

T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

**METİN MADENCİLİĞİ TEKNİKLERİ
KULLANILARAK TÜRKÇE MÜŞTERİ
YORUMLARININ SINIFLANDIRILMASI**

Yüksek Lisans Tezi

Tezi Hazırlayan
VELİ CENGİZ

İstanbul, 2023

T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

**METİN MADENCİLİĞİ TEKNİKLERİ
KULLANILARAK TÜRKÇE MÜŞTERİ
YORUMLARININ SINIFLANDIRILMASI**

Yüksek Lisans Tezi

Tezi Hazırlayan

VELİ CENGİZ

2120003042

Orcid: 0000-0002-0833-3761

DANIŞMAN: Dr. Öğr. Üyesi Talat FİRLAR

İstanbul, 2023

YEMİN METNİ

Yüksek Lisans Tezi olarak sunduğum “Metin Madenciliği Teknikleri Kullanılarak Türkçe Müşteri Yorumlarının Sınıflandırılması” başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere uygun şekilde tarafımdan yazıldığını, yararlandığım eserlerin tamamının kaynaklarda gösterildiğini ve çalışmamın içinde kullanıldıkları her yerde bunlara atıf yapıldığını belirtir ve bunu onurumla doğrularım. 29.10.2023

Veli CENGİZ

T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ MÜDÜRLÜĞÜ
TEZLİ YÜKSEK LİSANS SINAV TUTANAĞI

28.11.2023

Enstitümüz **Bilgisayar Mühendisliği** Anabilim Dalı **Bilgisayar Mühendisliği** Programı yüksek lisans öğrencilerinden **2120003042** numaralı **Veli CENGİZ'in** “İstanbul Beykent Üniversitesi Lisansüstü Eğitim – Öğretim Yönetmeliği”nin ilgili maddesine göre hazırlayarak, Enstitümüze teslim ettiği “**Metin Madenciliği Teknikleri Kullanılarak Türkçe Müşteri Yorumlarının Sınıflandırılması**” konulu tezini, Yönetim Kurulumuzun 14.11.2023 tarih ve 2023/45 sayılı toplantısında seçilen On-Line toplanan biz jüri üyeleri huzurunda, İstanbul Beykent Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 29. maddesinin 3. fıkrası gereğince 45 dakika süre ile Zoom programı aracılığıyla on-line olarak aday tarafından savunulmuş ve sonuçta adayın tezi hakkında “**OYBİRLİĞİ**” ile “**KABUL**” kararı verilmiştir.

İşbu tutanak, 2 nüsha olarak hazırlanmış ve Enstitü Müdürlüğü’ne sunulmak üzere tarafımızdan düzenlenmiştir.

DANIŞMAN
Dr. Öğr. Üyesi Ta*** Fİ***
(İstanbul Beykent Üniversitesi)

ÜYE
Dr. Öğr. Üyesi Ke*** Gö*** NA***
(İstanbul Beykent Üniversitesi)

ÜYE
Dr. Öğr. Üyesi Ke*** ŞA***
(İstanbul Medipol Üniversitesi)

Adı Soyadı : Veli CENGİZ
Danışmanı : Dr. Öğr. Üyesi Talat FİRLAR
Türü ve Tarihi : Yüksek Lisans (Tezli), 2023
Alanı : Bilgisayar Mühendisliği
Anahtar Kelimeler : Metin Madenciliği, E-Ticaret, Müşteri Yorumları,
Yapay Sinir Ağları, Erişimsel Sinir Ağları

ÖZ

METİN MADENCİLİĞİ TEKNİKLERİ KULLANILARAK TÜRKÇE MÜŞTERİ YORUMLARININ SINIFLANDIRILMASI

Günümüzde teknolojinin hızlı gelişmesi ile insanların satın alma ve satma faaliyetlerini e-ticaret platformlarında yapmaya neden olmuştur. Bu durum, e-ticaret platformlarındaki ürünlere yazılan müşteri yorumlarının önemini artırmıştır. İnsanların ürün veya hizmet hakkında doğru bilgi edinmelerine yardımcı olmuştur. Satıcılar için de müşteri memnuniyetini ölçme ve iyileştirme imkânı sağlamıştır. Özellikle, şirketler yapılan olumsuz yorumları en hızlı şekilde tespit etmek ve bu sayede iş iyileştirme yapmak için çeşitli metin madenciliği tekniklerini kullanmaktadırlar. Bu tez çalışması, Türkçe dilindeki müşteri yorumlarının analiz edilmesi ve 32 farklı sınıfa ayrılması amacıyla metin madenciliği yöntemlerini kullanmayı hedeflemektedir. İki farklı makine öğrenimi modeli olan yapay sinir ağları (ANN) ve evrişimli sinir ağları (CNN) kullanılarak bu sınıflandırma gerçekleştirilmiştir. Her iki model de öncelikle eğitim veri setiyle eğitilmiş ve daha sonra test veri seti üzerinde değerlendirilmiştir. Elde edilen sonuçlara göre, ANN modeli için %92.34 doğruluk (Accuracy) ve %92.40 F1-skoru elde edilmiştir. CNN modeli ise %92.78 doğruluk ve %92.83 F1-skoru ile daha yüksek performans göstermiştir. Her iki model de oldukça yüksek başarı oranları elde etmiştir. CNN modelinin daha yüksek performans sergilemesi, özellikle metin verileri üzerinde evrişimli yapıların daha etkili olduğunu göstermiştir.

Name Surname : Veli CENGİZ
Supervisor : Asst. Prof. Dr. Talat FİRLAR
Degree and Date : Master's (Thesis) /2023
Area : Computer Engineering
Key Words : Text Mining, E-Commerce, Customer Reviews,
Artificial Neural Networks, Access Neural Networks

ABSTRACT

CLASSIFICATION OF CUSTOMER COMMENTS IN TURKISH USING TEXT-MINING TECHNIQUES

Today, the rapid development of technology has led people to carry out their buying and selling activities on e-commerce platforms. This has increased the importance of customer reviews written on products on e-commerce platforms. It has helped people to get accurate information about the product or service. It has also allowed sellers to measure and improve customer satisfaction. In particular, companies use various text mining techniques to identify negative reviews as quickly as possible and thus make business improvements. This thesis aims to use text mining methods to analyze customer reviews in the Turkish language and classify them into 32 different classes. Two other machine learning models, artificial neural networks (ANNs) and convolutional neural networks (CNNs), are used for this classification. Both models were first trained on the training dataset and then evaluated on the test dataset. According to the results, 92.34% Accuracy and 92.40% F1-score were obtained for the ANN model. The CNN model performed better with 92.78% accuracy and 92.83% F1-score. Both models achieved very high success rates. The higher performance of the CNN model shows that convolutional structures are more effective, especially on text data.

İÇİNDEKİLER

Sayfa No.

ÖZ

ABSTRACT

TABLolar LİSTESİ	v
ŞEKİLLER LİSTESİ	vi
KISALTMALAR	viii
SÖZLÜK.....	ix
GİRİŞ	1

BİRİNCİ BÖLÜM

1. LİTERATÜR ARAŞTIRMASI	4
2. VERİ	8
2.1. Büyük Veri	8
2.2. Veri Madenciliği	10
2.3. Veri Madenciliği Uygulama Süreci	10
2.4. Veri Madenciliği Teknikleri.....	11
2.5. Metin Madenciliği	13
2.5.1. Metin Madenciliğinin Kullanım Alanları	14
2.5.2. Metin Madenciliği Süreci.....	15
2.5.3. Metin Madenciliği Teknikleri	16
3. DOĞAL DİL İŞLEME	17
3.1. Doğal Dil İşlemenin Kullanım Alanları	17
3.2 Doğal Dil İşleme Süreci	18
3.3. Doğal Dil İşlemede Veri Hazırlama Süreci	18
4. METİN SINIFLANDIRMA.....	19
4.1. Metin Sınıflandırma Algoritmaları	20
4.1.1..Karar ağaçları	20
4.1.2. Destek vektör makinesi	21
4.1.3. K-en yakın komşu (K-nearest neighbors)	22
4.1.4. Naive Bayes	22
4.1.5. Yapay Sinir Ağları	22

4.2. Metin Sınıflandırmada Performans Ölçütleri.....	23
4.2.1. Doğruluk (Accuracy)	24
4.2.2. Kesinlik Ölçütü	25
4.2.3. Duyarlılık Ölçütü	26
4.2.4. Özgüllük.....	26
4.2.5. Negatif Prediktif Değer	26
4.2.6. F-Ölçeği	27
4.2.7. ROC Eğrisi.....	27
5. METİN MADENCİLİĞİ TEKNİKLERİ KULLANILARAK TÜRKÇE	
MÜŞTERİ YORUMLARININ SINIFLANDIRILMASI.....	28
5.1. Veri setinin elde edilmesi.....	29
5.2. Veri setinin incelenmesi.....	29
5.3. Veri setine istatistiksel yöntemlerin uygulanması.....	30
5.4. Verilerin ön hazırlık aşamalarının hazırlanması	31
5.5. Doğal Dil İşleme Süreci	35
5.6. Çoklu sınıf sınıflandırma yöntemi	36
5.7. Sınıflandırma yöntemlerine uygun yapay sinir ağları ve makine öğrenmesi modellerin tasarlanması	38
5.7.1. ANN Modelinin Tasarlanması	39
5.7.2. Tasarlanan ANN modellerin eğitim veri setine uygulanması	41
5.7.3. Eğitilen ANN modelinin test veri seti ile değerlendirilmesi.....	42
5.7.4. CNN Modelinin Tasarlanması	44
5.7.5. Tasarlanan CNN modelinin eğitim veri setine uygulanması	45
5.7.6. Eğitilen CNN modelinin test veri seti ile değerlendirilmesi	47
5.8. En iyi performans sonuçları için modellerin karşılaştırılması	49
5.9. En iyi performansı gösteren modelin karşılaştırılması	53
SONUÇ	55
KAYNAKÇA.....	56

TABLÖLAR LİSTESİ

	Sayfa No
Tablo 1: Metin Madenciliği Teknikleri.....	16
Tablo 2: Veri Ön İşleme Kütüphane ve Fonksiyonları.....	35
Tablo 3: Model Sonuçlarının Karşılaştırılması-1	50
Tablo 4: Model Sonuçlarının Karşılaştırılması-2	51
Tablo 5 : Model Sonuçlarının Karşılaştırılması-3.....	52



ŞEKİLLER LİSTESİ

	Sayfa No
Şekil 1: Veri Madenciliği İşlem Süreci	11
Şekil 2: Veri Sınıflandırması	12
Şekil 3: Veri Kümelenmesi	12
Şekil 4: Metin Madenciliği İşlem Süreci.....	16
Şekil 5: Doğal Dil İşleme Bileşenleri.....	17
Şekil 6: Doğal Dil İşleme Süreci.....	18
Şekil 7: Veri Hazırlığı Adımları.....	19
Şekil 8: Metin Sınıflandırma Algoritmaları	20
Şekil 9: Karar Ağaçları Şeması	21
Şekil 10: Karmaşıklık Matrisi	24
Şekil 11: Roc Eğrisi.....	28
Şekil 12: Veri seti içeriği.....	29
Şekil 13: Veri seti data istatistiği.....	30
Şekil 14: Kategori bağlı yorum sayısı	31
Şekil 15: Benzersiz kayıt istatistiği	32
Şekil 16: Kategori Yorum Sayısı Grafiği	32
Şekil 17: Kategori Yorum yüzdelik Dağılımı	33
Şekil 18: Yorum Kelime Sayısı.....	33
Şekil 19: Kelime İstatistiği	34
Şekil 20: Yorumlarda en çok geçen 20 kelime.....	34
Şekil 21: En Çok Geçen Kelimler	35
Şekil 22: Veri ön işleme öncesi ve sonrası.....	36
Şekil 23: Önerilen Yapay Zekâ Çoklu Sınıf Sınıflandırma Mimarisi	37
Şekil 24: Önerilen Yapay Sinir Ağı Mimarisi.....	39
Şekil 25: Model Eğitim Geçmişi	42
Şekil 26: Model Performans Metrikleri.....	42
Şekil 27: Karmaşıklık Matrisi Isı Haritası.....	43
Şekil 28: ANN Modeli Performans Grafikleri	43
Şekil 29: Önerilen CNN Mimarisi.....	44

Şekil 30: Model Eğitim Geçmişi	47
Şekil 31: Model Performans Metrikleri.....	47
Şekil 32: Karmaşıklık Matrisi Isı Haritası.....	48
Şekil 33: CNN Modeli Performans Grafikleri	49
Şekil 34: Belirli Bir Metnin Kategori Tahmin Sonucu	53
Şekil 35: Modellerin Karşılaştırılması	54



KISALTMALAR

ANN : Yapay Sinir Ağları (Artificial Neural Network)

CNN : Evrişimli Sinir Ağları (Convolutional Neural Network)

ÇSS : Çoklu sınıf sınıflandırma

DM : Veri Madenciliği (Data Mining)

DT : Karar Ağaçları (Decision Trees)

GBM : Gradyan Artırma Makinesi (Gradient Boosting Machine)

K-NN : K-En Yakın Komşu (K-Nearest Neighbors)

LDA : Doğrusal Diskriminant Analizi (Linear Discriminant Analysis)

LGBM: Hafif Gradyan Artırma Makinesi (Light Gradient Boosting Machine)

LR : Lojistik Regresyon (Logistic Regression)

LSTM : Uzun-Kısa Süreli Bellek (Long Short Term Memory)

NB : Naive Bayes

NLP : Doğal Dil İşleme (Natural Language Processing)

NLTK : Natural Language Toolkit

RF : Rastgele Orman (Random Forest)

RF : Rastgele Orman (Random Forest)

SVM : Destek Vektör Makineleri (Support Vector Machines)

SVM : Destek Vektör Makineleri (Support Vector Machines) Ekstrem

TF-IDF: Terim Frekansı-Ters Belge Frekansı

TSA : Ağaç Tohum Algoritması (Tree-Seed Algorithm)

XGB : Gradyan Artırma (Extreme Gradient Boosting)

XGB : Ekstrem Gradyan Artırma (Extreme Gradient Boosting)

SÖZLÜK

Veri: Bilgi parçalarının genel adıdır ve sayılar, metinler, semboller, görüntüler, sesler gibi çeşitli biçimlerde bulunur.

Büyük Veri (Big Data): Geleneksel veritabanlarının işlem kapasitesini aşan, büyük hacimli, çeşitli, hızlı ve doğruluğu önemli veri kümeleridir.

Veri Madenciliği (Data Mining): Veri kümelerinde gizli bilgileri ve desenleri keşfetmek amacıyla istatistiksel ve matematiksel tekniklerin kullanıldığı süreçtir.

Metin Madenciliği (Text Mining): Metin verilerini analiz ederek gizli desenleri, duygusal tonları ve önemli bilgileri çıkarmayı hedefleyen bir veri madenciliği alt dalıdır.

Doğal Dil İşleme (NLP - Natural Language Processing): Bilgisayarların doğal dil formlarını anlayabilmelerini sağlayan yapay zeka alanıdır.

Metin Sınıflandırma (Text Classification): Metin belgelerini belirli kategorilere sınıflandırmak amacıyla yapay zeka ve makine öğrenme kullanılan bir işlemdir.

Yapay Sinir Ağları (Artificial Neural Networks - ANN): İnsan beyninin biyolojik sinir ağlarından ilham alarak tasarlanmış ve bilgisayarlar aracılığıyla öğrenme ve veri işleme yetenekleri sağlayan yapay zeka modelleridir.

Erişimsel Sinir Ağları (Convolutional Neural Networks - CNN): Görsel verilerle çalışmak için özel olarak tasarlanmış yapay sinir ağlarıdır ve nesne tanıma, görüntü işleme gibi görevlerde kullanılır.

GİRİŞ

Günümüzde, teknolojinin hızla gelişmesi ve üretilen veri miktarının katlanarak artması verinin depolanma, işleme ve bilgiye dönüşme ihtiyacını ortaya çıkarmaktadır. Bu ihtiyaçları ortadan kaldırmak için geliştirilen teknolojilerden biri büyük veridir. Geçmişte ve şu anda olmak üzere farklı zaman dilimlerinde kullandığımız, paylaştığımız ve ürettiğimiz verilerin tamamı büyük veriyi oluşturmaktadır.

Bu çalışmada, Türkçe müşteri yorumları yeni özgün kelime sözlükleri, teknoloji alanına özgü yeni sözlük, Türkçe için üretilen yeni dil kurallarından oluşan yeni bir Çoklu sınıf sınıflandırma (ÇSS) yöntemi önerilmiştir. Bu algoritma, bir e- ticaret platformu üzerinden alınan kullanıcı yorumları üzerinde uygulanmıştır. Algoritmanın performansını değerlendirmek için algoritma tahminleri beş farklı makine öğrenmesi yönteminin tahminleriyle karşılaştırılmıştır. Önerilen yaklaşımın Türkçe metin madenciliği, doğal dil işleme ve metin sınıflandırma çalışmalarında kullanılması hedeflenmiştir.

E-ticaretin hızla yaygınlaşmasıyla birlikte, web üzerinde giderek daha fazla ürün satılmakta ve giderek daha fazla insan web üzerinden ürün satın almaktadır. Müşteri memnuniyetini ve alışveriş deneyimlerini artırmak için, müşterilerinin satın aldıkları ürünleri incelemelerini veya görüşlerini ifade etmelerini sağlamak, çevrimiçi satıcılar için yaygın bir uygulama haline gelmiştir. Gittikçe daha fazla sayıda kullanıcının internette rahat etmesiyle birlikte, daha fazla insan yorum yazıyor. Sonuç olarak, bir ürünün aldığı yorumların sayısı hızla artıyor. Bazı popüler ürünler, bazı büyük satıcı sitelerinde yüzlerce yorum alabilir. Bu, potansiyel bir müşterinin ürünü satın alıp almama konusunda karar vermesine yardımcı olmak için onları okumasını çok zorlaştırmaktadır. Bu durum, ürün veya hizmet hakkında yazılmış yorumların analiz edilme ihtiyacını ortaya çıkarmıştır. Çünkü müşterilerin bir satın alma işlemi anında o ürün hakkında yaşanan deneyimleri önceden incelemesi, ürün hakkında fikir sahibi olmalarını sağlamaktadır. Böylece, satın alma konusunda yaşanan karar problemlerinin giderilmesi mümkün olmaktadır. Benzer şekilde, daha iyi hizmet sunmak isteyen kurumlar müşteri yorumlarını analiz ederek şikâyet ve memnuniyet başlıklarını tespit edebilir. Bu analiz sayesinde, kurumlar iş iyileştirme ve geliştirme yapmak için bir imkâna sahip olmaktadır. Bu kapsamda, müşteriler ve satıcılar

tarafından ilgili ürün hakkında yazılmış tüm yorumların okunması her zaman mümkün olmayabilir. Örneğin; bir ürün hakkında yazılmış binlerce yorumun tamamının okunması hem iş yüküne hem de zaman maliyetine sebep olmaktadır. Bu durumları ortadan kaldırmak için müşteri yorumlarını analiz eden sistematik yaklaşımlara gerek duyulduğu görülmektedir

Çalışmanın birinci bölümünde literatür araştırması, ikinci bölümünde büyük veri başlığı altında veri madenciliği, metin madenciliği, doğal dil işleme, makine öğrenmesi algoritmaları, performans metrikleri ve vektörizasyon yöntemleri anlatılmaktadır. Üçüncü bölümde; veri toplama ve etiketleme işlemleri, veri ön işleme adımları, geliştirilen yeni sözlükler, dil kuralları, puanlama yöntemi ve metin sınıflandırmayı içeren ÇSS yöntemi önerilmiştir. Dördüncü bölümde, bir e-ticaret platformu üzerinden elde edilen müşteri yorumlarının üzerinde önerilen ÇSS yönteminin uygulanması anlatılırken beşinci bölümde önerilen ÇSS yönteminin makine öğrenmesi yöntemleriyle karşılaştırılması yer almaktadır. Son bölümde algoritmanın performansı ve Türkçe metin madenciliği ekosistemine sağladığı katkı tartışılmaktadır.

Özgün kelime sözlükleri, teknoloji alanına özgü yeni bir sözlük, yeni dil kuralları ve puanlama sistemine sahip olan ÇSS yönteminin Türkçe metinleri olumlu, olumsuz ve tarafsız sınıflandırmak için önerilmiştir. Çalışma, Türkçe metin madenciliği alanında yeni bir yaklaşım sunması, kaynak niteliği taşıması ve farklı alanlarda kullanılabilir olma potansiyeline sahiptir. Çalışmanın literatüre katkıları aşağıda gösterilmektedir:

- Özgün sıfat ve fiil sözlükleri önerilmiştir.
- Teknoloji alanında kullanılmak üzere ve kök ayırmaya yardımcı olan teknoloji alanına özgü yeni sözlük önerilmiştir.
- Cümle ya da kendisinden önce gelen kelimenin anlamını etkileyen ifadeleri tespit etmek için yeni Türkçe dil kuralları önerilmiştir.
- Üretilen sözlüklere ve kurallara dayalı olarak Türkçe metinleri puanlayan bir puanlama sistemi önerilmiştir.
- Geliştirilen sözlükler, dil kuralları ve puanlama sistemini içeren Türkçe metinleri olumlu, olumsuz ve tarafsız olarak sınıflandıran yeni bir yaklaşım önerilmiştir.

BİRİNCİ BÖLÜM

METİN MADENCİLİĞİ TEKNİKLERİ

Büyük veride kullanılan teknolojiler, hızlı algılama keşif ve analiz yaparak büyük verilerden ekonomik olarak değer üretmek için yapılan yeni nesil yap ve mimarilerdir (Villars, Olofson, & Eastwood, 2011). Tüm teknik karmaşıklığına rağmen, büyük veri basit bir fikirden oluşur: geçmiş hakkında yeterli ayrıntıyı toplayıp, doğru analitik araçları uygulayıp bu sayede gelecek hakkında alışılmadık derecede doğru tahminler yapmanıza yardımcı olabilecek beklenmedik bağlantılar ve korelasyonlar bulabilirsiniz. Bu bağlantılara ve ilişkilere erişebilmek için veri madenciliği tekniklerinden faydalanmak gerekmektedir.

Metin madenciliği, yapılandırılmamış metinden ilginç, önemsiz olmayan bilgilerin keşfi ve çıkarılmasıdır. Bu, bilgi alımından (yani, belge veya web üzerinden alınan veri) metin sınıflandırmasına ve kümelemeye, (son zamanlarda) varlık, ilişki ve olay çıkarmaya kadar her şeyi kapsar (Kao & Poteet, 2007) . Metin madenciliği, metinleri analiz ederek kelimelerin ve cümlelerin anlamını tespit etmeye yararken yazının olduğu her alanda kullanılma potansiyeline sahiptir. Bu durum, kullanım alanlarının çok geniş olduğunu ve çok sayıda çalışma konusunu meydana getirdiğini göstermektedir. Bu doğrultuda, metin madenciliği çalışmalarında, metin analizi gerçekleştirmek için doğal dil işleme gereksinim duyulmaktadır.

Doğal dil işleme (NLP), bilgisayarların doğal dil kullanarak iletişim kurmasını ve anlamasını sağlamak için kullanılan bir alandır. Makine, insan dilinin sözdizimini ve anlamını öğrenir, işler ve çıktısını kullanıcıya verir. Doğal dil işleme alanı, bilgisayar sistemlerinin doğal ve insan tarafından anlaşılabilir bir dille anlamlı görevleri yerine getirmesini içerir (Raina & Krishnamurthy, 2022). Doğal dil işleme ile metinde yer alan kelimelerin anlamını tespit etmek mümkündür. Bu durum, müşteri yorumlarını analiz etmek ve sundukları hizmeti geliştirmek isteyen kurumlar için bir fırsat oluşturmaktadır. Özellikle, son dönemde dünya çapında dijital ürün kullanımlarının artması, alışveriş ve satın alma eğilimlerinin dijital platformlara taşınmış olmasından dolayı e-ticaret sektörü popülarlığını fazlasıyla artırmış durumdadır.

Bu bölümde detaylı olarak literatür araştırması ele alınacaktır.

1. LİTERATÜR ARAŞTIRMASI

Literatürde, metin sınıflandırma konusunda sözlük tabanlı yaklaşımlar ve makine öğrenmesi algoritmaları yoğun olarak kullanılmıştır. Bu kapsamda, literatür araştırması metin madenciliğinde sözlük tabanlı yaklaşımlar, makine öğrenmesi tabanlı yaklaşımlar ve müşteri yorumları üzerine yapılan çalışmalar olarak üç başlıkta incelenmiştir.

Sözlük Tabanlı Yaklaşımlar: Sözlük tabanlı metin sınıflandırma çalışmaları için Türkçe metinlerde kullanılmak üzere önerilen kelime sözlüğü ile metin içindeki kelimelerin türleri tespit edilmiştir (Demir, Chawai, & Doğan, 2019, s. 58-66). Sıfat, isim, fiil, zarf, ön ve son eklerden oluşan bu sözlüklerde toplamda 30.000 adet kelime bulunurken sözlükler İngilizce için oluşturulmuş 3 farklı sözlüğün Türkçe'ye çevrilmesiyle oluşturulmuştur. Sözlüklerin etkinliği farklı veri setleriyle ölçülmüştür. Bununla birlikte, duygu çıkarımı yapmak için olumlu ve olumsuz metinleri sınıflandırmada kullanılan sözlükler %74.11'lik doğruluk değerine sahip olmuştur. Bu çalışma, literatüre yeni sözlükler üretmesiyle katkı sağlarken performans metriği olarak sadece doğruluk oranının ölçülmesi yönünden de istatistiksel analizde eksik kalmıştır.

Sözlük tabanlı başka bir çalışmada (Pilavcılar, 2007), sözlüğe bağlı olarak frekans hesabı yapılmıştır. Buradan hareketle, K-NN ve NB yöntemleriyle makaleler politika, ekonomi, sağlık ve spor olmak üzere dört kategoriye ayrılmıştır. Kelimeleri analiz etmek için Joker (Wild Card) yöntemi kullanıldığından dolayı yumuşama ses olayına sahip sözcüklerin (ilaç-ilacı, grip-gribi vb.) analizi daha kolay hale gelmiştir. Çalışma, metin sınıflandırma işlemleri için otomatik bir sistem sunması yönünden katkı sağlarken sözlükte yer alan 412 kelimenin yeterli sayıda olmaması ve sadece doğruluk metriğinin kullanılması açısından eksik kalmıştır.

İngilizce-Türkçe çevirisiyle yeni sözlükler oluşturan başka bir çalışmada (Uçan, 2014) , Google'da dahil olmak üzere dört farklı çeviri programı hibrit olarak kullanılmıştır. Oluşturulan sözlüklerde kelimelere duygu derecelerine göre skor ataması yapılırken metinlerin polarite değerleri hesaplanmıştır. Kelimelerin ekli ve eksiz halleri incelenerek eklerin metindeki önemleri analiz edilmiştir. Algoritma, otel yorumları için makine öğrenmesi yöntemleriyle benzer başarıma sahip olsa da film yorumlarında yeterli başarıma ulaşamamıştır. Çalışma, farklı yöntemlerin bir arada

kullanılması ve otomatik bir yaklaşım sunması yönüyle katkı sağlarken farklı alanlarda kullanılma potansiyelinin düşük olması yönüyle eksik kalmıştır.

İngilizce’ den Türkçe’ ye çevirme yoluyla sözlük tabanlı metin sınıflandırma işlemi yapan diğer çalışmada (Türkmenoğlu, 2015) , kelimeler isim/sıfat veya fiil olmak üzere sözlükte iki ayrı gruba ayrılmıştır. Kelimelerde yer alan durum ekleri incelenerek Twitter’ dan alınan tweet metinleri (kısa metinler) ve film yorumları (uzun metinler) duygu tabanlı sözlük yardımıyla olumlu, olumsuz ve tarafsız olarak sınıflandırılmıştır. Çalışma, kelimelerin farklı özelliklerini incelemesi, kuralsız ve hatalı yazılmış olma ihtimali yüksek metinleri analiz etmesi yönüyle katkı sağlarken makine öğrenmesi tabanlı sınıflandırıcılara kıyasla yaklaşık %10 oranlara daha düşük başarıma sahip olması yönüyle eksik kalmıştır.

Polarite odaklı sözlük tabanlı bir yaklaşımda (Yurtalan, Koyuncu, & Turhan, 2019), Twitter’ dan alınan veri kümesi kelime tabanlı analiz edilerek cümle analizi gerçekleştirilmiştir. Kelimeler -3 ile 3 arasında 1.0 ölçeğinde sınıflandırılarak sıfatlar negatif ve pozitif skor değeri, diğer türlerde sıfır skor değerini almıştır. Çalışma; kök ve eklerin bir arada incelenmesi, bağlaçların metinlerin anlamlarına yaptıkları etki, kendisinden önce gelen ifadenin anlamını değiştiren “yok”, “değil” gibi kelimelerin analiz edilmesi bakımından katkı sağlamıştır.

Makine Öğrenmesi Tabanlı Yaklaşımlar: Kelimelerdeki yapım-çekim eklerini, edat ve bağlaçları çıkararak terim ağırlıklandırma (TF-IDF) haber metinlerini sınıflandıran çalışmada (Başkaya & Aydın, 2017, s. 1-5), sınıflandırıcı olarak NB, RF ve SVM kullanılmıştır. Kelime kökleri Zemberek ile oluşturulmuştur. Çalışmada, RF algoritmasının en iyi sonucu verdiği görülürken sadece 80 haber metninin kullanılması veri sayısının yetersiz olduğunu göstermektedir. Bu durum çalışmanın eksik kaldığı yönü göstermiştir.

ANN ve SVM’ nin daha iyi sonuç verdiği görülmüştür. Türkçe tweet metinlerini sınıflandıran çalışmada (Onan, 2017) , sınıflandırıcı olarak NB, SVM ve LR yöntemleri kullanılmıştır. F1-Skoru değerine göre üç yöntemin de birbirine yakın (%79, %75 ve %78) sonuç verdiği görülmüştür.

Bir havayolu řirketi iin metinleri olumlu, olumsuz ve tarafsız olarak sınıflandıran alıřmada (Rane & Kumar, 2018), sınıflandırıcı olarak DT, RF, SVM, K-NN, LR, NB ve AdaBoost kullanılmıřtır. alıřma sonunda; RF, SVM, LR ve AdaBoost birbirine yakın ve en iyi sonuları etmiřtir. BBC haber metinlerini terim ağırlıklandırma (TF-IDF) yntemiyle vektrize ederek makine ğrenmesi yntemleriyle sınıflandıran alıřmada (Shah, Patel, Sanghvi, & Shah, 2020); LR, RF ve k-NN sınıflandırıcıları kullanılmıřtır. Bu iřlem sonucunda, LR ve RF en iyi sonuca ulařmıřtır. Amazon'daki kullanıcı mesajlarını olumlu ve olumsuz olarak sınıflandıran bir alıřmada (Al Amrani, Lazaar, & El Kadiri, 2018, s. 511-520), RF ve SVM algoritmaları hibrit olarak kullanılmıřtır. Bu yaklařım, sınıflandırıcıların kullanıcı yorumlarını negatif ve pozitif olarak sınıflandırmada tekli kullanımlarından daha iyi sonu vermiřtir. Bu durum, geleneksel yaklařımların birlikte kullanımlarıyla yntemlerin modern hale dnüşme olasılıklarının olduėunun gstergesidir. Facebook ve Twitter'da yazılan yorumları inceleyen bir alıřmada (Chawan, Krishna, & Krishna, 2017), SVM sınıflandırıcısıyla kesinlik, duyarlılık ve F1-Skoru metrikleri altında deėerlendirilmiřtir. alıřmada, metinler olumlu ve olumsuz olmak zere iki sınıfa ayrılmıřtır. Ayrıca, kurumların daha iyi hizmet vermesi adına müşteri yorumları zerinde metin madenciliėi tekniklerinin kullanılabilir olduėu grlmüştür. Covid-19 pandemi salgını srecinde, sosyal medyada yer alan metinlerden anksiyete veya kaygı bozukluėu yařayan bireyleri tespit etmeye alıřan alıřmada (Fauziah, Saifullah, & Aribowo, 2020), RF ve XGBoost sınıflandırıcılarının performansları karřılařtırılmıřtır. Metinleri olumlu ve olumsuz sınıflandırmada doėruluk metriėinde RF daha iyi sonu verirken kesinlik ve duyarlılıkta XGBoost daha iyi sonu vermiřtir. Bu durum, XGBoost'un metin sınıflandırma alıřmalarında kullanılabilir olduėunu gstermiřtir. Metin tabanlı duygu sınıflandırma iin LR kullanan alıřmada (Alotaibi, 2019, s. 31-37), LR ynteminin szlk tabanlı yaklařımlardan daha iyi sonu verdiėi grlmüştür.

Müşteri Yorumları zerine alıřmalar: Literatrde, kullanıcı yorumlarını metin madenciliėi teknikleriyle analiz eden alıřmalar da bulunmaktadır. Hepsiburada zerinde yer alan kullanıcı yorumlarını sıfat szlė ve derinlik ncelikli arama algoritma yardımıyla sınıflandıran alıřma (Ergn, 2012), kurumlara iř zekâsı alanında kullanılabilir bir yaklařım sunmuřtur. Bu yaklařım ile hem zaman tasarrufu saėlamak

hem de müşterilerin öneri ve şikâyet gibi geri bildirim mesajlarının otomatik olarak tespit edilmesi hedeflenmiştir. Şikayetvar web sitesi üzerinden bir telekomünikasyon dağıtıcısı için yazılmış yorumları analiz eden çalışmada (Kiraz, Eski, & Geçici, 2020), en sık kelimelerin kullanımına bağlı olarak frekans hesaplaması yapılmıştır. Bu doğrultuda, en çok şikâyet edilen alanlar belirlenmiştir. Çalışma, müşteri kazanmak ve daha iyi hizmet sunmak isteyen kurumlar için katkı sağlamıştır. Havayolu müşteri memnuniyetini, web üzerinde yer alan yorumlar yardımıyla inceleyen çalışmada (Lucini, Tonetto, Fogliatto, & Anzanello, 2020), sıfatlar derecelendirilerek metinlerin LR ve LDA algoritmalarıyla analizi gerçekleştirilmiştir. Çalışma, kullanıcılardan geri bildirim toplamak için sistematik yaklaşımların gerekliliğini ortaya çıkarmıştır. Amazon Music, Deezer, Spotify, Tidal ve YouTube olmak üzere beş farklı müzik yayıncısı için yazılmış kullanıcı yorumlarını analiz eden çalışmada (Chung, Lee, & Yoon, 2022), kullanıcı memnuniyetinin çevreye uyumu, fiyat politikası ve içerikle olmak üzere üç başlıkta toplandığını göstermiştir. Çalışma, kurumların rekabet analizi ve müşteri memnuniyeti konuları üzerinde iyileştirme yapması için uygun bir analiz ve görüş sunmuştur. Hepsiburada'daki kullanıcı yorumlarını analiz eden diğer çalışmada (Rumelli, Akkuş, Kart, & Isik, 2019), morfolojik analiz ve polarite ağırlıklandırma yöntemleriyle duygu çıkarımı yapılmıştır. Turizm sektöründe metin madenciliği tekniklerini kullanan diğer çalışmada (Ergüt, 2019, s. 103-114), yurt dışından gelen turistlerin tripadvisor.com üzerindeki yorumları analiz edilerek işletmeler için memnuniyet ve şikâyet başlıkları tespit edilmiştir. Yorum analizi yapan başka bir çalışmada (Işık, 2019); NB, LR, ve SVM kullanılarak karşılaştırma yapılırken LR yönteminin daha iyi sonuç verdiği görülmüştür.

Literatürde, Türkçe metin sınıflandırma çalışmalarının çok az sayıda olduğu görülmektedir. Mevcut çalışmalarda; performans metriği için sadece doğruluk değerinin kullanılması, hatalı yazılan metinler için yöntem geliştirilmiş olmaması, kelimelerin anlam bakımından incelenmemesi, örneklem sayısının yeterli olmaması, eş anlamlı sözcüklerin dikkate alınmaması, kelime türleri için ayrı değerlendirme yapılmaması, tarafsız yorumların sınıflandırılmaması, kullanım alanlarının kısıtlı olması, kelime ve metinlerin anlamlarını değiştiren sözcüklerin analiz edilmemesi, biçimsel açıdan bir analiz sunmamasından ve yeterli başarımlar değerlerine sahip olmadıklarından dolayı eksiklikler tespit edilmiştir. Benzer şekilde, Türkçe metin

sınıflandırma çalışmalarında kural ve sözlük tabanlı yaklaşımların da çok az sayıda olduğu görülmektedir. Bu durum, yeni sözlükler, dil kuralları, puanlama sistemleri ve yaklaşımların Türkçe metinler üzerinde uygulanması için bir ortam hazırlamıştır.

Buradan hareketle; Türkçe müşteri yorumlarının yapay zeka yöntemleriyle tahminlenmesi için ÇSS yöntemi önerilmiştir. Önerilen yöntem geleneksel yöntemlerden daha iyi sonuç vermesi yönüyle özgün değere sahiptir.

2. VERİ

Veri, belirli bir amaç için toplanan, kaydedilen, saklanabilen ve analiz edilebilen herhangi bir bilgi birimidir (Işıklı, 2014). Veri, çeşitli semboller, harfler, rakamlar ve diğer işaretler aracılığıyla ifade edilen ham ve işlenmemiş bilgilerdir. Veriler, anlamlı hale getirilebilmeleri için analiz edilmeli, işlenmeli ve yorumlanmalıdır. Verilerin doğru bir şekilde analiz edilmesi, işletmelerin karar alma süreçlerine yardımcı olabilir ve pazarlama stratejilerinin geliştirilmesine katkı sağlayabilir. Bu nedenle, veri analizi, işletmelerin rekabetçi avantajını korumak için önemli bir araçtır (Safran, 2015). Veri, düzensiz ve işlenmemiş bilgilerin bir araya gelmesiyle oluşan bir yapıdır. Bu bilgiler sayılar, semboller, harfler veya işaretler gibi farklı şekillerde ifade edilebilir. Ancak, bu verilerin anlamlı hale getirilebilmesi için analiz edilmesi, düzenlenmesi ve yorumlanması gerekir. Bu şekilde doğru veri analizi, işletmelerin karar alma süreçlerine yardımcı olabilir, bilim insanlarına ve diğer uzmanlara önemli bilgiler sağlayabilir ve pazarlama stratejilerinin geliştirilmesine katkıda bulunabilir. Temel amacı ise verilere bir düzen vermek ve analiz ederek anlamlı bilgiye ulaşmaktır. Genellikle amacı bilgiye dönüştürülmek olan işlenmemiş ham malzemeye “veri” denir.

2.1. Büyük Veri

Günümüzde hızla artan veri miktarı, bilgisayarların veri kaydetme ve saklama konusunda sunduğu kolaylıklarla birlikte sürekli bir artış trendi sergilemektedir. Bu durumda, veriyle ilgili kararların ertelenmesi söz konusu olabilirken, daha fazla bellek kullanımı sayesinde neredeyse her şey saklanabilmektedir. Bu verilerin etkili bir şekilde yönetilmesi ve analiz edilmesi, büyük veri olarak adlandırılan bu devasa bilgi miktarının değerlendirilmesi açısından büyük önem taşımaktadır.

Son yıllarda "büyük veri" kavramı, iş dünyası, bilgisayar bilimi, bilgi çalışmaları, bilgi sistemleri, istatistik gibi birçok alanda popüler hale gelmiştir. Bu, artan miktarda verinin bir sonucu olarak ortaya çıkmıştır ve verilerin toplanması, saklanması, işlenmesi ve analiz edilmesi konusunda yeni fırsatlar sunmaktadır (Maltby, 2011). Büyük veri, günümüzde hızla üretilen, çok çeşitli kaynaklardan gelen ve büyük boyutlarda olan verileri ifade eder. Bu veriler, geleneksel yöntemlerle işlenmeleri mümkün olmayacak kadar büyük boyutlara sahiptir ve özel teknolojiler gerektirirler. Büyük veri, işletmeler, bilgisayar bilimi, istatistik ve diğer birçok alanda büyük ilgi görmektedir.

Büyük veri, genellikle yüksek hacim, çeşitlilik ve hızda biriken dijital verilerdir. Bu veriler, 5V modeli olarak adlandırılan hacim (volume), çeşitlilik (variety), hız (velocity), değer (value), ve gerçeklik (veracity) gibi farklı özelliklere sahiptir (Lomotey & Deters, 2014). Büyük veri teknikleri ve teknolojileri, bu verilerin etkili bir şekilde işlenmesine ve analiz edilmesine yardımcı olur.

Büyük veri için kullanılan 5V modeli, verinin farklı özelliklerini tanımlamaktadır. Hacim özelliği, verinin büyük boyutlarda olduğunu ifade ederken, çeşitlilik özelliği farklı veri tiplerinin kullanılabilir olduğunu vurgular. Hız özelliği, verinin hızlı bir şekilde erişilebilir olduğunu ve boyutunun kısa zamanda arttığını belirtirken, değer özelliği verinin bilgiye dönüşüm sürecinde katma değer üretebildiğini ifade eder. Gerçeklik özelliği ise verinin geçerliliğine dair bir anlam taşır. Bu model, büyük veriyi daha anlaşılır bir şekilde ifade etmek için önemlidir.

Büyük veri, veri madenciliği teknikleri kullanılarak değerlendirilerek anlamlı bilgiye dönüştürülür. Bu bilgi çıkarımı, metinsel, kavramsal, tanısal, tahminsel, ses ve görüntüsel veriler üzerinde gerçekleştirilebilir. Metin madenciliği, özellikle metin verilerini analiz ederek sınıflandırma veya kümeleme gibi işlemler yapar. Bu nedenle, metin madenciliği çalışmalarında, doğal dil işleme araçları kullanılarak metinleri sadeleştirmek için çaba harcanır. Bu araçlar, kelime odaklı analiz yaparak metindeki anlamı doğru bir şekilde anlamayı, gereksiz sembollerin ve kelimelerin çıkarılmasını sağlar. Bu işlemler, tahmin modelleri oluşturmak için veri madenciliği ve metin madenciliği çalışmalarında gerçekleştirilir. Bu bağlamda, kendi kendine öğrenme

yeteneđi olan makine öğrenmesi yöntemleri, tahmin modelleri oluşturmak için veri setindeki verileri kullanarak sıklıkla tercih edilir.

2.2. Veri Madenciliđi

Veri, kaydedilmiş ve depolanmış bilgileri ifade etmektedir. Bu bilgiler, uydu görüntüleri, fiyat serileri, ankete katılanların cevapları gibi farklı kaynaklardan gelir. Veriler, iki tür bilgiyle ilişkilendirilmektedir. Birincisi, bireyler tarafından yapılan veri girişleri; örneđin, fiyatların kaydedilmesi veya anketteki soruların cevapları. İkincisi ise, meta verilerdir. Meta veriler, veriye anlam kazandırmaktadır ve verinin satır ve sütunlarının açıklamalarını göstermektedir. Meta veriler, kayıtlı veri tabanlarının geliştirilmesi ve sürdürülmesi için önemlidir. Veri madenciliđi için de, kayıtlı veri tabanlarının meta verileri göz önünde bulundurulmalıdır. Bu, verilerin analiz edilmesi ve anlamlı sonuçlar elde edilmesi için önemlidir (Mirkin, 2005).

Veri madenciliđi (Data mining), toplanan verilerin işlenerek anlamlı bilgiye dönüştürülmesini sağlar. Bu sayede, büyük veri depolarındaki yüksek boyutlu verilerin içerdikleri bilgi sistematik bir şekilde keşfedilebilir hale gelir.

Veri madenciliđi (Data Mining), büyük veri tabanlarından tahminlere dayalı bilgi çıkarma işlemidir. Veri ambarlarındaki saklı bilgiyi otomatik olarak keşfederek, kurumsal veri tabanlarında bulunan karmaşık ilişkileri ve kalıpları ortaya çıkarmak için istatistiksel analiz ve modelleme tekniklerini kullanır. Son 40 yılda, yapılandırılmış verileri işleme yönelik araçlar ve teknikler veri tabanlarından veri ambarlarına ve veri madenciliđine doğru gelişme göstermiştir. Veri ambarı uygulamaları iş açısından kritik önem taşımaktadır ve veri madenciliđi, bu büyük bilgi havuzlarından daha fazla değer çıkarabilir (Wang, 2003). Başka bir ifadeyle veri madenciliđi, verilerin içerisindeki desenlerin, ilişkilerin, deđişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların yarı otomatik olarak keşfedilmesidir, denilmektedir (Demiriz, 2006).

2.3. Veri Madenciliđi Uygulama Süreci

Veri madenciliđi, müşteri davranışlarının yönetimi, pazarlama faaliyetleri, satış tahminleri, kredi ve sigorta risk yönetimi, zararlı içeriklerin tespiti, müşteri memnuniyeti ve yeni müşteri kazanma gibi alanlarda yaygın olarak kullanılmaktadır. Bu yöntem, sınıflandırma, kümeleme, tahminleme ve aykırılık gibi problemlerde

verilerin analiz edilmesini sağlamaktadır. Ayrıca, veri madenciliği, veri setindeki birbirleriyle ilişkili verileri bir araya getirerek, aralarındaki bağlantıyı tespit etmeye çalışır. Bu bağlantılar gözle tespit edilemeyecek kadar karmaşık olabilir ve veri madenciliği, bu bağlantıların keşfedilmesine yardımcı olur. Bu nedenle, veri madenciliği sürecinin gerçekleştirilmesi, bu ilişkilerden faydalanmak için önemlidir. Bu süreç Şekil 1’de gösterilmektedir.



Şekil 1: Veri Madenciliği İşlem Süreci

Kaynak : Yazar tarafından oluşturulmuştur

2.4. Veri Madenciliği Teknikleri

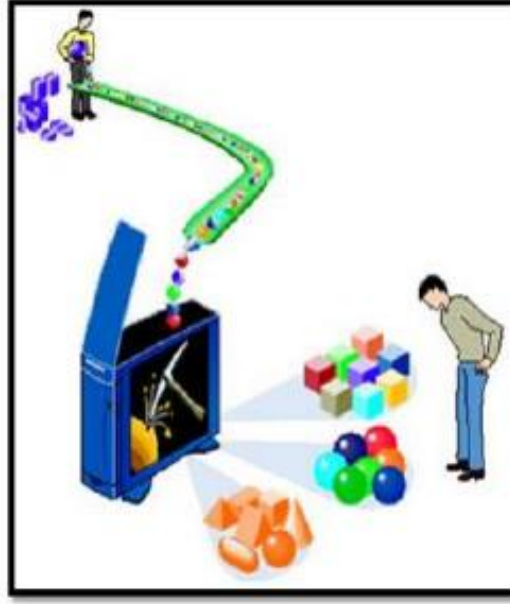
Veri madenciliği, büyük veri kümeleri içindeki kalıpları, ilişkileri ve trendleri keşfetmek için kullanılan bir dizi teknik ve yöntem içermektedir. Bu teknikler, veri ön işleme, veri görselleştirme, sınıflandırma, kümeleme, tahminleme, regresyon, derin öğrenme ve birliktelik kuralı gibi çeşitli konuları kapsamaktadır. Aşağıda, bu tekniklerin kısa bir açıklaması yer almaktadır:

Veri ön işleme: Veri madenciliği sürecinin ilk adımındır. Bu adımda, veri setleri temizlenir, düzenlenir, eksik veriler tamamlanır ve gereksiz veriler çıkarılır (Emeç & Özcanhan, 2023, s. 33-54).

Veri görselleştirme: Veri setlerinin görselleştirilmesi, verilerin daha iyi anlaşılmasına yardımcı olur. Bu adımda, tablolar, grafikler ve görseller kullanarak verilerin analizi yapılır.

Sınıflandırma: Veri setindeki örnekleri farklı sınıflara ayırmak için kullanılan bir tekniktir. Bu teknik, örneklerin etiketlenmesi veya sınıflandırılması için kullanılır.

Cinsiyetlere göre ilgi alanları farklılık göstermektedir; kadınlar genellikle alışveriş yapmaktan hoşlanırken, erkekler ise genellikle futbola ilgi duyarlar (Arslan , 2011).



Şekil 2: Veri Sınıflandırması

Kaynak: Arslan, Y (2011). "Veri Madenciliği". Çözumpark Web Sitesi,cozumpark.com E.T:01.11.23

Kümeleme: Benzer özelliklere sahip örnekleri gruplandırmak için kullanılan bir tekniktir. Bu teknik, veri setleri üzerinde yapılan analizler sonucunda farklı kümelere ayrılabilir (Puthier, 2008). Şekil 3’de kümeleme tekniği uygulanmış bir yapı bulabilirsiniz.



Şekil 3: Veri Kümelenmesi

Kaynak: Puthier, D (2008). Data Mining And Integration of Genomic Data For Gene regulatory Network Analysis.

Tahminleme: Gelecekteki olayların tahmin edilmesi için kullanılan bir tekniktir. Bu teknik, regresyon analizi gibi çeşitli yöntemlerle gerçekleştirilir.

Regresyon: Bağımlı ve bağımsız değişkenler arasındaki ilişkiyi inceleyen bir tekniktir. Bu teknik, doğrusal veya doğrusal olmayan bir ilişkiyi analiz etmek için kullanılabilir.

Derin öğrenme: Yapay sinir ağları kullanılarak, öğrenme süreci gerçekleştirilen bir tekniktir. Bu teknik, büyük veri kümelerindeki karmaşık desenleri tanımak ve anlamak için kullanılabilir.

Birliktelik kuralı: Veri setindeki birbiriyle ilişkili olan öğeleri analiz etmek için kullanılan bir tekniktir. Bu teknik, öğeler arasındaki bağlantıyı tespit ederek öngörülerde bulunmayı sağlar.

Bu teknikler, veri madenciliği sürecinin farklı aşamalarında kullanılabilir ve farklı veri türleri için farklı sonuçlar elde edilebilir.

2.5. Metin Madenciliği

Metin madenciliği (Text mining) doğal dil işleme (NLP) ve makine öğrenimi (ML) teknikleri kullanarak büyük miktarda metin verisini analiz eden bir veri madenciliği alt alanıdır. Metin verileri genellikle e-postalar, sosyal medya paylaşımları, müşteri yorumları, gazete makaleleri ve diğer belgeler gibi çeşitli kaynaklardan elde edilir. Metin madenciliği, bu metin verilerini otomatik olarak işleyerek anlamlı bilgileri çıkarmayı hedefler. Bu işlem sırasında, metin verileri ön işleme adımlarından geçirilir, yani kelime ayrıştırması, kelime kökleri çıkarımı ve stop word'lerin çıkarılması gibi işlemler yapılır. Daha sonra, metin verileri kelime frekansı, kelime dağılımı, duygusal analiz ve konu modelleri gibi teknikler kullanılarak analiz edilir.

Metin madenciliği, analiz edilmek üzere yapılandırılmamış metin verileri oluşturarak veri tabanlarındaki mevcut verileri arttırmak için kullanılabilecek yeni bir teknolojidir. Bu, klasik veri madenciliğinin bir uzantısından metinlere, “dünyanın kendisi hakkında yeni gerçekleri ve eğilimleri keşfetmek için büyük çevrimiçi metin koleksiyonlarının kullanılması” gibi daha karmaşık formülasyonlara kadar çeşitli

farklılıklara imkân sağlamaktadır. Metin madenciliği, veri madenciliği, doğal dil işleme, istatistik ve bilgisayar bilimi gibi çeşitli yaklaşımları içerisinde bulunduran disiplinler arası bir araştırma alanıdır.

Metin madenciliği, yazılı kaynaklardan (örneğin, belgeler, raporlar, e-postalar, bloglar ve sosyal medya paylaşımları gibi) elde edilen verilerin işlenerek anlamlı bilgi ve fikirlerin çıkarılması sürecidir. Bu süreçte, doğal dil işleme (NLP) teknikleri kullanılarak metinlerin anlaşılması, kategorize edilmesi, özetlenmesi ve bilgi çıkarma işlemleri gerçekleştirilir. Metin madenciliği, özellikle büyük veri setlerindeki bilgiyi hızlı ve verimli bir şekilde analiz edebilme imkânı sağlar. Bu analizler sonucunda, önceden keşfedilmemiş ilişkiler, trendler, kullanıcı davranışları ve müşteri geri bildirimleri gibi değerli bilgiler elde edilebilir. Metin madenciliği, web arama yöntemlerinden farklı olarak, önceden bilinmeyen bilgileri ortaya çıkarmayı hedeflemektedir ve sadece ilgili bilgiyi bulup gereksiz bilgileri ayıklama işlemi ile sınırlı değildir (Hearst, 2003).

Metin madenciliğinin uygulama alanları oldukça geniştir. Bu alanlar arasında sosyal medya izleme, müşteri memnuniyeti analizi, pazarlama araştırması, trend analizi, spam filtreleme, sözleşme analizi, tıp alanında hastalık teşhisi ve tedavi yönetimi gibi konular yer alır. Bu uygulamalar sayesinde, şirketler müşterilerinin ihtiyaçlarını ve isteklerini daha iyi anlayabilir, müşteri ilişkilerini yönetebilir ve iş süreçlerini optimize edebilirler. Bununla birlikte, metin madenciliği uygulamalarında da bazı zorluklar bulunmaktadır. Özellikle, dil çeşitliliği, duygu analizi, kelime anlamı, sentaks ve semantik yapılar gibi faktörler, doğru sonuçlar elde etmek için dikkate alınması gereken önemli konulardır.

2.5.1. Metin Madenciliğinin Kullanım Alanları

Metin madenciliği, yazılı metinlerden otomatik olarak bilgi çıkarmayı sağlayan bir yöntemdir. Bu yöntem, uzun süreler harcanması gereken görevleri sistematik hale getirerek avantaj sağlar. Metin madenciliği, yazılı iletişimin ve kaynakların kullanıldığı her alanda kullanılabilir ve çalışma alanı oldukça geniştir. Metin madenciliği, genellikle 6 ana başlık altında toplanmıştır (Seker, 2015, s. 30-32).

Bilgi Çıkarımı (Information Retrieval): Bilgi çıkarımı, ilgili konu veya sorun hakkında önemli bilgilerin toplandığı bir veri ön işleme aşaması yer alır. Bu aşamada, verilerin kaynağına göre farklı teknikler kullanılarak veriler toplanır, özetlenir, filtrelenir ve temizlenir. Bu sayede, analiz sürecinde kullanılacak daha kaliteli veri setleri oluşturulur ve sonraki aşamalarda daha doğru sonuçlar elde edilmesi hedeflenir.

Doğal Dil İşleme (Natural Language Processing): Bu aşamada, doğal dilde yazılmış metinlerin dilbilgisi yapısı, sözcük anlamları ve cümle yapıları gibi özellikleri analiz edilerek, bu metinlerin daha anlamlı ve işlevsel hale getirilmesi amaçlanır.

Adlandırılmış Varlık Tanıma (Named Entity Recognition): Bu aşamada, metindeki belirli varlıkların (örneğin, kişilerin, şirketlerin, yerlerin) tanınması ve etiketlenmesi hedeflenir. Bu tanımlama ve etiketleme işlemi, istatistiksel bazı özelliklerin çıkarılmasında ve metinden anlamsal bilgilerin elde edilmesinde önemli bir rol oynar.

Referans Çözümleme (Coreference Resolution): Bu süreçte bir metinde geçen farklı kelimeler veya kelime grupları, aynı varlığa referans veriyorsa bunların birbirleriyle eşleştirilmesi ve birleştirilmesi işlemidir. Bu sayede metindeki anlam daha net hale getirilebilir ve yapılan analizlerin doğruluğu artırılabilir.

Duygu Analizi (Sentiment Analysis): Bu aşamada metinde yer alan cümlelerin pozitif, negatif veya nötr duygusal içeriklerinin tespit edilmesi ve sınıflandırılması hedeflenmektedir.

2.5.2. Metin Madenciliği Süreci

Metin madenciliği süreci, çeşitli kaynaklardan elde edilen bir belge koleksiyonu ile başlar. Metin madenciliği aracı, belirli bir belgeyi alır ve önceden işlemek için biçim ve karakter kümelerini kontrol eder. Daha sonra belge, metin analizi aşamasından geçer. Bu aşamada, metinden anlamlı bilgi elde etmek için anlamsal analiz yapılır. Farklı amaçlara yönelik birçok metin analizi tekniği mevcuttur ve organizasyonun ihtiyaçlarına göre tekniğin kombinasyonları kullanılabilir. Bazı durumlarda, metin analizi teknikleri, bilgi çıkarılana kadar tekrarlanabilir. Sonuçta elde edilen bilgi, bir yönetim bilgi sistemi içine yerleştirilebilir ve bu sistem

aracılığıyla kullanıcılar için bol miktarda veri elde edilebilir. Metin madenciliği süreci şekil 4’te gösterilmektedir (Gaikwad, Chaugule, & Patil, 2014).



Şekil 4: Metin Madenciliği İşlem Süreci

Kaynak : Yazar tarafından oluşturulmuştur

Şekilde gösterilen süreç, metin madenciliği teknikleri yardımıyla gerçekleşmektedir. Bu teknikler, üzerinde çalışılan konunun problem ve ihtiyacına göre değişkenlik göstermektedir.

2.5.3. Metin Madenciliği Teknikleri

Metin madenciliği, önemli bir role sahip çeşitli sayıda teknik kullanır. Teknikler birbirinden farklılık göstermektedir. Metin madenciliği teknikleri ve bu tekniklerin amaçlarının karşılaştırılması aşağıdaki çizelgede gösterilmektedir (Dang & Ahmad, 2014, s. 22-25).

Tablo 1: Metin Madenciliği Teknikleri

Teknik	Amaç
Elde Etme	Yapılandırılmamış metinden değerli bilgiye ulaşmak
Çıkarım	Yapılandırılmış veri tabanından bilgi çıkarımı
Özetleme	Temel noktaları ve genel anlamları koruyarak uzunluğu azaltmak
Kategorizasyon	Belgeleri kategorilere ayırmak
Kümeleme	Metinleri kümelemek, sınıflandırmak ve analiz etmek

Bu tekniklerin kullanımı için konuşma dilini ve yazılı kaynakları bilgisayarın anlayacağı formata dönüştüren doğal dil işleme araçlarından faydalanılmaktadır. Doğal dil işleme yöntemleri, metin madenciliği çalışmaları için fazlasıyla önem taşımaktadır. Bu yöntemler, metinlerin daha yüksek doğrulukla analiz edilmesini sağlamak ve veriyi modele hazırlamak için kullanılan birincil araçlardır. Bu araçlar, metin madenciliği teknikleri ile hibrit olarak kullanılmaktadır.

3. DOĞAL DİL İŞLEME

Dil yeteneği, insan beyninin nasıl işlediğini anlamak için önemli bir araştırma alanı olan dilbilim, aynı zamanda bilgisayarlar için doğal dilin anlaşılması, yorumlanması ve üretilmesi gibi işlevleri yerine getiren bir mühendislik alanı olan Doğal Dil İşleme (DDİ) için de önemlidir. DDİ, yapay zekâ, biçimsel dilbilim, bilgisayar destekli dilbilim, bilişsel psikoloji gibi alanlardan faydalanarak doğal dilin işlenmesi için kuram, yöntem ve teknolojiler geliştirir. Önceleri yapay zekânın küçük bir alt alanı olarak görülen DDİ, günümüzde bilgisayar bilimlerinin temel bir disiplini haline gelmiştir.

Doğal dil işlemenin amacı, doğal dillerin yapısını kurallı bir şekilde çözümleyerek anlaşılması ve yeniden üretilmesidir. Bu çözümlemenin sağladığı kolaylıklar arasında, yazılı dokümanların otomatik olarak çevrilmesi, soru-cevap sistemleri, otomatik metin özetleme, konuşma sentezi, otomatik konuşma ve komut anlama, konuşma üretimi, bilgi sağlama gibi birçok alan bulunmaktadır (Oğuz, 2009).

Doğal Dil İşleme (DDİ) süreci, genel olarak Doğal Dil Anlama (DDA) ve Doğal Dil Oluşturma (DDO) süreçlerinin bütünüdür. DDA, bilgisayarların okuduğunu anlamalarına odaklanırken, DDO bilgisayarların yazabilmesini sağlar. Doğal Dil İşleme süreci, veri ön işleme, morfoloji, sözdizim, anlamsal analiz, anlamsal ilişkilendirme, duygu analizi ve metin üretimi bileşenlerinden oluşur. Bu bileşenler, belirli bir doğal dil verisi üzerinde uygulanarak, dilin anlaşılması ve oluşturulması için kullanılabilir. Şekil 5'te Doğal Dil İşleme bileşenleri gösterilmektedir (Özmutlu, 2021).



Şekil 5: Doğal Dil İşleme Bileşenleri

Kaynak : Yazar tarafından oluşturulmuştur

3.1 Doğal Dil İşlemenin Kullanım Alanları

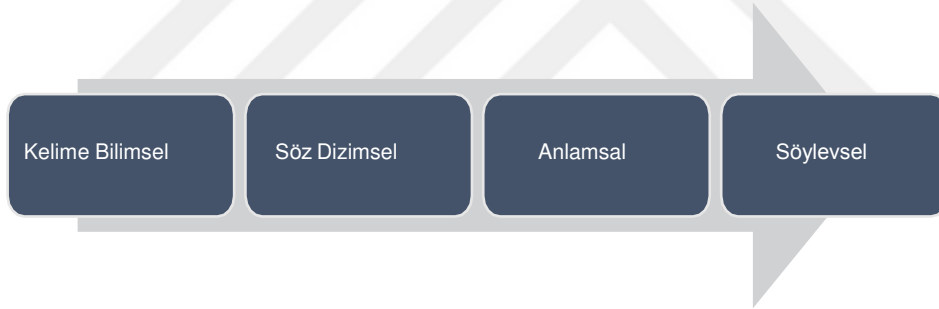
DDİ araçları, metin türündeki veri ile ilgili ihtiyaç duyulan tüm düzenleme, bilgiye erişme, anlam çıkarma ve diller arası çeviri işlemleri için kullanılmaktadır.

Doğal dil işleme kullanım alanları aşağıdaki gibidir (Adalı, 2012):

- Soru-cevap sistemleri
- Chatbotlar
- Dijital asistanlar (Amazon Alexa, Google Home gibi)
- Duygu analizi (sentiment analysis)
- İçerik filtreleme
- Bilgi çıkarma ve sınıflandırma
- Sosyal medya analizi
- Tıbbi dil işleme (Electronic Health Records gibi)
- Konuşma tanıma ve doğal dil anlama
- Doğal dilde arama motorları
- Finansal dil işleme (ticaret stratejileri gibi)

3.2 Doğal Dil İşleme Süreci

Doğal dil işleme, kelime bilimsel, söz dizimsel, anlamsal veya söylevsel açıdan metin verilerini analiz etmek amacıyla kullanılan bir teknolojidir. Bu analiz süreci, doğal dil işleme sürecini oluşturur. Şekil 6'da doğal dil işleme sürecinin aşamaları gösterilmektedir.

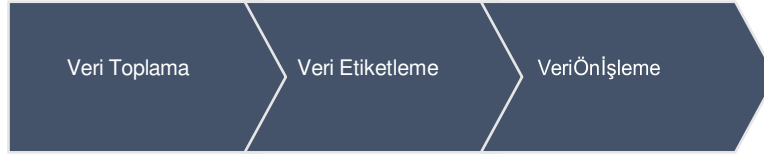


Şekil 6: Doğal Dil İşleme Süreci

Kaynak : Yazar tarafından oluşturulmuştur

3.3 Doğal Dil İşlemede Veri Hazırlama Süreci

Doğal dil işlemede kullanılacak verinin uygun hale getirilmesi, bir sınıflandırıcı modelin eğitilmesi için gereklidir. Bu süreç, veri toplama, etiketleme ve ön işleme aşamalarından oluşmaktadır. Veri hazırlığı için izlenen adımlar, Şekil 7'de gösterilmektedir. Bu adımlar arasında veri toplama, verinin düzenlenmesi, anlamsal ve sözdizimsel analiz, tokenizasyon, duraklama işareti kaldırma, kelime köklerinin belirlenmesi, kelime dağarcığı oluşturma ve veri bölümlendirme gibi işlemler yer almaktadır. Bu işlemler verinin kalitesini artırır ve sonuçta daha iyi bir doğal dil işleme modeli elde edilmesine yardımcı olur.



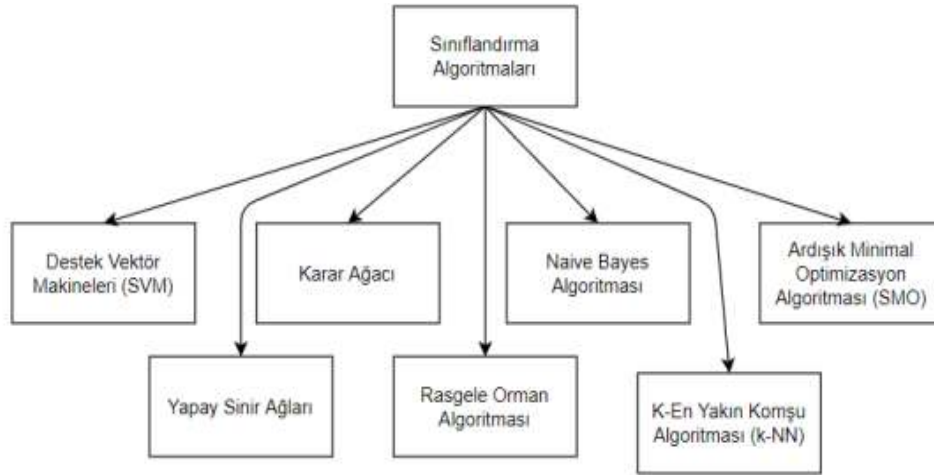
Şekil 7: Veri Hazırlığı Adımları

Kaynak : Yazar tarafından oluşturulmuştur

4. METİN SINIFLANDIRMA

Bilgiye erişimdeki artış, veri kaynaklarının artması ve veri toplama teknolojilerinin gelişmesi sonucu büyük veri setleri ortaya çıkmıştır. Bu veri setleri, işlenmesi, analizi ve anlamlı bilgi çıkarımı için önemli kaynaklar olabilir. Ancak, büyük veri setleri içinde kaybolmak kolaydır ve doğru bilgiye erişim zorlaşır. Bu nedenle, verileri uygun kategorilere göre sınıflandırmak ve etiketlemek, veri yönetimi sürecinin önemli bir parçasıdır. Verilerin sınıflandırılması, veri erişimini kolaylaştırır, veriler arasındaki ilişkileri ortaya çıkarır ve veri analizinde doğru sonuçlar elde etmeyi mümkün kılar. Ayrıca, veri sınıflandırması, veri yönetimi ve güvenliği için de önemlidir, çünkü doğru şekilde sınıflandırılmış veriler, veriye erişim haklarını sınırlamak ve gizliliği korumak için kullanılabilir (Erten, 2015).

Sınıflandırma, belirli özelliklere sahip verileri gruplamak ve yeni gelen verileri bu gruplara atamak için kullanılan bir makine öğrenimi yöntemidir. Bu yöntem genellikle, önceden etiketlenmiş (yani sınıflandırılmış) bir veri kümesi kullanarak modelin eğitilmesini ve ardından bilinmeyen verilerin bu modele göre sınıflandırılmasını içerir. Eğitim ve test veri kümesi, sınıflandırma algoritmalarının kullanıldığı modelin öğrenme ve test aşamalarında kullanılmaktadır. Farklı sınıflandırma algoritmaları arasında karar ağaçları, destek vektör makineleri, rastgele ormanlar ve yapay sinir ağları gibi yöntemler yer almaktadır (Demirel, 2019).



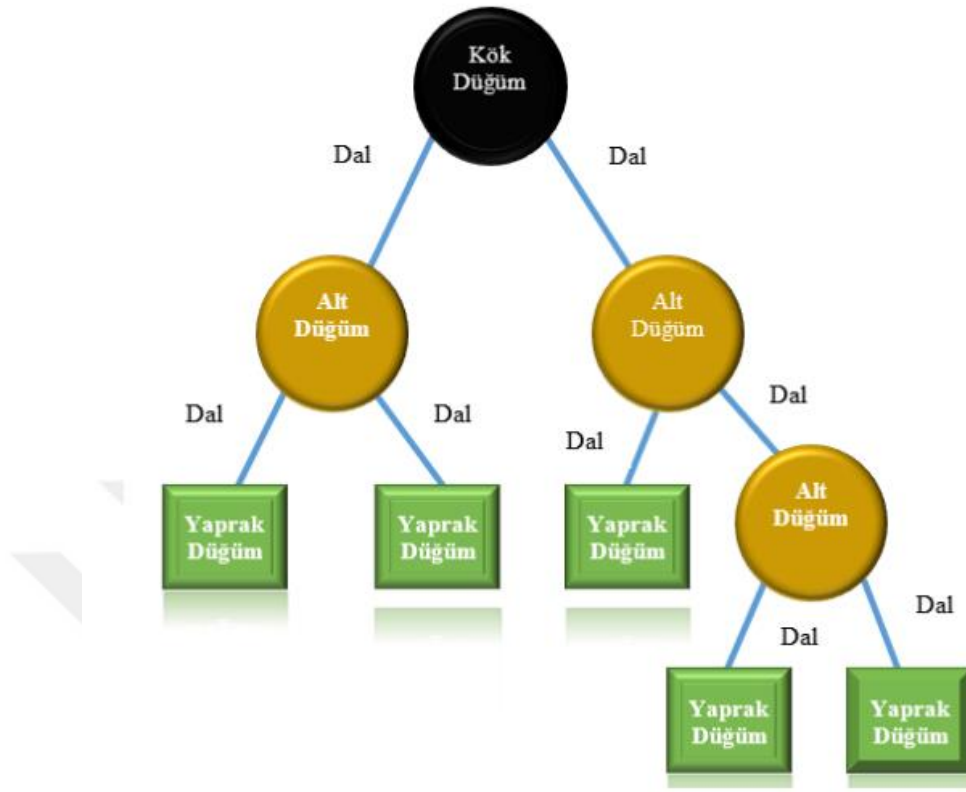
Şekil 8: Metin Sınıflandırma Algoritmaları

Kaynak : Yazar tarafından oluşturulmuştur

4.1 Metin Sınıflandırma Algoritmaları

4.1.1 Karar ağaçları

Karar ağaçları, sınıflandırma ve regresyon problemlerinde kullanılan popüler bir makine öğrenimi algoritmasıdır. Her bir düğümde, bir öznelik değerine göre ayrılan alt dallar oluşturulur ve sonunda yapraklarda sınıflandırma veya tahmin değerleri belirlenir. Bu nedenle, karar ağaçları sınıflandırma problemlerinde sınıfları açıklayıcı kurallar halinde sunar ve yorumlanması kolaydır. Ayrıca, diğer makine öğrenimi algoritmalarına göre daha hızlı bir şekilde eğitilebilirler ve genel olarak overfitting'e (aşırı uyum) eğilimli değildir. Ancak, aşırı dallanma durumunda karar ağaçları karmaşık hale gelebilir ve bu durum, modelin genelleme performansını olumsuz etkileyebilir (Demirel, 2019).



Şekil 9: Karar Ağaçları Şeması

Kaynak: Demirel, Ş (2019). Karar Ağacı Algoritmaları Ve Çocuk İşçiliği Üzerine Bir Uygulama.Yüksek Lisans Tezi. Marmara Üniversitesi Sosyal Bilimler Enstitüsü konometri Anabilim Dalı, 115, İstanbul.

Karar ağaçları, veri sınıflandırma algoritmalarından biridir ve genellikle sınıflandırma problemlerinin çözümünde kullanılır. Ağacın yapısı, kök düğüm, alt düğüm ve yaprak düğüm olmak üzere üç düğümden oluşur. Her düğüm, verinin bir özelliğini ifade eder. Yaprak düğümleri sınıf etiketlerini taşırken, yapraklara giden dallar özellikler üzerinde yapılan işlemleri gösterir. Ağacın bütün düğümleri, belli bir özelliğe ait kuralları ifade eder ve bu kurallara göre karar aşamasından sonra seçilecek dallar belirlenir. Karar ağaçları, kolay entegrasyonu ve yorumlana bilirliliği nedeniyle sıklıkla kullanılır (Amanet, 2017).

4.1.2 Destek vektör makinesi

Destek Vektör Makineleri (SVM), verileri sınıflandırmak veya regresyon analizi yapmak için kullanılan bir makine öğrenimi algoritmasıdır. 1990'ların başında Vladimir Vapnik ve ekibi tarafından geliştirilmiştir (Akşehirli, Ankaral, Aydın, &

Saraçlı, 2013, s. 19-28). SVM, veri noktalarını, sınıflar arasındaki bir ayrım çizgisine (hiper düzlem) en uzak olan noktalara dayanarak sınıflandırır. Bu nedenle, SVM, yüksek boyutlu veri kümelerinde etkilidir ve diğer sınıflandırma yöntemlerine kıyasla daha az veri gerektirir. Ayrıca, SVM, çok sayıda özellikle de iki sınıflı verilerde yüksek doğruluk oranları sağlar. Bunun yanı sıra, sağlık sektöründe de kullanılan SVM, hastalıkların teşhisinde ve tıbbi görüntüleme tekniklerinde de etkili bir yöntem olarak kullanılmaktadır.

4.1.3 K-en yakın komşu (K-nearest neighbors)

K-NN algoritması, sınıflandırma problemlerinde kullanılan etkili bir algoritmadır. Bu algoritma verileri düzlem üzerinde yerleştirerek yeni verinin sınıfını en yakın k komşusuna bakarak belirler. Uzaklık hesaplaması yapılırken farklı yöntemler kullanılabilir. Bu sayede büyük veri setleri üzerinde hızlı ve doğru sınıflandırma işlemleri gerçekleştirilebilir. K-NN algoritması özellikle yoğunluğa dayalı metin sınıflandırma problemlerinde kullanılmaktadır. Hızlı ve etkili sonuçları nedeniyle diğer sınıflandırma yöntemleriyle karşılaştırıldığında tercih edilmektedir.

4.1.4 Naive Bayes

Naive Bayes algoritması veri boyutunun yüksek olduğu durumlarda, kriterlerin bağımsız olduğu veya bağımsız varsayılabilir olduğu durumlarda kullanılır. Ayrıca büyük veritabanlarına uygulandığında yüksek doğruluk ve hız sağlaması tercih edilmesinde etkili bir faktördür. Naive Bayes, düşük sapma ile çalışabilen, yeni eğitim verileri ile hızlı güncellenebilen, veri kümesindeki gürültüyü işleme yeteneği yüksek ve eksik değerleri işleme yeteneği olan bir sınıflandırma yöntemi olarak diğer yöntemlerle kıyaslandığında yüksek hesaplama verimliliği gibi avantajlara sahiptir (Cherian & Bindu, 2017, s. 68-73).

4.1.5 Yapay Sinir Ağları

Yapay sinir ağları (ANN), insandaki sinir ağlarından esinlenerek geliştirilmiş bir makine öğrenmesi yöntemidir. Bu yöntem, veriden öğrenebilme, genelleme yapabilme ve sınırsız sayıda değişkenle işlem yapabilme gibi özellikleri içermektedir. YSA'nın en küçük birimi nöron olarak adlandırılır ve nöronlar birleşerek katmanları,

katmanlar ise bir ağı oluşturur (Arslan, Kaynar, & Yüksek, 2017, s. 381-388). YSA en az iki katmana sahip olmalıdır: girdi katmanı ve çıktı katmanı. Girdi katmanı, verilerin ağ tarafından okunmasını sağlayan ve özellik sayısı kadar nöron içeren katmandır. Çıktı katmanı ise ağ tarafından işlenerek üretilen sonuçların dışarıya aktarıldığı katmandır. Gizli katman, girdi ve çıktı arasında birtakım hesaplamaları birleştirerek sonuçlandırır.

4.2 Metin Sınıflandırmada Performans Ölçütleri

Metin sınıflandırma görevinde performans ölçümü, modelin doğru bir şekilde etiketlediği metinlerin sayısını ve yanlış etiketlenen metinlerin sayısını ölçmek için kullanılan bir yöntemdir (Lipton, Elkan, & Narayanaswamy, 2014). Bu yöntem, modelin yetkinliğini ölçmek için farklı metriklerin kullanılmasını gerektirir. Bu doğrultuda, doğruluk (accuracy), kesinlik (precision), duyarlılık (sensitivity), özgüllük (Specificity), negatif prediktif değer (NPD), F-ölçeği, ROC (Receiver Operating Characteristics) eğrisi gibi metrikleri kullanılmaktadır (Amanet, 2017). Bu metriklerin hesaplanabilmesi için karmaşıklık matrisi (confusion matrix) kullanılır. Karmaşıklık matrisi, sınıflandırma modellerinin performansını objektif bir şekilde değerlendirmek için kullanılan bir araçtır. Bu matris, modelin tahmin ettiği sınıflar ile gerçek sınıflar arasındaki benzerlik ve farklılıkları gösterir. Bu sayede, modelin hangi sınıfları doğru tahmin ettiği ve hangi sınıfları yanlış tahmin ettiği açıkça görülebilir.

Bir sınıflandırma modelinin performansını ölçmek için, karmaşıklık matrisi kullanılarak doğru pozitif, yanlış pozitif, doğru negatif ve yanlış negatif kavramlarına dayalı metrikler hesaplanır. Bu metrikler, modelin sınıflandırma doğruluğu, kesinliği, duyarlılığı ve F1 skoru gibi performans ölçümlerini sağlar. Bu ölçümler, modelin performansını daha anlamlı bir şekilde değerlendirmemizi sağlar.

<i>Confussion Matrix</i>		Gerçek (Actual) Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahminlenen (Predicted) Sonuçlar	Pozitif (1)	TP [1, 1] True Pozitif	FP [1, 0] False Pozitif
	Negatif (0)	FN [0, 1] False Negatif	TN [0, 0] True Negatif

Şekil 10: Karmaşıklık Matrisi

Kaynak: Amanet, H (2017). Türkçe sosyal medya metinlerinde duygu analizi. Yüksek Lisans Tezi, Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Ve Bilgisayar Bilimleri Anabilim Dalı, 95, Trabzon.

Karmaşıklık matrisinde yer alan kavramlar şu şekilde açıklanabilir:

TP (True Positive); gerçek pozitiflerin sayısı, yani doğru olarak sınıflandırılmış pozitif etiketlenen örneklerin sayısıdır.

FN (False Negative); yanlış negatiflerin sayısı, yani yanlış olarak sınıflandırılmış negatif etiketlenen örneklerin sayısıdır.

FP (False Positive); yanlış pozitiflerin sayısı, yani yanlış olarak sınıflandırılmış pozitif etiketlenen örneklerin sayısıdır.

TN (True Negative); gerçek negatiflerin sayısı, yani doğru olarak sınıflandırılmış negatif etiketlenen örneklerin sayısıdır.

Bu kavramlar, sınıflandırma modelinin performansını değerlendirmek için kullanılan çeşitli metriklerin hesaplanmasında kullanılır.

4.2.1 Doğruluk (Accuracy)

Doğruluk oranı (accuracy), bir sınıflandırma modelinin doğru bir şekilde sınıflandırdığı örneklerin oranını ifade eder. Bu metrik, tüm sınıfların eşit derecede önemli olduğu durumlarda yaygın bir şekilde kullanılır. Ancak, sınıflar arasında

dengelesizlik varsa (örneğin bir sınıfın diğerlerine göre daha az örneği varsa), doğruluk oranı yanıltıcı sonuçlar verebilir. Bu nedenle, doğruluk oranı tek başına yeterli değildir ve kesinlik, duyarlılık ve F1 skoru gibi diğer performans ölçüleri de göz önünde bulundurulmalıdır.

Ayrıca, doğruluk oranı yalnızca modelin eğitim setindeki performansını ölçer ve aşırı uyum (overfitting) durumlarında yanıltıcı sonuçlar verebilir. Bu nedenle, modelin performansını gerçek dünya verilerinde test edebilmek için ayrı bir test seti kullanılmalıdır. Test seti, eğitim setinden bağımsız ve rastgele örneklerden oluşmalıdır (Goyal & Mehta, 2012).

Bir sınıflandırma modelinin doğruluk oranı, modelin tüm test örneklerini doğru bir şekilde sınıflandırma yüzdesini ifade eder. Bu oran, modelin genel başarı seviyesini belirlemek için kullanılır ve aynı zamanda modelin performansını diğer metriklerle karşılaştırmak için de temel bir ölçüt olabilir. Doğruluk oranı, doğru sınıflandırılan örneklerin toplam örnek sayısına bölünmesiyle hesaplanır (Gündüz, 2013).

$$Doğruluk = \frac{TP+TN}{TP+FP+TN+FN} \quad (3.1)$$

4.2.2 Kesinlik Ölçütü

Kesinlik ölçütü, sınıflandırma modelinin pozitif olarak sınıflandırdığı örnekleri temel alarak, gerçek pozitiflerin sayısını pozitif olarak sınıflandırılan örneklerin sayısına bölerek hesaplanan bir performans metriğidir. Yani kesinlik, modelin pozitif olarak sınıflandırdığı örneklerin ne kadarının gerçek pozitif olduğunu gösterir. Kesinlik ölçütü, modelin yanlış pozitifler üretmesinin ne kadar önemli olduğu durumlarda kullanılır (Gündüz, 2013).

$$Kesinlik = \frac{TP}{TP+FP} \quad (3.2)$$

4.2.3 Duyarlılık Ölçütü

Duyarlılık (recall) ölçütü, gerçek pozitif örneklerin ne kadarının model tarafından pozitif olarak sınıflandırıldığını ölçer. Yani, toplam pozitif örnek sayısından kaç tanesinin doğru olarak tanındığını ölçer. Bu ölçüt, bir sınıflandırma modelinin gerçek pozitifleri kaçırmaması ve pozitif olarak sınıflandırılan örneklerin gerçekten pozitif olması gereken durumlarda yanlış pozitif vermemesi için önemlidir. Duyarlılık ölçütü, aynı zamanda hassasiyet (sensitivity) olarak da adlandırılır.

$$Duyarlılık = \frac{TP}{TP+FN} \quad (3.3)$$

4.2.4 Özgüllük

Özgüllük, modelin gerçek negatifleri doğru bir şekilde sınıflandırma yeteneğini ölçen bir performans metriğidir. Özgüllük, TN (Doğru Negatif) değerinin toplam negatif örnek sayısına bölünmesi ile hesaplanır. Yani, modelin gerçek negatiflerin tespitindeki doğruluğunu gösterir.

$$Özgünlük = \frac{TN}{TN+FP} \quad (3.4)$$

4.2.5 Negatif Prediktif Değer

NPD ölçütü, negatif sınıflandırılmış örneklerin gerçek negatif örneklerle ne kadar uyumlu olduğunu ölçer. Doğru olarak sınıflandırılmış negatif örnek sayısının toplam negatif örnek sayısına bölünmesiyle hesaplanır. Bu ölçüt, sınıflandırma modelinin yanlış pozitiflerin sayısını minimize etme konusunda ne kadar başarılı olduğunu ölçer.

$$NPD = \frac{TN}{TN+FN} \quad (3.5)$$

4.2.6 F-Ölçeği

F-ölçütü, sınıflandırma modelinin performansını değerlendirmek için kullanılan bir ölçüttür ve kesinlik ve duyarlılık ölçütlerinin birleşiminden oluşur. Kesinlik ölçütü, pozitif olarak sınıflandırılmış örneklerin gerçek pozitif örneklerden oranını ifade ederken, duyarlılık ölçütü, gerçek pozitif örneklerin tespit edilme oranını ifade eder. F-ölçütü, bu iki ölçütün ağırlıklı harmonik ortalamasının alınması sonucu hesaplanır ve sınıflandırma performansının iyiliğini ifade eder. Yüksek bir F-ölçütü değeri, modelin hem kesinlik hem de duyarlılık açısından iyi bir performans sergilediğini gösterir.

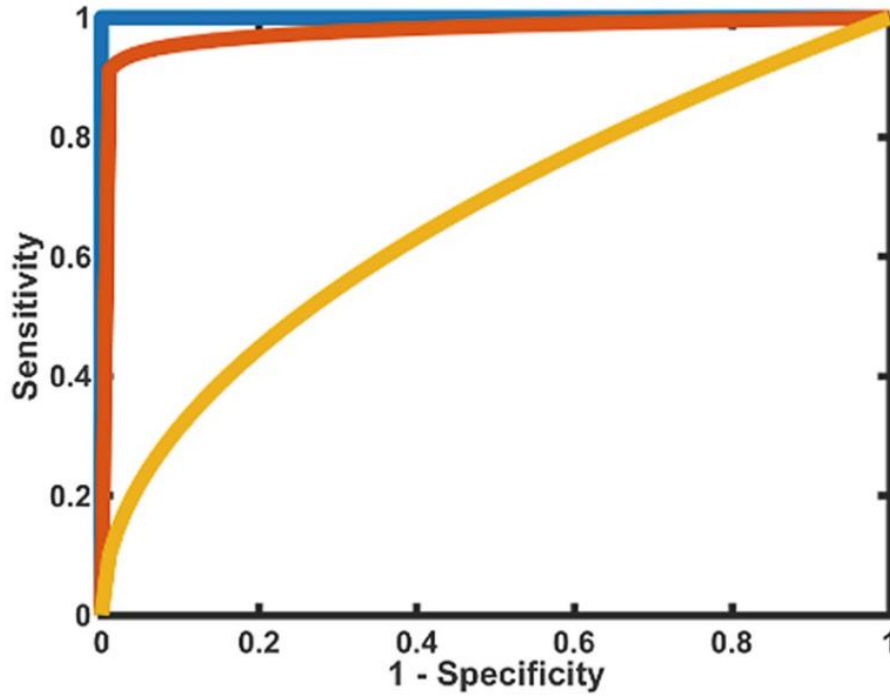
$$F - \text{Ölçeği} = \frac{2 * \text{Kesinlik} * \text{duyarlılık}}{TP + TN + FN + FP} \quad (3.6)$$

4.2.7 ROC Eğrisi

ROC (Receiver Operating Characteristic) eğrisi, sınıflandırma modellerinin performansını değerlendirmek için kullanılan bir grafiksel araçtır. Genellikle dengesiz veri setleriyle çalışırken, sınıflandırma modellerinin hassasiyetini (sensitivite) ve özgüllüğünü (specificity) değerlendirmek için kullanılır.

ROC eğrisi, doğru pozitif oranının (TPR) yanlış pozitif orana (FPR) karşı çizilmesiyle oluşur. FPR, gerçek negatiflerin ne kadarının yanlış pozitif olarak sınıflandırıldığını gösterirken, TPR, gerçek pozitiflerin ne kadarının doğru pozitif olarak sınıflandırıldığını gösterir. ROC eğrisi, farklı sınıflandırma eşik değerlerinde TPR ve FPR değerlerini çizerek oluşur.

ROC puanı, ROC eğrisinin altında kalan alan (AUC) olarak hesaplanır. ROC eğrisinin altında kalan alan, 0 ile 1 arasında bir değer alır. AUC değeri 1'e yaklaştıkça, sınıflandırma modelinin performansı o kadar yüksek olarak değerlendirilir. AUC değeri 0.5'e yaklaştıkça, modelin performansı rastgele tahmin etmekle aynı seviyededir. AUC değeri 0'dan küçük olamaz.



Şekil 11: Roc Eğrisi

Kaynak : Yazar tarafından oluşturulmuştur

5. METİN MADENCİLİĞİ TEKNİKLERİ KULLANILARAK TÜRKÇE MÜŞTERİ YORUMLARININ SINIFLANDIRILMASI

Günümüzde dijital platformlarda satılan ürünlere yapılan yorumlar önemli bir veri kaynağıdır. Bu çalışmada, Türkçe yorumların analizleri ele alınacak ve Kaggle platformunda 32 ana konu üzerinde yapılmış olan 400 bin kayıt içeren bir veri seti kullanılacaktır. Veriler, makine öğrenmesi yöntemleri ile eğitilmeden önce ön işleme aşamalarından geçirilecektir. Daha sonra verilerin istatistiksel analizleri yapılacaktır. Son olarak, öğrenme yöntemleri kullanılarak elde edilen sonuçlar karşılaştırılacak ve en iyi performans gösteren algoritma tespit edilecektir. Önerilen modele test verileri uygulanarak, Türkçe müşteri yorumları kategorize edilecektir. Bu çalışma, dijital platformlarda yapılan yorumların analizi ve sınıflandırılması konusunda önemli bir katkı sunacaktır.

5.1 Veri setinin elde edilmesi

Bu çalışma için kullanılan veri seti, dijital platformlarda satılan ürünlere yapılan yorumlardan oluşmaktadır. Toplamda 430 bin kayıt içeren veri seti, 32 farklı kategoriye ayrılmıştır ve her kategori yaklaşık 13 bin yorum içermektedir. Veriler, Türkçe web sitelerinden toplanmıştır ve ürünlerin yanı sıra ürün kategorilerine yapılan yorumları da içermektedir.

5.2 Veri setinin incelenmesi

Bu çalışmada kullanılan veri seti, dijital platformlarda satılan ürünlere yapılan Türkçe yorumlardan oluşuyor. Toplamda 430 bin kayıt içeren bu veri seti, 32 farklı kategoriye ayrılmış ve her bir kategori yaklaşık 13 bin yorum içeriyor. Veri seti dengeli bir yapıya sahip. Türkçe web sitelerinden toplanan veriler, ürünlerin yanı sıra ürün kategorilerine yapılan yorumları da içeriyor. Veri seti csv formatında indirilerek işlemlere başlandı. Veri setinde, her yorumun kategori bilgisini gösteren "kategori" adında bir kolon ve yorumun kendisini gösteren "text" adında bir kolon bulunmaktadır. Veri seti 431306 satır ve 2 kolondan oluşmaktadır.

category	text
0	Altus Hırdavat Yapı Malzemeleri Drone Diye Kargodan Lastik Ayakkabı Çıktı."Instagram'da dolanırken sponsorlu bir bağlantı gördüm. Drone satışı yapılıyor. Normalde böyle şeylere inanmam ancak takipçi sayısının fazla olması, numaralarının olması, ödemeyi peşin değil karşı ödemeli ödenmesi, fotoğraflara yapılan yorumlar vs.... Az da olsa güvenernek ben de sipariş vermek istedim...Devamını oku
1	Albay Bilgisayar Garanti Yalanı İle Yanılttı, Garanti kapsamında yer alan Casper bilgisayarım garanti belgesi ile birlikte izmit Casper bilgisayar yetkilisi servisine albay bilgisayara bıraktım. Önce almak istemedi uzun ikna cabası ve uğraş sonunda zorla garanti dahiline bıraktım bilgisayar açılmıyordu. Sonrasında ertesi gün bilgisayarın yapıldı...Devamını oku
2	Okyanus Petshop Ürün İade Sorunu! Satıcıdan GittiGidiyor üzerinden 385 TL tutarında Frolicat Interactive Kedi Lazer oyuncacı satın aldım. Ürünü açıp bir kez çalıştırdığımda çok gürültü çıkardığını fark ettim ve ürünü iade etmeye karar verdim. Hatam iade sürecinin 14 gün içinde olması gerektiğini bilmemektir. Ürünü teslim aldıktan son...Devamını oku
3	Anda İhtiyaç Ve Tüketim Maddeleri Knee Torik Dizlik Esnemiyor! Anda ihtiyaç ve tüketim maddeleri pazarlama san. Lim. Şti. Den 139 TL kapıda teslim knee torik adli ürünü satın aldım. Ürün 0815774588452 takip no... ile Aras Kargo tarafından tarafıma teslim edildi. Ürün kargoya firma tarafından 26.05.2020 tarihinde kargoya verilmiş. Ürünün üst dizlik bedeni bacağım...Devamını oku
4	BB Kargo Dağıtım'dan Aldığım Ürün Bozuk Çıktı Ve Muhafaz Bulamıyorum! Fotoğrafını göndermiş olduğum ürünün kablosunda temassızlık var ve ne fatura ne numara sadece gönderen kargonun adı var ve onlara da ulaşamıyorum ayrıca kargo daha bugün elimde geçti Denizli Marmaris gezmiş ürün. Beni arayan numarayı buldum ona da ulaşamıyorum.Devamını oku
...	...
431301	Ulasim Pegasus Kabin Boyu Bagajim İçin 20 Euro Aldılar,Arkadaşlarımla Prag'dan İstanbul'a dönüş seferinde ilk defa Pegasus hava yollarını kullandım ve çok kötü bir havayolu. Pegasus'un kuralları gereği ekstra bagaj parası ödememek için valiziniz kabin içi valizi olmalı kuralına uyarak sürekli kabin içine aldığım 7.8 kg'lık valizimi aldım. Her uçağa sığ...Devamını oku
431302	Ulasim Pegasus Sefer Saat Değişikliği,"05.02.2020 tarihinde Amsterdam'dan 16:25'de kalkması planlanan uçuşumuz, 3 kişinin ölümüyle sonuçlanan kaza sebebiyle pistin kapatılması sonucu 23:00'da kalkıp İstanbul Yeni Havalimanına saat 04:00'da iniş yapmıştır. Bu süre zarfında tarafımıza hiçbir bilgi verilmemiş, kabin bagajlarımızı uçaktan al...Devamını oku
431303	Ulasim THY Saygısız Müşteri Temsilcisi!,"THY'den bileti aldığım biletim isteğe bağlı iptal iade ve değişiklik yapılmasını istedim, AMA buna istinaden ödediğim paranın sadece sağlık sigortası olduğunu iddia ettiler.
431304	Ulasim KLM Hava Yolları Bilgisizlikleri,Kraliyet firması imiş bizim Topkapı Sultançiftliği dolmuş hattı bile daha iyi hizmet veriyor biletim değişiyor ne T** nin haberi var ne de Kim ye ulaşabiliyorum. Aslında ulaştım ingilizce devam edelim diyen adama ben türkçe biliyorum dedim. Adam Türkçe için 10 dakika sonra arayın dedi. Ama çok bekle...Devamını oku
431305	Ulasim Pegasus Çağrı Merkezi,"Pegasus internet sayfasından aldığım biletime evcil hayvanımı ekletmek için 0850 250 67 02, 0888 228 12 12 numaralı hatları en az 5 kere aradım her birinde ortalama 20 dk bekledim dakikası 1,5 TL den faturalandırıldığı bilgisi paylaşıyorlar. Ben böyle bir şey görmedim uçuşuma saatler kaldı ve ben ...Devamını oku

431306 rows x 2 columns

Şekil 12: Veri seti içeriği

Kaynak : Yazar tarafından oluşturulmuştur

5.3 Veri setine istatistiksel yöntemlerin uygulanması

Veri Seti 431306 satır ve 2 kolondan oluşan Kategori ve yorumlardan oluşmaktadır.

Bu çalışmada kullanılan veri seti, dijital platformlarda satılan ürünlere yapılan Türkçe yorumları içermektedir. Toplamda 431306 satır yorum ve 32 farklı kategoriye ayrılmıştır. Benzersiz yorum sayısı ise 427231'dir. Bu veri setinde aynı yorumların farklı kayıtlar altında tekrarlanması söz konusu olduğundan, bu tekrarlanan 4075 yorum veri setinden çıkarılmıştır.

category		text
count	431306	431306
unique	32	427231
top	kamu-hizmetleri	Akman Nikah Ve Balo Salonu Sözleşme Bedeli Kapora İadesi Yapmıyor,20.07.2019 tarihinde 13.06.2020 tarihli kına gecesi için Sincan harikalar diyarı akman prestij salonuna kapora bedeli 2600 TL ödeyerek anlaşma yaptık. Korona virüsü nedeniyle iptal etmek için mayıs ayından itibaren defalarca görüşüm ve sonra dilekçe verdim sözlü olarak kapora bedeli ödenmez ama dil...Devamını oku
freq	14009	154

Şekil 13: Veri seti data istatistiği

Kaynak : Yazar tarafından oluşturulmuştur

Veri setinde en fazla yoruma sahip kategori "kamu hizmetleri"dir ve bu kategoriye ait 14.009 adet yorum bulunmaktadır.

Kategori ve kategoriye bağı yorum sayısı aşağıdaki şekilde verilmiştir.

kamu-hizmetleri	14009
finans	13982
cep-telefon-kategori	13981
enerji	13979
ulasim	13957
kargo-nakliyat	13918
medya	13910
mutfak-arac-gerec	13870
alisveris	13835
mekan-ve-eglence	13810
elektronik	13772
beyaz-esya	13763
kucuk-ev-aletleri	13733
internet	13694
giyim	13690
etkinlik-ve-organizasyon	13655
icecek	13564
saglik	13560
sigortacilik	13488
spor	13449
mobilya-ev-tekstili	13434
otomotiv	13379
turizm	13318
egitim	13269
gida	13150
temizlik	13113
hizmet-sektoru	13031
mucevher-saat-gozluk	12964
bilgisayar	12964
kisisel-bakim-ve-kozmetik	12659
anne-bebek	12382
emlak-ve-insaat	12024

Şekil 14: Kategori bağı yorum sayısı

Kaynak : Yazar tarafından oluşturulmuştur

5.4 Verilerin ön hazırlık aşamalarının hazırlanması

Veri setindeki benzersiz olmayan yorumları kaldırmak için Pandas kütüphanesinde bulunan "duplicated" ve "drop_duplicates" fonksiyonları kullanıldı. "duplicated" fonksiyonu bir DataFrame veya Series nesnesi olarak, True-False ifadeleriyle bir Series nesnesi döndürür. "drop_duplicates" fonksiyonu ise,

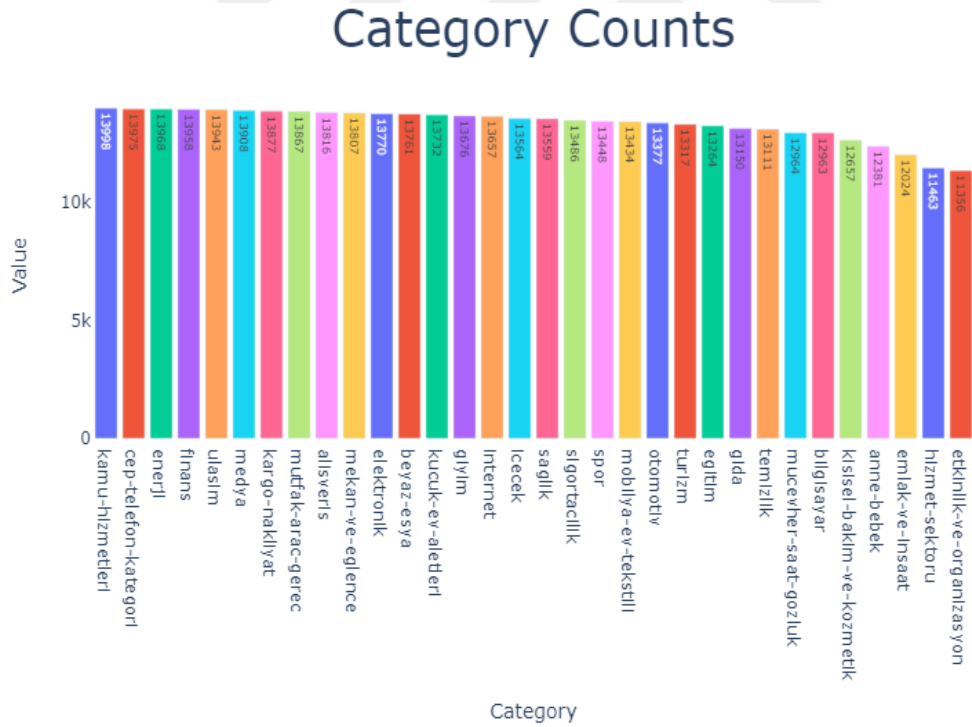
inplace=True veya False parametreleriyle ve subset= {kolon etiketi} parametresiyle çalışır. Varsayılan olarak ignore_index=False'dur. Bu, bir kopyayı kaldırır, dizinin değişmeyeceği anlamına gelir. ignore_index parametresinin True olması durumunda, dizi indexleri yeniden düzenlenmiştir. Varsayılan olarak "ilk" tutulmuştur.

	category	text
count	427231	427231
unique	32	427231
top	kamu-hizmetleri	Altus Hırdavat Yapı Malzemeleri Drone Diye Kargodan Lastik Ayakkabı Çıktı,"Instagram'da dolanırken sponsorlu bir bağlantı gördüm. Drone satışı yapıyor. Normalde böyle şeylere inanmam ancak takipçi sayısının fazla olması, numaralarının olması, ödemeyi peşin değil karşı ödemeli ödenmesi, fotoğraflara yapılan yorumlar vs... Az da olsa güvenernek ben de sipariş vermek istedim...Devamını oku"
freq	13998	1

Şekil 15: Benzersiz kayıt istatistiği

Kaynak : Yazar tarafından oluşturulmuştur

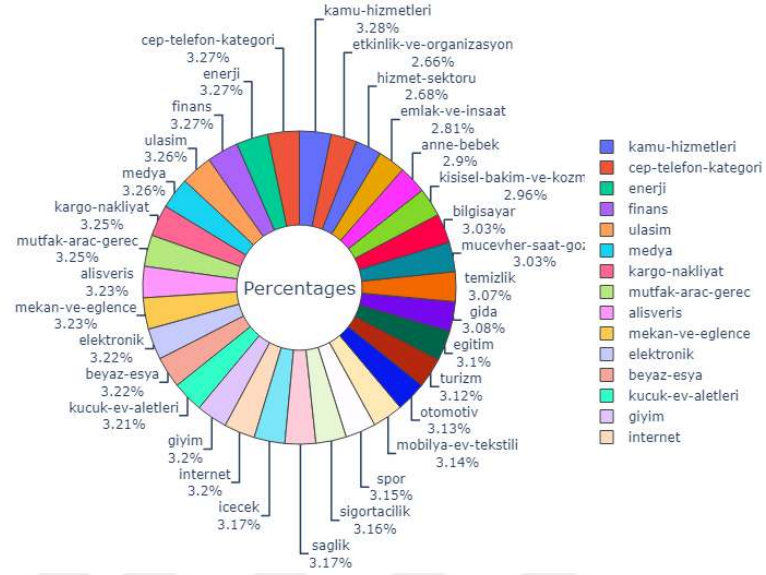
Veri setinden tekrarlayan kayıtların çıkarılması sonucunda, benzersiz kayıt sayısı 427231'e düşmüştür. Kategorilerin sayısı değişmedi ve hala 32 olarak kalmıştır. En çok yorum içeren kategori "kamu hizmetleri" olup, bu kategoride 13998 adet yorum bulunmaktadır.



Şekil 16: Kategori Yorum Sayısı Grafiği

Kaynak : Yazar tarafından oluşturulmuştur

Veri setindeki örnek sayısı her konuda yaklaşık olarak 13 bin olarak dağılmıştır, bu nedenle veri setinin dengeli olduğunu söylenebilir.



Şekil 17: Kategori Yorum yüzdelik Dağılımı

Kaynak : Yazar tarafından oluşturulmuştur

Veri setinde girilmiş olan metni incelediğimizde öncelikle kaçır kelimedenden oluştuğunu inceleyeceğiz. Aşağıdaki şekilde “words” kolonunda her bir yorumda kaç kelime olduğu bilgisi tutulmaktadır.

words	text
0 49	Altus Hırdavat Yapı Malzemeleri Drone Diye Kargodan Lastik Ayakkabı Çıktı."Instagram'da dolanırken sponsorlu bir bağlantı gördüm. Drone satışı yapıyor. Normalde böyle şeylere inanmam ancak takipçi sayısının fazla olması, numaralarının olması, ödemeyi peşin değil karşı ödeme ödenmesi, fotoğraflara yapılan yorumlar vs... Az da olsa güvenerek ben de sipariş vermek istedim...Devamını oku"
1 44	Albay Bilgisayar Garanti Yalanı İle Yanılttı.Garanti kapsamında yer alan Casper bilgisayarın garanti belgesi ile birlikte izmit Casper bilgisayar yetkili servisine albay bilgisayara bıraktım. Önce almak istemedi uzun ikna cabası ve uğraş sonunda zorla garanti dahiline bıraktım bilgisayar açılmıyordu. Sonrasında ertesi gün bilgisayarın yapıldı...Devamını oku
2 47	Okyanus Petshop Ürün İade Sorunu).Satıcıdan Gittigidiyor üzerinden 385 TL tutarında Frolicat Interactive Kedi Lazer oyuncakı satın aldım. Ürünü açıp bir kez çalıştırdığımda çok gürültü çıkardığını fark ettim ve ürünü iade etmeye karar verdim. Hatam iade sürecinin 14 gün içinde olması gerektiğini bilmemektir. Ürünü teslim aldıktan son...Devamını oku
3 53	Anda İhtiyaç Ve Tüketim Maddeleri Knee Tork Dizlik Esnemiyor! Anda ihtiyaç ve tüketim maddeleri pazarlama san. Lim. Şti. Den 139 TL kapıda teslim knee tork adlı ürünü satın aldım. Ürün 0815774588452 takip no... ile Aras Kargo tarafından tarafıma teslim edildi. Ürün kargoya firma tarafından 26.05.2020 tarihinde kargoya verilmiş. Ürünün üst dizlik bedeni bacağım...Devamını oku
4 48	BB Kargo Dağıtım'dan Aldığım Ürün Bozuk Çıktı Ve Muhafaz Bulamıyorum! Fotoğrafını göndermiş olduğum ürünün kablosunda temassızlık var ve ne fatura ne numara sadece gönderen kargonun adı var ve onlara da ulaşamıyorum ayrıca kargo daha bugün elime geçti Denizli Marmaris gezmiş ürün. Beni arayan numarayı buldum ona da ulaşamıyorum.Devamını oku

Şekil 18: Yorum Kelime Sayısı

Kaynak : Yazar tarafından oluşturulmuştur

Veri setindeki yorumların kelime istatistiği incelendiğinde, her bir yorumun ortalama 44 kelime içerdiği, en az 2 kelime, en fazla 183 kelime kullanıldığı ve yorumların %75'inin 49 kelimeden oluştuğu görülmektedir.

count	427231.000000
mean	44.408624
std	8.108499
min	2.000000
25%	42.000000
50%	46.000000
75%	49.000000
max	183.000000
Name: words, dtype: float64	

Şekil 19: Kelime İstatistiği

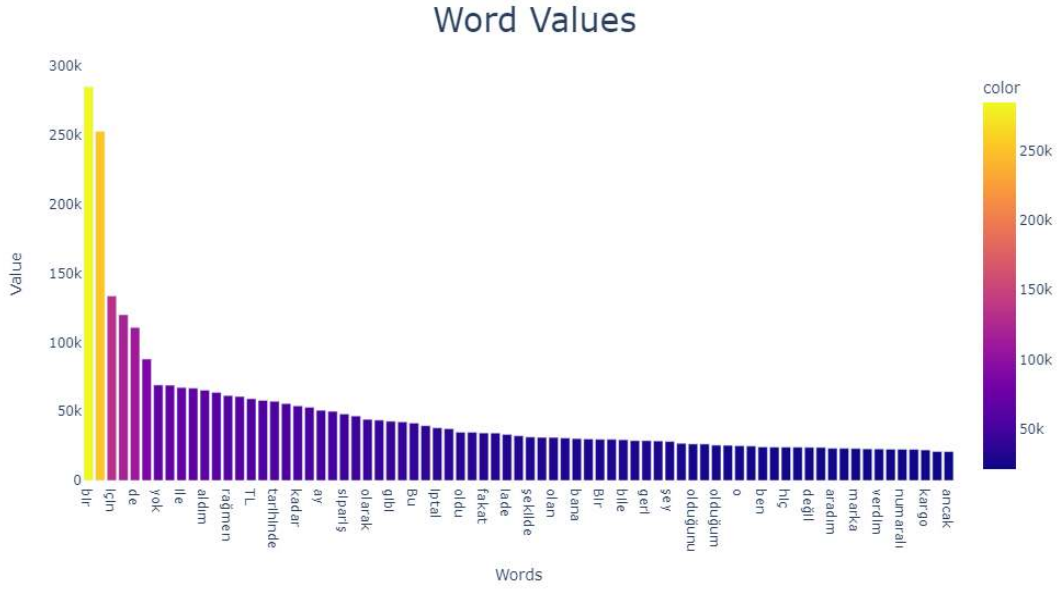
Kaynak : Yazar tarafından oluşturulmuştur

Most Used 20 Words	
oku :	382925
bir :	290203
ve :	254437
için :	134568
bu :	123079
de :	111211
da :	88801
ama :	70306
yok :	67821
ile :	67620
gün :	67342
aldım :	65740
sonra :	64183
ne :	62723
rağmen :	61891
TL :	59720
önce :	58567
tarihinde :	57490
çok :	56586
kadar :	55290

Şekil 20: Yorumlarda en çok geçen 20 kelime

Kaynak : Yazar tarafından oluşturulmuştur

Yorumları incelediğimizde, sıklıkla kullanılan "oku" kelimesinin, yorumların kategorize edilmesinde bir etkisinin olmadığı tespit edilmiştir. "Devamını oku" ifadesi nedeniyle yorumların sonunda sıklıkla kullanıldığı görülmüştür. Dolayısıyla bu kelime kategori seçimine etki etmeyeceği için tüm yorumlardan silinmiştir. Silme işleminden sonra yorumlarda geçen en çok kelimeler aşağıdaki şekilde gösterilmiştir.



Şekil 21: En Çok Geçen Kelimler

Kaynak : Yazar tarafından oluşturulmuştur

5.5 Doğal Dil İşleme Süreci

Veri ön işlemede sırasıyla; noktalama işaretlerinin temizlenmesi, büyük-küçük harf dönüşümü, kelimelere ayırma, durak kelimelerin çıkarılması işlemleri yapılmaktadır (Emeç & Özcanhan, 2023, s. 33-54). Bu işlemler için Python’da kullanılan kütüphaneler ve fonksiyonlar Tablo 2’de verilmiştir.

Tablo 2: Veri Ön İşleme Kütüphane ve Fonksiyonları

İşlem	Kütüphane	Fonksiyon
Noktalama İşaretlerinin Silinmesi	string	<i>punctuation()</i>
Büyük-Küçük Harf Dönüşümü	string	<i>lower()</i>
Kelimelere Ayırma	NLTK	<i>word_tokenize()</i>
Durak Kelimelerin Çıkarılması	NLTK	<i>stopwords()</i>

Bu işlemler yapıldıktan sonra aşağıdaki örnek yorumun öncesi ve sonrası gösterilmektedir. Şekilde görüldüğü üzere noktalama işaretleri temizlenmiş büyük harfler küçük harflere dönüştürülmüş rakam ve kategorize etme işlemine etkisi olmayacak kelimeler çıkarılmıştır.

Öncesi
Okyanus Petshop Ürün İade Sorunu!,Saticıdan GittiGidiyor üzerinden 385 TL tutarında Frolicat Interactive Kedi Lazer oyuncacı satın aldım Ürünü açıp bir kez çalıştırdığımda çok gürültü çıkardığını fark ettim ve ürünü iade etmeye karar verdim Hatam iade sürecinin 14 gün içinde olması gerektiğini bilmemektir Ürünü teslim aldıktan sonra
Sonrası
okyanus petshop ürün iade sorunu satıcıdan gittigidiyor üzerinden tl tutarında frolicat interactive kedi lazer oyuncacı satın aldım ürünü açıp bir çalıştırdığımda gürültü çıkardığını fark ettim ürünü iade etmeye karar verdim hatam iade sürecinin gün içinde olması gerektiğini bilmemektir ürünü teslim aldıktan sonra
Time Passed
54.9051194190979

Şekil 22: Veri ön işleme öncesi ve sonrası

Kaynak : Yazar tarafından oluşturulmuştur

5.6 Çoklu sınıf sınıflandırma yöntemi

Çoklu sınıf sınıflandırma (Multiclass Classification), makine öğrenimi ve veri madenciliği alanlarında kullanılan bir sınıflandırma yöntemidir. Bu yöntem, bir nesnenin birden fazla sınıf veya kategoriye ait olabileceği durumları ele alır. Özellikle sınıflandırılacak verilerin birden fazla sınıfa ait olabileceği durumlarda kullanılır.

Çoklu sınıf sınıflandırma, iki ana sınıflandırma türünden biridir. Diğer tür ise ikili sınıflandırmadır (binary classification). İkili sınıflandırmada, nesneler iki farklı sınıfa ayrılırken, çoklu sınıf sınıflandırmada nesneler birden fazla sınıfa veya kategoriye atanabilir.

Örneklerle açıklamak gerekirse:

E-posta Sınıflandırma: E-postaların spam (istenmeyen) veya ham (istenen) olarak sınıflandırılması bir ikili sınıflandırma örneğidir. Ancak e-postaların farklı kategorilere (örneğin iş, kişisel, reklam, haber) ayrılması çoklu sınıf sınıflandırma gerektirir.

Hastalık Tanısı: Bir hastanın birçok farklı hastalığa sahip olup olmadığını belirlemek için çoklu sınıf sınıflandırma kullanılabilir. Hastaların birçok farklı hastalığa sahip olma olasılığı vardır ve bu nedenle sonuç birden fazla kategoriye ait olabilir.

Yüz Tanıma: Bir görüntünün sadece bir kişiye ait olmaması durumunda (örneğin bir grup fotoğrafı) yüz tanıma çoklu sınıf sınıflandırma gerektirebilir. Çünkü görüntüde birden fazla kişi bulunabilir ve her kişi farklı bir sınıfa ait olabilir.

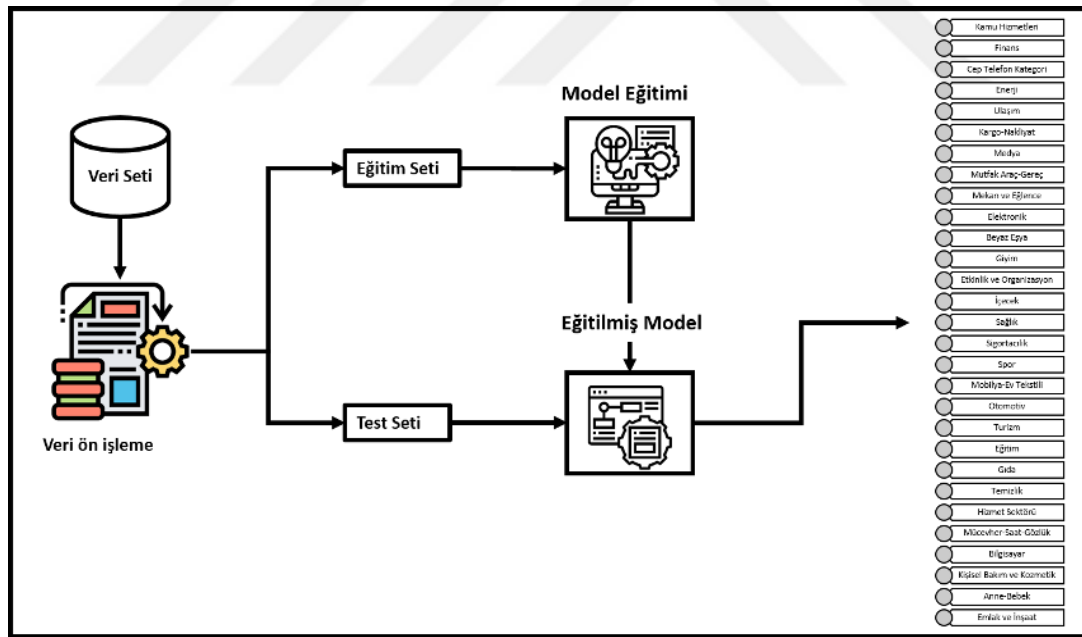
Çoklu sınıf sınıflandırma algoritmaları, verileri birden fazla sınıf arasında ayırmak için kullanılır. Bu algoritmaların başlıca özellikleri şunlardır:

Softmax Fonksiyonu: Çoklu sınıf sınıflandırmada sıklıkla kullanılan bir aktivasyon fonksiyonudur. Softmax, her sınıfın olasılığını hesaplar ve sonuçları normalize eder, böylece toplam olasılık 1'e eşittir.

Karar Sınırları: Çoklu sınıf sınıflandırma, karar sınırlarının nasıl çizileceği ve birden fazla sınıfın nasıl tahmin edileceği konularında karmaşık bir yapıya sahiptir. Bu nedenle, doğru karar sınırlarını çizmek için çeşitli algoritmalar ve yöntemler kullanılır.

Eğitim ve Değerlendirme: Çoklu sınıf sınıflandırma modelleri, genellikle çoklu sınıf doğruluk, hassasiyet, duyarlılık ve F1 puanı gibi metriklerle değerlendirilir.

Çoklu sınıf sınıflandırma, bir nesnenin birden fazla kategoriye ait olabileceği karmaşık sınıflandırma problemlerinde oldukça önemlidir ve geniş bir uygulama alanına sahiptir.



Şekil 23: Önerilen Yapay Zekâ Çoklu Sınıf Sınıflandırma Mimarisi

Kaynak : Yazar tarafından oluşturulmuştur

Şekil 23’de önerilen yapay zekâ uygulamasının ve modelinin tasarımını ve işleyişini açıklar. Bu açıklamada birçok bileşen ve kavram bulunmaktadır. Önerilen yöntem bileşenleri:

Veri Seti: Çalışma için kullanılan verilerdir. Bu veri seti, farklı kategorilere ait metin verilerini içerir ve her metin örneği bir veya daha fazla sınıfa ait olabilir. Veri seti, modelin eğitimi ve testi için kullanılır.

Eğitim Seti: Veri setinin bir bölümüdür ve modelin öğrenmesi için kullanılır. Eğitim seti üzerindeki veriler, modelin sınıflandırma görevini öğrenmesine yardımcı olacak şekilde kullanılır.

Test Seti: Veri setinin ayrılmış bir bölümüdür ve modelin performansını değerlendirmek için kullanılır. Model eğitildikten sonra test seti üzerindeki veriler kullanılarak modelin doğruluğu ve performansı ölçülür.

Model Eğitimi: Veri seti üzerindeki eğitim seti kullanılarak bir yapay sinir ağı veya makine öğrenimi modeli eğitilir. Bu model, metin verilerini analiz edip sınıflandırma yapabilen bir yapay zeka sistemidir.

Eğitilmiş Model Nesneleri: Modelin eğitimi tamamlandıktan sonra, eğitilmiş model nesneleri oluşturulur. Bu nesneler, modelin eğitilmiş ağırlıklarını ve parametrelerini içerir. Eğitilmiş model, artık yeni metin verilerini sınıflandırabilir.

Çıktı Sınıfları: Sınıflandırma görevinde kullanılan kategorileri temsil eder. Bu durumda, 32 farklı sınıf bulunur ve her bir sınıf, metin verilerinin hangi kategoriye ait olduğunu gösterir. Örneğin, "Finans," "Sağlık," "Gıda" gibi sınıflar metinleri sınıflandırmak için kullanılır.

Şekil 23'deki yapı, metin madenciliği veya metin sınıflandırma uygulamaları için kullanılan bir temel mimariyi temsil eder. Veri setinin eğitim ve test olarak ayrılması, modelin öğrenme süreci ve sonuçların sınıflandırma kategorileriyle temsil edilmesi, bu tür projelerde yaygın olarak kullanılan adımlardır. Bu mimari, metin tabanlı verileri etkili bir şekilde sınıflandırmak ve kategorize etmek için kullanılabilir.

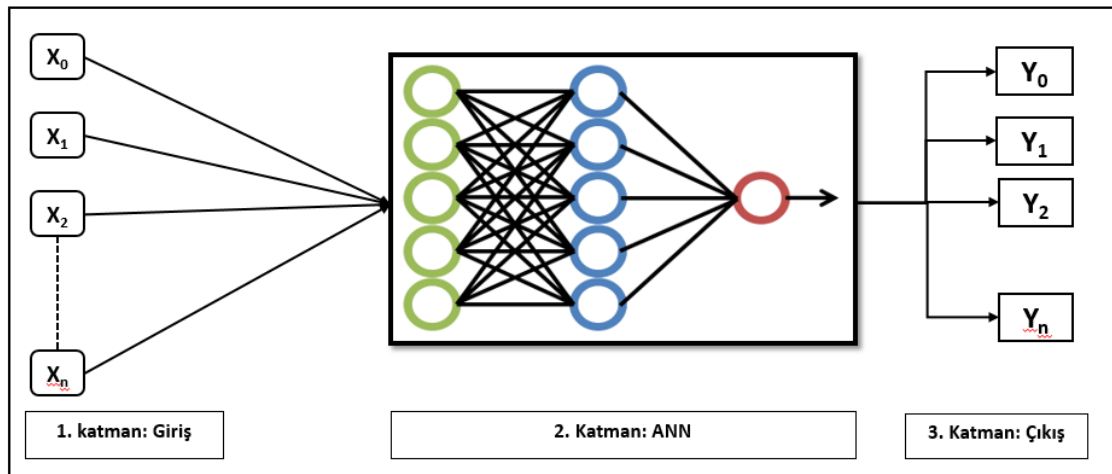
5.7 Sınıflandırma Yöntemlerine Uygun Yapay Sinir Ağları ve Makine Öğrenmesi Modellerin Tasarlanması

Çeşitli veri sınıflandırma problemlerini çözmek amacıyla özelleştirilmiş yapay sinir ağı ve makine öğrenmesi modellerinin tasarımını ele alır. Bu çalışma, veri analitiği, metin madenciliği, görüntü işleme, ses tanıma veya çeşitli veri sınıflandırma görevlerini içerebilir. Tasarlanan modeller, verilerin özelliklerini öğrenme, karmaşık

ilişkileri tanımlama ve belirli bir problemi çözme yeteneğine sahip olacak şekilde inşa edilir. Bu başlık altında yapılan araştırmalar, veri tabanlı problemlerin çözümünde gelişmiş performans ve sonuçlar elde etmek için çeşitli sınıflandırma yöntemleri ile yapay sinir ağı ve makine öğrenmesi modellerinin uygun bir şekilde birleştirilmesini vurgular. Bu çalışmalar, endüstri, sağlık, finans, doğal dil işleme ve daha birçok alanda kullanılabilen özelleştirilmiş ve etkili sınıflandırma modellerinin geliştirilmesine katkı sağlar.

5.7.1 ANN Modelinin Tasarlanması

Doğal dil işleme süreci sonucunda oluşan metin verilerini Sklearn kütüphanesindeki TfidfVectorizer sınıfı kullanarak içeri aktarılmıştır. Bu sınıfı, metin verilerini TF-IDF (Terim Sıklığı-Ters Belge Frekansı) matrisine dönüştürmek için kullanılmıştır. TfidfVectorizer sınıfının bir örneğini oluştururken, en fazla 20,001 kelime içerecek şekilde yapılandırılmıştır. TF-IDF matrisi oluştururken, veri setinizdeki kelimelerin önemini değerlendirmek için kullanılmıştır. Modeli eğitirken ve değerlendirirken, verilerinizi eğitim ve test setlerine ayrılmıştır. Bu, modeli eğitirken verilere ne kadar iyi uyduğunu ve genelleme yeteneğini değerlendirmenize yardımcı olacaktır. Bu, verilerin yüzde 20'sini test etmek ve modelinizin performansını değerlendirmek için kullanılmıştır. Yüzde 80'nini modeli eğitmek için kullanılmıştır.



Şekil 24: Önerilen Yapay Sinir Ağı Mimarisi

Kaynak : Yazar tarafından oluşturulmuştur

Şekil 24'deki önerilen ANN mimarisi olarak belirtilen bu mimari, bir Yapay Sinir Ağı (Artificial Neural Network - ANN) modelini tanımlar. Yapay Sinir Ağları, çeşitli makine öğrenimi ve yapay zeka uygulamalarında kullanılan bir model türüdür. İşte bu önerilen mimarinin temel bileşenlerini yorumlayan bir açıklama:

Giriş Katmanı (X_0, X_1): Bu katman, ANN modeline veri girişini temsil eder. X_0 ve X_1 , giriş verilerinin özelliklerini ve boyutlarını ifade eder. Örneğin, bir veri örneği, iki özellikten oluşuyorsa $X_0 = 2$ ve $X_1 = 1$ olabilir. Giriş katmanı, modelin veriyi almasını ve içsel işleme tabi tutmasını sağlar.

ANN Modeli (İkinci Katman): İkinci katman, ANN modelini temsil eder. ANN'lar, veriyi işlemek ve karmaşık ilişkileri öğrenmek için geliştirilmiş matematiksel modelledir. ANN'lar farklı katmanlardan (giriş katmanı, gizli katmanlar ve çıkış katmanı) oluşabilir. Gizli katmanlar, verilerin içsel özelliklerini çıkarmak ve karmaşıklığı işlemek için kullanılır. Her gizli katmanın altındaki düğümler (nöronlar), verileri işleyen ve ileten birimlerdir. Bu katmanların sayısı ve düğümlerin sayısı, modelin karmaşıklığını belirler.

Çıkış Katmanı (Y_0, Y_1): Bu katman, ANN modelinin çıkışını temsil eder. Y_0 ve Y_1 , çıkış verilerinin boyutunu ve formatını belirtir. Bu modelde çıkışlar iki farklı sınıfa temsil edebilir veya farklı sınıflar için olasılıkları içerebilir. Özellikle sınıflandırma problemlerinde, çıkış katmanı genellikle softmax aktivasyon fonksiyonu ile sonlandırılır ve her sınıf için bir olasılık dağılımı üretir.

Bu önerilen ANN mimarisi, genel olarak veri işleme ve sınıflandırma problemleri için kullanılan bir yapay zeka modelini temsil eder. Giriş katmanı verileri alır, ANN modeli bu verileri işler ve çıkış katmanı sonuçları üretir. Bu tür modeller, çeşitli uygulama alanlarında kullanılır ve veri özelliklerine ve problemin karmaşıklığına bağlı olarak farklı şekillerde özelleştirilebilir.

5.7.2 Tasarlanan ANN modellerin eğitim veri setine uygulanması

Keras Kütüphanesi kullanılarak Yapay Sinir Ağı (ANN) modeli oluşturulmuştur. 1024,512,216 Nörondan oluşan üç giriş ve bir çıkış katmanı oluşturulmuştur.

İlk katman (Dense katmanı): 1024 nörona sahiptir, ReLU (Rectified Linear Activation) aktivasyon fonksiyonu kullanır ve giriş boyutu, `x_train` verisinin sütun sayısıdır.

İkinci katman (Dense katmanı): 512 nörona sahiptir ve ReLU aktivasyon fonksiyonu kullanır.

Üçüncü katman (Dense katmanı): 256 nörona sahiptir ve ReLU aktivasyon fonksiyonu kullanır.

Çıkış katmanı (Dense katmanı): Sınıf sayısı kadar nörona sahiptir ve sınıflandırma problemlerinde yaygın olarak kullanılan softmax aktivasyon fonksiyonunu kullanır. Optimizasyon algoritması olarak Adam optimizer kullanılır. Öğrenme hızı (lr), beta değerleri gibi hiperparametreler belirtilmiştir.

Modelin eğitilmesi için gerekli parametreleri ve kayıp fonksiyonunu belirledim. Çoklu Sınıf sınıflandırma kullandığım için kayıp fonksiyonu olarak "categorical_crossentropy" kullanılmıştır.

Oluşturduğunuz yapay sinir ağı modelinin yapısını görselleştirerek, modelin hangi katmanlardan ve hangi boyutlardan oluştuğunu daha iyi anlamana yardımcı olmasını sağladım. Bu görselleştirme, modelinizin karmaşıklığını ve yapılandırmasını anlamak için çok faydalıdır ve modelinizi daha iyi ayarlamak veya optimize etmek için kullanılacaktır.

Eğitim verilerini 250 parçalık topluluklara böldük. Her topluluk, model güncellemelerinin hesaplandığı birimdir. Bu, eğitim işlemi hızlandırmak ve hafızayı daha iyi kullanmak için kullanılır. Eğitim verilerinin model tarafından 10 kez geçilmesini belirtilmiştir. Her bir geçiş, bir eğitim döngüsünü temsil eder. Eğitim verilerini aynı sıradaki verilerle alışmasını önlemek için her eğitim geçişi öncesinde verilerin karıştırılmasını sağlanmıştır.

Eğitim sırasında kayıp (loss) ve doğruluk (accuracy) gibi metriklerin her eğitim döngüsünde nasıl değiştiğini takip etmek edilmiştir. Model eğittikten sonra test verileri ile tahminlemeye başlanılmıştır.

```
Epoch 1/10
1336/1336 [=====] - 194s 144ms/step - loss: 2.6661 - accuracy: 0.5473
Epoch 2/10
1336/1336 [=====] - 190s 142ms/step - loss: 0.8217 - accuracy: 0.8499
Epoch 3/10
1336/1336 [=====] - 187s 139ms/step - loss: 0.4584 - accuracy: 0.8989
Epoch 4/10
1336/1336 [=====] - 196s 146ms/step - loss: 0.3444 - accuracy: 0.9187
Epoch 5/10
1336/1336 [=====] - 197s 147ms/step - loss: 0.2862 - accuracy: 0.9307
Epoch 6/10
1336/1336 [=====] - 195s 146ms/step - loss: 0.2486 - accuracy: 0.9387
Epoch 7/10
1336/1336 [=====] - 192s 144ms/step - loss: 0.2210 - accuracy: 0.9451
Epoch 8/10
1336/1336 [=====] - 193s 144ms/step - loss: 0.1993 - accuracy: 0.9501
Epoch 9/10
1336/1336 [=====] - 197s 147ms/step - loss: 0.1812 - accuracy: 0.9550
Epoch 10/10
1336/1336 [=====] - 201s 150ms/step - loss: 0.1657 - accuracy: 0.9590
2671/2671 [=====] - 126s 47ms/step
```

Şekil 25: Model Eğitim Geçmişi

Kaynak : Yazar tarafından oluşturulmuştur

5.7.3 Eğitilen ANN modelinin test veri seti ile değerlendirilmesi

Modelin doğruluğunu ve sınıflandırma performansını anlamak için Doğruluk, Hassasiyet, duyarlılık ve F1 puanı metrikleri kullanılmıştır. Doğruluk (accuracy) puanı hesaplanmıştır. Gerçek sınıflar ile model tarafından tahmin edilen sınıflar arasındaki doğru tahmin oranını ölçer. Hassasiyet (precision), duyarlılık (recall), ve F1 puanı gibi sınıflandırma performans metrikleri hesaplanmıştır.

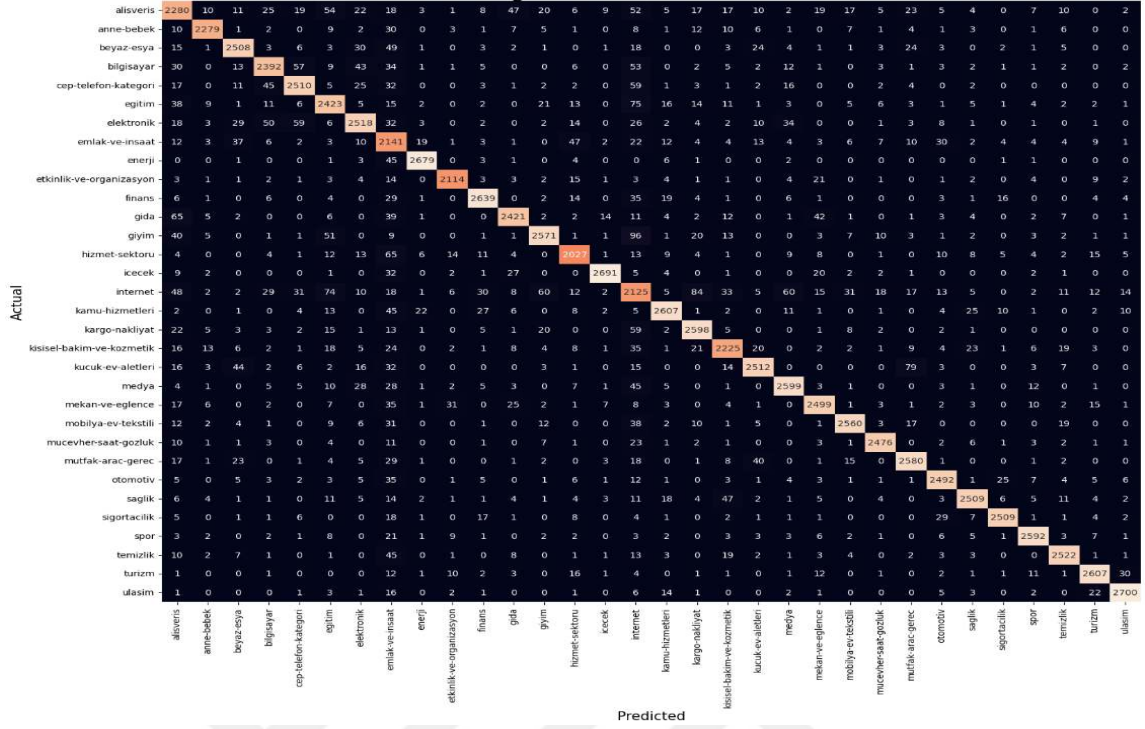
```
ANN Accuracy: 0.9234379205823493
ANN Precision: 0.9254049704064015
ANN Recall: 0.9234379205823493
ANN F1-score: 0.9240819426028434
```

Şekil 26: Model Performans Metrikleri

Kaynak : Yazar tarafından oluşturulmuştur

Gerçek ve tahmin edilen sınıflar arasındaki karşılaştırmaları gösteren Karmaşıklık matrisini hesaplanmıştır. Karmaşıklık matrisi ısı haritası kullanarak görselleştirilmiştir.

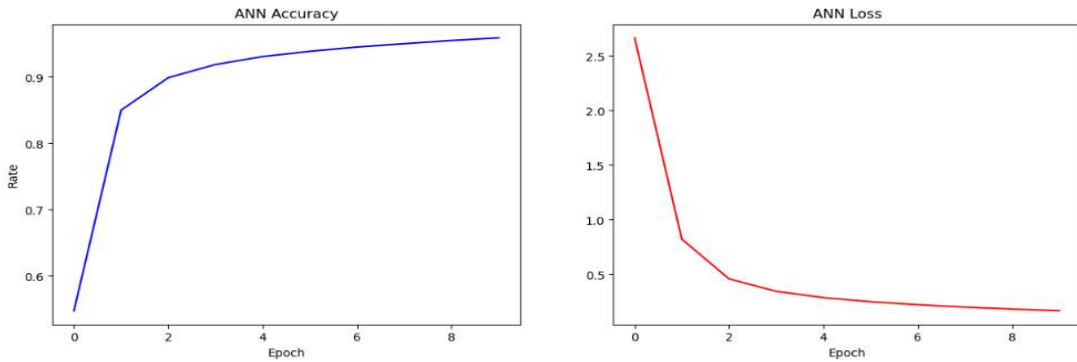
ANN Accuracy: 0.9234379205823493



Şekil 27: Karmaşıklık Matrisi Isı Haritası

Kaynak : Yazar tarafından oluşturulmuştur

ANN modelinin eğitim sürecindeki performansını görsel olarak analiz etmek için grafik oluşturulmuştur. Doğruluk (accuracy) grafiği, modelin her eğitim dönemi boyunca ne kadar iyi performans gösterdiğini gösterirken, kayıp (loss) grafiği, modelin ne kadar iyi öğrendiğini ve ne zaman yaklaşık bir dengeye ulaştığını gösterir. Bu grafikler, modelin eğitimini izlemek ve gerektiğinde ayarlamalar yapmak için önemlidir.



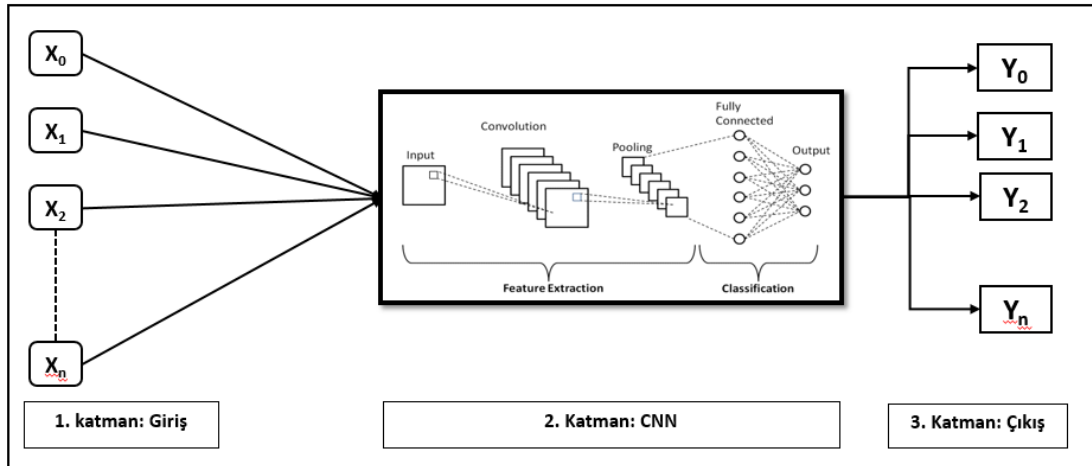
Şekil 28: ANN Modeli Performans Grafikleri

Kaynak : Yazar tarafından oluşturulmuştur

5.7.4 CNN Modelinin Tasarlanması

Keras kütüphanesinden metin işleme için gerekli modülleri içe aktarılmıştır. En fazla 20001 kelime kullanılacak şekilde ayarlanmıştır. Metin verilerini one-hot kodlama (one-hot encoding) ile temsil etmek için kullanılan bir liste oluşturulmuştur. Her metin, (one_hot) işlevi kullanılarak one-hot kodlamaya dönüştürülür ve sonuçlar onehot_corpus listesine eklenmiştir. One-hot kodlanmış metinleri, maksimum kelime sayısına göre dolgu yaparak aynı boyuta getirdin. Bu, metinlerin aynı boyutta olması gerektiği için gereklidir.

Metin verilerini CNN modeline uygun hale getirip ve eğitim ve test setlerine bölünmüştür. % 20 test ve % 80 eğitim verisi olacak şekilde ayarlanmıştır. CNN modelinin metin verilerini işleme ve öğrenmesi için gereken ön işlemi gerçekleştirilmiştir.



Şekil 29: Önerilen CNN Mimarisi

Kaynak : Yazar tarafından oluşturulmuştur

Şekil 29 daki önerilen CNN mimarisi olarak belirtilen bu mimari, bir Evrişimli Sinir Ağı (Convolutional Neural Network - CNN) modelini tanımlar. CNN'ler özellikle görüntü işleme ve metin sınıflandırma gibi alanlarda başarılı olan derin öğrenme modelleridir. İşte bu önerilen mimarinin temel bileşenlerini yorumlayan bir açıklama:

Giriş Katmanı (X_0, X_1): Bu katman, CNN modeline veri girişini temsil eder. X_0 ve X_1 , bu giriş verilerinin özelliklerini ve boyutlarını ifade eder. CNN modelleri genellikle görüntü işleme için kullanıldığından, X_0 ve X_1 , bir görüntünün genişliği ve

yüksekliği gibi özellikleri temsil edebilir. Örneğin, bir görüntünün boyutları 128x128 piksel ise, $X_0 = 128$ ve $X_1 = 128$ olabilir. Ancak bu, sadece bir örnek veri boyutu olarak düşünülebilir ve gerçek uygulamalarda daha büyük veya farklı boyutlar da kullanılabilir.

CNN Modeli (İkinci Katman): İkinci katman, CNN modelini temsil eder. CNN, özellikle görüntü işleme için tasarlanmış bir derin öğrenme mimarisidir. CNN, özellikle görüntülerdeki özellikleri (kenarlar, köşeler, desenler vb.) algılamak için kullanılır. Bu katman, bir dizi evrişim katmanı (convolutional layer), havuzlama katmanı (pooling layer) ve tam bağlantılı katmanlardan (fully connected layers) oluşabilir. Evrişim katmanları, görüntülerin özelliklerini çıkarmak için çeşitli filtreleri kullanır. Havuzlama katmanları, özellik haritalarını küçültmek ve önemli bilgileri korumak için kullanılır. Tam bağlantılı katmanlar, sonuçları sınıflandırmak ve çıkışları üretmek için kullanılır. CNN'nin katmanları, verinin özelliklerini öğrenme yeteneğine sahiptir ve bu nedenle model, giriş verilerini daha iyi temsil edebilir.

Çıkış Katmanı (Y_0, Y_1): Bu katman, CNN modelinin çıkışını temsil eder. Y_0 ve Y_1 , çıkış verilerinin boyutunu ve formatını belirtir. Bu mimariye göre, çıkışlar iki farklı sınıfı temsil edebilir veya farklı sınıflar için olasılıkları içerebilir. Özellikle çoklu sınıf sınıflandırma problemlerinde, çıkış katmanı genellikle softmax aktivasyon fonksiyonu ile sonlandırılır ve her sınıf için bir olasılık dağılımı üretir.

Özetle, bu önerilen CNN mimarisi, görüntü işleme veya başka bir özellik tabanlı problemin çözümü için kullanılabilen bir derin öğrenme modelini temsil eder. Giriş katmanı verileri alır, CNN modeli bu verileri işler ve çıkış katmanı sonuçları üretir. Bu tür modeller, özellikle görüntü sınıflandırma, nesne tanıma ve benzeri görevlerde kullanılır.

5.7.5 Tasarlanan CNN modelinin eğitim veri setine uygulanması

Gerekli Keras katmanları içe aktarılmıştır. Bu katmanlar, CNN modelini oluşturmak için kullanılacaktır.

İlk katman, gömme (embedding) katmanıdır. Bu katman, metin verilerini önce gömme vektörlerine dönüştürür. MAX_FEATURES belirtilen en fazla kelime

sayısını, 100 boyutlu gömme vektörlerini ve maxTextLen ile belirtilen maksimum kelime sayısını kullanır.

Evrişimli katman (Conv1D) eklenir. Bu katman, metin verilerinin özelliklerini çıkarmak için bir filtre kullanır. 64 filtre ile çalışır, filtre boyutu 2'dir, 'same' dolgu yöntemi kullanılır ve ReLU aktivasyon fonksiyonu kullanılır.

Maksimum havuzlama katmanı (MaxPooling1D) eklenir. Bu katman, evrişimli katmanın çıkardığı özellikleri özetler.

Düzleştirme katmanı eklenir. Bu, CNN katmanlarının çıkışlarını tek boyutlu hale getirir, böylece geleneksel yoğun katmanlara beslenebilir.

Yoğun (fully connected) katmanlar eklenir. Her ikisi de ReLU aktivasyon fonksiyonu ile tanımlanmıştır.

Çıkış katmanı eklenir. Bu, çoklu sınıf sınıflandırma için softmax aktivasyon fonksiyonunu kullanır ve sınıf sayısı kadar nörona sahiptir.

Optimizasyon algoritması olarak Adam optimizier kullanılır. Öğrenme hızı (lr) ve diğer hiperparametreler belirtilir.

Model, optimizier, metrikler ve kayıp fonksiyonu ile derlenir. Bu, modelin eğitim öncesi yapılandırılmasını tamamlar.

Oluşturduğunuz CNN modelinin yapısını görselleştirdim. Bu görselleştirme, modelin katmanlarını ve veri akışını daha iyi anlamana yardımcı olur ve modelin nasıl çalıştığını daha iyi anlamana katkıda bulunmuştur.

CNN modelini eğitip ve ardından test verileri üzerinde tahminler yaptım. Tüm eğitim verilerinin model tarafından 10 kez geçilmesini belirledik. Her bir geçiş, bir eğitim döngüsünü temsil eder. Eğitim verilerini küçük topluluklara (batch) böler. Her topluluk, model güncellemelerinin hesaplandığı birimdir. Bu, eğitim işlemini hızlandırmak ve hafızayı daha iyi kullanmak için kullanılır. Eğitim verilerinin her döngü öncesi karıştırılmasını sağlar. Bu, modelin her döngüde aynı sıradaki verilere alışmasını önler ve genellemeyi teşvik eder.

Eğitim sırasında kayıp (loss) ve doğruluk (accuracy) gibi metriklerin döngü bazında nasıl değiştiği takip edilmiştir.

Modelin eğitimini tamamlayıp test verileri üzerinde tahminler yapılmıştır. Tahmin sonuçları, modelin test verileri üzerindeki performansını değerlendirmenize ve modelin doğruluğunu ölçmenize yardımcı olur.

```
Epoch 1/10
268/268 [=====] - 180s 669ms/step - loss: 3.2603 - accuracy: 0.1731
Epoch 2/10
268/268 [=====] - 178s 664ms/step - loss: 1.6776 - accuracy: 0.6002
Epoch 3/10
268/268 [=====] - 177s 661ms/step - loss: 0.8351 - accuracy: 0.7978
Epoch 4/10
268/268 [=====] - 178s 662ms/step - loss: 0.5411 - accuracy: 0.8713
Epoch 5/10
268/268 [=====] - 190s 707ms/step - loss: 0.4049 - accuracy: 0.9035
Epoch 6/10
268/268 [=====] - 185s 689ms/step - loss: 0.3302 - accuracy: 0.9209
Epoch 7/10
268/268 [=====] - 177s 661ms/step - loss: 0.2809 - accuracy: 0.9324
Epoch 8/10
268/268 [=====] - 179s 668ms/step - loss: 0.2454 - accuracy: 0.9408
Epoch 9/10
268/268 [=====] - 178s 663ms/step - loss: 0.2184 - accuracy: 0.9471
Epoch 10/10
268/268 [=====] - 183s 684ms/step - loss: 0.1964 - accuracy: 0.9527
2671/2671 [=====] - 331s 124ms/step
```

Şekil 30: Model Eğitim Geçmişi

Kaynak : Yazar tarafından oluşturulmuştur

5.7.6 Eğitilen CNN Modellerinin Test Veri Seti ile Değerlendirilmesi

Modelin doğruluğunu ve sınıflandırma performansını anlamak için Doğruluk, Hassasiyet, duyarlılık ve F1 puanı metrikleri kullanılmıştır. Doğruluk (accuracy) puanı hesaplanmıştır. Gerçek sınıflar ile model tarafından tahmin edilen sınıflar arasındaki doğru tahmin oranını ölçer. Hassasiyet (precision), duyarlılık (recall), ve F1 puanı gibi sınıflandırma performans metrikleri hesaplanmıştır.

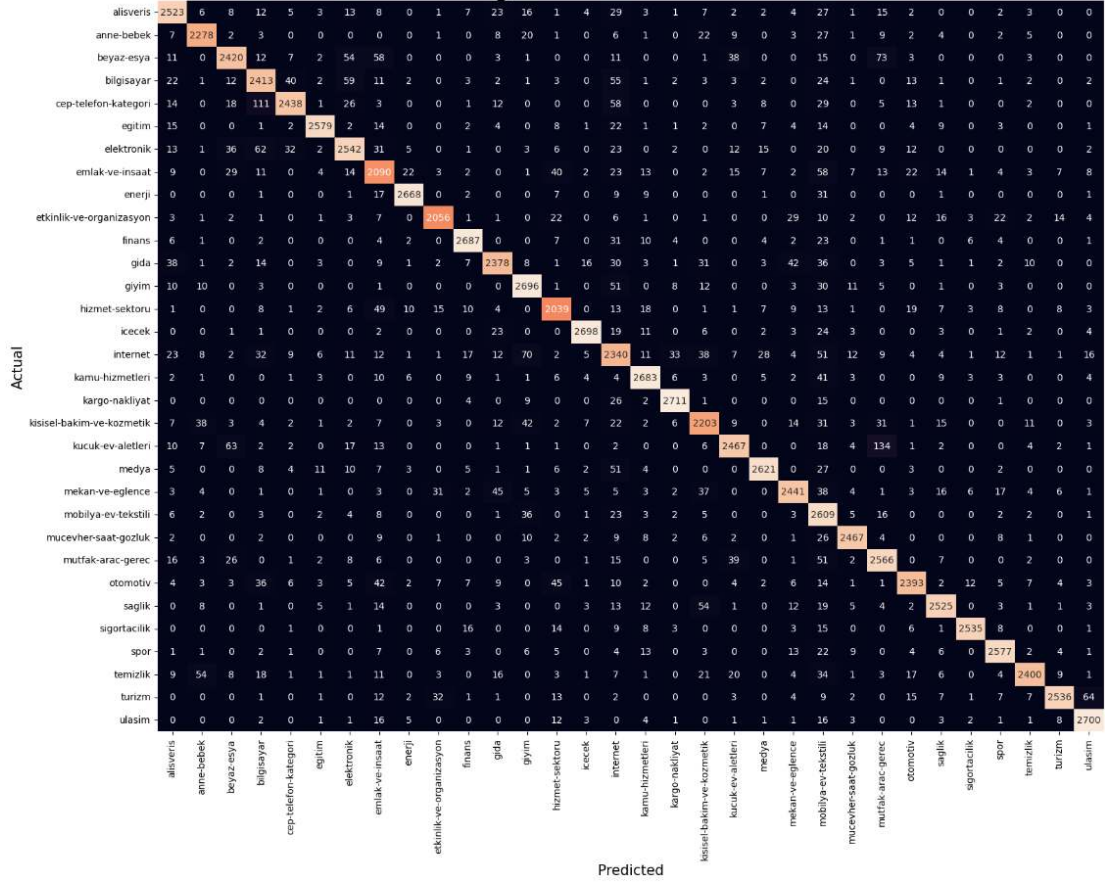
```
CNN Accuracy: 0.9278149028052477
CNN Precision: 0.9298980331113499
CNN Recall: 0.9278149028052477
CNN F1-score: 0.928308238353192
```

Şekil 31: Model Performans Metrikleri

Kaynak : Yazar tarafından oluşturulmuştur

Gerçek ve tahmin edilen sınıflar arasındaki karşılaştırmaları gösteren Karmaşıklık matrisi hesaplanmıştır. Karmaşıklık matrisi ısı haritası kullanarak görselleştirilmiştir.

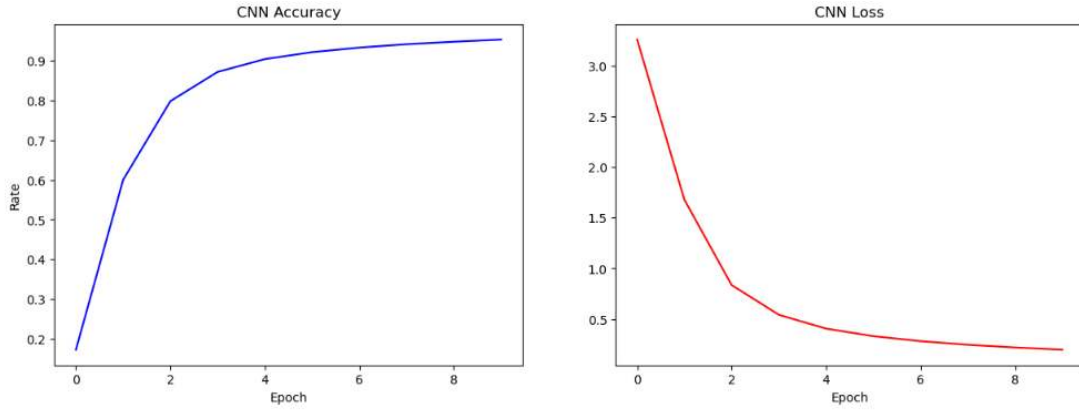
CNN Accuracy: 0.9278149028052477



Şekil 32: Karmaşıklık Matrisi Isı Haritası

Kaynak : Yazar tarafından oluşturulmuştur

CNN modelinin eğitim sürecindeki performansını görsel olarak analiz etmek için grafik oluşturulmuştur. Doğruluk (accuracy) grafiği, modelin her eğitim dönemi boyunca ne kadar iyi performans gösterdiğini gösterirken, kayıp (loss) grafiği, modelin ne kadar iyi öğrendiğini ve ne zaman yaklaşık bir dengeye ulaştığını gösterir. Bu grafikler, modelin eğitimini izlemek ve gerektiğinde ayarlamalar yapmak için önemlidir.



Şekil 33: CNN Modeli Performans Grafikleri

Kaynak : Yazar tarafından oluşturulmuştur

5.8 En iyi performans sonuçları için modellerin karşılaştırılması

Rastgele seçilen 10 metin örneği için gerçek kategorileri ile yapay sinir ağı (ANN) ve evrişimli sinir ağı (CNN) modelleri tarafından tahmin edilen kategorileri karşılaştırılmıştır.

Verilen metin örneği için ANN modelini kullanarak tahmin yapar. Önce metindeki gereksiz kelimeleri kaldırır, ardından metni TF-IDF vektörüne dönüştürür ve son olarak ANN modeli ile tahmin yapar.

Verilen metin örneği için CNN modelini kullanarak tahmin yapar. Önce metindeki gereksiz kelimeleri kaldırır, metni one-hot kodlamaya dönüştürür, ardından CNN modeli ile tahmin yapar.

Modellerin rastgele metin örnekleri üzerindeki performansını kontrol etmek için kullanılır. Her bir metin örneği için gerçek ve tahmini kategorileri görsel olarak görüntüler. Bu, modellerin ne kadar iyi çalıştığını anlamak için kullanışlıdır ve model tahminlerinin gerçek verilere ne kadar yakın olduğunu değerlendirmenize yardımcı olur. Tablo3, Tablo4 ve Tablo 5’de performans sonuçları sunulmaktadır.

Tablo 3: Model Sonuçlarının Karşılaştırılması-1

Text	Gerçek Kategori	ANN Kategorisi Tahmini	CNN Kategorisi Tahmini
Cinemaximum Kötü Salon ve Projesiyon!,"02.09.2017 tarihinde Aqua Florya şubenizde Hitman's Bodyguard filmi izledik (izlemek denirse).	etkinlik-ve-organizasyon	etkinlik-ve-organizasyon	etkinlik-ve-organizasyon
Kahta Petrol Turizm 1 Saat Geç Geldi!,Bir firma bu kadar kötü olabilir 1 saat geç geldi ve Kırşehir de 22:45 de olan otobüsüne yetişirsin diyerek güvendirip bindirdiler Mucurda 20 dakika mola verdiler ve başka sefer yok Yarın nöbetime yetişememe durumum var Güvendirip bindirdiler bir daha asla bu firmadan bilet almam Böyle bir mağd	ulasim	ulasim	ulasim
Electrolux El Süpürgesi Fiyaskosu,"Şarjlı el süpürgesinin yılda 1 kere şarjı bitiyor, 300 TL'ye yeni şarj takıyorlar Etrafımdaki herkes bunun delilik olduğunu söylese de bu sene de 300 TL verip şarjı değiştirdim Aradan 1 hafta geçti alet yine çalışmıyor Yarın servise götüreceğim ama biliyorum yine para isteyecekler Gidin Arçelik	kucuk-ev-alemleri	kucuk-ev-alemleri	kucuk-ev-alemleri

Tablo 4: Model Sonuçlarının Karşılaştırılması-2

Text	Gerçek Kategori	ANN Kategori Tahmini	CNN Kategori Tahmini
Damla Su Kötü Servis, Bugün Bisu apli kasyonu vasıtasıyla ön ödemeli olarak Damla Su'dan eve su söyledim İçinde bulunduğumuz korona günlerinde suyu getiren kişi kapının önüne atıp kaçmış Her kesin ayakkabısıyla bastığı yere suyu bırakmış Biz hijyene bu kadar dikkat ediyoruz üstelik Bunun üzerine eşim bayiye bu durumu a	icecek	icecek	icecek
VFS Global Tam Bir Fiyasko!"Geçtiğimiz hafta, son 5 yıl içerisinde aldığım 8 vize sonrasında (son ikisi İsveç ve Almanya'dan birer senelik olan) Schengen Vizesi başvurusu için VFS Global Yunanistan Ankara Vize Başvuru merkezinden randevu aldım Öncelikle, randevu almak tam anlamı ile bir işkenceydi zira her gün sistemleri hata	hizmet-sektoru	hizmet-sektoru	hizmet-sektoru
İçim Kaşar Peyniri Benzin Tadı Var,17 122019 tarihinde Kadıköy A101 'den kaşar peyniri aldım Sabah çocuğuma tost yaparken tadına bakayım dedim iyi ki bakmışım tarihi mi geçti diye baktım tarihi geçmemiş kokladım bozuk da kokmuyordu ama tadında benzin gibi kolonya gibi tarif edemediğim bir tat var İyi ki tadına bakmışım çocuk yese	gıda	gıda	gıda

Tablo 5 : Model Sonuçlarının Karşılaştırılması-3

Text	Gerçek Kategori	ANN Kategori Tahmini	CNN Kategori Tahmini
Trendyol 18 Gündür Teslim Edilmeyen Ürünler,22 Mayıs da verdiğim siparişim teslimatı 2 ile 9 Haziran arasında olduğ u yazıyor ama bugün 9 Haziran siparişi m daha kargoya bile verilmemiş Bu nas ıl bir şeydir ya müşteri hizmetlerine söy lüyorum yoğunluk deyip duruyorlar bir sipariş 18 gündür kargoya verilmiyorsa pes yani	internet	internet	internet
Hades Group Parayı Geri Vermiyor!"15 0 TL kura giriş ücreti diyorlar 80 TL'de kendi anlaştıkları darülaceze'de de bulu nan özel Sadabad Hastanesinden sağlık raporu istiyor 3 aydır kayıtlıyım Sözde i şe yönlendiriyorlar Ama verdikleri num araları arayınca çoğu eleman ihtiyacımı z yok diyor, bazılarında Hades Group'la ilişkimizi kestik	hizmet-se ktoru	hizmet-sektor u	hizmet-sekt oru
Koçtaş Ürün Rengi Farklı Geldi!,"20 gü ndür ürün gönderilmesi için çaba harca dım bir şey demiyorum Duvar kağıdı ik i ayrı renk, bu kadar olmaz artık Usta fa rk etti yaparken Yaptığınız hizmetten ut anın Nasıl firmasınız gerçekten ayıp	alisveris	alisveris	alisveris
Enerjisa Toroslar Elektrik Perakende Pe r Abn K/B Bedeli Yansıtma Sorunu, Fa turama yansıttığınız yasal olmayan per Abn K/B bedelini faturamdan düşürme nizi istiyorum Zaten depozito bedeli öd üyoruz bu neyin ödemesidir anlamıyoru m Yansıttığınız bedeli faturamdan düşü rün Daha önce farklı illerde ikamet etti m bu ödemenin alındığını ilk defa sizde gördüm çözüm bulunmazsa Cİ	enerji	enerji	enerji

Belirli bir metin örneği için, metin örneğinin kendisi ve ANN ile CNN modelleri tarafından tahmin edilen kategoriler görüntülenir. Bu, modellerin belirli metinlere nasıl tepki verdiğini anlamak için kullanışlıdır. Tahmin sonuçları, metinlerin hangi kategorilere ait olduğunu gösterir.

```
Text
-----
Yemeğin içinden kıl çıktı, gitmenizi önermiyorum.
-----
1/1 [=====] - 0s 13ms/step
ANN Predicted Category: mekan-ve-eglenence
1/1 [=====] - 0s 22ms/step
CNN Predicted Category: gida
*****

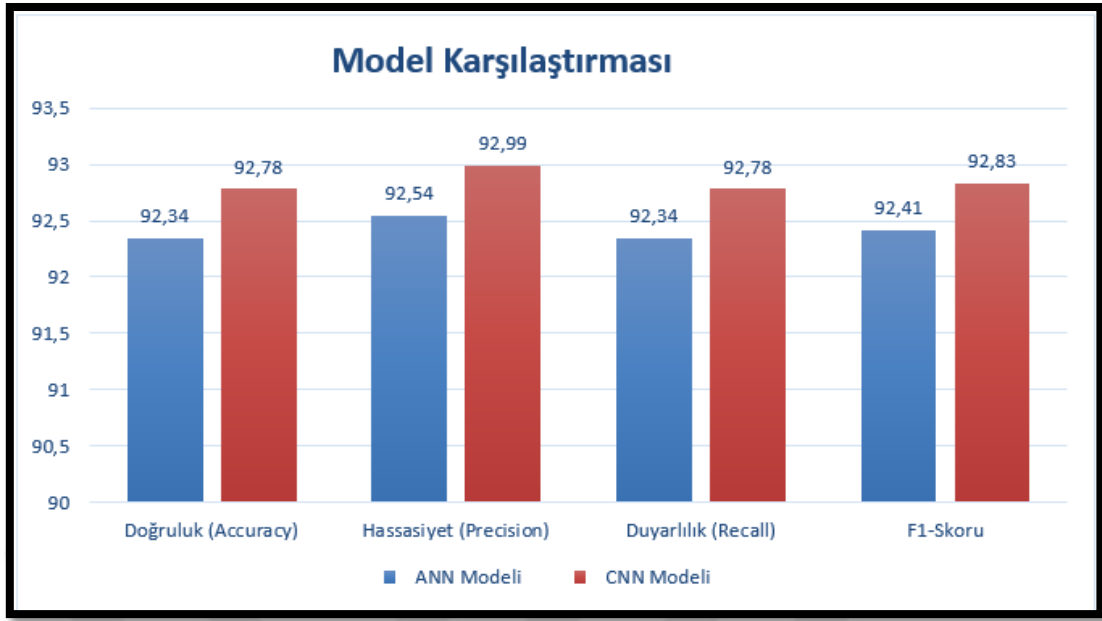
Text
-----
Tuş bozuk.
-----
1/1 [=====] - 0s 13ms/step
ANN Predicted Category: bilgisayar
1/1 [=====] - 0s 22ms/step
CNN Predicted Category: medya
*****
```

Şekil 34: Belirli Bir Metnin Kategori Tahmin Sonucu

Kaynak : Yazar tarafından oluşturulmuştur

5.9 En iyi performansı gösteren modelin karşılaştırılması

Bu sonuçlara göre, her iki model de oldukça yüksek performans göstermektedir. Hem yapay sinir ağı (ANN) hem de evrişimli sinir ağı (CNN) modelleri için doğruluk (accuracy), hassasiyet (precision), duyarlılık (recall) ve F1-score değerleri oldukça yakındır. Bu, her iki modelin de veri setini başarıyla sınıflandırdığını gösterir.



Şekil 35: Modellerin Karşılaştırılması

Kaynak : Yazar tarafından oluşturulmuştur

Her iki modelin de doğruluk değeri yaklaşık olarak aynıdır ve yaklaşık olarak %92.3 ile %92.8 arasında değişir. Aynı şekilde, hassasiyet, duyarlılık ve F1-score değerleri de oldukça benzerdir.

Bu sonuçlara dayanarak, her iki model de benzer bir performans sergiliyor gibi görünüyor. Ancak, her iki model arasındaki farklar oldukça küçük olduğu için, hangi modelin "en iyi" olduğunu kesin olarak belirlemek için başka faktörler de göz önüne alınmalıdır. Örneğin, uygulanacak olan kullanım senaryosu, hesaplama kaynakları, eğitim süresi ve modelin karmaşıklığı gibi faktörler de değerlendirilmelidir.

Sonuç olarak, her iki model de başarılıdır ve hangi modelin tercih edileceği, projenin gereksinimlerine ve kısıtlamalarına bağlı olacaktır.

SONUÇ

Bu tez çalışması, Türkçe müşteri yorumlarının metin madenciliği teknikleri ile sınıflandırılması üzerine odaklanmıştır. Çalışma, yapay sinir ağları (ANN) ve evrişimli sinir ağları (CNN) gibi makine öğrenimi modellerini kullanarak 32 farklı sınıfa ayrılan müşteri yorumlarını analiz etmeyi amaçlamıştır. Elde edilen sonuçlar, ANN modeli için %92.34 doğruluk ve %92.40 F1-skoru, CNN modeli için ise %92.78 doğruluk ve %92.83 F1-skoru olarak belirlenmiştir. Her iki model de yüksek performans sergilemiştir, ancak CNN modelinin biraz daha yüksek doğruluk ve F1-skoru elde ettiği gözlemlenmiştir. Bu sonuçlar, metin madenciliği tekniklerinin ve özellikle evrişimli sinir ağlarının, Türkçe müşteri yorumlarını başarılı bir şekilde sınıflandırmada etkili olduğunu göstermektedir. Bununla birlikte, ANN modelinin de başarılı bir performans sergilediği ve farklı metin madenciliği tekniklerinin uygun şekilde kullanılmasıyla elde edilen sonuçların önemli olduğu vurgulanmalıdır.

Bu çalışmanın bulguları, müşteri geri bildirimlerinin otomatik olarak sınıflandırılması ve analiz edilmesinde makine öğrenimi modellerinin kullanımının potansiyelini göstermektedir. Bu, pazarlama stratejilerinin geliştirilmesi, ürün geliştirme süreçlerinin yönlendirilmesi ve müşteri memnuniyetinin artırılması gibi alanlarda değerli bir araç olabilir. Ayrıca, bu çalışma gelecekteki araştırmalara yol göstermektedir. Özellikle, daha karmaşık yapıdaki modellerin veya farklı metin madenciliği tekniklerinin kullanılmasıyla sınıflandırma performansının nasıl etkilenebileceği ve daha geniş veri kümeleri üzerinde modellerin genelleme yeteneğinin nasıl olacağı gibi konularda derinlemesine analizler yapılabilir.

Sonuç olarak, bu tez çalışması, metin madenciliği tekniklerinin müşteri yorumlarının sınıflandırılması için güçlü bir araç olduğunu ortaya koymaktadır. Elde edilen yüksek performans, bu tekniklerin gerçek dünya uygulamalarında kullanılabilirliğini ve müşteri geri bildirimlerinin otomatik analizindeki değerini vurgulamaktadır.

KAYNAKÇA

- Adalı, E. (2012). Doğal Dil İşleme. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi.
- Akşehirli, Ö. Y., Ankaral, H., Aydın, D., & Saraçlı, Ö. (2013). Tıbbi Tahminde Alternatif Bir Yaklaşım: Destek Vektör Makineleri. *Turkiye Klinikleri Journal of Biostatistics*. 19-28.
- Al Amrani, Y., Lazaar, M., & El Kadiri, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*. 511-520.
- Alotaibi, F. M. (2019). Classifying text-based emotions using logistic regression. *Vawkum Transactions on Computer Sciences*. 31-37.
- Amanet, H. (2017). Türkçe sosyal medya metinlerinde duygu analizi. Yüksek Lisans Tezi, Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Ve Bilgisayar Bilimleri Anabilim Dalı, 95, Trabzon.
- Amasyali, M. F. (2005). Automatic construction of turkish wordnet. In *Proceedings of the IEEE 13th Signal Processing and Communications Applications Conference*, Kayseri.
- Arslan, H., Kaynar, O., & Yüksek, A. G. (2017). Kurumsal kolektif süreçler için e-posta iletilerinden görev keşfi ve gerçek zamanlı görev yönetim sisteminin geliştirilmesi. *Bilişim Teknolojileri Dergisi*. 381-388.
- Arslan, Y. (2011). "Veri Madenciliği". Çözümpark Web Sitesi.
- Başkaya, F., & Aydın, İ. (2017). Haber metinlerinin farklı metin madenciliği yöntemleriyle sınıflandırılması. *IEEE International Artificial Intelligence and Data Processing Symposium (IDAP)*. 1-5.
- Boser, B., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *5th Annual ACM Workshop on COLT*. Pennsylvania: ACM Press. 144–152.

- Chawan, R. K., Krishna, K. V., & Krishna, S. V. (2017). Text mining based on tax comments as big data analysis using XGBOOST and feature selection. *European Journal of Molecular & Clinical Medicine*.
- Cherian, V., & Bindu, M. S. (2017). Heart Disease Prediction Using Naive Bayes Algorithm and Laplace Smoothing Technique. *International Journal of Computer Science Trends and Technology (IJCST)*. 68-73.
- Chung, J., Lee, J., & Yoon, J. (2022). Understanding music streaming services via text mining of online customer reviews. *Electronic Commerce Research and Applications*.
- Dang, S., & Ahmad, P. H. (2014). Text mining: Techniques and its application. *International Journal of Engineering & Technology Innovations*. 22-25.
- Dehkharghani, R., Saygin, Y., Yanikoglu, B., & Oflazer, K. (2016). SentiTurkNet: a Turkish polarity lexicon for sentiment analysis. *Language Resources and Evaluation*. 667-685.
- Demir, Ö., Chawai, A. I., & Doğan, B. (2019). Türkçe metinlerde sözlük tabanlı yaklaşımla duygu analizi ve görselleştirme. *International Periodical of Recent Technologies in Applied Engineering*. 58-66.
- Demirel, Ş. (2019). Karar Ağacı Algoritmaları Ve Çocuk İşçiliği Üzerine Bir Uygulama. Yüksek Lisans Tezi. Marmara Üniversitesi Sosyal Bilimler Enstitüsü konometri Anabilim Dalı, 115, İstanbul.
- Demiriz, A. (2006). Veri Madenciliğine Giriş, Veri Madenciliği Ders Notları, <http://web.sakarya.edu.tr/~ademiriz/dms2006.htm>, 2006.
- Emeç, M., & Özcanhan, M. (2023). Veri Ön İşleme ve Öznitelik Mühendisliğinin Yapay Zekâ Yöntemlerine Uygulanması. *MÜHENDİSLİKTE ÖNCÜ VE ÇAĞDAŞ ÇALIŞMALAR*. 33-54.
- Ergün, K. (2012). Metin madenciliği yöntemleri ile ürün yorumlarının otomatik değerlendirilmesi. Doktora Tezi. Sakarya: Sakarya Üniversitesi, Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı.

- Ergüt, Ö. Ü. (2019). Otel yorumlarının metin madenciliği teknikleri ile incelenmesi. International Congress of Management Economy and Policy, İstanbul. 103-114.
- Erten, F. (2015). Metin Madenciliği Tabanlı Bir Web Sitesi Sınıflandırma Aracı Tasarımı.Yüksek Lisans Tezi. Maltepe Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı, İstanbul.
- Eryiğit, G. (2014). ITU Turkish NLP web service. In Proceedings of Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Angeles. 1-4.
- Fauziah, Y., Saifullah, S., & Aribowo, A. S. (2020). Design Text Mining for Anxiety Detection using Machine Learning based-on Social Media Data during COVID- 19 pandemic. In Proceeding of LPPM UPN “Veteran” Yogyakarta Conference Series– Engineering and Sc.
- Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. International Journal of Computer Applications.
- Goyal, A., & Mehta, R. (2012). Performance Comparison of Naïve Bayes and J 48 Classification Algorithms, International Journal of Applied Engineering Research.
- Gündüz, H. (2013). Borsa İstanbul (BIST) 100 endeksi yönünün ekonomi haberleri ile tahmin edilmesi. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı, 61, İstanbul.
- Hearst, M. (2003). What is text mining. [<https://www.jaist.ac.jp/~bao/MOT-Ishikawa/FurtherReadingNo1.pdf>]. California: University of California Berkeley, School of Information.
- Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. AAAI. 755-760.
- Huang, Y., Murphey, Y. L., & Ge, Y. (2015). Intelligent typo correction for text mining through machine learning. International Journal of Knowledge Engineering and Data Mining. 115-142.

- Işık, M. (2019). Müşteri yorumları üzerinde metin analitiği çalışmaları ve yorumların makine öğrenmesi algoritmaları ile modellenmesi. Yüksek Lisans Tezi. İstanbul: İstanbul Medeniyet Üniversitesi, Fen Bilimleri Enstitüsü Uygulamalı Matematik ve Hesaplama.
- Işıklı, Ş. (2014). Büyük Veri, Epistemoloji ve Etik Tartışmalar. AJIT-E: Online Academic Journal of Information Technology, 5(17), 89-122. Erişim adresi: <https://dergipark.org.tr/tr/pub/ajit-e/issue/54437/740935>.
- Kao, A., & Poteet, S. (2007). *Natural language processing and text mining*. London: Springer-Verlag Limited.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30(1).
- Kiraz, A., Eski, H., & Geçici, A. Ö. (2020). Analysis of online customer complaints of mobile telecommunication sector with text mining. *Turkish Online Journal of Educational Technology* (pp. 22-33).
- Kleinbaum, D., & Klein, M. (1994). *Logistic Regression: A self-learning text*. *Statistics for Biology and Health*, 3(1).
- Kohavi, R., & Provost, F. (1998). Machine Learn. Glossary of terms *journal of machine learning*, 30(1) , 271-275.
- Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). Thresholding classifiers to maximize F1 score. *Archive of Cornell University*, 1402, 1892.
- Lomotey, R. K., & Deters, R. (2014). Analytics-as-a-service framework for terms association mining in unstructured data. *International Journal of Business Process Integration and Management (IJBPIIM)*, 7(1), 49-61.
- Lucini, F. R., Tonetto, L. M., Fogliatto, F. S., & Anzanello, M. J. (2020). Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews. *Journal of Air Transport Management*, 83(1), 101760.
- Maltby, D. (2011). Big data analytics. In 74th Annual Meeting of the Association for Information Science and Technology (ASIST) (pp. 1-6).

- Mirkin, B. (2005). Clustering for Data Mining: A Data Recovery Approach, Chapman & Hall – CRC.
- Oğuz, B. (2009). Metin Madenciliği Teknikleri Kullanılarak Kulak Burun Boğaz Hasta Bilgi Formlarının Analizi, Akdeniz Üniversitesi, Yüksek Lisans Tezi, Sağlık Bilimleri Enstitüsü, Yrd. Doç. Dr. Uğur Bilge, Antalya.
- Onan, A. (2017). Twitter mesajları üzerinde makine öğrenmesi yöntemlerine dayalı duygu analizi. Yönetim Bilişim Sistemleri Dergisi, 3(2), 1-14.
- Özçelik, M., Arıcan, B. N., Bakay, Ö., Sarmış, E., Ergelen, Ö., Bayezit, N. G., & Yıldız, O. T. (2021). HisNet: a polarity lexicon based on wordnet for emotion analysis. In Proceedings of the 11th Global Wordnet Conference (pp. 157-165), Pretoria.
- Özmutlu, A. E. (2021). Doğal Dil İşleme. Bilgisayar Bilimlerinde Teorik ve Uygulamalı Araştırmalar.
- Pilavcılar, İ. F. (2007). Metin Madenciliği ile Metin Sınıflandırma. Yüksek Lisans Tezi. İstanbul: Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü Matematik Mühendisliği Anabilim Dalı.
- Porter, M. F. (1980). An Algorithm for Suffix Stripping, Program: Electronic Library and Information Systems, 14(3), 130-137.
- Puthier, D. (2008). Data Mining And Integration of Genomic Data For Gene regulatory Network Analysis.
- Raina, V., & Krishnamurthy, S. (2022). *Natural language processing. Building an Effective Data Science Practice* (pp. 63-73). California: Apress Publishing.
- Rane, A., & Kumar, A. (2018). Sentiment classification system of twitter data for US airline service analysis. IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), 1(41), 769-773.
- Rumelli, M., Akkuş, D., Kart, Ö., & Isik, Z. (2019). Sentiment analysis in Turkish text with machine learning algorithms. IEEE Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-5).

- Safran, B. (2015). Bilgi Yönetilebilir mi? Kavramsal Bir Yaklaşım. Türk Kütüphaneciliği. 29(3), 550-565. Erişim adresi: <https://dergipark.org.tr/tr/pub/tk/issue/48741/620093>.
- Seker, S. E. (2015). Metin Madenciliği (Text Mining). Yönetim Bilişim Sistemleri Ansiklopedisi. 30-32.
- Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 5(1), 1-16.
- Şahin, E. (2018). Makine öğrenme yöntemleri ve kelime kümesi tekniği ile istenmeyen e-posta/e-posta sınıflaması. Yüksek Lisans Tezi. Ankara: Hacettepe Üniversitesi, Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı.
- Terán, L., & Mancera, J. (2019). Dynamic profiles using sentiment analysis and twitter data for voting advice applications. Government Information Quarterly, 37(1), 1.
- Türkmenoğlu, C. (2015). Türkçe metinlerde duygu analizi. Yüksek Lisans Tezi. İstanbul: İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı.
- Uçan, A. (2014). Otomatik duygu sözlüğü çevirimi ve duygu analizinde kullanımı. Yüksek Lisans Tezi. Ankara: Hacettepe Üniversitesi, Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı.
- Ü, Ç. (2008). Türkçe İçin Doğal Dil İşleme Çalışmaları – 1 Biçim bilimsel ve Morfolojik Analize Çözüm Önerileri, <http://www.teknoturk.org/docking/yazilar/tt000135-yazi.htm>, 25.02.2008.
- Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks, 5(1), 7-16.
- Villars, R. L., Olofson, C. W., & Eastwood, M. (2011). Big data: What it is and why you should care. Massachusetts: International Data Corporation.

- W., B. D., S., S., L., O.-M., A., S., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7), 1123-1131.
- W., B. D., S., S., L., O.-M., A., S., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7), 1123-1131.
- Wang, J. (. (2003). *Data mining: opportunities and challenges*. Pennsylvania: Idea Group Publishing, 7.
- Yang, S., & Zhang, H. (2018). Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis. *International Journal of Computer and Information Engineering*, 12(7), 525-529.
- Yeşilyurt, S., & Dalkılıç, H. (2021). Xgboost ve Gradient Boost Machine ile Günlük Nehir Akımı Tahmini. 3rd International Symposium of Engineering Applications on Civil Engineering and Earth Sciences (IACES) (pp. 36-44), Karabük.
- Yurtalan, G., Koyuncu, M., & Turhan, Ç. (2019). A polarity calculation approach for lexicon-based Turkish sentiment analysis. *Turkish Journal of Electrical Engineering & Computer Sciences*, 27(2), 1325-1339.