiz eğitim için bir veri seti almaktadır. Bu eğitim veri seti çapraz doğrulama yönteminden ayrılan kısım olarak gelmektedir. Buradan elde edilen model ile yine çapraz doğrulama da test olarak ayrılan veri seti bir etiketleme ile tahminlemede bulunur ve bu tahminleme sonucu olarak tüm yöntemlerde bahsettiğimiz karışıklık matrisi mevcuttur.

Şekil 4.7: RapidMiner derin öğrenme ekran görüntüsü

Derin Öğrenme yönteminde parametrelerin başarı oranına etkisini gözlemlemek için epoch, dağıtım fonksiyonu paremetrelerinin etkisi incelenmiştir. Tablo 4.24'teki karışıklık matrisi ve Tablo 4.25'te görüldüğü üzere doğruluk oranımız %87.77 çıkmıştır. Buna bağlı olarak kesinlik değerlerimiz %87.93, hassasiyet değerlerimiz ise %87.20 çıkmıştır. Bu sonuçlardan, ilgili veri setine ait örneklem kümesi ve seçilen öznitelikler için en başarılı algoritmanın derin öğrenme algoritması olduğu görülmektedir. Derin öğrenme algoritmasında kullanılan gizli katman sayısı 40'tır. Epoch değerleri tablolarda gösterimi sağlanmıştır.

Tablo 4.24: Derin öğrenme algoritması için karışıklık matrisi (Epoch10)

	DOĞRU P1	**DOĞRU P2**	**DOĞRU P5**	**DOĞRU P4**	**DOĞRU P3**	**SINIF KESİNLİK**
TAH. P1	113	20	18	7	2	70.62%
TAH. P2	8	1103	79	95	7	85.37%
TAH. P5	0	30	447	9	0	91.98%
TAH. P4	1	35	6	598	2	93.15%
TAH. P3	0	2	0	3	64	92.75%
SINIF HASSSASİYET	92.62%	92.69%	81.27%	83.99%	85.33%	

Tablo 4.25: Derin öğrenme algoritması için doğruluk,kesinlik ve hassasiyet değerleri (Epoch10)

	DOĞRULUK	KESİNLİK	HASSASİYET
DEĞERLER	%87.77 (+/-) 1.86%	% 87.93 (+/-) 4.32%	87.20% (+/-) 2.63%

Tablo 4.26'da derin öğrenme algoritmasındaki devir(epoch) 3 seçilmiş olup k-fold için değerimiz k=10 seçilmiştir. Elde edilen değerlere göre doğruluk oranımız Tablo 4.27'de görüldüğü üzere ortalama %83.16 başarı elde edildiği gözlenmiştir.

Tablo 4.26: Derin öğrenme algoritması için karışıklık matrisi (Epoch3)

	DOĞRU P1	**DOĞRU P2**	**DOĞRU P5**	**DOĞRU P4**	**DOĞRU P3**	**SINIF KESİNLİK**
TAH. P1	107	36	31	15	4	55.44%
TAH. P2	12	1103	97	166	14	79.24%
TAH. P5	1	28	419	9	0	91.68%
TAH. P4	2	20	2	517	0	95.56%
TAH. P3	0	3	1	5	57	86.36%
SINIF HASSSASİYET	87.70%	92.69%	76.18%	72.61%	76.00%	

Tablo 4.27: Derin öğrenme algoritması için doğruluk,kesinlik ve hassasiyet değerleri (Epoch3)

	DOĞRULUK	KESİNLİK(pre	HASSASİYET(re
DEĞERLER	%83.16 (+/-) 2.79%	% 83.95 (+/-) 4.94%	81.17% (+/-) 3.52%

Tablo 4.28'de derin öğrenme algoritmasındaki dağıtım fonksiyonu bernoulli seçişmiş olup k-fold için değerimiz *k*=10 seçilmiştir. Elde edilen değerlere göre doğruluk oranımız Tablo 4.29'da görüldüğü üzere ortalama %87.39 başarı elde edildiği gözlenmiştir.

Tablo 4.28: Derin öğrenme algoritması için karışıklık matrisi (Bernoulli)

	DOĞRU P1	**DOĞRU P2**	**DOĞRU P5**	**DOĞRU P4**	**DOĞRU P3**	**SINIF KESİNLİK**
TAH. P1	105	13	12	4	0	78.36%
TAH. P2	14	1107	80	110	11	83.74%
TAH. P5	0	33	454	9	0	91.53%
TAH. P4	3	36	4	587	2	92.88%
TAH. P3	0	1	0	2	62	95.38%
SINIF HASSSASİYET	86.07%	93.03%	82.55%	82.44%	82.67%	

Tablo 4.29: Derin öğrenme algoritması için doğruluk,kesinlik ve hassasiyet değerleri (Bernoulli)

	DOĞRULUK	KESİNLİK	HASSASİYET
DEĞERLER	%87.39 (+/-) 2.39%	% 89.33 (+/-) 3.87%	85.44% (+/-) 3.31%

5. SONUÇ VE ÖNERİLER

Yapılan araştırmalarda verilerin elde edilmesi sürecinden analiz çalışmalarına kadar siyasi eğilim tahminleme işleminin birçok parametereye bağımlı zor bir süreç olduğu gözlemlenmiştir. Bu yüzden, veri seti daraltılarak siyasi parti yönetimlerinin Twitter üzerinde aktif veya siyasi eğilimi belli olan yazar, gazeteci ve akademisyenlerin alındığı 2650 kişilik veri setinde kişilerin birbiri ile olan ilişkileri (arkadaşlık matrsilerinin), Twitter'in sağladığı bazı veriler ve bu verilerden elde edilen uzman verisine ait özel veriler ile başarılı sayılabilecek bir sınıflandırma işlemi yapıldığı gözlenmiştir.

Tezin veri setindeki girişlerin çok ve derin öğrenme sistemlerinin yapay sinir ağları için uygun olmasından dolayı en başarılı sonuçlar derin öğrenme sistemi ile elde edilmiştir. Veri setimize göre en kötü sonuçlar ise lazy algoritmalarından k-NN ile veri setimizin normalizasyon yapılmadan önceki durumu ile elde edilmiştir. Tüm veriler normalizasyon işlemleri dahil optimum parametrelere göre tekrar değerlendirildiğinde en kötü doğruluk oranı Rassal Orman yöntemi ile elde edildiği gözlenmiştir.

Veri setinin algoritmaların uygulanma aşamasından önce ön işlemlerin ne kadar başarılı olduğu k-NN algoritmasındaki yaptığımız normalizasyon gibi ön işlemler sayesinde doğruluk %29 oranında artırılmıştır. Tablo 5.1, Tablo 5.2, ve Tablo 5.3'te doğruluk, kesinlik, hassasiyet gösterimi sunulmuştur.

Tablo 5.1: Normalizasyon öncesi k-nn sonucu

Algoritma	Doğruluk	Kesinlik	Hassasiyet
k-NN	%50.84 (+/-) 2.06%	%27.57 (+/-) 2.35%	%28.50 (+/-) 1.54%

Tablo 5.2: Normalizasyon sonrası k-nn sonucu

Algoritma	Doğruluk	Kesinlik	Hassasiyet
k-NN	%79.31 (+/-) 2.13%	%74.76 (+/-) 3.55%	80.36% (+/-) 3.56%

Tablo 5.3'teki gösterimler k-NN'deki *k* değerleri arasındaki en başarılı *k* için gösterim sağlanmış olup, ağaç yapılarındaki ağaç sayısı ve diğer girdilerin parametrelerin en optimum sonuca göre elde ettiğimiz sonuçların karşılaştırılması sağlanmıştır. Böylelikle parametlerin optimum değerleri için en yüksek sonuçlar karşılaştırılarak yorumlanmıştır. Tablo 5.3'te Algoritmaların karşılaştırılmaları incelendiğinde en başarılı algoritma derin öğrenme olmuş ve %87.77 başarı oranı gözlenmiştir.

Tablo 5.3: Algoritmaların karşılaştırmaları

Algoritmalar	**Doğruluk**	**Hassasiyet**	**Kesinlik**
k-NN	%79.31 (+/-) 2.13%	% 74.76 (+/-) 3.55%	80.36% (+/-) 3.56%
Naive Bayes	%79.37 (+/-) 2.74%	% 74.70 (+/-) 3.51%	74.40% (+/-) 4.81%
Rassal Orman	%72.52 (+/-) 2.29%	% 89.65 (+/-) 2.70%	59.01% (+/-) 2.59%
Derin Öğrenme	%87.77 (+/-) 1.86%	% 87.93 (+/-) 4.32%	87.20% (+/-) 2.63%

Makine öğrenmesi algoritmalarında, yapay zekâ sistemlerinde kullanılan algoritmalarda veriler eğitim ve test olarak ayrılmaktadır. Burada eğitim verisi yazılım öğrenme işlemi yapılırken kullanılırken, test verisi ile birlikte bu işlemin başarı oranı gözlemlenir. Burada test verisinin daha homojenik olarak değerlendirilmesi için daha önceki bölümde önerdiğimiz k-Fold yöntemi ile değerlendirmeyi biraz daha homojenik hale getirmeye çalıştık. Fakat bu yapılan işlem eğitim verisinin ne kadar doğru seçildiği ile ilgili bir bilgi veya yöntem içermemektedir. Bu sebeple verilerimizde eğitimde kullandığımız verilerin niteliklerinin ise sisteme verdiği duyarlılığını analiz etmemiz en önemli etkenlerden bir tanesidir. Örnek ile açıklamak gerekirse boy ve kilo ile bir sınıflandırma yaptığımızı varsayalım ve bu sınıflandırma işlemini Bölüm 4.4.1'de kullanılan k-NN ile yaptığımızı varsayalım bu işlem sonucunda elde edeceğimiz sınıflandırma sonuç başarı oranı boy ve kilo yani bilgilerin ilişkisi ile önemlidir. Bunun yerine boy ve yaş unsurunu ortaya koyduğumuzda aynı algoritma aynı sistem ile test ettiğimizde tahminlememizin daha doğru veya yanlış olduğunu görüntüleyebiliyoruz.

Bu sebeple girdilerin duyarlılık analizinin yapılması veri analiz çalışmalarında önemli bir yere sahiptir. Fakat tezimizin asıl amacı farklı olması ve bu duyarlılık analizin tüm çalışmalarla ilgili sonuçları ayrı bir tez veya makale çalışması olması sebebi ile gelecek çalışmalarımıza bırakılmıştır.

Toplam 5 parti analiz için veri setimize dahil edilmiştir. Bu sayının artması tüm partilerin dahil edilmesi gelecek çalışmalara bırakılan diğer bir konudur. Ayrıca 4 farklı yöntem ile yaptığımız analiz çalışmaları daha farklı yöntemler ve veri setindeki değişiklikler ile sonuçların incelenmesi gelecek çalışmalara bırakılmıştır.

6. KAYNAKLAR

Abel, F., Gao, Q., Houben, G.J., and Tao, K., "Analyzing temporal dynamics in Twitter profiles for personalized recommendations in the social web", *Proceedings of the 3rd ACM International Web Science Conference*, (2011).

Aggarwal, C.C., An introduction to social network data analytics. (ed: Aggarwal C.), *Social Network Data Analytics,* Boston: Springer, 1-15, (2011).

Al-Deen, H.S.N. and Hendricks, J.A., *Social media: usage and impact*, Plymouthi U.K.: Lexington Books, (2011).

Alshammari, A., Kapetanakis, S., Evans, R., Polatidis, N., and Alshammari, G., "User modelling on Twitter with exploiting explicit relationships for personalized recommendations", (eds: Madureira A.M., Abraham A., Gandhi, N., Varela, M.L.), *Hybrid Intelligent Systems, Advances in Intelligent Systems and Computing*, 923, Charm: Springer, 135-145, (2018).

America, K., "Twitter as an influence on the quality of online interpersonal relationships and language use", Magister Artium in the Department of Linguistics Thesis, *Department of Linguistics*, University of Western Cape, Cape,Town, South Africa, (2013).

Anderson, P.J., and McLeod, A., "The great non-communicator? The mass communication deficit of the European Parliament and its press directorate", *Journal of Common Market Studies*, 42(5), 897–917, (2004).

Argamon, S., Dhawle, S., Koppel, M., and Pennebaker J.W., "Lexical predictors of personality type", *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*, (2005).

Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L. and Hughes, M. "Sentiment analysis of political tweets: towards an accurate classifier", *Proceedings of the NAACL Workshop on Language Analysis in Social Media*, Atlanta, Georgia, June 13, 49-58, (2013).

Bastem, K. and Şeker, Ş.E., "Veri madenciliği yöntemleri ile Twitter üzerinden MBTI kişilik tipi analizi", *Yönetim Bilişim Sistemleri Ansiklopedisi,* 4(2), (2017).

Bektaş Şeker, T., *İnternet ve Bilgi Açığı*, Konya: Çizgi Kitapevi, (2005)

Bhola, A. "Twitter and polls: analyzing and estimating political orientation of Twitter users in India general #elections2014", Master of Technology in Computer Science, *Indraprastha Institute of Information Technology*, New Delhi, India, (2014).

Bilal, M., Gani, A., Lali, M.I.U., Marjani, M. and Malik, N., "Social profiling: a review, taxonomy, and challenges", *Cyberpsychology, Behavior, and Social Networking*, 22(7), 433-450, (2019).

Blackshaw, P. and Nazzaro, M., "Consumer-generated media (CGM) 101: word-of-mouth in the age of the web-fortified consumer [online]", (21/06/2019), http://www.nielsenbuzzmetrics.com/whitepapers, (2004).

Blais, A. and Rubenson, D., "The source of turnout decline: new values or new contexts?", *Comparative Political Studies*, 46,(1), 95-117, (2013).

Boyd, D.M. and Ellison, N.B. "Social network sites: definition, history, and scholarship", *Journal of Computer-Mediated Communication,* 13, 210-230, (2007).

Boyd, D., Golder, S. and Lotan, G. "Tweet, tweet, retweet: conversational aspects of retweeting on Twitter", *Proceedings of the 43rd IEEE Hawaii International Conference on System Sciences*, 05-08 January, 1-10, (2010).

Briatte, F. and Gallic, E., "Recovering the French Party Space from Twitter Data [online]", (04/06/2019), https://halshs.archives-ouvertes.fr/halshs-01511384/document, (2015).

Briggs-Myers, I. and Briggs, K.C. *Myers-Briggs Type Indicator: MBTI*, Palo Alto, California: Consulting Psychologists Press, (1988).

Çetin, M., Uğur, A., ve Bayzan, Ş., "İleri beslemeli yapay sinir ağlarında backpropagation (geriye yayılım) algoritmasının sezgisel yaklaşımı", *Akademik Bilişim Kongresi, Pamukkale Üniversitesi,* Denizli, (2006).

Chang, T., Chopra, V., Zhang, C. and Woolford, S.J., "The role of social media in online weight management: systematic review", *Journal of Medical Internet Research*, 15(11), e262, (2013).

Chen, J., Nairn, R., Nelson, L., Bernstein, M. and Chi, E. "Short and tweet: experiments on recommending content from information streams", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1185–1194, (2010).

Chen, M., Mao, S. and Liu, Y., "Big data: a survey", *Mobile Networks and Applications*", 19(2), 171–209, (2014).

Chin, D.N. and Wright, W.R., "Social media sources for personality profiling", *Proceedings of the 2nd Workshop on Emotions and Personality in Personalized Services*, Aalborg, Denmark, (2014).

Clement, J., "Google's ad revenue from 2001 to 2018 (in billion U.S. dollars) [online]", (01/07/2019), https://www.statista.com/statistics/266249/advertising-revenue-of-google/, (2019).

Clemons, E. K., Barnett, S. and Appadurai, A. "The future of advertising and the value of social network websites: some preliminary examinations",

Proceedings of the 9th International Conference on Electronic Commerce, Minneapolis, USA, 267-276, (2007).

Conover, M.D., Gonçalves, B., Ratkiewicz, J., Flammini, A. and Menczer, F., "Predicting the political alignment of Twitter users", *Proceedings of the IEEE Third International Conference on Social Computing*, Boston, MA, USA, 9-11 October, (2011).

Dalton, R., *Democratic Challenges, Democratic Choices: The Erosion of Political Support in Advanced Industrial Democracies*, Oxford: Oxford University Press, (2004).

Demirel, Y., Seçkin, Z. ve Özçınar, M.F., "Örgütsel iletişim ile örgütsel vatandaşlık davranışı arasındaki ilişki üzerine bir araştırma", *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 20(2), 33-48, (2011).

Dennis, J., "#stopslacktivism: why clicks, likes, and shares matter", (ed: Dennis, J.), *Beyond Slacktivism: Political Participation on Social Media, Interest Groups, Advocacy and Democracy Series*, Cham: Palgrave Macmillan, 25-69, (2019).

Drury, G., "Opinion piece: social media: should marketers engage and how can it be done effectively?", *Journal of Direct Data and Digital Marketing Practice*, 9(3), 274–277, (2008).

Dube, R., "Five major characteristics of social networks [online]", (15/07/2019), https://socialnetworking.lovetoknow.com/Characteristics_ of_Social_Networks, (2011).

Dwi Prasetyo, N. and Hauff, C., "Twitter-based election prediction in the developing world", *Proceedings of the 26th ACM Conference on Hypertext and Social Media*, 149-158, (2015).

Esparza, S.G., O'Mahony, M.P. and Smyth, B., "CatStream: categorising tweets for user profiling and stream filtering", *Proceedings of the 2013 International Conference on Intelligent User Interfaces, Santa Monica*, California, USA, *25-36,* (2013).

Fan, W. and Gordon, M.D. "The power of social media analytics", *Communications of the ACM*, 57(6), 74-81, (2014).

Farook, F.S. and Abeysekara, N., "Influence of social media marketing on customer engagement", *International Journal of Business and Management Invention*, 5(12), 115-125, (2016).

Fernandes de Mello Araújo E. and Ebbelaar D., "Detecting Dutch political tweets: a classifier based on voting system using supervised learning", *Proceedings of the 10th International Conference on Agents and Artificial Intelligence* , Volume 2, 462-469, (2018).

Friedman, J.H., "Greedy function approximation: a gradient boosting machine", *Annals of Statistics*, 29(5), 1189–1232, (2001).

Genç, H., "İnternetteki etkileşim merkezi sosyal ağlar ve e-iş 2.0 Uygulamaları", *XII. Akademik Bilişim Konferansı,* Muğla Üniversitesi, Muğla, (2010).

Ghazi, M.R. and Gangodkar, D., "Hadoop, MapReduce and HDFS: a developers perspective", *Procedia Computer Science*, 48, 45-50, (2015).

Gomes, J.E.A., Prudencio, R.B.C. and Nascimento, A., "A comparative study of group profiling techniques in co-authorship networks", *Proceedings of the 5th IEEE Brazilian Conference on Intelligent Systems*, 373-378, (2016).

Hannon, J., Bennett, M., and Smyth, B., "Recommending Twitter users to follow using content and collaborative filtering approaches", *Proceedings of the 4th ACM Conference on Recommender Systems*, 199–206, (2010).

Hartwell, C.J., "The use of social media in employee selection: prevalence, content, perceived usefulness, and influence on hiring decisions", PhD Thesis, *Purdue University*, Indiana, USA, (2015).

Hochreiter, S., and Schmidhuber, J. "Long short-term memory", *Neural Computation*, 9(8), 1735–1780, (1997).

Hoffman, D. and Fodor, M., "Can you measure the ROI of your social media marketing?", *Sloan Management Review*, 52(1), (2010).

Honeycutt, C. and Herring, S., "Beyond microblogging: conversation and collaboration in Twitter", *Proceedings of the 42nd IEEE Hawaii International Conference on System Sciences*, 1-10, (2009).

Hu, X. and Liu, H., "Social media, mining and profiling in", (eds: Mansell, R. and Ang, P.H.), *The International Encyclopedia of Digital Communication and Society*, John Wiley & Sons, 1–6, (2015).

Ikeda, K., Hattori, G., Ono, C. and Asoh, H., Higashino, T., "Twitter user profiling based on text and community mining for market analysis", *Knowledge-Based Systems*, 51, 35-47, (2013).

Islam, J. and Zhang, Y., "Visual sentiment analysis for social images using transfer learning approach", *Proceedings of the IEEE International Conference on Big Data Cloud Computing (BDCloud), Soc. Comput. Netw. (SocialCom), Sustain. Comput. Commun.*, 124-130, (2016).

Ituski, H., Matsubara, H., Arita, K. and Omi, K. "Effective clusterization of political tweets using kurtosis and community duration", *Proceedings of the IEEE International Conference on Social Computing*, Alexandria, VA, USA, 928-931, (2013).

Johnson, K. and Goldwasser, D., "Identifying stance by analyzing Political discourse on Twitter", *Proceedings of the EMNLP Workshop on Natural Language Processing and Computational Social Science*, Austin, Texas, USA, 66–75, (2016).

Kaplan, A.M. and Hainlein, M., "Users of the world, unite! The challenges and opportunities of social media", *Business Horizons*, 53(1), 59-68, (2010).

Kıraç, E., "Örgütsel iletişimin örgütsel bağlılık algılaması üzerindeki etkileri ve bir araştırma", Yüksek Lisans Tezi, *Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü*, Denizli, (2012).

Krizhevsky, A., Sutskever, I., and Hinton, G.E., "Imagenet classification with deep convolutional neural networks", *Proceedings of the Advances in neural Information Processing Systems Conference*, 1097-1105, (2012).

Lashkari, A.H., Chen, M. and Ghorbani, A.A., "A survey on user profiling model for anomaly detection in cyberspace", *Journal of Cyber Security and Mobility*, 8(1), 75-112, (2019).

LeCun, Y., Bengio, Y. and Hinton, G.E. "Deep learning", *Nature*, 521, 436– 444, (2015).

Mangold, W.G. and Faulds, D.J., "Social media: the new hybrid element of the promotion mix", *Business Horizons*, 52, 357-365, (2009).

Mayfield, A., "What is social media, iCrossing [online]", (05.01.2019), http://www.icrossing.co.uk/fileadmin/uploads/eBooks/What_is_Social_Media_iCrossing_ebook.pdf, (2008).

Nash, K.M., "Social media in the workplace: new technology, old problems [online]", (10/05/2019), http://www.thehrspecialist.com, (2009).

Nguyen, P. and Halem, M. "Deep learning models for predicting CO2 flux employing multivariate time series", *Proceedings of the 5th KDD Workshop on Mining and Learning from Time Series*, August 5th, 2019, Anchorage, Alaska, USA.

Oikonomou, L. and Tjortjis, C., "A method for predicting the winner of the USA presidential elections using data extracted from Twitter", *Proceedings of the IEEE South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conference*, Kastoria, Greece, (2018).

Özay, M.A., "Siyasi parti liderlerinin 2017 yılında paylaştıkları tweetlerin nitel analizi", *Van Yüzüncü Yıl Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 3(5), 181-193, (2018).

Özkan, Y., *Veri Madenciliği Yöntemleri*, İstanbul: Papatya Yayıncılık, (2016).

Pear Analytics, "Twitter study – august 2009 [online]", (13/05/2019), https://pearanalytics.com/wp-content/uploads/2012/12/Twitter-Study-August-2009.pdf, (2009).

Pennacchiotti, M., Popescu, A.M., "A machine learning approach to Twitter user classification", *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 281-288, (2011).

Pennebaker, J., Chung, C. and Ireland, M., "The development and psychometric properties of LIWC 2007", Austin, Texas, USA, (2007).

Picard, R.W., *Affective Computing*, Massachusetts, USA: The MIT Press, (2000).

Pine, D.J., *Introduction to Python for Science and Engineering*, Boca Raton: CRC Press, (2019).

Pla F. ve Hurtado L.F., "Political tendency identification in Twitter using sentiment analysis techniques", *Proceedings of the 25th International Conference on Computational Linguistics*, Dublin, Ireland, August 23-29, 183–192, (2014).

Qing, Q., "Factors that influence interpersonal communication: culture, power and technology [online]", (15/07/2019), https://numerons.files.wordpress.com/2012/04/05-factors-that-influence-interpersonal-communication-imp.pdf, (2007).

Rao, D., Yarowsky, D., Shreevats, A., Gupta, M., "Classifying latent user attributes in Twitter*", Proceedings of the 10th International Workshop on Search and Mining User-generated Contents*, 710–718, (2010).

Razak, R.A., Fahrurazi, F.R., "Agile testing with Selenium", *Proceedings of the IEEE 5th Malaysian Conference in Software Engineering,* Malezya, 217-219, (2015).

Rhim, J., Lee, S., Doh, Y.Y., "Discovery of smartphone user group profiling based on user's motivations and usage behaviors through focus group interviews", (ed: P.-L.P. Rau), *Cross-Cultural Design, Lecture Notes in Computer Science*, 9741, 426–435, Cham: Springer International Publishing, (2016).

Rigolin, V.H., "What is Twitter? How do I get started? Why should I become a user?", *Journal of the American Society of Echocardiography*, 31(3), A31-A32, (2018).

Ross, W. and Slovensky, R., "Using the Internet to attract and evaluate job candidates", (ed: Z. Yan), *Encyclopedia of Cyber Behavior*, Vol.2, 537-549. Hershey, PA: IGI Global, (2012).

Sanders, E., de Gier, M., van den Bosch, A., "Using demographics in predicting election results with Twitter", (eds: Spiro E., Ahn YY.), *Social Informatics, Lecture Notes in Computer Science*, 10047, Cham: Springer, (2016).

Sarkar, D., *Text Analytics with Python*, California: Apress, 69–114, (2019).

Sayımer, İ., *Sanal Ortamda Halkla İlişkiler*, İstanbul: Beta Yayınları, (2008).

Schiaffino, S., and Amandi, A., "Intelligent user profiling", (ed: Bramer M.), *Artificial Intelligence An International Perspective, Lecture Notes in Computer Science*, 5640, 193-216, Heidelberg: Springer, (2009).

Scott, J. and Carrington, P.J. *The SAGE Handbook of Social Network Analysis*, SAGE Publishing, London:UK, (2011).

Scott, J., "Social network analysis*", Sociology*, 22(1), 109–127, (1988).

Seethalakshmi, V., Govindasamy, V. and Akila, V., "Job scheduling in big data: a survey", *Proceedings of the IEEE International Conference on Computation of Power, Energy, Information and Communication*, 23-31, (2018).

Segalin, C., Cheng, D. S., and Cristani, M. "Social profiling through image understanding: Personality inference using convolutional neural networks", *Computer Vision and Image Understanding*, 156, 34–50, (2017).

Selenium Home Page [online], (11/07/2019), https://www.seleniumhq. org/, (2019).

Social Report, "The ultimate guide to social media post lengths in 2019 [online]", (01/07/2019), https://www.socialreport.com/help-center/article/360020940251-The-Ultimate-Guide-to-Social-Media-Post-Lengths-in-2019#h_39578906767515447650941 68, (2019).

Solmaz, B., Tekin, G., Herzem, Z. and Demir, M., "İnternet ve sosyal medya kullanımı üzerine bir uygulama", *Selçuk İletişim Dergisi*, 7(4), 23-32, (2013).

Srinivas, K. and Prakash, C., "A comparative study of testing framework with special emphasis on Selenium for financial applications", *International Journal of Soft Computing,* 12(3), 148-155, (2017).

Stewart, M.C., and Arnold, C.L. "Defining social listening: recognizing an emerging dimension of listening", *International Journal of Listening*, 32, 85–100, (2018).

Şahin, E., "Sosyal medya hesaplarının kural tabanlı profil çıkarımı: kullanıcı siyasi eğilimlerinin sınıflandırılması ve araştırılması", Yüksek Lisans Tezi, *Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Ana Bilim Dalı*, Denizli, (2018).

Şeker, S.E., "Bilgi getirimi ve çıkarımı [online]", (02/07/2019), http://bilgisayarkavramlari.sadievrenseker.com/2010/10/01/bilgi-getirimi-ve-cikarimi-information-retrieval-and-extraction/, (2010).

Şeker, S.E., "Sosyal ağlarda veri madenciliği", *Yönetim Bilişim Sistemleri Ansiklopedi*, 2(2), 30-39, (2015).

Tang, D., Wei, F., Qin, B., Liu, T., Zhou, M. "Coooolll: A Deep Learning System for Twitter Sentiment Classification", *Proceedings of the 8th International Workshop on Semantic Evaluation*, 208–212, Dublin, Ireland, August 23-24, (2014).

Tang, L., Wang, X. and Liu, H., "Group profiling for understanding social structures", *ACM Transactions on Intelligent Systems and Technology*, 3(1), 1-25, (2011).

Troussas, C., Virvou, M., Caro, J. and Espinosa, K.J., "Language learning assisted by group profiling in social networks", *International Journal of Emerging Technologies in Learning*, 8(3), 35-38, (2013).

Tumasjan, A., Sprenger, T.O., Sandner, P.G. and Welpe, I.M., "Predicting elections with Twitter: What 140 characters reveal about political sentiment", *Proceedings of the Fourth AAAI International Conference on Weblogs and Social Media*, 178-185, (2010).

Twitter Yardım Merkezi, "Retweetleme [online]", (01/07/2019), https://help.twitter.com/tr/using-twitter/how-to-retweet, (2019b).

Twitter Yardım Merkezi, "Takip etme ile ilgili sık sorulan sorular [online], (01/07/2019), https://help.twitter.com/tr/using-twitter/following-faqs, (2019a).

Uzunkaya, C., Ensari, T. and Kavurucu, Y., "Hadoop ecosystem and its analysis on tweets", Procedia- Social and Behavioral Sciences, 195, 1890–1897, (2015).

Vergeer, M., "Twitter and political campaigning", Sociology Compass, 9(9), 745-760, (2015).

Vosoughi, S., "Automatic detection and verification of rumors on Twitter", Doctoral dissertation, *Massachusetts Institute of Technology,* Massachusetts, USA, (2015).

Vural, Z., *Bilgi İletişim Teknolojileri ve Yansımaları* Ankara: Nobel Yayınları, (2006).

Web-Sadi-Seker, "K Fold Cross Validation (K Katlamalı Çarpraz Doğrulama) [online]", (01/05/2019), http://bilgisayarkavramlari.sadievrenseker.com/2013/03/31/k-fold-cross-validation-k-katlamali-carpraz-dogrulama/ , (2013)

Web-Bayes, "Makine Öğrenmesi 4.Hafta [online]", (01/05/2019), http://bmb.cu.edu.tr/uorhan/DersNotu/Ders04.pdf , (2019)

Wegrzyn-Wolska, K. and Bougueroua, L. , "Tweets mining for French presidential election", *Proceedings of the 4th International Conference on Computational Aspects of Social Networks,* Sao Carlos, Brezilya, 138-143, (2012).

Wei, R. and Xu, L.Z., "New media and politics: a synopsis of theories, issues, and research", *Oxford Research Encyclopedia of Communication*, doi: 10.1093/acrefore/ 9780190228613.013.104, (2019).

Weinberg, T., *The New Community Rules: Marketing on the Social Rules*, Sebastapol, California, USA: O'Reilly Media, (2009).

Xue D, Wu L, Hong Z, Guo, S., Gao, L., Wu, Z., Zhong, X., Sun, J., " Deep learning-based personality recognition from text posts of online social networks", *Applied Intelligence*, 48(11), 4232–4246, (2018).

Yılmaz, B., "Toplumsal iletişim ve kütüphane", *Hacettepe Üniversitesi Edebiyat Fakültesi Dergisi*, 20(2), 11-29, (2003).

Zeng, D., Chen, H., Lusch, R. and Li, S-H. , "Social media analytics and intelligence", IEEE Intelligent Systems, 25(6), 13-16, (2010).

7. ÖZGEÇMİŞ

Adı Soyadı	:	Vasfi TATAROĞLU
Doğum Yeri ve Tarihi	:	Denizli - 1992
Lisans Üniversite	:	Pamukkale Üniversitesi
Elektronik posta	:	vasfi@vasfi.me
İletişim Adresi	:	Çamlaraltı Mah. Hüseyin Yılmaz Caddesi Pamukkale Teknokent D-Blok Z01 Alternet Yazılım Ltd. Şti.