T.C.
NECMETTİN ERBAKAN ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
YÖNETİM BİLİŞİM SİSTEMLERİ ANABİLİM DALI
YÖNETİM BİLİŞİM SİSTEMLERİ BİLİM DALI

TWİTTER VERİLERİ İLE MAKİNE ÖĞRENMESİ KULLANILARAK DUYGU ANALİZİ: TORKU ÖRNEĞİ

MUHAMMED ALİ BAHAR

YÜKSEK LİSANS

DANIŞMAN:
DR. ÖĞR. ÜYESİ HASAN ALİ AKYÜREK

KONYA-2023

T.C.

NECMETTİN ERBAKAN ÜNİVERSİTESİ

Sosyal Bilimler Enstitüsü

Bilimsel Etik Sayfası

Öğrencinin	Adı Soyadı	MUHAMMED ALİ BAHAR		
	Numarası	19081031031		
	Ana Bilim / Bilim Dalı	YÖNETİM BİLİŞİM SİSTEMLERİ / YÖNETİM BİLİŞİM SİSTEMLERİ		
	Programı	Tezli Yüksek Lisans	X	
		Doktora		
	Tezin Adı	TWİTTER VERİLERİ İLE MAKİNE ÖĞRENMESİ KULLANILARAK DUYGU ANALİZİ: TORKU ÖRNEĞİ		

BİLİMSEL ETİK SAYFASI

Bu tezin hazırlanmasında bilimsel etiğe ve akademik kurallara özenle riayet edildiğini, tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada başkalarının eserlerinden yararlanılması durumunda bilimsel kurallara uygun olarak atıf yapıldığını bildiririm.

Muhammed Ali BAHAR

T.C.

NECMETTİN ERBAKAN ÜNİVERSİTESİ

Sosyal Bilimler Enstitüsü

ÖZET

Öğrencinin	Adı Soyadı	Muhammed Ali BAHAR		
	Numarası	19081031031		
	Ana Bilim / Bilim Dalı	Yönetim Bilişim Sistemleri / Yönetim Bilişim Sistemleri		
	Programı	Tezli Yüksek Lisans	X	
		Doktora		
	Tez Danışmanı	Dr. Öğr. Üyesi Hasan Ali AKYÜREK		
	Tezin Adı	TWİTTER VERİLERİ İLE MAKİNE ÖĞRENMESİ KULLANILARAK DUYGU ANALİZİ: TORKU ÖRNEĞİ		

Teknolojinin gelişmesiyle birlikte insanlar duygu ve düşüncelerini sosyal medya platformları üzerinden paylaşmaya başlamışlardır. Bu paylaşımlarda gündemde olan konular, firmalar ve ürünler ile ilgili görüşler, eleştiriler ve öneriler yer almaktadır. Bu paylaşımlar sayesinde pek çok alanda incelenebilir büyük bir veri kümesi oluşmaktadır. Bu veri kümesi makine öğrenmesi ve modelleri kullanılarak duygu analizi yapılmasını kolaylaştırmaktadır. Makine öğrenmesi modelleri ile bu veri kümeleri işlenerek herhangi bir konu veya başlık ile ilgili görüşlerin ağırlıkları olumlu-olumsuz şeklinde sınıflandırılabilmektedir.

Bu çalışmada sosyal medya platformlarından Twitter üzerinde "Torku" kelimesi ile ilgili yapılan paylaşımlar kullanılarak kullanıcıların duygu analizleri TextBlob ve VaderSentiment kütüphaneleri kullanılarak etiketlendirilip 4 farklı makine öğrenmesi modeli eğitilerek duygu ağırlıkları belirlenmiştir. Sonrasında bu modeller karşılaştırılarak belirlenen veri seti için en iyi model tespit edilmeye çalışılmıştır. Bu sayede kullanıcıların "Torku" kelimesi ile ilgili genel görüşlerinin oranları çıkartılmıştır.

Anahtar Kelimeler: Makine Öğrenmesi, Duygu Analizi, Torku Kelimesi Duygu Analizi, Twitter

T.C.

NECMETTİN ERBAKAN ÜNİVERSİTESİ

Sosyal Bilimler Enstitüsü

ABSTRACT

Author's	Name and Surname	Muhammed Ali BAHAR		
	Student Number	19081031031		
	Department	Management Information Systems / Management Information Systems		
	Study Programme	Master's Degree (M.A.)	X	
		Doctoral Degree (Ph.D.)		
	Supervisor	Asst. Prof. Hasan Ali AKYÜREK		
	Title of the Thesis/Dissertation	SENTIMENT ANALYSIS USING MACHINE LEARNING WITH TWITTER DATA: CASE OF TORKU		

With the development of technology, people have started to share their feelings and thoughts on social media platforms. These posts include opinions, criticisms and suggestions on current issues, companies and products. Thanks to these posts, we have a large dataset that can be analyzed in many areas. This dataset facilitates sentiment analysis using machine learning and models. By processing these datasets with machine learning models, the weights of opinions on any subject or topic can be classified as positive-negative-neutral.

In this study, using the posts about the word "Torku" on Twitter, one of the social media platforms, the emotional analyzes of the users were tagged using the TextBlob and VaderSentiment libraries, and the emotion weights were determined by training 4 different machine learning models. Afterwards, these models were compared and the best model was tried to be determined for the determined data set. In this way, the ratios of the general opinions of the users about the word "Torku" were determined.

Keywords: Machine Learning, Sentiment Analysis, Torku Word Sentiment Analysis, Twitter

İÇİNDEKİLER

ÜÇÜNCÜ BÖLÜM

MATERYAL VE YÖNTEM

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

FORMÜLLER LİSTESİ

KISALTMALAR LİSTESİ

ML	Makine Öğrenmesi (Machine Learning)
DA	Duygu Analizi (Sentiment Analysis)
GPU	Grafik İşlemci Birimi (Graphics Processing Unit)
TP	Gerçek Pozitif (True Positive)
FP	Yanlış Pozitif (False Positive)
TN	Gerçek Negatif (True Negative)
FN	Yanlış Negatif(False Negative)
RBF	Radyal Tabanlı İşlevler (Radial Basis Function)

ÖNSÖZ

Bu çalışmada sosyal medya platformu twitter üzerinden içerisinde “Torku” kelimesi geçen tweetler toplanarak bu tweetlerin duygu analizleri yapılmıştır.

Çalışmanın yapılmasında ve planlanmasında ilgi ve desteğini benden esirgemeyen, gerekli yönlendirmeleri sağlayan ve çalışmanın takipçisi olan değerli hocam Sn. Dr. Öğr. Üyesi Hasan Ali AKYÜREK’e ve çalışmam sırasında sürekli olarak beni destekleyen değerli aileme sonsuz teşekkürlerimi sunarım.

Muhammed Ali BAHAR
KONYA-2023

GİRİŞ

Teknolojinin gelişmesiyle birlikte insanlar duygu ve düşüncelerini sosyal medya platformları üzerinden paylaşmaya başlamışlardır. Bu paylaşımlarda gündemde olan konular, firmalar ve ürünler ile ilgili görüşler, eleştiriler ve öneriler yer almaktadır. Bu paylaşımlar sayesinde pek çok alanda incelenebilir büyük bir veri kümesi oluşmaktadır. Bu veri kümesi ML (Machine Learning) ve modelleri kullanılarak DA (Duygu Analizi) yapılmasını kolaylaştırmaktadır. Makine öğrenmesi modelleri ile bu veri kümeleri işlenerek herhangi bir konu veya başlık ile ilgili görüşlerin ağırlıkları olumlu-olumsuz-nötr şeklinde sınıflandırılabilmektedir.

Dijital çağda insanlar günlük yaşamlarında pek çok metinle karşılaşmaktadır. Bu metinler sosyal medya gönderilerini, e-postaları, ürün incelemelerini, müşteri hizmetleri konuşmalarını ve daha fazlasını içerir. Bu metinleri incelemek, insanların duyguları, düşünceleri ve davranışları hakkında değerli bilgiler sağlayabilir. Bu nedenle DA, doğal dil işleme ve ML tekniklerini kullanarak metinlerdeki duygusal ifadeleri belirlemeyi amaçlayan bir araştırma alanıdır.

Makine öğrenimi, doğal dil işlemede çok önemli bir rol oynar. Duygu analizi de bu alandaki en popüler uygulamalardan biridir. Duygu analizi, metni otomatik olarak tarayarak içerdiği duygusal içeriği tahmin etmeyi amaçlar. Bu teknik birçok alanda kullanılabilmektedir. Örneğin, müşteri memnuniyeti analizi, ürün inceleme puanlaması, sosyal medya paylaşım analizi ve daha fazlası için kullanılabilmektedir. Duygu analizi, insanların büyük miktarda veriyi hızlı ve doğru bir şekilde analiz etmesine yardımcı olmaktadır. Duygu analizi, doğal dil işlemede en popüler uygulamalardan biridir. Bu teknik, içerdikleri duygusal içeriği tahmin etmek için metinleri otomatik olarak taramayı amaçlar (Pang & Lee, 2008). Makine öğreniminde DA yapmak, doğal dil işleme problemlerini çözmek için çeşitli teknikler gerektirir. Örneğin, kelime gömme, sinir ağı modelleri ve destek vektör makineleri gibi teknikler kullanılabilmektedir. Bu teknikler, makine öğrenimi algoritmalarında kullanılmak üzere metin verilerini sayısal verilere dönüştürür. Makine öğrenimi ile DA yapmak, doğal dil işleme problemini çözmek için birçok farklı teknik içerir. Kelime gömme yöntemleri, sinir ağı modelleri ve destek vektör makineleri gibi teknikler

kullanılabilmektedir (Kim, 2014). Duygu analizi hem işletmeler hem de bireyler için oldukça faydalıdır. İşletmeler, müşterilerinin ürünleri ve hizmetleri hakkında ne düşündüğünü anlamak için duygu analizini kullanabilir. Bu alanda yapılan çalışmalara bir örnek, DA için Twitter verilerini kullanan "Twitter Sentiment Analysis: A Machine Learning Approach" makalesidir (Pak & Paroubek, 2010). Makalede, duygu analizinde kelime gömme yöntemlerinin ve sinir ağı modellerinin etkili olduğu belirtilmektedir. Makine öğrenimi ile DA, işletmelerin ve bireylerin büyük miktarda veriyi hızlı ve doğru bir şekilde analiz etmesine yardımcı olmaktadır (Cambria & White, 2014). Bu alandaki çalışmaların artmasıyla birlikte DA teknikleri ve algoritmaları da gelişmektedir. Bireyler ise sosyal medya paylaşımlarının duygusal içeriklerini analiz ederek anlayışlarını derinleştirebilirler. Makine öğrenimi DA, doğal dil işlemedeki en önemli uygulamalardan biridir. Bu teknik, şirketlerin ve bireylerin büyük miktarda veriyi hızlı ve doğru bir şekilde analiz etmesine yardımcı olmaktadır.

BİRİNCİ BÖLÜM

GİRİŞ VE KAVRAMLAR

1.1. Duygu Analizi

Metin, görüntü, ses gibi verilerin içindeki duygusal içeriği belirlemek için kullanılan bir yapay zekâ tekniği olan DA günümüzde birçok alanda kullanılmaktadır (Cambria, Schuller, Xia ve Havasi, 2013; Kim & Yoon, 2018; Liu, 2012; Pan ve Lee, 2008). Duygu analizi, müşteri hizmetleri, pazarlama, sosyal medya, sağlık, eğitim gibi birçok alanda kullanılmakta ve müşteri memnuniyeti, marka bilinirliği, pazarlama stratejileri, duygusal bozuklukların tespiti, öğrenme süreçlerinin iyileştirilmesi gibi birçok faktörden etkilenmektedir.

Duygu analizi, doğal dil işleme alanında çok önemli bir konudur. Bu teknik, metni otomatik olarak taramayı ve içerdiği duygusal içeriği tahmin etmeyi amaçlar. İnsan duyarlılığı analizi, büyük veri kümelerine kıyasla oldukça zahmetli ve zaman alıcıdır. Bu nedenle, makine öğrenimi DA, büyük veri kümelerinin hızlı ve doğru analizi için çok önemlidir.

Duygu analizi teknikleri, metin içindeki duygusal içeriği belirleyerek olumlu, olumsuz veya nötr duyguları belirlemek için kullanılmaktadır (Liu, 2012). Müşteri geri bildirimlerinden, sosyal medya gönderilerinden, haberlerden ve diğer veri kaynaklarından gelen duygusal içeriği analiz ederek insanların ne söylediğinin ve hissettiğinin anlaşılmasına yardımcı olmaktadır.

Sosyal medyayı analiz etmek için DA teknikleri de kullanılmaktadır. Kullanıcılar sosyal medya platformlarında duygusal içerikler paylaşılmaktadır. Kullanıcıların ne hakkında konuştuğunu ve nasıl hissettiklerini anlamak amacıyla bu gönderilerin duygusal içeriğini analiz etmek için DA teknikleri kullanılmaktadır (Cambria ve diğerleri, 2013). Bu şekilde işletmeler, marka farkındalığını yönetmek ve müşteri ilişkilerini geliştirmek için kullanabilirler.

Duygu analizi teknikleri pazarlama endüstrisinde de kullanılmaktadır. Duygu analizi teknikleri, pazarlama kampanyaları, marka algıları ve ürünlerdeki duygusal içeriği analiz etmek için kullanılmaktadır (Kim & Yoon, 2018). Bu sayede işletmeler, müşterilerinin duygusal tepkilerini anlayabilir ve pazarlama stratejilerini buna göre formüle edebilir.

Sağlık sektöründe, ruh hali analizi teknolojisi, depresyon ve kaygı gibi duygusal bozuklukları tespit etmek ve tedavi etmek için kullanılabilmektedir. Duygu analizi teknolojisi, çeşitli duygusal faktörlerin analizi yoluyla hastanın duygusal durumunu belirleyebilmekte ve uygun tedavi yöntemleri önermektedir (Liu, 2012).

Eğitimde, öğrencilerin duygusal durumlarını belirlemek ve öğrenme sürecini iyileştirmek için DA teknolojisi kullanılmaktadır. Örneğin, sınıf içi ve çevrimiçi dersler sırasında öğrenci tepkilerini analiz ederek, öğrencilerin öğrenme sürecinde ne kadar ilgili olduklarını ve ne kadar başarılı olduklarını belirleyebiliriz (Kim & Yoon, 2018). Bununla birlikte, DA teknolojisi aynı zamanda bazı zorlukları da beraberinde getirir. Duygu analizi, dilsel çeşitlilik, kelime oyunları ve iğneleme gibi faktörlerden etkilenebilir. Ayrıca DA teknolojisi, insanın duygusal ifadelerini tam olarak anlayamamaktadır. Bu nedenle DA sonuçları, insanların yargılarını ve deneyimlerini tam olarak yansıtmamaktadır (Liu, 2012).

Duygu analizi teknolojisi bu nedenle müşteri hizmetleri, pazarlama, sosyal medya analitiği, sağlık ve eğitim sektörü gibi birçok alanda kullanılan güçlü bir araçtır. Ancak DA sonuçları dilsel çeşitlilik, kelime oyunu, iğneleme gibi faktörlerden etkilendiği ve insanların duygusal ifadelerini tam olarak anlayamadıkları için ihtiyatlı yaklaşılmalıdır.

Duygu analizi birçok alanda kullanılabilmektedir. Örneğin işletmeler, müşterilerin ürün ve hizmetlerini nasıl gördüklerini anlamak için duygu analizini kullanabilir. Sosyal medya hisse senetleri, DA için başka bir iyi veri kaynağıdır. Ek olarak, insanların sağlık, eğitim ve diğer alanlarda verdiği geri bildirimleri analiz etmek için DA kullanılabilmektedir.

Özetle, DA doğal dil işlemedeki en önemli uygulamalardan biridir. DA gerçekleştirmek için makine öğrenimini kullanmak, büyük veri kümelerini hızlı ve doğru bir şekilde analiz etmek için kritik öneme sahiptir. Bu teknik, işletmelerin müşterilerinin ürün ve hizmetleri hakkındaki görüşlerini anlamasına, kişisel sosyal medya gönderilerinin duygusal içeriğini analiz etmesine ve diğer birçok alanda veri analizini kolaylaştırmasına yardımcı olmaktadır (Pang ve Lee, 2008).

1.1.1. Duygu Analizi Tarihçesi

Duygu analizi, bir metnin duygusal tonunu belirleme sürecidir. Bu, doğal dil işleme (NLP) alanında önemli bir konudur ve birçok uygulamada yararlıdır. B. Sağlık sektörü için müşteri geri bildirim analizi, sosyal medya takibi (Balahur & Turchi, 2015; Pang & Lee, 2008).

Duygu analizi üzerine ilk araştırma 1980'lerin sonunda başladı. Bu dönemde DA ağırlıklı olarak kelime bazlı yöntemlerle gerçekleştirilmiştir. Ancak bu yöntemler yetersizdir ve doğrulukları düşüktür. Bu nedenle, DA daha karmaşık teknikler kullanılarak geliştirilmiştir (Taboada ve diğerleri, 2011).

2000'lerin başında, DA için makine öğrenimi teknikleri kullanıldı. Bu teknikler, bilgisayarların doğal dili öğrenmesini ve anlamasını sağlamak için geliştirilmiştir. Bu dönemde DA yöntemleri daha sofistike hale gelmiş ve bu yöntemlerin kullanımı yaygınlaşmıştır (Pang ve Lee, 2008).

2010'lu yıllarda sosyal medya platformlarının yaygınlaşmasıyla birlikte DA önemli bir konu haline geldi. Sosyal medya platformları, kullanıcılara duygularını paylaşabilecekleri bir alan sağlamıştır. Hal böyle olunca da insanlar DA tekniklerini sosyal medya verilerine uygulamaya başlamıştır (Balahur ve Turchi, 2015).

Duygu analizi günümüzde aktif bir araştırma konusu olmaya devam etmekte ve NLP alanında önemli bir yer tutmaktadır. Duygu analizi yöntemleri, yapay zeka ve makine öğrenimi teknikleriyle birlikte daha doğru sonuçlar üretmek için sürekli olarak gelişmektedir (Taboada vd., 2011; Pang & Lee, 2008).

1.1.2. Duygu Analizi Türleri

Duygu analizi, metindeki duygusal ifadeleri belirleme ve sınıflandırma işlemidir. Bu işlem tipik olarak doğal dil işleme ve yapay zeka teknikleri kullanılarak gerçekleştirilir. Duygu analizi birçok alanda kullanılmaktadır. Bu alanlar arasında sosyal medya analitiği, müşteri hizmetleri, pazarlama, sağlık hizmetleri, bilgi güvenliği ve daha fazlası yer alır. Cambria, Schuller, Xia ve Havasi (2013) tarafından yapılan bir araştırma, duygu analizinin sosyal medya analizi, pazarlama kampanyaları, müşteri memnuniyeti, marka farkındalığı, duygusal bozuklukları tespit etme ve öğrenmeyi iyileştirme gibi birçok faktörü etkilediğini tespit edilmiştir. Olumlu, Olumsuz ve Tarafsız Duygu Analizi, metindeki ifadelerin olumlu, olumsuz veya tarafsız olarak sınıflandırılmasıdır. Bu tür bir analiz, ürün incelemeleri, müşteri geri bildirimi ve sosyal medya gönderileri gibi çeşitli veri kaynaklarıyla yapılabilir (Liu, 2012). Metinde yer alan ifadelerin ne kadar duygusal olduğunu belirlemek için duygu yoğunluğu analizinden yararlanılır. Bu tür bir analiz, metindeki ifadelerin duygusal yoğunluğunu ölçerek duygusal tepkilerin gücünü belirleyebilir (Kim & Yoon, 2018).

Duygusal yönelim analizi, bir metindeki ifadelerin duygusal yönelimini belirlemeye yardımcı olmaktadır. Bu tür bir analiz, öfke, mutluluk, üzüntü, korku ve neşe gibi belirli duyguların yanı sıra olumlu, olumsuz ve tarafsız ifadeleri tanımlayabilir (Cambria ve diğerleri, 2013). Metinde yer alan ifadelerin anlam ve duygu içeriklerini belirlemek için anlatım ölçeği analizinden yararlanılır. Bu tür bir analiz, belirli kelimelerin ne sıklıkta kullanıldığına bağlı olarak metnin duygusal tonunu ve anlamını ölçebilmektedir (Pang & Lee, 2008).

Duygu akışı analizi, metindeki duygusal ifadelerin zaman içinde nasıl değiştiğini belirlemek için kullanılmaktadır. Bu tür bir analiz, metin içindeki duygusal ifadenin zamanla değişen tonunu gözlemleyerek bir metnin zaman içindeki duygusal durumun belirlenmesini sağlar (Cambria ve diğerleri, 2013). Duygu kümesi analizi, metindeki farklı duygusal ifadeleri gruplandırmak için kullanılmaktadır. Bu tür bir analiz, benzer duyguları birleştirerek bir metnin duygusal tonunu daha iyi anlamak için kullanılabilmektedir (Liu, 2012).

Bu, metninizin duygusal içeriği hakkında daha doğru bilgiler elde etmek için farklı türde DA kullanmanıza olanak tanır. Duygu analizi, pozitif, negatif ve nötr DA, duygu yoğunluğu analizi, duygu yönelimi analizi, yüz ifadesi ölçeği analizi, duygu akışı analizi ve duygu kümesi analizi içerir. Bu tür analizler birçok alanda kullanılmakta ve müşteri memnuniyeti, marka algısı, pazarlama stratejileri, duygusal bozuklukların tespiti ve öğrenmenin iyileştirilmesi gibi birçok faktörü etkilemektedir (Cambria vd., 2013; Kim & Yoon, 2018; Liu, 2012; Pang & Lee, 2008).

1.1.3. Duygu Analizinde Kullanılan Veri Setleri

Duygu analizi, metinlerin duygusal içeriğini anlama yöntemidir (Cambria, Schuller, Xia ve Havasi, 2013). Bu yöntem, metindeki duygusal ifadeleri tanımlamak ve sınıflandırmak için doğal dil işleme ve makine öğrenimi tekniklerini kullanır. Duygu analizi birçok alanda kullanılmaktadır. Bu alanlar arasında sosyal medya analitiği, müşteri hizmetleri, pazarlama, sağlık hizmetleri, bilgi güvenliği vb. yer alır (Cambria ve diğerleri, 2013). Duygu analizi için kullanılan veri seti, bu yöntemin doğruluğunu ve etkinliğini belirlemede kritik öneme sahiptir.

A. Sosyal medya kayıtları

Sosyal medya veri kümeleri, DA için en yaygın kullanılan veri kümelerinden biridir. Bu kayıtlar Twitter, Facebook, Instagram ve diğer sosyal medya platformlarından alınan verileri içerir. Sosyal medya kayıtları, kullanıcıların görüşlerini, düşüncelerini ve duygusal tepkilerini gösterir. Bu veriler ürün incelemelerinde, müşteri geri bildirimlerinde, politik kampanyalarda vb. kullanılabilmektedir (Kim & Yoon, 2018).

B. Film ve müzik kayıtları

Filmler ve müzik veri kümeleri, DA için yaygın olarak kullanılan diğer veri kümeleridir. Bu veri kümeleri, filmlerde ve müzikte duygusal ifadeler içerir. Bu veriler film incelemelerinden, şarkı sözlerinden ve diğer sanat eserlerinden elde edilebilir. Film ve müzik veri kümeleri, genellikle güçlü duygusal ifadeler içerdikleri için ruh hali analizi için çok yararlıdır (Pang & Lee, 2008).

C. Ürün inceleme kaydı

Ürün Derecelendirmeleri veri kümesi, DA için kullanılan başka bir veri kümesidir. Bu kayıtlar, çeşitli ürünlerle ilgili müşteri yorumlarını içerir. Bu incelemeler, ürün kalitesi, performans, kullanılabilmektedirlik ve diğer faktörler hakkında duygusal ifadeler içerir. Ürün inceleme belgeleri, pazarlama stratejisi ve müşteri memnuniyeti için kritik öneme sahiptir (Liu, 2012).

D. Sağlık kayıtları

Sağlık kayıtları, DA için kullanılan başka bir veri kümesidir. Bu kayıtlar hasta yorumlarını ve sağlıklarıyla ilgili duygusal ifadeleri içerir. Bu veriler, sağlık hizmetlerinin etkinliği, hasta memnuniyeti ve sağlık hizmetlerindeki gelişmeler hakkında bilgi sağlar. Sağlık kayıtları sağlık sektörü için önemlidir (Kim & Yoon, 2018).

Duygu analizi veri seti, bu yöntemin doğruluğunu ve etkinliğini belirlemek için önemlidir. Sosyal medya veri kümeleri, film ve müzik veri kümeleri, ürün incelemeleri veri kümeleri ve sağlık veri kümeleri, DA için en sık kullanılan veri kümeleridir (Cambria vd., 2013; Kim & Yoon, 2018; Liu, 2012; Pang & Lee, 2008). Bu veri setlerini müşteri memnuniyeti, pazarlama stratejileri, sağlık hizmetleri ve daha birçok alanda kullanarak bu alanlarda iyileştirmeler yapılabilir.

1.1.4. Veri Önişleme ve Temizleme

Duygu analizi, metindeki duyguları, hisleri ve düşünceleri tanımlamaya yönelik bir makine öğrenimi uygulamasıdır. Ancak, doğru sonuçlar elde etmek için verilerin ön işlenmesi ve temizlenmesi gerekir. Bu süreç, istenmeyen bilgilerin düzenlenmesini, filtrelenmesini ve kaldırılmasını içerir.

Veri ön işleme ve temizleme, duygu analizinde önemli adımlardır. Bu adımların amacı, verileri daha anlamlı ve kullanışlı, daha az gürültülü ve daha tutarlı hale getirmektir (Liu, 2012). Veri ön işleme ve temizleme için adımlar aşağıdaki gibidir:

Veri toplama:

Veri toplama, duygu analizinin uygulanması için gerekli verilerin elde edilmesidir. Bu adım, veri kaynaklarının tanımlanmasını ve verilerin toplanıp kaydedilmesini içerir.

Veri Düzenleme:

Verileri düzenlemek, verileri daha sonra çalışmak daha kolay olacak şekilde düzenlemek anlamına gelir. Bu adım, verileri doğru bir şekilde kaydeder ve daha sonra kolayca işlenmek üzere düzenler.

Veri filtreleme:

Filtreleme adımı, verilerden gereksiz bilgileri kaldırır. Bu, verileri daha az gürültülü ve daha tutarlı hale getirir. Bu adım, verilerdeki istenmeyen kelimeler, semboller ve özel karakterler gibi gürültü üreten unsurları temizler (Pang & Lee, 2008).

Veri dönüşümü:

Bir dönüştürme adımı, verilerin başka bir biçime dönüştürülmesini içerir. Bu adım, verilerin işlenmesini kolaylaştırır ve sonuçların daha doğru olmasını sağlar. Örneğin, tüm metni küçük harfe dönüştürmek veri tutarlılığı için önemlidir (Liu, 2012).

Verileri birleştirmek:

Birleştirme adımı, farklı veri kaynaklarından gelen verilerin birleştirilmesini içerir. Bu adım, verileri daha kapsamlı ve anlamlı kılar (Turney, 2010).

Veri analizi:

Analiz adımları, verilerin istatistiksel veya görsel analizini gerçekleştirmeyi içerir. Bu adım, verileri daha anlamlı hale getirebilir. Örneğin, kelime sıklığı analizi

yapmak, yüksek duygusal yankı uyandıran kelimeleri tanımlayabilir (Pang & Lee, 2008).

Veri ön işleme ve temizleme, DA gibi makine öğrenimi uygulamaları için çok önemlidir. Çünkü bu uygulamalar, doğru sonuçlar almak için temiz verilere ihtiyaç duyar. Verilerdeki gürültü, yanlış veri etiketleri, eksik veriler ve diğer hatalar sonuçları çarpıtabilir (Turney, 2010).

Veri ön işleme ve temizlemenin önemi birçok çalışmada vurgulanmıştır (Liu, 2012; Pang ve Lee, 2008; Turney, 2010). Örneğin, Pang ve Lee (2008), doğru sonuçlar elde etmek için DA uygulamalarında veri ön işleme ve temizleme ihtiyacını vurgulamıştır. Liu (2012), DA uygulamalarında verileri düzenlemenin, temizlemenin ve filtrelemenin önemini vurgulamıştır. Turney (2010), DA uygulamalarının doğru sonuçlar için verilerin önceden işlenmesini gerektirdiği konusunda uyardı.

Bu nedenle, DA uygulamalarında doğru veri ön işleme ve temizleme, doğru ve güvenilir sonuçlar sağlar.

1.1.5. Öznitelik Çıkarımı ve Seçimi

Duygu analizi, metnin duygusal içeriğini belirlemek için kullanılan doğal bir dil işleme yöntemidir. Bu yöntem, doğru sonuçlar elde etmek için doğru bir özellik çıkarma ve seçim süreci gerektirir. Nitelikler, duygusal özellikleri yansıtan metin verilerinden çıkarılan özelliklerdir. Özellik çıkarımı ve seçimi, duygu analizinde önemli adımlardır. Bunun nedeni, doğru özelliklerin seçilmesinin analiz sonuçlarını doğrudan etkilemesidir.

Özellik çıkarma, metin verilerinin analize uygun bir biçimde düzenlenmesi ve gereksiz bilgilerin çıkarılmasıdır. Nitelikler, verilerinizdeki duygusal öğeleri yansıtan sözcükler veya tümceciklerdir. Bu özellikler pozitif, negatif veya nötr olarak sınıflandırılabilir.

Duygu analizi, metin verilerindeki duygusal ifadeleri tespit etmek için kullanılan doğal bir dil işleme tekniğidir. Bu süreçte özellik çıkarma ve seçme, metin

verilerindeki önemli özellikleri tanımlar ve bu özellikleri duygusal etkiyi belirlemek için kullanır.

Duygu analizi için özellik çıkarımı ve seçimi, çeşitli yöntemler kullanılarak gerçekleştirilebilir. Örneğin, kelime frekansı, n-gramlar, kelime duyarlılığı analizi, kelime dağılımı, kelime yoğunluğu ve kelime sıklığı gibi nitelikler kullanılabilmektedir (Manning ve diğerleri, 2008). Bu fonksiyonlar, duygu analizinde etkili sonuçlar veren yaygın olarak kullanılan fonksiyonlardan biridir. Kullanılan fonksiyonların DA için en uygun olması önemlidir. Bu nedenle, niteliklerin önemini belirlemek için çeşitli yöntemler kullanılmaktadır. Örneğin, özellik seçimi için karar ağaçları, korelasyon analizi, özellik önem değerlendirmesi ve özellik çapraz doğrulama gibi yöntemler kullanılmaktadır.

Karar ağacı teknikleri, sınıflandırma sürecindeki etkilerini ölçerek en önemli özellikleri belirler. Bu yöntem, özelliklerin sınıflandırma aşamasındaki etkisini ölçmek için bir ağaç yapısı kullanır. Ağacın kökünden başlayarak, her dal en önemli nitelikleri belirlemek için bir karar verme sürecinden geçer (Breiman ve diğerleri, 1984).

Özellikler arasındaki ilişkileri belirlemek için korelasyon analizi teknikleri kullanılmaktadır. Bu yöntem, özellikler arasındaki ilişkiyi ölçer ve en düşük korelasyona sahip özelliği seçer. Korelasyon analizi, özellikler arasındaki ilişkileri hesaplamak için Pearson korelasyon katsayısını kullanır. Pearson korelasyon katsayısı, iki özellik arasındaki doğrusal ilişkiyi ölçen istatistiksel bir tekniktir (Pearson, 1896). Özelliklerin sınıflandırma üzerindeki etkisini ölçmek için özellik önem puanlama yöntemleri kullanılmaktadır. Bu yöntem, özelliklerin sınıflandırma doğruluğu üzerindeki etkisini belirlemek ve en önemli özellikleri belirlemek için kullanılmaktadır. Bu yöntem, sınıflandırıcı modellerin performansına en çok katkıda bulunan özellikleri belirlemek için kullanılmaktadır (Breiman, 2001).

Öznitelikler arasındaki ilişkileri belirlemek için öznitelik çapraz doğrulama yöntemleri kullanılmaktadır. Bu yöntem, sınıflandırma doğruluğu üzerindeki etkilerini ölçerek en uygun özellikleri belirler. Bu yöntem, veri setini bir eğitim setine ve bir test

setine ayırır ve işlevi çapraz doğrulama yoluyla test eder. Bu yöntem, özelliklerin sınıflandırma doğruluğu üzerindeki etkisini ölçmek için kullanılmaktadır ve genellikle daha güvenilir sonuçlar elde etmek için kullanılmaktadır (Kohavi ve John, 1997).

Özetle, duygu analizinde özellik çıkarma ve seçme, verilerdeki önemli özellikleri belirlemek için kullanılan tekniklerdir. Kullanılan özelliklerin DA için en uygun olması için özellik seçimi önemlidir. Özellik seçimi için karar ağaçları, korelasyon analizi, özellik önem değerlendirmesi ve özellik çapraz doğrulama gibi yöntemler kullanılabilmektedir.

Duygu analizi için uygun özelliklerin seçilmesi ve çıkarılması, analiz sonuçlarını doğrudan etkiler. Bu nedenle, analiz için uygun öznitelikleri seçmek için öznitelik çıkarma ve seçme prosedürleri doğru bir şekilde uygulanmalıdır.

1.1.6. Duygu Analizi Uygulamaları ve Örnekleri

Duygu analizi, metnin duygusal içeriğini belirlemek için kullanılan doğal bir dil işleme yöntemidir. Duygu analizi birçok farklı alanda kullanılabilmektedir ve farklı kullanım durumlarına sahiptir. Duygu analizinin bazı kullanımları ve örnekleri şunlardır:

a) Sosyal medya analitiği:

Sosyal medya platformları, insanların duygusal tepkilerini ve düşüncelerini ifade etmek için sıklıkla kullandıkları yerlerdir. Duygu analizi, sosyal medya platformlarında paylaşılan mesajları, yorumları ve tweetleri analiz etmek için kullanılabilmektedir. Örneğin, Twitter'da marka kampanyaları ile ilgili paylaşılan tweetlerin duygusal tonunu analiz ederek kampanyaların etkisini anlaşılabilmektedir.

b) Pazar araştırması:

Duygu analizi, pazar araştırması için de kullanılabilmektedir. İşletmeler, müşterilerinin ürünleri ve hizmetleri hakkında ne düşündüğünü anlamak için duygu analizini kullanabilir. Bu sayede müşteri geri bildirimleri daha iyi anlaşılabilmekte ve işletmeler müşteri memnuniyetini artırmak için gerekli adımları atabilmektedir.

c) Sağlık Hizmetleri Bölümü:

Duygu analizi sağlık alanında da kullanılabilmektedir. Örneğin, depresyon, anksiyete veya diğer duygusal sorunları olan hastaların metinlerini analiz etmek, onların duygusal durumları hakkında bize fikir verebilir. Bu sayede hastaların tedavisinde daha iyi sonuçlar alınabilir.

d) Politik analiz:

Duygu analizi, politikacıların konuşmaları ve sosyal politik tartışmalar için de kullanılabilmektedir. Örneğin, politikacıların konuşmalarının DA, etkileri hakkında bilgi edinmeye yardımcı olabilir.

e) Müzik ve film endüstrisi:

Duygu analizi, müzik ve film endüstrilerinde de kullanılabilmektedir. Örneğin, bir filmin veya şarkının duygusal tonunun DA, izleyicilerin ve dinleyicilerin nasıl tepki verdiğine dair fikir edinmeye yardımcı olabilir. Bu uygulama örnekleri, duygu analizinin çok çeşitli alanlarda ve endüstrilerde kullanılabileceğini göstermektedir. Ayrıca duygu analizinin farklı kaynaklardan gelen verileri kullanabileceğini ve bu verileri analiz etmenin daha iyi sonuçlar verebileceğini gösterir.

1.1.7. Duygu Analizi Uygulamalarında Karşılaşılan Zorluklar

Duygu analizi, metnin duygusal içeriğini belirlemek için kullanılan doğal bir dil işleme yöntemidir. Ancak, DA uygulamaları bazı sorunlarla karşılaşabilir. Bu sorunlar, doğal dilin karmaşıklığından, farklı kültürlerdeki farklılıklardan ve dil kullanımından, metnin uzunluğundan ve diğer faktörlerden kaynaklanabilir. Duygu analizi uygulamalarında karşılaşılan sorunlar şunlardır:

a) Belirsizlik:

Doğal dilin karmaşıklığı, DA uygulamalarında belirsizliğe neden olabilir. Kelimelerin birden çok anlamı vardır ve bu da doğru DA sonuçlarının alınmasını zorlaştırır.

b) Kültürel ve dil farklılıkları:

Farklı kültürler ve dil kullanımı, DA uygulamalarında karşılaşılan diğer bir zorluktur. Örneğin, bir kelime veya deyim bir kültür tarafından olumlu, ancak başka bir kültür tarafından olumsuz olarak algılanabilir.

c) Uzunluk:

Metin uzunluğu, DA uygulamaları için de zorlayıcı olabilir. Özellikle aşağıdakiler gibi kısa mesajlar (tweetler, yorumlar vb.) için: Sosyal medya platformlarında duygu ifadeleri kısa ve öz olabilir, bu nedenle doğru DA sonuçları elde etmek zor olabilmektedir.

d) Sınıflandırma:

Duygu analizi uygulamalarında sınıflandırma da zor olabilir. Örneğin, bir kelimenin veya cümlenin duygusal tonunu belirlerken, üç sınıftan birine girer: nötr, pozitif ve negatif. Bununla birlikte, bazı durumlarda, duygusal bir ton birden fazla duygusal sınıf içerebilir.

e) Eğitim veri kalitesi:

Duygu analizi uygulamaları için kullanılan eğitim verilerinin kalitesi de sorun olabilir. Eğitim verileriniz doğru ve temsili örnekler içermelidir. Aksi takdirde hatalı sonuçlar alabilirsiniz.

Bu sorunlar, DA uygulamalarının doğru sonuçlar almasını zorlaştırabilir. Ancak, bu zorlukların üstesinden gelmek için bazı çözümler var. Örneğin, eğitim verilerinin kalitesini artırmak, farklı kültürleri ve kullanım farklılıklarını hesaba katmak ve sözcük belirsizliğini çözmek için sözcük vektörleri gibi yöntemler kullanılabilmektedir. Bu bir uzunluk sorunuysa, verilen kelimelerle ifade edilen duygusal ifadeleri (örneğin ifadeleri) özetlemek veya kullanmak gibi başka çözümler de vardır. Duygu analizi uygulamalarının sorunları vardır ancak doğru yöntem ve çözümlerle bu sorunların üstesinden gelinebilir. Bu size doğru ve güvenilir sonuçlar verecektir.

1.2. Makine Öğrenmesi

Özetle, DA doğal dil işlemedeki en önemli uygulamalardan biridir. Duygu analizi gerçekleştirmek için makine öğrenimini kullanmak, büyük veri kümelerini hızlı ve doğru bir şekilde analiz etmek için kritik öneme sahiptir. Bu teknik, işletmelerin müşterilerinin ürün ve hizmetleri hakkındaki görüşlerini anlamasına, kişisel sosyal medya gönderilerinin duygusal içeriğini analiz etmesine ve diğer birçok alanda veri analizini kolaylaştırmasına yardımcı olmaktadır (Pang & Lee, 2008).

1.2.1. Makine Öğrenmesi Tarihçesi

Yapay zekanın bir alt alanı olan makine öğrenimi, bilgisayarlara kendi kendine öğrenme yeteneği kazandırmayı amaçlar. Makine öğrenimi 1950'lerde doğdu, ancak son yıllarda büyük veriler, bulut bilgi işlem ve GPU (Graphics Processing Unit)'lardaki gelişmeler sayesinde muazzam bir ilerleme kaydetti.

İlk makine öğrenimi tekniği olan denetimli öğrenme, 1950'lerde ortaya çıktı ve Arthur Samuel tarafından IBM'de geliştirilen bir satranç programında tanıtıldı (Samuel, 1959). Denetimli öğrenmede, öğrenme, eğitim verileri aracılığıyla bir modelin parametrelerini bir algoritmaya uydurarak yapılır. 1970'lerde denetimsiz veya yarı denetimli öğrenme yöntemleri ortaya çıkmaya başladı. Bunlardan en önemlisi k-means kümeleme algoritmasıdır (Steinhaus, 1956). Makine öğrenimi, sinir ağlarının yeniden canlanması ve derin öğrenme algoritmalarının geliştirilmesiyle 1990'larda ivme kazandı (LeCun, Bengio ve Hinton, 2015). 1997'de Deep Blue, o zamanki dünya satranç şampiyonunu yendi (Campbell, Hoane & Hsu, 2002).

2000'lerin başında denetimli öğrenme algoritmalarını kullanan destek vektör makineleri (SVM'ler) popüler hale geldi (Cortes & Vapnik, 1995). SVM'ler, metin sınıflandırmasında ve nesne tanımada başarıyla kullanılmıştır. 2006 yılında Geoffrey Hinton tarafından yapılan bir araştırma, derin sinir ağlarının önemini vurguladı (Hinton ve diğerleri, 2006). Derin bir sinir ağı modeli olan evrişimli sinir ağları (CNN'ler), büyük veri kümelerinde büyük bir başarıyla kullanılmıştır (Krizhevsky, Sutskever & Hinton, 2012).

Makine öğrenimi günümüzde hemen hemen her alanda kullanılmaktadır. Daha sofistike modeller ortaya çıktıkça ve daha fazla veri türü kullanıldıkça, makine öğrenimi hayatımızın her alanında giderek daha alakalı hale gelecektir.

1.2.2. Makine Öğrenmesi Türleri

Makine öğrenimi, bilgisayarların veri analizinde kalıpları tanımasını ve bunları gelecekteki olayları tahmin etmek için kullanmasını sağlayan bir yapay zeka biçimidir. Makine öğrenimi, verileri analiz etmek, kalıpları belirlemek ve gelecekteki olayları tahmin etmek için çeşitli yöntemler kullanır (Pang & Lee, 2008).

a) Denetimli öğrenme

Denetimli öğrenme, algoritmaların önceden tanımlanmış etiketli verilerden öğrenmesine olanak tanır. Bu veriler, doğru cevaplar (etiketler) için eğitim verileri olarak kullanılmaktadır. Algoritmalar bu verileri kullanarak yeni verileri analiz eder ve doğru yanıtları tahmin eder. Denetimli öğrenme, sınıflandırma ve regresyon gibi birçok uygulamada kullanılmaktadır (Pang & Lee, 2008).

b) Denetimsiz öğrenme

Denetimsiz öğrenme, etiketlenmemiş verileri kullanarak öğrenme sürecidir. Bu verinin doğru yanıtı ve kategorisi yoktur. Algoritmalar bu verileri analiz eder ve kalıpları tanımaya çalışır. Denetimsiz öğrenme, veri kümelerindeki kalıpları ve yapıları tanımak, verilerdeki özellikleri belirlemek ve veri kümelerindeki benzer örnekleri eşleştirmek için kullanılmaktadır (Bishop, 2006).

c) Yarı denetimli öğrenme

Yarı denetimli öğrenme, hem etiketlenmiş hem de etiketlenmemiş verileri kullanan bir öğrenme sürecidir. Bu yöntem, etiketli veriler eksik olduğunda kullanılmaktadır. Etiketlenmemiş veriler, algoritmaların kalıpları tanımasına ve doğru yanıtları tahmin etmesine yardımcı olmaktadır (Zhu, Ghahramani, & Lafferty, 2003).

d) Takviyeli öğrenme

Takviyeli öğrenme, bir ortam içinde belirli görevleri gerçekleştirmeyi öğrenen algoritmalara dayanır. Bir algoritma belirli bir durumda bir eylem gerçekleştirdiğinde, sonuçlar değerlendirilir ve doğru sonuçlar için ödüllendirilir. Yanlış sonuçlar için cezalandırılacak. Bu yöntem, oyunlar ve robotik gibi uygulamalar için kullanılmaktadır (Hastie, Tibshirani & Friedman, 2009; Bishop, 2006).

e) Aktarım öğrenmesi

Transfer öğrenimi, algoritmaların daha önce öğrenilen bilgileri yeni görevler için kullanmasına izin verir. Bu yöntem, sınırlı veri kümeleri veya yeni görevler için kullanılabilmektedir. Bu yöntem, önceden öğrenilen bilgilere dayanarak daha hızlı ve daha doğru tahminler yapmak için kullanılmaktadır.

Sonuç olarak makine öğrenimi, veri analizinde kalıpları belirlemek ve gelecekteki olayları tahmin etmek için çeşitli yöntemler kullanır. Denetimli, denetimsiz, yarı denetimli, takviyeli ve aktarım öğrenmesi bu yöntemlerden bazılarıdır. Bu teknikler birçok uygulamada kullanılmakta ve yapay zekanın geliştirilmesinde önemli rol oynamaktadır (Taylor & Stone, 2009)

1.2.3. Makine Öğrenmesi Modelleri

Makine öğrenimi, büyük veri kümelerini analiz etmek için kullanılan yapay zeka tekniklerinden biridir. Makine öğrenimi modelleri, verileri işleyerek sonuç üretir ve çeşitli uygulama alanlarında kullanılmaktadır (Alpaydin, 2010).

a) Regresyon modelleri

Regresyon modelleri, bir değişken bir veya daha fazla başka değişkene bağlı olduğunda kullanılmaktadır. Bu modeller, sürekli çıktı verilerini (sıcaklık ve fiyat gibi) tahmin etmek için kullanılmaktadır. Regresyon modelleri, doğrusal regresyon, lojistik regresyon ve çoklu regresyon gibi alt modelleri içerir (Hastie, Tibshirani ve Friedman, 2009).

b) Sınıflandırma modelleri

Bir sınıflandırma modeli, bir veri kümesindeki örnekleri belirli bir sınıfa veya kategoriye atamayı amaçlar. Bu modeller önceden etiketlenmiş verileri öğrenir ve yeni verileri analiz edip sınıflandırır. Sınıflandırma modellerinin karar ağaçları, KNN (k-en yakın komşu), Naive Bayes ve SVM (destek vektör makinesi) gibi alt modelleri vardır (Bishop, 2006).

c) Kümeleme modelleri

Kümeleme modelleri, benzer özelliklere veya niteliklere sahip örnekleri birleştirerek bir veri kümesini kümelere ayırmayı amaçlar. Bu modeller, etiketlenmemiş veriler ve benzer özelliklere sahip toplu veriler üzerinde eğitilir. Kümeleme modelleri, hiyerarşik kümeleme, K-means ve DBSCAN gibi alt modelleri içerir (Murphy, 2012).

d) Derin öğrenme modelleri

Derin öğrenme modelleri, insan beynindeki sinir ağlarına benzer şekilde işlev gören yapay sinir ağlarını kullanarak eğitim verir. Bu modeller, büyük veri kümelerini işleyebilir ve yüksek doğruluk elde edebilir. Derin öğrenme modelleri, Konvolüsyonel Sinir Ağları (CNN), Tekrarlayan Sinir Ağları (RNN) ve Uzun-Kısa Süreli Bellek (LSTM) gibi alt modelleri içerir (Goodfellow, Bengio ve Courville, 2016).

e) Güçlü öğrenme modeli

Ortam içinde belirli görevleri gerçekleştirmek için güçlü öğrenme modelleri kullanılmaktadır. Bu modeller çevredeki verilerden öğrenir ve sonuçlarını optimize eder. Güçlü öğrenme modelleri, Q-learning, SARSA ve DDPG (Deep Deterministic Gradient Climbing) gibi alt modelleri içerir (Sutton & Barto, 2018).

f) Doğrusal olmayan model

Doğrusal olmayan modeller, veriler arasında doğrusal olmayan ilişkiler elde etmek için kullanılmaktadır. Bu modellerin çekirdek yöntemleri, ağaç yöntemleri, RBF'ler (radyal tabanlı işlevler) ve sinir ağları gibi alt modelleri vardır (Hastie, Tibshirani ve Friedman, 2009).

Sonuç olarak makine öğrenimi modelleri, toplanan verileri analiz ederek, kalıpları belirlemek ve gerçekleşecek olayları tahmin etmek için kullanılan çeşitli teknikleri içerir. Regresyon, sınıflandırma, kümeleme, derin öğrenme, güçlü öğrenme ve doğrusal olmayan modeller dahil olmak üzere birçok model türü vardır. Bu modeller, çeşitli uygulama alanları ve veri türleri için kullanılabilmektedir.

1.2.4. Makine Öğrenmesi için Veri Seti

Veri kümeleri, makine öğrenimi algoritmalarının üzerinde eğitildiği ve sonuçlarının doğrulandığı temel yapı taşlarıdır (Alpaydin, 2010). Bir veri kümesi genellikle bir veya daha fazla bağımsız değişken (özellikler) ve bir bağımlı değişken (hedef) içerir. Veri kümeleri, çeşitli veri türleri ve kaynaklarından gelebilir (Hastie, Tibshirani ve Friedman, 2009).

Veri kümeleri, özellikleri çıkarmak için önceden işlenebilir veya ham verilerden oluşabilir (Guyon & Elisseeff, 2003). Veri kümesi boyutu, özellikleri ve kalitesi, makine öğrenimi algoritmalarının performansını etkileyebilir (Kotsiantis, 2007). Veri kümeleri, veri madenciliği teknikleri kullanılarak oluşturulabilir (Han ve Kamber, 2006).

Veri kümeleri, kategorik veriler, sayısal veriler ve metin dahil olmak üzere çeşitli veri türlerini içerebilir (Aggarwal, 2015). Sınıflandırma problemlerinde kategorik veriler, regresyon problemlerinde sayısal veriler kullanılmaktadır. Metin verileri, dil işleme teknikleri kullanılarak analiz edilebilir (Manning, Raghavan & Schütze, 2008).

Makine öğrenimi algoritmaları, veri kümelerini analiz eder ve sonuçları optimize eder (Bishop, 2006). Veri kümeleri, makine öğrenimi algoritmalarının

doğruluğunu etkileyebilecek çeşitli veri türleri ve kaynaklarından gelebilir (Hastie ve diğerleri, 2009). Veri kümeleri, veri madenciliği teknikleri kullanılarak oluşturulabilir (Han ve Kamber, 2006).

1.2.5. Makine Öğrenmesi Kullanım Alanları

Makine öğrenimi, günümüzde birçok alanda kullanılan bir yapay zeka tekniğidir. Makine öğrenimi algoritmaları, büyük veri kümelerindeki kalıpları ve ilişkileri keşfederek sonuçlar üretir. Uygulama alanına bağlı olarak, makine öğrenimi farklı türde verileri kullanabilir ve farklı amaçlara hizmet edebilir.

Sağlık sektörü, birçok alanda makine öğrenimi teknolojisini kullanmaktan yararlanabilir. B. Hastalığı teşhis etme, tedavi rejimlerini optimize etme ve hastalığın yayılma riskini tahmin etme (Shickel ve ark., 2018). Makine öğrenimi finans sektöründe de yaygın olarak kullanılmaktadır. Bu alan, hisse senetleri, emtia fiyatları ve diğer finansal araçlardaki fiyat değişikliklerini tahmin etmek için makine öğrenimi algoritmalarını kullanır (Géron, 2019).

Makine öğrenimi ayrıca otomotiv endüstrisinde araç sürülebilirliğini ve yakıt verimliliğini iyileştirmek için kullanılmaktadır (Rajaraman & Ullman, 2011). Makine öğrenimi, müşteri davranışını analiz etmek ve pazarlama stratejilerini optimize etmek için de kullanılabilmektedir (Provost & Fawcett, 2013).

Eğitim, makine öğrenimi için başka bir uygulama alanıdır. Makine öğrenimi, öğrenci performansını analiz etmek, öğrenci performansını tahmin etmek ve öğrenci ihtiyaçlarını belirlemek için kullanılabilmektedir (Romero ve diğerleri, 2010). Ek olarak, çevrimiçi materyallerle öğrenci etkileşimlerini analiz ederek öğrenme deneyimini geliştirmek için makine öğrenimi algoritmaları kullanılabilmektedir (Koedinger & Corbett, 2006).

1.2.6. Makine Öğrenmesi Performans ve Doğruluk Ölçümleri

Makine öğrenimi algoritmalarının performansı, doğruluk ölçütleriyle belirlenir. Doğruluk ölçümleri, bir makine öğrenimi modelinin ne kadar iyi performans

gösterdiğini belirlemek için kullanılmaktadır. Doğruluk ölçümleri, modelin tahminlerinin gerçek değerlerle ne kadar iyi eşleştiğini gösterir.

Makine öğrenimi algoritmalarının performansı, çeşitli doğruluk ölçütleri kullanılarak belirlenebilir. Doğruluk, yanlış pozitifleri, yanlış negatifleri ve doğruluğu açıklayan bir ölçüdür. Spesifiklik, gerçek olumsuz sonuçları tanımlayan bir ölçüdür (Flach, 2015).

Bir başka doğruluk ölçüsü de ROC eğrisidir. Bir ROC eğrisi, bir sınıflandırma modelinin farklı eşikleri için duygu ve özgüllük ölçümlerini gösteren bir grafiktir (Fawcett, 2006). Model performansını ölçmek için ROC eğrisi altındaki alan (AUC) kullanılmaktadır.

Karışıklık matrisi, sınıflandırma problemlerinde kullanılan bir metrik olup, gerçek ve tahmin edilen sınıfların sayısal değerlerini gösteren bir tablodur (Scikit-learning, 2021). Bu matris, sınıflandırma algoritmasının performansını ölçmek için kullanılmaktadır (Chandola, 2018).

Karışıklık matrisi dört farklı değer içerir ve aşağıdaki gibi olmaktadır:

Tahmin / Gerçek	Pozitif	Negatif
Pozitif	TP	FN
Negatif	FP	TN

Tablo – 1: Karışıklık Matris Tablosu

Gerçek pozitif (TP), yanlış pozitif (FP), gerçek negatif (TN) ve yanlış negatif (FN). TP gerçek bir pozitifi (doğru sınıflandırılmış bir pozitif örneği) temsil ederken, FP yanlış bir pozitifi (yanlış sınıflandırılmış bir pozitif örneği) temsil eder. TN gerçek bir negatifi (doğru sınıflandırılmış bir negatif örneği) temsil ederken, FN yanlış bir negatifi (yanlış sınıflandırılmış bir negatif örneği) temsil eder (Scikit-learn, 2021; Chandola, 2018). Karışıklık matrisini hesaplamak için, önce bir sınıflandırma algoritması kullanılarak test verileri üzerinde tahminler yapılır. Bu tahminler gerçek

sınıflarla karşılaştırılır ve her numune için TP, FP, TN ve FN değerleri hesaplanır. Bu değerler daha sonra karışıklık matrisi tablosuna eklenir.

Örneğin, bir sınıflandırma algoritması örnekleri iki sınıfa ayırır: olumlu ve olumsuz. Karışıklık matrisi, gerçek sınıf ile tahmin sınıfı arasındaki ilişkiyi temsil eden bir tablodur. Örneğin, bir sınıflandırma algoritması 10 örnek için;

Gerçek veri sonuçları	1	1	0	1	0	1	1	0	0	0
Tahmine edilen sonuçlar	1	0	0	1	1	0	1	1	1	0

Tablo – 2: Karışıklık Matris Tablosu İçin Örnek Veri Seti

Bu durumda;

TP için iki veride de 1 olanları sayılır: 3 olarak belirlenir.

FP için gerçek veride 0 olan fakat tahmin verisinde 1 olarak işaretlenmiş veriler sayılır: 3

TN için iki veride de 0 olarak işaretlenmiş veriler sayılır: 2

FN için gerçek veride 1 olan fakat tahmin verisinde 0 olarak işaretlenmiş veriler sayılır: 2

Bu durumda karışıklık tablosu (Confusion Matrix) aşağıdaki gibi olmaktadır.

Tahmin / Gerçek	**Pozitif**	**Negatif**
Pozitif	**3**	**3**
Negatif	**2**	**2**

Tablo – 3: Örnek Verilerle Karışıklık Matrisi

Doğru model performansı hapishanelerini ve beklentileri kullanmak için kullanılan ölçüm aralığını ölçün. İşte bazı popüler performans ölçümleri (Scikit-learn, 2021; Brownlee, 2019; Géron, 2019):

Accuracy (Doğruluk): Bu, doğru sonuçlara sahip numunelerin toplam numune oranıdır. Örneğin hayal gücü 100 örneği sınıflandırıp 85 örneği doğru sınıflandırırsa accuary (doğruluk) oranı %85'tir (Géron, 2019).

$$\frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Formül – 1: Accuary Değeri Hesaplanması

Precision (Kesinlik): Pozitif sonuçlara sahip numuneler arasında gerçek pozitif testlerin oranı. Örneğin, bir sınıflandırma algoritması 100 pozitif örneği sınıflandırdı ve 80 örneği gerçekte pozitifse, precision (kesinlik) oranı %80'dir (Brownlee, 2019).

$$\frac{TP}{TP + FP}$$

Formül – 2: Precision Değeri Hesaplanması

Recall (Geri Çağırma): Gerçek pozitif numuneler arasında pozitif testlerin oranı. Örneğin, bir sınıflandırma algoritması 100 gerçek pozitif örneği sınıflandırdı ve 75 örneği pozitif olarak sınıflandırdıysa, recall (geri çağırma) oranı %75'tir (Scikit-learning, 2021).

$$\frac{TP}{TP + FN}$$

Formül – 3: Recall Değeri Hesaplanması

F1 Score (F1 Skoru): Precision (Doğruluk) ve recall (geri çağırma) oranının harmonik ortalamasıdır. Bu metrik, hem doğruluğu hem de gözetlemeyi izler ve davranışsal performanslarını ölçmek için kullanılmaktadır. Örneğin, bir sınıflandırma algoritması için precision (kesinlik) oranı %80 ve recall (geri çağırma) oranı %75 ise, f1 score (F1 skoru) 2 * (0.8 * 0.75) / (0.8 + 0.75) = 0.77'dir (Brownlee 2019).

$$2 * \frac{PRECISION * RECALL}{PRECISION + RECALL}$$

Formül – 4: F1 Score Değeri Hesaplanması

ROC eğrisi: Algoritma bunun gerçek bir pozitif (hassasiyet) olduğunu varsayar ve aşağıda çizilen bir eğri ile farklı yanlış pozitif oranlar (özgüllük 1) görüntüler. Bu istatistik, Algı Becerisi yeteneğini kullananlar tarafından kullanılmak üzere tasarlanmıştır ve ROC Dış Kıvrım Alanı (AUC) değeri (Scikit-learning, 2021) olarak ifade edilir.

Makine öğrenimi algoritmalarının performansı, k-katlı çapraz doğrulama kullanılarak da belirlenebilir. Bu yöntem, veri setini parçalara ayırır ve modeli her parça için ayrı ayrı eğitir. Bu şekilde modelin genel performansı belirlenebilir (Kohavi, 1995).

Sonuç olarak, makine öğrenimi algoritmasının performansı bir doğruluk ölçüsü kullanılarak belirlenir. Duygu, özgüllük ve ROC eğrileri gibi doğruluk ölçümlerini kullanarak model performansı ölçülebilir. Modelinizin genel performansını belirlemek için K-katlamalı çapraz doğrulamayı da kullanabilirsiniz.

1.2.7. Makine Öğrenmesi Uygulamaları ve Örnekleri

Makine öğrenimi, günümüzde birçok alanda kullanılan bir yapay zeka tekniğidir. Makine öğrenimi algoritmaları, büyük veri kümelerindeki kalıpları ve ilişkileri keşfederek sonuçlar üretir. Makine öğreniminin birçok uygulaması vardır.

Makine öğrenimi birçok alanda avantajlar sunar. B. Sağlık sektöründe kullanım için hastalık teşhisi ve tedavi planlarının optimizasyonu. Örneğin, makine öğrenimi algoritmaları, kanser teşhisi için kullanılabilecek tümör görüntü analizini gerçekleştirebilir (Wang ve diğerleri, 2016). Makine öğrenimi, hasta sağlık verilerini analiz ederek hastalık riskini tahmin etmek için de kullanılabilmektedir (Lasko ve diğerleri, 2013).

Makine öğrenimi finans sektöründe de yaygın olarak kullanılmaktadır. Hisse senetleri, emtia fiyatları ve diğer finansal araçlardaki fiyat değişikliklerini tahmin etmek için makine öğrenimi algoritmaları kullanılmaktadır (Géron, 2019). Makine öğrenimi, kredi riskini değerlendirmek için de kullanılabilmektedir.

Makine öğrenimi, otomotiv endüstrisinde araç sürülebilirliğini ve yakıt verimliliğini iyileştirmek için kullanılmaktadır (Rajaraman & Ullman, 2011). Makine öğrenimi, müşteri davranışını analiz etmek ve pazarlama stratejilerini optimize etmek için de kullanılabilmektedir (Provost & Fawcett, 2013).

Eğitim, makine öğrenimi için başka bir uygulama alanıdır. Makine öğrenimi, öğrenci performansını analiz etmek, öğrenci performansını tahmin etmek ve öğrenci ihtiyaçlarını belirlemek için kullanılabilmektedir (Romero ve diğerleri, 2010). Ek olarak, materyalleri kişiselleştirmek ve öğrenci öğrenme deneyimini iyileştirmek için makine öğrenimi algoritmaları kullanılabilmektedir (Koedinger & Corbett, 2006).

1.2.8. Makine Öğrenmesi Dezavantajları ve Zorlukları

Makine öğrenimi, günümüzde birçok alanda kullanılan bir yapay zeka tekniğidir. Bununla birlikte, makine öğrenme algoritmalarının da bazı dezavantajları ve zorlukları vardır.

Makine öğrenimi algoritmaları, büyük veri kümelerini analiz etmek için tasarlanmıştır, ancak veri kümesindeki yanlış veya eksik veriler, doğru sonuçların alınmasını zorlaştırabilir. Bu da modelin yanlış öğrenmesine ve sonuçların yanıltıcı olmasına neden olabilir. Makine öğrenimi algoritmaları bile insanın karar verme sürecini açıklayamaz. Makine öğrenimi uygulamaları bu nedenle yanlış kararlar verebilir (Shalev-Shwartz & Ben-David, 2014).

Makine öğrenimi algoritmalarının bir başka dezavantajı, fazla uydurma sorunudur. Fazla uydurma, modelin eğitim verilerine gereğinden fazla uymasına ve yeni veriler üzerinde kötü performans göstermesine neden olabilir. Fazla uydurma sorunları, model karmaşıklığı ve veri kümesi boyutu gibi faktörlerden etkilenebilir (Hastie ve diğerleri, 2009).

Makine öğrenimi algoritmalarının bir başka dezavantajı, yetersiz uyum sorunudur. Yetersiz uyum, zayıf model performansına yol açabilir çünkü veri kümesindeki örüntüler doğru bir şekilde öğrenilmemiştir (Hastie ve diğerleri, 2009).

Makine öğrenimi algoritmalarını kullanmak etik ve sosyal sorunları gündeme getirebilir. Örneğin makine öğrenimi algoritmaları cinsiyet, ırk ve yaş gibi faktörleri göz ardı ederek taraflı kararlar verebilmektedir (Buolamwini & Gebru, 2018). Ek olarak, makine öğrenimi algoritmaları, insanların gizliliğini tehlikeye atabilecek kişisel verileri analiz edebilir.

Makine öğrenimi algoritmalarının bazı dezavantajları ve zorlukları vardır. Hatalı veriler, aşırı başarı, düşük başarı, etik ve sosyal sorunlar, makine öğrenimi uygulamalarında karşılaşılan en yaygın zorluklardır. Bu nedenle, makine öğrenimi algoritmaları geliştirilirken veri kalitesi kontrolü, veri kümesi boyutu, model karmaşıklığı ve etik sorunlar gibi faktörler dikkate alınmalıdır.

İKİNCİ BÖLÜM

TEORİK ÇERÇEVE

2.1. Ön Hazırlık

2.1.1. Veri Seti

Duygu analizi veri defterleri genellikle insan tarafından etiketlenmiş veriler içerir (Baccianella, Esuli & Sebastiani, 2010). Bu koruma, bir metin belgesindeki sözcüklerin pozitif, negatif veya nötr olduğunu belirtmek için etiketler kullanır. Bu veri algılama, makine öğrenimi ve derin öğrenme teknikleri kullanılarak DA modellerini eğitmek için kullanılabilmektedir. Duygu Analizi Veri Kitabı çeşitli versiyonlarda mevcuttur. Örneğin, DA için veri oluşturmak üzere film incelemeleri, ürün incelemeleri ve sosyal medya gönderileri gibi kaynaklar kullanılabilmektedir. Burada veri toplama zaman alıcı olabilir, ancak doğru ve etiketlenmiş bilgiler alırsınız.

Duygu analizi için bir veri kümesi oluşturmanın ilk adımı, veri kaynağını belirlemektir. Bir veri kaynağı, bir metin belgesi veya bir web sitesi veritabanı olabilir. Veri seçimi, belirli bir amaç için kullanılan veri kümesine bağlıdır.

Veri toplama süreci, belirli kaynaklardan veri toplamayı içerir. Bu işlem otomatik veya manuel olabilir. Bu, otomatik veri toplama yöntemleri, web tarama araçları veya API'ler kullanılarak yapılabilir. Öte yandan, manuel veri toplama yöntemi, kullanıcının verileri manuel olarak toplamasını gerektirir.

Ortaya çıkan bölüm önce temizlenmelidir. Bu süreç, metin belgelerinden gereksiz maliyetlerin kaldırılmasını ve verilerin ön işleme adımlarından çıkarılmasını içerir. Ön işleme, metin belgelerinin kelime bölümlemesinden, blok kelime işlemesinden, kök kaldırmasından veya lematizasyondan sonra gerçekleşir (Turney, 2002).

Lematizasyon, doğal dil işleme (NLP) alanında kullanılan bir tekniktir ve sözcükleri köklere göre bölümlere ayırarak analiz etmeye olanak tanır. Bu teknik, bir

koleksiyona farklı varyasyonlar veya eklemelerle oluşturulan örüntülerin tek bir ortak kök içeriğine dayanmaktadır (Bird, Klein ve Loper, 2009; Manning ve Schütze, 1999).

Örneğin "gitti", "gidiyor", "gidecek" gibi farklı zaman ve çekimlere sahip kelimelerin kökleri "git" olarak tanımlanmaktadır.

Veri etiketleme, veri kümesi oluşturmanın en önemli adımlarından biridir. Etiketleme işlemi, metin belgelerini pozitif, negatif veya nötr olarak etiketler. Bu işlem manuel veya otomatik olarak yapılabilir. Otomatik çalıştırma yöntemleri, makine öğrenimi yapıları kullanılarak çalıştırılabilir.

Son olarak, veri setini eğitim veri seti ve test veri seti olarak ayırmak gerekmektedir. Eğitim veri kümesi, DA modelini eğitmek için kullanılan verileri içerir. Test veri seti, modeli test etmek için kullanılmaktadır.

2.1.2. Google Colab

Google Colab, Google tarafından sağlanan ücretsiz bir Jupyter Notebook hizmetidir (Google, 2021). Hizmet, kullanıcıların internet bağlantısı olan herhangi bir cihazdan Python kodu oluşturmasına, çalıştırmasına ve paylaşmasına olanak tanır (Gupta, 2021). Google Colab, kullanıcı tarafından sağlanan Python kodunu çalıştırmak için bulut tabanlı hesaplar kullanır ve kullanıcılara yüksek performanslı GPU'lar ve TPU'lar gibi kaynaklar sağlar (Google, 2021).

Google Colab özellikle araştırmacılar, öğrenciler ve veri bilimcileri için kullanışlıdır. Örneğin, veri analizi, makine öğrenimi ve derin öğrenme gibi veri yoğun işlemler için kullanılabilmektedir. Google Colab, kullanıcıların bilgisayarlarında yeterli kaynak olmadığında bulut tabanlı kaynakları kullanarak bu işlemleri gerçekleştirmesine olanak tanır (Gupta, 2021).

Google Colab, kullanıcının Python kodunu buluta yükleyip çalıştırabileceği bulut tabanlı bir hesap kullanır. Bu sayede kullanıcılar, bilgisayarlarında yeterli kaynak olmadığında istedikleri işlemleri gerçekleştirmek için bulut tabanlı kaynakları kullanabilirler (Gupta, 2021).

Google Colab, kullanıcılara güçlü GPU ve TPU kaynakları sağlar. Bu kaynaklar özellikle veri yoğun işlemler için kullanışlıdır ve kullanıcıların daha hızlı sonuç almasına yardımcı olmaktadır (Google, 2021). Google Colab, kullanıcıların Python kodunu paylaşmasını kolaylaştırır. Kullanıcılar kodlarını Google Drive'a kaydederek başkalarıyla paylaşabilir (Gupta, 2021).

Google Colab, Google Cloud Platform ile birlikte çalışır. Bu, kullanıcıların Google Cloud Platform tarafından sağlanan hizmetleri kullanarak daha zengin projeler geliştirmesine olanak tanır (Google, 2021).

Google Colab, kullanıcıların veri analizi, makine öğrenimi ve derin öğrenme gibi veri yoğun işlemler için ihtiyaçlarını karşılayan bir hizmettir. Google Colab'ın ücretsiz olması, bulut tabanlı kaynaklar kullanması ve kullanıcılara güçlü GPU ve TPU kaynakları sağlaması, onu özellikle veri bilimi alanında popüler bir hizmet haline getirdi (Gupta, 2021; Google, 2021).

2.2. Twitter üzerinden veri toplama

Twitter API (Application Programming Interface), Twitter tarafından sağlanan ve geliştiricilerin Twitter verilerine erişmesine, Twitter hesaplarını yönetmesine ve Twitter platformunda uygulamalar geliştirmesine olanak tanıyan bir hizmettir (Twitter, 2023a). API'ler, farklı yazılım sistemleri arasındaki iletişim için programlama arayüzleri olarak hizmet eder. Twitter API, geliştiricilerin Twitter verilerini kullanarak çeşitli uygulamalar ve hizmetler oluşturmasına olanak tanır (Twitter, 2023a).

Twitter API, tweet koleksiyonu için çok önemlidir. API'ler sayesinde belirli kullanıcılardan gelen tweetleri, konularla ilgili tweetleri, belirli konumlardan tweetleri vb. toplamak kolaydır (Ghosh, 2021).

Twitter API'sini kullanarak tweet toplamak için öncelikle Twitter geliştirici portalında bir hesap oluşturulmalıdır. Bu hesap, API anahtarına erişmeyi ve API'yi kullanarak Tweet toplanmasını sağlamaktadır (Twitter, 2023b).

API anahtarınızı aldıktan sonra Python gibi bir programlama dilinde birkaç satır kod yazarak tweet toplama işlemini gerçekleştirebilirsiniz.

Kimlik doğrulama değişkenleri aracılığıyla API anahtarınızın kimlik doğrulanır ve API değişkenleri aracılığıyla Twitter API'sine erişim sağlanır. search() yöntemi, "Python" kelimesiyle ilgili İngilizce tweetleri arar ve toplanacak tweet sayısı için count parametresini kaullanılır. (Ghosh, 2021).

Twitter API, tweetleri toplamayı çok kolay ve hızlı hale getirir. API anahtarı Twitter Geliştirici Portalından alındıktan sonra, Python veya başka bir programlama dili kullanarak tweet toplanabilir. Bu, belirli kullanıcılardan belirli konular veya tweetler hakkında veri toplanmasına olanak tanımaktadır (Twitter, 2023a).

2.3. Veri Ön İşleme (Veri Temizleme)

Bugünlerde sosyal medya platformlarında milyarlarca tweet paylaşılmaktadır. Bu Tweetlerde yer alan duygu ve düşünceler analiz edilerek işletmelerin ve bireylerin ilgi alanlarına göre farklı amaçlar için kullanılabilmektedir. Bu hedefler, ürün geliştirme, pazarlama stratejisi, fikir araştırması ve politika analizini içerir (Saif, Fernandez, He & Alani, 2013).

Bununla birlikte, tweet DA ile ilgili en büyük sorunlardan biri, tweetlerin doğal dilde yazılmasıdır. Bu nedenle tweetlerin anlamını doğru anlayabilmek ve DA yapabilmek için öncelikle veri temizleme işlemi yapmak gerekmektedir. Veri temizleme, gereksiz bilgileri Tweetlerden kaldırmaktadır. İşlem, tweetlerden özel karakterleri, sayıları, noktalama işaretlerini ve diğer anlamsız bilgileri kaldırmakla başlar. Tweetteki tüm harfleri küçük harfe dönüştürmek de önemlidir. Bu süreçler, tweetlerin anlamını etkilemeden duygu analizinin doğruluğunu artırabilir (Davidov, Tsur, Rappoport, 2010; Ding, Liu ve Yu, 2008).

Stopwords, aynı zamanda duraklama kelimeleri olarak da bilinir, DA için önemli olmadıkları için tweetlerden çıkarılması gereken kelimelerdir. Bu kelimeler "ve", "ama", "veya", "ben", "sen" gibi yaygın sözcükleri içerebilir. Bu sözcüklerin kaldırılması, DA sürecinin doğruluğunu artırır (Yu & Hatzivassiloglou, 2003).

Tweet duyarlılığını analiz ederken veri temizleme ve duraklatma sözcükleri kaldırılmalıdır. Bu adımlar, tweetlerdeki gereksiz bilgileri kaldırır ve DA sürecinin doğruluğunu artırır.

Duygu analizi için mevcut yöntemler arasında sözcük tabanlı yöntemler ve makine öğrenimi tabanlı yöntemler bulunur. Kelime tabanlı yöntemler, tweetlerdeki kelimelerin sıklığına dayalı olarak DA gerçekleştirir. Makine öğrenimi tabanlı teknikler, tweetleri önceden etiketlenmiş verilerle eğitir ve bu verilere dayalı olarak DA gerçekleştirir. Bu yöntemlerin her birinin avantaj ve dezavantajları vardır (Pak ve Paroubek, 2010).

Duygu analizi yöntemleri genel olarak iki kategoriye ayrılır: denetimli ve denetimsiz yöntemler.

Denetimli yöntemler, eğitim verilerini kullanan makine öğrenimi tekniklerini kullanır. Bu yöntemler tipik olarak, özellikle duygularla etiketlenmiş veri örneklerini kullanarak bir model eğitir. Bu model daha sonra belirli bir metnin içeriğini belirlemek için kullanılabilmektedir (Pang & Lee, 2008).

Denetimsiz yöntemler, belirli duygusal etiketler kullanmadan metnin duygusal içeriğini tespit etmek için kullanılmaktadır. Bu yöntemler genellikle cümle veya kelime sıklık analizini içerir. Bu yöntemler, belirli kelimelerin bir metinde veya belirli kelimelerin kombinasyonlarında geçiş sıklığına bağlı olarak duyguların yoğunluğunu belirleyebilir (Mishra ve Yadav, 2019).

Bu nedenle, tweetler üzerinde DA yapmak için öncelikle veri temizleme işlemini gerçekleştirmelisiniz. Bu işlem, gereksiz bilgileri tweetlerden kaldırır ve duygu analizinin doğruluğunu artırır.

2.4. Çeviri

Twitter, dünya çapında milyonlarca insan tarafından fikir, düşünce ve duygularını ifade etmek için kullanılan popüler bir sosyal medya platformudur (Gao & Zhang, 2019). Ancak farklı dillerde yazılmış tweetleri anlamak ve analiz etmek dil engelleri nedeniyle zor olabilir. Bu nedenle tweetlerin otomatik olarak İngilizceye

çevrilmesi tweetlerin anlaşılmasını ve analiz edilmesini kolaylaştırmaktadır (Gao & Zhang, 2019). Özellikle Türkçe gibi karmaşık morfolojik yapılara sahip dillerde DA daha da zorlaşmaktadır.

Türkçe tweetlerin DA, Türkçe dil bilgisi ve yapısının anlaşılmasını gerektirir. Bu nedenle Türkçe tweetlerin DA, İngilizce tweetlere göre daha zordur. Ayrıca, DA kitaplıkları genellikle İngilizce'yi destekler ve diğer dillerde DA için daha az araç ve kaynak sağlanır.

Google Cloud Translate API, yazılı metni çeşitli dillere çevirmeye yönelik bir Google Cloud hizmetidir. Bu hizmet, tweetleri otomatik olarak İngilizce'ye çevrilmesini sağlamaktadır (Google Cloud, 2021). Tweetlerin İngilizceye çevrilmesi, İngilizce bilgisinin grekli olduğu durumlarda tweetlerin anlaşılmasına yardımcı olmaktadır. Tweetlerin İngilizceye çevrilmesi için Google Cloud Translate API'nin kullanılması gerekliliği, İngilizce bilgisinin gerekli olduğu durumlarda farklı dillerde yazılan Tweetlerin anlaşılabilmesi için önem arz etmektedir. Google Cloud Translate API, tweetleri otomatik olarak İngilizce'ye çevirmek için etkili bir araçtır (Gao & Zhang, 2019).

Google Cloud Translate API'yi kullanarak tweetlerin İngilizce'ye çevrilmesi için öncelikle bir Google Cloud hesabına ve bir API anahtarına ihtiyaç duyulmaktadır. Ardından, tweetler Twitter API'sini kullanılarak İnglizce’ye çeviri işlemi yapılmaktadır. İngilizceye çevrilen Tweetler daha analiz edilebilir ve anlaşılması daha kolaydır (Google Cloud, 2021). Sonuç olarak, tweetleri İngilizce'ye çevirmek için Google Cloud Translate API'yi kullanmak, bunların anlaşılmasını ve analiz edilmesini kolaylaştırır. Bu süreç, farklı dillerde tweet atmanın yarattığı dil engelini yıkmanın etkili bir yoludur (Gao & Zhang, 2019).

2.5. Veri Setinin Etikenlendirilmesi

Twitter'da paylaşılan tweetlerin DA, markalaşma, kampanya performansı ve kamu yararı gibi birçok alanda önemli rol oynamaktadır. Bunun için VADER, TextBlob gibi doğal dil işleme araçlarını kullanarak tweetler etiketlendirilebilir.

VADER, sosyal medya metinleri için bir duygu analiz aracıdır (Hutto & Gilbert, 2014). VADER, tweetlerin olumlu, olumsuz veya nötr olarak etiketlendirilmesini sağlamaktadır. Hızlı ve etkili bir DA aracı olan VADER, sosyal medya metinlerinde yaygın olarak kullanılan konuşma özelliklerine uyum sağlar.

TextBlob, Python programlama dili için doğal bir dil işleme kitaplığıdır (Loria, 2018). TextBlob, tweetleri etiketlemek için kullanılabilmektedir ve olumlu, olumsuz ve nötr olarak etiketlenebilir. Doğal dil işleme ve DA için etkili bir araç olan TextBlob, diğer DA kitaplıklarına kıyasla kullanımı kolaydır.

Tweetleri VADER ve TextBlob ile etiketlemek, doğru sonuçlar almanın etkili bir yoludur. Tweetler, VADER ve TextBlob gibi doğal dil işleme araçları kullanılarak etiketlenebilir. Bu araçlar, farklı özelliklere sahip tweetleri doğru bir şekilde etiketleyebilir.

2.5.1. TextBlob ile Etikenlendirme

TextBlob, Python programlama dilinde yazılmış bir NLP kütüphanesidir. Bu kitaplık, metin kitaplıklarını düzenlemek için çeşitli araçlar sağlar. TextBlob, metin verilerini işlerken doğal dil işleme tekniklerini kullanır. Metin belgelerindeki kelimeleri (kelimeler, kelimeler, cümleler, paragraflar) tanımlayabilir ve bu öğeler üzerinde çeşitli işlemler gerçekleştirebilirsiniz. TextBlob, metin verilerinin DA, anahtar kelime işlevleri, konuşma tanıma, veri tanımlama vb. için kullanılabilmektedir (Bird, Klein ve Loper, 2009).

Polarity değerini belirlemek için metin içerisinde geçen her kelimenin textblob duygu sözlüğü içerisindeki puanlaması alınarak ortalaması hesaplanıp duygu etiketi belirlenmektedir;

$$\frac{x_1 + x_2 + x_3 + x_n}{n}$$

Formül – 5: TextBlob Duygu Etiketlendirme

2.5.2. Vader ile Etiketlendirme

Vader (Valence Aware Dictionary ve sEntiment Reasoner) bir DA aracıdır. Vader, doğal dil işleme için bir Python kitaplığıdır. Bu kitaplık, metin verilerinin duyarlılığını belirlemek için kullanılmaktadır. Vader, metindeki olumlu, olumsuz ve tarafsız etkileri belirlemek için sözlük tabanlı bir yaklaşım kullanır. Bu sözlük, kelime listelerinin yanı sıra olumlu, olumsuz ve nötr duyguları ayırt eden bir sistem içerir. Vader, metin verilerindeki duyarlılığı belirlemek için metin belgesindeki her kullanım için duygu puanlarını toplayarak çıkarımlar yapar. Bu sonuç, algılamanın pozitif, negatif veya nötr olup olmadığını belirler. Vader ayrıca metin belgelerindeki ruh hallerini tanımlamak için emoji kullanır (Hutto ve Gilbert, 2014).

Vader her kelime için duygu puanını -4 ile +4 arasında bir ölçekte ölçer; burada -4 en olumsuz ve +4 en olumludur. Tek tek kelimelerin -4 ile 4 arasında bir duyarlılık puanı vardır, ancak bir cümlenin döndürülen duygu puanı -1 ile 1 arasındadır. Bir cümlenin duygu puanı, duygu içeren her kelimenin duygu puanının toplamıdır. Ancak, -1 ile 1 arasında bir değere eşlemek için toplama bir normalleştirme işlemi uygulanır. Uygulanan normalleştirme işlemi;

$$\frac{x}{\sqrt{x^2 + \alpha}}$$

Formül – 6: Vader Sentiment Duygu Etiketlendirme

Burada x, cümleyi oluşturan kelimelerin duygu puanlarının toplamıdır ve alfa, 15 olarak ayarlanan bir normalleştirme parametresidir. Burada x büyüdükçe -1 veya 1'e yaklaşmaktadır. Bu nedenle, VADER duygu analizi, büyük belgelerde değil, tweetler ve cümleler gibi kısa belgelerde en iyi sonucu vermektedir.

2.6. Modellerin Oluşturulması

Duygu analizi için bir model oluşturmak için öncelikle uygun bir veri seti seçmemiz gerekir. Veri seti doğru bir şekilde etiketlenmeli ve kullanılan özellikler duygu analizine uygun olmalıdır. Ardından veri setinin eğitim, doğrulama ve test setlerine ayrılması gerekmektedir. Eğitim seti, model eğitimi için kullanılmaktadır ve

doğrulama seti, modeli aşırı uyum açısından kontrol etmek ve hiperparametreleri optimize etmek için kullanılmaktadır. Modelin etkinliğini arttırmak için bir test seti kullanılmaktadır. Model oluşturma için en popüler teknik, makine öğrenme oranıdır. Bu teknikler, modeller oluşturmak için veri kümesindeki öznitelikleri kullanır. Örnekler arasında Naive Bayes, Karar Ağaçları,Random Forestlar ve Destek Vektör Makineleri (SVM'ler) bulunur (Pang & Lee, 2008). Yeni teknolojiler arasında derin öğrenme araştırması, özellikle sinir ağlarına dayalı modeller popülerlik kazanmaktadır (Kim, 2014).

2.7. Modellerin Eğitilmesi

Modelleri eğitmek için birçok makine öğrenimi tekniği kullanılabilmektedir. Örnekler arasında Naive Bayes, Karar Ağaçları,Random Forestlar ve Destek Vektör Makineleri (SVM'ler) bulunur (Pang & Lee, 2008). Yeni teknolojiler arasında derin öğrenme araştırması, özellikle sinir ağlarına dayalı modeller popülerlik kazanmaktadır (Kim, 2014). Model eğitimi, eğitim merkezinde kuruluş. Eğitim seti, öğrenmeyi modellemek için kullanılmaktadır. Model aşırı uyumunu kontrol etmek ve hiperparametreleri ayarlamak için yönlendirme ayarları kullanılmaktadır. Modelin etkinliğini arttırmak için bir test seti kullanılmaktadır.

Model eğitimi sırasında aşırı uyum sorununa dikkat etmelisiniz. Overfitting, model eğitim çalışmalarının sınırlarını ifade eder. Bu, model güncellemeleri sırasında yavaş performansa neden olabilir. Ayrıca, modelin monitör düzeninin performansını korur ve fazla takmayı önler. Model performansı, accuary (doğruluk), precision (kesinlik), recall (geri çağırma) ve F1 score (F1 puanı) gibi metrikler kullanılarak değerlendirilmelidir.

2.8. Veri Özellik Seçimi

Duygu analizi için veri özelliklerini seçmek için veri seti özellikleri değil, belirli bir özellikler seti kullanılmaktadır. Veri kümesindeki özelliklerin sayısı çok fazlaysa, modelin eğitilmesi çok uzun sürer ve bu da hareket ve aşırı uyum sorunlarına

yol açar. Dengesizlik, ortam, eksik veri ve veri boyutu gibi kurallar da dikkate alınmalıdır. Bu nedenle doğru fonksiyonun öğrenilmesi, modelin tamamlanması ve çalıştırılması gerekmektedir.

Model eğitimi için doğru işlevi seçmek çok önemlidir. Model performansını önemli ölçüde artıran benzersiz bir seçim. Özel seçim tanıtımı, özellik seçimini, özellik çıkarımını ve özellik tanıtımını içerir (Mikolov ve diğerleri, 2013). Özellik seçimi, veri kümenizdeki özellikleri inceleyip analiz etmek ve model eğitimi için en iyi özellikleri seçmekle başlar. Veri setindeki özelliklerin sayısı ve dengesi, gürültü, eksik veri ve veri boyutu gibi hususlar da bu aşamada dikkate alınmalıdır.

Özellik seçiminin farklı beklentileri vardır. Üç farklı kategoriye girenler. B. Filtreleme, Paketleme ve Yerleştirme. Filtreleme teknikleri, veri kümesi içinde niteliklerden bağımsız olarak çalışır (Pang & Lee, 2008). Örneğin, ki-kare ve korelasyona dayalı özellik seçimi (CFS) gibi yöntemler bu kategoridedir. Sarma yöntemi, modeli sağlayan özellikleri kapsar (Manning ve diğerleri, 2008). Örneğin, Özyinelemeli Özellik Eleme (RFE) ve Sıralı Özellik Seçimi (SFS) bu kategoriye girer. Katıştırılmış kontroller, modeller oluştururken özellikler bekler (Aggarwal & Zhai, 2012). Decision Trees, Random Forest ve Support Vector Machines (SVM) gibi karar ağaçları da bu kategoriye girer.

2.9. Modellerin Sonuç Karşılaştırmaları

Duygu analizi için birçok model mevcuttur. Bu, Naive Bayes, Karar Ağaçları, Random Forestlar, Destek Vektör Makineleri (SVM) ve derin öğrenme teknikleri gibi teknikleri içerir. Bu modellerin performansı, f1 score (f1 puanı) ve accuary (doğruluk) gibi metrikler kullanılarak değerlendirilir.

Accuary (Doğruluk), bir sınıflandırma modelinin doğruluğunu hesaplamak için kullanılan bir metriktir (Witten, Frank & Hall, 2016). Accuary (Doğruluk), modelin doğru tahmin ettiği tüm numuneler içindeki tüm numunelerin oranını ifade eder. Ancak, kararsız veri kümeleri için yanıltıcı olabilir. Örneğin, veri kümesindeki pozitif örnek sayısı, negatif örnek sayısından çok daha fazlaysa, modelin doğru tahmin ettiği

tüm örnekler yüksek accuary (doğrulukla) ile negatif olarak sınıflandırılacaktır. Dolayısıyla accuary (doğruluk) metriği tek başına yeterli değildir.

F1 score (F1 puanı), bir sınıflandırma modelinin doğruluğunu hesaplamak için kullanılan bir ölçüdür. F1 score (F1 puanı), duygu ve özgüllük metriklerinin harmonik ortalaması alınarak hesaplanır. Accuary, modelin tüm pozitif numuneler arasından doğru bir şekilde tahmin ettiği pozitif numunelerin oranını ifade eder. Spesifiklik ise, modelin tüm negatif örnekler arasında doğru bir şekilde tahmin ettiği negatif örneklerin oranını ifade eder. F1 score (F1 puanı), model performansını daha iyi ölçmek için hem duyarlılığı hem de özgüllüğü hesaba katar (Powers, 2020).

Örneğin, bir çalışmasında ürün incelemelerinin duygusal tonunu belirlemek için SVM modelini Naive Bayes modeliyle karşılaştırdı (Pang & Lee, 2008). Bu çalışma, SVM modelinin daha yüksek f1 puanları ve doğruluk değerleri verdiğini göstermiştir. Benzer şekilde, başka bir çalışmada derin öğrenme teknikleri kullanılarak Twitter verileri üzerinde DA yapılmıştır (Kim, 2014). Bu çalışma, evrişimli sinir ağı (CNN) ve tekrarlayan sinir ağı (RNN) modellerini karşılaştırmaktadır. Bu çalışmada, CNN modellerinin daha yüksek f1 score (f1 puanları) ve daha yüksek accuary (doğruluk) sağladığı gösterilmiştir.

Birçok model, katsayılar ve ağırlıklar gibi parametreler kullanır. Bu parametreler, modelin hangi özelliklere odaklandığını ve hangilerinin daha önemli olduğunu gösterir. Örneğin, doğrusal bir regresyon modeli, her parametrenin bağımsız değişken üzerindeki etkisini açıklayan katsayılar içerir. Bu katsayılar, model için hangi bağımsız değişkenlerin daha önemli olduğunu gösterir.

Diğer bir yöntem ise özellik önem sıralamasıdır. Bu yöntem, modelin karar vermek için hangi özellikleri kullandığını anlamak için kullanılmaktadır. Özellik önem sıralaması, özelliğin sınıflandırmadaki önemini gösterir. Özellik önem düzeyi, modelin hangi özelliklere daha fazla önem verdiğini gösterir.

Ayrıca karışıklık matrisi, modellerin tahminlerini karşılaştıran bir tablodur. B. Gerçek pozitif, gerçek negatif, yanlış pozitif, yanlış negatif. Bu tablo, modelin hangi

sınıfları doğru veya yanlış olarak sınıflandırdığını gösterir. Bu, modelin neden yanlış olduğunu anlamayı veya doğru tahmin etmeyi kolaylaştırır (Powers, 2020).

ÜÇÜNCÜ BÖLÜM

MATERYAL VE YÖNTEM

Bu çalışmada sosyal medya platformu Twitter üzerinde atılan gönderiler (tweetler) üzerinden makine öğrenmesi yöntemleri kullanılarak duygu analizi yapılmıştır. Denetimli makine öğrenmesi yönteminde daha önceden etiketlendirilmiş veriler ile makine öğrenmesi modelleri eğitilir ve sonuçları karşılaştırılır. Fakat verinin büyüklüğü ve çalışmanın tek bir kişi tarafından hazırlanması sebebiyle farklı bir yaklaşım uygulanmıştır. Etiketlendirme işlemleri Vader ve TextBlob kütüphaneleri ile yapılmış ve bu etiketlendirmeler üzerinden modeller ayrı ayrı eğitilerek sonuçları karşılaştırılmıştır. Çalışmada veri seti "torku" kelimesi geçen tweetleri toplayacak şekilde filtrelenerek oluşturulmuştur. Torku kelimesinin aynı zaman da Konyaspor'un resmi isim sponsoru olmasından dolayı futbol ile ilgili tweetlerin de veri setine girmesi sebebiyle hem veri toplama adımlarında hem de veri ön işleme adımları sırasında veri setinde düzenlemeler gerçekleştirilmiştir. Çalışma veri seti Türkçe olmasına karşın kullanılan veri etiketlendirme kütüphanelerini ya Türkçe hiç desteklememesi yada verimli etiketlendirmeler yapamaması sebebiyle veri seti İngilizce diline çevrilerek işlemler çeviri yapılmış veri seti üzerinden gerçekleştirilmiştir. Kullanılan farklı etiketlendirme yöntemlerinin yanında karşılaştırmalar yapabilmek ve çalışmanın genişletilebilmesi için birden çok makine öğrenmesi modeli ile çalışılmıştır. Çalışmanın devamında bu konulardan ve yöntemlerden detaylı şekilde bahsedilecektir.

3.1. Ön Hazırlık

Çalışmanın başlangıcında ortam seçimi, veri toplama işlemleri için kullanılacak yöntem, kullanılacak dil seçimi ihtiyacı bulunmaktadır. Öncelikle çalışmanın hangi programlama dili ile yapılacağı belirlenmiştir. Python programlama dilinin açık kaynak olması sebebiyle kaynak dökümanın fazla olmasından dolayı dil olarak Python dili seçilmiştir. Sonrasında çalışma ortamı belirlenmesi gerekmiştir. Mevcut son kullanıcı cihazları yeterli işlem gücüne sahip olsa da yedeklilik ve kesintilere karşı etkisiz olmasından ve bulut sistemlerin donanımlarının çok daha güçlü olmasından dolayı bulut bir sistem üzerinde çalışma yürütülmüştür. Bulut sistem

olarak Google Colab platformu kullanılmıştır. Twitter platformundan toplanacak veriler ve nasıl toplanacağı araştırılarak API çözümü ile veri toplamasına karar verilmiştir.

Torku kelimesi için müşteri düşüncülerinin duygu analizi için Google Colab platformu kullanılmıştır.

“Torku” kelimesi geçen tweetler toplanarak veri seti oluşturulmuştur. Veri işlemeleri için python pandas kütüphanesi kullanılmıştır. Pandas kütüphanesi, veri bilimi/veri analizi ve makine öğrenimi görevleri için en yaygın şekilde kullanılan açık kaynaklı bir Python paketidir. Çok boyutlu diziler için destek sağlayan Numpy adlı başka bir paketin üzerine inşa edilmiştir. Veri işlemler sırasında ihtiyaç duyulan görevleri gerçekleştirilmesi için kolaylık sağlamaktadır.

3.2. Veri Seti Oluşturulması

Sosyal medya platformu Twitter üzerinden Twitter API, Twitter geliştirici portalı aracılığıyla alınan API bilgileri kullanılarak başlatılır.

Daha sonra maxTweets adlı bir değişkene 10 değeri atanır ve all_tweets adlı boş bir liste oluşturulur. Daha sonra Atık kullanımı ile ilgili özel bir arama sorgusu gerçekleştirilir. Bu sorgu, belirli bir tarih aralığında ('benden beri' ve 'bitiş göstergesi' ile belirlenir) Türkçe tweetlerin 'torku' bildirimlerini içerir. Ancak yanıtlar ve terimler hariç tutulur (-filter:Bağlantı filtresi:cevaplar) ve kelime açıklamaları Konyaspor ve Fenerbahçe (-konyaspor – fenerbahçe) -galatasaray -Beşiktaş) tweetlerini filtrelemek için kullanılmıştır. Her tweette sayısal bir dizin oluşturur ve beklendiği gibi 50.000'e kadar tweeti bir araya getirir. Tweet tarihi ve biçimlendirilmiş tarih (twtdate.strftime("%Y-%m-%d")) her tweet için all_tweets içerir. Ayrıca tweet içeriği ve URL, add() yöntemi kullanılarak aynı listeye eklenir. Sonrasında, all_tweets listesi, toplanan tweetlerin tarihi, içeriği ve URL'si gibi bilgileri içerir.

Twitter kendi API’ları üzerinden tweet toplama, gönderme vb. işlevleri desteklemektedir. Fakat API için bir twitter hesabı üzerinden proje oluşturulduktan sonra bu proje Twitter tarafından incelenerek onaylandıktan sonra kullanım

açılmaktadır. Bu onay süreci hem uzun sürmekte hem de projenizin firma tarafından kabul görmesi gerektiği için zorlu bir süreç olmaktadır. Proje onaylandıktan sonrasında ise ücretsiz API projeleri süre ve sayı kısıtlı olarak çalışmaktadır. Bu sorunun çözümü için SNSCrape isimli açık kaynak kodlu python kütüphanesi kullanılmıştır.

Tweetlerin toplanması için python ile yazılmış twitter API veya http protokolünü kullanan açık kaynak kodlu SNSCrape kütühanesi kullanılmıştır. Öncelikle bir twitter hesabı ile API keyine ihtiyaç duyulmaktadır.

Tweetlerin toplanması sırasında "torku" kelimesi ile ilgili kelime benzerliğinden dolayı ilgisiz tweetlerin de listelendiği tespit edilmiştir. Bu sebeple belirli filtrelemeler ile bu sorun aşılmıştır. Filtreler kullanılarak bu filtrelerin geçtiği tweetler liste dışı tutulmuştur. Bunlar;

- Reklam tweetlerini ve urlleri elemek için link filtresi,
- Konyaspor ile ilgili tweetleri elemek için Konyaspor, spor vb.
- Araçlar ile ilgili (araç torku) tweetleri engellemek için araba, lastik, viraj vb.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 38010 entries, 0 to 38009
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   id         38010 non-null  int64
 1   tweet_en   38010 non-null  object
 2   tweet_tr   38010 non-null  object
 3   tweet_orj  38010 non-null  object
 4   date       38010 non-null  object
 5   url        38010 non-null  object
 6   username   38010 non-null  object
dtypes: int64(1), object(6)
memory usage: 2.0+ MB
```

Resim 1: Veri Seti Boyutu

Veri setinin geniş tutulabilmesi için tarih sınırlaması yapılmamıştır. Sayı sınırlaması olarak 50.000 seçilmiştir. Fakat Twitter üzerinde toplamda 38010 tweet bulunmuştur. Aşağıda toplanan veri setine dair görüntü yer almaktadır.

	tweet-tr	tarih
5000	#nutella da neymiş Torku on basar buna Fiskobi...	2020-04-10
5001	Sırf şeker yenecek bir şeyi yok. Torku bitter...	2020-04-10
5002	kahvaltıda kendimizi şımartalım dedik.. torku ...	2020-04-10
5003	#nutella alacağımı Türk malı alırım NUGA ve TO...	2020-04-10
5004	101 den aldığım 350gr Fermente dana sucukla il...	2020-04-10
...	...	...
10006	Pınar alma,Torku alma derken sıradaki ne olaca...	2017-08-08
10007	Torku Konya Spor'un golünden sonraki sesleri i...	2017-08-08
10008	Torku dan 1 tl lik alışveriş yaparsam elim kır...	2017-08-08
10009	Tinere karşı Torku ayranı 💪 \n #AnadoluSeninle...	2017-08-08
10010	torku yeşil şerbet = cool lime, artık sıtarbak...	2017-08-08

Resim 2: Veri Seti İçeriği

3.3. Veri Temizleme

Twitter verileri, kullanıcı adları, bağlantılar, sayılar, özel karakterler ve diğer dış etkenlerden oluşur. Bu nedenle, analizden önce tweetleri önceden işlemek önemlidir.

Bu işleve clean_tweets denir. Bu işlev bir Tweet parametresi alır. Bu parametre, temizlenecek Tweet içeriğini içerir. İlk olarak, tweetin metni küçük harfe dönüştürülür ("lower_case_tweet" değişkeni).

Daha sonra "..." değerleri boşluklarla kümelenir (tree_dot değişkenleri). Bu işaretçilerin grupları Tweetlerde kullanılmaktadır, ancak genellikle gereksizdir ve analitiğin medya bileşeni olarak kabul edilir.

Bir sonraki bölümde, Remodule kullanıcı adları, bağlantılar, özel karakterler ve sayılar (user_removed değişkeni) kaldırılmıştır. Bu bölümde kullanılan normal ifadeler şunlardır:

(@[A-Za-z0-9]+): Kullanıcı adları

(http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\\(\\),]|(?:%[0-9a-fA-F][a-fA-F]))+): Linkler

([^\\w\\s]): Özel karakterler

(\\d): Sayılar

([^0-9A-Za-z \t]): Özel karakterler

(\w+:\/\/\S+): Linkler

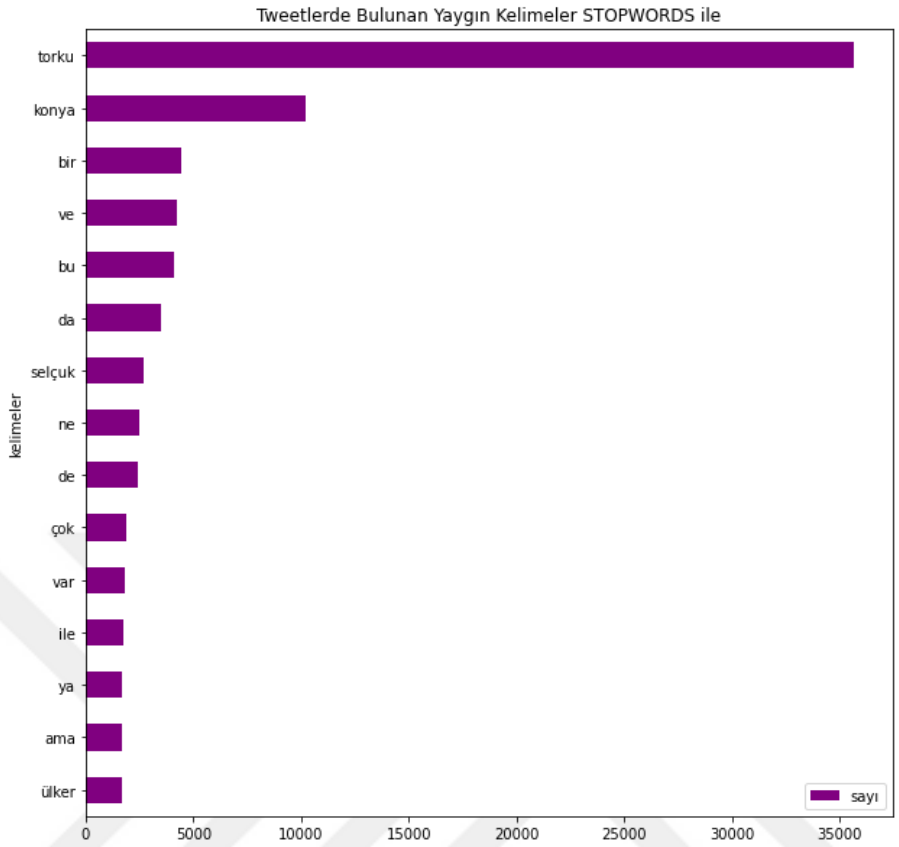
Sonraki bölümler, özel karakterler, sayılar ve Türkçe karakterler (number_removed değişkeni) kaldırılmıştır. Bu hareketlerde kullanılan normal ifade [^a-zA-Z0-9workforceö] şeklindedir.

@ tanımları ve özel karakterler (hasag_removed değişkeni) kaldırılır.

Bir sonraki bomba, nltk kitaplığından WordPunctTokenizer sınıfını kullanarak tweet cümlelerinin (kelime değişkenleri) hecelenmiş ve sayılar kaldırılmıştır.

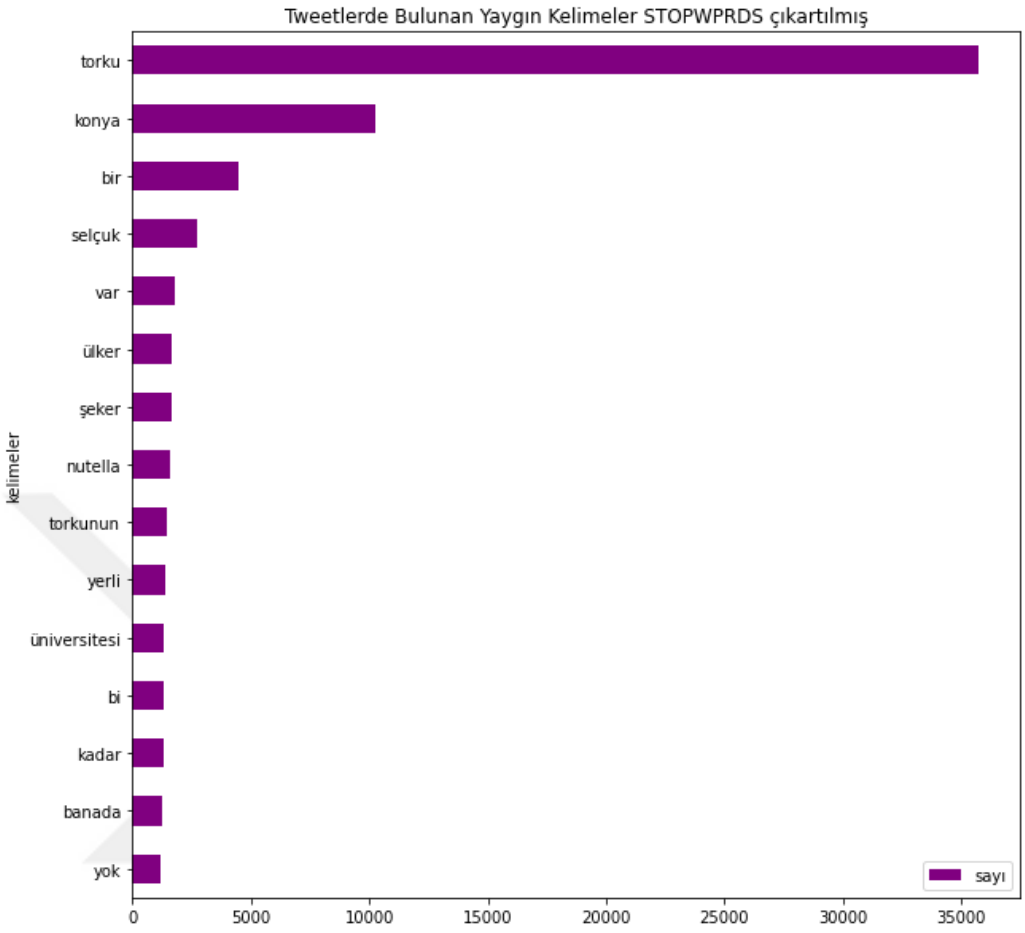
Sonrasında veri setinde bulunan fakat duygu durumuna etki etmeyen sonlandırıcı kelimeleri (stopwords) çıkarılmıştır.

Aşağıda bulunan grafikte sonlandırıcı kelimeler filtrelenmeden öncesinde tüm tweetler içerisinde geçen en yaygın kelimeler ve sayıları gösterilmiştir.



Grafik 1: Sonlandırma Kelimeleri Filtrelenmeden Tweetlerde Bulunan Yaygın Kelimeler

Stopwords kelimeler filtrelendikten sonra ise en yaygın kelime ve sayıları aşağıdaki gibi olmuştur.



Grafik 2: Sonlandırma Kelimeleri Filtrelenerek Tweetlerde Bulunan Yaygın Kelimeler

Yukarıdaki grafikte de görüldüğü gibi sonlandırma kelimeleri filtrelendikten sonra en çok geçen kelimelerin konu ile ilgili kelimeler olduğu belirlenmiştir.

Fakat kelime havuzu içerisinde "torku" kelimesinin aynı zaman Konyaspor'a isim sponsorluğu sebebiyle futbol ile alakalı kelimelerin de olduğu belirlenmiştir. Yukarıda bahsedilen tweet filtreleme işlemi sırasında belirli filtreler kullanılmasına rağmen bazı tweetlerin bu filtrede etkilenmediği gözlemlenmiştir.

Toplanan tweetler üzerinde API ile veri setinin oluşturulması sırasında filtrelenmeyen fakat veri seti incelendiğin halen "torku" kelimesi ile ilgisi olmayan tweetlerin olduğu gözlemlenmiştir. Toplanan veri seti excel tablosuna kaydedildiği

için tablo üzerinden aşağıda bulunan kelimeler geçen tweetler tekrar filtreleme işlemi ile veri setinden çıkartılmıştır;

'sport', 'fenev', 'takım', 'takim', 'gol', 'torku arena', 'şekerspor', 'spor', 'trabzon', 'beşiktaş', 'galatasaray', 'besiktas', 'selçuk','üniversite','basketbol','basketball','futbol', 'football', 'kupa', 'maç', 'besiktasta', 'forex', 'stad','stadyum','sarı kart', 'kırmızı kart', 'hakem','gol','gool','basketball','kupa', 'beko', 'cup' kelimeler içeren tweetler filtrelenerek veri setinden çıkartılmıştır.

Bu işlemlerin ardından veri setinde bulunan toplam tweet sayısı 10278 olmuştur.

Son olarak, birleştirilmiş temiz tweet metni ("clean_tweet" değişkeni) bir değişkene atanmıştır.

3.4. Çeviri

Türkçe tweet, Google Cloud Platform Speech API kullanılarak İngilizce'ye çevrilmiştir. Öncelikle google.cloud ve googleapiclient kitaplıkları içe aktarılır. Ardından, Google Cloud Platform'dan bir API anahtarı alınmıştır ve build() yöntemini kullanarak Çeviri API'sini başlatılmıştır.

Translate() adlı fonksiyon ile çevrilecek metni (Metin), kaynağı (Kaynak) ve hedef başarısını (Hedef) parametre olarak gönderilmiştir.. İşlev içinde, build() yöntemi Translate API ile başlatılır. Metin daha sonra service.translations().list() yöntemi kullanılarak çevrilir. Bu yöntem çevrilecek metni q parametresinde, kaynak dil kodunu kaynak parametresinde ve hedef dil kodunu hedef parametresinde alır. Çeviri sonucu bir sözlük yapısı döndürür. Son olarak, çevrilmiş metin (a["translations"][0]["translatedText"]) işlevden döndürülür ve bir değişkene aktarılmıştır.

3.5. Veri Setinin Etikenlendirilmesi

Veri etiketlem işlemi veri setinde bulunan verilerin pozitif negatif veya nötr olarak işarelenmesi demektir. Bu, bir metni olumlu, olumsuz veya nötr olarak

yorumlama sürecidir. Bu çalışmada veri etiketlendirme için iki kütüphane kullanılmıştır; Bunlar TextBlob ve Vader kütüphaneleridir.

TextBlob, metnin duygusal tüketimini hesaplamak için kullanılan kutupsal değerleri hesaplayan doğal bir dil işleme kitaplığıdır. Polarite değerleri -1 ile 1 arasındadır, burada -1 negatif ve 1 pozitiftir. TextBlob etiketlendirme işlemi, her kelimenin kutupsal gözlemini dikkate alır ve değerine bağlı olarak algının olumlu veya olumsuz olduğunu belirler.

Başka bir doğal dil işleme kitaplığı olan Vader, metnin duygusal durumunu belirlemen için kullanılan bileşik değerleri hesaplar.

Bileşik değer, -1 ile 1 arasında bir değerdir; burada -1 negatif ve 1 pozitiftir. Vader'ın mekanizması, her algının bileşik gözlemlerini gözlemlemesi ve anlayışı değerine göre olumlu veya olumsuz olarak işaretlemesidir.

3.6. Veri Öznitelik Seçimi

Veri öznitelik seçimi, bir veri kümesi içinde bir özelliğin kullanılıp kullanılmayacağına karar verme sürecidir. Bu işlem, veri kümesindeki metni vektörleştirir ve her denemeyi bir özellik olarak kullanır. Bu çalışmada, veri niteliklerini seçmek için TfidfVectorizer kullanılmıştır.

TfidfVectorizer, doğal dil işleme uygulamaları için sıklıkla kullanılan bir özellik çıkarma aracıdır (Chen vd., 2020). Bu araç, kelimelerin metinde geçiş sıklığını hesaplar ve her kelime için bir ağırlık değeri üretir (Chen vd., 2020). Bu ağırlık değerleri daha sonra özellik matrisinde kullanılabilmektedir.

Öznitelik seçimi, TfidfVectorizer'da kullanılan kelime dağarcığının boyutunu belirlemek anlamına gelir (Liu ve diğerleri, 2021). Kelime dağarcığı, bir metindeki tüm sözcüklerin bir listesidir. Liste ne kadar uzun olursa özellik matrisi de o kadar büyük olmakta ve bu da model eğitim sürecini yavaşlatabilmektedir (Liu ve diğerleri, 2021).

TfidfVectorizer, metni vektörleştirmeye yönelik bir yöntemdir. Cümlelerdeki kelimelerin ağırlığına odaklanan bir tekniktir. Bu, metin içindeki anlamı belirler ve her kelimeyi bir özellik olarak kullanır. Bu, veri özniteliklerini azaltır ve model performansını artırır.

3.7. Modellerin Oluşturulması

Model oluşturulurken, veri kümesindeki metinleri olumlu, olumsuz veya nötr olarak sınıflandıran bir sınıflandırma algoritması seçilir ve uygulanır. Bu çalışma farklı sınıflandırma modelleri kullanır. Bunlar:

Random Forest:

Random forest, karar ağaçlarından oluşan bir makine öğrenimi modelidir ve bir topluluk öğrenme tekniğini temsil eder. Bu model, daha güçlü ve sağlam öngörü yetenekleri sağlamak için birden fazla karar ağacını bir araya getirir. Her ağaç rastgele seçilen özellikleri öğretir ve sonuçlar nihai tahmini yapmak için toplanır.

Random forest, çıkarım ve regresyon problemlerinde yaygın olarak kullanılmaktadır. Çalışma prensibi, her ağacı tahmin etmek ve oylama sonuçlarını veya tüm ağaçların tahminlerinin ortalama bağımsız yapısını belirlemektir. Böylece aşırı yüke dayanıklı bir model oluşturulur.

Linear SVM:

Linear SVM (Destek Vektör Makinesi), what-if problemleri için yaygın olarak kullanılan bir makine öğrenme modelidir. SVM'ler, veri noktalarını uzaydaki sınıflarla sınırlamak için karar sınırları veya hiper düzlemler oluşturur. Linear SVM, devam eden kitapları keşfetmek için harikadır.

Linear SVM, maksimum sınır bileşeni ilkesine dayanmaktadır. Veri noktaları, sınıflar arasındaki marjı en üst düzeye çıkarmak için optimize edilmiştir. Bu marj, sınıflar arasındaki boşluğu en üst düzeye çıkarmak ve yeni pencereler için doğru sınıf tahminleri yapmak için kullanılmaktadır.

Naive Bayes:

Naive Her Bayes, özellikle doğal dil işleme gibi metin tabanlı problemlerde etkili olan saplantılı bir düşünme tekniğidir. Bu model, Bayes'in teorisine dayalı konfigürasyon gerçekleştirir.

Naive Bayes, özellikler arasında ölçümler yapmaktadır. Yani özelliklerin birbirinden bağımsız olduğu varsayılmaktadır. Bu düşünme biçimi sayesinde, duygu için gerekli olan olasılık hesapları kolaylaşmaktadır.

Gradient Boosting:

Gradient boosting, temel olarak zayıf tahmincilerden güçlü tahminciler oluşturmayı amaçlayan bir topluluk tekniği olan bir makine öğrenimi tekniğidir. Bu modelde, zayıflamış birimler hata düzeltme işlemi tahminine birer birer girer.

Bir gradyan artırma işlemi, her tahmincinin önceki tahmin edicinin hatasını korumaya çalıştığı gradyan (gradyan) tabanlı bir yaklaşımdır. Bu şekilde, her tahmin edici, genel hatayı azaltmak için önceki tahmin edicinin hatasını düzeltir ve son tahmin edici, genel hatayı en aza indirecek şekilde okur.

3.8. Modellerin Eğitilmesi

Modellerin eğitilmesi, veri setindeki metinlerin kullanılan sınıflandırma algoritmalarına göre sınıflandırılması işlemidir. Bu işlem GridSearchCV ile gerçekleştirilmiştir.

Model eğitimi için veri seti %30 test, %70 eğitim verisi olarak bölünmüştür. Sonrasında veri setinin %70 i ile modeller eğitilerek kalan %30 model testleri için kullanılmıştır.

GridSearchCV bir hiperparametre optimizasyonudur. Bu bağlamda, belirli hiperparametrelerin farklı değerlerini test ederek hiperparametre ölçümlerini mümkün olan en iyi performansla gerçekleştirmeye çalışılmıştır. Bu, model limitlerini artırır ve fazla öğrenmeyi önler. Her model seçilen optimal hiperparametrelerle eğitilmiştir.

Bu hiperparametreler;

Random Forest Modeli:

n_estimators = 200: Oluşturulacak ağaç sayısını ifade eder. Daha yüksek bir n_estimators değeri, daha iyi performans elde edilmesine yardımcı olabilir, ancak hesaplama süresi arttıracaktır.

max_depth = 3: Oluşturulacak maksimum ağaç derinliği. Daha yüksek bir maksimum derinlik değeri, daha ayrıntılı bir model üretecektir, ancak fazla uyuma riski artar.

Gradient Boosting Modeli:

n_estimators = 200: Oluşturulacak ağaç sayısını ifade eder. Daha yüksek bir n_estimators değeri, daha iyi performans elde edilmesine yardımcı olabilir, ancak hesaplama süresini arttıracaktır.

max_depth = 3: Oluşturulacak maksimum ağaç derinliği. Daha yüksek bir maksimum derinlik değeri, daha ayrıntılı bir model üretecektir, ancak fazla uyuma riski artar.

LinearSVM Modeli:

C=1.0: Hata toleransını kontrol eden bir parametredir. Daha yüksek bir C değeri, yanlış hizalamalar için daha büyük cezalara ve daha az fazla takma riskine yol açar. Daha düşük bir C değeri, daha fazla başlangıç kapasitesi ve daha fazla fazla takma riski sunar.

max_iter=1000: Algoritmanın maksimum yineleme sayısı.

Naive Bayes Modeli:

alfa = 1.0: Laplace düzenlemesinin bir parametresidir. Laplace sort, Naive Bayes sınıflandırıcısında kullanılan bir sıralama yöntemidir. Bu yöntem, Bayes teoremindeki sonucun sıfır olma olasılığını ortadan kaldırmak için kullanılmaktadır. Daha yüksek alfa değerleri daha fazla düzeltme uygular ve gereksiz olma olasılığı daha düşüktür. Daha düşük alfa değerleri daha fazla hareket sağlar ve fazla takma riski daha yüksektir.

fit_prior = True: Bu, sınıf keşfi gerçekleştiğinde sınıf frekansının korunmasını sağlar. Sınıf frekansı, bir veri kümesindeki örneklerin sınıf etiketleri arasındaki dağılımını ifade eder. Örneğin, iki sınıflı bir sınıflandırma probleminde, sınıfın frekansı, veri kümesindeki sınıf etiketlerinin örnek sayısına bağlıdır. Sınıf frekansları, Naive Bayes sınıflandırıcısının sınıflar arasındaki dengesizlikleri hesaba katmasına izin verir.

3.9. Modellerin Sonuç Karşılaştırmaları

Classification_report kullanarak model performansını ölçme. Bu işlev, modelin doğruluğunu, kesinliğini, hatırlamasını ve F1 puanını hesaplar. Ayrıca, her model için pozitif veya negatif olarak işaretlenen metinsel modaliteleri de ortadan kaldırır. Bu, hangi modelin en iyi performansı gösterdiğini belirlemektedir.

Bu çalışma, veri kümesindeki metni etiketleme, veri özniteliklerini seçme, beklenen değerleri seçme, modelleri eğitme, sonuçları karşılaştırma gibi birçok teknik işlemi gerçekleştirmektedir.

Bu veriler üzerindeki performans, VADER ve TextBlob DA kitaplıkları kullanılarak ölçülmüştür. Farklı mekanik yapı modelleri kullanılarak yapılandırılmış yapı sonuçları, precision (kesinlik), recall (geri çağırma), f1 score (f1 puanı) ve accuary (doğruluk) metrikleri kullanılarak değerlendirilmiştir.

Bir karışıklık matrisi, sınıflandırma problemleri için bir performans göstergesidir. Bir sınıflandırma problemi, bir veri kümesindeki örnekleri farklı sınıflara bölen bir makine öğrenimi problemidir.

Bir karışıklık matrisi, gerçek ve tahmin edilen sınıflar arasındaki ilişkiyi gösteren bir kare matristir. Bir sınıflandırma modeli, her örneği belirli bir sınıfa atar. Karışıklık matrisi, gerçek ve tahmin edilen sınıflar arasındaki farkı gösterir ve modelin doğruluğunu değerlendirir. Karışıklık matrisi, dört farklı değere sahip 2x2'lik bir matristir.

Gerçek pozitif (TP), yanlış pozitif (FP), yanlış negatif (FN) ve gerçek negatif (TN). TP gerçek pozitiflerin sayısı, FP yanlış pozitiflerin sayısı, FN yanlış negatiflerin sayısı ve TN gerçek negatiflerin sayısıdır.

Aşağıda modellerin her iki etiketlendirmeye göre karışıklık matrisleri ve sınıflandırma performans tabloları verilmiştir.

- **Random Forest**

Random Forest				
	precision	recall	f1-score	support
negative	0.69	1.00	0.82	2131
positive	0.00	0.00	0.00	953
accuracy			0.69	3084
macro avg	0.35	0.50	0.41	3084
weighted avg	0.48	0.69	0.56	3084

Tablo – 4: Random Forest Sınıflandırma Performansı (TextBlob)

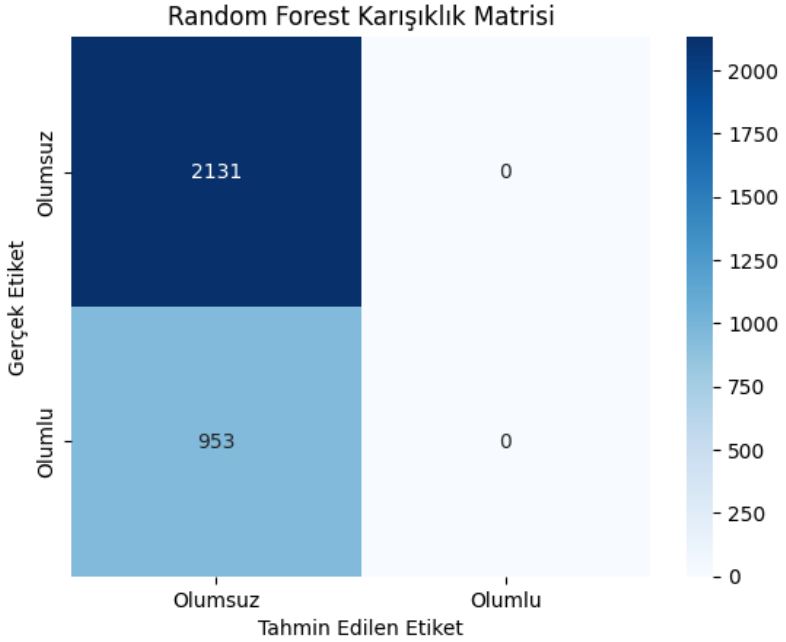
Random forest orman modelinin negatif sınıfının precision değeri 0.69 idi. Bu, negatif olarak tahmin edilen verilerin yaklaşık %69'unun aslında negatif olduğunu göstermektedir. Ancak pozitif sınıf için 0.00 değeri belirlendi. Bu, pozitif olduğu tahmin edilen testlerin hiçbirinin doğru olmadığı anlamına gelir. Bu, modelin pozitif sınıflar arasında ayrım yapamayacağını gösterir. Ek olarak, negatif sınıf için, yani 1.00'lik bir recall değeri belirlendi. Gerçek negatif numunelerin %100'ü doğru şekilde sınıflandırılmıştır. Ancak pozitif sınıf, 0 recall değeri hesaplandı. Gerçekten pozitif

olan testlerin hiçbiri doğru bir şekilde tanınamadı. Benzer şekilde F1 score değeri negatif sınıf için 0.82, pozitif sınıf için 0.00 olarak hesaplanmıştır. Genel olarak, random forest modeli pozitif sınıfı sınıflandırmakta başarısız olmuştur ve accuary değeri zayıftır (0.69).

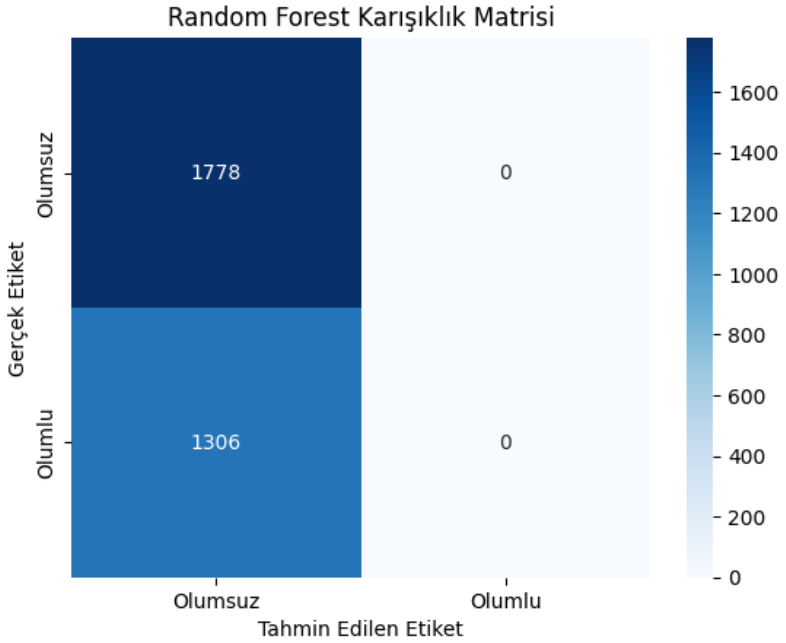
Random Forest				
	precision	recall	f1-score	support
negative	0.58	1.00	0.73	1778
positive	0.00	0.00	0.00	1306
accuracy			0.58	3084
macro avg	0.29	0.50	0.37	3084
weighted avg	0.33	0.58	0.42	3084

Tablo – 5: Random Forest Sınıflandırma Performansı (VADER)

Random forest modelinin negatif sınıfının precision değeri 0.58 idi. Bu, negatif olarak tahmin edilen verilerin yaklaşık %58'inin aslında negatif olduğunu gösterir. Ancak pozitif sınıf için 0.00 değeri belirlendi. Bu, pozitif olduğu tahmin edilen testlerin hiçbirinin doğru olmadığı anlamına gelir. Recall değerleri de benzer şekilde hesaplandı, negatif sınıf için 1.00 ve pozitif sınıf için 0.00 değerleri hesaplandı. F1 score değerleri de, negatif sınıfta 0.73 ve pozitif sınıfta 0.00 idi. Random forest modeli, pozitif sınıfı doğru bir şekilde tahmin edemedi ve düşük bir accuary oranına (0.58) sahipti.



Grafik 3: Random Fores Karışıklık Matrisi (TextBlob)



Grafik 4: Random Forest Karışıklık Matrisi (VADER)

- **Linear SVM**

Linear SVM:				
	precision	recall	f1-score	support
negative	0.97	0.99	0.98	2131
positive	0.97	0.92	0.94	953
accuracy			0.97	3084
macro avg	0.97	0.95	0.96	3084
weighted avg	0.97	0.97	0.97	3084

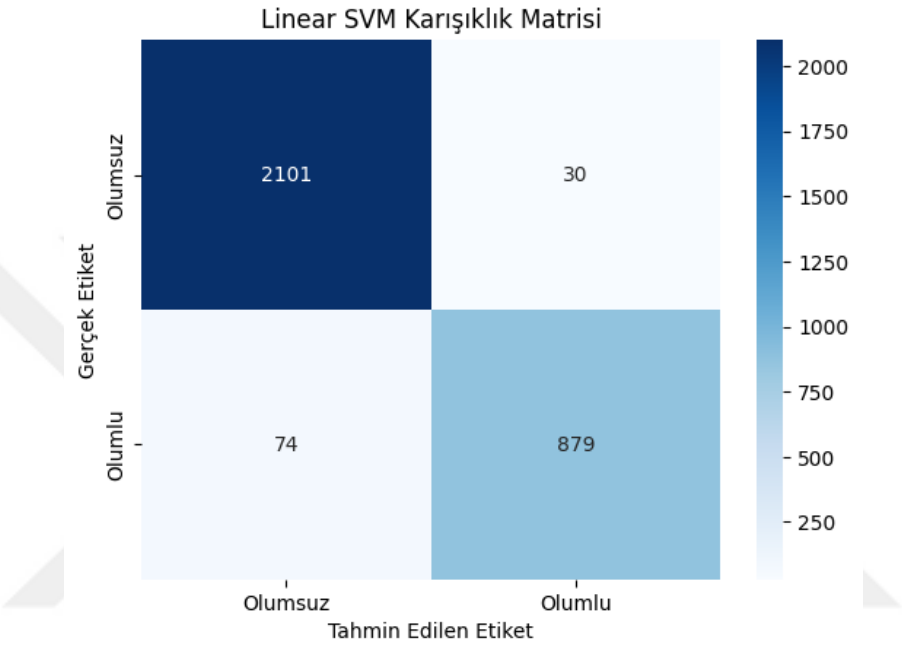
Tablo – 6: Linear SVM Sınıflandırma Performansı (TextBlob)

Linear SVM modeli hem negatif hem de pozitif sınıflarda yüksek precision değerlerine (0.97 ve 0.97) ve recall değerleri (0.99 ve 0.92) elde etti. Bu hem negatif hem de pozitif sınıfların doğru bir şekilde tanındığını gösterir. Precision değeri, kavramsal modelin doğru bir şekilde tanımlandığını gösterir ve precision tahmini, doğru şekilde tanımlanan gerçek pozitif testlerin sayısını gösterir. F1 score değeri de negatif sınıfta 0.98, pozitif sınıfta 0.94 gibi yüksekti. Linear SVM modelleri genellikle yüksek accuary değeri ile (0.97) iyi performans gösterdi.

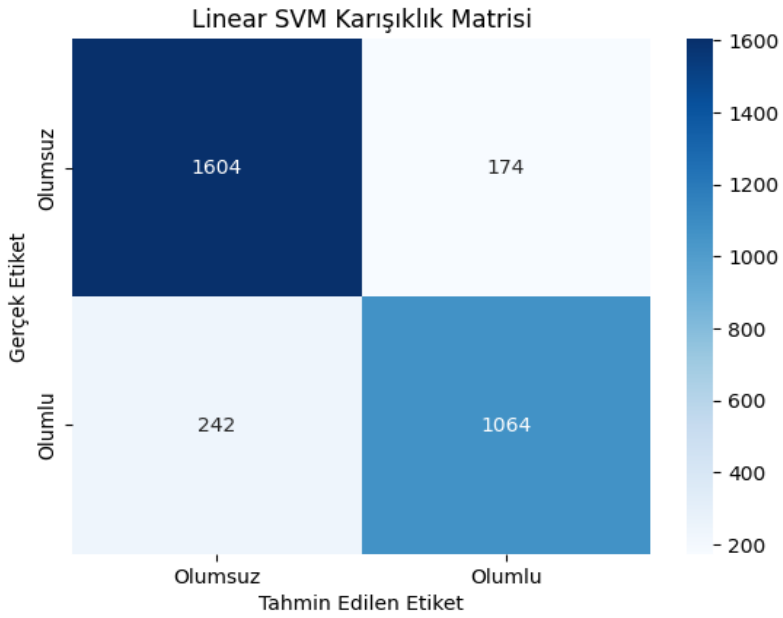
Linear SVM:				
	precision	recall	f1-score	support
negative	0.87	0.90	0.89	1778
positive	0.86	0.81	0.84	1306
accuracy			0.87	3084
macro avg	0.86	0.86	0.86	3084
weighted avg	0.86	0.87	0.86	3084

Tablo – 7: Linear SVM Sınıflandırma Performansı (VADER)

Linear SVM modeli, negatif ve pozitif bölmeler için yüksek precision değerleri (0.87 ve 0.86) ve recall değerleri (0.80 ve 0.81) elde etti. Bu hem negatif hem de pozitif sınıfların doğru bir şekilde tahmin edilebileceğini göstermektedir. F1 score değeri de negatif sınıfta 0.89, pozitif sınıfta 0.84 gibi yüksekti. Linear SVM modeli yüksek accuary değeri ile (0.87) iyi performans gösterdi.



Grafik 5: Linear SVM Karışıklık Matriksi (TextBlob)



Grafik 6: Linear SVM Karışıklık Matrisi (VADER)

- **Naive Bayes**

Naive Bayes:				
	precision	recall	f1-score	support
negative	0.90	0.90	0.90	2131
positive	0.77	0.76	0.77	953
accuracy			0.86	3084
macro avg	0.83	0.83	0.83	3084
weighted avg	0.86	0.86	0.86	3084

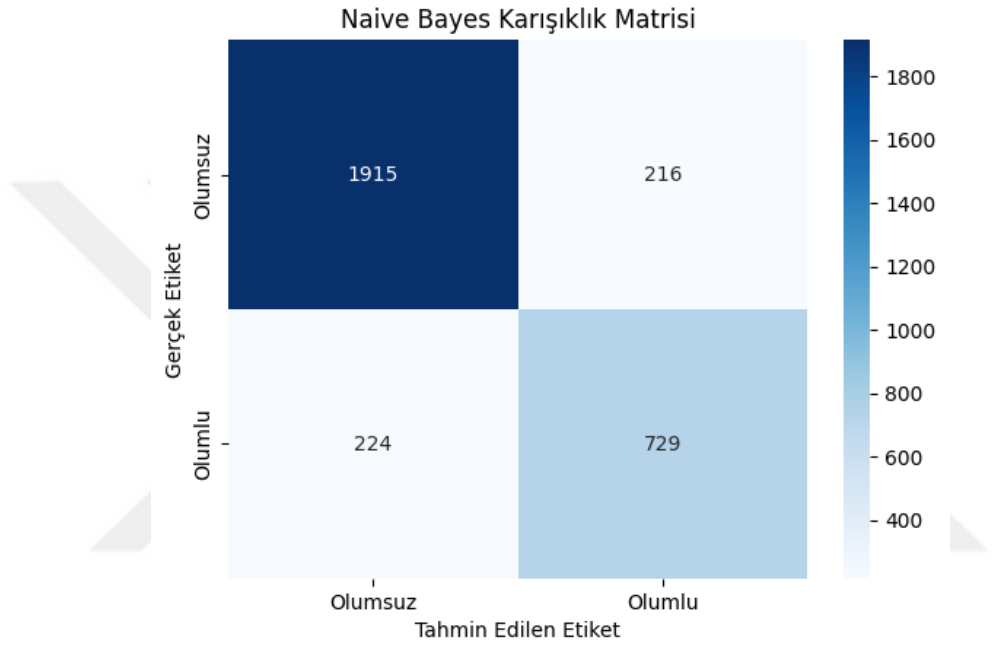
Tablo – 8: Naive Bayes Sınıflandırma Performansı (TextBlob)

Naive Bayes modeli, negatif ve pozitif sınıflarda kabul edilebilir precision değerleri (0.90 ve 0.77) ve recall değerlerine (0.90 ve 0.76) ulaşmıştır. Bu, modelleme örneğinin doğru tahminlerle iyi çalıştığını gösterir. Benzer şekilde F1 score değeri negatif sınıf için 0.90, pozitif sınıf için 0.77 olarak hesaplanmıştır. Naive Bayes modeli genel olarak kabul edilebilir bir accuary oranına (0.86) ulaştı, ancak pozitif sınıf beklentileri açısından diğer modellerden biraz daha kötü performans gösterdi.

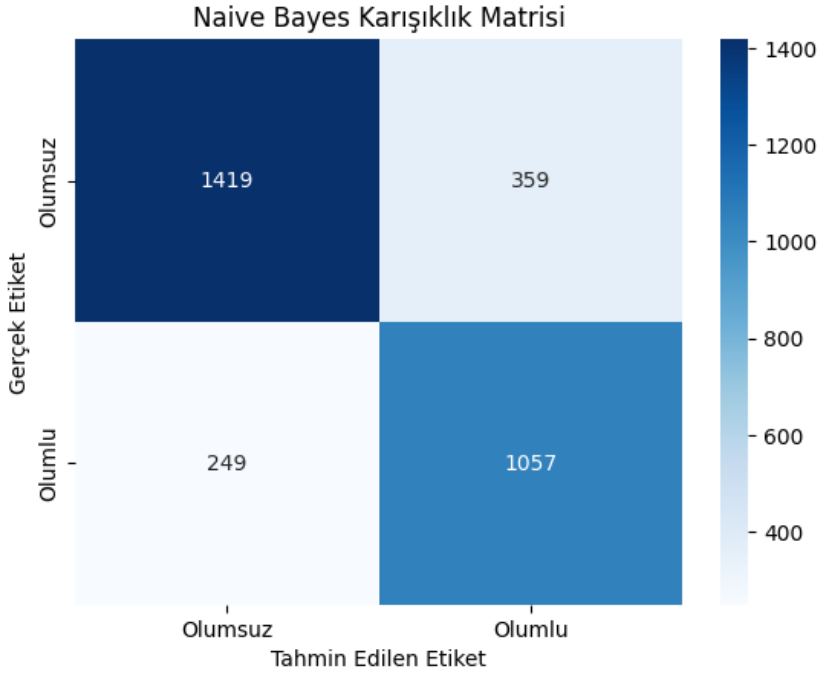
Naive Bayes:				
	precision	recall	f1-score	support
negative	0.85	0.80	0.82	1778
positive	0.75	0.81	0.78	1306
accuracy			0.80	3084
macro avg	0.80	0.80	0.80	3084
weighted avg	0.81	0.80	0.80	3084

Tablo – 9: Naive Bayes Sınıflandırma Performansı (VADER)

Naive Bayes modeli, negatif ve pozitif sınıflarda kabul edilebilir precision değerleri (0.85 ve 0.75) ve recall değerlerine (0.80 ve 0.81) ulaştı. Bu, modelleme örneğinin doğru tahminlerle iyi çalıştığını gösterir. Benzer şekilde F1 score değeri negatif sınıf için 0.82, pozitif sınıf için 0.78 olarak hesaplanmıştır. Naive Bayes modeli, genel olarak kabul edilebilir accuary oranlarıyla (0.80) pozitif sınıf varsayımında diğer modellerden biraz daha iyi performans gösterdi.



Grafik 7: Naive Bayes Karışıklık Matrisi (TextBlob)



Grafik 8: Naive Bayes Karışıklık Matrisi (VADER)

- **Gradient Boosting**

Gradient Boosting				
	precision	recall	f1-score	support
negative	0.93	0.97	0.95	2131
positive	0.93	0.83	0.88	953
accuracy			0.93	3084
macro avg	0.93	0.90	0.91	3084
weighted avg	0.93	0.93	0.93	3084

Tablo – 10: Gradient Boosting Sınıflandırma Performansı (TextBlob)

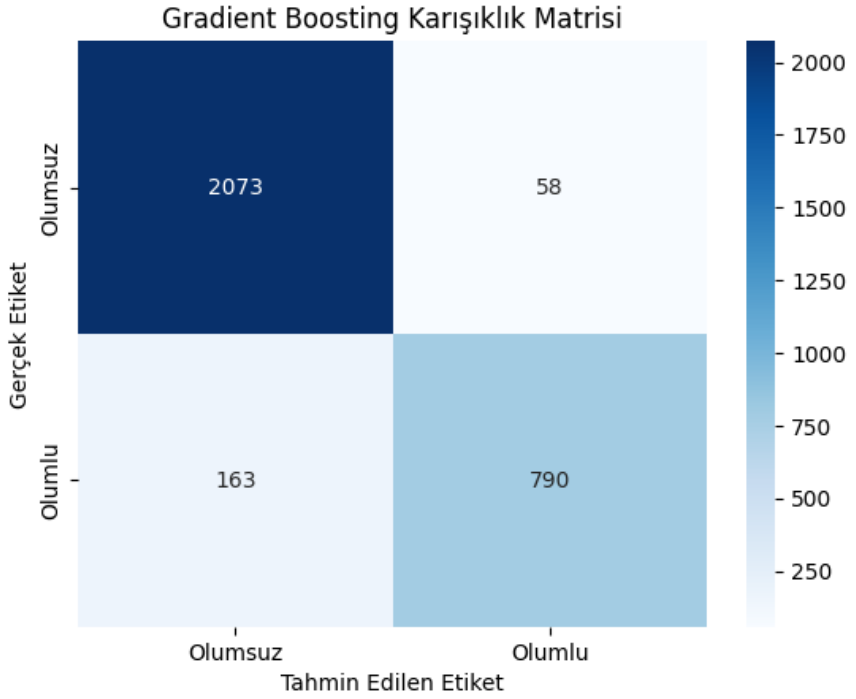
Gradient boosting modeli, negatif ve pozitif sınıflarda yüksek precision değerleri (0.93 ve 0.93) ve recall değerleri (0.97 ve 0.83) elde etti. Bu, modelin hem negatif hem de pozitif sınıfları doğru bir şekilde tahmin edebildiğini göstermektedir.

F1 score da negatif sınıfta 0.95 ve pozitif sınıfta 0.88 gibi yüksekti. Gradient Boosting modeli genellikle yüksek bir accuary oranıyla (0,93) başarılı olmaktadır.

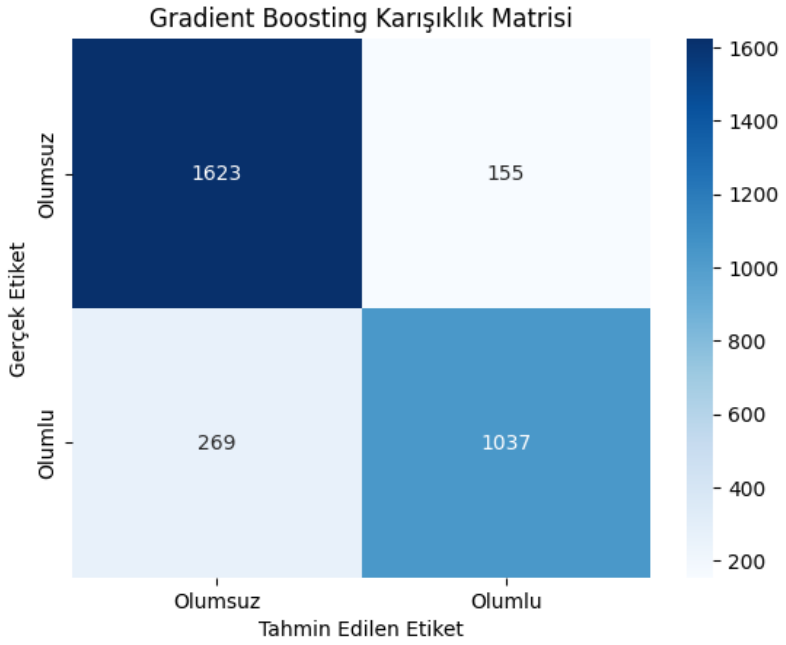
Gradient Boosting				
	precision	recall	f1-score	support
negative	0.86	0.91	0.88	1778
positive	0.87	0.79	0.83	1306
accuracy			0.86	3084
macro avg	0.86	0.85	0.86	3084
weighted avg	0.86	0.86	0.86	3084

Tablo – 11: Gradient Boosting Sınıflandırma Performansı (VADER)

Gradient boosting modeli, negatif ve pozitif sınıflarda yüksek precision değerleri (0.86 ve 0.87) ve recall değerleri (0.91 ve 0.79) elde etti. Bu, modelin hem negatif hem de pozitif sınıfları doğru bir şekilde tahmin edebildiğini göstermektedir. F1 score değeri de negatif sınıfta 0.88, pozitif sınıfta 0.83 gibi yüksektir. Gradient Boosting modeli genellikle yüksek accuary değeri ile (0,86) başarılı olduğu söylenebilir.



Grafik 9: Gradient Boosting Karışıklık Matrisi (TextBlob)



Grafik 10: Gradient Boosting Karışıklık Matrisi (VADER)

	Textblob Accuary	Vader Sentiment Accuary
Random Forest	0.69	0.58
Linear SVM	0.97	0.87
Naive Bayes	0.86	0.80
Gradient Boosting	0.93	0.86

Tablo – 12: Modellerin Accuary (Doğruluk) Karşılaştırmaları

Sonuç olarak TextBlob için, Linear SVM modeli ve Gradient Boosting modeli en yüksek precision, recall ve F1 score değerlerine sahipti ve genel olarak iyi tahmin performansı sergiledi. Naive Bayes modeli kabul edilebilir performans gösterdi, ancak Random Forest modeli pozitif sınıfta başarısız oldu.

Vader için ise, Linear SVM ve Gradient Boostin modellerinin en yüksek precision, recall, F1 score değerlerine ve iyi bir tahminleme performansına sahip olduğunu gösterdi. Naive Bayes modeli kabul edilebilir performans gösterdi, ancak Random Forest modeli pozitif sınıfta başarısız oldu.

Accuary (Doğruluk) oranlarına bakıldığı zaman 3 modelin iki etiketlendirme yönteminde de yüksek oranlar verdiği modellerin başarılı olduğu gözlemlendi. Accuary(Doğruluk) değerlerine bakıldığı zaman yukarıda ki tabloda da görüldüğ gibi en yüksek oran Textblob etiketlendirme ile Lienar SVM modellinde olduğu görülmüştür. Ayrıca karışıklık matrislerine bakıldığı zaman ise en doğru tahminleme yapan modelin Linear SVM (TextBlob) olduğu görülmektedir.

3.10. Kelime Ağırlıkları

N-gram, doğal dil işlemede kullanılan bir kavramdır. Bir mesajın veya parolanın n-gramı, n ardışık tümceden oluşur. Örneğin, "merhaba dünya" ifadesi için 1 gram "merhaba", "dünya" ve 2 gram "merhaba dünya" dır.

N-gram oranı, metindeki belirli bir n-gramın davranışını belirlemek için kullanılan bir ölçüdür. N-gram frekansları metnin gramer yapısını, kelime yapısını ve diğer özelliklerini analiz etmek için kullanılmaktadır. 1 gramlık bir oran, metindeki

her geçiş olarak kabul edilir. Örneğin, bir metinde en sık kullanılan kelimeleri bulmak için bir gramlık kelime sıklığı yöntemi kullanılabilmektedir.

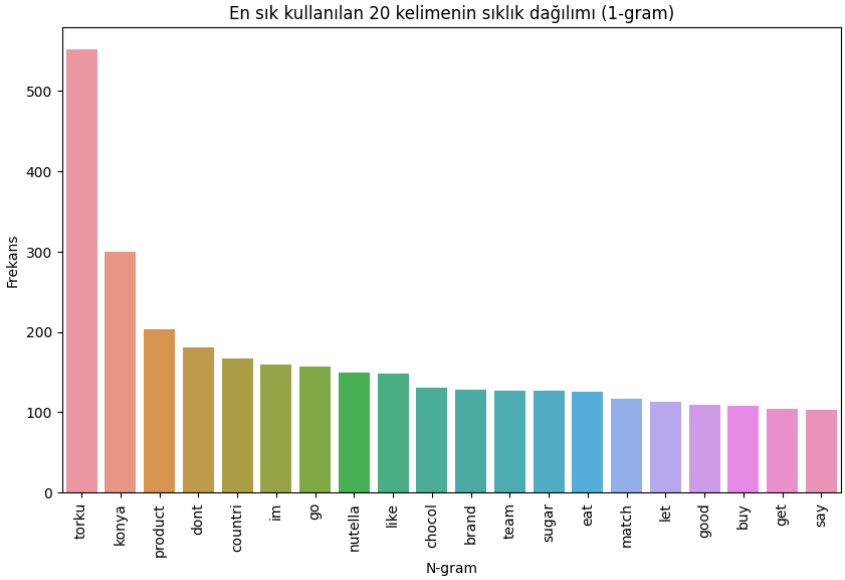
N-gram, bir metinde ardışık kelime gruplarını ifade eden ve duygu analizi yöntemlerinde sıklıkla kullanılan bir terimdir (Ghosh ve Veale, 2016). N-gram, metindeki kelimelerin sırasını dikkate alarak bir metnin duygusal tonunu belirlemeye yardımcı olan bir kümeleme tekniğidir (Pang ve Lee, 2008).

N-gramlar, bir metindeki farklı boyutlardaki kelime gruplarını ifade eder. Örneğin 2 gramlık bir cümle iki kelimeden oluşurken, 3 gramlık bir cümle üç kelimeden oluşur (Liu, 2012). Bu gruplar, metindeki kelimelerin sırasına bakarak metnin duygusal tonunu belirlemeye yardımcı olmaktadır. N-gramlar, sözlük tabanlı bir duygu analizi yönteminde bir kelime listesi oluşturmak için kullanılmıştır (Pang & Lee, 2008). Cümlelerin boyutu, metnin duygusal tonunu belirlemek için kullanılan yöntemlere bağlı olarak değişebilir. Makine öğrenimi tabanlı yöntemlerde, eğitim verileri ve algoritmik olarak belirlenmiş duygusal ton oluşturmak için n-gramlar kullanılmaktadır (Akbik vd., 2018).

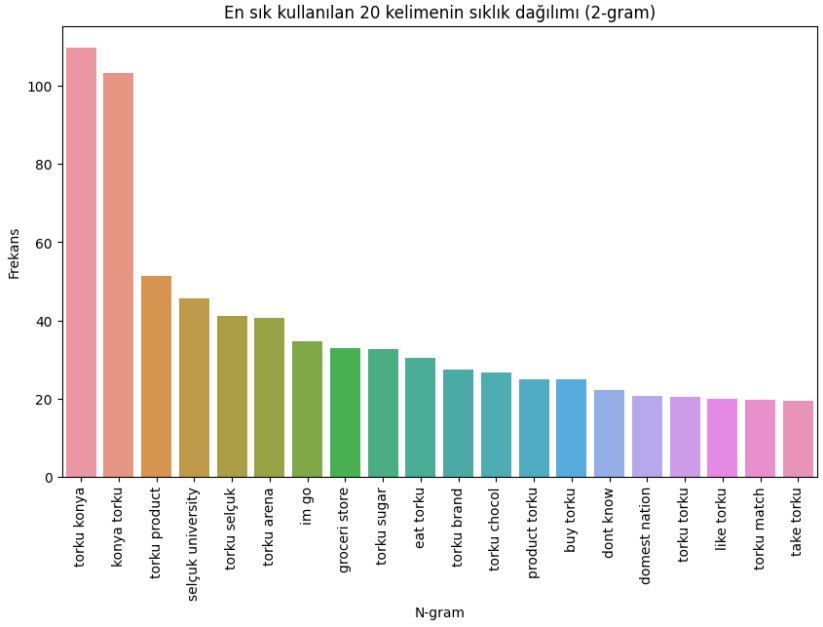
Örneğin 2 gramlık bir örnek olarak "iyi film" ifadesi olumlu bir duyguyu, "kötü film" ifadesi ise olumsuz bir duyguyu ifade etmektedir. 3 gramlık bir örnek olarak “çok iyi film” ifadesi daha güçlü olumlu duyguları ifade ederken, “hiç iyi değil” ifadesi daha güçlü olumsuz duyguları ifade etmektedir (Ghosh ve Veale, 2016). Özetle n-gram, duygu analizi yöntemlerinde metnin duygusal tonunu belirlemeye yardımcı olan bir kümeleme tekniğidir. Cümlelerin boyutu, metnin duygusal tonunu belirlemek için kullanılan yöntemlere bağlı olarak değişebilir. N-gramlar, sözlük tabanlı duygu analizi ve makine öğrenimi yöntemlerinde kullanılabilmektedir (Liu, 2012).

2-gram, metinde iki ardışık noktayı tamamlamak için hesaplama. Burada deyimler ve cümle yapıları hakkında bilgi bulabilirsiniz. Örneğin metin içerisinde en sık kullanılan iki kelime öbeği veya ardışık ikili kelimeleri 2 gram hızında bulabilirsiniz.

N-gram frekansları, doğal dil işlemenin birçok alanında kullanılmaktadır. Örneğin metin yapısı, kelime önerileri ve benzeri görevler için kullanılmaktadır. N-gram vericileri, metindeki önemli kelimeleri tanımlamak için de kullanılabilmektedir. Örneğin metinde en çok kullanılan 2 gram, metindeki en önemli cümleleri birbirine bağlar.



Grafik – 11 : En sık kullanılan 20 kelimenin (1-gram) sıklığı



Grafik – 12 : En sık kullanılan 20 kelimenin (2-gram) sıklığı

Yukarıda çalışma veri setinde oluşan 1'li ve 2'li n-gramlar gösterilmiştir.

SONUÇ VE ÖNERİLER

Bu çalışma sonucunda ML tabanlı DA modellerinin yüksek accuary (doğruluk) puanlarına sahip olduğunu gözlemlenmiştir. Modeller genellikle %90 veya %100 accuary (doğruluk) elde eder, bu da makine öğreniminin DA için etkili bir şekilde kullanılabileceğini düşündürür.

Ancak, gerçek verilerin ve metnin farklı betimlemelerle temsil edilmesinin genel uygulamada bazı sorunlara yol açabileceğine dikkat çekilmiştir. Bu nedenle, daha büyük ve daha çeşitli veri kitapları için modellemeyi dikkate almak önemlidir. Bu, metnin biriktirdiği nüansları ve dilsel karmaşıklığı ve metin temizleme sürecini dikkate almayı içerir. Alay ve yanlış anlamalar da dahil olmak üzere DA yapmak ve metin olmadan duygusal ifadeleri doğru bir şekilde belirlemek zor olabilir. Bu gibi durumlarda, modelin yanlış yönlendirmeyi veya yorumu gerçekmiş gibi göstermemesi anlaşılır bir durumdur. Bu, daha gelişmiş dil işleme teknikleri ve metin temizleme yöntemleri kullanarak bu bölümlerin kümelerini oluşturmaya yönelik bir kılavuzdur.

Bu çalışma DA uygulamalarında güvenilir sonuçlar elde etmenin mümkün olduğunu göstermektedir. Özellikle sosyal, pazarlama analitiği ve müşteri geri bildirim analitiği alanlarında bu modeller, kullanıcılara ve medya müşterilerine önemli avantajlar sunabilir. Modelin üst düzey sistemi, daha iyi kararlar almanıza yardımcı olmak için müşteri hizmetlerini ve geri bildirimleri etkili bir şekilde analiz edebilir.

Özetle, bu çalışma, makine öğrenimi tabanlı DA modellerinin yüksek accuary (doğruluk) değerlerine sahip olduğunu ve oldukça güvenilir DA sonuçları elde edebildiğini göstermektedir. Ancak, gerçek veriler üzerindeki performansını doğrulamak ve metnin yapamadığı zorlukları ele almak için daha fazla araştırma ve geliştirmeye ihtiyaç vardır.

Ancak duygu analizinde Türkçe verilerin çeviri işlemi ile analiz edilmesinin sorunlara sebebiyet verdiği çeviri işleminin Türkçe'nin yapısından dolayı yeterli olmadığı çeviri işlemi sırasında duygu durumunun yanlış yorumlamaya sebep olabileceği gözlemlenmiştir. Ayrıca "torku" kelimesinin spor kulübü sponsorluğu sebebiyle ve araç torku anlamında da kullanılmasından dolayı veri temizleme

işlemlerinde veri setinin içerisinde analizleri etkileyen tweetlerin de yer almasına sebep olmuştur. Bu kelimeleri içeren tweetler veri temizleme işlemleri sırasında veri setinden çıkartılmıştır. Bu sebeple Twitter verileri ile duygu analizi yapılması için veri temizleme işlemlerinin önemli olduğu fakat temizleme işlemleri sırasında duygu analizi için belirlenen konuyla ilgisi olmayan tweetlerinde analize dahil olabileceği bu sebeple veri setinin kontrol edilmesinin önemi anlaşılmıştır.

KAYNAKÇA

Aggarwal, C. C. (2015). Data classification: Algorithms and applications. Boca Raton, FL: CRC Press.

Aggarwal, C. C., & Zhai, C. (2012). Mining text data (Vol. 7). Springer Science & Business Media.

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. arXiv preprint arXiv:1805.04867.

Alpaydin, E. (2010). Introduction to machine learning (2nd ed.). Cambridge, MA: MIT Press.

Balahur, A., & Turchi, M. (2015). Sentiment analysis in social media. In Multilingual Information Access Evaluation II. Multimedia Experiments (pp. 441-453). Springer, Cham.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

Bishop, C. M. (2006). Pattern recognition and machine learning. New York, NY: Springer.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

Brownlee, J. (2019). How to Calculate Precision, Recall, F1, and More for Deep Learning Models. Machine Learning Mastery.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 77-91). New York, NY: ACM.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. IEEE Computational Intelligence Magazine, 9(2), 48-57.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2), 15-21.

Campbell, M., Hoane, A. J., & Hsu, F. (2002). Deep Blue. Artificial intelligence, 134(1-2), 57-83.

Chandola, V. (2018). Confusion Matrix in Machine Learning. Towards Data Science.

Chen, Y., Li, T., & Wang, W. (2020). A comparative study of feature selection methods for text classification. Neural Computing and Applications, 32(23), 16773-16785.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. Proceedings of the 23rd International Conference on Computational Linguistics, 241-249.

Ding,X., Liu, B., & Yu, P. (2008). A Holistic Lexicon-Based Approach to Opinion Mining. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 340-348.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.

Flach, P. (2015). Machine learning: The art and science of algorithms that make sense of data. Cambridge, UK: Cambridge University Press.

Gao, F., & Zhang, J. (2019). Sentiment Analysis of Twitter Data Using Machine Learning Techniques. IEEE Access, 7, 72584-72591.

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (2nd ed.). Sebastopol, CA: O'Reilly Media.

Ghosh, A. (2021). How to Scrape Tweets with Python and Tweepy. Towards Data Science. https://towardsdatascience.com/how-to-scrape-tweets-with-python-and-tweepy-95ededa3e4ab

Ghosh, S., & Veale, T. (2016). A novel n-gram based approach to sentiment classification. Expert Systems with Applications, 61, 161-171.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. Cambridge, MA: MIT Press.

Google Cloud. (2021). Cloud Translation. Retrieved from https://cloud.google.com/translate.

Google. (2021). Colaboratory. Google. https://research.google.com/colaboratory/

Gupta, A. (2021). Google Colab - Introduction to Colab. GeeksforGeeks. https://www.geeksforgeeks.org/google-colab-introduction-to-colab/

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182.

Han, J., & Kamber, M. (2006). Data mining: Concepts and techniques (2nd ed.). San Francisco, CA: Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). New York, NY: Springer.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. Eighth International Conference on Weblogs and Social Media.

Kim, D., & Yoon, S. (2018). A systematic review of sentiment analysis research in hospitality and tourism. Journal of Travel Research, 57(8), 1063-1083.

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Koedinger, K. R., & Corbett, A. T. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In K. Sawyer (Ed.), The Cambridge handbook of the learning sciences (pp. 61-78). Cambridge, UK: Cambridge University Press.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (pp. 1137-1143). Montreal, Quebec, Canada.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2),273-324.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31, 249-268.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

Lasko, T. A., Denny, J. C., & Levy, M. A. (2013). Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PLoS ONE, 8(6), e66341.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

Liu, X., Zhang, X., & Cai, H. (2021). A hybrid feature selection method for text classification. Journal of Ambient Intelligence and Humanized Computing, 12(3), 3205-3214.

Loria, A. (2018). TextBlob: Simplified text processing. https://textblob.readthedocs.io/en/dev/.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge

Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mishra, S., & Yadav, A. K. (2019). A Review of Sentiment Analysis Techniques. International Journal of Advanced Research in Computer Science, 10(2), 61-66.

Murphy, K. P. (2012). Machine learning: A probabilistic perspective. Cambridge, MA: MIT Press.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. LREC, 10(2010), 1320-1326.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 187, 253-318.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. Sebastopol, CA: O'Reilly Media.

Rajaraman, A., & Ullman, J. D. (2011). Mining of massive datasets. Cambridge, UK: Cambridge University Press.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. J. d. (2010). Handbook of educational data mining. Boca Raton, FL: CRC Press.

Saif, H., Fernandez, M., He, Y., & Alani, H. (2013). Evaluation Datasets for Twitter Sentiment Analysis: A Survey and a New Dataset, the STS-Gold. Language Resources and Evaluation, 47(1), 1-38.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

Scikit-learn. (2021). Confusion Matrix. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html

Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. New York, NY: Cambridge University Press.

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589-1604.

Steinhaus, H. (1956). Sur la division des corps materiels en parties. Bulletin de l'Academie Polonaise des Sciences, C1. III, VIII, 1957.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). Cambridge, MA: MIT Press.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

Taylor, M. E., & Stone, P. (2009). Transfer learning for reinforcement learning domains: A survey. Journal of Machine Learning Research, 10(Aug), 1633-1685.

Thakur, A. (2020). Understanding Confusion Matrix.Towards Data Science.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 417-424.

Turney, P. D. (2010). Sentiment analysis and subjectivity. Handbook of Natural Language Processing, 2, 627–666.

Twitter. (2023a). About Twitter API. https://developer.twitter.com/en/docs/twitter-api

Twitter. (2023b). Getting started with the Twitter API.

https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api

Wang, J., Yang, M., Zhu, Y., Wang, R., & Su, X. (2016). A convolutional neural network approach to predict lung cancer incidence. Neurocomputing, 191, 88-97.

Witten, I. H., Frank, E., & Hall, M. A. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Yu, H., & Hatzivassiloglou, V. (2003). Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 129-136.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. In Proceedings of the 20th International Conference on Machine Learning (ICML-03) (pp. 912-919).