BILGISAYAR BILIMLERI VE TEKNOLOJILERI DERGISI
JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGIES
BIBTED JCST
http://dergipark.org.tr/tr/pub/bibted

Araştırma Makalesi

Sosyal Medya Kullanıcılarının Cumhur İttifakı Hakkındaki Görüşlerinin Makine Öğrenmesi Teknikleri ile Sınıflandırılması

Yaşar Safalı[*1]

[1] *Mersin Üniversitesi, Bilgisayar Mühendisliği Bölümü, Mersin, Türkiye*

Anahtar Kelimeler:
Cumhur İttifakı, Sosyal medya analizi, Metin Madenciliği, Kappa testi, Makine öğrenmesi

ÖZ

Sosyal ağların kullanım oranlarının artması ile sosyal medya kullanıcıları tarafından oluşturulan verilerin sayısı günden güne artmaktadır. Bu veriler ile duygu analizi, kişilik tespiti vb. akademik çalışmalar yapılabilmektedir. Bu çalışmada popüler sosyal ağlardan olan Facebook, Twitter, Instagram ve Linkedin kullanıcılarının "Cumhur İttifakı" hakkındaki görüşleri sınıflandırılmıştır. Bu amaçla, Facebook, Twitter, Instagram ve Linkedin kullanıcılarının Cumhur ittifakı hakkında paylaştıkları olumlu ve olumsuz görüşlere göre veri kümesi oluşturulmuştur. Bu veri kümesi üzerinde terim frekansı yöntemi uygulanarak öz nitelik çıkarımı yapılmıştır. Öz nitelik çıkarımı yapıldıktan sonra verilere sınıf etiketleri eklenerek eğitim kümesi oluşturulmuştur. Sınıf etiketine sahip eğitim kümesi denetimli makine öğrenmesi yöntemine uygun hale getirilmiştir. Oluşturulan eğitim kümesi üzerine makine öğrenmesi algoritmaları kullanılarak dört farklı model inşa edilmiştir. K En Yakın Komşu, Karar Ağacı, Sıralı Minimum Optimizasyon ve Bayes Sınıflandırma algoritmaları ile oluşturulan modeller kullanıcıları Cumhur İttifakı'nı destekleyen ve desteklemeyen kullanıcılar olarak sınıflandırmıştır. Oluşturulan dört farklı modelin doğruluk oranları hesaplanırken Kappa istatistik testi kullanılmıştır. Yapılan test sonucunda K En Yakın Komşu sınıflandırma algoritmasının kullanıldığı model %97 oranında başarı vermiştir. Sosyal medya kullanıcılarının Cumhur İttifakı hakkındaki görüşleri başarılı bir şekilde sınıflandırılmıştır.

Classification of Social Media Users' Opinions on the Cumhur Alliance with Machine Learning Techniques

Keywords:
Cumhur Alliance, Social media analysis, Text Mining, Kappa test, Machine Learning

ABSTRACT

With the increase in the usage of social networks, the number of data generated by social media users is increasing day by day. With these data, emotion analysis, personality detection, etc. academic studies can be done. In this study, the opinions of Facebook, Twitter, Instagram and Linkedin users, which are popular social networks, about the "People's Alliance" were classified. For this purpose, a data set was created according to the positive and negative opinions shared by Facebook, Twitter, Instagram and Linkedin users about the People's alliance. The term frequency method was applied on this data set and the attribute inference was made. After the self-attribute extraction, class labels were added to the data and the training set was created. The training set with the class label has been adapted to the supervised machine learning method. Four different models were built on the created training set using machine learning algorithms. Models created with K Nearest Neighbor, Decision Tree, Ordered Minimum Optimization and Bayes Classification algorithms classified users as those who support and do not support the Alliance of the People. Kappa statistical test was used to calculate the accuracy rates of four different models. As a result of the test, the model using the K Nearest Neighbor classification algorithm has achieved 97% success. Social media users' views on the People's Alliance have been successfully classified.

[*]**Sorumlu Yazar**

[*](yasarsafali.01@gmail.com) ORCID ID 0000–0001–9717–9892

e-ISSN: 2717-8579
Geliş Tarihi: 07/04/2020; Kabul Tarihi: 07/10/2020

1. GİRİŞ

Ülkemizde internet kullanım oranı her geçen yıl giderek artmaktadır. 2019 yılında Türkiye İstatistik Kurumu (TÜİK) tarafından yayınlanan Hanehalkı Bilişim Teknolojileri Kullanım Araştırması'na göre internetin Türkiye'deki kullanım oranı %75'dir. ("URL-1")

İnternet kullanım oranının artması ile sosyal ağların kullanım oranı da giderek artmıştır. Kullanıcılar günün önemli bir kısmını sosyal ağlarda geçirmektedir. Türkiye'de 58.000.000 kişi sosyal ağları aktif olarak kullanmaktadır. ("URL-2") Bu çoğunlukta kullanıcının paylaştığı görüşler sosyal ağlarda büyük miktarda veri oluşturmaktadır. Bu veriler aracılığıyla kullanıcıların duygu analizi, kullanıcılara ait siyasi görüşlerin çıkarılması gibi çeşitli analizler yapılabilmektedir. Ülkemizde sosyal ağlar etkin olarak siyasi görüşlerin, siyasi veya güncel olaylar hakkındaki fikirlerin paylaşıldığı alanlardır. Bu çalışmada, kullanıcıların Cumhur İttifakı hakkındaki görüşleri sınıflandırılmıştır.

Sosyal medya kullanıcılarının Cumhur İttifakı hakkındaki görüşlerinin sınıflandırılması işlemi temel olarak veri toplama, veri ön işleme, öz nitelik çıkarımı, modellerin oluşturulması ve deneysel sonuçlar olmak üzere 5 adımdan oluşmaktadır.

Veri kümesi, sosyal medya kullanıcılarının Cumhur İttifakı hakkındaki görüşlerinden oluşmaktadır. Sosyal ağlarda herkese açık profillerden yararlanılmıştır. Veri kümesi Cumhur İttifakını destekleyen ve desteklemeyen her platformda 500'er adet toplamda 4.000 kullanıcıdan oluşmaktadır. Her kullanıcı için 200 adet olmak üzere toplamda 800.000 paylaşım verisi kullanılmıştır.

Sosyal ağlardaki kullanıcı mesajları içerisinde sembol ve özel karakterler barındığından üzerinde çalışma yapmayı zorlaştırmakta, veriler üzerinde gürültü ve bozulmalara sebep olmaktadır. Verilerin çalışmaya uygun hale getirilebilmesi için üzerinde yer alan gürültülerin ve bozulmaların giderilmesi gerekmektedir. Bozulmaları ve gürültülü verileri temizlemek için veri kümesi üzerinde doğal dil işleme yöntemleri uygulanır (Akgül ve ark., 2016). Kullanılan doğal dil işleme yöntemleri sayesinde gereksiz veriler temizlenir, bozulmuş veriler onarılır. Veriler üzerinde işlem yapmaya hazır hale getirilir.

Veri kümesi üzerinde iyileştirme işlemi yapıldıktan sonra "terim frekansı" (TF) yöntemiyle öz nitelik çıkarımı yapılır. Terim frekansı yöntemi uygulanırken Cumhur İttifakı'nı destekleyen ve desteklemeyen kullanıcıların paylaşımlarından elde edilen terimler, "terimler listesi" olarak tanımlanır. Bu listeler Cumhur ittifakını destekleyen ve desteklemeyen paylaşımlara ait anahtar kelimelerden oluşmaktadır. Bu anahtar kelimeler seçilirken kullanıcılara ait olumlu ve olumsuz verilerden tekrar değerleri en yüksek 200 adet kelime referans alınmıştır. Anahtar kelimeler iki sınıf içinde eşit sayıda 100'er adet seçilmiştir. Terimler listesi oluşturulduktan sonra veri seti bilgisayarın üzerinde işlem yapabileceği sayısal değerlere dönüştürülmüştür (Seker, 2014).

Öz nitelik çıkarımı yapıldıktan sonra verilere sınıf etiketleri eklenerek eğitim kümesi oluşturulmuştur. Eğitim kümesine sınıf etiketi eklenerek bilgisayara verilerin nasıl sınıflandırılacağı öğretilmek istenmiştir. Veri madenciliği yöntemlerinde sıkça başvurulan bu duruma supervised learning (denetimli öğrenme) denir (Çetin ve Amasyali, 2015). Oluşturulan eğitim kümesi üzerine makine öğrenme~~si~~ algoritmaları kullanılarak dört farklı model inşa edilmiştir. K En Yakın Komşu, Karar Ağacı, Sıralı Minimum Optimizasyon ve Bayes Sınıflandırma algoritmaları ile oluşturulan modeller vasıtasıyla kullanıcılar, Cumhur İttifakı'nı destekleyenler ve desteklemeyenler olarak sınıflandırılmıştır. Oluşturulan dört farklı modelin doğruluk oranları hesaplanırken Kappa istatistik testi kullanılmıştır. Kappa istatistik testi sonuçlarına göre K En Yakın Komşu (KNN) algoritmasının kullanıldığı model %97 oranında başarı vererek en başarılı model olmuştur.

Sosyal ağlarda yer alan kullanıcıların Cumhur İttifakı hakkındaki görüşleri olumlu ve olumsuz sınıflandırılarak çalışma tamamlanmıştır. Sosyal ağ kullanıcılarının paylaşımları referans alınarak yapılan bu çalışmada kullanıcılar, Cumhur İttifakı hakkında herhangi bir paylaşım yapmamış veya belli bir eşik değerin altında paylaşım yapmışlarsa bu kullanıcılar nötr yani tarafsız olarak değerlendirilmiştir.

Çalışmanın ilerleyen bölümlerinde işleyiş şu şekilde düzenlenmiştir. İkinci bölümde Sosyal ağlar üzerinde metin madenciliği, veri madenciliği teknikleri kullanılarak yapılan çalışmalara, üçüncü bölümde veri toplanmasına, dördüncü bölümde toplanan~~a~~ veriler üzerinde temizleme çalışmalarının yapıldığı ön işleme adımlarına yer verilmiştir. Beşinci bölümde ön işleme adımından geçirilen veri seti üzerinde öz nitelik çıkarılması adımına, altıncı bölümde ise modellerin inşa edilip makine öğrenmesi algoritmaları ile sınıflandırıldığı sınıflandırma ve deneysel sonuçlar adımına yer verilmiştir. Çalışmanın deneysel sonuçlarının tartışıldığı sonuç adımına ise yedinci bölümde değinilmiştir.

2. LİTERATÜR TARAMASI

Twitter erişilmesi kolay ve büyük verileri bünyesinde barındırdığı için sosyal medya madenciliğinde, veri madenciliği uygulamalarında, metin madenciliği çalışmalarında ve doğal dil işleme gibi alanlarda yapılan çalışmalarda tercih sebebi olmuştur. Akgül ve arkadaşları tarafından yapılan çalışmada Twitter verileri üzerinde duygu analizi yapılmıştır. Bu makalede veriler olumlu, olumsuz ve nötr sınıf etiketlerine göre el yordamıyla elde

edilmiştir. Veri seti üzerinde sözlük tabanlı model ve n-gram modeli uygulanarak sınıflandırma işlemi gerçekleştirilmiştir. Modellerin başarı oranları incelendiğinde n-gram modeli daha başarılı bir sonuç vermiştir (Akgül ve ark., 2016).

Nizam ve Akın tarafından yapılan çalışmada ise makine öğrenmesi yöntemlerinden denetimli öğrenme yöntemi kullanılarak Twitter üzerinde dengeli ve dengesiz veri kümelerinin performanslarının karşılaştırılması yapılmıştır. Twitter üzerinde bir gıda firmasına ait ürünlere yapılan yorumlardan veri kümesi oluşturulmuştur. Bu veri kümesi üzerinde yorumlar pozitif, negatif ve nötr olarak sınıflara ayrılmıştır. Bu sınıflara ait veri dağılımının Weka kütüphanesinde yer alan sınıflandırma algoritmalarına etkisi incelenmiştir (Nizam ve Akın, 2014).

Türkmen ve Cemgil tarafından yapılan çalışmada 2013 yılında başlayan Gezi Parkı olaylarına Twitter kullanıcılarının destek verip vermediği incelenmiştir. Bu çalışmada destek verenler gösteri yanlısı, destek vermeyenler gösteri karşıtı ve tarafsız olan mesajlar nötral olarak değerlendirilmiştir. 1351 tweet üzerinden el ile oluşturulan veri kümesi üzerinde özellik çıkarımı yapılırken Ki-Kare İstatistik metodu kullanılmıştır. Özelik çıkarımı yapıldıktan sonra SVM ve RO sınıflayıcısı ile sınıflandırma işlemi gerçekleştirilmiştir (Türkmen ve Cemgil., 2014). 2011 yılında Pennacchiotti ve Popescu tarafından gerçekleştirilen çalışmada Twitter kullanıcılarının mesajlarındaki dilsel içerik bilgileri ve ağ yapıları kullanılarak makine öğrenmesi yaklaşımı ile etnik kökenleri ve siyasi yaklaşımları sınıflandırılmıştır (Pennacchiotti ve Popescu 2011). Nikfarjam ve arkadaşları ise Twitter üzerinde ilaçların gösterdiği yan etkilere yapılan hasta yorumlarını incelemişlerdir. Veri seti oluşturulurken DailyStrength ve Twitter üzerinde yer alan yorumlar kullanılmıştır. SVM yönteminin MetaMap yönteminden daha başarılı olduğu sonucu çıkarılmıştır (Nikfarjam ve ark., 2015).

Kılınç ve arkadaşları tarafından yapılan çalışmada metin madenciliği teknikleri kullanılarak akademik çalışmaların makine öğrenmesi algoritmaları ile sınıflandırılması yapılmıştır. Veri kümesi Research Gate üzerinde yer alan akademik bildirilerden oluşturulmuştur. Materials Science, Engineering ve Social Sciences & Humanities kategorilerine ait iki farklı sınıf etiketine sahip akademik bildirilerden oluşturulmuştur. Oluşturulan veri kümesi üzerinde sınıflandırma işlemi yapılırken K En Yakın Komşu algoritması kullanılarak %96,67 başarı oranı elde edilmiştir (Kılınç ve ark., 2016).

Çalış ve arkadaşları tarafından yürütlen çalışmada reklam içerikli epostaların tespiti yapılmıştır. Bu tespit işleminde metin madenciliği tekniklerinden terim frekansı ve TF-IDF yöntemleri kullanılmıştır. Veri kümesi 400 reklam içerikli 400 tanede reklam içerikli olmayan Türkçe e-postalardan oluşturulmuştur. Veri kümesi içerinde yer alan sınıf verileri eşit sayıda seçilmiştir. Daha sonra makine öğrenmesi algoritmalarından Naïve Bayes ve K En Yakın Komşu sınıflandırıcıları kullanılarak epostaların sınıflandırılması gerçekleştirilmiştir. En doğru sonucu %96,5 başarı oranı ile K En yakın Komşu algoritması vermiştir (Çalış ve ark., 2013).

Kaynar ve arkadaşları tarafından yapılan çalışmada IMDB üzerinde yer alan yorumlardan veri kümesi oluşturulmuştur. Veri kümesi 1000 adet olumlu yorumdan ve 1000 adet olumsuz yorumdan oluşturulmuştur. Oluşturulan veri kümesinin %75'i eğitim, %25'i ise test verisi olarak kullanılmıştır. Veri kümesi üzerinde TF-IDF yöntemi kullanılarak veriler makine öğrenmesi algoritmalarının çalışabileceği sayısal değerlere dönüştürülmüştür. Daha sonra makine öğrenmesi algoritmalarından Naïve Bayes, Merkez Tabanlı Sınıflayıcı, Destek Vektör Makineleri, Çok katmanlı Yapay Sinir Ağları kullanılarak veri kümesi üzerinde sınıflandırma işlemi gerçekleştirilmiştir. Kullanılan algoritmaların başarı oranları birbiri ile karşılaştırıldığında yapay sinir ağları ve destek vektör makineleri algoritmaları diğer algoritmalara göre daha başarılı olmuştur (Kaynar ve ark., 2016).

3. VERİ TOPLAMA

Veri kümesi, sosyal medya kullanıcılarının Cumhur İttifakı hakkındaki görüşlerinden oluşmaktadır. Sosyal ağlarda herkese açık profillerden yararlanılmıştır. Veri kümesi Cumhur İttifakını destekleyen ve desteklemeyen her platformda 500'er adet toplamda 4000 kullanıcıdan oluşmaktadır. Her kullanıcı için 200 adet toplamda 800000 paylaşım verisi kullanılmıştır. Veri kümesine ait bilgiler Tablo 1' de verilmiştir. Verilerin her sınıf için eşit sayıda seçilmesine özen gösterilmiştir. Bu sayede modellere ait başarı oranlarının artması hedeflenmiştir.

Tablo 1. Veri kümesinde kullanılan sosyal ağlar ve kullanıcı sayısı

Sosyal Ağ	Toplam Kullanıcı Sayısı	Cumhur İttifakını Destekleyen Kullanıcı Sayısı	Cumhur İttifakını Desteklemeyen Kullanıcı Sayısı	Paylaşım Sayısı
Facebook	1000	500	500	200000
Twitter	1000	500	500	200000
İnstagram	1000	500	500	200000
Linkedin	1000	500	500	200000

4. VERİ ÖN İŞLEME

Sosyal ağlardaki kullanıcı mesajları içerisinde sembol ve özel karakterler barındırdığından üzerinde çalışma yapmayı zorlaştırmakta, veriler üzerinde gürültü ve bozulmalara sebep olmaktadır. Verilerin çalışmaya uygun hale getirilebilmesi için üzerinde yer alan gürültülerin ve bozulmaların giderilmesi gerekmektedir.

Bozulmaları ve gürültülü verileri temizlemek için veri kümesi üzerinde doğal dil işleme yöntemleri uygulanır (Akgül ve ark., 2016). Kullanılan doğal dil işleme yöntemleri sayesinde gereksiz veriler temizlenir, bozulmuş veriler onarılır. Veriler üzerinde işlem yapmaya hazır hale getirilir. Sık kullanılan doğal dil işleme yöntemlerinden birisi de cümlenin anlamına etki etmeyen kelimelerin cümleden çıkarılmasıdır. Türkçe dili için kullanılan en önemli doğal dil işleme kütüphanelerinin başında Zemberek Kütüphanesi gelmektedir. (Müngen ve Kaya, 2018) Bu çalışmada da Türkçe doğal dil işleme kütüphanelerinden birisi olan Zemberek Kütüphanesi kullanılmıştır. Zemberek Kütüphanesi kelimeler üzerinde yapılan yazım yanlışlarını düzeltebilmesinin yanı sıra kelimeleri köklerine ayırabilmektedir. Bu kütüphane sayesinde elde edilen veri seti, çalışmada kullanılmak üzere köklerine ayrılmış ve üzerindeki bozulmalar giderilmiştir. Veri ön işleme adımı tamamlandıktan sonra veriler özellik çıkarımı için hazır hale getirilmiştir.

Veri ön işleme adımı tamamlandıktan sonra veriler öz nitelik çıkarımı için hazır hale getirilmiştir.

5. ÖZ NİTELİK ÇIKARIMI

Veri kümesi üzerinde iyileştirme işlemi yapıldıktan sonra terimler listesi oluşturulmuştur. Cumhur İttifakı'nı destekleyen ve desteklemeyen kullanıcıların paylaşımlarından elde edilen terimler terimler listesi olarak tanımlanır. Bu listeler Cumhur İttifakı'nı destekleyen ve desteklemeyen paylaşımlara ait anahtar kelimelerden oluşur. Bu anahtar kelimeler seçilirken kullanıcılara ait olumlu ve olumsuz verilerden tekrar değerleri en yüksek 200 adet kelime referans alınmıştır. Anahtar kelimeler iki sınıf içinde eşit sayıda 100'er adet seçilmiştir. Cumhur İttifakı'nı destekleyen terim listesi Tablo 2'de, Cumhur İttifakı'nı desteklemeyen terim listesi Tablo 3'de gösterilmiştir.

Tablo 2. Cumhur İttifakı destekleyen terim listesi

Sıra	Kelime	Frekans
1	Cumhur	1087
2	Erdoğan	965
3	Vatan	942
4	Başkan	911
…	…	…
200	Millet	245

Tablo 3. Cumhur İttifakı desteklemeyen terim listesi

Sıra	Kelime	Frekans
1	Muharrem	1112
2	Laik	1075
3	Atatürk	980
4	Halk	850
..	..	..
200	İnce	350

Terim frekansı yöntemi kullanılarak terimler listesi oluşturulduktan sonra veriler bilgisayarın üzerinde işlem yapabileceği sayısal verilere dönüştürülmüştür.

Öz nitelik çıkarımı yapılırken veri madenciliği yöntemlerinden olan terim frekansı yöntemi kullanılmıştır. Terim frekansı yöntemi kullanılarak oluşturulan listeler içerisinde yer alan anahtar kelimelerin sosyal ağ kullanıcılarının mesajları içerisindeki tekrar sıklığı hesaplanmıştır. Hesaplama sonucu elde edilen değer sosyal ağ kullanıcılarının mesajlarında yer alan toplam frekans (TF) sayısına oranı hesaplanmıştır. Cumhur İttifakını destekleyenler için Cdestek, desteklemeyenler için Cdesteklemeyen değerleri hesaplanmıştır. Terimler listesinde her bir sınıf için 100 terim yer aldığından her terimin değeri ayrı ayrı hesaplanarak toplanan sonuç Cdestek ve Cdesteklemeyen değerleri ayrı ayrı hesaplanmıştır.

$$C_{destek} = \sum_{1}^{100} \left(\frac{frekans(k)}{TF} \right) \quad (1)$$

$$C_{desteklemeyen} = \sum_{1}^{100} \left(\frac{frekans(k)}{TF} \right) \quad (2)$$

Öz nitelik çıkarımı yapıldıktan sonra verilere sınıf

etiketleri eklenerek eğitim kümesi oluşturulmuştur. Eğitim kümesine Cdestekleyen ve Cdesteklemeyen sınıf etiketi eklenerek bilgisayara verilerin nasıl sınıflandırılacağı öğretilmek istenmiştir. Veri madenciliği yöntemlerinde sıkça başvurulan bu duruma Supervised Learning (Denetimli Öğrenme) denir (Çetin ve Amasyali, 2015).

Cumhur İttifakı'nı destekleyen kullanıcılar için 2000 adet, desteklemeyen kullanıcılar için 2000 adet toplamda 4000 adet kullanıcıya ait paylaşımlardan öz nitelik çıkarımı yapılmıştır. Çıkarılan öz nitelikler sayesinde eğitim kümesi oluşturulmuştur. Çıkarılan öz niteliklerden %80'i eğitim kümesi için kullanılmıştır. %20'si ise test verisi olarak seçilmiştir. Eğitim kümesindeki verilere sınıf etiketi eklenmiştir. Eğitim kümesine ait veriler Tablo 4 'de gösterilmiştir.

Tablo 4. Eğitim kümesi verileri

Paylaşım	Cdestek	Cdesteklemeyen	Sınıf
1	0.368717	0.047328	Cdestekleyen
2	0.021874	0.589650	Cdesteklemeyen
3	0.928461	0.410733	Cdestekleyen
4	0.319645	0.522839	Cdesteklemeyen
…	..		
640000	0.874429	0.946351	Cdesteklemeyen

6. MODELLERİN OLUŞTURULMASI VE DENEYSEL SONUÇLAR

Oluşturulan eğitim kümesi üzerine makine öğrenmesi algoritmaları kullanılarak dört farklı model inşa edilmiştir. K En Yakın Komşu, Karar Ağacı, Sıralı Minimum Optimizasyon ve Bayes Sınıflandırma algoritmaları ile oluşturulan modeller kullanıcıları Cumhur İttifakı'nı destekleyenler ve desteklemeyenler olarak sınıflandırmıştır. Oluşturulan dört farklı modelin doğruluk oranları hesaplanırken Kappa istatistik testi kullanılmıştır. Bu test istatistik alanında gözlemciler arası uyumda kullanılabildiği gibi farklı testlerin uyumunda da kullanılmaktadır. Bu nedenle bu test sınıflandırma problemlerinde, asıl (gerçek) sonuç ile sınıflandırıcının verdiği kararın arasındaki uyumun göstergesidir (Nizam ve Akın,2014).

Kappa İstatistik Testi sınıflandırma işlemi sonucunda modelin doğruluk oranı analiz edilir. Yapılan analiz sayesinde doğru ve yanlış sınıflandırılmış veriler incelenir. Kappa değerinin sınır aralığı -1 ile 1 aralığıdır. Veriler %100 başarılı sınıflandırıldığında Kappa değeri 1 olur. Kappa değerinin bir diğer yorumlanan durum, 0 ile 1 arasında olduğu durumdur. Bu durumda veri setinin güvenilirliği söz konusudur. Kappa değerinin sıfırdan küçük olduğu durumlarda ise veri setinin güvenilirliği yoktur yorumu yapılabilir (Nizam ve Akın,2014). Kappa değeri hesaplanırken kullanılan matematiksel formül aşağıda verilmiştir. Kappa değeri *K* hesaplanırken kullanılan Po değeri asıl sonucun ve sınıflandırıcının verdiği kararın aynı sonuç olma oranı, Pc oranı ise aynı sonucu elde etmenin beklenen olasılığını verir.

$$K = \left(\frac{Po - Pc}{1 - Pc} \right) \tag{3}$$

Çalışmada Kappa değerleri hesaplanırken Weka Kütüphanesi'nden yararlanılmıştır. Weka Kütüphanesi, veri madenciliği uygulamalarında kullanılan, içerisinde birçok sınıflandırma, denetleme, bağıntı fonksiyonları, yapay sinir ağı algoritmaları ve veri ön işleme yöntemleri barındıran bir programdır. Makine öğrenme algoritmaları yardımıyla veri analizi, grafiksel değerlendirme, görsel sonuç üretme yapabilmektedir.

K En Yakın Komşu algoritması, yaygın kullanılan bir sınıflandırma algoritmasıdır. Uzaklık hesabına göre sınıflandırma yapan bir algoritmadır. Öklid uzaklığına göre hesaplama yapan bu algoritma en yakın k komşuyu incelemektedir. Tespit ettiği bütün gözlemlere küme olarak davranmaktadır. Gözlemler sonucu oluşan kümelerin birleşmesiyle kümeler oluşturulur. Oluşturulan bu kümeler arasında mesafe tespiti yapılır. Algoritma kullanılırken dikkat edilmesi gereken nokta komşuluk değerinin (k değeri) çift sayıda değil de tek sayıda seçilmesidir. K değerinin çift sayıda seçildiği durumlarda eşitlik durumu ortaya çıkabilir. Bu durumda sınıflandırma sonucunda herhangi bir sonuç alınmayabilir. Öklid mesafe tespitine göre hesaplama işlemi şu şekilde gerçekleşir.

$$mesafe(x, y) = \sqrt{\sum_{1}^{i} \left(x_i - y_i \right)^2} \tag{4}$$

Karar ağaçları yaygın olarak kullanılan bir sınıflandırma algoritmasıdır. Enformasyona dayalı olarak oluşturulan verilerin özelliklerini otomatik olarak işleme özelliğine sahip sınıflandırma algoritmasıdır (Daş ve Türkoğlu, 2014). Kurallar kök ve yapraklar arasında oluşturulur. Veri kümesindeki en ayırt edici özellik kök olarak belirilenir. Veriler yinelemeli (recursive) olarak sınıflandırılarak ideal sonuç aranır. Zayıf dallar bu sınıflandırma tekniği ile budanarak bir sınıflandırma oluşturulmaya çalışır.

Bayes Sınfılandırma (Naive Bayes) makine öğrenmesi çalışmalarında sınıflandırma amacı ile kullanılan bir algoritmadır. Yüksek performanslı, kolay uygulaması olan yöntemde sınıflandırılacak

kümelerin ve örnek verilerin sınıfı belirlenir. Naive Bayes yaklaşımında, n boyutlu bir uzayda olan y vektörü (y1,..,y2), m tane, sınıf bulunan KC (K1,…Kn) data sınıfının son olasılığını maksimum eden bir sınıf etiketi K arar (Nizam ve Akın,2014).

Sıralı Minimum Optimizasyon algoritması, fazladan bir matrise ihtiyaç duymadan, sayısal Quatratic Programming (QP) optimizasyon tekniklerini kullanmadan SVM QP problemlerini normalden yüksek bir performansta çözüme ulaştırır. Global olarak tüm kayıp değerleri yeni değerlerle değiştirir ve nominal nitelikleri ikili olanlara dönüştürür. Ayrıca tüm nitelikleri önceden tanımlanmış değerlerle normalize eder (Daş ve Türkoğlu, 2014).

Tablo 5'de kullanılan modeller ve bu modellere ait başarı oranları ifade edilmiştir. Sonuçlar incelendiğinde Kappa İstatistik Testi'ne göre en başarılı model K En Yakın Komşu algoritmasının kullanıldığı modeldir. En doğru sınıflandırmayı yapan model seçilirken doğru sınıflandırılmış verilerin oranı ve Kappa İstatistik Testi değeri referans alınmıştır. Bu sebeple çalışmada K En Yakın Komşu algoritması kullanılarak %97 oranında başarı elde edilmiştir.

Tablo 5. Modellere ait başarı oranları

Modeller	Doğru Sınıflandırılan Verilerin oranı	Yanlış Sınıflandırımış Verilerin oranı	Kappa İstatistik Testi değeri
Bayes Sınıflandırma Modeli	%95	%5	0.9523
Sıralı Minimum Optimizasyon Modeli	%89	%11	0.8993
Karar Ağacı Modeli	%92	%8	0.9129
K En Yakın Komşu Modeli	%97	%3	0.9638

Analizde kullanılan sınıflandırma modellerinin başarı oranlarının yüksek olduğu görülmüştür. Veri madenciliği çalışmalarında başarı oranını etkileyen en önemli faktörlerden birisi veri kümesinin dengeli bir şekilde dağılmasıdır. Bu nedenle Veri toplamada Cumhur İttifakı'nı destekleyen ve desteklemeyen kullanıcı sayısı eşit alınmıştır.

Sosyal ağlarda yer alan kullanıcıların Cumhur İttifakı hakkındaki görüşleri olumlu ve olumsuz sınıflandırılarak çalışma tamamlanmıştır. Sosyal ağ kullanıcılarının paylaşımları referans alınarak yapılan bu çalışmada kullanıcılar Cumhur İttifakı hakkında herhangi bir paylaşım yapmamış veya belli bir eşik değerin altında paylaşım yapmışlarsa bu kullanıcılar nötr yani tarafsız olarak değerlendirilmiştir.

7. SONUÇ

Sosyal ağlar bünyesinde büyük miktarda veri barındıran platformlardır. Kullanıcıların paylaşımları toplanarak veri kümesi oluşturulmuştur. Veri kümesi oluşturulurken sosyal ağ kullanıcılarının Cumhur İttifakı hakkındaki görüşlerinden yararlanılmıştır. Oluşturulan veri kümesi ön işleme adımından geçirilerek öz nitelik çıkarımı yapılmıştır. Öz nitelik çıkarımı sonucu elde edilen sayısal değerlere sınıf etiketi eklenerek eğitim kümesi oluşturulmuştur. Eğitim kümesi oluşturulduktan sonra dört farklı sınıflandırma modeli üzerinden başarı oranları elde edilmiş ve sonuçlar karşılaştırılmıştır. Kappa İstaistik Testi ile Yapılan karşılaştırma sonucunda en başarılı sonucu veren K En Yakın Komşu sınıflandırma algoritması sosyal ağ kullanıcılarının Cumhur İttifakı hakkındaki görüşlerini %97 başarı oranı ile sınıflandırmıştır.

KAYNAKÇA

Akgül, E. S., Ertano, C., & Diri, B. (2016). Twitter verileri ile duygu analizi. Pamukkale University Journal of Engineering Sciences, 22(2).

ÇALIŞ, K., GAZDAĞI, O., & YILDIZ, O. (2013). Reklam İçerikli Epostaların Metin Madenciliği Yöntemleri ile Otomatik Tespiti. International Journal Of Informatics Technologies, 6(1), 1-7.

Çetin, M., & Amasyalı, M. F. (2013, April). Supervised and traditional term weighting methods for sentiment analysis. In 2013 21st Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE...

Daş, B., & Türkoğlu, İ. (2014). DNA dizilimlerinin sınıflandırılmasında karar ağacı algoritmalarının karşılaştırılması. Elektrik-Elektronik-Bilgisayar ve Biyomedikal Mühendisliği Sempozyumu (ELECO 2014), 381-383.

Kaynar, O., Yıldız, M., Görmez, Y., & Albayrak, A. (2016). Makine öğrenmesi yöntemleri ile Duygu Analizi. In International Artificial Intelligence and Data Processing Symposium (IDAP'16) (pp. 17-18).

KILINÇ, D., BORANDAĞ, E., YÜCALAR, F., TUNALI, V., ŞİMŞEK, M., & ÖZÇİFT, A. (2016). KNN algoritması ve r dili ile metin madenciliği kullanılarak bilimsel makale tasnifi.

Müngen, A. A., & Kaya, M. (2018). Extracting abstract and keywords from context for academic articles. Social Network Analysis and Mining, 8(1), 45.

Nikfarjam, A., Sarker, A., O'connor, K., Ginn, R., & Gonzalez, G. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association, 22(3), 671-681.

Nizam, H., & Akın, S. S. (2014). Sosyal medyada makine öğrenmesi ile duygu analizinde dengeli ve dengesiz veri setlerinin performanslarının karşılaştırılması. XIX. Türkiye'de İnternet Konferansı.

Pennacchiotti, M., & Popescu, A. M. (2011, July). A machine learning approach to twitter user classification. In Fifth international AAAI conference on weblogs and social media.

Seker, S. E. (2014). Sosyal Ağlarda Akan Veri Madenciliği. YBS Ansiklopedi, 1(3), 21-25.

Türkmen, A. C., & Cemgil, A. T. (2014, April). Political interest and tendency prediction from microblog data. In 2014 22nd Signal Processing and Communications Applications Conference (SIU) (pp. 1327-1330). IEEE.

URL-1:
http://www.tuik.gov.tr/PreHaberBultenleri.do?id=33679
[Erişim Tarihi: 25.08.2019]

URL-2:
https://wearesocial.com/uk/blog/2019/01/digital-in-2019-global-internet-use-accelerates
[Erişim Tarihi: 31.01.2019]