****

**MARMARA ÜNİVERSİTESİ**

**TEKNOLOJİ FAKÜLTESİ**

**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

**BİTİRME PROJESİ**

Twitter Verilerinden Makine Öğrenmesi ve Derin Öğrenme Yöntemleri ile Duygu Analizi

****

**PROJE YAZARI**

Elif YANIK

Mustafa Eren GÜLBAHAR

**DANIŞMAN**

Prof. Dr. Serhat ÖZEKES

**İSTANBUL, 2024**

**MARMARA ÜNİVERSİTESİ**

**TEKNOLOJİ FAKÜLTESİ**

**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

Marmara Üniversitesi Teknoloji Fakültesi Bilgisayar Mühendisliği Öğrencisi Elif YANIK ve Mustafa Eren GÜLBAHAR’ın “Twitter Verilerinden Makine Öğrenmesi ve Derin Öğrenme Yöntemleri ile Duygu Analizi” başlıklı bitirme projesi çalışması, 03/06/2024 tarihinde sunulmuş ve jüri üyeleri tarafından başarılı bulunmuştur.

**Jüri Üyeleri**

Prof. Dr. Serhat ÖZEKES (Danışman)

Marmara Üniversitesi ......................................................................... (İMZA) ..................

Dr. Öğr. Üyesi Eyüp Emre ÜLKÜ (Üye)

Marmara Üniversitesi ......................................................................... (İMZA) ..................

Dr. Öğr. Üyesi Timur İNAN (Üye)

Marmara Üniversitesi ......................................................................... (İMZA) ..................

# **İÇİNDEKİLER**

**SEMBOLLER LİSTESİ** **5**

**KISALTMALAR LİSTESİ** **6**

**ŞEKİL LİSTESİ** **8**

**TABLO LİSTESİ** **9**

**ÖZET** **10**

**ABSTRACT** **11**

**1. GİRİŞ** **12**

**2. LİTERATÜR TARAMASI****14**

**3. MATERYAL VE YÖNTEM** **….27**

3.1.Doğal Dil İşleme…………………………………………………………...……27

3.2.Metin Madenciliği …………………………………………………………….…29

3.2.1.Veri Ön İşleme ..………………………………………………………………..31

3.2.2.Veri Etiketleme ……………………………………………………………..….32

3.3.Derin Öğrenme ……………………………………………………………..……32

3.3.1.Derin Öğrenme Algoritmaları …………………………………………………34

3.3.1.1.Yapay Sinir Ağları …………………………………………………...………35

3.3.1.2.Konvolüsyonel Sinir Ağları …………………………………………………35

3.3.1.3.Uzun Kısa Vadeli Hafıza Ağları ………………………….…..……………...36

3.3.1.4.Embedding ……………………………………………………..……………36

3.4.Makine Öğrenmesi ………………………………………………………………37

3.4.1.Makine Öğrenmesi Algoritmaları ……………………………………………..37

3.4.1.1.Naive Bayes Sınıflandırması ……………………………………………...…37

3.4.1.2.Destek Vektör Makineleri ……………………………………………………38

3.4.1.3.Aşırı Gradyan Arttırma Algoritması ……………………………………..….39

3.4.1.4.Kategorik Arttırma Algoritması ……………………………………..………39

3.4.2.Performans Ölçütleri …………………………………………………..………40

3.5.Duygu Analizi ………………………………………………………...…………41

3.6.Twitter ve Twitter API, Selenium ile Veri Çekme ………………….……………41

3.7. Speech-to-Text Fonksiyonu………………………………………...……………43

**4. BULGULAR VE TARTIŞMA****44**

4.1.Çalışmanın Mimari Yapısı ……………………………………………………….44

4.2.Çalışmada Kullanılan Veri ve Sesli Verinin Kullanımı ,…………………………45

4.3. Çalışma Ortamı ve Kullanılan Kütüphaneler ……………………………...……46

4.4. Twitter Verilerine Erişim ………………………………………………..………48

4.5.Veri Ön işleme …………………………………………………………...………49

4.5.1. İstenmeyen Öğelerin Temizlenmesi …………………………………………..51

4.6. Veri Etiketleme ……………………………………………………………….…51

4.7. Verilerin Ayrılması ve Modelleme ………………………………………….…..52

4.7.1.Modellerin Doğruluklarının Karşılaştırılması …………………………………57

4.8.Arayüz Oluşturma ……………………………...…………………………..……58

4.9. Duygu Analizi ve Şekilleştirme……………...…………………………….……60

**5. SONUÇLAR** **62**

**6 .KAYNAKÇA** **64**

**7. ÖZGEÇMİŞ** **75**

**SİMGELER LİSTESİ**

**@:** Twitter kullanıcı adı öneki

**#:** Hashtag

**P(c|x):** Sonraki olasılık (posterior probability)

**x:** Veri

**c:** Hipotez

P(x|c): Hipotezin doğru olduğu varsayıldığında verinin olasılığı

**P(x):** Verinin olasılığı

**P(c):** Hipotezin doğru olma olasılığı

# **KISALTMALAR LİSTESİ**

**API:** Application Programming Interface

**DDİ:** Doğal Dil İşleme

**GPU:** Graphic Processing Unit

**RNN:** Recurrent Neural Network

**GRU:** Gated Recurrent Unit

**TF-IDF:** Term Frequency-Inverse Document Frequency

**NLTK:** Natural Language Toolkit

**ANN:** Artificial Neural Network

**CNN:** Convolutional Neural Network

**LSTM:** Long Short Term Memory

**SVM :** Support Vector Machine

**XGBOOST:** Extreme Gradient Boosting

**CATBOOST:** Category Boosting

**DT:** Decision Trees

**MLP:** Multi-layer Perceptron

**XGB:** Extreme Gradient Boosting

**SVM:** Support Vector Machine

**MLR:** Multiple Linear Regression

**NB:** Naive Bayes

**KNN:** K-Nearest Neighbor

**LR:** Linear Regression

**POS:** Parts of Speech Tagging

**KNIME:** Konstanz Information Miner

**BoW:** Bag of Words

**URL:** Uniform Resource Loader

**YSA:** Yapay Sinir Ağları

**GBM:** Gradient Boost Machine

**GBDT:** Gradient Boosting Decision Trees

**TP:** True-Positive

**FP:** False-Positive

**FN:** False-Negative

**TN:** True-Negative

**HTML:** HyperText Markup Language

**HTTP:** Hiper Metin Transfer Protokolü

**CSV:** Comma Separated Values

**XML:** Extensible Markup Language

**XPath:** Extensible Markup Language Path

**KVKK/GDPR:** Kişisel Verilerin Korunması Kanunu/Genel Veri Koruma Tüzüğü

**RT:** Retweet

# **ŞEKİL LİSTESİ**

**Şekil 3.1.** Derin sinir ağı …………………………………………………………………… 33

**Şekil 3.2.** Derin Öğrenme Süreçleri …………………………………………………………34

**Şekil 3.4.** Karışıklık Matrisi …………………………………………………………………40

**Şekil 4.1.** Çalışma sisteminin mimari yapısı. ………………………………………..………44

**Şekil 4.2.** Sınıflandırma modelinin mimari yapısı. …………………………………….……45

**Şekil 4.3.** Etkisiz kelimelerin (stopwords) veri setinden çıkartılması.………………….……50

**Şekil 4.4.** Veri setinden özel karakter, emoji, url ve rakamların temizlenmesi. ……………..51

**Şekil 4.5.** ANN derin öğrenme algoritmasıyla oluşturulan modele ait karışıklık matrisi.………………………………………………………………………………………..53

**Şekil 4.6.** Embedding derin öğrenme algoritmasıyla oluşturulan modele ait karışıklık matrisi………………………………………………………………………………………...54

**Şekil 4.7.** CATBOOST makine öğrenmesi algoritmasıyla oluşturulan modele ait karışıklık matrisi.……………………………………………………………………………….….……55

**Şekil 4.8.** XGBOOST makine öğrenmesi algoritmasıyla oluşturulan modele ait karışıklık matrisi.………………………………………………………………………………….…….55

**Şekil 4.9.** SVM makine öğrenmesi algoritmasıyla oluşturulan modele ait karışıklık matrisi.………………………………………………………………………………….…….56

**Şekil 4.10.** Naive Bayes makine öğrenmesi algoritmasıyla oluşturulan modele ait karışıklık matrisi.………………………………………………………………………………….…… 57

**Şekil 4.11.** Makine öğrenmesi algoritmaları başarı karşılaştırması.…………………………57

**Şekil 4.12.** Derin öğrenme algoritmaları başarı karşılaştırması.……………………………..58

**Şekil 4.13.** Giriş ekranı.……………………………………………………………………...59

**Şekil 4.14.** Olumlu cümle girildiğindeki ekran.……………………………………………...60

**Şekil 4.15.** Olumsuz cümle girildiğindeki ekran.………………………………….…………60

**Şekil 4.16.** Çalışmada kullanılan duygu analizi mimarisi.……………………..….…………61

**Şekil 4.17.** Duygu sınıflarının dağılım oranı.…………………………………..….…………62

# **TABLO LİSTESİ**

**Tablo 3.1.** Metin Madenciliği ile Veri Madenciliğinin Karşılaştırılması …………………... 29

# **ÖZET**

Teknolojinin ilerlemesi, sosyal medya platformlarının gelişmesine ve geniş kullanıcı kitlelerine ulaşmasına olanak sağlamıştır. Bu platformlar aracılığıyla insanlar hem birbirleriyle iletişim kurabilir hem de toplumsal olaylar, ürünler veya konular hakkında ortak başlıklar altında bir araya gelerek duygu ve düşüncelerini paylaşabilirler. Sosyal medya paylaşımları, duygu analizi çalışmaları için değerli bir veri kaynağı oluşturur. Bu çalışmalar sayesinde paylaşılan veriler, analiz edilerek belirli bir konu hakkında olumlu, olumsuz veya tarafsız duygu ifadeleri tespit edilebilir.

Bu çalışmada sosyal medya platformu Twitter‘da 1 Ocak 2020 ve 31 Aralık 2021 arasında paylaşılan çeşitli konulu 30000 adet Türkçe tweet toplanmıştır. Ham veri setinin etiketleme aşamasında 10.853 adet Türkçe tweet kullanılmıştır. Her bir tweet, olumlu veya olumsuz olarak etiketlenmiş ve iki sınıflı bir veri seti hazırlanmıştır. Oluşturulan veri seti, ön işleme tabi tutulmuş ve ayrıca NLTK kütüphanesi ile normalleştirilerek işlenebilir bir hale getirilmiştir.

Sınıflandırma aşamasında girdi olarak kullanılacak veri seti için manuel etiketleme işlemi yapılarak modeller oluşturulmuştur. Bu modeller TF-IDF ve CountVectorizer veri sayısallaştırma yöntemleri ile ANN, CNN, LSTM, Embedding derin öğrenme metodları ve Naive Bayes, SVM, XGBOOST, CATBOOST makine öğrenmesi algoritmaları kullanılarak sınıflandırma modelleri üzerinden yapılan paylaşımların duygu analizi gerçekleştirilmiştir. Türkçe metinlerin manuel etikete sahip olduğu ve tarafsız olarak işaretlenmiş metinler veri setinden çıkarıldığında başarı oranının arttığı çıkarımına varılmıştır.

Makine öğrenmesi algoritmaları arasında TF-IDF – SVM ile 0.844‘lük, derin öğrenme algoritmaları arasında LSTM ile 0.841’lik oranla en iyi sonucu verdiği görülmüştür. Araştırmanın sonuçları, yapay zeka algoritmalarıyla yapılan duygu analizi açısından önemli bilgiler sunmakla kalmayıp, gelecekte çeşitli sektörlerde uygulanabilirliği açısından da büyük değer taşımaktadır.

**Anahtar Kelimeler:** Duygu Analizi, Doğal Dil İşleme, Makine Öğrenmesi, Derin Öğrenme, Twitter

# **ABSTRACT**

The advancement of technology has led to the development of social media platforms, reaching large user bases. Individuals can communicate with others on social media platforms, and in the face of societal events, they can come together under a common topic to share their feelings and thoughts about a product or an issue. These shares create a vast data source that can be used in many areas for sentiment analysis studies. By processing and analyzing this data through sentiment analysis studies, positive, negative, or neutral sentiment expressions regarding the relevant topic can be determined.

In this study, 30,000 Turkish tweets on various topics shared on the social media platform Twitter between January 1, 2020, and December 31, 2021, were collected. During the labeling phase of the raw dataset, 10.853 Turkish tweets were used. Each tweet was labeled as either positive or negative, and a binary-labeled dataset was prepared. The created dataset was preprocessed and further normalized using the NLTK library to make it processable.

Manual labeling was performed on the dataset, which would be used as input in the classification phase, to create models. These models were used to perform sentiment analysis on the shares through classification models using TF-IDF and CountVectorizer data digitization methods with ANN, CNN, LSTM, Embedding deep learning methods, and Naive Bayes, SVM, XGBOOST, CATBOOST machine learning algorithms. It was inferred that the success rate increased when neutral-marked texts were removed from the dataset where Turkish texts had manual labels.

Among the machine learning algorithms, TF-IDF - SVM gave the best result with a rate of 0.844, and among the deep learning algorithms, LSTM gave the best result with a rate of 0.841. The results of the research contain significant information in terms of sentiment analysis with artificial intelligence algorithms and are valuable for potential applications in different sectors in the future.

**Keywords:** Sentiment Analysis, Natural Language Processing, Machine Learning, Deep Learning, Twitter

**1.GİRİŞ**

Teknolojinin hızla gelişmesiyle birlikte internet ortamı ve sosyal medya platformları da önemli bir çeşitlilik ve büyüme göstermiştir. Özellikle geniş bir kullanıcı kitlesine sahip olan Twitter artık sadece iletişim aracı olmaktan öte; sağlık, siyaset, spor, teknoloji, eğitim, e-ticaret ve sinema gibi çeşitli alanlarda insanların görüşlerini paylaştığı bir platform haline gelmiştir. Bu platformlar; bir ürün, etkinlik veya konuyla ilgili toplumun genel görüşlerini yansıtan önemli bir veri kaynağı olarak da değer kazanmıştır.

Özellikle Twitter gibi sosyal medya platformları üzerinde paylaşılan Tweet'ler, genellikle halkın belirli bir ürün, etkinlik veya konu hakkındaki düşüncelerini yansıtmaktadır. Bu platformlardaki içerikler, büyük miktarda yapılandırılmamış veriyi içerir ve bu veriler, işletmelerden kamuoyuna kadar geniş bir kitle için önemli bir bilgi potansiyeli taşır. Bu verilerin analiz edilmesi; sağlık hizmetlerinden pazarlama stratejilerine, ticaretten güvenliğe kadar birçok alanda fırsatlar sunar. Her bir paylaşım; kullanıcının belirli bir konu, ürün, hizmet veya olay hakkındaki duygu, düşünce ve tutumunu yansıtır.

Bu nedenle duygu analizi çalışmaları, sosyal medya verilerinin değerini ortaya çıkarmak ve anlamak için önemli bir araç haline gelmiştir. Bu çalışmalar, bir ürün veya konuyla ilgili paylaşımların olumlu, olumsuz veya tarafsız olduğunu belirlemeyi amaçlar ve bu bilgiler, şirketlerin pazarlama stratejilerini geliştirmesinden siyasi partilerin seçmenler hakkında bilgi edinmesine kadar birçok alanda kullanılabilir.

Duygu analizi; politika ve kamuoyu araştırmaları, müşteri hizmetleri, sosyal medya monitörü, finansal piyasalar, insan kaynakları ve işe alım, pazar araştırması, sağlık hizmetleri gibi birçok alanda kullanılabilecek bir tekniktir. Bu alanlarda nasıl kullanılabileceği ise aşağıda sıralanmıştır:

**Politika ve Kamuoyu Araştırmaları:** Duygu analizi politika ve kamuoyu araştırmalarında kullanılabilir. Duygusal analiz, halkın politikacılara veya politikalara nasıl baktığını belirlemek için kullanılır.

**Müşteri Hizmetleri:** Müşteri memnuniyetini artırmak için müşteri hizmetleri personeli duygu analizi kullanabilir. Müşteri hizmetleri çağrıları, e-postalar ve sosyal medya mesajları bu kategoriye girer. İletişim kanalları aracılığıyla alınabilecek geri bildirimleri kapsar.

**Sosyal Medya Monitörü:** Duygu analizi, bir markanın, ürünün veya hizmetin sosyal medyada nasıl algılandığını izlemek ve anlamanıza yardımcı olabilir. Bu, bir markanın imajını ve pazarlama planlarını yönetmesine yardımcı olabilir.

**Finansal Piyasalar:** Finansal piyasaları anlamak ve öngörmek için duygu analizi kullanılabilir. Hisse senedi fiyatları ve diğer finansal göstergeler, yatırımcıların duygusal durumları ve piyasa hakkındaki görüşleri tarafından büyük ölçüde etkilenir.

**İnsan Kaynakları ve İşe Alım:** İnsan kaynakları ve işe alma süreçlerinde duygu analizi de kullanılabilir. Örneğin, iş başvurusu metinlerini inceler veya çalışanların motivasyonunu ve iş yerindeki duygusal durumlarını değerlendirir.

**Pazar Araştırması:** Pazar araştırmasında duygu analizi kullanılabilir. Örneğin, bir işletme, bir ürünün veya hizmetin piyasadaki algısını belirlemek için tüketicilerin görüşlerini ve incelemelerini inceleyebilir.

**Sağlık Hizmetleri:** Hastaların duygularını ve duygusal durumlarını anlamak için duygu analizi kullanılabilir. Bu, hastaların tedavileri hakkındaki görüşlerini anlamak için kullanılabilecek online incelemeler ve geri bildirimler içerir.

Duygu analizi, insanların yazdıkları metinlerden bu duygu, düşünce ve tutumlarını ortaya çıkarmayı hedeflemektedir [1]. Duygu sınıflandırma ise olumlu, olumsuz veya belirsiz gibi önceden belirlenen duygu kategorilerine göre metin verilerinin sınıflandırması alt işlemlerinden oluşmaktadır [2]. Tweet verileri üzerinde duygu analizi yapmak, diğer veri kümelerine göre daha zordur [3]. Twitter'in karakter sınırlaması nedeniyle metinler yoğunlaştırılmış anlamlar taşır. Tweet’ler karakter kısıtlamasından dolayı emoji ve özel kullanımlar içermektedir.

Bu doğrultuda Twitter‘de yer alan çeşitli konulardaki Türkçe paylaşımlar tez kapsamında ele alınarak makine öğrenmesi ve derin öğrenme algoritmaları ile duygu analizi gerçekleştirilmiştir.

Duygu analizi uzun zamandır çalışmalara konu olan bir alan olmuştur. Bu anlamda çalışmanın ikinci bölümünde daha önce yapılan ve literatürde yer almış Türkçe ve İngilizce çalışmalardan bahsedilmiştir.

Üçüncü bölümde çalışma kapsamında kullanılan yöntem ve analiz sürecinin adımları anlatılmış ayrıca Twitter ve Twitter verilerine erişimde kullanılan Twitter uygulama programlama arayüzü (Application Programming Interface-API) hakkında bilgilere yer verilmiştir.

Dördüncü bölümde deneysel çalışmalar ile sistem mimarisi üzerinden uygulanan adımlar tek tek ele alınmıştır. Twitter‘den elde edilen veriler literatürde yer alan birçok veri ön işleme adımlarına tabi tutularak işlenebilir bir hale getirilmiştir. Ayrıca veri ön işleme aşamasında kullanılan kütüphaneler ile veri setinde yer alan Türkçe kelimeler için yazım denetimi ve kelime tahmini ile normalleştirme işlemine bu bölümde yer verilmiştir.

Dördüncü bölümün devamında manuel olarak etiketlenmiş veri setleri ile oluşturulan model, farklı sayısallaştırma yöntemleri ile temsil edilerek makine öğrenmesi ve derin öğrenme algoritmalarına girdi olarak kullanılmış ve en iyi performans gösteren modele karar verilmiştir.

**2. LİTERATÜR TARAMASI**

Yerli ve yabancı literatür incelendiğinde sosyal medya üzerindeki verilerden yararlanılarak yapılan birçok çalışmaya rastlanmıştır. Bu kısımda makine öğrenmesi ile Twitter verilerinin duygu analizi konusu ile ilgili bazı literatür çalışmalarına yer verilmiştir.

Osmanoğlu vd.(2020) öğrenme teknikleri kullanarak materyalin kişiler üzerindeki etkisini analiz etmiş ve bu etkileri pozitif, negatif veya nötr olarak sınıflandırmıştır. Bu sayede, negatif geri dönüş alan makalelerin geliştirilmesine yönelik öneriler sunmayı amaçlamışlardır. Çalışmada, ilk olarak metin ön işlemede adımlarından, metnin sadeleştirilmesi yazım denetimi, durak kelime çıkarma uygulanmıştır. Makine öğrenmesi, denetimli öğrenme tekniklerinden olan; DT, MLP, XGB, SVM, MLR, Gaussian NB ve KNN algoritmaları kullanılmıştır. Eğitim için, 6059 adet etiketli veri kullanılmıştır. Yüzde 77.5 başarı oranıyla LR algoritması en başarılı sınıflandırma algoritması olmuştur [4].

Pervan ve Keleş (2017) çalışmalarında e-ticaret sitelerinden toplanan müşteri yorumları üzerinde kelime modellerini Word2Vec algoritması ile oluşturarak, RF sınıflandırma algoritması 18 ile ürün inceleme verileri üzerinde ikili (pozitif/negatif) sınıflandırma yapmışlardır. Veri ön işleme adımlarında, metin sadeleştirme, kelime bölütleyici ve yazım denetimi işlemlerinden sonra öznitelik seçimi için Word2Vec modeli uygulanmıştır. Bu çalışmada yüzde 84.23 doğruluk oranı elde edilmiştir [5].

Yurtalan ve arkadaşları (2019), Türkçe Tweet’ler için sözlük tabanlı duygu analizi gerçekleştirmişlerdir. Bu sözlük, pozitif ve negatif kelime köklerinin POS etiketleriyle etiketlenmesiyle oluşturulan 1181 veri öğesi ve bunların polarite değerlerinden oluşmaktadır. Türkçe tweet veri setini kelime, kelime grubu ve deyim/atasözü seviyelerinde inceleyerek, Türkçe metinlerin sözlük tabanlı analizinde kelime grubu seviyesindeki analizin en etkili yöntem olduğu ve sistem performansını geliştirdiği tespit edilmiştir [6].

Karçıoğlu ve Aydın [7], İngilizce ve Türkçe Twitter veri kümelerinden BoW ve Word2Vec yöntemleriyle çıkartılan özellikleri kullanarak Tweet’leri MÖ yöntemleriyle sınıflandırmışlar. Çalışmalarında, kelime köklerinin alınmasının duygu sınıflandırma performansı üzerindeki etkisi incelenmiştir. Sonuç olarak kök alma işlemi, hem Türkçe hem de İngilizce veri kümeleri için önerilmemiştir.

Rumelli ve arkadaşları [8], çalışmalarında e-ticaret ürün yorumlarını kullanmış ve bir duygu analizi modeli geliştirmişlerdir. Hepsiburada.com e-ticaret sitesinden 272 bin ürün yorumu çekilerek bir veri seti oluşturulmuştur. Bu veri setinde, 1 ve 2 yıldızlı yorumlar olumsuz, 3 yıldızlı yorumlar nötr, 4 ve 5 yıldızlı yorumlar ise olumlu olarak etiketlenmiştir. Çalışmada kullanılmak üzere 13 bin olumlu ve 13 bin olumsuz yorum seçilerek veri seti hazırlanmıştır. NB, RF, SVM ve K-En Yakın Komşu (K-Nearest Neighbors – KNN) algoritmaları kullanılarak analizler gerçekleştirilmiş ve çalışma sonucunda ortalama %73 doğruluk oranı elde edilmiştir.

Üçükkartal [9], 2020 yılında Twitter’de paylaşılan İngilizce tweetler üzerinde duygu analizi ve metin şekilleştirme yapmıştır. Çalışmada Twitter API üzerinden veriler çekilmiş, veriler ön işlemden geçirilerek TexBlob kütüphanesi ile veriler pozitif, negatif ve nötr olarak etiketlenmiştir. WordCloud kütüphanesi ile kelimeler şekilleştirilmiştir.

İlhan ve Sağaltıcı [10], daha önceden sınıflandırılmış tweet verileri üzerinde farklı makine öğrenmesi teknikleri uygulayarak duygu analizi çalışması yapmışlardır. Olumlu ve olumsuz duygu analizi için N-gram tekniği kullanılmış, NB, DVM gibi yöntemlerden faydalanılarak sınıflandırıcıların performans karşılaştırılması yapılmıştır. Sonuç olarak, sınıflandırıcılardan en yüksek değeri Destek Vektör Makinelerinin verdiği tespit edilmiştir.

Tuzcu [11], çevrimiçi bir kitap satış sitesinden elde ettiği kullanıcı yorumları üzerinden sınıflandırma algoritmaları ile duygu analizi gerçekleştirerek algoritmaların sınıflandırma başarısını ölçmeyi amaçlamıştır. Veri setini 47 adet kitap için yazılmış yorumlar arasından duygu sınıflandırılması yapılmış toplam 1400 yorum oluşturmuştur. Yorum sınıflandırması pozitif ve negatif olarak ikili kutup ile yapılmış, ilk önce Python dilinde Çok Katmanlı Algılayıcı daha sonra da RapidMiner programında Naive Bayes, Destek Vektör Makineleri ve Lojistik Regresyon algoritmaları kullanılarak gerçekleştirilmiştir. Sonuç olarak Çok Katmanlı Algılayıcı çalışma kapsamında en iyi sonucu veren algoritma olarak belirlenmiştir.

Beşkirli ve arkadaşları [12] yapmış oldukları çalışmada, Twitter uygulamasında kullanıcıların aşı duyurusu öncesi ve sonrasında yazılmış olan tweet’ler üzerinde duygu analizi yapılarak aralarında duygu farkı olup olmadığına bakılmıştır. Çalışma sonucunda aşı duyurusu öncesinde olumlu ve olumsuz cümleler birbirine yakın iken, aşı duyurusundan sonra olumlu cümlelerde büyük bir artış olup olumsuz cümlelerde ise azalma olduğu tespit edilmiştir.

Atılgan ve Yoğurtçu [13], 2021 yılında yapmış oldukları çalışmalarında bir kargo firmasının Twitter üzerindeki yorumlarını incelemiş ve cümlelere pozitif, negatif ve nötr duygu analizi sınıflandırmaları uygulamışlardır. 1138 tweet verisi üzerinden yapılan çalışmanın sonucunda en yüksek paya sahip duygu sınıfının negatif olduğu ortaya çıkmış ve en çok kullanılan kelimelerin “engel”, ”uzak” ve ”kal” olduğu tespit edilmiştir.

Eyipınar ve arkadaşları [14], Youtube platformundaki sporcu beslenmeleriyle alakalı 6 video belirleyip bu videoların Türkçe paylaşılmış yorumlarından elde ettikleri veri seti üzerinden sözlüğe dayalı yöntemle duygu analizi çalışması yapmışlardır. Elde edilen veri seti alanında uzman 2 kişi tarafından değerlendirilerek olumlu, olumsuz ve nötr olarak etiketlenmiştir, verilerin ön işleme adımları Orange 3 programıyla yapılmıştır. Araştırma sonucunda yorumların pozitif, negatif ve nötr yüzdeliklerinin sırasıyla %27.62, %17.3 ve %55.08 olduğu tespit edilmiştir.

2021 yılında Kumaş [15], çalışmasında Türkçe Twitter verilerine metin madenciliği yöntemleri uygulayarak duygu analizi yapmıştır. Naive Bayes (NB), KNN, Destek Vektör Makinesi, Lojistik Regresyon (LR) ve Karar Ağacı sınıflandırma algoritmalarını kullanarak pozitif ve negatif olarak sınıflandırmıştır. Sınıflandırma sonuçlarını f1 skoru ile değerlendirmiş ve sırasıyla NB, KNN, Destek Vektör Makineleri, LR ve Karar Ağacı sınıflandırıcılarıyla elde ettiği f1 skorları %70, %65, %73, %71 ve %69 olarak bulmuştur.

Temizhan ve Mendeş [16], 2021 yılında yayımladıkları çalışmalarında Covid-19 pandemisi ile ilgili Twitter uygulamasında paylaşılan 40.000 Türkçe tweeti metin madenciliği teknikleri ile incelemiş ve duygu analizlerini yapmışlardır. R yazılım dilinde Twitter API aracılığı ile çekilen veriler gerekli ön işlemden geçirilip temizlendikten sonra Syuzhet kütüphanesi aracılığıyla 10 adet duygu skoruna göre duygu sınıflandırması hesaplanmıştır. Araştırma sonucunda kişilerin covid- 19 kelimesinden bahsederken negatif duygular içeren metinler paylaştıkları görülmüştür.

Uyaroğlu, Akdeniz ve Cebeci [17], 2021 yılında Sakarya Büyükşehir Belediyesi ve bazı ilçe belediyeleri hakkında Twitter kullanıcılarının paylaşmış olduğu tweetleri toplayarak makine öğrenmesi ve derin öğrenme yöntemlerinden Turkish Bert yöntemini kullanarak bir duygu analizi belirleme çalışması yapmışlardır. Twitter platformundan en çok takipçisi olan 5 ilçe belediyesini kapsayan ve Türkçe atılmış tweetler Python dilinde çekilmiş ve veri ön işleme adımlarına tabi tutulmuştur. %20’si test, %80’i eğitim olmak üzere örnek veri seti oluşturulmuş ve Naive Bayes, Decision Tree, Gradient Boosting ve Random Forest makine öğrenmesi algoritmaları kullanılmış, en iyi sonucu Naive Bayes algoritmasının verdiği tespit edilmiştir. Çalışma sonucunda belediyelerin vatandaşlardan gelen olumlu ya da olumsuz geri bildirimleri görüp iletilen taleplere daha hızlı cevap verebilmeleri sağlandığı belirtilmiştir.

Karaahmetoğlu ve arkadaşları [18], korona virüs salgını ile ilgili paylaşılan 4.575 adet Türkçe paylaşım üzerinden duygu analizi çalışması yapmış ve sözlük tabanlı bir yöntem kullanmışlardır. Çalışma kapsamında Python dilinde Twitter uygulamasında kullanıcılar tarafından paylaşılan “korona” etiketli yorumlar toplanmış ve veri ön işleme adımlarından geçirildikten sonra SentiTurkNet duygu kelime sözlüğü kullanılarak metinler olumlu, olumsuz ve tarafsız olarak etiketlenmiş, aynı zamanda veriler test ve eğitim verisi olarak ayrıldıktan sonra Bert temelli algoritma ile duygu analizleri gerçekleştirilmiştir. Çalışma sonucunda iki yöntemde de olumlu duyguların olumsuz duygulara göre 7 kat fazla çıktığı tespit edilmiştir.

Köksal ve arkadaşları tarafından yapılan çalışmada [19], Twitter platformunda Bitcoin ile ilgili paylaşılan Türkçe yorumlar Python dili kullanılarak çekilmiş ve yorumlar pozitif, negatif, nötr olarak ayrıştırılarak etiket bulutunda toplanmıştır. Naive Bayes ve Lojistik Regresyon algoritmaları kullanılarak model başarı oranları test edilmiş ve sırasıyla %75,53 ile %72,19 bulunmuştur. Ardından günlük toplam pozitif tweet oranıyla Bitcoin günlük açılış değerleri birlikte kullanılmış ve Doğrusal Regresyon, Rastgele Orman Regresyon yöntemleriyle modeller oluşturularak Bitcoin kapanış değeri tahminlemesi yapılmıştır. Sonuç olarak sırasıyla 𝑟² değeri %88,97 ve %94,16 değerleri bulunmuştur.

Dikkaya Kavak 2022 yılında yapmış olduğu tez çalışmasında [20], 2011-2021 tarihleri arasında Twitter platformunda “torku” etiketi ile yazılmış olan tweet’leri netnografi ve metin madenciliği yöntemiyle incelemiştir. BERT modeli ile duygu analizleri yapılmış, kelimeler pozitif ve negatif olarak ayrıştırılmıştır. Çalışma sonucunda olumlu, olumsuz ve bütün kategorilerde en çok tekrar eden kelimeler tespit edilmiştir.7212 kelimeden 4283 tanesi pozitif ve 2929 tanesi de negatif olarak tespit edilmiştir.

Dinçer ve arkadaşları [21], 2022 yılında siber zorbalığın tespiti için Twitter’da paylaşılan 3200 adet tweet Twitter API kullanılarak çekilmiş, kullanıcıların paylaşmış olduğu yorumlarda siber zorbalık olup olmadığı araştırılmıştır. Analiz sürecinde yapay zekâ yöntemlerinden Destek Vektör Makinesi, Naive Bayes, Lojistik Regresyon kullanılmış, performans analizi için f1-skor, doğruluk, hassasiyet ve kesinlik değerleri kullanılmış, analiz sonucunda en yüksek değeri %87 ile LR algoritması verdiği gözlemlenmiştir.

Özoran’ın 2022 yılında yaptığı çalışmada [22], Twitter’da Dünya Sağlık Örgütü tarafından paylaşılan 1 ay süre zarfındaki tweet’leri incelenmiş, bu çalışmada DSÖ’nün paylaşımlarının hangi kategoride olduğu, etkileşim oranı, duygu analizi gibi konular tespit edilmeye çalışılmıştır. Çalışmada hibrit yöntem kullanılmış olup içerik analizi ile 7 farklı ana kategori belirlenmiş ve kategorilerin duygu analizi yapılmıştır. Çalışma sonucunda Covid19 ve bilgilendirme kategorileri negatif duygu sınıfında, teşekkür ve misyon gibi diğer kategoriler ise pozitif duygu sınıfında olduğu ortaya çıkmıştır.

Günyaktı ve Bursa [23], 2022 yılında covid-19 pandemi döneminde sosyal medya kullanıcılarının öğretmenler ve sağlık personelleri hakkında paylaşmış oldukları gönderilerin duygu analizlerini belirlemek amacıyla çalışma yapmışlardır. Twitter kullanıcılarının paylaşmış olduğu Türkçe tweetler veri seti olarak kullanmışlardır. Veriler ön işlemlerden geçirildikten sonra duygu analizi için sözlük tabanlı yaklaşım uygulanmış ve SWNetTR++ sözlüğü kullanılmıştır. Çalışma sonucunda paylaşılan gönderilerin öğretmenler için %70’i pozitif, sağlık çalışanları için %61’inin pozitif olduğu tespit edilmiştir.

Torun ve Şengül [24], Twitter platformundaki kullanıcıların kripto para ile ilgili paylaşımları üzerinden bir duygu analizi çalışması yapmıştır. “Kriptopara” etiketi ile paylaşılmış Türkçe tweetler R programlama dili ile çekilmiş, polarite ağırlıkları hesaplanarak duygu analizleri bulunmuştur. Çalışma sonucunda duygu polaritesi sıfıra yakın olduğundan dolayı duygu analizi nötr olarak tespit edilmiştir.

Turan ve arkadaşları [25], Youtube ve Twitter platformlarındaki Türkçe paylaşımları Python dili vasıtasıyla çekmişler, bu veri seti üzerinden duygu analizi çalışması yapmışlardır. Elde edilen veri seti ön işlemden geçirildikten sonra gönüllü kişilerce elle olumlu ve olumsuz etiketlemesi yapılmış ardından her veri setinden eşit sayıda veri alınıp birleştirilmiştir. Duygu analizinde makine öğrenmesi algoritmalarından Naive Bayes, Lojistik Regresyon, Destek Vektör Makineleri ve Rassal Orman sınıflandırıcıları kullanılmıştır. Araştırma sonucunda Youtube platformunda olumlu görüş fazlayken Twitter uygulamasında olumsuz görüşün daha fazla olduğu görülmüş, aynı zamanda en başarılı algoritmanın %88 ile Naive Bayes olduğu belirtilmiştir.

Demirci ve Kömeçoğlu [26], WhatsApp uygulamasının gizlilik sözleşmesini güncellediğini duyurmasından sonra kullanıcıların Twitter platformunda bu konu ile ilgili paylaşmış oldukları yorumlarını toplayarak bir duygu analizi çalışması yapmışlardır. Bu kapsamda NodeXL yazılımı paketi ile veriler çekilmiş ve ön işlemeye tabi tutularak veriler temizlenmiştir. Duygu analizi için BERTürk dil modeli kullanılmış, veriler pozitif ve negatif olarak etiketlenmiştir. Araştırma sonucunda yorumların %60’ı olumlu ve %30’unun ise olumsuz olduğu tespit edilmiştir.

Sel tarafından 2022 yılında yapılan çalışmada [27], Twitter platformunda koronavirüs salgını ile ilgili dört farklı anahtar kelime seçilerek Türkçe veriler toplanmış, veriler ön işlemden geçirildikten sonra Google E-tablolar yardımıyla İngilizceye çevrilmiş ardından TexBlob kütüphanesi yardımı ile cümlelerin pozitif ve negatif olma durumlarına göre etiketleme işlemi yapılmıştır. Kelime sıklıklarının tespiti için n-gram yöntemi uygulanmış, araştırma sonucunda pozitif duygu içeren ifadelerde; hastane, yoğun bakım, test gibi kelimeler, negatif duygu içeren kelimelerde ise; ev, salgın, kayıp edilen insan gibi başlıkların görüldüğü tespit edilmiştir.

Şahinaslan ve arkadaşları [28], çok dilli duygu analizi çalışması yapmışlardır. Youtube uygulamasında kullanıcıların farklı dillerde paylaşmış oldukları yorumlar PHP dili ile toplanmış, ardından veri ön işleme adımlarına tabi tutulmuştur. Ön işlemden geçirilen veri seti Google API aracı ile İngilizce diline çevrilmiş, Naive Bayes algoritması kullanılarak pozitif, negatif ve nötr olarak sınıflandırılmıştır. WEKA programı ile yapılan doğru sınıflandırma oranı %65,56 olarak bulunmuştur.

Kandıran ve arkadaşları [29], Covid-19 pandemi sürecinde Twitter’de uzaktan eğitim ile ilgili paylaşılmış Türkçe yorumları, 28 etiket üzerinden Python dilinde çekerek veri setini oluşturmuşlardır. Duygu analizi için makine öğrenmesi yaklaşımlarından BERTürk modeli kullanılmış, verileri pozitif ve negatif olarak sınıflandırılmıştır. Çalışma sonucunda kullanıcıların pozitif paylaşımlarda bulundukları değerlendirilmiştir.

Akdeniz [30], uzaktan eğitimle ilgili Twitter verileri üzerinde yapmış olduğu duygu analizi çalışmasında duygu durumlarını belirlemek için manuel etiketleme, TextBlob, Vader ve Bert modellerini kullanmıştır. TF-IDF, Word2Vec gibi farklı sayısallaştırma yöntemleri ile LR, NB gibi farklı makine öğrenmesi yöntemleri kullanılarak duygu analizleri yapılmıştır. Manuel etiketleme yapılan Türkçe verilerde en iyi başarı oranı 0,79 ile TF-IDF ve LR ikilisi ile elde edilmiştir.

Zhao ve Cao [31], önerdikleri MÖ tabanlı modelde n-gram, anlamsal ve polarite skoru özellikleri birleştirerek daha iyi sınıflandırma başarımı elde etmişlerdir.

Çoban ve diğerleri [32], Türkçe tweet’lerden kelime çantası ve n-gram özellikleri çıkarıp, MÖ yöntemleriyle olumlu veya olumsuz olarak sınıflandırmışlardır.

Siddiqua ve diğerleri [33], kural tabanlı bir sınıflandırıcıyı, denetimli sınıflandırıcıların çoğunluğuna dayalı bir oylama topluluğu ile birleştirmişlerdir.

Chamansingh ve Hosein [34], her tweet için bir kelime skoru elde etmiş ve cümlenin skoruna göre tweet’leri duygu bakımından sınıflandırmışlardır.

Jianqiang ve Xiaolin [35], altı farklı Twitter veri kümesi üzerinde, veri önişlemlerin sınıflandırma başarımına olan etkisini incelemişlerdir. Her bir veri seti için n-gram ve sözlük tabanlı özellikler elde etmiştir ve MÖ teknikleriyle sınıflandırmışlardır.

Riyadh ve diğerleri [36], önişlemden geçirilen Twitter veri kümesinden unigram ve unigram ile sözcük etiketleme özelliklerini çıkarmışlardır.

Multinomial Naive Bayes (MNB) algoritması kullanılarak, beş farklı duygu sınıfı (mutlu, üzgün, nefret, şaşırmış ve belirsiz) doğrultusunda sınıflandırma yapılmıştır.

Tsapatsoulis ve Djouvas [37], Twitter paylaşımlarından manuel özellik çıkarımının, klasik metin özellik çıkarma yöntemlerine kıyasla performansını analiz etmişlerdir. Manuel özellik çıkarımının n-gram modellerine göre daha iyi sınıflandırma sağladığı sonucuna varmışlardır.

Kaur ve diğerleri [38], tweet’lerden n-gram yöntemiyle özellik çıkarımı yapmış ve tweet’leri K-En Yakın Komşu algoritmasıyla pozitif, negatif veya belirsiz olarak sınıflandırmışlardır.

Permatasari ve diğerleri [39], Twitter'a özgü özellikler, sözdizimsel özellikler, PoS özellikleri, sözlük tabanlı özellikler ve BoW özelliklerini birleştirerek tweet'leri duygusal açıdan sınıflandırmışlardır. Ancak, yalnızca BoW özellikleri kullanıldığında, birleşik özelliklere kıyasla daha iyi bir performans elde etmişlerdir.

Aydın ve diğerleri [40], tweet’lere önişlem uyguladıktan sonra öznitelik mühendisliği yaparak 10 özellik çıkarmışlardır. Bu özellikler, çoklu popülasyon tabanlı parçacık sürü optimizasyon yöntemi kullanılarak duygu bakımından sınıflandırılmıştır.

Salur ve diğerleri [41], Türkiye’deki turizm merkezleri hakkında yapılan Twitter paylaşımlarından elle özellik çıkarımı yapmış olup, duygu bakımından sınıflandırıldıktan sonra her turizm merkezi hakkında oluşturulan duygu sonuçları mobil uygulama yardımıyla kullanıcılara sunulmuşlardır.

Ahuja ve diğerleri [42], TF-IDF ve n-gram özellik çıkarma yöntemlerinin Twitter veri kümeleri üzerindeki etkisini araştırmışlardır. Yapılan sınıflandırmada TF-IDF yöntemiyle elde edilen özelliklerin n-gram özelliklerine göre daha iyi sonuç verdiği gözlemlenmiştir.

Saad ve Yang [43], TF-IDF özellikleri üzerine bir dengeleme ve puanlama modeli oluşturarak, MÖ yöntemleriyle tweet’leri sınıflandırmışlardır.

Topçu ve Erkaya [44], Müşteri İlişkileri Yönetimi (Customer Relations Management – CRM) yazılımlarında toplanan talep mesajlarının makine öğrenmesi yöntemleri kullanılarak kategorize edilmesi amaçlı bir çalışma yapmışlardır. Çalışmada vatandaşlar tarafından gönderilen 3 bin farklı konudan oluşan 220 bin talep mektubu kullanılmıştır. Veriler ön işlemlerden geçirildikten sonra TFIDF algoritması kullanılarak özellik seçimi yapılmıştır. Veri seti 5 farklı algoritma kullanılarak test edilmiştir. En verimli algoritmanın %85 doğruluk oranıyla LSTM olduğu görülmüştür. Çalışma sonucunda oluşturulan model örnek olarak “1314. sokakta bozuk yollar var. Bir an önce asfalt atılmasını ve yolun onarılmasını talep etmekteyim. Gereğini arz ederim.” şeklinde gelen talebi “Asfalt Bakım Onarım” olarak kategorilendirmiştir.

Mengutayci ve Temurtas [45], Türkçe otel yorumlarının yapay sinir ağları ile sınıflandırılması konulu bir çalışma yapmıştır. Çalışmalarında, çevrimiçi otel rezervasyonu yapılan bir platformdaki müşteri yorumlarını analiz etmişlerdir. Bu yorumların 4-5 yıldız olanları olumlu, 1-2 yıldız olanları olumsuz olarak etiketlenmesini sağlayarak iki sınıflı bir çalışma gerçekleştirmişlerdir. Veri setini gerekli ön işlemlerden geçirdikten sonra TF-IDF modeli ile yorumlar bir matrise dönüştürülmüş ve her terimin sınıflandırmadaki önem derecesinin belirlenmesi sağlanmıştır. Veri seti iki farklı yöntemle eğitim ve test verilerine ayrılmış ve bu veri setlerinin başarısı test edilmiştir. Testlerini yapay sinir ağı mimarileriyle gerçekleştirmişlerdir.

Mayda ve Korkmaz [46], ortak çalışmalarında müşteri yorumlarının sınıflandırılması için sözlük tabanlı bir yaklaşım kullanmışlardır. Bu çalışma kapsamında bir sıfat sözlüğü hazırlanmıştır. Kitap satışının yapıldığı bir e-ticaret platformundan topladıkları okuyucu yorumlarından bir veri seti hazırlamışlar ve sıfat sözlüğünü kullanarak olumlu, olumsuz, nötr sınıflar için bir duygu analizi çalışması yapmışlardır. Ayrıca, çalışmalarında Türkçe diline özgü kurallar geliştirerek çalışmanın başarısını artırmışlardır. Çalışma sonucunda yaklaşık %62’lik doğruluk oranı elde etmişlerdir.

Gezici ve Yanıkoğlu [47], çalışmalarında makine öğrenmesi algoritmalarıyla Türkçe film incelemelerinin duygu analizini gerçekleştirmişlerdir. Bu çalışmalarında denetimli öğrenme ve sözlük tabanlı yaklaşımları birleştirerek yeni bir yöntem sunmuşlardır. Çalışmalarında “beyazperde.com” adlı film sitesinden önceden oluşturulan bir veri seti kullanılmış ve bu veri setindeki incelemelerin 4 ve 5 yıldızlı olanları olumlu, 1 ve 2 yıldız olanları olumsuz olarak etiketleyerek ikili sınıflandırma çalışması yapmışlardır. NB ve SVM algoritmalarının sonuçları karşılaştırılmıştır. Bu karşılaştırma esnasında sözlük tabanlı yaklaşım ile denetimli öğrenme yaklaşımlarının sonuçları da göz önünde bulundurulmuştur. Sonuçlarda NB algoritmasının doğruluk oranının %75 olduğu görülmüştür.

Onan [48], bir çalışmasında denetimli ve denetimsiz terim ağırlıklandırma modellerini kullanarak kapsamlı bir duygu analizi karşılaştırması gerçekleştirmiştir. Veri seti olarak Türkçe Twitter mesajları kullanılmıştır. Mesajlar pozitif ve negatif olarak iki kişi tarafından etiketlenmiş ve gerekli ön işlemlerden geçirilmiştir. Ardından KNN, LR, SVM ve NB algoritmalarıyla test etmişlerdir. Bu testlere AdaBoost, Bagging ve Random SubSpace topluluk öğrenme yöntemleriyle devam edilmiştir. Testler sonucunda denetimli terim ağırlıklandırma modellerinin denetimsiz terim ağırlıklandırma modellerine göre daha başarılı olduğu gözlemlenmiştir.

Demircan ve ark. [49], ise çalışmalarında Türkçe metinler üzerinde duygu analizi yöntemlerini ele almışlardır. Veri seti olarak eticaret ürün yorumlarını kullanmışlar ve veri etiketlenmesi için ürün yorumlarındaki puanlardan faydalanmışlardır. DT, LR, SVM, RF ve KNN algoritmalarıyla modeller geliştirmişler ve test etmişlerdir. Testler sonucunda SVM ve RF modellerinin daha iyi sonuçlar verdiğini gözlemlemişlerdir.

Toçoğlu ve ark. [50], Türkçe tweetler üzerinden duygu analizi çalışması yapmışlardır. Çalışmada sözlük tabanlı bir yaklaşım kullanarak her tweete altı duygu kategorisi için otomatik olarak açıklama eklemişlerdir. Çalışmalarında ANN, CNN, LSTM ve RNN mimarilerini inceleyerek bunlar üzerinden testleri gerçekleştirmişlerdir. Testler sonucunda sözlük tabanlı yaklaşım ile otomatik eklenen açıklamaların etkisini incelemişlerdir. Deneysel çalışmalar sonucu elde ettikleri doğruluk oranlarını karşılaştırdıklarında en yüksek puanı %74 ile CNN mimarisinin verdiğini gözlemlemişlerdir.

Çataltaş ve ark. [51], bir çalışmalarında e-ticaret olumsuz ürün yorumlarından ürünlerin kusurlu özelliklerini bulmak için bir metin analizi yöntemi önermişler ve bulunan ürün kusurlarından da yapılandırılmış bir özet oluşturmaya odaklanmışlardır. Amazon.com üzerinden ayakkabı kategorisine ait 10 bin olumsuz yorum çekerek bir veri seti oluşturmuşlardır. Ön işlemlerin ardından veri sayısı 6190’a inmiş ve çalışmayı bu verilerle gerçekleştirmişlerdir. Ürün kusurlarını bulmak için DBSCAN algoritması kullanılarak yorumlar kümelere ayrılmıştır. Bulunan her kusur daha önceden tanımlanmış Part-of-Speech (Konuşma Parçası) kalıplarını arayarak kendisinin görüş sözcüklerini bulmak için kullanılmıştır. Bulunan kusurlar ve onların görüş sözcükleri kullanılarak yapılandırılmış bir özet oluşturulmuştur.

Cheong vd. [52], 2010-2011 Avustralya sel baskınları sırasındaki önemli olayları kümelemek için sosyal medya metriklerinden yararlanmışlardır.

Mandel vd. [53], Irene kasırgası esnasında atılan tweet’ leri makine öğrenmesi teknikleri ile sınıflandırarak, Twitter mesaj sayılarının kasırganın şiddetlendiği dönemler ile korele olduğunu ve mesajlardaki kaygı düzeyinin coğrafik konuma bağlı olarak değiştiğini ortaya koymuşlardır.

Caragea vd. [54], Sandy kasırgası sırasında Twitter’ da kullanıcı mesajlarının duygu sınıflandırmasını gerçekleştirerek, duyguları kasırga merkezli bir coğrafik haritada Şekilleştirmişlerdir.

Guevara vd. [55], daha önce meydana gelen felaketlerden yola çıkarak oluşturdukları veriler ile alan uyarlaması yoluyla yeni bir afet konusunda atılan tweet’leri sınıflamaya yönelik bir çalışma önermişlerdir.

Imran vd. [56], makine öğrenmesi teknikleri kullanarak afetler hakkında atılan tweet’leri dikkat, tavsiye, hasar, yardım gibi kategorilere sınıflandırmışlardır.

Ashktorab vd. [57], Twitter’ın kurulduğu 2006 yılından itibaren 2014 yılına kadar Kuzey Amerika’da gerçekleşen 12 farklı doğal felaket ile ilgili tweet’leri analiz etmişlerdir. Müdahale ekiplerinin çalışmalarına yardımcı olacak çıkarımlar yapmak için sınıflandırma ve kümeleme yöntemlerinin bir kombinasyonunu kullanmışladır.

Kireyev vd. [58], konu tabanlı kümeleme ve Şekilleştirme, kavram seçimi, ağırlıklandırma ve dinamik korpus arıtımı ismini verdikleri yeni bir teknik kullanarak sosyal medyada paylaşılan felaket haberleri üzerinde sınıflama yapmışlardır.

Vieweg vd. [59], Twitter kullanımındaki artış ve platformdan bağımsız şekilde hızlı iletişim olanağı sunmasının afet mağdurları ile yardım ekipleri arasındaki etkileşimi kolaylaştırdığını bildirmişlerdir.

Kumar vd. [60], afetler sırasında acil müdahaleye yardımcı olmak için tweet’ler üzerinden konum ve belirli anahtar kelimelerin sınıflandırılmasını gerçekleştirmişlerdir.

Go ve ark. çalışmalarında [61], 1-gram, 2-gram ve n-gram gibi farklı veri temsili yapılarıyla, Naive Bayes, MAxEnt ve Destek Vektör Makinaları (DVM) makine öğrenmesi sınıflandırıcılarını kullanarak duygu analizi gerçekleştirmiştir. Çalışmalarında Destek Vektör Makinalarının diğer sınıflandırıcılardan daha iyi sonuçlar verdiğini bildirmişlerdir.

Barkur ve Vibha [62], Hindistan’da Covid-19 haberlerine ilişkin topladıkları 2400 tweetten oluşan bir veri setini incelemişlerdir. Bu çalışmada, R programla dili kullanarak paylaşılan tweetlerin duygu analizini gösteren kelime bulutu çıkartılmıştır.

Li ve ark. [63] ise Covid-19’un insan davranışı üzerindeki psikolojik etkisine odaklanmıştır. Covid-19 haberleri nedeniyle insanların gergin olduğunu ve depresyon düzeylerinin arttığı bildirilmiştir.

Süral ve ark. [64], sürekli duygusal zekâ olarak tanımladığı zekânın sorunlu sosyal medya kullanımı ile ilgili olduğunu ve doğrudan doğruya sosyal medyada popüler bir taraf sunmak ve zaman geçirme güdüleriyle ilişkili olduğunu göstermiştir. Bu çalışma, sürekli duygusal zekâ puanı düşük olan bireylerin gerçek hayattaki sorunlarıyla başa çıkmak için sosyal medya platformlarını bir başa çıkma stratejisi olarak kullandıklarını savunmuştur.

Xue ve ark. [65], 1 Mart 2020’den 21 Nisan 2020’ye kadar 25 farklı etiket kullanarak Covid-19 ile ilgili atılan twitleri toplamışlardır. Çalışma, 4 milyon tweetten oluşan veri seti üzerinde gerçekleştirilmiştir. Gizli Dirichlet Ayırımı (GDA) algoritmasını kullanarak tweetlerdeki popüler konu başlıklarını tespit etmişlerdir.

Feng ve Zhou [66], Twitter’da paylaşılan tweetlerdeki duyguları incelemek amacıyla tekli ve çoklu etiketlere odaklanan BERT modeli kullanmışlardır. Önerilen modelin farklılığı, duyguları ifade etmede etkili olan emojiler dikkate alınmaktadır.

Hornung ve ark. [67], duygusal zekâ ve Facebook kullanımı arasında orantılı bir ilişki olduğunu savunan bir çalışma sunmuştur. Önerilen çalışmada farklı yaş gruplarından oluşan bir veri seti toplanmıştır. Bu veri seti üzerinde yapılan duygu analizi çalışmaları sonucu duygusal zekâ ve Facebook kullanımı arasındaki ilişkinin veri setindeki genç grup için olumlu yaşlı grup için olumsuz olduğunu belirtilmiştir.

Depoux ve ark. [68], koronavirüs salgınıyla mücadelede ortaya çıkan sosyal medya paniğiyle ilgili alınması gereken bir takım önlemleri tartışmıştır. Birçok toplumun maruz kaldığı stresli karantina günlerinde artan sosyal medya kullanımıyla birlikte sosyal medyada panik etkisi yaracak paylaşımlarda artmıştır. Bu tarz paylaşımların sağlık bakanlığının ve kurumlarının alacağı önlemlerle engellenip sosyal medyanın akıllıca kullanılması bireylerin ruhsal dengelerini korumalarına yardımcı olacağı vurgulanmıştır.

Jelodar ve ark. [69], sosyal medyada ifade edildiği şekliyle Covid-19 ile ilgili konu modellemesi için NLP kullandılar. Daha sonra çalışmalarında LSTMRNN (LSTM- Recurrent Neural Netwroks) modelini kullanarak sınıflandırma yapmışlardır.

Lu ve Zhanh [70], duygu analizinde uzun eğitim süresini kısaltmak amacıyla AT-BiGRU modeline dayalı dikkat mekanizmalı bir model önermiştir.

Yılmaz ve Orman, gerçekleştirdikleri araştırmada [71], Twitter verilerini kullanarak duygu analizi yapmak için LSTM derin öğrenme tekniğini benimsemişlerdir. Farklı veri kümeleri üzerinde %85 - %97 arasında doğruluk başarısı elde edilmiştir. Bu başarım oranının veri kümesindeki veri sayısıyla ilişkili olduğu sonucuna varmışlardır.

Sel ve Hanbay çalışmalarında [72], ön eğitimli olan modellerden faydalanarak 5.292 adet Türkçe Twitter gönderileri üzerinden cinsiyet tespiti yapılmıştır. Çalışmada makine öğrenmesi yöntemleri, derin öğrenme modelleri ve ön eğitimli dil modelleri kullanılmıştır. Yaş, meslek ve cinsiyet gibi çeşitli parametreleri değerlendirmişlerdir. Çalışmada, SVM makine öğrenmesi tekniği, Bi-LSTM, CNN derin öğrenme yöntemleri ile Bert, Bert-128k, Distilbert, Electra Discriminator ve Electra Generator yöntemleri kullanılmıştır. Çalışmada sonucunda bu modellerde (%64,52 - %80,12) arasında doğruluk başarım oranı elde edilmiştir. En yüksek başarım doğruluk oranı Bert-128k modelinde bulunmuştur. Bert modelinin kullanılmasıyla n=1 için %70, n=2 için %73’lük bir başarı elde etmişlerdir.

Samuel ve ark. [73], Twitter verileri üzerinden benzer bir çalışmada Naive Bayes yöntemi ve Lojistik Regresyon modeli kullanılmıştır. Naive Bayes yöntemi kullanıldıklarında 77 karakterden küçük Twitter verilerinde başarı oranı %91.43 gibi yüksek seviyede bulunurken 120 karakterden küçük olan veriler üzerinde test edildiğinde bu doğruluk başarı oranının %57.14’e düştüğü gözlemlenmiştir.

Aygün vd. [74], bu çalışma ile benzer çalışmalar yapmışlardır. Ancak COVID-19 aşıları ülkemizde uygulanmaya başlamadan önceki tarihlere ait İngilizce tweetleri kullanmaları, duygu sınıflandırması için makine öğrenme veya derin öğrenme yöntemlerini kullanmaları ile bu çalışmadan farklıdırlar. Bu çalışmada, ülkemizde aşıların uygulanmaya başladığı tarihten itibaren Türkçe tweetler kullanılarak sözlük tabanlı bir duygu analizi yapılmıştır.

Lyu vd. [75], pandemi başlangıcından itibaren ilk 10 ay içerisinde kamuoyunun COVID-19 aşılarına yönelik tutumlarını ve aşıyı kabullenmelerini veya tereddütlerini öğrenmek için tweetleri kullanarak konu modelleme ve duygu analizi gerçekleştirmişlerdir. Çalışmadaki veriyi 11 Mart 2020-31 Ocak 2021 tarihleri arasındaki İngilizce tweetler oluşturmaktadır. Kanada Ulusal Araştırma Konseyi Duygu Sözlüğü kullanılarak duygu analizi yapılmış ve konu modelleme için gizli Dirichlet tahsisi kullanılmıştır. Konu modelleme sonucunda 16 konu belirlenmiştir. Aşılamayla ilgili görüşler en çok tweet atılan ve en çok tartışılan konu olarak belirlenmiştir. Rusya'nın dünyanın ilk COVID-19 aşısını onayladığı 11 Ağustos 2020 tarihinde aşı konusundaki ilerlemeler dünya çapında en çok tartışılan konu haline gelmiştir. Aşı yaptırma talimatı konusu Ocak 2021'in ilk haftasından sonra en çok tartışılan konu haline gelmiştir. Haftalık ortalama duygu puanları, duyguların olumlu yönde arttığını göstermiştir. Duygu analizinin sonucunda güvenin en baskın duygu olduğunu, ardından korku, üzüntü, beklenti gibi duyguların geldiğini tespit edilmiştir. Güven duygusunun, Pfizer'in aşısının %90 etkili olduğunu açıkladığı 9 Kasım 2020'de zirveye ulaştığı görülmüştür.

Liu ve Liu [76], Pfizer'in ilk COVID19 aşısını duyurusunun ardından kamuoyu duyarlılığını ölçmek ve konuma göre duyarlılığın değişimini tespit etmek için duygu analizi gerçekleştirmişlerdir. Çalışmada 1 Kasım 2020-31 Ocak 2021 tarihleri arasındaki COVID19 aşılarıyla ilgili İngilizce tweetler kullanılmıştır. Tweetlerde geçen duygunun gerçek olup olmadığını belirlemek amacıyla Valence Aware sözlüğü ve duygu çıkarımcısı kullanılmıştır. Tweetlerde geçen başlıca konuları çıkarmak için gizli Dirichlet tahsisi kullanılmıştır. Hem zamansal hem de konumsal analiz gerçekleştirerek farklı konumlarda gönderilen tweetlerdeki duygu farklılıkları belirlenmiştir. Sonuçta analiz edilen tweetlerin yüzde 42,8 olumlu, yüzde 26,9 nötr ve yüzde 30,3 olumsuz duygular içerdiği belirlenmiştir. Olumlu duygu içeren tweetler için deneme sonuçları, yönetim, yaşam, bilgi, etkililik ve olumsuz duygu içeren tweetler için deneme sonuçları, komplo, güven, etkililik ve yönetim olmak üzere 5’er konu belirlenmiştir.

Lyu vd. , Liu ve Liu , İngilizce tweetleri kullanmaları ve 31 Ocak 2021’den sonraki tweetleri analize dahil etmemeleri nedeniyle bu çalışmadan farklılık göstermektedirler.

Marcec ve Likic [77], Twitter’da COVID-19 aşıları ile ilgili atılan İngilizce tweetlere duygu analizi uygulamışlardır. 1 Aralık 2020-31 Mart 2021 tarihi aralığındaki 4 aylık döneme ait tweetleri kullanmışlardır. Çalışma sonuçları Pfizer ve Moderna aşılarına ilişkin duyguların 4 ay boyunca olumlu ve istikrarlı göründüğünü buna karşılık, AstraZeneca/Oxford aşısına ilişkin olumlu duyguların Aralık ayı ile Mart ayı karşılaştırıldığında önemli bir düşüş göstererek zamanla azaldığını göstermiştir. İlgili çalışma duygu sınıflandırmasında AFINN sözlüğünü kullanması, İngilizce tweetler kullanması ve 31 Mart 2021 tarihinden sonraki tweetleri analiz etmemesi yönleriyle bu çalışmadan farklıdır.

Villavicencio vd. [78], Filipinler'de COVID-19 aşılarına yönelik duygular üzerine olumlu, olumsuz ve tarafsız olmak üzere analiz yapmışlardır. Naive Bayes sınıflandırma modelini kullanmışlardır. Çalışmada 1-31 Mart 2021 tarihlerindeki İngilizce ve Filipince tweetler analiz edilmiştir. Analiz sonuçlarına göre tweetlerin %83,38’i olumlu, %8,26’sı olumsuz, %8,36’sı nötr duygu içermektedir.

Ansari ve Khan [79], COVID-19 aşıları ile ilgili kamuoyu tepkisini ölçmek için dünya çapında atılan İngilizce tweetleri kullanarak duygu analizi gerçekleştirmişlerdir. Çalışmada 15 Mayıs-25 Haziran 2021 tarihleri arasındaki tweetler kullanılmıştır. Naïve Bayes sınıflandırıcısı kullanılarak duygu analizi gerçekleştirilmiştir. Araştırma sonucunda, Türkiye hakkında atılan tweetlerin olumsuz duygu ağırlığında olduğu çıkarılmıştır.

Villavicencio vd. , Ansari ve Khan, İngilizce ya da Filipince tweetler kullanmaları, makine öğrenme yöntemlerini kullanmaları ve 2021 yılının sadece belli aylarındaki tweetleri analiz etmeleri sebebiyle bu çalışmadan farklıdırlar.

Çılgın vd. [80], Türkiye’de COVID-19 aşılarına ilişkin kamu duyarlılığını aşıyla ilgili tweetlere makine öğrenmesi algoritmalarını uygulayarak analiz etmişlerdir. Çalışmada 1 Nisan-31Ağustos 2021 tarihleri arasında toplam 412.588 adet Türkçe tweet kullanılmıştır. Sonuçta incelenen tüm tweetlerin %17,70’inin pozitif, %41,97’sinin nötr ve %40,34’ünün negatif duygu polaritesinde olduğu tespit edilmiştir.

Aslan [81], derin öğrenme tabanlı doğal dil işleme tekniklerini kullanarak Türk halkının aşılama süreciyle ilgili duygularını ve tepkilerini araştırmak için duygu analizi gerçekleştirmişdir. Çalışmada, Twitter’da 16 Haziran-18 Eylül 2021 tarihleri arasında paylaşılan Türkçe tweetler kullanılmıştır. Analiz sonucunda tweetlerin %34,9’unun olumlu, %49,6’sının nötr, %15,5’inin olumsuz duygu dağılımına sahip olduğu görülmüştür.

Çılgın vd. , Aslan, Türkçe tweetlerle bu çalışmaya benzer duygu analizi gerçekleştirmişlerdir. İlgili çalışmalar 2021 yılının belirli aylarındaki tweetleri kullanmışlardır. Ancak bu çalışmada 2021 yılının tamamı da analize dahil edilmiştir.

Mermer ve Özsezer [82], COVID-19 aşı tartışmalarını analiz etmek için Twitter verisine duygu analizi uygulamışlardır. Çalışmada 10 Mart 2020- 18 Nisan 2022 aralığındaki 10.308 adet Türkçe tweet kullanılmıştır. Çalışma sonucunda COVID19 aşıları ile ilgili tweet kullanıcılarının %7,5’inin olumlu, %0,59’unun olumsuz, %91,91’inin nötr duyguda olduğu belirlenmiştir. Çalışmada duygu analizi hem sözlük tabanlı hem de makine öğrenme yöntemleri kullanılarak gerçekleştirilmiştir.

Özyurt ve Kısa [83], pandemi sürecinde uzaktan eğitim ile ilgili Twitter paylaşımlarında duygu analizi gerçekleştirmişlerdir. Kaggle veri paylaşım platformundan elde ettikleri verilerden rastgele 999 kaydı manuel olarak pozitif ve negatif olarak etiketlendirmişlerdir. Veri seti KNIME üzerinde kurulan model ile uygun düğümler kurarak önce ön işleme ardından duygu analizi aşamalarından geçirilerek başarı hesaplaması yapmışlardır. Temelde sözlük tabanlı yaklaşımı esas aldıkları çalışmada %88,4 oranında başarıya ulaşmışlardır.

**3. MATERYAL VE YÖNTEM**

**3.1. Doğal Dil İşleme**

Doğal diller; düşünceleri ifade etmek, iletişim kurmak için insan toplulukları tarafından kullanılan sözcükler ve gramer kurallarını içinde barındırır. İnsanlar kavramlarla düşünür, kavramlar üzerinden iletişim kurar. Oysa bilgisayarlar matematiksel kurallarla çalışan makinelerdir. Kısa zamanda çok sayıda hesaplama yapabilseler de insanların kolaylıkla anlayabildiği karmaşık dilsel yorumlamalardan yoksundurlar. Ancak bilgisayarlar doğal dilleri insanlar gibi yorumlayamasalar da sınıflandırma ve çeviri işlemlerini başarı ile gerçekleştirebilirler.

Bu doğrultuda; bilgisayarlara, iletişim dilinin aktarılması ve dilin doğru bir şekilde çözümlenmesi ihtiyacı oluşmuştur. Bunun sonucunda da bilgisayar bilimi, yapay zekâ ve bilişimsel dil biliminin birleşiminin bir araya gelmesiyle Doğal Dil İşleme (Natural Language Processing) doğmuş olup bu alanın asıl kaynağını bilgisayar kullanarak iki farklı dili birbirine çevirme işlemi oluşturmaktadır (Çoban, 2016; Gordin, 2015; Yıldırım, 2018).

Son yıllarda oldukça önem kazanmış bir çalışma alanı ve bilgisayar bilimi olan DDİ, esasen 20. yüzyılın ikinci yarısının başlarında yapay zekânın küçük bir alt dalı olarak meydana gelmiştir (Oflazer, 2006). 1954 yılında Georgetown deneyi olarak adlandırılan Georgetown Üniversitesi ve IBM‘in birlikte geliştirdiği bilgisayarlı çeviri deneyinde 60‘dan fazla Rusça cümlenin başarılı bir şekilde İngilizce’ye çevrilebildiği görülmüştür. 1980‘li yıllarda ise hesaplama gücünün hızla artması ile bu alandaki gelişmeler hızlanmıştır (Gordin, 2015). Zamanla yapılan araştırmalar ve gerçekleştirilen uygulamalar sonucunda elde edilen başarılar ile de DDİ nihayetinde bilgisayar bilimlerinin temel bir disiplini olarak kabul edilmeye başlanmıştır.

Doğal dil işleme kısaca; birçok kaynaktan elde edilen doğal dilde yazılmış metinler ile ilgilenen bir alan olarak tanımlanabilir (Korkusuz, 2018). DDİ, doğal dillerin kurallı olan yapısını detaylıca incelerken aynı zamanda çözümleyerek işlemekte ve anlaşılması yahut yeniden üretilmesi amacını taşıyarak otomatik çeviri, konuşma, ses tanıma, üretme ve duygu analizi gibi pek çok konuyu içeren çalışmada kullanılmaktadır (Yelmen, 2016). Genellikle metin tabanlı çalışmalarda kullanılmakta olup istatistiksel olarak metnin üzerinden sonuçlar üretmeyi kapsarken çoğunlukla yapay zekâ altındaki dil bilim bilgisine dayalı çalışmaları içermektedir (Şeker, 2014).

DDİ yöntemleri, insanların el ile yapmak zorunda olduğu bazı metin sınıflandırmaların otomatik olarak yapılabilmesine olanak vermekte ve birçok spesifik uygulamada ciddi kolaylıklar sağlamaktadır. DD‘nin metinsel verinin yoğun olduğu soru cevaplama, bilginin çıkartılması, duygu analizi, makine diline çeviri, sözcük anlamı açıklaştırma ve özetleme alanlarında kullanılabilecek tekniklerden oluştuğu söylenebilir.

Bilgisayarların doğal dili anlamlandırabilmesi için, dilin modellenebilen matematiksel bir çerçeveye dönüştürülmesi gerekir. Bu amaçla kullanılan başlıca teknikler şunlardır:

**Dizgelere ayırma (Tokenization):** Bir metni kelimelere bölme işlemidir. Herhangi bir ayraç kullanarak bölme yapılabileceği gibi genelde boşluk karakterine göre bölme yapılır. Metin bölme işleminden sonra, bütün metin çok öğeli bir sözcük dizisine dönüşür.

**Durak ifadelerin (stop words) kaldırılması:** Metin içerisindeki bağlaçların, niteleyicilerin, tek başına anlamsal katkısı olmayan sözcüklerin temizlenmesi işlemidir. Bu işlem esnasında kullanılmayan rakamlar, noktalama işaretleri, yeni satır (\n), tab (\t), başa dönme (\r) gibi özel karakterlerin temizliği de gerçekleştirilir.

**N-Gram:** Belirli bir N sayısına bağlı olarak sıralı sözcükleri gruplamak için kullanılır.

**Kelime vektörü, ağırlıklandırma ve skorlama:** Metin içerisinde n-gram yöntemleri ile oluşturulan vektörlerin kodlanması için belirli skorlama ve ağırlıklandırma teknikleri kullanılır. Bu tekniklerden olan Terim sıklığı (Term Frequency –TF), bir doküman içerisinde belirli terimlerin ne kadar sık yer aldığını hesaplamak için kullanılan yöntemdir. Çeşitli frekans ve normalizasyon formülleri kullanılarak hesaplama yapılır.

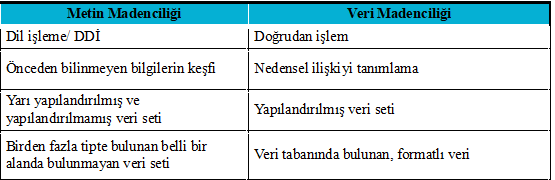
**3.2.METİN MADENCİLİĞİ**

Son yıllarda donanım ve yazılım teknolojisindeki büyük gelişmelere bağlı olarak hızlı ilerlemeler gösteren veri madenciliği, farklı veri türlerine uygulanabilen bir alan haline gelmiştir. Donanım ve yazılım platformlarının geliştirilmesiyle bu durum, özellikle web ve sosyal ağlar için büyük miktardaki metin türündeki veriler için de gerçekleşmiştir. Farklı uygulamalar sonucu elde edilen metin verilerinin miktarının artmasıyla verilerin dinamik ve ölçülebilir bir şekilde öğrenilebilmesini sağlayan algoritmik tasarımların geliştirilmesine ihtiyaç duyulmaya başlanmıştır (Bender ve ark., 2003). İşte bu noktada metin madenciliği devreye girmektedir.

Metin madenciliği; yapılanmamış ya da yarı yapılı metinlerden anlamlı sonuçlar çıkarabilmek amacıyla metinlerin belirli süreçlerden geçirilip yapılandırılmış hale getirilerek çıktıların analiz edilmesini ifade etmektedir (Karamanlı, 2019). Metin madenciliği uygulaması temelde insan beynindeki en karmaşık analitik işleme sistemine sahip anlayışları, yazı dilinde analiz etmeyi amaçlamaktadır.

Metin türündeki verilerde kelime ve cümlelerin her zaman doğal sıralanışında olmamasından dolayı metin madenciliğinde kullanılacak istatistiksel yöntemlerin doğru sonuçlar verebilmesi, DDİ teknikleriyle kelime ve cümlelerin ön işlemesinin dikkatli bir şekilde yapılması ile mümkündür (Berger ve ark., 1996).

**Tablo 3.1.** Metin Madenciliği ile Veri Madenciliğinin Karşılaştırılması.

****

Tablo 3.1.‘de metin madenciliği ile veri madenciliğinin karşılaştırılması yapılmıştır. Mevcut verinin türüne göre metin madenciliği ya da veri madenciliği tercih edilebilir. Genel olarak, veri madenciliğinde veri hazır bir veri tabanından alınır ve formatı bellidir. Yapılandırılmamış veri seti kullanılarak nedensel ilişkiyi tanımlama amacı güdülür. DDİ çalışmaları ile birlikte yürütülen metin madenciliğinde ise, metin kaynaklı literatürdeki diğer bir çalışma alanı olan birçok veri türü aynı anda daha dağınık veri kaynaklarında bulunmaktadır. Ayrıca yarı yapılandırılmış ve yapılandırılmamış veri seti kullanılarak önceden bilinmeyen bilgilerin keşfedilmesi amaçlanır.

Metin madenciliğinin bilgisayar bilimi, istatistik, yapay zekâ, makine öğrenmesi, yönetim bilimi ve diğer bilimler ile iç içe olduğu söylenebilir (Amanet, 2017). Esasında metin madenciliği kavram çıkarma, bilgi çıkarma, bilgi geri getirme, web madenciliği, sınıflandırma, kümeleme ve DDİ alanlarında kullanılır.

Metin madenciliği teknikleri, temelde dört kategoriye ayrılmaktadır (Tunalı, 2009):

* **Sınıflandırma (Classification):** Nesnelerin önceden bilinen sınıflara yahut kategorilere dahil edilmesi işlemini ifade eder.
* **Birliktelik Analizi (Association Analysis):** Sıklıkla birlikte yer alan yahut gelişen sözcük veya kavramların belirlenerek doküman içeriğinin veya doküman kümelerinin anlaşılmasını sağlar.
* **Bilgi Çıkarımı (Information Extraction):** Dokümanların içerisindeki yararlı veri ya da ifadelerin bulunması işlemini ifade eder.
* **Kümeleme (Clustering):** Doküman kümelerinin temelini oluşturan yapıların keşfedilmesi işlemini ifade eder.

Veri kaynağı olarak metinlerin ele alındığı metin madenciliğinde, ilk olarak metin kaynaklarından amaca uygun olan veri seçimi yapılırken sonrasında tekrar eden ve etkisiz kelimeler filtrelenmekte ve metinler istenilen özelliklere göre parçalara ayrılmaktadır. Bunu yaparken de aynı zamanda metinler noktalama işaretlerinden, sayısal verilerden arındırılarak metinlerdeki büyük harfler küçük harflere de dönüştürülmektedir. Böylece metin analizi için bir ön işleme yapılmaktadır. Metin verilerinin ön işleme aşamasını geçmesinin ardından gösterim aşamasına geçilir ve bu aşamada bir kelimenin metin içerisinde ne kadar önemli olduğunun istatistiki olarak değerlendirilmesi için metnin sayısal olarak ifade edilmesi işlemi gerçekleştirilmektedir (Aninditya ve ark., 2019; Beşkirli ve ark., 2021). Son olarak da pek çok aşamadan geçerek metin madenciliği için kullanıma hazır hale getirilen veriler metin kaynaklarından bilgi keşfi için sınıflandırma, birliktelik analizi, bilgi çıkarımı ve kümeleme teknikleri uygulanarak analiz edilmektedir.

**3.2.1.Veri Ön İşleme**

Veri seti oluşturma aşamasında elde edilen metinleri kullanabilmek için öncelikle metinlerin düzenlenmesi gerekmektedir. Bu düzenleme ile metinlerde bulunan yazım yanlışlarını, kullanılmayacak olan internet sayfası linkleri, kullanıcı adı gibi verileri kısaltma olarak kullanılan kelimelerde düzenlemeler yaptıktan sonra kullanıma hazır hale getirilir (Kuzucu, 2015).

Yapılandırılmamış metin formatındaki verilerin, metin madenciliği süreçlerinde işlenerek yapılandırılmış hale getirilebilmesi için gereken ilk adım ön işleme aşamasıdır. Bu çalışma kapsamında, DÖ yöntemlerine giriş için verilecek vektörlerin elde edilmesinden önce, veri kümesi üzerinde gerçekleştirilen ön işlemler şu şekilde sıralanabilir:

* **URL Bilgisi Silme:** Tweet'lerde bulunan URL'ler, paylaşımın duygu analizine katkı sağlamadığından, yalnızca "URL" kelimesiyle değiştirilmiştir.
* **Boşluk Silme:** Tweet'lerde başta ve sonda bulunan gereksiz boşluklar kaldırılmış, kelimeler arasındaki boşluklar ise tek boşluğa indirilmiştir.
* **Noktalama İşareti Silme:** Tweet’ler içerisinde geçen tüm noktalama işaretleri silinmiştir.
* **Kullanıcı Adı Bilgisi Silme:** Tweet’lerin içerisindeki kullanıcı adı bilgisi silinmiştir.
* **Anlamsız Kelime Silme:** Tweet’ler içerisinde geçen kelimelerdeki yazım yanlışları NLTK[84] yardımıyla düzeltilmiştir. Kelime görünümünde yazılmış anlamsız karakter dizilimleri tespit edilmiş ve silinmiştir.
* **Rakam Silme:** Tweet’ler içerisinde geçen rakamlar silinmiştir.
* **Tek Karakter Silme:** Tweet'lerde yer alan tek karakterlik içerikler kaldırılmıştır.
* **Kelime Düzeltme:** “MErhabaaaaa” gibi karakter tekrarı içeren kelimeler, NLTK kütüphanesi yardımıyla Türkçe dil bilgisi kuralına göre düzeltilmiştir.

Sınıflandırmada kullanılacak veriler, bazen eksik ya da tutarsız olabilmektedir. Veri setinde bulunan eksik ya da hatalı verilere gürültü adı verilmektedir. Veri setinde gürültülü verilerin bulunması halinde, bu sorunun giderilmesi beklenirken bu gibi durumlarda aşağıdaki yöntemler kullanılmaktadır (Özkan, 2008):

* Gürültülü verilerinin ilgili veri setinden silinmesi yahut yerine yenisinin eklenmesi gerekmektedir.
* Gürültülü verinin yerine sabit bir değer kullanılabilir.
* Tüm verilerin ya da bir kısım verilerin ortalaması hesaplanıp gürültülü verilerin yerine bu değer kullanılabilir.
* Gürültülü verilerin yerine, veri setinde yer alan verilerin tamamı ya da belli bir kısmı kullanılarak gürültülü veriler tahmin edilir ve elde edilen bu veriler gürültülü verilerin yerine kullanılabilir.

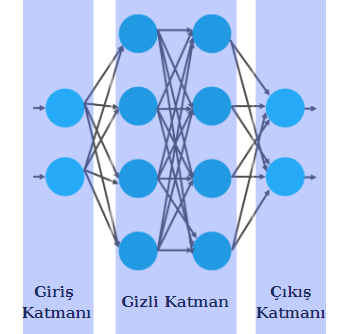
**3.2.2.Veri Etiketleme**

Veri etiketleme temel olarak Şekil, metin veya işitsel verileri insanların algıladığı özelliklerle işaretlemek ve anlamlandırmaktır. Veriler üzerinde birçok farklı öğe etiketlenebilse de genellikle belirtilen alt kümeye odaklanması beklenmektedir. Bu çalışmalar, 3 temel aşamadan oluşmaktadır: veri hazırlama, modelin eğitimi ve modelin uygulanmasıdır. Bu çalışmalarda veri etiketleme, verilerin hazırlanmasında büyük rol oynamaktadır. Bu aşamada, modelin eğitilmesi için etiketlenmiş veriler kullanılmaktadır (Andırın, 2020). Etiketleme işlemi kişiler tarafından manuel olarak değerlendirilip yapılabildiği gibi geliştirilen hazır modeller ile etiketlendirme işlemi yapmak mümkündür.

**3.3.DERİN ÖĞRENME**

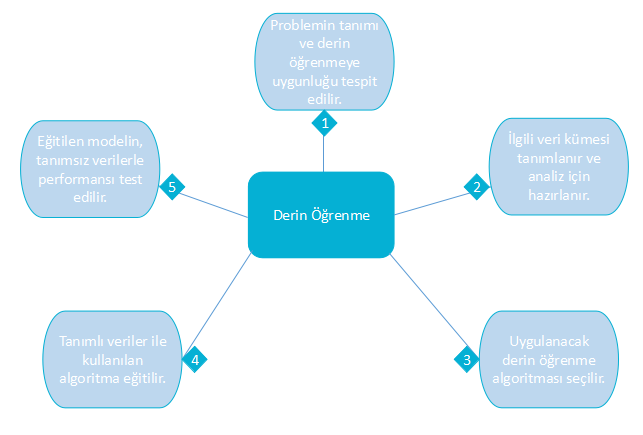
Derin Öğrenme son yılların en önemli konuları arasındadır. Derin öğrenme, doğrusal olmayan işlem birimlerinin birçok katmanını kullandığı, birbirini izleyen her katmanın girdi olarak bir önceki katmanların çıktısını aldığı bir makine öğrenmesidir.Derin öğrenme algoritmaları, yapay sinir ağlarının (YSA) daha karmaşık bir yapısal versiyonu olarak kabul edilebilir. Derin öğrenmenin tek katmanlı yapay sinir ağlarından farkı, lineer olmayan problemlerin çözülmesine olanak tanımasıdır [85]. Derin öğrenme yaklaşımında gizli katmanlar bulunmaktadır.

Derin öğrenmenin kullanım alanları her geçen gün artmaktadır. Günümüzde GPU hızlarının artmasıyla birlikte, derin ağlar çok daha hızlı ve etkin olarak eğitilebilmekte ve bu sayede artan başarı oranları ile hemen hemen her alanda yaygınlaşarak kullanılmaktadır. Derin öğrenme modeli, makine öğreniminin birçok uygulamasını sağlamış ve yapay zeka alanındaki başarıyı artırmıştır. Derin sinir ağının hücre yapısı aşağıdaki Şekil 3.1'de gösterilmektedir [86].



**Şekil 3.1.** Derin sinir ağı

Geleneksel makine öğrenmesi algoritmaları doğrusal yapıdayken, derin öğrenme algoritmaları problemin karmaşıklığına göre değişen bir hiyerarşi modeline sahiptir. Derin öğrenme süreci başarı sonucu belirli bir seviyeye gelene kadar devam eder. Bu süreçte verilerin geçmesi gereken genel adımlar Şekil 3.2.’te gösterilmektedir [87].



**Şekil 3.2.** Derin Öğrenme Süreçleri

**3.3.1.Derin Öğrenme Algoritmaları**

Günümüz derin öğrenme algoritmalarının temeli insan beynindeki bir nöronun bilgisayar benzetmesi yapılmaya çalışılmasıyla ortaya çıkmış bir süreçtir. Donald Hebb'in sinir hücrelerinin incelenmesiyle başlayan süreç, daha sonra bilgisayar ortamında matematiksel olarak modelinin oluşturulmasıyla yapay sinir ağlarının temelini oluşturmuştur (Hebb ve ark. 2005). Yapay sinir ağlarının gelişimi ve ilerleyişi sürecinde yıllar geçmiş ve son yıllarda derin öğrenme terimi ortaya atılmış, araştırmalar bu alanda yoğunlaşmıştır.

Derin öğrenme algoritmaları, geçmiş verilere dayanarak gelecekteki değerleri tahmin etme yeteneği sunar. Özellikle ileriye dönük tahmin işlemlerinde kullanılan derin öğrenme, öğrenme verilerini ifade etmek ve istenen profili oluşturmak için geçmiş verileri kullanma kabiliyetine sahiptir.

Derin öğrenme, özellikle sınıflandırma, tanıma ve tespit gibi alanlarda kullanılmaktadır (Deng ve Yu, 2014). Yapılan çalışmalar bu yönde ağırlık kazanmıştır. Son zamanlarda, derin öğrenme algoritmaları çeşitli alanlarda yaygın olarak kullanılmakta ve günlük yaşamımızda birçok teknolojik yapı içerisinde yer almaktadır. Bu algoritmaların başarılı sonuçlar elde etmesi, bu alana olan ilgiyi artırmıştır. Yapılan pek çok sınıflandırma, tanıma çalışmaları mevcuttur.

**3.3.1.1.Yapay Sinir Ağları**

YSA; girdi, çıktı ve ağırlık değerleri bir algoritmaya göre kendi kendine öğrenebilen, çıkarımlar yapabilen ve ayrıca kararlar alabilen doğrusal olmayan problemlerin de çözülebilmesine olanak sağlayan insan beynindeki biyolojik sinir yapısını modellemektedir [88, 89, 90]. Tek katmanlı algılayıcıların ilk sinir ağlarından biri olduğu söylenebilir.Bu yapı, birden fazla girdi alıp bu girdilerden çıktı üreten bir sinir hücresine dayanır. Ve (and), Veya (or), Değil (not) durumları gibi doğrusal fonksiyonlarla ilgili problemlerde kullanılabilirler.

Bu ağ yapısında öğrenmeyi sağlamak için eğitim seti adı verilen ve örneklerden oluşan bir veri seti kullanılmaktadır. Tek bir nöron için birçok girdi yeterli olmayacağından, paralel işlem için birden fazla nörona ihtiyaç vardır. Veriler giriş katmanından alınır ve gizli katmana aktarılır. Ara katman, en az bir katman olarak farklılık gösterebilir. Her katmanın çıkışı, bir sonraki katmanın girişi olmaktadır. Her nöron, bir sonraki katmandaki nöronlara bağlanır. Çıktı katmanı, önceki katmandan alınan verileri işleyerek ağın çıktısını belirler. Çıktıların sayısı çıktı katmanındaki elemanların sayısına eşittir. Ağın öğrenmesi için örnek girdi ve çıktılardan oluşan bir eğitim seti gereklidir. Aktivasyon fonksiyonuna girecek girdi, ağırlıklar, transfer fonksiyonu ve net girdidir. Yani ağın giriş değerleri, ağırlıkları ve ağda uygulanan aktivasyon işlemi kullanılarak ağın çıktısı elde edilir. Bir çıktı katmanı, önceki katmandaki işlem öğelerine bağlı birden fazla işlem öğesi ve işlem öğesi içerir. Çok katmanlı ağlarda çeşitli aktivasyon fonksiyonları kullanılabilir. Sigmoid işlevi bunlardan birisidir. Etkinleştirme işlevi, girdi katmanları ve çıktı katmanları tarafından oluşturulan eğri eşitlemesini etkinleştirir. Ağın performansını etkilediği için uygun aktivasyon işlevinin seçilmesi önemlidir. Aktivasyon fonksiyonunu doğrusal olarak seçmek de mümkündür. Sigmoid aktivasyon fonksiyonu 0 ile 1 arasındaki değerleri hesapladığından, bir olayın meydana gelme olasılığını bulan modellerde kullanılır. Yapay sinir ağlarında, girdiyi işleyerek bu girdi sonucunda üreteceği çıktıyı belirleyen aktivasyon fonksiyonları kullanılmaktadır. Derin öğrenme modelinde, hata değerini minimize etmek için kullanılan geriye besleme işleminde türev hesabı kullanılmaktadır. Bu sebeple genellikle türevi rahatlıkla alınabilen aktivasyon fonksiyonları seçilmektedir. Bir aktivasyon fonksiyonu, yapay bir nörondaki, girdilere dayalı bir çıkış sağlayan fonksiyondur [91].

**3.3.1.2.Konvolüsyonel Sinir Ağları**

Konvolüsyonel Sinir Ağları (Convolutional Neural Network – CNN) mimarisi, özellikle görüntü tanıma, doğal dil işleme ve konvolüsyonel çekirdek ile yerel özellikleri tanıma yeteneğine sahip diğer alanlarda kullanılan bir derin öğrenme mimarisidir (Yan ve ark., 2015). CNN mimarisi evrişim katmanlarından oluşur ve her katmanda filtreler kullanılarak verilerin özelliklerinin çıkarılmasına dayanır [92]. CNN modelleri girdi olarak alınan görüntülerin özelliklerini çıkartır ve bu işlemi yaparken filtreler kullanır. Örneğin bir görüntü için bu filtreler kenarlar, köşeler veya şekiller olabilir. Ardından bir veya birden fazla tam bağımlı katman kullanarak bu özellikleri sınıflara ayırır. Bu katmanlar son çıktıyı hesaplayarak modelin tahmin ettiği sınıfın olasılığını hesaplar. CNN mimarinin avantajı verilerin özelliklerini otomatik öğrenebilmesidir. Bu nedenle verilerin özelliklerinin elle çıkarılmasına gerek yoktur. Ayrıca, eğitim verilerindeki küçük değişikliklere karşı dirençlidir ve genelleştirme yeteneği yüksektir.

**3.3.1.3. Uzun Kısa Vadeli Hafıza Ağları**

Uzun-Kısa Süreli Bellek Ağları (Long-Short Term Memory – LSTM) mimarisi, aslında RNN’nin bir sorununa çözüm olarak literatüre girmiş ve RNN’nin alt türü olarak görülmektedir. RNN mimarisindeki “uzun süreli bellek kaybı” sorununu çözmek üzere gizli durum hücresine ek olarak birçok kapı mekanizması içermektedir. Bu kapılar, ağın hangi bilgileri saklayıp hangilerini saklamayacağına karar vermesine yardımcı olur [93]. LSTM mimarisinde üç ana kapı bulunmaktadır. Bu kapılar, unutma (forget) kapısı, giriş (input) kapısı ve çıkış (output) kapısıdır. Unutma kapısı, hafıza hücresindeki bilginin ne kadarını unutulacağını belirler. Bu kapı önceki hafıza hücresindeki bilgi ve şimdiki girdiye dayanarak belirlenir. Giriş kapısı, hafıza hücresine ne kadar yeni bilgi ekleneceğine karar verir. Bu kapı şimdiki girdi ve önceki saklı durum arasındaki ilişkiye dayanarak belirlenir. Çıkış kapısı ise, hafıza hücresindeki bilgiyi kullanarak ne kadar çıktı üretileceğini belirler. Bu kapı hafıza hücresindeki bilgi ve şimdiki girdiye dayanarak belirlenir.

**3.3.1.4. Embedding**

Kelime yerleştirme teknikleri ya da diğer bilinen adıyla kelime vektörleri, dokümanlar içerisindeki benzer anlamı olan kelimelerin ve kelime öbeklerinin benzer bir gösterim ile sunulmasını sağlayan bir tür temsil yöntemidir. Bu yöntemler ile kelimeler, bir vektör uzayında temsil edilir. Bir kelime yerleştirme tekniği, bir dokümanda yer alan kelimelerin her birinin bir vektör uzayını çıkarıp tüm sayısal değerlerini bu vektör uzayının üzerine taşır ve bir dizi sayısal veri üretir. Kısaca kelime yerleştirme teknikleri, metinleri sayısal verilere dönüştüren yöntemlerdir.

**3.4.Makine Öğrenmesi**

Makine öğrenmesi eldeki mevcut veri setleri üzerinden eğitilerek, veriler üzerinde tespit edilen ilişkiler ile yeni karşılaşılan durumlar hakkında çıkarımlar yapabilen teknikler bütünüdür. Bu tekniklerden biri olan doğal dil işleme insanlar tarafından konuşulan dillerin makineler tarafından anlamlandırılmasına yönelik çalışmaları kapsar. Doğal dil işleme süreci bir cümle içerisindeki kelimelerin ilişki ağını ortaya koyarak, kategorizasyon, kavram çıkarımı, özetleme, çeviri gibi dilsel işlemleri makineler aracılığı ile gerçekleştirmeye çalışır. Twitter gibi sosyal ağlar, doğal afetler konusunda müdahale hızını ve verimliliğini arttırmaya yardımcı olabilecek değerli bilgiler içerir [94].

Makine öğrenmesi, veri madenciliği ve doğal dil işleme, çeşitli sınıflandırma kriterlerine göre tweet'leri önceliklendirerek verilerin filtrelenmesine yardımcı olur.

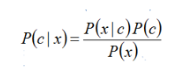
**3.4.1.Makine Öğrenmesi Algoritmaları**

Makine öğrenmesi algoritmaları, karmaşık veri kümelerinin keşfedilerek analiz edilmesine ve anlamlandırılmasına yardımcı olan kod parçacıkları olarak tanımlanabilir. Her algoritma, bir makinenin belirli bir hedefe ulaşmak için takip ettiği belirli ve sınırlı adımlardan oluşan bir yönergeler serisidir. Makine öğrenmesi modelinin hedefi, tahmin yapmak ya da bilgileri kategorilere ayırmak amacıyla kullanılabilecek desenler ortaya çıkarmaktır (Anonim, 2021). Literatürde pek çok farklı alanda pek çok farklı makine öğrenmesi algoritması bulunmaktadır. Bu bölümde bu çalışmada kullanılan makine öğrenmesi algoritmaları açıklanmaya çalışılmıştır.

**3.4.1.1.Naive Bayes Sınıflandırması**

NB sınıflandırma algoritması, Bayes teoremine dayalı, olayların gerçekleşme sıklığının hesaplandığı koşullu olasılık yöntemini kullan istatistiksel bir sınıflandırma yöntemidir. NB teoreminde, önerme sınıflandırmada kullanılacak olan her özniteliğin istatistiksel olarak bağımsız olduğu üzerine kuruludur ve her bir nitelik verilen sınıf içinde diğer niteliklerden bağımsız olarak kabul edilmektedir. Bir başka ifade ile tüm özellikler sonuç olasılığını birbirinden bağımsız olarak etkilemektedir. Yani, bir özniteliğin sınıfta yarattığı etki, diğer özniteliklerin varlığına veya yokluğuna bağlı değildir (Elmas, 2019; Karaöz, 2018). NB sınıflandırma algoritması basitçe, önceki bilgiler ışığında hipotez olasılığının hesaplanması için iyi bir yoldur.

NB sınıflandırma algoritmasının işleyişinde öncelikle sınıfların olasılıkları ve özniteliklerin tüm durumları için sınıf değerlerine bağlı koşullu olasılıkları bulunur. Yeni gelen veriler ise ilgili sınıflar için tüm olasılıklar çarpılması ile elde edilen değerlerden hangisi daha yüksekse o sınıfa atanmaktadır (Dean, 2014).

 (3.1)

Bayes teoremi, yukarıda yer alan formülle ifade edilmektedir (3.1). Formülde görülen P(c|x) sonraki olasılığı (posterior probability), x veriyi, c ise hipotezi ifade etmektedir. Sonuç ise, x‘ ye göre c‘nin yani veriye göre hipotezin olasılığıdır. Hipotezin doğru olduğu varsayıldığında verinin olasılığı P(x|c) ile ifade edilir. P(x) ve P(c), birbirlerinden bağımsız olarak sırayla verinin olasılığı ve hipotezin doğru olma olasılığıdır. Bu olasılığın adı, önceki olasılık (prior probability)’dir. (Brownlee, 2020).

NB algoritmaları öneri sistemleri, duygu analizi, spam filtreleme gibi alanlarda oldukça sık kullanılmaktadır. NB hızlı ve kolay uygulanabilir olmasına rağmen, tahmin edicilerin bağımsız olması gerekliliği en büyük dezavantajıdır. Çoğu gerçek yaşam durumlarında, tahmin ediciler bağımlıdır ve bu da sınıflandırıcının performansını olumsuz etkilemektedir (Gandhi, 2018).

**3.4.1.2.Destek Vektör Makineleri**

Destek Vektör Makineleri (Support Vector Machine – SVM) algoritması sınıflandırma ve gruplandırma problemlerinde kullanılmaktadır.Veri noktalarını birbirinden ayıran bir çizgi oluşturarak sınıflandırma işlemini gerçekleştirir. SVM algoritması genellikle çözümlenmesi zor olan büyük veri setlerinde etkili olduğu görülmüştür [95]. SVM algoritması, verileri birbirinden ayıran en iyi hiper düzlemi bulmak için çalışır. Hiper düzlem, verilerin tamamen ayrılamadığı durumlarda da kullanılabilir.Bu durumda bir hata payı belirlenir ve hiper düzlem en düşük hata payını sağlayacak şekilde oluşturulur. Bu algoritma özellikle doğrusal olmayan veri setleri için etkilidir. Bu veri setlerinde verileri daha yüksek boyutlu bir uzaya taşır ve orada daha kolay sınıflandırılabilir hale getirir. Bu işlem çekirdek fonksiyonları kullanılarak yapılır.

**3.4.1.3. Aşırı Gradyan Arttırma Algoritması**

Aşırı Gradyan Artırma (XGBoost) algoritması, karar ağaçları ve makine öğrenmesinde sıkça tercih edilen bir araç haline gelmiştir. Bu yaklaşım, sınıflandırma, regresyon ve sıralama gibi görevlerde yüksek performans sağlayan önemli bir denetimli öğrenme aracı olarak kabul edilmektedir (Mitchell ve Frank, 2017: 127). Denetimli öğrenme, bir dizi etiketli eğitim örneğinden tahmin edici bir model çıkarım görevi anlamına gelir. Bu öngörücü model ile bir üretim hatasının meydana gelip gelmediğini belirleme ya da belirli bir günde sıcaklık veya nem tahmini gibi sorunlara çözüm üretilebilmektedir. Bu bağlamda XGBoost algoritması, doğru modeller üretmek için artırma adı verilen bir işlemi uygulayan, denetimli bir öğrenme algoritması olarak ifade edilebilmektedir (Mitchell ve Frank, 2017: 129).

XGBoost algoritmasındaki artırılmış ağaçlar, regresyon ve sınıflandırma ağaçlarına bölünmüştür. Bu algoritmanın özü, amaç fonksiyonu değerinin optimize edilmesine dayanmaktadır (Zheng vd., 2017: 1171). XGBoost algoritmasının en önemli özelliği, tüm senaryolarda ölçeklenebilirliğidir. Sistem, tek bir makinedeki mevcut popüler çözümlerden 10 kattan daha hızlı çalışır ve dağıtılmış veya hafıza sınırlı ayarlarda milyarlarca örneğe ölçeklenir (Chen ve Guestrin, 2016: 785).

**3.4.1.4. Kategorik Arttırma Algoritması**

Yandex firması tarafından 2017 yılında geliştirilen CatBoost, gradyan arttırma tekniğini kullanan, hızlı eğitim süresine sahip ve yüksek performansıyla bilinen bir GBM (Gradyan Arttırma) algoritmasıdır. CatBoost, ardışık olarak ağaçları birleştirerek, bir önceki ağacın hatalarını gidermeye çalışır, böylelikle modelin genel hata oranını azaltarak tahmin performansını artırır (Patrous, 2018).

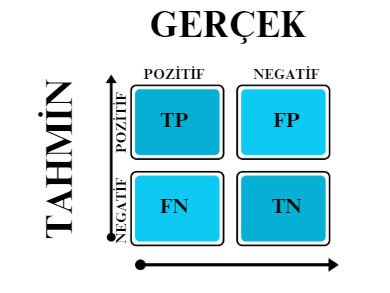
Bununla birlikte, CatBoost kategorik değişkenleri doğrudan işleyebilir ve bu değişkenlerin etkisini daha iyi yakalayabilir. Ayrıca, overfitting (aşırı uydurma) sorunlarına karşı dirençlidir ve daha az hiper parametre ayarı gerektirir, bu da kullanımını kolaylaştırır (Ibrahim ve ark., 2020).

CatBoost, kategorik öznitelikleri hızlı bir şekilde işleyen Gradyan Arttırma Karar Ağacı (GBDT, Gradyan Boosted Decision Tree) algoritmasıdır. Az veriyle bile başarılı sonuçlar elde edebilen bu algoritma, geleneksel GBDT algoritmalarından farklı olarak ön işleme süresi yerine eğitim süresi boyunca kategorik öznitelikleri ele almaktadır (Muratlar, 2020).

Gradyan arttırmada kategorik özniteliklerin kullanılmasında çeşitli yöntemler bulunmasına rağmen, bu yöntemler tahminlerde kaymalara neden olabilmektedir. Bu nedenle, CatBoost algoritması önerilmektedir. CatBoost, tahminleri iyileştirmek ve aşırı uyum sorununu çözmek için etkili bir çözüm sunmaktadır.

**3.4.2.Performans ölçütleri**

Doğruluk, duyarlılık, kesinlik, F-ölçütü gibi kavramlar sınıflandırma algoritmalarının performans ölçümünde kullanılmaktadır. Bir sınıflandırma algoritması tarafından gerçekte olan ve tahmin edilen sınıflandırmalar hakkında bilgi edinmek ve bu algoritmanın performansını özetlemek için karışıklık matrisi kullanılır. Matriste yer alan veriler ile bu algoritmaların karşılaştırılması ve performans ölçüsü değerlendirilir. Aşağıdaki Şekil 3.3.‘te iki sınıflı bir sınıflandırıcı için karışıklık matrisi gösterilmektedir (Albayrak, 2018).



**Şekil 3.4.** Karışıklık Matrisi

Şekil 3.4.‘teki TP, FP, FN, TN girdilerinin ifadesi şöyledir:

**TP (True Pozitif):** Gerçekte pozitif olan ve tahmin edilenin de pozitif olduğu örnek sayısı.

**FP (False Pozitif):** Gerçekte pozitif olan ama tahmin edilenin pozitif olmadığı örnek sayısı.

**FN (False Negatif):** Gerçekte negatif olan ama tahmin edilenin negatif olmadığı örnek sayısı.

**TN (True Pozitif):** Gerçekte negatif olan ve tahmin edilenin de negatif olduğu örnek sayısı.

Sınıflandırma performansını karşılaştırmak için kullanılan metrikler ve bunlara ait formüller aşağıdaki gibidir (Nalçakan ve ark., 2015):

**Doğruluk (Accuracy):** Doğru olarak sınıflandırılmış örneklerin (TP + TN) toplam örnek sayısına (TP + TN + FP + FN) oranıdır.

**Duyarlılık (Recall):** Pozitif olarak etiketlenmiş (TP) örnek sayısının, toplam pozitif (TP+FN) örnek sayısına oranıdır.

**Kesinlik (Precision):** Pozitif olarak etiketlenmiş (TP) örnek sayısının, pozitif olarak ön görülen tüm örnek sayısına (TP+FP) oranıdır.

**F-Ölçütü (F-Measure):** Duyarlılık ve kesinlik metrikleri tek başına anlamlı bir performans ölçütü çıkarmakta yeterli değildir. F-Ölçütü, duyarlılık ve kesinlik metrikleri kullanılarak hesaplanmaktadır.

**3.5.Duygu Analizi**

Duygu analizi, insanların duygularının ve düşüncelerinin uygun bilgisayar algoritmalarıyla analiz edilmesidir. Duygu analizi, insanların duygularını olumlu, olumsuz veya nötr olarak sınıflandırmayı amaçlar. Bu sınıflandırma süreci sonucunda, bir ya da birden fazla yazarın belirli bir konudaki düşüncesi belirlenir. Duygu analizi çalışmaları günümüzde yaygın bir şekilde kullanılmaktadır ve elde edilen bilgiler son derece değerlidir. Bu tür çalışmalarla elde edilen sonuçlar, birçok şirketin veya devletin insanların görüşlerini anlamasına ve karar verme süreçlerini doğrudan etkilemesine yardımcı olur.

Duygu analizi yapabilmek için başlangıçta bir veri kümesine ihtiyaç duyulur. Bu veri kümeleri etiketlenmiş verilerden oluşmalıdır. Yani bir yazının olumlu, olumsuz ya da nötr duygu belirttiği belirtilmelidir.Bu aşamadan sonra veriler, gerekli teknikler ve algoritmalar kullanılarak ön işleme sürecine tabi tutulur. Bu şekilde verilerde genel olarak yazım hatası, noktalama işaretleri, gereksiz kelimeler gibi sorunlar tespit edilir. Ham veriler bu hatalardan arındırılmak için ön işleme sürecine girer. Veri temizleme işlemi bittikten sonra sınıflandırma işlemi gerçekleştirilir. Bu işlemler için daha çok makine öğrenmesi algoritmaları kullanılır [96]. Eğitim verileriyle sınıflandırma işleminden sonra test verileriyle karşılaştırılır ve sistem başarısının ne kadar olduğu belirlenir.

**3.6.Twitter, Twitter Api, Selenium ile Veri Çekme**

Kullanıcıların 280 karakter sınırı ile "tweet" adı verilen gönderiler yazabildiği, güncel ve trend konular hakkında duygu ve düşüncelerini paylaşabildiği popüler bir sosyal medya platformu olan Twitter, ilk kez 2006 yılında kullanılmaya başlanmış, 2011 yılında da Türkçe desteği ile kullanıcılara sunulmuştur.

Her geçen gün kullanıcı sayısı giderek artan Twitter markaların, kamu kuruluşlarının, devlet adamlarının, sporcu ve birçok ünlü kişinin de içinde olduğu 1.3 milyar kullanıcıya ulaşarak büyük bir bilgi üretim, dağıtım ve tüketim platformu haline gelmiştir. Platform üzerinde aylık ortalama 330 milyon aktif kullanıcı yer almakta ve her gün 500 milyon Tweet atılmaktadır (Ahlgren ve ark. 2021). Bu yapısı ile önemli bir haber kaynağı ve dağıtım kanalı haline gelmiştir. Birçok şirket ürün veya hizmetleri hakkında fikir sahibi olmak adına Twitter’i dijital pazar olarak kullanmakta ve her gün Twitter‘de ortalama 164 milyon reklam gösterilmektedir. Yine politikacılar ya da ünlü isimler platform sayesinde geniş kitlelere ulaşma imkânı bulmaktadır (Akyul, 2019).

Günümüzde artık vazgeçilmez hale gelen Twitter insanların bilgiye ulaşma, eğlence, sosyalleşme, bir durum veya ürün hakkında fikir beyan etme, toplumda meydana gelen olaylar karşısında duygu ve düşüncelerini diğer insanlarla paylaşma gibi ihtiyaçlarını karşılamaktadır.

Twitter API Search, Rest ve Stream kütüphanelerini içeren farklı özelliklere sahiptir. Bu özellikler dışında program geliştiriciler tarafından kullanılan başka yazılımlarda bulunmaktadır. Twitter bünyesindeki verilerin yani Tweet’lerin elde edilmesi bu tür yazılımlarla oldukça kolay hale gelmiştir. Bunun yanı sıra farklı parametreler (tarih, dil, konum vb.) ile sorgu gönderme, Twitter mesaj paylaşımı, diğer kullanıcıların gönderdiği mesajları liste olarak görme, gibi birçok adımı sıralamak mümkündür. Yapılan bu sorgular Twitter jargonunda var olan hashtag(#), mention(@), kelime ve sözcük içerebilmektedir. Fakat gizlilik ve güvenlik politikası gereği Twitter yalnızca herkesin erişimine açık olan tweetlere erişim izni vermektedir (Çoban, 2016).

Büyük verilerin hızlı, otomatik ve sürekli bir şekilde elde edilmesini sağlayan internetten veri kazıma işlemi iki adımda yapılabilir. Bu adımlardan ilki internet kaynaklarının elde edilmesi, diğeri ise elde edilen verilerden istenen bilgilerin çıkarılması işlemidir (Zhao, 2017).

BeautifulSoup kütüphanesi, HTML kodundaki bilgileri ayıklamak ve çıkarmak için kullanılan bir modüldür. Selenium ise, web tarayıcısı ile etkileşim kurarak kimlik doğrulama, yönlendirme, çerezler gibi HTTP isteklerini gerçekleştirmeye ve sosyal medya platformlarına otomatik giriş yapmaya olanak tanıyan bir Python kütüphanesidir. Temelde tarayıcı sürücüleri (web driver) üzerinden komutları takip ve kabul eden, bunları bir tarayıcıya gönderen otomasyon çerçevesidir (Zhao, 2017). Bu çalışmada, tarayıcı sürücülerinden Google Chrome seçilmiştir. Bu bağlantılarda önemli olan bir diğer faktör ise kullanılan tarayıcının güncel olmasıdır.

**3.7. Speech-to-Text Fonksiyonu**

Speech to Text fonksiyonu, konuşulan dilin yazılı metne dönüştürülmesini sağlayan ileri bir teknolojidir. Bu fonksiyon, sesli komutları, konuşmaları veya diğer sesli girdileri gerçek zamanlı olarak yazılı metin formatına çevirir. Fonksiyon, kullanıcının sesini algılar, tanır ve bunu yazıya döker. Sesin tanınması ve doğru şekilde yazıya aktarılması, ses tanıma algoritmalarının karmaşıklığı ve doğruluğu sayesinde mümkün olur.

Bu teknolojinin temel amacı, kullanıcıların konuşarak bilgisayar veya mobil cihazlarla etkileşimde bulunmalarını sağlamaktır. Kullanıcılar, metin yazma gereksinimini ortadan kaldırarak hızlı not alma, mesaj yazma veya belge oluşturma gibi işlemleri kolaylaştırır. Örneğin, bir öğrenci ders sırasında önemli notları hızlıca kaydedebilir veya bir iş profesyoneli, toplantı sırasında konuşulanları anında metin haline getirebilir.

Ayrıca, Speech to Text fonksiyonu, işitme engelli veya görme engelli bireyler için büyük bir erişilebilirlik sağlar. İşitme engelli bireyler, konuşmaları metne dönüştürerek kolayca okuyabilirken, görme engelli bireyler, konuşarak bilgisayar veya mobil cihazları etkin bir şekilde kullanabilirler. Bu sayede, teknolojinin engelli bireyler için sunduğu erişilebilirlik çözümleri, onların günlük yaşamlarını ve iş hayatlarını önemli ölçüde kolaylaştırır.

Bu teknoloji, müşteri hizmetlerinde de geniş çapta kullanılmaktadır. Çağrı merkezlerindeki müşteri hizmetleri temsilcileri ile yapılan görüşmeler yazıya dökülerek analiz edilebilir ve geri bildirim için kullanılabilir. Bu, müşteri memnuniyetini artırmak ve hizmet kalitesini yükseltmek için değerlidir. Ayrıca, Speech to Text teknolojisi, ders kayıtları, toplantı notları ve diğer birçok alanda da kullanılabilir. Örneğin, bir konferans sırasında yapılan konuşmalar kaydedilip yazıya dökülebilir ve daha sonra referans alınabilir.

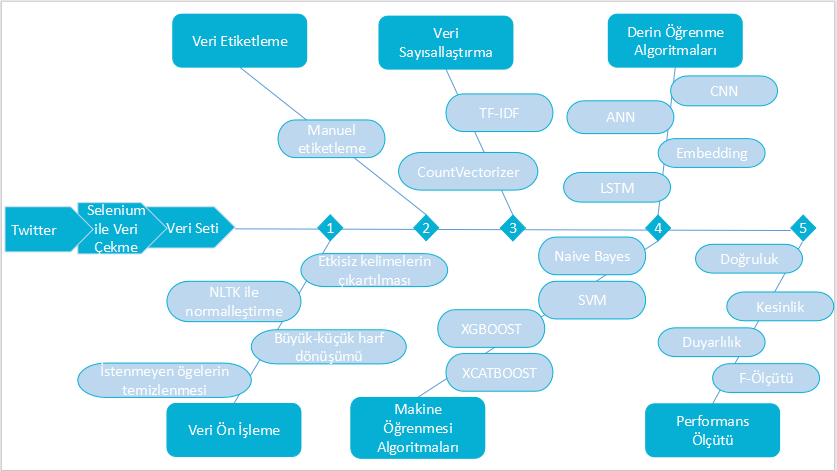
Diğer yandan, dil öğrenme süreçlerinde de bu teknoloji oldukça faydalıdır. Dil öğrenen bireyler, konuşmalarını yazılı metne dönüştürerek dil hatalarını ve telaffuzlarını gözden geçirebilirler. Bu, dil öğrenme sürecini hızlandırır ve geliştirir. Ayrıca, multimedya ve yayıncılık alanında da önemli bir rol oynar. Video ve podcast içerikleri yazılı metne dönüştürülerek, altyazı veya transkript olarak kullanılabilir. Bu, içeriğin daha geniş bir kitleye ulaşmasını sağlar ve arama motorları tarafından daha iyi indekslenmesine yardımcı olur.

Büyük miktarda sesli veri, yazılı metin olarak daha kolay aranabilir ve organize edilebilir. Bu, özellikle büyük veri setleri ile çalışan araştırmacılar ve analistler için oldukça değerlidir. Özetle, Speech to Text teknolojisi, konuşulan dilin doğru ve hızlı bir şekilde metne dönüştürülmesini sağlayarak, çeşitli alanlarda verimliliği ve erişilebilirliği artırır.

**4. BULGULAR VE TARTIŞMA**

**4.1.Çalışmanın Mimari Yapısı**

Bu çalışmada, çeşitli konularda atılan Türkçe Tweet’ler elde edilerek, yapılan paylaşımlar olumlu ve olumsuz olarak manuel olarak etiketlendirilmiştir. Etiketleme işlemleriyle oluşturulan bu model ile, farklı veri sayısallaştırma ( TF-IDF, CountVectorizer ) yöntemleri uygulanmıştır. Ardından farklı makine öğrenmesi algoritmaları (Naive Bayes, SVM, XGBOOST, CATBOOST) ve farklı derin öğrenme algoritmaları (ANN, CNN, LSTM, Embedding) kullanılarak en iyi performansı gösteren modeli üzerinden duygu analizi yapılması amaçlanmıştır.



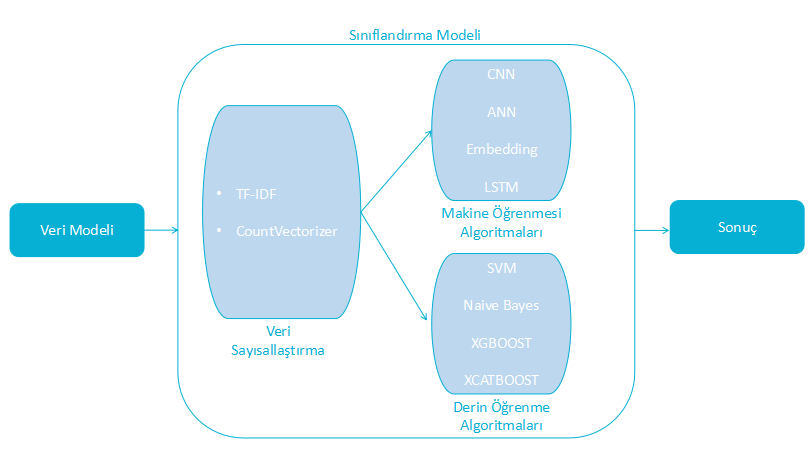
**Şekil 4.1.** Çalışma sisteminin mimari yapısı.

Geliştirilen sisteminin ilk aşamasında elde edilen veri seti çeşitli ön işlem adımlarına tabi tutularak daha anlamlı ve ölçülebilir bir hale gelmesi amaçlanmıştır. Bu doğrultuda ham hali ile elde edilmiş Tweet’ler ilk olarak küçük harfe dönüştürülmüş ardından özel karakter, emoji, rakam gibi istenmeyen öğeler temizlenmiştir. Veri ön işlemenin son adımı olarak tekrar eden satırlar ve 40 karakterden küçük tweetler veri setinden çıkartılmıştır.

İkinci aşamada, makine öğrenmesi ve derin öğrenme algoritmalarına girdi olarak kullanılacak veri seti için sistemin ikinci aşamasında atılan Tweet’in hangi duyguyu ifade ettiği belirlenerek manuel etiketleme işlemi gerçekleştirilmektedir.

Üçüncü aşamada, TF-IDF ve CountVectorizer ile metinler sayısal ifadelere dönüştürülerek temsil edilmektedir. Bu aşamada farklı sayısallaştırma yöntemlerinin makine öğrenmesi ve derin öğrenme algoritmaları ile kullanımında sınıflandırmaya etkisi amaçlanmıştır.

Sistemin dördüncü adımında, oluşturulan modelin farklı sayısallaştırma yöntemleri ve farklı makine ve derin öğrenmesi algoritmaları (Naive Bayes, SVM, XGBOOST, CATBOOST, ANN, CNN, LSTM, Embedding) ile birlikte kullanılarak karşılaştırılması ve beşinci adımda da bu karşılaştırma sonucunda performans ölçütlerine bakılarak en iyi sınıflandırma modelinin tespit edilmesi amaçlanmıştır.



**Şekil 4.2.** Sınıflandırma modelinin mimari yapısı.

**4.2.Çalışmada Kullanılan Veri ve Sesli Verinin Kullanımı**

Çalışmada veri kaynağı olarak, 2021 yılının ilk çeyreği itibarıyla dünyanın en büyük sosyal medya platformu olan Twitter seçilmiştir. Twitter, bu dönemde günlük 199 milyon aktif kullanıcıya sahiptir (Twitter, 2021).

Veriler Twitter’de herkese açık olarak paylaşılan Tweet verilerinden toplanmıştır. Bu veri seti için alana göre herkese açık olarak paylaşılan Tweetler’de sadece Türkçe paylaşılan belirli konulardaki veriler saklanmıştır. Bu platformda 1 Ocak 2020 ve 31 Aralık 2021 tarihleri arasında arasında yayınlanan “olimpiyat”, “sanat”, “futbol”, “salgın”, “telefon”, “yapay zeka”, “Türkiye”, “ekonomi”, “voleybol”, “savaş” konularıyla ilgili Tweet’ler, Retweet’ler ve bu Tweet’lere yapılan yorumlar toplanmıştır.

Son olarak, toplanan veriler daha sonra ön işleme ve duygu sınıflandırma adımlarında kullanılmak üzere CSV dosyasına dönüştürüldü.

Sonuç olarak, 1 Ocak 2020 ve 31 Aralık 2021 tarihleri arasında 30.000 Türkçe Tweet’ten oluşan bir veri seti toplanmıştır. Bu veri setinin doğruluğunu sağlamak amacıyla bir sonraki aşamada detayları verilen ön işleme aşaması gerçekleştirildi.

Speech to Text fonksiyonu, konuşulan dili yazılı metne dönüştüren gelişmiş bir teknolojidir. Bu fonksiyon, sesli komutları, konuşmaları veya diğer sesli girdileri anında yazılı metin formatına çevirir. Kullanıcının sesini algılar, tanır ve bu sesi metne dönüştürür. Sesin tanınması ve doğru şekilde yazıya aktarılması, kullanılan ses tanıma algoritmalarının karmaşıklığı ve doğruluğu sayesinde gerçekleştirilir. Çalışmada, Speech-to-Text fonksiyonu kullanılarak anlık olarak alınan sesli verinin kullanımı da sağlanmıştır. Alınan bu sesli veri fonksiyon sayesinde metne dönüştürülür. Alınan bu metin projede girdi olarak kullanılacaktır.

**4.3. Çalışma Ortamı ve Kullanılan Kütüphaneler**

Son yıllarda makine öğrenimi ve veri biliminde oldukça popüler olan R, python, java gibi birçok programlama dili kullanılmaktadır. Bu çalışmada dinamik, hızlı, yerleşik kütüphaneleri ve veri madenciliği kabiliyetleri ile programlama dili olarak Python tercih edilmiştir. Genel amaçlı bir programlama dili olan Python, 1991 yılında Guido Van Rossum tarafından ilk sürümü ortaya çıkarılmıştır. Kullanım kolaylığı ve geniş kütüphane desteği ile günümüzde oldukça popülerleşmiş ve geniş bir kullanıcı kitlesine sahip olmuştur (Malkoç, 2012).

Python kodlaması için localde (PyCharm, SPYDER, PYDEV vb. ideler) veya bulut sistemlerde (Jupyter Notebook, Google Colab vb.) kullanabilecek birçok geliştirme ortamı mevcuttur. Geliştirme ortamı olarak da Jupyter Notebook tercih edilmiştir. Ortam kurulumu gerektirmeden önceden yapılandırılmış hazır paketler ile Jupyter Notebook, makine öğrenimi ve derin öğrenme gibi çalışmalarda oldukça yaygın olarak kullanılmaktadır. Aynı zamanda ücretsiz bir bulut hizmeti olan Jupyter Notebook, Drive ve Github gibi ortamlara kolay paylaşım imkanı da sunmaktadır.

Python içerisinde kolaylıkla kullanılabilen ve geliştiricilerin veri madenciliği, makine öğrenmesi gibi çalışmalarda sıklıkla kullandığı Tensorflow, Keras, Pandas, Numpy, Snscrape, SciKit-Learn, NLTK, Matplotlib gibi açık kaynaklı birçok kütüphane mevcuttur. Bu tez çalışmasında kullanılan bazı python kütüphaneleri aşağıdaki gibidir:

* **NLTK:** DDİ ve metin madenciliğinde popüler olarak kullanılan bir Python kütüphanesidir. Kütüphane içerisinde barındırdığı paketler sayesinde köke indirgeme, durak kelimeler, metni kelimelere ayırma gibi işlemler kolaylıkla uygulanabilmektedir. Duygu analizi, sınıflandırma, kaynak oluşturma, ayrıştırma, etiketleme, anlamsal akıl yürütme ve otomatik özet oluşturma gibi insan diliyle ilgili diğer birçok işlem için ayrıca NLTK kütüphanesinden faydalanılabilmektedir (Mehta, 2021).
* **Numpy:** Matrisler ve çok boyutlu dizi nesneleri üzerinde sıralama, şekil işleme, matematiksel ve mantıksal işlemlerde çeşitli kolaylıklar sağlayan, bilimsel hesaplamalar için temel bir Python kütüphanesidir. Kütüphane ayrıca dizi formatındaki matematiksel işlemlerin yürütme süresini hızlandırır, vektörlere dönüştürülme performansını artırır (Anonymous, 2021).
* **Pandas:** İlişkisel veya etiketlenmiş verilerle kolay ve sezgisel olarak çalışabilmeyi sağlamak için geliştirilen Pandas, veri yapılarının dataframe nesnelerine dönüştürülmesine, dataframe üzerinde sütunların oluşturulması ve kaldırılmasına, eksik verilerin işlenmesine olanak tanır. Veri ön işleme, veri yönetimi ve analizi için hızlı ve kolay bir araçtır (Anonymous, 2021).
* **Matplotlib:** Veri Şekilleştirmelerinde kullanılan standart bir Python kütüphanesidir. Gelişmiş Şekilleştirmeler için yetersiz kalan Matplotlib sağladığı nesne yönelimli API ile iki boyutlu diyagram ve grafikleri uygulamalara gömmeyi sağlar (Custer, 2020).
* **Scikit-learn:** Makine öğrenimi algoritmalarını (doğrusal regresyon, LR, karar ağaçları, rastgele orman gibi) içeren bir Python kütüphanesidir. Bu kütüphanenin bu kadar popüler olmasının nedeni ihtiyaç duyulan temel yöntemlerin büyük çoğunluğunu içeriyor olmasıdır. Scikit-learn ile veride yer alan eksik değerleri doldurmak, özniteliklere karar vermek, çapraz doğrulama yapmak, sonuç çıktılarını değerlendirmek gibi veri analitiği çalışmalarının baştan sona yürütülmesini mümkün kılmaktadır (Yüceoğlu, 2017).
* **Snscrape:** Twitter geliştirici hesabına ihtiyaç duyulmadan, herhangi bir istek sınırı ya da kısıtlama olmaksızın Tweet’lerin toplanabileceği bir kütüphanedir. Bu kütüphaneyi kullanabilmek için Python 3.8 veya üzeri sürüm gerekmektedir (Beck, 2020).
* **Seaborn:** İstatistiksel modelleri Şekilleştirmek, verileri özetlemek ve genel dağılımlarını göstermek için makine öğrenimi aracı olarak hizmet eden bir Python kütüphanesidir. Matplotlib'i temel alan bu kütüphane geniş bir Şekilleştirme galerisi sunmaktadır (Custer, 2020).

**4.4. Twitter Verilerine Erişim**

Twitter üzerinden verilerin çekilebilmesi için Twitter geliştirici hesabı ile Twitter API‘ye erişim sağlamak gerekmektedir. 2018 yılına kadar API‘ye erişip Tweet’leri çekmek mümkündü ancak 2018 yılından sonra geliştirici hesabı için Twitter’e başvuruda bulunup başvurunun onaylanması için de belirli bir süre beklemek gerekmektedir. Onay işleminden sonra API‘ya erişimde Python dili için geliştirilmiş “tweepy” kütüphanesi kullanılarak uygulamaya özel Consumer Key, Consumer Screet, Access Token ve Acces Token Secret değerleri ile Twitter bağlantısı sağlanarak veriler çekilebilmektedir. Twitter API anlık olarak son yedi günün verilerine erişim izni vermekle birlikte her 15 dakikada 180 istek gibi bazı kısıtlamalar da koymaktadır.

Bu tez çalışmasında veri setinin büyüklüğü, çekilmek istenen verinin zaman aralığı göz önüne alındığında bahsedilen kısıtlamalardan bağımsız dilediğimiz kadar Tweet’in çekilebildiği Selenium kütüphanesi tercih edilerek, istenilen tarihler arasında anahtar kelimeler girilip alıntı olmayan Türkçe Tweet’ler kullanılmıştır.

Twitter’den veri çekme işleminde tarayıcı sunucu üzerinden işlemlerin optimizasyonu için Selenium kütüphanesi, sayfa kaynağını indirmek için BeautifulSoup ve kazıma sonucu elde edilen verileri bir veri çerçevesine aktarmak için Pandas kütüphaneleri eklenir. Tarayıcı sürücünün başlatılması işlemi için Google Chrome sunucularına bağlanılarak Twitter uygulamasına ait internet sayfasına bağlanılır. Oturum açma işlemi için kullanıcı adı ve şifrenin koda eklenmesi ile internet sayfasında uygun girişlere gerekli bilgilerin girilmesi işlemi gerçekleştirilir. Voleybol anahtar kelimesi (örneğin, “voleybol” geocode:39.1070000,39.5472000,30km and “voleybol”, since:2020-01-01 until:2021-12-31) enlem, boylam, arama çapı ve tarih aralığı gibi bilgilerle birlikte kullanılarak kodun arama sonuçlarını açması işlemi gerçekleştirilir. Sayfa kaynağını yüklemek amacıyla BeautifulSoup kütüphanesi kullanılarak XML formatında sayfanın içeriği indirilir. Bu şekilde, istenilen anahtar kelime ile elde edilen arama sonuçlarının sayfa kaynağına ulaşılır. Sayfa kaynağı içerisinden XPath’lar sayesinde istenilen metin ve konum bilgilerinin elde edilmesi için ortam hazırlanır.

Twitter’den alınacak iletilerin bir liste içine eklenmesi için ve döngülerde kullanılan döngü tekrar başlangıç sayısı ve maksimum döngü tekrar sayısının tanımlanması işlemi gerçekleştirilir. İlk sayfadaki iletilerin XPath’ları alınır ve sayfa kaynağındaki iletiler listeye eklenerek bir gönderi listesi oluşturulur. İletilerin gönderi listesine eklenmesi işlemi gerçekleştirilir. İlk gönderiden başlayarak sayfadaki tüm iletilerin teker teker listeye eklenmesi için bir döngü oluşturulur. İlk sayfadan alınan iletilerin ardından, diğer sayfalardan da iletilerin alınması için otomatik kaydırma işlemi gerçekleştirilir. Tekrarlayan gönderilerden arındırılmış bir liste oluşturmak amacıyla yeni bir gönderi listesi oluşturulur.

Yeni gönderi listesine, eski gönderi listesinde birden fazla aynı olan iletilerden yalnızca biri eklenir. Böylece tekrarlı iletilerden arındırılır.Gönderi listesinin 200 iletiyi aşıp aşmadığı kontrol edilir. 200 adet ileti listeye alındıktan sonra kod çalışmayı durdurur.Eğer 200'den az ileti varsa döngü 10 kez çalıştırılır. Döngü sayısı maksimuma ulaştığında kod çalışmayı durdurur. Döngüden her gelen ileti gönderi listesine eklenir. Tekrarlanan gönderilerden arındırılmış gönderi listesi bir veri çerçevesine yazdırılır.

Veri setinin oluşturulmasında Selenium kütüphanesi ile Twitter‘e erişim sağlanarak, “olimpiyat”, “sanat”, “futbol”, “salgın”, “telefon”, “yapay zeka”, “Türkiye”, “ekonomi”, “voleybol”, “savaş” anahtar kelimeleri girilerek alıntı yapılmamış tweetler elde edilmiştir. Veri seti içerisinde konu ile alakalı olmayan, gündem oluşturmak için atılan ilgili etiketli tweetler filtrelenerek veri setinin dışında tutulmuştur. 1 Ocak 2020 ve 31 Aralık 2021 tarihleri arasında atılan toplamda 30.000 Tweet KVKK/GDPR kapsamında sadece id, tweets, date başlıkları altında toplanarak veri seti oluşturulmuştur.

**4.5.Veri Ön İşleme**

Twitter sosyal medya platformundan toplanan veriler genellikle temiz değildir. Metin temizleme, metnin anlamının çıkarılması veya analiz edilmesi zor olan kelimeleri veya diğer bileşenleri temizlemek için yapılan metin madenciliği işlemlerinden biridir. Analiz aşamasında, deneysel sonuçları olumsuz yönde etkileyebilecek gereksiz birçok özel karakter, ifade, link, etiket ve emoji bulunur. Bu karakterler fazla bilgi sağlamadıkları gibi duygu analizi için de işlenmesi zor unsurlardır. Bu aşamada aşağıda verilen işlemler uygulanarak veri seti üzerinde ön işleme yapılmıştır:

• Tekrarlanan Tweet’ler veri setinden çıkartılır.

• Numaralar kaldırılır.

• Noktalama işaretleri kaldırılır.

• Twitter RT, @ ve cümlelerdeki bağlantılar temizlenir.

• Metin küçük harfe dönüştürülür, böylece “aşı” ve “Aşı” gibi kelimeler analizde aynı kelime olarak kabul edilir.

• Özel karakterler ve duygu ifade eden yüz ifadeleri (emojiler), kendilerine özgü kodları kullanılarak metin içinde tespit edildikten sonra kaldırılır.

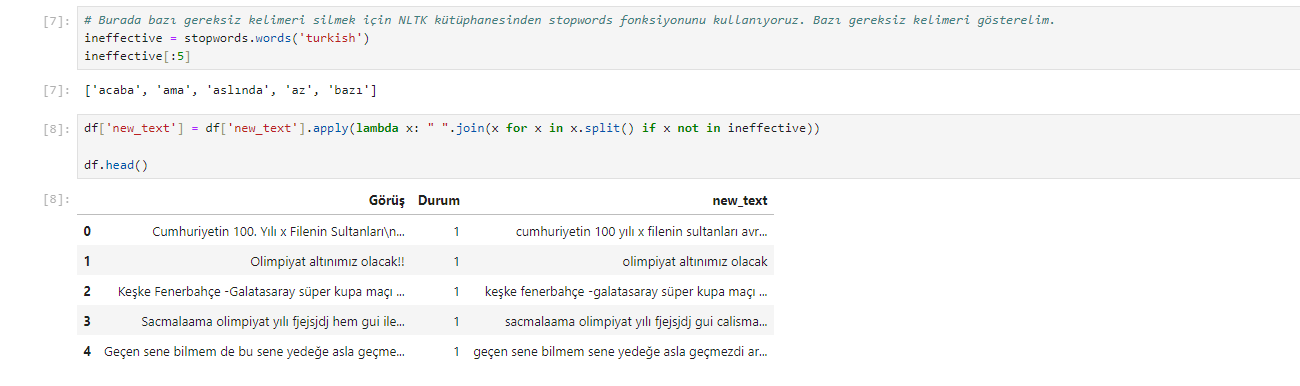
• Her bir tweet Google Çeviri Python kütüphanesi Googletrans kullanılarak İngilizce metne çevrilir. Googletrans, Google Translate API'sını uygulayan ücretsiz ve sınırsız bir Python kütüphanesidir. Bu kütüphane, metni anlama ve çevirme gibi yöntemler için Google Translate Ajax API'sını kullanır.

• Her bir Tweet için tokenleştirme işlemi uygulanır. Tokenleştirme, metin içindeki kelimelerin boşluklara göre ayrıştırılması işlemidir.

• Daha sonra Tweet’lerden “bir”, “bu”, “şu”, “filan” vb. gibi etkisiz kelimeler (stopwords) çıkartılır.

• Her bir Tweet içerisindeki kelimeler köklerine ayrılır. Lemmatizasyon kelimelerin morfolojik analizlerini kullanır. Örneğin, “doktorum” kelimesi “doktor” olarak güncellenir.

Önişleme uygulamasının amacı, metnin temel anlamını koruyarak metin içerisindeki kelime sayısını azaltmaktır. Toplanan ham veri setinde duygu analizi aşamasında kullanılmayacak gereksiz kelime ve ifadeler bulunmaktadır. Bu nedenle, herhangi bir veri analizi yapmadan önce ön işleme gerçekleştirilmesi gerekir.



**Şekil 4.3.** Etkisiz kelimelerin (stopwords) veri setinden çıkartılması.

**4.5.1. İstenmeyen öğelerin temizlenmesi**

Metinleri istenilen biçimde parçalanıp dizilere kaydetme işleminin yapıldığı bu adımda, NLTK kütüphanesi ile Tweet’ler; kelime kelime, tüm özel karakter ve noktalama işaretleri ile birlikte ayrılmıştır.

Bu adımda noktalama işaretleri ile birlikte Tweet içerisinde etiketlenen konu başlıkları ve bahsedilen kişiler (#uzaktanegitim #meb @alican @ktun gibi) metinlerden tamamen kaldırılmıştır. Böylece metinlerin daha sade ve anlamlı hale gelmesi sağlanmıştır.

Noktalama işareti, kişi ve etiketlerin ön işleminden sonra Tweet içerisinde yer alan rakam, özel karakter, emoji, link gibi ifadelerin temizlenmesi sağlanmıştır.



**Şekil 4.4.** Veri setinden özel karakter, emoji, url ve rakamların temizlenmesi.

**4.6. Veri Etiketleme**

Twitter‘den elde edilen veriler ön işlemden geçirilmiş ve 10.853 Tweet duygu analizi için kullanılabilir hale getirilmiştir. Bu aşamada atılan Tweetler’den 8.683 adet rastgele seçilerek makine öğrenmesi modellerinde kullanılabilmesi için etiketleme işlemi ile hangi duyguyu ifade ettiği belirlenmiştir. Etiketleme işlemi; atılan Tweet olumluysa olumlu, olumsuzsa olumsuz olarak yapılmıştır.

**4.7. Verilerin Ayrılması ve Modelleme**

Tezin önceki bölümlerinde veriler uygun forma sokularak gereken ön modelleme aşamaları tamamlanmıştır. Bu bölümde de en iyi sınıflandırma modelini bulabilmek için etiketlendirilen veri setleri TF-IDF ve CountVectorizer ile ayrı ayrı sayısal olarak temsil edilerek Python içerisinde bulunan “sklearn” kütüphanesinde yer alan “train\_test\_split” metodu ile iki parçaya ayrılmıştır. Birinci parça ile veriler eğitilirken ikinci parça ile doğrulu test edilmiştir. Çalışmada “random\_state” değeri 42, “test\_size” değeri 0.30 alınarak veri setinin %80‘i eğitim, %20‘u test için uygulanmıştır.

Tez kapsamında sınıflandırma işlemi için etiketlendirilen metinlerin olumlu, olumsuz veya nötr duygu dağılımları ile birlikte ayrıca manuel olarak işaretlenmiş nötr etikete sahip metinlerin veri setinden çıkartılması ile sadece olumlu ve olumsuz olmasına bakılarak makine öğrenmesi ve derin öğrenme algoritmalarının performansları karşılaştırılmıştır. Olumlu, olumsuz veya nötr gibi çoklu sınıflandırma problemlerinde ikili sınıflandırma algoritmalarını kullanmak için elimizdeki veri setini çoklu ikili sınıflandırma veri setine bölmek ve her birini bir ikili sınıflandırma modeline sığdırmak gerekmektedir. Bire karşı hepsi (OneVsRest) bu yaklaşım için kullanılan bir modeldir. Model çok sınıflı bir sınıflandırmayı sınıf başına bir ikili sınıflandırma problemine böler daha sonra her bir ikili sınıflandırma problemi için bir ikili sınıflandırıcı eğitilir, en güvenilir model kullanılarak tahminler yapılır.

TF-IDF ve CountVectorizer veri sayısallaştırma yöntemleri ile ANN, CNN, LSTM, Embedding derin öğrenme metodları ve Naive Bayes, SVM, XGBOOST, CATBOOST makine öğrenmesi algoritmaları kullanılarak sınıflandırma modelleri oluşturulmuştur. Manuel etiketlendirilerek oluşturulan veri setleri bu aşamada girdi olarak kullanılmış, her birinin sınıflandırma işlemlerindeki performansları değerlendirilmiştir. Sınıflandırma modellerinden hangisinin en yüksek başarıya sahip olduğuna karar verebilmek için doğruluk oranları karşılaştırılmıştır.

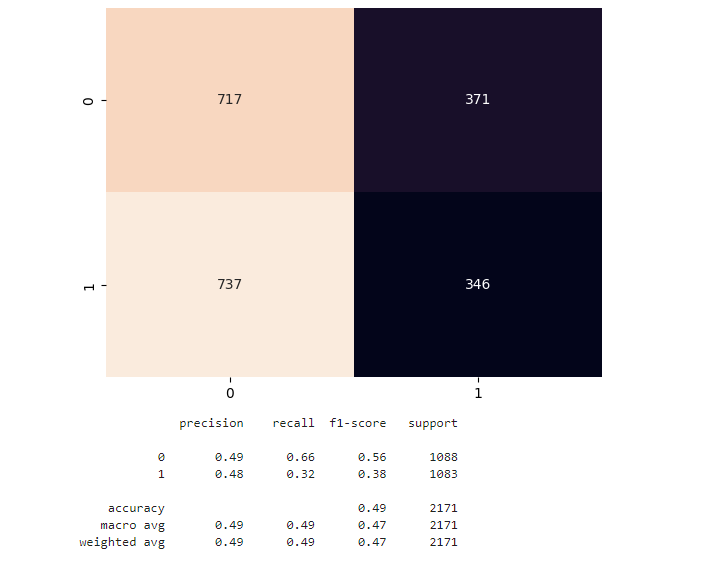
ANN derin öğrenme algoritması kullanılarak oluşturulan sınıflandırma modelinin kesinliği 0.49, duyarlılığı 0.66, F-ölçütü 0.56 ve doğruluğu 0.49’dur (Şekil 4.5.).

CNN derin öğrenme algoritması kullanılarak oluşturulan sınıflandırma modelinin kesinliği 0.49, duyarlılığı 0.66, F-ölçütü 0.56 ve doğruluğu 0.49’dur.

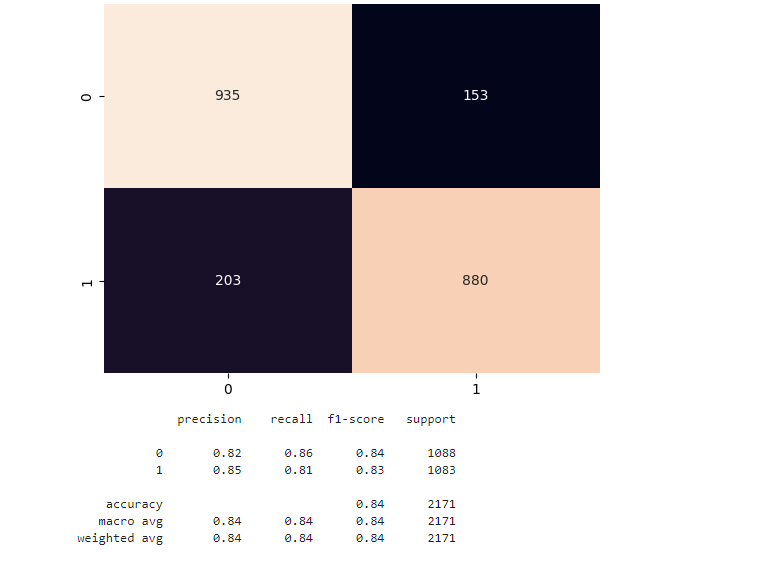
LSTM derin öğrenme algoritması kullanılarak oluşturulan sınıflandırma modelinin kesinliği 0.49, duyarlılığı 0.66, F-ölçütü 0.56 ve doğruluğu 0.49’dur.

Embeddding derin öğrenme algoritması kullanılarak oluşturulan sınıflandırma modelinin kesinliği 0.85, duyarlılığı 0.86, F-ölçütü 0.84 ve doğruluğu 0.84’tür (Şekil 4.6.).

Oluşturulan her bir modelin karışıklık matrisine aşağıda yer verilmiştir.



**Şekil 4.5.** ANN derin öğrenme algoritmasıyla oluşturulan modele ait karışıklık matrisi.

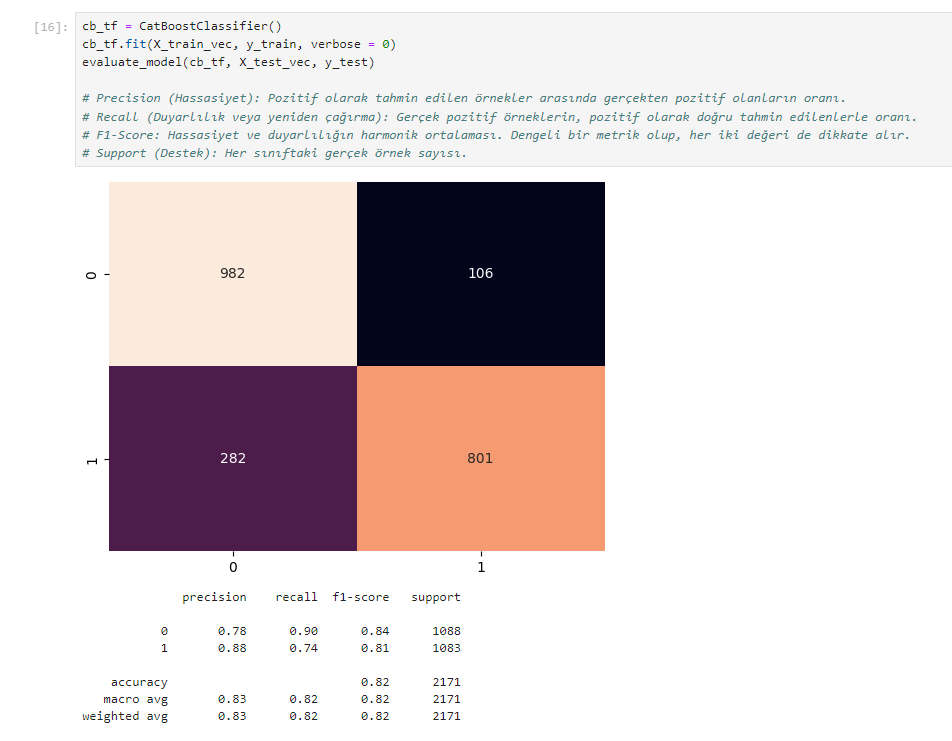


**Şekil 4.6.** Embedding derin öğrenme algoritmasıyla oluşturulan modele ait karışıklık matrisi

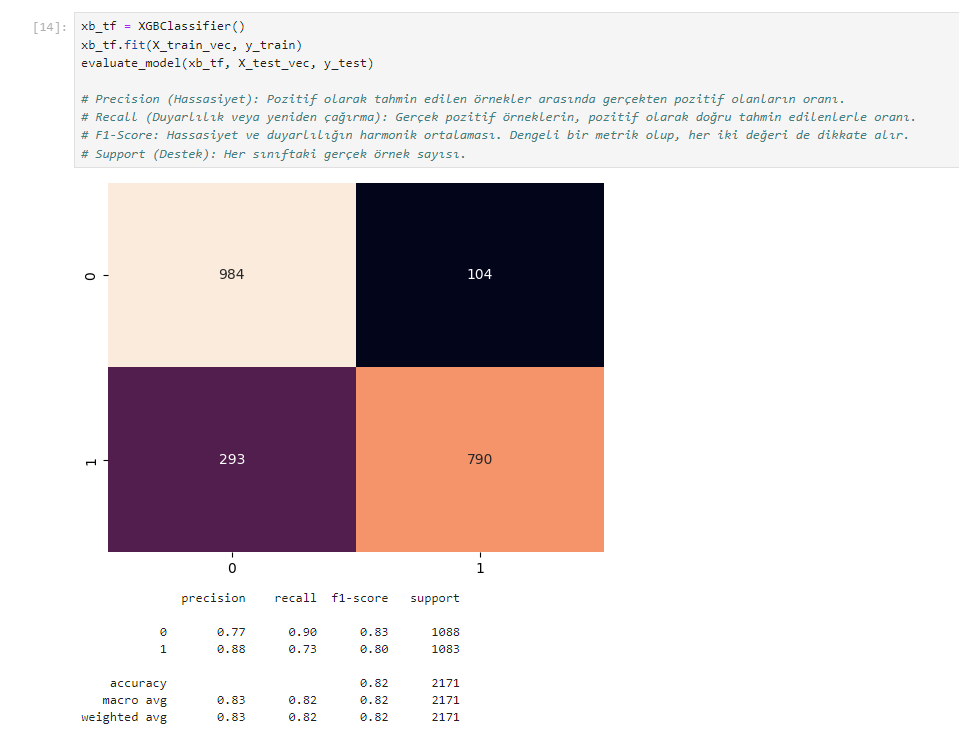
CATBOOST makine öğrenmesi algoritması kullanılarak oluşturulan sınıflandırma modelinin kesinliği 0.88, duyarlılığı 0.90, F-ölçütü 0.84 ve doğruluğu 0.82’dir (Şekil 4.7. ).

XGBOOST makine öğrenmesi algoritması kullanılarak oluşturulan sınıflandırma modelinin kesinliği 0.88, duyarlılığı 0.90, F-ölçütü 0.83 ve doğruluğu 0.82’dir (Şekil 4.8. ).

Oluşturulan modellere ait karışıklık matrisine aşağıda yer verilmiştir.



**Şekil 4.7.** CATBOOST makine öğrenmesi algoritmasıyla oluşturulan modele ait karışıklık matrisi.

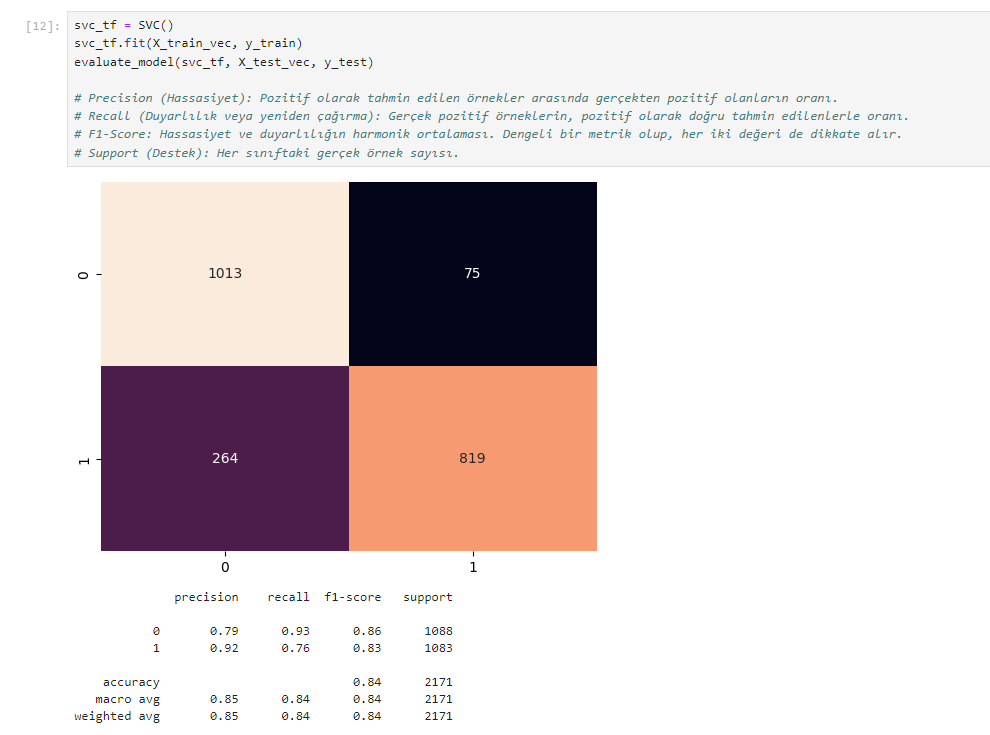


**Şekil 4.8.** XGBOOST makine öğrenmesi algoritmasıyla oluşturulan modele ait karışıklık matrisi.

SVM makine öğrenmesi algoritması kullanılarak oluşturulan sınıflandırma modelinin kesinliği 0.92, duyarlılığı 0.93, F-ölçütü 0.86 ve doğruluğu 0.84’dur (Şekil 4.9.).

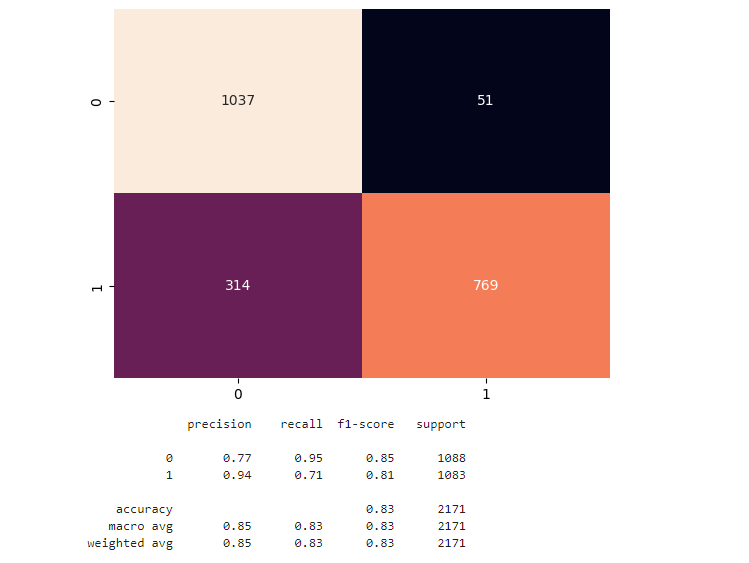
Naive Bayes makine öğrenmesi algoritması kullanılarak oluşturulan sınıflandırma modelinin kesinliği 0.94, duyarlılığı 0.95, F-ölçütü 0.85 ve doğruluğu 0.83’dur (Şekil 4.10. ).

Oluşturulan modellerin karışıklık matrisine aşağıda yer verilmiştir.



**Şekil 4.9.** SVM makine öğrenmesi algoritmasıyla oluşturulan modele ait karışıklık matrisi.

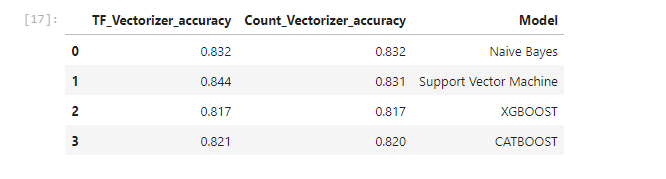
.



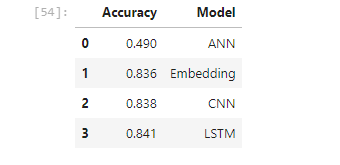
**Şekil 4.10.** Naive Bayes makine öğrenmesi algoritmasıyla oluşturulan modele ait karışıklık matrisi.

**4.7.1.Modellerin doğruluklarının karşılaştırılması**

Türkçe metinlerden oluşan ve manuel olarak etiketlendirilen modeller, TF-IDF ve CountVectorizer ile sayısallaştırılarak Naive Bayes, SVM, XGBOOST, CATBOOST makine öğrenmesi algoritmalarının yanı sıra ANN, CNN, LSTM, Embedding gibi derin öğrenme metodları ile her bir model sınıflandırma işlemine tabi tutulmuştur. Sınıflandırma işlemi sonrasında modellere ait doğruluk oranlarını elde edilmiştir (Şekil 4.11. , Şekil 4.12.). Tablo üzerinde ondalık sayılar sıfırdan sonra üç basamağa yuvarlanmış ve modellere ait doğruluk oranları sıralanmıştır.



**Şekil 4.11.** Makine öğrenmesi algoritmaları başarı karşılaştırması.



**Şekil 4.12.** Derin öğrenme algoritmaları başarı karşılaştırması.

Şekil 4.11. ve Şekil 4.12’da yer alan doğruluk oranları incelendiğinde makine öğrenmesi algoritmaları arasında TF-IDF – SVM ile 0.844‘lük, derin öğrenme algoritmaları arasında LSTM ile 0.841’lik en yüksek doğruluk oranı elde edilmiştir.

Veri sayısallaştırma işlemleri karşılaştırıldığında TF-IDF ile sayısallaştırma işleminde makine öğrenmesi algoritmalarının başarı oranının CountVektorizer’e göre kısmen daha iyi sonuçlar verdiği gözlemlenmiştir.

**4.8. Arayüz Oluşturma**

Bu proje, makine öğrenmesi ve derin öğrenme modelleri kullanılarak duygu analizi gerçekleştirmek amacıyla geliştirilmiştir. Projeye entegre edilen Speech-to-Text fonksiyonu sayesinde, hem sesli hem de yazılı metin girdileri kullanılarak geliştirilen modelin test edilmesi mümkündür.

Kullanıcı dostu arayüzü ile bu sistem, farklı duygu analizlerini hızlı ve etkili bir şekilde gerçekleştirmeye olanak tanır. Arayüzde, "Girilen Cümle" ve "Modelin Çıktısı" olmak üzere iki temel girdi alanı bulunmaktadır.

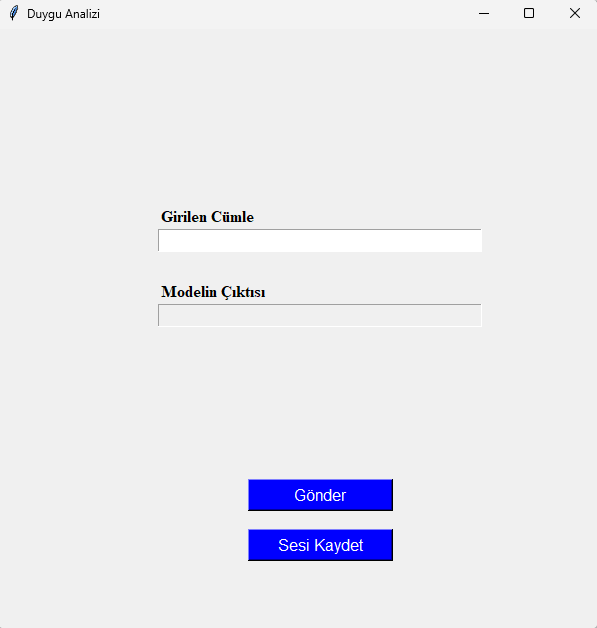
**Girilen Cümle:** Bu alan, kullanıcının analiz edilmek üzere metin girişi yapabileceği bölümdür. Kullanıcılar, bu alana doğrudan yazılı metin girebilir veya sesli olarak metin girişi yapabilirler.

**Modelin Çıktısı**: Bu bölümde, analiz edilen cümlenin duygu durumu sonucu görüntülenir. Model, girdi olarak alınan cümleyi işler ve duygu sonucunu “Olumlu” veya “Olumsuz” olarak burada gösterir.

Arayüzde ayrıca iki ana buton bulunmaktadır:

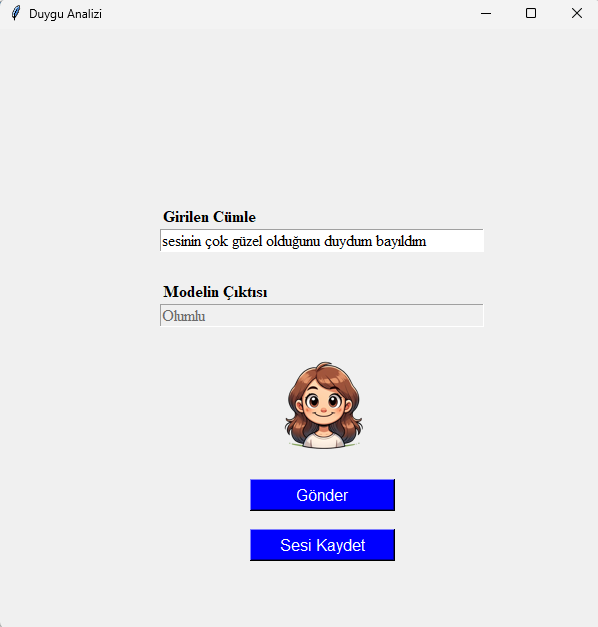
**Gönder Butonu:** Kullanıcı tarafından girilen yazılı metnin analiz edilmesi için kullanılır. Bu butona basıldığında, sistem girilen cümleyi model üzerinden analiz eder ve sonucu "Modelin Çıktısı" alanında “Olumlu” veya “Olumsuz” olrak gösterir.

**Sesi Kaydet Butonu:** Kullanıcılar sesli girdilerini bu butona basarak kaydedebilirler. Kaydedilen sesli veri, Speech-to-Text fonksiyonu kullanılarak yazılı metne dönüştürülür ve ardından duygu analizi için işlenir. Bu sayede, kullanıcılar sesli komutlarını metne çevirerek analiz yapabilirler.

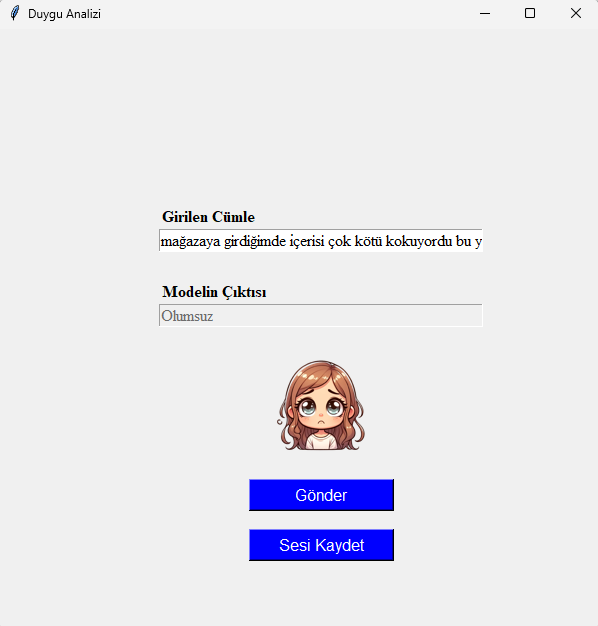
****

**Şekil 4.13.** Giriş ekranı.

Arayüzdeki görseller, analiz edilen duygu durumuna göre değişiklik göstermektedir. Bu, kullanıcıya görsel bir geri bildirim sağlayarak, analiz sonucunun daha anlaşılır ve etkili olmasını sağlar. Metin girilmesinin ardından “Gönder” butonuna basılmasıyla sistemin arayüzlerine ait görsellere aşağıda yer verilmiştir:

****

**Şekil 4.14.** Olumlu cümle girildiğindeki ekran.

****

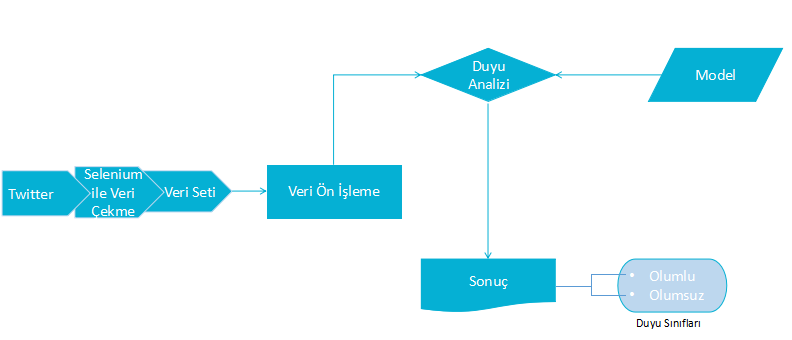
**Şekil 4.15.** Olumsuz cümle girildiğindeki ekran.

**4.9. Duygu Analizi ve Şekilleştirme**

Duygu analizi, temelde, metinlerin içerisinde gizlenmiş olan görüşlerin ve duyguların anlaşılması, bu görüşlerin uygun yöntemler ile olumlu veya olumsuz olarak etiketlenmesini amaçlamaktadır. Çalışmada, 1 Ocak 2020 ve 31 Aralık 2021 tarihleri arasında Twitter ortamından paylaşılan Türkçe yorumlar toplanmış ve veri kümesi üzerinde duygu analizi gerçekleştirebilmek için birçok modelden yararlanılmıştır.

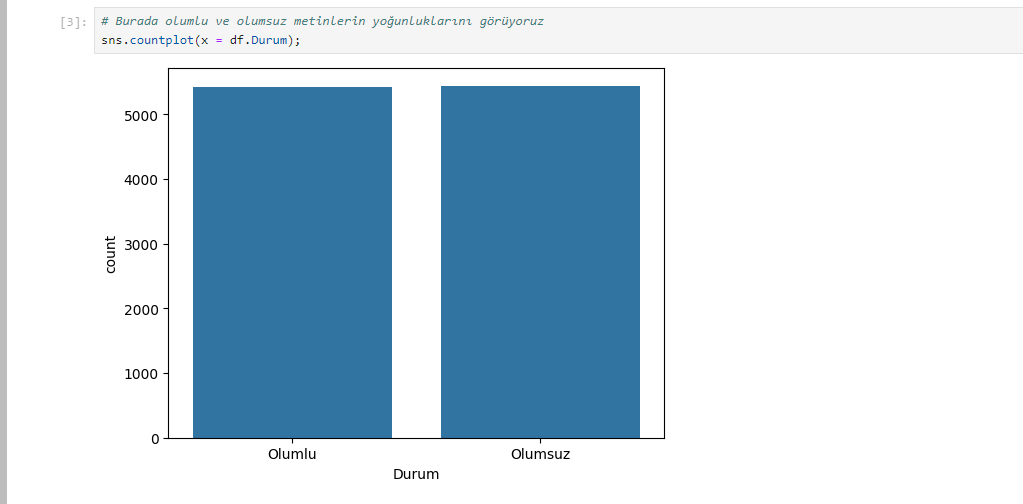
Modelin eğitilmesinden önce tüm veri kümesi alıntılama ve etiket temizleme, noktalama işaretlerinin temizlenmesi, kaçış karakterlerinin temizlenmesi, bağlantıların ve diğer web sitesi adreslerinin temizlenmesi, HTML öğelerinin temizlenmesi, yüz ve diğer ifade öğelerinin temizlenmesi, elektronik posta adreslerinin temizlenmesi gibi ön işleme aşamalarından geçerek temizlenmiş ve etiketlenmek üzere hazır hale getirilmiştir.

Türkçe metinlerden oluşan ve manuel olarak etiketlendirilen veri seti için TF-IDF – SVM sınıflandırma modelinin performansı en yüksek olmuştur. Bu model kullanılarak 10.853 tweet için tahminleme yapılmıştır. Paylaşılan tweetlerin olumlu veya olumsuz olarak hangi duyguyu ifade ettiği belirlenerek analiz gerçekleştirilmiştir. Çalışmada kullanılan duygu analizi mimarisi Şekil 4.16.‘daki gibidir.



**Şekil 4.16.** Çalışmada kullanılan duygu analizi mimarisi.

TF-IDF – SVM ile tüm veri setinde tahminleme yapıldığında 5413 adet olumlu ve 5440 adet olumsuz Tweet atıldığı sonucuna varılmıştır. Duygu sınıflarının dağılım oranı Şekil 4.16.’da gösterilmiştir.



**Şekil 4.17.** Duygu sınıflarının dağılım oranı.

**5.SONUÇLAR**

İnsanlar tarafından büyük ilgi gören ve oldukça aktif bir şekilde kullanılan sosyal medya platformları birçok alanda kullanılabilir büyük bir veri kaynağı haline gelmiştir. Ortaya çıkan bu büyük veri duygu analizi gibi çalışmaların temel veri kaynağı olmuştur. Duygu analizi uzun süredir çalışmalara konu olan bir alan olsa da Türkçe verilerle yapılmış çok fazla uygulamasının bulunmadığı, İngilizce dili ile yapılan çalışmaların literatürde oldukça büyük bir yere sahip olduğu gözlemlenmiştir.

Bu tez kapsamında sosyal medya platformu Twitter‘de paylaşılan uzaktan eğitim konulu Türkçe Tweet’ler çalışmanın veri setini oluşturmuştur. Bu veri seti çeşitli ön işlemlere tabi tutulmuş ve NLTK kütüphanesi ile normalleştirme işlemi gerçekleştirilmiştir. Verilerin ön işleme sonrasında anlamlı ve işlenebilir olması, NLTK kütüphanesi ile imla hataları ve anlam denetiminin yapılması sınıflandırma başarısında önemli bir rol oynadığı görülmüştür.

Çalışma kapsamında TF-IDF ve CountVectorizer yöntemleri kullanılarak veri sayısallaştırma işlemi gerçekleştirilmiştir. Verilerin farklı sayısallaştırma yöntemleri ile temsil edilmesi kullanılan makine öğrenmesi algoritmalarına bağlı olarak sınıflandırma performanslarında etkili olmuştur. Sayısal olarak ifade edilen ve sınıf etiketine sahip olan veri seti eğitim ve test için ayrılarak Naive Bayes, SVM, XGBOOST, CATBOOST makine öğrenmesi algoritmalarının yanı sıra ANN, CNN, LSTM, Embedding gibi derin öğrenme metodlarının performansları karşılaştırılmıştır. Kullanılan makine öğrenmesi ve derin öğrenme algoritmaları ve bu algoritmaların sahip olduğu parametrelerin doğru değerlerde seçimi performanslarının değerlendirilmesinde doğrudan etkili olduğu görülmüştür.

Bu tez çalışmasının sonunda, en iyi sınıflandırma modelinin tespit edilerek devamında bu model üzerinden duygu analizi gerçekleştirmek için önceki bölümlerde bahsedilen manuel olarak veri etiketleme işlemi gerçekleştirilmiş, bu model için farklı sayısallaştırma yöntemi ile farklı makine öğrenmesi ve derin öğrenme algoritmaları kullanılarak performansları karşılaştırılmıştır. Manuel olarak işaretlenmiş nötr etiketlerin veri setinden çıkartılarak oluşturulduğu modelde ikili sınıflandırma işleminde başarının arttığı gözlemlenmiştir.

Türkçe metinler ve manuel etiketlere sahip veri seti için en iyi sınıflandırma performansını gösteren sayısallaştırma ve makine öğrenmesi algoritmasının TF-IDF – SVM olduğu, en iyi derin öğrenme algoritması performansının LSTM olduğu sonucu çıkarılmıştır.

Sonuç olarak sayısallaştırma yöntemleri, makine öğrenmesi ve derin öğrenme algoritmaları kullanılarak Türkçe doğal dil öğrenimine etkisi mevcut araştırma kapsamında incelenmiş ve sonuçlar analiz edilmiştir. Bu çalışmadan elde edilen bulguların ve tanımlamaların, Türkçe metin işleme ve doğal dil işleme alanlarında derin öğrenme ve makine öğrenmesi konusunda ileride yapılacak çalışmalara yardımcı olacağı düşünülmektedir.

Bu tez kapsamında olumlu ve olumsuz duygu ifadeleri kullanılarak çalışılmıştır. Gelecek çalışmalarda üzgün, mutlu, kızgın, şaşkın gibi daha ayrıntılı duygu ifadeleri belirten çalışmalar yapılabilir.

İngilizce metinler için geliştirilmiş ve yüksek başarı oranlarına sahip duygu çıktıları veren birçok hazır model bulunmaktadır. Türkçe çalışmaların çok az olduğu bu alanda bahsedilen hazır modeller Türkçe metinler için de geliştirilebilir.

Veri setinin kalitesi ve verinin doğru etikete sahip olması makine öğrenmesi ve derin öğrenme algoritmalarının sınıflandırma performanslarını değerlendirmede önemli rol oynamaktadır. Türkçe metinlerde ön işleme adımlarının geliştirilip normalleştirilmesi, aynı zamanda etiketli veri örneklerinin sayısının çoğaltılması başarı oranları artırılabilir.

**6. KAYNAKÇA**

**[1]** Yue L, Chen W, Li X, Zuo W, Yin M. “A survey of sentiment analysis in social media”, *Knowledge and Information Systems*, 60(2): 617–663, 2019.

**[2]** Singh NK, Tomar DS, Sangaiah AK., “Sentiment analysis: a review and comparative analysis over social media”, *J. Ambient Intell. Hum. Comput*, 11(1): 97–117, 2020.

**[3]** Zhao J, Cao X., “Combining semantic and prior polarity for boosting twitter sentiment analysis”, *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, 832–837, 2015.

**[4]** Osmanoğlu UÖ, Atak ON, Çağlar K, Kayhan H, Can T., “Sentiment Analysis for Distance Education Course Materials: A Machine Learning Approach”, *Journal of Educational Technology and Online Learning*, 3, 31–48, 2020.

**[5]** Pervan N, Keleş HY., “Sentiment Analysis Using A Random Forest Classifier On Turkish Web Comments”, *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering*, 59(2), 69–79, 2020.

**[6]** Yurtalan G, Koyuncu M, Turhan Ç., “A polarity calculation approach for lexicon-based Turkish sentiment analysis”. *Turkish Journal of Electrical Engineering and Computer Science,* 27, 1325–1339, 2019.

**[7]** Karcioglu AA, Aydin T., ” Sentiment Analysis of Turkish and English Twitter Feeds Using Word2Vec Model”, *27th Signal Processing and Communications Applications Conference (SIU)*, 1–4, 2019.

**[8]** Rumelli M, Akkuş D, Kart Ö, Isik Z., “Türkçe Metinlerde Makine Öğrenmesi Algoritmaları ile Duygu Analizi”, *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, (IEEE, 1-5), 2019.

**[9]** Üçükkartal, H., “Twitter'daki verilere metin madenciliği yöntemlerinin uygulanması”, *ESTUDAM Bilişim Dergisi*, 1(2), 10-13, 2020.

**[10]** İlhan, N., & Sağaltıcı, D., “Twitter’da duygu analizi”, *Harran Üniversitesi Mühendislik Dergisi*, 5(2), 146-156, 2020.

**[11]** Tuzcu, S., “Çevrimiçi kullanıcı yorumlarının duygu analizi ile sınıflandırılması”, *ESTUDAM Bilişim Dergisi*, 1(2), 1-5, 2020.

**[12]** Beşkirli, A., Gülbandılar, E., & Dağ, İ., “Metin madenciliği yöntemleri ile twitter verilerinden bilgi keşfi” *ESTUDAM Bilişim Dergisi*, 2(1), 21-25, 2021.

**[13]** Atılgan, K., & Yoğurtcu, H., “Kargo firması müşterilerinin twitter gönderilerinin duygu analizi”, *Çağ Üniversitesi Sosyal Bilimler Dergisi*, 18(1), 31-39, 2021.

**[14]** Eyipınar, C., Büyükkalkan, F., & Semiz, K., “Sporcu Beslenmesi ile ilgili youtube video yorumlarının duygu analizi”, *Uluslararası Beden Eğitimi Spor ve Teknolojileri Dergisi*, 2(2), 27-39, 2021.

**[15]** Kumaş, E., “Türkçe twitter verilerinden duygu analizi yapılırken sınıflandırıcıların karşılaştırılması”, *ESTUDAM Bilişim Dergisi*, 2(2), 1-5, 2021.

**[16]** Temizhan, E., & Mendeş, M., “Covıd-19 pandemisi ile ilgili twitter mesajlarının metin madenciliği tekniği ile değerlendirilmesi”, *Türkiye Klinikleri Biyoistatistik Dergisi*, 13(2), 185-200., URL: <https://doi.org/10.5336/biostatic.2020-79992>, 2021.

**[17]** Uyaroğlu Akdeniz, F., & Cebeci, H., “Belediye hizmetlerin değerlendirilmesinde duygu analizi yaklaşımı: Sakarya ili örneği”, *Zeki Sistemler Teori ve Uygulamaları Dergisi*, 4(2), 127-135, URL: <https://doi.org/10.38016/jista.932762>, 2021.

**[18]** Karaahmetoğlu, E., Ersöz, S., & Karaahmetoğlu, O., “Sosyal ağ tabanlı verilerden faydalanarak korona virüs konulu duygu analizi çalışması”, *Ergonomi Dergisi*, 4(1), 47-54, URL: <https://doi.org/10.33439/ergonomi.824333>, 2021.

**[19]** Köksal, B., Erdem, G., Türkeli, C., & Kamışlı Öztürk, Z., “Twitter'da duygu analizi yöntemi kullanılarak bitcoin değer tahminlemesi”, *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, 9, 280-297, URL: <https://doi.org/10.29130/dubited.792909>, 2021,

**[20]** Dikkaya Kavak, F., “Sosyal medyada gıda sektörünün netnografi ve metin madenciliği yöntemi ile incelenmesi: torku markasının twitter analizi”, Yayımlanmamış yüksek lisans tezi, Ankara Üniversitesi, 2022.

**[21]** Dinçer, E., Kayaoğlu, D., & Safarli, S., “Metin Madenciliği ve duygu analizi ile siber zorbalık tespiti”, *ESTUDAM Bilişim Dergisi*, 3(2), 38-45, URL: <https://doi.org/10.53608/estudambilisim.1070884>, 2022.

**[22]** Özoran, B., “Bir halkla ilişkiler aracı olarak twitter: dünya sağlık örgütü paylaşımlarının içerik analizi ve metin madenciliği ile incelenmesi”, *Manisa Celal Bayar Üniversitesi Sosyal Bilimler Dergisi*, 20(4), 125-146, 2022.

**[23]** Günyaktı, R., & Bursa, N., “Covid-19 pandemisinde sağlık çalışanları ve öğretmenlere yönelik algının duygu analizi ile twitter verileri üzerinden incelenmesi”, *Selçuk İletişim Dergisi*, 15(1), 264-285, URL: <https://doi.org/10.18094/josc.980334>, 2022.

**[24]** Torun, N., & Şengül, A., “Kripto para birimlerinin twitter verileri ile metin madenciliği kapsamında incelenmesi”, *Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi*, 6(1), 54-65, URL: <https://doi.org/10.33461/uybisbbd.1089670>, 2022.

**[25]** Turan, U., Emre, İ., & Kıran, S., “Metaverse ile ilgili Türkçe dilindeki çeşitli sosyal medya platformu verileri ile duygu analizi”, *Bilişim Sistemleri ve Yönetim Araştırmaları Dergisi*, 4(2), 1-16, 2022.

**[26]** Gülpınar Demirci, V., & Buluz Kömeçoğlu, B., “Sosyal Ağlarda veri gizliliği: Türkiye’de whatsapp gizlilik sözleşmesine gösterilen tepkilerinin duygu analizi”, *Kafkas Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 13(26), 710-742, URL: <https://doi.org/10.36543/kauiibfd.2022.030>, 2022.

**[27]** Sel, A., “Pandemi sürecinde toplum görüşünün duygu analizi yöntemiyle incelenmesi: Türkiye örneği”, *Beykoz Akademi Dergisi*, 10(2), 134-154, URL: <https://doi.org/10.14514/beykozad.882353>, 2022.

**[28]** Şahinaslan, Ö., Dalyan, H., & Şahinaslan, E., “Naive Bayes sınıflandırıcısı kullanılarak Youtube verileri üzerinden çok dilli duygu analizi”, *Bilişim Teknolojileri Dergisi*, 15(2), 221-229, URL: <https://doi.org/10.17671/gazibtd.999960>, 2022.

**[29]** Kandıran, E., Gümüş, B., & Özer, M., “Covid-19 pandemi sürecinde uzaktan eğitimin twitter yansımalarının duygu analizi”, *International Journal of Social Sciences and Education Research*, 8(3), 228-242, URL: <https://doi.org/10.24289/ijsser.1102248>, 2022.

**[30]** Akdeniz, A., “Makine öğrenmesi yöntemleri kullanılarak uzaktan eğitim konulu Türkçe tweetlerin duygu analizi”, Yayımlanmamış yüksek lisans tezi, Konya Teknik Üniversitesi, 2022.

**[31]** Zhao J, Cao X, “Combining semantic and prior polarity for boosting twitter sentiment analysis”, *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity),* 832–837, 2015.

**[32]** Çoban Ö, Özyer B, Özyer GT, “Sentiment analysis for Turkish Twitter feeds”, *2015 23nd Signal Processing and Communications Applications Conference (SIU),* 2388–2391, 2015.

**[33]** Siddiqua UA, Ahsan T, Chy AN, “Combining a rule-based classifier with ensemble of feature sets and machine learning techniques for sentiment analysis on microblog”, *19th International Conference on Computer and Information Technology (ICCIT),* 304–309, 2016.

**[34]** Chamansingh N, Hosein P., “Efficient sentiment classification of Twitter feeds”, *3rd International Conference on Knowledge Engineering and Applications (ICKEA)*, 78–82, 2016.

**[35]** Jianqiang Z, Xiaolin G., “Comparison research on text pre-processing methods on twitter sentiment analysis”, *IEEE Access*, 5: 2870–2879, 2017.

**[36]** Riyadh AZ, Alvi N, Talukder KH, “Exploring human emotion via Twitter”, *20th International Conference of Computer and Information Technology*, 1–5, 2017.

**[37]** Tsapatsoulis N, Djouvas C., “Feature extraction for tweet classification: Do the humans perform better?”, *12th International Workshop on Semantic and Social Media Adaptation and Personalization*, 53–58, 2017.

**[38]** Kaur S, Sikka G, Awasthi LK., “Sentiment Analysis Approach Based on N-gram and KNN Classifier”, *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC),* 13–16, 2018.

**[39]** Permatasari PI, Fauzi MA, Adikara PP, Sari EDL, “Twitter Sentiment Analysis of Movie Reviews using Ensemble Features Based Naïve Bayes”, *3rd International Conference on Sustainable Information Engineering and Technology*, 92–95, 2018.

**[40]** Aydın İ, Salur MU, Başkaya F., “Duygu Analizi için Çoklu Populasyon Tabanlı Parçacık Sürü Optimizasyonu”, *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 52–64, 2018.

**[41]** Salur MU, Aydin I, Alghrsi SA., “SmartSenti: A Twitter-Based Sentiment Analysis System for the Smart Tourism in Turkey”, *2019 International Conference on Artificial Intelligence and Data Processing Symposium (IDAP)*, 1–5, 2019.

**[42]** Ahuja R, Chug A, Kohli S, Gupta S, Ahuja P., “The impact of features extraction on the sentiment analysis”, *Procedia Comput. Sci*, 152: 341–348, 2019.

**[43]** Saad SE, Yang J., “Twitter Sentiment Analysis Based on Ordinal Regression”, *IEEE Access*, 7: 163677–163685, 2019.

**[44]** Topçu, A. E. ve Erkaya, A. E., “Recent Research and Assessments for Computer Engineering”, *Livre de Lyon*, 1-14 s, 2020.

**[45]** Mengutayci, Ü. ve Temurtas, H., “Yapay Sinir Ağları ile Türkçe Otel Yorumlarının Sınıflandırılması”, *International Black Sea Coastline Countries Scientific Research Symposium – VI*, 683-687, 2021.

**[46]** Mayda, I. ve Korkmaz, M., “Sentiment Analysis with Turkish Adjective Dictionary”, *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, (IEEE, 1-6), 2018.

**[47]** Gezici, G. ve Yanıkoğlu, B., “Sentiment Analysis in Turkish”, *Theory and Applications of Natural Language Processing*, 255–271, 2018.

**[48]** Onan, A., “Ensemble of Classifiers and Term Weighting Schemes for Sentiment Analysis in Turkish”, *Scientific Research Communications*, 1(1), 1-12, 2021.

**[49]** Demircan, M., Seller, A., Abut, F. ve Akay, M. F., “Developing Turkish sentiment analysis models using machine learning and e-commerce data”, *International Journal of Cognitive Computing in Engineering*. 2, 202-207, 2021.

**[50]** Toçoğlu, M. A., Öztürkmenoğlu, O. ve Alpkoçak, A., “Emotion Analysis from Turkish Tweets using Deep Neural Networks”, *IEEE Access*. 7, 183061-183069, 2019.

**[51]**  Çataltaş, M., Doğramacı, S., Yumuşak, S. ve Öztoprak, K., “Extraction of Product Defects and Opinions from Customer Reviews by Using Text Clustering and Sentiment Analysis”, *2020 IEEE International Conference on Big Data (Big Data)*, 4529-4534, 2020.

**[52]** F. Cheong and C. Cheong, “Social media data mining: A social network analysis of tweets during the Australian 2010-2011 floods”, *15th Pacific Asia Conference on Information Systems (PACIS)*, 1–16, 2011.

**[53]** B. Mandel, A. Culotta, J. Boulahanis, D. Stark, B. Lewis, and J. Rodrigue, “A demographic analysis of online sentiment during hurricane irene,”, *Proceedings of the second workshop on language in social media*, 27–36, 2012.

**[54]** C. Caragea, A. Squicciarini, S. Stehle, K. Neppalli, and A. Tapia, “Mapping moods: Geo-mapped sentiment analysis during hurricane sandy”, 642–651, 2014.

**[55]**  H. Li , “Twitter mining for disaster response: A domain adaptation approach”, 2015.

**[56]** M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, “Extracting Information Nuggets from Disaster- Related Messages in Social Media”, 2013.

**[57]** Z. Ashktorab, C. Brown, M. Nandi, and A. Culotta, “Tweedr: Mining twitter to inform disaster response.,”, *ISCRAM*, 2014.

**[58]** K. Kireyev, L. Palen, and K. Anderson, “Applications of topics models to analysis of disaster-related twitter data,”, *NIPS workshop on applications for topic models: text and beyond*, vol. 1, 2009.

**[59]** S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, “Microblogging during two natural hazards events: what twitter may contribute to situational awareness,”, *Proceedings of the SIGCHI conference on human factors in computing systems*, 1079–1088, 2010.

**[60]** S. Kumar, G. Barbier, M. A. Abbasi, and H. Liu, “Tweettracker: An analysis tool for humanitarian and disaster relief,”, *Fifth international AAAI conference on weblogs and social media*, 2011.

**[61]** Go, A., Bhayani, R. ve Huang, L., “Twitter sentiment classification using distant supervision”, *CS224N project report*, 1(12), 2009.

**[62]** Barkur, G. ve Vibha, G. B. K., “Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India”, *Asian journal of psychiatry*, 51, 102089, 2020.

**[63]** Li S, Wang Y, Xue J, Zhao N ve Zhu, T., “The impact of COVID-19 epidemic declaration on psychological” 2020.

**[64]** Süral, I., Griffiths, M. D., Kircaburun, K. ve Emirtekin, E., “Trait emotional intelligence and problematic social media use among adults: The mediating role of social media use motives”, *International Journal of Mental Health and Addiction*, 17(2), 3, 2019.

**[65]** Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y. ve Zhu, T., “Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach”, *Journal of medical Internet research*, 22(11), e20550, 2020.

**[66]** Feng, Y. ve Zhou, W., “Is working from home the new norm? an observational study based on a large geo-tagged covid-19 twitter dataset”, *arXiv preprint arXiv:2006*, 08581, 2020.

**[67]** Hornung, O., Dittes, S. ve Smolnik, S., “When emotions go social–understanding the role of emotional intelligence in social network use”, 2018.

**[68]** Depoux, A., Martin, S. ve Karafillakis, E., Raman Preet, Annelies Wilder-Smith, and Heidi Larson., “The pandemic of social media panic travels faster than the covid-19 outbreak”, 2020.

**[69]** Jelodar, H., Wang, Y., Orji, R. ve Huang, S., “Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach”, *IEEE Journal of Biomedical and Health Informatics,* 2020.

**[70]** Lu, X. ve Zhang, H., “Sentiment analysis method of network text based on improved at-bigru model”, *Scientific Programming*, 2021.

**[71]**  M. C. Yılmaz ve Z. Orman, "LSTM Derin Öğrenme Yaklaşımı ile Covid-19 Pandemi Sürecinde Twitter Verilerinden Duygu Analizi", *Acta Infologica*, URL: https://doi.org/10.26650/acin.947747.

**[72]** İ. Sel ve D. Hanbay, "Ön Eğitimli Dil Modelleri Kullanarak Türkçe Tweetlerden Cinsiyet Tespiti", *Fırat Üniversitesi Müh. Bil. Dergisi*, 33, 2, 675-684, 20211.

**[73]**  J. Samuel, G. G. M. Nawaz Ali, M. Rahman, E. Esawi ve Y. Samuel, "COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification", *Information*, cilt 11(6), 314, 2020.

**[74]** Aygün, I., Kaya, B., & Kaya, M., “Aspect based twitter sentiment analysis on vaccination and vaccine types in covid-19 pandemic with deep learning”, *IEEE Journal of Biomedical and Health Informatics*, 26(5), 2360-2369, 2021.

**[75]** C. J. Lyu, L. E. Han, K. G. Luli, “COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis”, *J Med Internet Res 2021*, 23(6), e24435, 2021.

**[76]** S. Liu and J. Liu, “Public attitudes toward COVID19 vaccines on English-language Twitter: A sentiment analysis”, *Vaccine 39*, 5499–5505, 2021.

**[77]**  Marcec, R., & Likic, R., “Using twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines”, *Postgraduate medical journal*, 98(1161), 544- 550, 2022.

**[78]** C. Villavicencio, J. J. X. Macrohon, X. A. Inbaraj, J. H. Jeng, J. G. Hsieh, “Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Naïve Bayes”, *Information*, 12, 204, URL: <https://doi.org/10.3390/info12050204>, 2021.

**[79]** Ansari, M. T. J., & Khan, N. A., “Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content”, *Electronic Journal of General Medicine*, 18(6), 2021.

**[80]** Çılgın, C., Gökçen, H., & Gökşen, Y., “Twitter’da COVID-19 aşılarına karşı kamu duyarlılığının çoğunluk oylama sınıflandırıcısı temelli makine öğrenmesi ile duygu analizi”, *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, 38(2), 1093- 1104, 2022.

**[81]** Aslan, S., “BiGRU-CNN Tabanlı Derin Öğrenme Modeliyle Türkiye’deki Covid-19 Aşılarına Yönelik Twitter Duygu Analizi”, *International Journal of Pure and Applied Sciences*, 8(2), 312-330, 2022.

**[82]** Mermer, G., & Özsezer, G., “Discussions About COVID-19 Vaccination on Twitter in Turkey: Sentiment Analysis”, *Disaster Medicine and Public Health Preparedness*, 17, e266, 2023.

**[83]** Özyurt, Ö., & Kısa, N., “Covıd-19 salgını sürecinde uzaktan eğitime ilişkin tweetlerin duygusal analizi”, *Journal of Computer and Education Research*, 9(18), 853-868, URL: <https://doi.org/10.18009/jcer.950790>, 2021.

**[84]** Akın AA, Akın MD., “Zemberek, An Open Source Nlp Framework for Turkic Languages”, *Structure 2007*, 10: 1–5, 2021.

**[85]** Brownlee J., “Develop Sequence Prediction Models With Deep Learning”, *Machine Learning Mastery*, 2017.

**[86]** Berman D., “A survey of deep learning methods for cyber security”, *Information*, 10(4), 122, 2019.

**[87]** Kayaalp K, Süzen AA., “Derin öğrenme ve Türkiye’deki uygulamalar”,. *Iksad International Publishing House*, 6-21, 2018.

**[88]** Sucu İ., “Yapay Zekânın toplum üzerindeki etkisi ve yapay zekâ (AI) filmi bağlamında yapay zekâya bakış”, *Uluslararası Ders Kitapları ve Eğitim Materyalleri Dergisi*, 2(2), 203-215, 2019.

**[89]** Ömrüuzun B, Saldanlı A., “Yapay Sinir Ağları ile Kripto Paraların Fiyat Modellemesi”, MSc Thesis, Istanbul University, 2019.

**[90]** Öztemel E., “Yapay Sinir Ağları”, *Üçüncü baskı, Papatya Yayıncılık*, 2012.

**[91]** Şenkal S, Emeksiz C., “Gauss fonksiyonları ile pencerelenmiş polinom aktivasyon fonksiyonu kullanan dalgacık sinir ağı ile yapay sinir ağlarının, çok kısa vadeli rüzgâr hızı tahmininde, tahmin başarımlarının karşılaştırılması”, *3nd International Symposium on Innovative Approaches in Scientific Studies*, 2019.

**[92]** LeCun Y, Bengio Y, Hinton G., “Deep learning”, *Nature*, 2015; 521: 436–444, 2021.

**[93]** Yang S, Yu X, Zhou Y., “LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example”, *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI),* 98-101, 2021.

**[94]** Terpstra T, De Vries A, Stronkman R, Paradies GL., “Towards a realtime Twitter analysis during crises for operational crisis management”, *Simon Fraser University Burnaby*, 2012.

**[95]** Mahesh B., “Machine Learning Algorithms - A Review”, *International Journal of Science and Research (IJSR),* 9(1), 381-386, 2020.

**[96]** Mehmet Can Y, Zeynep O., “LSTM Derin Öğrenme Yaklaşımı ile Covid-19 Pandemi Sürecinde Twitter Verilerinden Duygu Analizi”, *Acta Infologica*, 359-372, 2021.

**7. ÖZGEÇMİŞ**

**ELİF YANIK**

**Phone Number:**  +90 ( 551 ) 554-4258

**E-mail :**  yanikelif13@gmail.com

**GitHub Link:** https://github.com/elifynk

**Address:**  Ataşehir / İstanbul

**EDUCATION**

**Computer Engineering**

Marmara University / 2020- Expected June 2024 (2,88 / 4)

Hayrullah Kefoğlu Anatolian High School / 2015-2019 (86,29 / 100)

**EXPERIENCES**

**IT Governance Intern /** Sompo Japan Insurance Inc. / *April 2024 - Present*

**Back-End Developer Intern /** Vakıf Katılım Bankası A.Ş. / *October 2023 - January 2024*

**Front-End Developer Intern /** Inooster Information Technologies / *July 2023 - August 2023*

**Artificial Intelligence Intern /** Sigorta Bilgi ve Gözetim Merkezi / *August 2022 - September 2022*

**Intern /** Denizbank */ June 2021 –July 2021*

**MUSTAFA EREN GÜLBAHAR**

**Phone Number:**  +90 ( 530 ) 047 43 34

**E-mail :**  erenngulbahar@gmail.com

**GitHub Link:** https://github.com/erengulbahar

**Address:**  Üsküdar / İstanbul

**EDUCATION**

**Computer Engineering**

Marmara University / 2020- Expected June 2024 ( 3,15 / 4)

Ahmet Keleşoğlu Anatolian High School / 2015-2019 ( 90,82 / 100)

**EXPERIENCES**

**Software Engineer /** IstAnalytics **/** *July 2023 - Present*

**Data Engineer Intern /** *VakıfBank* **/** *October 2023 - January 2024*

**Data Engineer Intern /** *Koçtaş* **/** *July 2023 - September 2023*

**Software Engineer /** *DappLabs* **/** *May 2023 - August 2023*

**Software Engineer /** *Türkiye Finans Katılım Bankası* **/** *February 2023 - July 2023*

**Software Engineer Intern /** *IstAnalytics* **/** *June 2022 - August 2022*