ROBOTİK SÜREÇ OTOMASYONU SÜREÇLERİNDE YAPISAL OLMAYAN VERİLERİN DOĞAL DİL İŞLEME VE MAKİNE ÖĞRENMESİ TEKNİKLERİYLE SINIFLANDIRILMASI

Zeynep ÖRPEK
241402204

YÜKSEK LİSANS TEZİ
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı
Danışman: Dr. Öğr. Üyesi Önder ŞAHİNASLAN

İstanbul
T.C. Maltepe Üniversitesi
Lisansüstü Eğitim Enstitüsü
Haziran, 2026

ROBOTİK SÜREÇ OTOMASYONU SÜREÇLERİNDE YAPISAL OLMAYAN VERİLERİN DOĞAL DİL İŞLEME VE MAKİNE ÖĞRENMESİ TEKNİKLERİYLE SINIFLANDIRILMASI

Zeynep ÖRPEK
241402204
ORCID: 0000-0001-7130-9118

YÜKSEK LİSANS TEZİ
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı
Danışman: Dr. Öğr. Üyesi Önder ŞAHİNASLAN

İstanbul
T.C. Maltepe Üniversitesi
Lisansüstü Eğitim Enstitüsü
Haziran, 2026



JÜRİ VE ENSTİTÜ ONAYI

ZEYNEP ÖRPEK'ın "Robotik Süreç Otomasyonu Süreçlerinde Yapısal Olmayan Verilerin Doğal Dil İşleme ve Makine Öğrenmesi Teknikleriyle Sınıflandırılması" başlıklı tezi 25.06.2026 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Maltepe Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliği" nin ilgili maddeleri uyarınca Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans/Doktora/Sanatta Yeterlik tezi oy birliğiyle/oy çokluğuyla, başarılı/başarısız olarak kabul edilmiştir.

	Unvanı. Adı ve Soyadı	İmza
Üye (Tez Danışmanı)	Dr. Öğr. Üyesi Önder ŞAHİNASLAN Maltepe Üniversitesi	.....
Üye	Doç. Dr. Emin BORANDAĞ Manisa Celal Bayar Üniversitesi	.....
Üye	Dr. Öğr. Üyesi Erdal YAŞLICA Maltepe Üniversitesi	.....

Prof. Dr. Selva ERSÖZ KARAKULAKOĞLU
Enstitü Müdürü

TEŞEKKÜR

Bu tez çalışmam süresince bilgi ve tecrübeleriyle bana yol gösteren değerli danışman hocam Dr. Öğr. Üyesi Önder ŞAHİNASLAN'a teşekkürlerimi sunarım.

Hayatımda her zaman yanımda olan, desteği, sevgisi ve inancıyla bana güç veren sevgili eşim Adem'e en içten teşekkürlerimi sunarım. Hayatım boyunca bana her konuda güvenen, destek olan, emeklerini ve sevgilerini her zaman hissettiren anneme, babama ve kardeşlerime sonsuz teşekkür ederim.

Zeynep ÖRPEK

Haziran, 2026

ÖZET

ROBOTİK SÜREÇ OTOMASYONU SÜREÇLERİNDE YAPISAL OLMAYAN VERİLERİN DOĞAL DİL İŞLEME VE MAKİNE ÖĞRENMESİ TEKNİKLERİYLE SINIFLANDIRILMASI

Zeynep Örpek
Yüksek Lisans Tezi
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı
Danışman: Dr. Öğr. Üyesi Önder Şahinaslan
Maltepe Üniversitesi Lisansüstü Eğitim Enstitüsü, 2026

Robotik süreç otomasyonu sistemleri, yapılandırılmış ve kural tabanlı verilerle yürütülen işlemlerde yüksek verimlilik sağlamaktadır. Ancak doğal dilde yazılmış ve standart bir formata sahip olmayan e-postaların işlenmesinde sınırlı kalabilmektedir. Bu durum manuel kontrol ihtiyacını artırabilmektedir. Bu tez çalışmasında, robotik süreç otomasyonu süreçlerine gelen yapısal olmayan e-posta içeriklerinin doğal dil işleme ve makine öğrenmesi teknikleriyle otomatik olarak sınıflandırılması ele alınmıştır. Çalışmada, bankacılık sektöründe faaliyet gösteren bir şirketin robotik süreç otomasyonu robotlarına iletilen 5.243 e-posta kaydı kullanılmıştır. İlk deneylerde bazı modellerin başarı değerlerinin 1.00'a yakın çıkması, benzer e-posta içeriklerinin eğitim ve test kümelerine aynı anda düşmesinden kaynaklanabilecek veri sızıntısı riski olarak değerlendirilmiştir. Bu nedenle benzer içerikler kosinüs benzerliği kullanılarak gruplandırılmış ve model performansları grup temelli çapraz doğrulama yöntemiyle ölçülmüştür. Modelleme aşamasında destek vektör makinesi, k-en yakın komşu, rastgele orman ve BERTurk modeli karşılaştırmalı olarak değerlendirilmiştir. Deneysel sonuçlara göre BERTurk modeli yüzde 94 doğruluk, 0,93 makro ortalama F1 skoru ve 0,96 ağırlıklı ortalama F1 skoru ile en başarılı model olmuştur. Çalışma kapsamında ayrıca, sınıflandırma modelinin robotik süreç otomasyonu sistemlerine bağımsız bir servis olarak entegre edilebilmesine yönelik sistem mimarisi önerilmiştir. Elde edilen bulgular, e-posta içeriklerinin robotik süreç otomasyonu süreçlerine göre otomatik olarak sınıflandırılabileceğini göstermektedir.

Anahtar Sözcükler: Robotik Süreç Otomasyonu, Doğal Dil İşleme, Makine Öğrenmesi, Yapısal Olmayan Verilerin Anlamlandırılması

ABSTRACT

CLASSIFICATION OF UNSTRUCTURED DATA IN ROBOTIC PROCESS AUTOMATION PROCESSES USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING TECHNIQUES

Zeynep Örpek
Master Thesis
Department of Computer Engineering
Master Program in Computer Engineering with Thesis
Thesis Advisor: Assist. Prof. Dr. Önder Şahinaslan
Maltepe University Graduate School, 2026

Robotic process automation systems provide high efficiency in operations carried out with structured and rule-based data. However, they may remain limited in processing e-mails written in natural language and lacking a standard format. This situation may increase the need for manual control. This thesis examines the automatic classification of unstructured e-mail contents received by robotic process automation processes using natural language processing and machine learning techniques. The study uses 5,243 e-mail records sent to the robotic process automation robots of a company operating in the banking sector. In the initial experiments, some models produced performance values close to 1.00, which was considered a potential data leakage risk caused by similar e-mail contents being included simultaneously in both training and test sets. Therefore, similar contents were grouped using cosine similarity, and model performances were evaluated through a group-based cross-validation method. In the modeling phase, support vector machine, k-nearest neighbor, random forest, and the BERTurk model were comparatively evaluated. According to the experimental results, the BERTurk model achieved the best performance with an accuracy of 94%, a macro average F1-score of 0.93, and a weighted average F1-score of 0.96. In addition, within the scope of the study, a system architecture was proposed to enable the classification model to be integrated into robotic process automation systems as an independent service. The findings indicate that e-mail contents can be automatically classified according to robotic process automation processes.

Keywords: Robotic Process Automation, Natural Language Processing, Machine Learning, Unstructured Data Classification

İÇİNDEKİLER

JÜRİ VE ENSTİTÜ ONAYI ii

TEŞEKKÜR iii

ÖZET iv

ABSTRACT v

İÇİNDEKİLER vi

TABLOLAR LİSTESİ ix

ŞEKİLLER LİSTESİ x

KISALTMALAR xi

1. GİRİŞ 1

1.1 Problem 10

1.2 Amaç 12

1.3 Çalışmanın Önemi 13

1.4 Varsayımlar 13

1.5 Sınırlılıklar 13

1.6 Tanımlar 14

2. LİTERATÜR ARAŞTIRMASI 17

2.1 Yapısal Olmayan Verilerin Analizi ve Sınıflandırılmasına Yönelik Çalışmalar 17

2.2 RPA Süreçlerinde Yapısal Olmayan Verilerin ML ve NLP ile Anlamlandırılmasına Yönelik Çalışmalar 22

3. YÖNTEM 29
3.1 Araştırma Tasarımı ve Metodolojik Çerçeve 29
3.2 CRISP-DM Aşamalarına Göre Yöntemsel Süreç 31
3.2.1 İş anlayışı 33
3.2.2 Veri anlayışı 33
3.2.3 Veri hazırlama 39
3.2.3.1 Kullanılmayan kolonların çıkarılması 40
3.2.3.2 E-posta başlıklarında yer alan uyarı metinlerinin temizlenmesi 40
3.2.3.3 E-posta Footer bölümlerinin temizlenmesi 40
3.2.3.4 E-posta meta verilerinin ön işleme sürecinde kaldırılması 40
3.2.3.5 Süreçle ilgili olmayan metinlerin temizlenmesi 41
3.2.3.6 Regex tabanlı temizlik işlemleri 41
3.2.4 Modelleme 45
3.2.4.1 Destek vektör makinesi 49
3.2.4.2 k-En yakın komşu 50
3.2.4.3 Rastgele orman 51
3.2.4.4 Transformer tabanlı derin öğrenme modelleri 51
3.2.5 Değerlendirme 52
3.2.6 Uygulama 57
3.2.7 Önerilen sistem mimarisi 57
4. BULGULAR VE TARTIŞMA 61
4.1 Bulgular 61

4.1.1 Klasik ML modellerinin sonuçları 62
4.1.2 Transformer tabanlı model (BERTurk) sonuçları 62
4.2 Tartışma 63
4.3 Deney Sürecinde Karşılaşılan Sorunlar ve Çözüm Yöntemleri 67
5. SONUÇ ve ÖNERİLER 69
KAYNAKÇA 73
ÖZ GEÇMİŞ 78

TABLOLAR LİSTESİ

Tablo 1. İlgili Çalışmaların Özeti 26

Tablo 2. Veri Setinde Yer Alan Alanlar 34

Tablo 3. Veri Setinde Yer Alan Süreç Bilgileri 36

Tablo 4. Veri Setinden Anonimleştirilmiş Örnek Kayıtlar 37

Tablo 5. Ön İşleme Çalışmalarında Kullanılan Kütüphaneler 41

Tablo 6. Veri Ön İşleme Sonucu Veri Sayıları 42

Tablo 7. Çalışmada Kullanılan Modellerin Kod Tanımları 48

Tablo 8. Ön Deneylerde Kullanılan Eğitim-Test Veri Seti Ayrım Oranı 52

Tablo 9. Karışıklık Matrisi Bileşenleri ve Açıklamaları 55

Tablo 10. Klasik ML Modellerinin Sonuçlarının Karşılaştırılması 62

Tablo 11. BERTurk Modeli Sonuçları 63

ŞEKİLLER LİSTESİ

Şekil 1. Yapay Zekanın Alt Alanları 6

Şekil 2. ML Yöntemlerinin Öğrenme Türlerine Göre Sınıflandırılması 7

Şekil 3. CRISP-DM Metodolojisi..... 30

Şekil 4. Çalışmanın Yöntemsel İş Akış Diyagramı 32

Şekil 5. Benzerlik Kavramı İçin Kosinüs Derecesi 43

Şekil 6. E-posta Metinlerinin İşlenmesi Adımları ve İlgili Kütüphaneler..... 44

Şekil 7. Çalışmada Geliştirilen Model Eğitim ve Değerlendirme Kodları..... 47

Şekil 8. Çalışmada Kullanılan Modelleme Yaklaşımlarının Şematik Gösterimi 49

Şekil 9. Önerilen Sistem Mimarisi 58

KISALTMALAR

Kısaltma		Türkçe	İngilizce
AI	:	Yapay Zeka	Artificial Intelligence
BERT	:	Çift Yönlü Kodlayıcı Temsilleri	Bidirectional Encoder Representations from Transformers
CRISP-DM	:	Veri Madenciliği için Endüstriler Arası Standart Süreç Modeli	Cross-Industry Standard Process for Data Mining
DL	:	Derin Öğrenme	Deep Learning
k-NN	:	k-En Yakın Komşu	k-Nearest Neighbor
ML	:	Makine Öğrenmesi	Machine Learning
NLP	:	Doğal Dil İşleme	Natural Language Processing
RF	:	Rastgele Orman	Random Forest
RPA	:	Robotik Süreç Otomasyonu	Robotic Process Automation
SVM	:	Destek Vektör Makinesi	Support Vector Machine
TF-IDF	:	Terim Frekansı-Ters Doküman Frekansı	Term Frequency-Inverse Document Frequency

1. GİRİŞ

Teknolojinin gelişmesiyle birlikte iş süreçleri önemli bir dönüşüm geçirmiştir ve otomasyon sistemlerinin kullanımı yaygınlaşmıştır. Otomasyon sistemleri, iş süreçlerinde insan müdahalesini en aza indirerek uygun görevlerin veya süreçlerin makineler, yazılımlar ya da teknolojik sistemler aracılığıyla yapılmasıdır. Otomasyon sistemleri, iş süreçlerinin daha hızlı ve hatasız bir şekilde yapılmasına olanak sağlayarak kurumsal verimliliği artırmayı amaçlamaktadır.

Günümüz rekabet koşulları, dijitalleşme, endüstri 4.0, yapay zeka gibi bilgi teknolojilerindeki ilerlemeler, kullanım alanlarının yaygınlaşması, teknolojik yenilik ve dönüşümler iş süreçlerinin yeniden ele alınmasını, sürekli iyileştirmelerle optimize edilmesini zorunlu kılmaktadır (Şahinaslan, İş Süreci Optimizasyonu: Yöntem, Teknoloji, Riskler ve Fırsatlar, 2023). Çoğu sektörde, iş akışlarında yer alan belirli bir düzende sürekli olarak gerçekleştirilen rutin işlemler yer almaktadır. Teknolojinin ilerlemesine ve dijitalleşmenin yaygınlaşmasına rağmen, bu tür tekrar eden rutin görevlerin sayısı gün geçtikçe artmaktadır. Bu tekrar eden rutin işlemler, işletmelerin iş süreçlerinde önemli bir yere sahiptir. Çalışanların tekrar eden rutin işler ile ilgilenmesi, çalışanların zaman ve enerjilerini daha yaratıcı, analitik düşünme gerektiren, stratejik veya nitelikli görevlere zaman ayıramamalarına neden olmaktadır. Bu durum çalışanları, katma değeri daha düşük olan bu rutin görevleri yapmaya itmektedir. Bu nedenle, çalışanların potansiyeli verimli bir biçimde kullanılamamaktadır. Bu durum, çalışanların bireysel verimliliğini ve işletmelerin kurumsal performansını olumsuz etkilemektedir. Bu nedenle, iş süreçlerinin otomasyonu ve dijital çözümlerle desteklenmesi, çalışanların daha nitelikli işlere yönlendirilmesi açısından büyük öneme sahiptir.

Teknolojinin gelişmesiyle birlikte otomasyon sistemleri de önemli bir evrim geçirmiş ve birçok sektörde iş yapış şekillerinde köklü bir değişikliğe neden olmuştur. Bu dönüşüm sürecinde özellikle RPA (Robotik Süreç Otomasyonu, Robotic Process Automation), dikkat çeken çözümlerden biri olarak öne çıkmaktadır. RPA, yapılandırılmış ve tekrarlayan iş süreçlerinin RPA robotları aracılığıyla otomatikleştirilmesini sağlayan bir teknolojidir. Operasyonel işlerde RPA'nın kullanımıyla birlikte süreçler daha verimli hale

gelirken iş kapasitesinde bazı çalışmalarda %20'ye varan artışlar rapor edilmiştir (Aguirre ve Rodriguez, 2017). İşletmeler, bu teknolojiyi kullanarak operasyonel maliyetlerini düşürürken insandan kaynaklanan hataları en aza indirerek süreçlerini daha güvenilir hale getirmektedir. Bu nedenle, RPA çözümleri günümüzde farklı ölçeklerdeki birçok kuruluş için stratejik bir araç haline gelmiştir.

RPA, birçok sektörde yaygın olarak kullanılmaktadır. Bu sektörlerden bazıları aşağıdaki gibidir:

- Bankacılık ve finans,
- Sigortacılık,
- Sağlık ve ilaç,
- İnsan kaynakları,
- Bilgi teknolojileri,
- Haberleşme ve telekomünikasyon,
- Perakende,
- Kamu hizmetleri,
- Gayrimenkul,
- Ulaşım ve lojistik,
- İmalat,
- Enerji sektörleri (Özdem ve Bora, 2022).

RPA finans sektöründe yaygın olarak kullanılmakta olup, yinelenen ve kural tabanlı işlemleri otomatikleştirme konusunda önemli avantajlar sunmaktadır. Bu teknoloji veri girişi, fatura işleme, mutabakat işlemleri, varlık muhasebesi ve iç denetim gibi işlemlerde insan müdahalesini azaltarak hız, doğruluk ve maliyet tasarrufu sağlamaktadır (Aguirre ve Rodriguez, 2017).

Sigortacılık sektöründe poliçe yenileme, hasar tespiti, fiyat teklifi hesaplama ve müşteri destek hizmetlerinde kullanılmaktadır (Kumar, 2022).

Sağlık sektöründe hasta kayıt işlemleri, fatura düzenleme, randevu planlama ve sigorta taleplerinin işlenmesi gibi süreçlerde kullanılarak sağlık sektörü çalışanları tekrar eden rutin işler yapmak yerine hastaların sağlığı ile ilgili daha kritik işler yapmasına olanak sağlamaktadır (Venigandla, 2022).

İnsan kaynakları departmanlarında RPA kullanılarak işe alım süreci, bordro yönetimi, performans değerlendirmesi, işe devam kayıtlarının tutulması, çalışan özlük işlemleri, çalışanlardan veri toplanması, depolanması ve yönetimi gibi süreçleri otomatikleştirebilmektedir (Mohamed ve arkadaşları, 2022).

RPA, bu sektörler ile sınırlı kalmayıp farklı birçok sektörde de yaygın olarak kullanılmaktadır. RPA, tekrarlayan ve kural tabanlı süreçlerin otomasyonu sayesinde organizasyonların daha hızlı ve etkin çalışmasına katkı sağlamaktadır.

RPA teknolojisi, kullanım biçimi ve görevlerin gerçekleştirilme düzeyine göre farklı türlerde sınıflandırılmaktadır. Bu sınıflandırma, otomasyonun süreç içindeki rolü ve insan etkileşimi gerekliliği gibi kriterler dikkate alınarak yapılmaktadır. Etkileşimli, bağımsız ve hibrit olmak üzere üç temel RPA türü bulunmaktadır (Axmann ve Harmoko, 2020).

Etkileşimli RPA (Attended RPA) türü, insan kullanıcı ile birlikte çalışan ve genellikle kullanıcı tarafından tetiklenen otomasyon robotlarını ifade etmektedir. Etkileşimli RPA robotları, masaüstü uygulamaları üzerinde çalışarak bir iş sürecinin belirli bir kısmını otomatikleştirmektedir. Kullanıcı bir işlem başlattığında RPA robotu aktifleşerek işlemi belirli bir noktaya kadar yürütmektedir.

Bağımsız RPA (Unattended RPA) türü, herhangi bir kullanıcı etkileşimi olmadan, belirlenen zamanlarda veya belirli olaylar gerçekleştiğinde otomatik olarak çalışmaktadır. Bu RPA robotları, iş süreçlerini uçtan uca kullanıcı müdahalesi olmaksızın kendi başlarına gerçekleştirme potansiyeline sahiptir. Bağımsız RPA robotları süreçlerin 7/24 çalışabilmesine olanak sağlamaktadır.

Hibrit RPA (Hybrid RPA) hem etkileşimli hem de bağımsız RPA robotlarının birlikte kullanıldığı bir otomasyon türüdür. Özellikle hem kullanıcı etkileşimi gerektiren hem de kullanıcı etkileşimi olmadan otomatik olarak yürütülebilecek işlemler içeren karmaşık iş süreçlerinde kullanılmaktadır. Hibrit RPA sistemleri, süreç verimliliğini artırma ve kullanıcı destekli karar alma noktalarında etkili çözüm sunmaktadır.

RPA'nın temelini RPA robotları oluşturmaktadır. RPA robotları, belirli görevleri yerine getirmek üzere özel olarak geliştirilmiş yazılımlardır. RPA, kopyalama, yapıştırma, veri çıkarma, birleştirme ve sistemler arası veri taşıma gibi tekrarlayan ve katma değeri düşük görevlerde insan davranışını taklit edebilen yazılım tabanlı bir otomasyon teknolojisidir (Aguirre ve Rodriguez, 2017). RPA robotları, insan kullanıcıların bilgisayar ortamında gerçekleştirdiği işlemleri taklit edebilecek şekilde geliştirilmektedir.

RPA sistemleri, sadece belirli görevleri yerine getiren RPA robotlarının yanı sıra iyi kurulmuş bir otomasyon mimarisi, süreçlerin tasarlanması, geliştirilmesi, devreye alınması, yürütülmesi ve izlenmesi gibi çeşitli işlevleri yerine getiren çok sayıda sürecin birbiriyle uyumlu olarak çalışmasıyla mümkün olmaktadır.

RPA robotlarının etkin bir şekilde yönetilebilmesi ve merkezi olarak kontrol edilebilmesi için Orchestrator adlı bir yönetim aracı kullanılmaktadır. Orchestrator, RPA robotlarına görev atama, süreçlerin zamanlanması, çalışma performansının izlenmesi, işlem kayıtlarının tutulması ve hata yönetiminin sağlanması gibi temel görevleri yerine getirmektedir. Özellikle büyük ölçekli şirketlerde, birden fazla RPA robotunun eşzamanlı çalışması durumunda süreçlerin yönetimi bu araç üzerinden yapılmaktadır.

Her RPA robotu, Orchestrator tarafından atanan bir veya birden fazla iş süreci üzerinde çalışmaktadır. Bu süreçler, UiPath Studio, Automation Anywhere, Blue Prism vb. gibi RPA geliştirme araçları kullanılarak tasarlanan, işlem adımlarını tanımlayan otomasyon senaryolarıdır. Süreçler, sistematik ve kural tabanlı şekilde yapılandırılmaktadır ve yapılandırılmış verilerle çalışmaya uygundur. Tanımlanan süreçler, paketlenerek uygun RPA robotlarına dağıtılarak yürütme süreci Orchestrator üzerinden kontrol edilmektedir.

Otomasyon süreçlerinin sürdürülebilirliği ve yeniden kullanılabilirliği konularında geliştirme ortamı (framework) yapıları önem arz etmektedir. Bu geliştirme ortamı

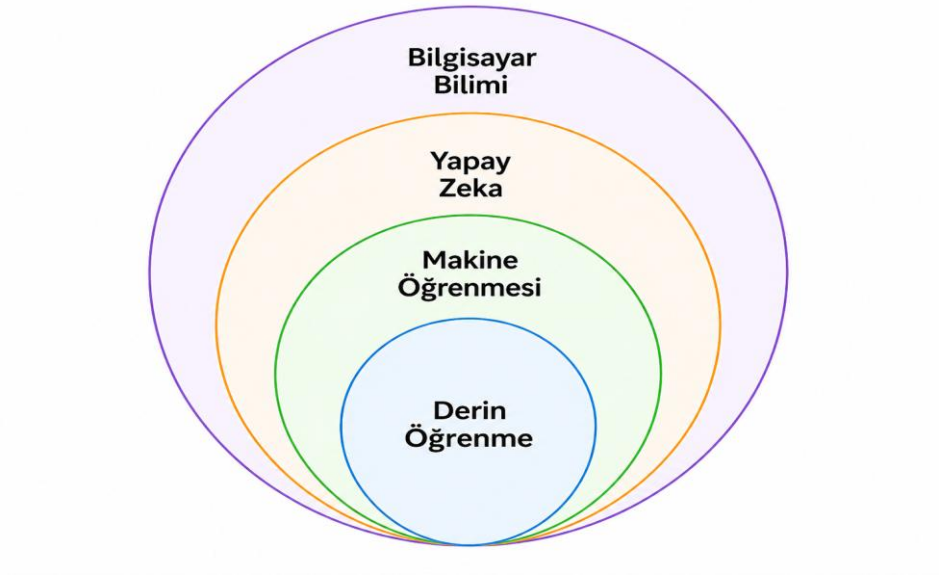
yapıları, otomasyon projelerinin standart bir mimari üzerinde geliştirilmesini sağlayarak hata yönetimi, loglama, bakım ve test gibi işlemleri sistematik biçimde yapılabilmesine ve yönetilmesine altyapı sağlamaktadır.

Geliştirilen RPA süreçlerinin test edilmesi, versiyonlanması ve canlı ortama geçişinin yapılması, dağıtım hattı (pipeline) üzerinde yapılandırılmaktadır. Dağıtım hatları, özellikle sürekli entegrasyon/sürekli dağıtım (CI/CD) süreçlerinin kullanıldığı büyük ölçekli kurumsal şirketlerde, geliştirilen süreçlerin otomatik olarak test edilmesini, hataların tespit edilmesini ve güncellemelerin sorunsuz bir biçimde canlı ortama geçişini sağlamaktadır.

RPA robotlarında, veri işleme süreçlerinde görevlerin dağıtılması ve önceliklendirilmesi için kuyruk yapısı kullanılmaktadır. Bu kuyruklar, işlemleri sıraya koyarak her bir işlemin RPA robotları tarafından bağımsız biçimde işlenmesini sağlamaktadır. Özellikle büyük hacimli işlemlerde iş yükünün dengelenmesi, işlemlerin paralel olarak yürütülmesi ve hata alan görevlerin yeniden çalıştırılabilmesi için altyapı sağlamaktadır.

Bütün bu bileşenler, RPA sistemlerinin verimli, güvenli ve sürdürülebilir bir şekilde işletilmesine altyapı sağlamaktadır. RPA sistemlerinin yapısal olmayan veriler karşısındaki sınırlılıkları, yapay zeka, makine öğrenmesi ve doğal dil işleme yaklaşımlarının bu alandaki rolünü ön plana çıkarmaktadır.

Bilgisayar bilimi, algoritmalar, veri yapıları ve matematiksel hesaplama işlemlerini içeren temel bir disiplindir. Bilgisayar bilimi alanı içerisinde yapay zeka, makinelerin veya bilgisayarların insan zekasına benzer biçimde karar verebilmesini amaçlayan bir araştırma alanıdır. Şekil 1'de yapay zeka, makine öğrenmesi ve derin öğrenme arasındaki hiyerarşik ilişki gösterilmektedir.



Şekil 1. Yapay Zekânın Alt Alanları

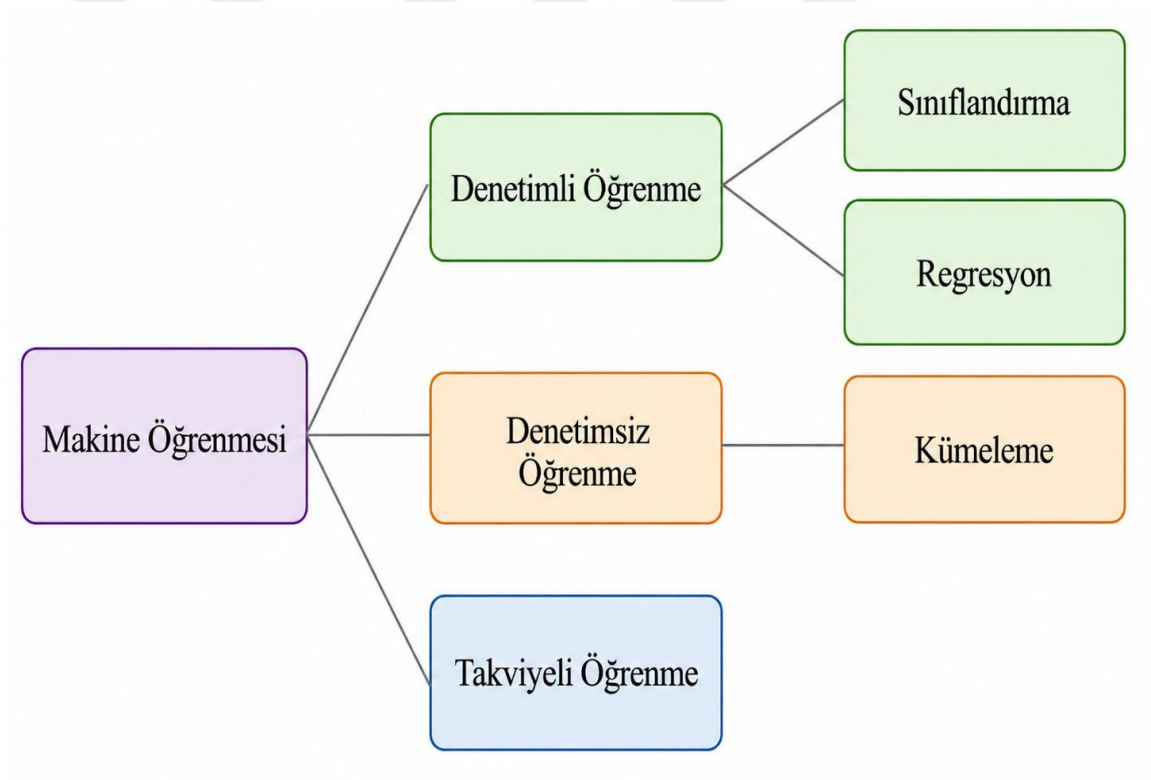
Yapay zekanın alt alanı olan ML (Makine Öğrenmesi, Machine Learning) alanında kullanılan yöntemler, sistemlerin önceden tanımlanmış kurallar yerine var olan verilerden öğrenerek karar verilmesini sağlamaktadır. Bu öğrenme yaklaşımlarının daha derin ve karmaşık yapılarda gerçekleştirilmesi ise derin öğrenme başlığı altında toplanmaktadır. Bu yöntemler çok katmanlı sinir ağlarıyla çalışmaktadır.

Bilgisayar bilimi, yapay zekâ aracılığıyla insan benzeri zekaya sahip sistemler geliştirirken, bu zekanın temelini ML oluşturur ve derin öğrenme bu sürecin en ileri düzey uygulamalarından biridir.

Yapay zeka, insan benzeri davranışları sergileme yeteneğine sahip yazılımlardır. Bu yazılımlar, temelde insan zekâsını ve davranışlarını taklit ederek ilgili süreci öğrenmektedir. Yapay zeka, bilgisayarlara veya makinelere insana benzer bir şekilde, görme, konuşma, anlama, öğrenme ve karar verme gibi karmaşık bilişsel yetenekler kazandırmaktadır.

ML, matematik ve istatistiksel yöntemler kullanılarak verilerden örüntüler çıkaran ve bu örüntülere dayalı tahmin veya sınıflandırma yapabilen modellerin geliştirilmesini sağlayan bir yapay zeka alt alanıdır.

Şekil 2'de ML yöntemlerinin öğrenme türlerine göre sınıflandırılması gösterilmektedir. Buna göre ML yöntemleri genel olarak denetimli öğrenme, denetimsiz öğrenme ve takviyeli öğrenme olmak üzere üç temel grupta ele alınmaktadır. Denetimli öğrenme, etiketli veri kullanılarak modelin girdiler ile çıktılar arasındaki ilişkiyi öğrenmesine dayanmakta ve sınıflandırma ile regresyon problemlerinde yaygın olarak kullanılmaktadır. Denetimsiz öğrenme ise etiket bilgisi bulunmayan verilerde benzer örüntülerin keşfedilmesini amaçlamakta ve kümeleme yöntemleriyle temsil edilmektedir. Takviyeli öğrenme ise bir ajanın çevresiyle etkileşime girerek ödül ve ceza mekanizması üzerinden en uygun davranışı öğrenmesine dayanmaktadır. Bu tez çalışması, e-posta içeriklerinin önceden tanımlı RPA süreçlerine atanmasını amaçladığı için denetimli öğrenme kapsamında yer alan çok sınıflı sınıflandırma problemi olarak ele alınmıştır.



Şekil 2. ML Yöntemlerinin Öğrenme Türlerine Göre Sınıflandırılması

ML yöntemleri, verilerden öğrenerek tahmin veya sınıflandırma yapabilen modeller geliştirmek amacıyla kullanılmaktadır ve öğrenme türleri üç gruba ayrılmaktadır. Bunlar:

1. Gözetimli (Denetimli) Öğrenme: Bu öğrenme şeklinde girdiler ve çıktılar arasında eşleme yapan fonksiyon regresyon veya sınıflandırma yöntemleriyle oluşturulmaktadır. Eğitim verileri hem girdileri hem de çıktıları içermektedir.
2. Gözetimsiz (Denetimsiz) Öğrenme: Bu öğrenme şeklinde, veri içerisindeki gizli yapılar bulunmaya çalışılmaktadır ve giriş verilerine ait sınıf etiketleri önceden bilinmez.
3. Takviyeli Öğrenme: Bu öğrenme şeklinde, girdilerin çevresiyle etkileşerek ödül ve ceza denilen geri bildirimler ile en uygun hareketi bulmayı amaçlamaktadır ve diğer öğrenme türlerinden farklıdır (Arslankaya ve Toprak, 2021).

ML problemleri temelde sınıflandırma ve kümeleme olmak üzere iki farklı yöntem ile verileri sınıflara ayırmaktadır. Bu yöntemler benzer amaçlara hizmet etse de öğrenme biçimleri ve kullanım alanları açısından farklılık göstermektedir.

Sınıflandırma yöntemi denetimli öğrenme kategorisine aittir. Bu yöntemde sınıflandırma algoritmaları veri setinde yer alan eski veriler üzerinde öğrenerek gelecek veri için sınıflandırma yapmaktadır. İkili, çok sınıflı ve çok etiketli olmak üzere üç farklı tipi bulunmaktadır. İkili sınıflandırma çıktı sınıfının iki adet olduğu durumdur. Doğru/yanlış, 0/1, kadın/erkek, spam/spam değil gibi örnekler ikili sınıflandırma olarak örnek olarak verilebilir. Çok sınıflıda ise çıktı sınıfı ikiden fazla olduğu durumdur. Köpek/kedi/kuş/balık, bebek/çocuk/genç/yaşlı gibi örnekler çok sınıflı sınıflandırma olarak örnek olarak verilebilir. Çok etiketli sınıflandırmada ise çıktının farklı etiketler alabilmesi durumudur. Görüntü sınıflandırmalarında aynı görüntü üzerinde kuş ve kedi olması durumu için çıktının aynı anda kuş ve kedi etiketlerini içermesi örnek olarak verilebilir.

Regresyon, çıktısı sürekli ve sayısal değerlerden oluşan problemleri çözmek amacıyla kullanılan denetimli bir öğrenme yöntemidir. Regresyon modellerinin temel amacı, girdi değişkenleri ile çıktı değişkeni arasındaki ilişkiyi öğrenerek yeni ve daha önce görülmemiş girdiler için çıktı değerlerini tahmin etmektedir. Bir evin oda sayısı ve bulunduğu kat gibi özelliklerinden satış fiyatının tahmin edilmesi regresyon problemlerine örnek olarak verilebilir.

Kümeleme yöntemi ise denetimsiz öğrenme kategorisine aittir. Bu yöntemde kümeleme algoritmaları veri setindeki örnekleri benzerliklerine göre kümelere ayırmaktadır. Bu sayede aynı özelliklere sahip verileri bu kümeler üzerinde ayrıştırmaktadır. Kümeleme yönteminde hedef değişken yoktur, kümeler üzerinden benzer özelliğe sahip veriler toplanmaktadır. Küme sayısı deneme yanılma, sezgisel ve algoritmalar ile belirlenebilmektedir.

Makine öğrenimi, derin öğrenme algoritmaları kullanarak insan zekâsını taklit eden bir teknolojidir (Şahinaslan ve arkadaşları, 2023). Derin öğrenme, büyük veri kümeleri üzerinde çok katmanlı yapay sinir ağları kullanarak karmaşık örüntüleri öğrenmeyi amaçlamaktadır. Her sektör elindeki büyük veriyi kullanarak öğrenme temelli daha zeki sistemler, mimariler kurmak zorundadır (Günerkan ve arkadaşları, 2019). Makine öğrenmesinden farklı olarak, verilerdeki özelliklerin önemli bir kısmını manuel olarak belirlemek yerine, bu özellikleri katmanlar aracılığıyla otomatik olarak çıkarabilmektedir. Bu nedenle görüntü işleme, ses tanıma, doğal dil işleme ve metin sınıflandırma gibi alanlarda yüksek başarı göstermektedir. Derin öğrenme modelleri, özellikle veri miktarının fazla olduğu durumlarda daha güçlüdür. Ancak bu modellerin eğitimi klasik makine öğrenmesi yöntemlerine göre daha fazla işlem gücü, bellek ve zaman gerektirmektedir.

Bu tez çalışması, e-posta içeriklerinin önceden tanımlı RPA süreçlerine atanmasını amaçladığı için etiketli verilerle öğrenme gerektiren denetimli öğrenme kapsamında ele alınmıştır. Problem yapısı gereği çok sınıflı metin sınıflandırma problemi olarak modellenmiştir.

NLP (Doğal Dil İşleme, Natural Language Processing), insan diliyle oluşturulan metinlerin, bilgisayarlar tarafından anlaşılabilir ve işlenebilir formata çevrilmesi için kullanılan yapay zeka alt alanlarından biridir. Bilgisayarların metin işleme, analiz etme ve üretme yeteneklerini güçlendirmektedir (Toprak, 2024).

NLP, yapay zekanın temel alt alanlarından biri olup, kullanıcıların doğal dilde ifade ettikleri girdilerin bilgisayar sistemleri tarafından anlaşılmasını ve uygun çıktılar üretilmesini sağlayan bir teknolojidir. Metin üretimi, duygu analizi ve metin sınıflandırma gibi birçok işlem NLP yöntemleri ile otomatik olarak yapılabilir.

Günümüzde yaygın olarak kullanılan sanal asistanlar, kullanıcıların sordukları soruları doğru şekilde yorumlayarak bağlama uygun yanıtlar üretebilmek için NLP tekniklerini kullanmaktadır. Bu sayede kullanıcılar, makineler ile daha doğal, hızlı iletişim kurma imkanına sahip olmaktadır.

NLP, ML ile kullanıldığında, metin verileri üzerinde öğrenme ve genelleme yeteneği kazandırmaktadır. Özellikle denetimli öğrenme yaklaşımlarında, etiketlenmiş metin verileri kullanılarak modeller eğitilmekte ve yeni metinler için doğru tahminler yapılabilmektedir. Bu yaklaşım, kurumsal ortamlarda yoğun olarak üretilen e-posta, rapor ve dokümanların otomatik olarak analiz edilmesini mümkün kılmaktadır.

Transformer mimarisi, NLP alanında kullanılan ve metinler arasındaki bağlamsal ilişkileri modellemeyi amaçlayan modern bir derin öğrenme yaklaşımıdır. Geleneksel sıralı modellerden farklı olarak transformer yapısı, metin içerisindeki kelimeleri sıralı biçimde işlemek yerine, tüm kelimeler arasındaki ilişkileri aynı anda değerlendirebilmektedir. Bu yaklaşım, özellikle uzun metinlerde bağlamın daha etkin şekilde temsil edilmesini sağlamaktadır.

Transformer mimarisinin temel bileşeni olan dikkat (attention) mekanizması, bir kelimenin anlamını belirlerken metindeki diğer kelimelerle olan ilişkisini dikkate almasına olanak tanımaktadır. Bu sayede, kelimelerin yalnızca tekil anlamları değil, bulundukları bağlama göre kazandıkları anlamsal değerler de modellenebilmektedir. Transformer tabanlı modeller, bu özellikleri sayesinde NLP görevlerinde yüksek başarı göstermektedir.

Transformer mimarisi üzerine inşa edilen modellerden biri olan BERT metinleri çift yönlü bağlamsal olarak temsil edebilen bir dil modelidir. BERT, büyük ölçekli metin koleksiyonları üzerinde ön eğitim (pre-training) süreci ile dilin genel yapısını öğrenmekte ve daha sonra belirli NLP görevleri için ince ayar (fine-tuning) yapılarak kullanılmaktadır (Devlin ve arkadaşları, 2019).

1.1 Problem

RPA sistemleri, belirli kural setleri ve algoritmalar doğrultusunda tekrarlayan görevleri yapması için programlanmaktadır. RPA sistemleri, genellikle kullanıcıların e-posta

adresleri, telefon numaraları veya mesajlaşma uygulamaları gibi çeşitli iletişim kanalları üzerinden süreçleri yürütmektedir. Bu kanalları kullanarak işlem yaptıkları süreçler hakkında bilgi alabilmektedir veya kullanıcılara bilgi verebilmektedir. Ancak, bu iletişim kanallarından gelen veriler büyük ölçüde önceden tanımlanmış bir formatı olmayan yapısal olmayan, düzensiz veya bozulmuş verilerden oluşmaktadır. RPA sistemlerinin karşılaştığı önemli bir zorluk bu verileri anlamlandıramamasıdır. Özellikle, e-posta ve mesajlaşma uygulamalarından gelen mesajlar, yapısal olmayan veriler içerebilmektedir. Bu verilerin doğru bir şekilde işlenmesi, anlaşılması ve uygun aksiyonların alınması RPA sistemleri için karmaşık bir problem haline gelmektedir.

RPA süreçlerinin kural tabanlı yapısı nedeniyle e-posta, hata bildirimi, müşteri talebi, çağrı merkezi kaydı ve serbest metin alanları gibi yapısal olmayan metin türlerini doğrudan işleyememektedir. Bu tür metinlerin bağlamsal ifadeler, yazım hataları ve biçimsel tutarsızlıklar içerebilmesi nedeniyle RPA süreçlerinde doğru iş akışının belirlenmesini zorlaştırmaktadır. Bu sorunun giderilmesinde makine öğrenmesi tabanlı metin sınıflandırma yaklaşımları kullanılabilir (Örpek ve arkadaşları, 2026).

RPA sistemleri, önceden tanımlanmış bir biçimde yapılandırılmış veriler üzerinde oldukça başarılı bir şekilde çalışabilmektedir. Ancak, doğal dilde yazılmış, anlam açısından belirsizlikler içeren, dil bilgisel yapılar bakımından çeşitli belirsizlikler içeren ve çeşitli örtük anlamlar taşıyan metinler ile çalışmak için NLP ve ML gibi yapay zeka bileşenlerine ihtiyaç duymaktadır. Bu durum, kullanıcıların RPA robotları ile etkileşiminde büyük bir engel teşkil etmektedir. Çünkü RPA sistemleri, yalnızca önceden belirlenmiş kural setlerine dayalı olarak çalışmaktadır. Kullanıcıların mesajları, bu kurallara uygun bir formatta iletilmediği takdirde, RPA robotları doğru bir şekilde etkileşime giremez veya istenilen sonucu veremez. Bu da kullanıcıların RPA robotları ile etkili bir şekilde iletişim kurmalarının önüne geçerek süreçlerin verimliliğini düşürmektedir.

Bu sorunun temel kaynağı, yapısal olmayan verilerin, yani doğal dilin, RPA sistemleri tarafından doğru bir şekilde işlenememesidir. Bu noktada, RPA robotlarının insan benzeri yorumlama yeteneğine yakın bir kapasiteye sahip olması gereklidir. Doğru bir şekilde yapısal olmayan içerikleri anlayabilecek ve gerektiğinde bu içerikleri uygun şekilde yorumlayarak ilgili aksiyonu alabilecek bir RPA sisteminin geliştirilmesi, sadece

verimlilik sağlamakla kalmayıp, aynı zamanda kullanıcı etkileşimlerinin kalitesini de artırma potansiyeline sahiptir. Bu bağlamda, RPA sistemlerinin daha verimli çalışabilmesi ve kullanıcılarla etkili bir iletişim kurabilmesi için, RPA robotlarına gelen mesajların anlamlandırılması ve uygun bir şekilde işlenmesi gereklidir.

1.2 Amaç

Bu tez çalışmasının temel amacı, RPA sistemlerine gelen yapısal olmayan mesajların sınıflandırılması ve anlamlandırılması için NLP ve ML yöntemleri ile çözüm geliştirilmesidir.

RPA sistemleri, genellikle yapılandırılmış verilerle etkin çalışmaktadır. Yapılandırılmış veriler, belirli formatlarda ve önceden tanımlanmış kurallar çerçevesinde olan verilerdir. Ancak, iş süreçlerinde kullanılan e-posta, mesajlaşma uygulamaları, sosyal medya gibi kanallardan gelen mesajlar genellikle yapısal olmayan, serbest formatta ve dilsel çeşitliliğe sahip içeriklerdir. Bu tür veriler, RPA sistemlerinin yeteneklerine uygun olmadığından dolayı RPA robotlarının aksiyonlarını sınırlandırmaktadır.

Yapısal olmayan verilerin anlamlandırılması için NLP teknikleri ve ML algoritmaları kullanılarak çözüm geliştirilebilir. NLP ve ML teknikleri, metinlerdeki anlamsal ilişkileri ortaya çıkarma, metinlerdeki ilişkileri belirleme ve dilsel yapıları analiz etme özellikleri ile RPA sistemlerinin mesajlardan anlam çıkarabilmesi için gerekli olan altyapıyı sağlamaktadır. Bu sayede, gelen veriler doğru şekilde sınıflandırılarak, daha sonra uygun aksiyon alınması sağlanabilir.

Örneğin, e-posta içeriği, gönderilen mesajın türüne göre (müşteri talebi, şikayet, bilgi sorgulama vb.) sınıflandırılabilir ve her bir kategoriye uygun ilgili aksiyon tetiklenebilir. Bu işlem, RPA robotlarının mesajları anlamlı bir şekilde işleyebilmesi ve ilgili aksiyonu alabilmesi için önemlidir.

Bu tez çalışmasında, RPA sistemlerine gelen yapısal olmayan mesajların sınıflandırılması ve anlamlandırılması için modellerin performansı karşılaştırmalı olarak analiz edilmiştir ve değerlendirilmiştir.

1.3 Çalışmanın Önemi

Bu tez çalışması, RPA sistemlerine entegre edilecek NLP ve ML yöntemlerinin, yapısal olmayan verilerin anlamlandırılması üzerindeki başarısını ve verimliliği artırma potansiyelini değerlendirmektedir. Yapısal olmayan veriler genellikle düzensiz, dilsel hatalar içeren ve karmaşık anlamlar barındıran içeriklerden oluşmaktadır. Bu tür yapısal olmayan verilerin işlenmesi, RPA sistemlerinin doğru, hızlı ve verimli bir şekilde çalışabilmesi için kritik bir öneme sahiptir. Bu kapsamda, NLP ve ML tekniklerinin kullanılması RPA sistemlerinin insan benzeri bir anlayış yeteneğine sahip olmasını sağlayarak, bu tür verilerle karşılaşıldığında sistemin doğru kararlar alabilmesini sağlayacaktır.

Bu tez çalışması, RPA sistemlerinin yapısal olmayan verilerle etkileşimde bulunma yeteneğini geliştirerek, işletmelerin dijital dönüşüm süreçlerine katkıda bulunmayı amaçlamaktadır.

1.4 Varsayımlar

Yapılan yapısal olmayan verilerin anlamlandırılması çalışması sonucunda, yapısal olmayan mesajlardan çıkarılan içeriğin, RPA robotları tarafından doğru bir şekilde işlenip ilgili süreç için aksiyon alacağı varsayılmaktadır. Bu varsayım, geliştirilen modelin sınıflandırma ve anlamlandırma aşamalarında yüksek doğruluk oranları sağlanması durumunda, RPA robotunun bu çıktıyı doğru bir şekilde işleyerek belirli iş akışları ile ilgili aksiyonları alacağı kabul edilmektedir. Dolayısıyla, yapısal olmayan verilerin anlamlandırılması çalışması sonucunda elde edilen sonucun RPA robotunun işlem süreçlerine doğru bir şekilde entegrasyonu ve bu süreçlerin verimli bir şekilde yönetilmesi beklenmektedir.

1.5 Sınırlılıklar

Bu çalışma, bankacılık sektöründe faaliyet gösteren bir firmanın belirli bir departmanda kullanılan RPA süreçlerine odaklanarak yapılmıştır. Çalışmanın kapsamı, yalnızca bankacılıkla ilgili belirli iş süreçlerine uygulanabilir ve bu süreçlerin gereksinimleri ile sınırlıdır. Dolayısıyla, elde edilen sonuçlar, farklı sektörlerdeki veya iş süreçlerindeki RPA uygulamalarına doğrudan genellenemez. Bankacılık sektörüne özgü dilsel yapıların,

düzenlemelerin ve iş akışlarının etkisi altında yapılan analizler, farklı sektörler ile uyumsuzluk gösterebilir. Bu nedenle, farklı sektörlerde yapılan RPA uygulamalarında, modelin başarı oranları, doğruluk seviyeleri ve etkinliği değişkenlik gösterebilir.

Çalışma boyunca kullanılan veri setleri ve metin örnekleri de bankacılık sektörüne özgü mesajlaşmalarla sınırlıdır. Bu, farklı sektörlerdeki veri çeşitliliği ve dilsel yapıların modelin başarısını etkileyebileceği anlamına gelmektedir. Sektörel farklılıklar ve veri farklılıkları, kullanılan NLP ve ML tekniklerinin performansını doğrudan etkileyebilir.

Bu çalışma kapsamında kullanılan veri seti Türkçe kurumsal e-posta içeriklerinden oluşmaktadır. Bu nedenle geliştirilen sınıflandırma modelleri, özellikle Türkçe metin yapıları, ifade biçimleri ve bankacılık sektörüne özgü terimler üzerinde değerlendirilmiştir. Çalışmada kullanılan BERTurk modeli de Türkçe metinler için önceden eğitilmiş bir transformer modeli olduğundan, İngilizce veya farklı dillerde gelen e-posta içeriklerinde model performansı düşebilir. Bu durum, çalışmanın dil açısından bir sınırlılığı olarak değerlendirilmektedir.

Bu kapsamda, bu çalışmanın bulgularının genellenebilirliği, yalnızca bankacılık sektöründe benzer iş süreçlerine sahip olan RPA uygulamalarıyla sınırlı kalacaktır. Bu durum, araştırmanın genellenebilirliği üzerinde sınırlayıcı bir etki oluşturmakta ve çalışmanın sonuçlarının farklı sektörlerde uygulanabilirliğinin ayrıca değerlendirilmesi gerektiğini göstermektedir.

1.6 Tanımlar

Bu tez kapsamında kullanılan temel kavram ve terimlerin açıklamaları aşağıda sunulmaktadır.

RPA: RPA, kuruluşların operasyonel verimliliğini artırmak amacıyla kullanılan bir otomasyon yaklaşımıdır. RPA kapsamında kullanılan robotlar ya da botlar, insan kullanıcıların yazılım sistemleri üzerinde gerçekleştirdiği işlemleri taklit edebilen yazılım ajanlarıdır. Bu sayede veri girişi, kontrol, yönlendirme ve sistemler arası işlem yürütme gibi tekrarlayan görevler yazılım robotları tarafından yerine getirilerek insan iş gücünün üzerindeki operasyonel yük azaltılmaktadır (Syed ve arkadaşları, 2020).

Yapısal Olmayan Veriler: Yapısal olmayan veriler, önceden tanımlanmış bir veri modeline sahip olmayan verilerdir. Bu verilere örnek olarak e-posta içerikleri, sosyal medya paylaşımları, ses kayıtları, videolar, belgeler (PDF, Word dosyaları), resimler ve formatsız metinler verilebilir (Aktan, 2018). Yapısal olmayan verilerin işlenmesi ve anlamlandırılması, genellikle gelişmiş analiz teknikleri gerektirmektedir.

Yapay Zeka: Yapay zeka, makineler tarafından insan benzeri davranışları sergileyebilmesini sağlayan teknolojilerin genel adıdır. Yapay zeka, bilgisayar sistemlerinin insan zekasına benzer şekilde algılama, öğrenme, çıkarım yapma, karar verme ve problem çözme gibi görevleri gerçekleştirebilmesini amaçlayan bir bilgisayar bilimi alanıdır (Lu ve arkadaşları, 2017). Yapay zeka uygulamaları, farklı veri türlerinden anlam çıkararak belirli görevlerin otomatikleştirilmesini ve karar destek süreçlerinin geliştirilmesini sağlamaktadır.

ML (Makine Öğrenmesi, Machine Learning): ML ise yapay zekanın bir alt alanıdır ve makinelerin daha önceki deneyimlerden süreçleri öğrenmesidir (Aylak ve arkadaşları, 2021). Bilgisayar sistemlerinin açıkça programlanmadan verilerden örüntüler öğrenmesini ve bu örüntülere dayanarak tahmin ya da sınıflandırma yapmasını sağlayan yöntemleri kapsamaktadır. ML algoritmaları, geçmiş verilerden öğrenerek yeni veriler üzerinde karar üretme yeteneği kazanmaktadır.

Metin Sınıflandırma: Metin sınıflandırma, metin tabanlı verilerin içeriklerine göre önceden belirlenmiş sınıf veya kategorilere ayrılması işlemidir. Bu süreçte e-posta, yorum, belge veya mesaj gibi metinsel veriler analiz edilerek ilgili sınıfa atanır. Bu tez kapsamında metin sınıflandırma, e-posta içeriklerinin ilgili robotik süreç otomasyonu alanlarına yönlendirilmesi amacıyla kullanılmaktadır.

NLP (Doğal Dil İşleme, Natural Language Processing): NLP, insan dili ile oluşturulan metin veya konuşma verilerinin bilgisayarlar tarafından analiz edilmesini, işlenmesini ve anlamlandırılmasını sağlayan yapay zeka alanıdır (Seker, 2015). Metin temizleme, normalizasyon, kök bulma, vektörleştirme ve sınıflandırma gibi işlemler NLP sürecinin temel adımları arasında yer almaktadır.

TF-IDF (Terim Frekansı-Ters Doküman Frekansı, Term Frequency-Inverse Document Frequency): TF-IDF, metinlerde yer alan kelimelerin önem derecesini belirlemek amacıyla kullanılan istatistiksel bir vektörleştirme yöntemidir. Bu yöntem, bir kelimenin ilgili metin içerisindeki kullanım sıklığını ve tüm veri kümesi içerisindeki ayırt ediciliğini birlikte değerlendirerek metinlerin sayısal olarak temsil edilmesini sağlar.

Derin Öğrenme: Derin öğrenme, yapay sinir ağları temelli çalışan ve büyük veri kümelerinden karmaşık örüntüleri öğrenebilen bir ML yaklaşımıdır (LeCun ve arkadaşları, 2015). Özellikle NLP, görüntü işleme ve ses tanıma gibi alanlarda yüksek başarı göstermektedir. Bu tez kapsamında derin öğrenme yaklaşımı, metin sınıflandırma performansının değerlendirilmesinde karşılaştırmalı olarak ele alınmaktadır.

Model Başarım Ölçütleri: Model başarım ölçütleri, geliştirilen ML modellerinin sınıflandırma performansını değerlendirmek için kullanılan metriklerdir. Doğruluk, kesinlik, duyarlılık ve F1 skoru sık kullanılan model başarım ölçütleri arasında yer almaktadır. Bu ölçütler ile modellerin genel başarısı ve sınıflara göre performansı analiz edilebilmektedir.

2. LİTERATÜR ARAŞTIRMASI

Literatürde, RPA uygulamalarının yaygınlaşmasıyla birlikte özellikle yapısal olmayan verilerin anlamlandırılması ve işlenmesi konusunun önemli bir problem olarak ortaya çıkmaktadır. Yapısal olmayan veriler metin, e-posta, sosyal medya içerikleri ve benzeri kaynaklardan elde edilen, formatsız veriler olarak tanımlanmaktadır. Yapısal olmayan veriler, veri işleme ve analiz süreçlerinde zorluklara neden olmaktadır. RPA süreçleri kural tabanlı çalışır. Kurallara uymayan bu yapılandırılmamış verilerin anlamlandırılması, insan-robot etkileşimini güçlendirecektir (Örpek ve arkadaşları, 2026).

Yapılan detaylı literatür araştırma çalışmaları, yapısal olmayan verilerin işlenmesinde karşılaşılan sorunların çözümü için çeşitli yöntem ve modellerin geliştirildiğini ortaya koymaktadır. Özellikle ML, derin öğrenme ve NLP tekniklerine dayanan yöntemler ile verinin anlamlandırılmasında önemli avantajlar sağlamaktadır. Akademik çalışmalarda, bu yöntemlerin farklı uygulama alanlarında performans değerlendirmeleri yapılarak, problem çözümüne en uygun yöntemlerin geliştirilmesi hedeflenmektedir. Ayrıca literatürde, teknolojinin hızlı gelişimiyle birlikte yeni model ve yöntemlerin kullanılması sayesinde daha başarılı çözümler geliştirilebileceği ve bu alanın gelecekte de araştırılmaya açık olduğu vurgulanmaktadır.

Bu kapsamda, yapısal olmayan verilerin anlamlandırılması ve sınıflandırılması RPA sistemlerinin verimliliği açısından kritik bir gereklilik olarak öne çıkmaktadır.

2.1 Yapısal Olmayan Verilerin Analizi ve Sınıflandırılmasına Yönelik Çalışmalar

Yapılandırılmış veriler ile çalışan araçlar ve yöntemler oldukça yaygındır. Ancak bu araçlar ve yöntemler yapısal olmayan veriler üzerinde başarı ile çalışmamaktadır. Blumberg ve Atre (2003), çalışmalarında, bilgi teknolojileri alanındaki en büyük ve çözülememiş sorunlardan biri olarak yapısal olmayan verilerin yönetimi olduğunu belirtmişlerdir. Çalışmalarında, yapılandırılmış veriler için geliştirilen araçların ve tekniklerin yapısal olmayan veriler üzerinde etkili olmadığı, bu nedenle yapısal olmayan veriler için yeni çözümlere ihtiyaç olduğunu vurgulamışlardır. Ayrıca, yapısal olmayan

verilerin kurumsal veri varlığı içinde önemli bir yer tuttuğunu ifade etmişlerdir. Çalışmada, metadata destekli arama, sınıflandırma, taksonomi oluşturma ve içerik zekası gibi yöntemlerin bu alanda çözüm sağlayabileceği belirtilmiştir (Blumberg ve Atre, 2003). Çalışma, yapısal olmayan verilerin yalnızca depolanması değil, aynı zamanda erişim, sınıflandırma ve yönetim açısından da özel çözümler gerektirdiğini ortaya koymaktadır.

Benzer biçimde, yapısal olmayan verilerin artan önemi ve geleneksel yöntemlerin bu veri türü karşısındaki yetersizliği literatürde farklı çalışmalarda da vurgulanmıştır. Eberendu (2016), sosyal medya, mobil bilişim ve çevrimiçi reklamcılık gibi yeni teknolojilerin ortaya çıkmasıyla birlikte, verilerin geleneksel yapısal formatlardan farklı olarak yarı yapısal veya yapısal olmayan formlarda olduğu ve bu verilerin klasik ilişkisel yöntemler ile işlenemediğini vurgulamıştır. Yapısal olmayan verilerin metin, e-posta, görüntü ve video gibi çeşitli biçimlerde bulunduğu ve bu çeşitlilik ile büyüklüğün arama ve analiz süreçlerinde karmaşıklığa neden olduğu belirtilmiştir. Bu çalışma, yapısal olmayan verilerin kaynaklarını ve bileşenlerini tanımlayıp, organizasyonların bu verileri analiz ederek rekabet avantajı sağladığını ortaya koymaktadır. Çalışma, yapısal olmayan verilerin analiz edilmesiyle kurumların iş süreçlerini iyileştirebileceğini ve yenilikçi çözümler geliştirebileceğini göstermektedir (Eberendu, 2016). Bu çerçevede, yapısal olmayan verilerin toplam verinin büyük bir bölümünü oluşturduğu ve bu verilerin etkili analizinin karar alma süreçleri açısından kritik bir rol oynadığı sonucuna ulaşılmıştır. Yapısal olmayan veriler yalnızca teknik bir veri işleme problemi değil, aynı zamanda kurumların iş süreçlerini geliştirmesi ve yenilikçi çözümler üretmesi açısından stratejik bir değer taşımaktadır.

Yapısal olmayan verilerin analizinde kullanılan yöntemler incelendiğinde, metin madenciliği ile NLP tekniklerinin öne çıktığı görülmektedir. Gharehchopogh ve Khalifehlou (2011), kurumsal bilgi kaynaklarının önemli bir bölümünün yapısal olmayan verilerden oluştuğunu vurgulamıştır. Yapısal olmayan verileri büyük hacme sahip olması söz konusu verilerden anlamlı bilgi çıkarılmasını kurumlar açısından kritik hale getirmektedir. Çalışmada, metin madenciliğinin daha çok istatistiksel yöntemler aracılığıyla büyük hacimli metinlerden desen ve örüntü çıkarımına odaklanmaktadır. NLP tekniklerinin dilin sözdizimsel ve anlamsal yapısını dikkate alarak daha derinlikli

analizler sunduğu belirtilmiştir. Bu yönüyle metin madenciliği ve NLP olmak üzere bu iki yaklaşımın birbirinden bağımsız değil, aksine bilgi çıkarımı süreçlerinde birbirini tamamlayan yapılar olduğu anlaşılmaktadır. Özellikle metin sınıflandırma, belge kümeleme, bilgi çıkarımı, özetleme ve anlam çıkarma gibi alanlarda bu iki yöntemin de birlikte kullanımıyla daha etkili sonuçlara ulaşılabileceği ifade edilmiştir (Gharehchopogh ve Khalifehlou, 2011). Bu çalışma, yapısal olmayan verilerin analizi çalışmalarında tek bir yönteme dayalı çözümler yerine iki tekniğin birlikte değerlendirilmesinin daha başarılı sonuçlar sağlayacağını ortaya koymaktadır.

Yapısal olmayan verilerin analizine yönelik uygulama odaklı çalışmalarda özellikle sosyal medya verilerinin önemli bir araştırma alanı olduğu görülmektedir. Selvi (2016) çalışmasında, büyük hacimli veri olan Twitter mesajlarının analiz edilmesine odaklanmıştır. Çalışmasında, NLP teknikleri ve büyük veri işleme yöntemlerini birlikte kullanmıştır. Twitter'dan elde edilen veriler mevcut yapısal veri setleri ile ilişkilendirilerek kullanıcı davranışlarının belirlenmesi amaçlanmıştır. Bu süreçte metin tabanlı analiz, NLP, olasılıksal hesaplama ve MapReduce gibi yöntemler kullanılarak büyük hacimli verilerin işlenebilir hale geldiği görülmektedir. Özellikle tweet'lerde sık kullanılan kelimelerin belirlenerek ve bu kelimelerin veri tabanında yer alan kategorilerle eşleştirilerek kullanıcıların ilgi alanları tespit edilmiştir. Elde edilen bu bulgular doğrultusunda, benzer ilgi alanlarına sahip kullanıcıları önermeye yönelik bir tavsiye sistemi geliştirilmiştir. Geliştirilen uygulama üzerinden iki binden fazla analiz gerçekleştirilmiştir. Ayrıca uygulama üzerinden kullanıcı geri bildirimleri toplanmıştır. Yapılan testler sonucunda uygulamanın %86 oranında başarılı olduğu sonucuna varılmıştır (Selvi, 2016). Bu çalışma, yapısal olmayan sosyal medya verilerinin NLP ve büyük veri teknolojileri ile analiz edilerek anlamlı kullanıcı davranışlarının ortaya çıkarılabileceğini göstermektedir.

Yapısal olmayan metin verilerinin analizine yönelik uygulama temelli çalışmaların önemli bir bölümü, görüş ve yorum içeriklerinin otomatik sınıflandırılmasına odaklanmaktadır. Göker ve Tekedere (2017) tarafından gerçekleştirilen çalışmada internet ortamında FATİH Projesi'ne ilişkin paylaşılan görüşlerin metin madenciliği yöntemleriyle analiz edilmesi ele alınmıştır. Bu kapsamda, yapısal olmayan metin verilerini işlenebilir hale getirmek amacıyla bir veri ön işleme yazılımı geliştirilmiş ve

elde edilen veriler ML algoritmaları kullanılarak sınıflandırılmıştır. Çalışmada 444 görüşten oluşan veri setinin TF-IDF (Terim Frekansı-Ters Doküman Frekansı, Term Frequency-Inverse Document Frequency) yöntemi ile sayısal vektörlere dönüştürüldüğü ve ardından farklı sınıflandırma algoritmalarının performanslarının karşılaştırıldığı görülmektedir. Elde edilen bulgular, en yüksek başarı oranının %88,73 ile Ardışık Minimal Optimizasyon algoritmasında sağlandığını göstermiştir. Bu sonuç, uygun ön işleme ve öznitelik çıkarımı çalışmalarının ardından yapısal olmayan görüş verilerinin ML yöntemleriyle etkili biçimde sınıflandırılabileceğini göstermektedir (Göker ve Tekedere, 2017). Bu çalışma, metin tabanlı kullanıcı görüşlerinin anlamlandırılması çalışmasında metin madenciliği ve sınıflandırma tekniklerinin uygulanabilir olduğunu ve başarılı sonuçlar sunduğunu göstermektedir.

Yapısal olmayan büyük veri kümelerinin analizine yönelik yapılan çalışmalarda, farklı ML algoritmalarının karşılaştırmalı olarak değerlendirilmesi önemli bir yere sahiptir. Punia ve arkadaşlarının (2021) yaptıkları çalışmada, yapısal olmayan büyük veri kümeleri üzerinde anlamlı bilgiler elde edebilmek amacıyla çeşitli sınıflandırma algoritmalarının performansları incelenmiştir. Veri seti, Twitter API aracılığıyla elde edilmiştir. Veri seti üzerinde karar ağaçları, yapay sinir ağları, SVM (Destek Vektör Makinesi, Support Vector Machine), Naive Bayes, doğrusal regresyon ve k-NN (k-En Yakın Komşu, k-Nearest Neighbor) gibi farklı modeller eğitilmiştir. Çalışma, yapısal olmayan verilerin sınıflandırılmasında tek bir yönteme bağlı kalmak yerine, farklı algoritmaların performanslarının karşılaştırmalı olarak ele alınmasının önemli olduğunu göstermektedir. Elde edilen bulgulara göre en yüksek sınıflandırma doğruluğunun %60 ile SVM algoritmasında sağlanmıştır. İkinci sırada ise %52,5 doğruluk oranı ile Naive Bayes algoritmasının yer almaktadır (Punia ve arkadaşları, 2021). Bu sonuçlar, özellikle büyük veri niteliği taşıyan büyük hacimli yapısal olmayan metinlerin analizinde SVM'in diğer yöntemlere göre daha başarılı performans gösterebildiğini göstermektedir. Bununla birlikte, doğruluk oranlarının sınırlı düzeyde kalması, yapısal olmayan büyük veri kümelerinin sınıflandırılmasının halen üzerinde çalışılması gereken bir alan olduğunu göstermektedir. Bu çalışma, ML algoritmalarının büyük veri analizindeki uygulanabilirliğini ortaya koymaktadır ve gelecekte yapılacak araştırmalar için iyileştirilmesi gereken yönleri vurgulamaktadır.

Metin sınıflandırma alanında gerçekleştirilen çalışmalar, bu problemin yalnızca belirli algoritmaların uygulanmasından ibaret olmadığını ve öznitelik çıkarımı, boyut indirgeme, model seçimi ve başarı değerlendirme süreçlerinin birlikte ele alınması gerektiğini göstermektedir. Kowsari ve arkadaşları (2019) çalışmalarında, metin sınıflandırma problemini incelemiştir. Veri boyutunun arttığı durumlarda metinlerin doğru biçimde sınıflandırılabilmesi için kullanılan temel yöntemleri değerlendirmiştir. Çalışmada, TF, TF-IDF, Word2Vec, GloVe ve FastText gibi öznitelik çıkarım yöntemleri değerlendirilmiştir. PCA, LDA, NMF, Autoencoder ve t-SNE gibi boyut indirgeme teknikleri değerlendirilmiştir. Ayrıca Naive Bayes, k-NN, SVM, Karar Ağaçları, Lojistik Regresyon, RF (Rastgele Orman, Random Forest) ve derin öğrenme temelli yaklaşımlar gibi farklı sınıflandırma algoritmaları ile karşılaştırmalı biçimde analiz edilmiştir (Kowsari ve arkadaşları, 2019). Çalışma, metin sınıflandırma başarısının yalnızca kullanılan algoritmaya bağlı olmadığını göstermektedir. Model başarısının metnin nasıl temsil edildiğine ve veri boyutunun nasıl yönetildiğine de bağlı olduğunu göstermektedir. Bu çalışma, metin sınıflandırma konusunda yöntemsel yapıları bütüncül biçimde değerlendirmesi bakımından önemli bir referanstır.

Metin sınıflandırma alanındaki çalışmalar, yalnızca sınıflandırma algoritmalarının karşılaştırılması çalışmaları değildir. Ek olarak, farklı metin temsil yöntemlerinin model performansı üzerindeki etkisine de odaklanmaktadır. Hendrawan ve arkadaşları (2022) tarafından gerçekleştirilen bu çalışmada, yerel ürün incelemelerinin duygu analizinde Naive Bayes ve XGBoost algoritmaları, TF-IDF ve Word2Vec vektörleştirme yöntemleri ile değerlendirilmiştir. Araştırmada 25.581 veri üzerinde noktalama işaretlerinin kaldırılması, sayı ve gereksiz kelimelerin temizlenmesi, etkisiz kelimelerin çıkarılması, kök bulma ve tokenizasyon gibi kapsamlı ön işleme adımlarının uygulandığı görülmektedir. Ön işleme sonrasında veri setinin %80 eğitim ve %20 test olarak ayrılmasıyla modellerin karşılaştırmalı performans analizi gerçekleştirilmiştir. Elde edilen sonuçlar, XGBoost ve Word2Vec kombinasyonunun 0,89 doğruluk ve 0,941 F1 skoru ile en başarılı performansı sağladığını ortaya koymuştur (Hendrawan ve arkadaşları, 2022). Bu bulgu, sözcükler arasındaki anlamsal ilişkileri daha güçlü biçimde temsil eden vektörleştirme yöntemlerinin sınıflandırma başarısını artırabildiğini göstermektedir. Ayrıca XGBoost algoritmasının, özellikle dengesiz veri kümelerinde Naive Bayes'e göre daha başarılı sonuçlar verebildiği anlaşılmaktadır. Bu çalışma, metin

sınıflandırma problemlerinde yalnızca algoritma seçiminin değil, aynı zamanda veri temsil yönteminin de model başarısı üzerinde belirleyici bir etkiye sahip olduğunu göstermektedir.

Literatürde yapısal olmayan verilerin analizine yönelik çalışmaların genel olarak üç eksende toplandığı görülmektedir. İlk olarak, yapısal olmayan verilerin yönetim ve analiz zorluklarını ele alan kuramsal çalışmalar bulunmaktadır. İkinci olarak, sosyal medya ve kullanıcı görüşleri gibi veri kaynakları üzerinde NLP ve ML tabanlı uygulama çalışmaları yer almaktadır. Üçüncü olarak ise metin sınıflandırmada öznitelik çıkarımı, vektörleştirme ve algoritma seçimi gibi yöntemsel boyutları inceleyen çalışmalar öne çıkmaktadır. Bununla birlikte, mevcut literatürde kurumsal süreçlerde kullanılan e-posta içerikleri gibi alan bağımlı ve operasyonel yapısal olmayan veriler üzerinde gerçekleştirilen çalışmaların daha sınırlı olduğu görülmektedir. Bu nedenle mevcut çalışma, yapısal olmayan verilerin kurumsal RPA süreçleri bağlamında sınıflandırılmasına odaklanarak literatürdeki bu boşluğa katkı sunmayı amaçlamaktadır.

2.2 RPA Süreçlerinde Yapısal Olmayan Verilerin ML ve NLP ile Anlamlandırılmasına Yönelik Çalışmalar

RPA sistemleri başlangıçta ağırlıklı olarak yapılandırılmış veriler ve kural tabanlı iş akışları üzerinde etkili olmuştur. Ancak kurumsal süreçlerde yapısal olmayan verilerin artmasıyla birlikte bu sistemlerin tek başına yetersiz kaldığı görülmüştür. Özellikle e-posta içerikleri, belgeler, müşteri mesajları ve taranmış dokümanlar gibi yapısal olmayan veri kaynaklarının süreçlere yoğun biçimde dahil olması, RPA sistemlerinin NLP ve ML teknikleriyle desteklenmesini gerekli hale getirmiştir. Bu doğrultuda literatürde, RPA'nın yapay zeka yöntemleriyle zenginleştirilmesi sonucunda belge analizi, metin sınıflandırma, bilgi çıkarımı ve süreç yönlendirme gibi daha karmaşık görevlerin otomatikleştirilebildiği görülmektedir.

Bu alandaki çalışmaların bir bölümü, NLP tekniklerinin RPA yetkinliklerini nasıl genişlettiğine odaklanmaktadır. Yadav ve arkadaşları (2025), geleneksel RPA sistemlerinin daha çok yapılandırılmış ve kural tabanlı görevlerde etkin olduğunu, buna karşılık yapısal olmayan veriler söz konusu olduğunda sınırlı kaldığını vurgulamıştır. Çalışmada, NLP'nin RPA sistemlerine entegre edilmesiyle birlikte belge analizi, e-posta

anlayışı ve müşteri etkileşimi gibi dil temelli süreçlerin otomatikleştirilebildiği belirtilmiştir. Ayrıca duygu analizi, sohbet robotları ve veri çıkarımı gibi uygulama alanlarında NLP destekli RPA çözümlerinin süreç verimliliğini artırdığı ve müşteri hizmetlerinde daha hızlı ve kişiselleştirilmiş yanıtlar sunulmasına katkı sağladığı ifade edilmiştir (Yadav ve arkadaşları, 2025). Bu yönüyle çalışma, NLP destekli RPA'nın yalnızca teknik bir otomasyon yaklaşımı olmadığını, aynı zamanda işletmelerin dijital dönüşümünde stratejik bir araç olarak değerlendirilebileceğini göstermektedir.

Benzer biçimde, yapısal olmayan belgelerin otomatik işlenmesinde yapay zeka tabanlı yöntemlerin rolü de literatürde önemli bir yer tutmaktadır. Baviskar ve arkadaşları (2021), belge sınıflandırma problemlerinde NLP, ML ve derin öğrenme tekniklerinin kullanımını incelemiştir. SVM, RF ve BERT (Çift Yönlü Kodlayıcı Temsilleri, Bidirectional Encoder Representations from Transformers) ile GPT gibi transformer tabanlı modellerin bağlamsal analiz ve yüksek doğruluk sağlama açısından öne çıktığı belirtilmiştir. Ayrıca RPA ile yapay zekanın birlikte kullanılmasının belge işleme süreçlerini hızlandırdığı, hata oranlarını azalttığı ve ölçeklenebilirliği artırdığı ifade edilmiştir (Baviskar ve arkadaşları, 2021). Bu bulgular, yapısal olmayan belge işleme süreçlerinde yalnızca otomasyonun değil, öğrenebilen modellerin de süreç başarısı üzerinde belirleyici olduğunu göstermektedir. Dolayısıyla çalışma, RPA ve yapay zeka entegrasyonunun belge tabanlı kurumsal iş akışlarında önemli bir gelişim alanı olduğunu ortaya koymaktadır.

Literatürde bazı çalışmalar ise RPA alanını daha geniş ve sistematik bir çerçevede değerlendirmektedir. Wewerka ve Reichert (2021) çalışmalarında, RPA alanındaki yayınları kapsamlı biçimde inceleyerek 63 çalışmayı yedi tematik kümede sınıflandırmış ve bu doğrultuda ANCOPUR adlı bir değerlendirme çerçevesi geliştirmiştir. Çalışmada RPA'nın tanımı, diğer teknolojilerden farkları, otomasyona uygun süreçlerin seçim kriterleri, kullanım alanları, etkileri, proje yöntemleri ve yapay zeka ile entegrasyonu gibi temel konular ele alınmıştır. Elde edilen bulgular, RPA'nın özellikle tekrarlayan, kural tabanlı ve yüksek manuel çaba gerektiren süreçlerde etkili olduğunu ancak yapay zeka ile entegrasyon alanının halen gelişmekte olan bir araştırma alanı olduğunu göstermektedir (Wewerka ve Reichert, 2021). Bu çalışma, RPA ve yapay zeka ilişkisinin yalnızca teknik bir uygulama meselesi olmadığını, aynı zamanda araştırma boşluklarının ve yöntemsel

ihtiyaçların sistematik biçimde değerlendirilmesini gerektiren bir alan olduğunu ortaya koymaktadır.

RPA ile yapay zeka ve ML'in birlikte kullanımı, özellikle veri bilimi iş akışları ve büyük veri ortamlarında daha görünür hale gelmiştir. Machireddy ve Devapatla (2022), RPA ile yapay zeka ve ML entegrasyonunun bulut tabanlı veri ambarı ortamlarında ölçeklenebilirlik ve verimlilik üzerindeki etkilerini incelemiştir. Çalışmada, yapay zeka modüllerinin veri sınıflandırma, NLP ve görüntü tanıma gibi görevleri yerine getirerek RPA robotlarının yapısal olmayan veya yarı yapılandırılmış verilerle yürütülen süreçleri otomatikleştirmesine olanak sağladığı belirtilmiştir. Bu entegrasyon sayesinde geleneksel RPA'nın ötesine geçilerek daha karmaşık ve bilişsel karar verme gerektiren işlemlerin otomasyonu mümkün hale gelmiştir. Ayrıca doğruluk, hız ve güvenilirlik açısından önemli iyileşmeler sağlandığı, operasyonel maliyetlerin azaldığı ve kaynak kullanımının daha verimli hale geldiği vurgulanmıştır. Bununla birlikte, yüksek kaliteli veri gereksinimi, model şeffaflığı ve çoklu bulut platformlarında entegrasyon gibi sınırlılıkların varlığı da ifade edilmiştir (Machireddy ve Devapatla, 2022). Bu durum, RPA ve yapay zeka entegrasyonunun yüksek potansiyel taşımasına rağmen, uygulama başarısının veri kalitesi ve sistem mimarisi gibi unsurlara da bağlı olduğunu göstermektedir.

RPA ile yapay zeka entegrasyonunu kavramsal ve uygulama boyutunda ele alan bu çalışmaların yanı sıra, yapısal olmayan metinlerin anlamlandırılmasında kullanılan dil modeli temelli yaklaşımlar da literatürde önemli bir yer tutmaktadır.

RPA süreçlerinde yapısal olmayan verilerin anlamlandırılmasında kullanılan yöntemler incelendiğinde, özellikle transformer tabanlı dil modellerinin önemli bir yere sahip olduğu görülmektedir. Devlin ve arkadaşları (2019) tarafından önerilen BERT modeli, transformer mimarisi üzerine inşa edilmiş çift yönlü bağlamsal dil temsilleri öğrenebilen ön eğitimli bir yaklaşım olarak NLP alanında önemli bir dönüm noktası oluşturmuştur. BERT'in metni yalnızca tek yönlü değil, her iki yönden bağlamı dikkate alarak işlemesi, modelin çok sayıda NLP görevinde dönemin en başarılı sonuçlarını elde etmesini sağlamıştır. Çalışmada, modelin hem tam ince ayar hem de özellik çıkarımına dayalı kullanım senaryolarında yüksek performans sergilediği gösterilmiştir (Devlin ve arkadaşları, 2019). Bu yönüyle BERT, yapısal olmayan metinlerin daha doğru temsil

edilmesi ve sınıflandırılması açısından sonraki çalışmalar için güçlü bir kuramsal ve teknik temel oluşturmuştur.

BERT'in metin sınıflandırma alanındaki kullanımını daha ayrıntılı biçimde ele alan çalışmalardan biri de Sun ve arkadaşları (2019) tarafından gerçekleştirilmiştir. Bu çalışmada, BERT'in ince ayar sürecinin metin sınıflandırma performansına etkisi kapsamlı biçimde değerlendirilmiştir. Özellikle [CLS] belirteci üzerinden elde edilen temsilin sınıflandırma katmanı ile uçtan uca eğitilebilmesi, uzun metinlerin 512 token sınırına göre işlenmesi, katman bazlı öğrenme oranı ayarlamaları, catastrophic forgetting riskinin azaltılması ve alan içi ek ön eğitim gibi stratejilerin performans üzerindeki etkileri sistematik olarak incelenmiştir. Elde edilen bulgular, uygun fine-tuning stratejileri ve ek ön eğitim adımlarının farklı veri kümelerinde hata oranlarını anlamlı biçimde düşürdüğünü göstermektedir (Sun ve arkadaşları, 2019). Bu sonuç, transformer tabanlı modellerin yalnızca güçlü bir başlangıç modeli sunmadığını, aynı zamanda görev ve alan özelliklerine göre uygun biçimde uyarlanmasının başarı üzerinde belirleyici olduğunu ortaya koymaktadır.

RPA ile derin öğrenme tabanlı çok kipli yaklaşımların birlikte kullanıldığı çalışmalar da son dönemde dikkat çekmektedir. Cho ve arkadaşları (2023), finans sektöründe yaygın olarak kullanılan yapısal olmayan dokümanların otomatik olarak işlenebilmesi amacıyla geleneksel RPA ile çok kipli akıllı doküman işleme çerçevesini birleştiren bütüncül bir yaklaşım önermiştir. Çalışmada taranmış finansal doküman görüntülerinin önce RPA aracılığıyla iş akışına dahil edildiği, ardından belge türü sınıflandırma ve anahtar bilgi çıkarımı adımlarının transformer tabanlı modeller ile gerçekleştirildiği belirtilmiştir. Özellikle metinsel içerik ile görsel ve yerleşim bilgisini birlikte değerlendiren modellerin, yalnızca tek bir veri türüne dayanan yaklaşımlara kıyasla daha yüksek doğruluk sağladığı vurgulanmıştır. Ayrıca RPA'nın, sınıflandırma sonuçlarına göre ilgili iş sürecini otomatik biçimde tetikleyerek manuel müdahaleyi azalttığı ve arka ofis işlemlerinde önemli zaman tasarrufu sağladığı gösterilmiştir (Cho ve arkadaşları, 2023). Bu bulgular, yapısal olmayan dokümanların otomatik anlaşılmasında RPA destekli çok kipli öğrenme yaklaşımlarının etkili ve uygulanabilir bir çözüm sunduğunu göstermektedir.

Literatür araştırması kapsamında incelenen çalışmalar, yapısal olmayan verilerin analizi, metin sınıflandırma yöntemleri, RPA sistemlerinde yapay zeka entegrasyonu ve

transformer tabanlı dil modelleri açısından değerlendirilmiştir. Bu kapsamda ilgili çalışmaların amaçları, kullanılan yöntemleri ve bu tez çalışmasıyla ilişkileri Tablo 1'de özetlenmiştir.

Tablo 1. İlgili Çalışmaların Özeti

Kaynak	**Odak Noktası**	**Yöntem**	**Bu Çalışma ile İlgisi**
Blumberg ve Atre (2003)	Yapısal olmayan verilerin yönetimi	Sınıflandırma, taksonomi, metadata	Yapısal olmayan verilerin özel yöntemlerle işlenmesi gerektiğini göstermektedir.
Eberendu (2016)	Yapısal olmayan büyük veri	Kavramsal analiz	E-posta gibi verilerin klasik yöntemlerle işlenmesinin zor olduğunu desteklemektedir.
Gharehchopogh ve Khalifehlou (2011)	Metin madenciliği ve NLP	Karşılaştırmalı analiz	NLP'nin yapısal olmayan verilerden anlam çıkarma sürecindeki önemini göstermektedir.
Göker ve Tekedere (2017)	Türkçe metin sınıflandırma	TF-IDF, ML algoritmaları	Türkçe metinlerde ML tabanlı sınıflandırmanın uygulanabilirliğini desteklemektedir.
Punia vd. (2021)	Büyük veri sınıflandırma	SVM, Naive Bayes, k-NN	Klasik ML algoritmalarının karşılaştırmalı kullanımına dayanak sağlamaktadır.
Kowsari vd. (2019)	Metin sınıflandırma yöntemleri	TF-IDF, Word2Vec, SVM, RF, DL	Metin temsil yöntemleri ve algoritma seçiminin önemini göstermektedir.
Hendrawan vd. (2022)	Ürün yorumlarının sınıflandırılması	TF-IDF, Word2Vec, XGBoost	Vektörleştirme yöntemlerinin model başarısına etkisini göstermektedir.

Yadav vd. (2025)	RPA ve NLP entegrasyonu	NLP destekli RPA	RPA süreçlerinde e-posta ve belge anlayışının otomatikleştirilebileceğini göstermektedir.
Baviskar vd. (2021)	Belge işleme	NLP, ML, BERT, RPA	RPA ve yapay zeka entegrasyonunun belge işleme süreçlerine katkısını desteklemektedir.
Wewerka ve Reichert (2021)	RPA literatürü	Sistematik haritalama	RPA ve yapay zeka entegrasyonundaki araştırma boşluğunu ortaya koymaktadır.
Devlin vd. (2019)	BERT modeli	Transformer tabanlı dil modeli	BERTurk modelinin teknik temelini oluşturmaktadır.
Cho vd. (2023)	Finansal doküman işleme	RPA, transformer modeller	Finans sektöründe RPA ve derin öğrenme entegrasyonuna örnek sunmaktadır.

Tüm bu çalışmalar birlikte değerlendirildiğinde, literatürde RPA'da yapısal olmayan verilerin anlamlandırılmasına yönelik araştırmaların üç temel eksende geliştiği görülmektedir.

– İlk olarak, NLP ve yapay zeka entegrasyonunun RPA sistemlerinin yetkinliklerini genişlettiğini ortaya koyan kavramsal ve uygulama odaklı çalışmalar bulunmaktadır.

– İkinci olarak, belge işleme, veri sınıflandırma ve bilgi çıkarımı gibi görevlerde ML ve derin öğrenme tabanlı yöntemlerin RPA ile bütünleştirilmesine yönelik çalışmalar öne çıkmaktadır.

– Üçüncü olarak ise BERT ve benzeri transformer tabanlı modellerin, yapısal olmayan metinlerin bağlamsal olarak daha doğru temsil edilmesini sağlayarak sınıflandırma başarısını artırdığı anlaşılmaktadır.

Bununla birlikte, literatürde RPA süreçlerine doğrudan entegre edilen metin sınıflandırma modellerinin karşılaştırmalı performans analizine dayalı deneysel çalışmalar sınırlıdır. Ayrıca çalışmaların büyük çoğunluğunun açık veri setleri üzerinde gerçekleştirildiği, gerçek kurumsal veri ile yapılan çalışmaların ise daha az olduğu görülmektedir.

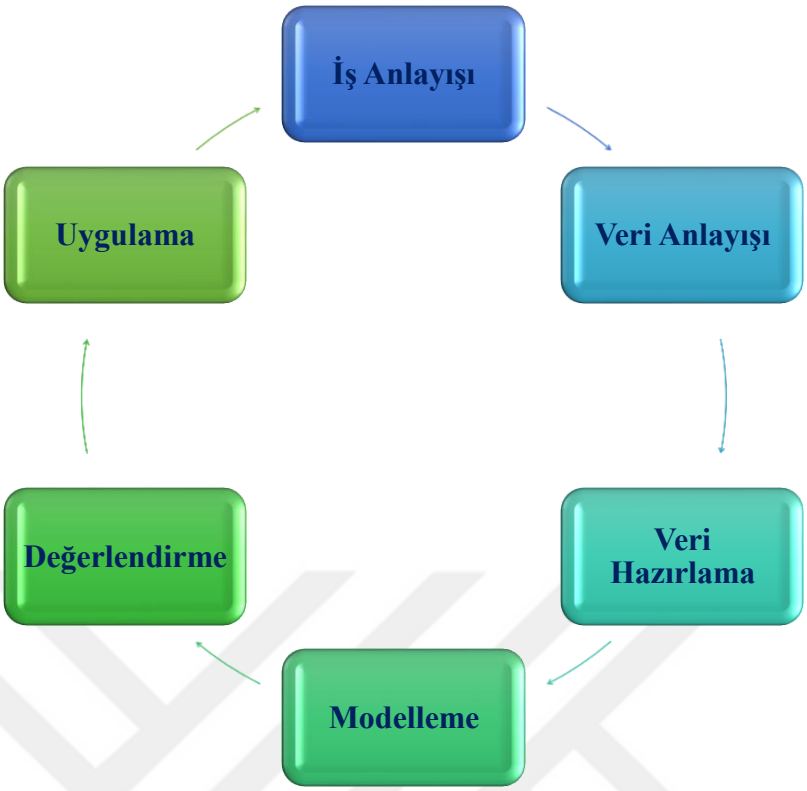
3. YÖNTEM

Bu bölümde, tez çalışmasında izlenen yöntemsel süreç açıklanmaktadır. Yapısal olmayan e-posta verilerinin RPA süreçlerine göre sınıflandırılması amacıyla CRISP-DM metodolojisi temel alınarak iş anlayışı, veri anlayışı, veri hazırlama, modelleme, değerlendirme ve uygulama aşamaları sistematik biçimde açıklanmıştır. Bu doğrultuda veri setinin hazırlanması, metinlerin ön işleme adımlarından geçirilmesi, sayısal temsillerin oluşturulması, benzerlik temelli gruplama yaklaşımı, kullanılan sınıflandırma modelleri ve performans değerlendirme metrikleri ayrıntılı olarak sunulmaktadır. Ayrıca, geliştirilen sınıflandırma modelinin kurumsal RPA süreçlerine bağımsız bir servis olarak entegre edilebilmesine yönelik önerilen sistem mimarisi de bu bölümde ele alınmaktadır.

3.1 Araştırma Tasarımı ve Metodolojik Çerçeve

Bu tez çalışmasında, yapısal olmayan e-posta verilerinin RPA süreçlerine göre sınıflandırılması problemi ele alınırken, veri madenciliği projelerinde yaygın olarak kabul gören CRISP-DM (Veri Madenciliği için Endüstriler Arası Standart Süreç Modeli, Cross Industry Standard Process for Data Mining) metodolojisi, bu çalışmanın metodolojik çerçevesi olarak benimsenmiştir. CRISP-DM, veri madenciliği projelerinin sistematik, tekrarlanabilir ve yönetilebilir biçimde yürütülmesini amaçlamaktadır. Sektör ve teknoloji bağımsız bir süreç modelidir (Wirth ve Hipp, 2000).

CRISP-DM metodolojisi İş Anlayışı, Veri Anlayışı, Veri Hazırlama, Modelleme, Değerlendirme ve Uygulama olmak üzere altı temel aşamadan oluşmaktadır. Bu aşamalar doğrusal bir yapıda ilerlemez. İteratif bir döngü içerisinde ilerlemektedir. Proje süresince önceki adımlara geri dönüşlere olanak tanımaktadır. Şekil 3'te, veri madenciliği projelerinin sistematik biçimde yürütülmesini sağlayan CRISP-DM metodolojisine ait süreç yapısı sunulmaktadır.



Şekil 3. CRISP-DM Metodolojisi

Bu tez kapsamında CRISP-DM süreci, çalışmanın problem yapısına ve uygulama alanı olan RPA süreçlerine uyarlanarak kullanılmıştır. Özellikle:

1. İş Anlayışı aşamasında, RPA robotlarına gelen e-posta içeriklerinin doğru iş sürecine yönlendirilememesi problemi tanımlanmıştır.

2. Veri Anlayışı aşamasında, bankacılık sektörüne ait gerçek operasyonel e-posta verileri incelenmiştir.

3. Veri Hazırlama aşamasında, metin temizleme, anonimleştirme ve ön işleme adımları uygulanmıştır.

4. Modelleme aşamasında, klasik ML algoritmaları ve transformer tabanlı derin öğrenme modelleri kullanılarak karşılaştırmalı analiz gerçekleştirilmiştir.

5. Değerlendirme aşamasında, modeller doğruluk, hassasiyet, geri çağırma ve F1 skoru gibi metrikler üzerinden değerlendirilmiştir.

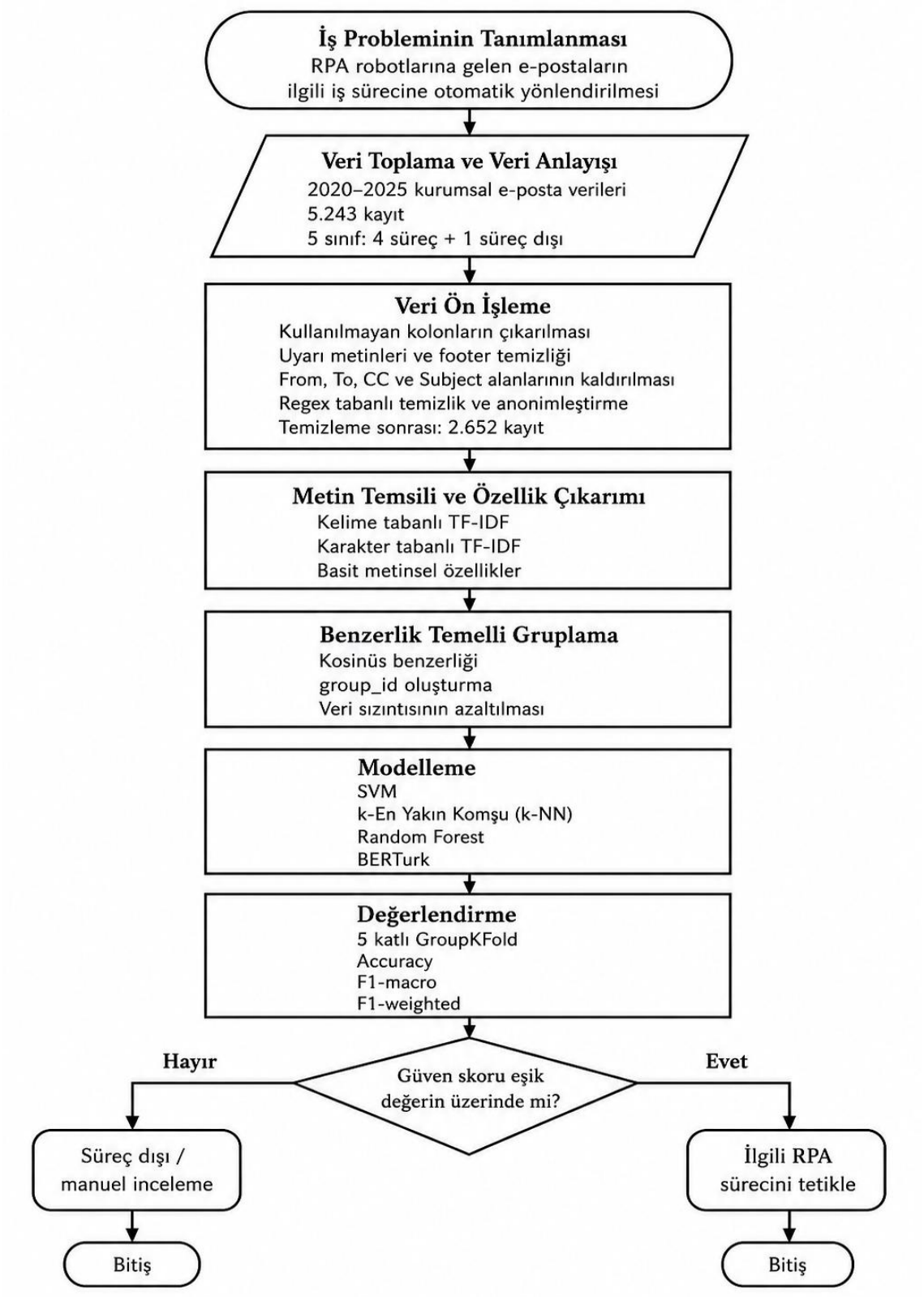
6. Uygulama aşaması ise, elde edilen model çıktılarının RPA süreçlerine entegre edilebilirliği açısından ele alınmıştır.

CRISP-DM'in sunduğu bu esnek yapı sayesinde, veri hazırlama ve modelleme aşamaları arasında sık sık geri dönüşler yapılmıştır. Bu sayede model performansını artırmaya yönelik iyileştirmeler sistematik biçimde gerçekleştirilmiştir.

3.2 CRISP-DM Aşamalarına Göre Yöntemsel Süreç

Bu tez çalışmasında izlenen yöntemsel süreç, CRISP-DM metodolojisi temel alınarak yapılandırılmıştır. Çalışma süreci, iş probleminin tanımlanması, veri setinin anlaşılması, e-posta metinlerinin ön işleme adımlarından geçirilmesi, metinlerin sayısal temsillere dönüştürülmesi, sınıflandırma modellerinin eğitilmesi, model performanslarının değerlendirilmesi ve önerilen sistem mimarisinin oluşturulması aşamalarından oluşmaktadır. Çalışmanın genel işleyişini bütüncül biçimde göstermek amacıyla yöntemsel iş akışı Şekil 4'te sunulmuştur.

Şekil 4'te gösterilen iş akışı mimarisi, tez kapsamında izlenen yöntemsel süreci göstermektedir. Süreç, RPA robotlarına iletilen e-posta verilerinin toplanması ile başlamaktadır. Ardından veri seti incelenmiş, kullanılmayan alanlar çıkarılmış, e-posta içerikleri temizlenmiş ve kişisel veri içerebilecek ifadeler anonimleştirilmiştir. Ön işleme sonrasında metinler TF-IDF tabanlı yöntemlerle sayısal temsillere dönüştürülmüş ve benzer e-posta içeriklerinin tespit edilmesi amacıyla kosinüs benzerliği kullanılarak grup kimlikleri oluşturulmuştur. Modelleme aşamasında SVM, k-NN, RF ve BERTurk modelleri kullanılmıştır. Son aşamada modeller doğruluk, makro F1 ve ağırlıklı F1 metrikleri ile değerlendirilmiştir. Elde edilen sonuçlar doğrultusunda RPA süreçlerine entegre edilebilir önerilen sistem mimarisi oluşturulmuştur.



Şekil 4. Çalışmanın Yöntemsel İş Akış Diyagramı

3.2.1 İş anlayışı

RPA robotlarında çalışan iş süreçlerinin kullanıcılar tarafından otomatik olarak tetiklenebilmesi için, robot kullanıcılarına ait e-posta adreslerine süreç talebini içeren iletiler gönderilmektedir. Bu e-postalar, doğal dilde yazılmış ve ilgili iş sürecini tanımlayan serbest metinlerden oluşmaktadır. Mevcut durumda bu e-posta içeriklerinin manuel olarak incelenmesi zaman kaybına ve operasyonel hatalara yol açabilmektedir. Bu nedenle, doğal dilde oluşturulan e-posta metinlerinden ilgili iş sürecinin otomatik olarak tespit edilmesi ve uygun RPA sürecine yönlendirilmesi gerekmektedir. Bu tez çalışmasının temel iş hedefi, söz konusu e-posta içeriklerinin otomatik ve doğru biçimde sınıflandırılarak RPA süreçlerinin iyileştirilmesidir.

3.2.2 Veri anlayışı

Bu çalışmada kullanılan veri seti, bankacılık sektöründe faaliyet gösteren ve operasyonel süreçlerinde RPA robotlarını aktif olarak kullanan bir finans kuruluşundan elde edilmiştir. Veri seti, 2020-2025 yılları arasında müşteriler ve kullanıcılar tarafından bankanın RPA robotlarına ait e-posta adreslerine gönderilen toplam 5.243 adet e-posta kaydından oluşmaktadır. Veriler kurum içindeki ilgili ekip tarafından belirlenen erişim yetkileri doğrultusunda sistemden dışa aktarılmış ve araştırma amacıyla şirket donanımları üzerinde çalışmalar yürütülmüştür.

Veri setinde yer alan e-posta içeriklerinin uzunlukları incelendiğinde, metinlerin karakter sayısı bakımından oldukça değişken bir yapıya sahip olduğu görülmektedir. En kısa e-posta içeriği 0 karakterden oluşurken, en uzun e-posta içeriği 5677 karakter uzunluğundadır. Bu durum, veri setinde boş veya anlamlı içerik taşımayan kayıtların yanı sıra oldukça uzun ve detaylı açıklamalar içeren e-postaların da bulunduğunu göstermektedir. E-posta uzunluklarındaki bu farklılık, metin sınıflandırma modellerinin performansını etkileyebilecek önemli bir faktördür. Çok kısa metinler sınıflandırma için yeterli bağlamsal bilgi sunmayabilirken, çok uzun metinler gürültü içeren ifadeler, imza alanları veya süreçle doğrudan ilişkili olmayan açıklamalar barındırabilir.

Veri setinde yer alan alanlar ve açıklamaları Tablo 2'de açıklanmaktadır.

Tablo 2. Veri Setinde Yer Alan Alanlar

Alan Adı	**Tanımı**
Proje Kodu	E-postanın ilişkilendirildiği proje, uygulama ya da iş sürecine ait benzersiz tanımlayıcı koddur. Bu alan, kayıtların belirli bir proje veya operasyon kapsamında izlenmesini ve sınıflandırılmasını sağlamaktadır.
Mail Box Owner	İlgili e-posta kutusunun sahibi olan RPA robotu kullanıcısı bilgisini göstermektedir. Bu alan, e-postanın hangi kurumsal posta kutusuna ait olduğunu belirlemek amacıyla kullanılmaktadır.
Mail From	E-postayı gönderen kişi, kullanıcı ya da sistemin adres bilgisini ifade etmektedir. Bu değişken, mesajın kaynağının belirlenmesi ve gönderici bazlı analizlerin yapılabilmesi açısından önem taşımaktadır.
Mail To	E-postanın gönderildiği alıcı kişi, grup veya sistem adresini göstermektedir. Bu alan, mesajın hedef kitlesini veya ulaştırıldığı işlem kanalını tanımlamaktadır.
Kural Durumu	E-postaya ilişkin önceden tanımlanmış iş kuralı, filtre veya yönlendirme mekanizmasının çalışıp çalışmadığını ya da hangi durumda olduğunu gösteren bilgidir. Bu alan, sistem tarafından uygulanan kural tabanlı değerlendirme sonucunun izlenmesini sağlamaktadır.
Mail Başlığı	E-postanın konu satırında yer alan kısa ve özet metinsel ifadedir. Bu alan, mesajın genel içeriği hakkında ilk düzeyde bilgi sunmakta ve metin sınıflandırma süreçlerinde yardımcı öznitelik olarak değerlendirilebilmektedir.

Mail İçeriği	E-postanın gövde kısmında yer alan asıl metinsel içeriği ifade etmektedir. Bu alan, çalışmanın temel analiz birimini oluşturmakta olup NLP ve metin sınıflandırma uygulamalarında ana veri kaynağı olarak kullanılmaktadır.
Ek Gönderimi	E-postada ek dosya bulunup bulunmadığını gösteren bilgidir. Bu değişken, mesajın yalnızca metin tabanlı mı yoksa dosya destekli mi iletildiğini belirlemek amacıyla kullanılmaktadır.
Ek Dosya Adı	E-postaya eklenen dosya veya dosyaların ad bilgisini ifade etmektedir. Bu alan, ek içeriklerin türü, bağlamı veya işlemle ilişkisi hakkında ek bilgi sağlayabilmektedir.
Ek Dosya Sayısı	E-postada yer alan toplam ek dosya adedini göstermektedir. Bu değişken, mesajın içerik yoğunluğu ve ek belge kullanım düzeyinin nicel olarak değerlendirilmesine olanak tanımaktadır.
İlgili Süreç Adı	E-postanın içerik veya bağlam bakımından ilişkilendirildiği iş süreci, operasyon veya RPA sürecinin adını ifade etmektedir. Bu alan, sınıflandırma çalışmalarında hedef değişken olarak kullanılmakta ve her bir e-postanın hangi sürece ait olduğunu göstermektedir.

Kaynak veri dosyasında yer alan alanlar: Proje Kodu, Mail Box Owner, Mail From, Mail To, Kural Durumu (Kural Var / Kural Yok), Mail Başlığı, Mail İçeriği, Ek Gönderimi, Ek Dosya Adı, Ek Dosya Sayısı ve İlgili Süreç Adı şeklindedir. Bu başlıklar, hem e-postaların yapısal özelliklerinin hem de içerik odaklı bilgilerin bütüncül bir şekilde değerlendirilmesine olanak sağlamaktadır.

Kurumsal iş akışları kapsamında RPA robotları, belirli kurallar çerçevesinde gelen e-postaları otomatik olarak işleyerek 4 temel süreç ve 1 süreç dışı olmak üzere toplam 5 sınıfa yönelik aksiyon almaktadır. Ancak ilgili e-postaların eksik, hatalı veya yanlış gönderilmesi durumunda robotlar süreci başlatamamakta, bu da operasyonel akışta kesintilere neden olmaktadır. Bu çalışma kapsamında amaçlanan, e-postanın gövde

metninden hareketle ilgili sürecin otomatik olarak tahmin edilmesidir. Böylece kullanıcıların belirli e-posta kurallarına tam uyum göstermediği durumlarda dahi sürecin doğru şekilde tetiklenmesi mümkün hale gelmektedir.

Veri seti yalnızca 5 sınıf içermektedir. Sınıflar, kurum içinde en sık kullanılan ve operasyonel olarak kritik olan süreçler arasından uzman görüşü ile belirlenmiştir ve bu sınıflar bankanın belirli RPA süreçleri ile doğrudan ilişkilidir. Bu durum, veri setinin sınırlılıkları arasında değerlendirilebilir. Çünkü model, yalnızca tanımlı 5 sınıfa yönelik tahmin yapabilmektedir. Ek olarak, süreç çeşitliliğinin sınırlı olması, sınıflandırma modelinin genelleme kapasitesini azaltabilmektedir. Bununla birlikte, veri seti bankacılık sektörüne özgü gerçek operasyonel verilerden oluştuğu için, çalışma pratik uygulama açısından yüksek derecede gerçeklik taşımaktadır.

Veri setinde yer alan süreçler ve her bir sürece ait detaylı süreç açıklamaları Tablo 3'te sunulmaktadır.

Tablo 3. Veri Setinde Yer Alan Süreç Bilgileri

No	Süreç Adı	Süreç Açıklaması
1	DBS Alıcı Kodu Tanımlanması	Bankanın Doğrudan Borçlandırma Sistemi (DBS) kapsamında, firmalara ait alıcı kodlarının tanımlanması ve güncellenmesine yönelik taleplerin işlenmesini sağlayan süreçtir.
2	Firma Maaş Ödemelerinin Yapılması	Kurumsal müşterilere ait personel maaş ödemelerinin toplu olarak hazırlanması, doğrulanması ve ilgili hesaplara aktarılmasını kapsayan otomatik işlem sürecidir.
3	Organizasyonel Değişiklik Bildirimlerinin Yapılması	Kurum içi birim, pozisyon veya yetki değişikliklerinin bildirimlerinin alınması ve ilgili sistem kayıtlarının güncellenmesini sağlayan süreçtir.

4	İptal Edilen Faturaların Takibinin Yapılması	Firmalar tarafından iptal edilen veya geçersiz kılınan faturaların kontrol edilmesi, ilgili kayıtların güncellenmesi ve gerekli bildirimlerin yapılmasını içeren süreçtir.
5	Süreç Dışı	Tanımlı RPA süreçlerinden herhangi biriyle ilişkilendirilemeyen e-postaların sınıflandırılması için kullanılan kategoridir.

Veri setinde yer alan süreç sınıflarının e-posta içerikleriyle nasıl temsil edildiğini göstermek amacıyla, anonimleştirilmiş örnek kayıtlar Tablo 4'te sunulmuştur. Kurumsal gizlilik ve kişisel verilerin korunması nedeniyle örneklerde gerçek kişi adı, firma adı, e-posta adresi, müşteri numarası, hesap bilgisi ve dosya adı gibi tanımlayıcı bilgilere yer verilmemiştir. Örnek kayıtlar, veri setinin genel yapısını göstermek amacıyla sadeleştirilmiştir.

Tablo 4. Veri Setinden Anonimleştirilmiş Örnek Kayıtlar

No	**Süreç Adı**	**Anonimleştirilmiş Örnek Veri**
1	DBS Alıcı Kodu Tanımlanması	Proje Kodu: PRJ_XX Kural Durumu: Kural Var Mail Başlığı: DBS alıcı kodu tanımlama talebi Mail İçeriği: Merhaba, *** için DBS alıcı kodu tanımlanmasını rica ederiz. Gerekli bilgiler ekte iletilmiştir. Ek Gönderimi: Var Ek Dosya Sayısı: 1 İlgili Süreç Adı: DBS Alıcı Kodu Tanımlanması
2	Firma Maaş Ödemelerinin Yapılması	Proje Kodu: PRJ_XX Kural Durumu: Kural Var Mail Başlığı: Maaş ödeme dosyası hk.

		Mail İçeriği: Merhaba *** Hanım, *** personel maaş ödeme dosyası ekte yer almaktadır. Ödemelerin işleme alınmasını rica ederiz. Ek Gönderimi: Var Ek Dosya Sayısı: 1 İlgili Süreç Adı: Firma Maaş Ödemelerinin Yapılması
3	Organizasyonel Değişiklik Bildirimlerinin Yapılması	Proje Kodu: PRJ_XX Kural Durumu: Kural Yok Mail Başlığı: Organizasyon değişikliği bildirimi Mail İçeriği: Merhabalar, *** biriminde yapılan *** değişikliği için organizasyon bilgisinin güncellenmesini rica ederiz. Ek Gönderimi: Yok Ek Dosya Sayısı: 0 İlgili Süreç Adı: Organizasyonel Değişiklik Bildirimlerinin Yapılması
4	İptal Edilen Faturaların Takibinin Yapılması	Proje Kodu: PRJ_XX Kural Durumu: Kural Var Mail Başlığı: İptal fatura kontrolü Mail İçeriği: Merhaba *** Bey, iptal edilen faturalara ilişkin listenin kontrol edilerek sistem kayıtlarının güncellenmesini rica ederiz. Ek Gönderimi: Var Ek Dosya Sayısı: 1 İlgili Süreç Adı: İptal Edilen Faturaların Takibinin Yapılması
5	Süreç Dışı	Proje Kodu: Yok Kural Durumu: Kural Yok Mail Başlığı: Yıllık izin bilgilendirmesi

Mail İçeriği: Merhaba, ***-*** tarihleri arasında yıllık izinde olacağım. Bu süre içerisinde acil konular için *** *** ile iletişime geçebilirsiniz. Bilgilerinize sunarım.

Ek Gönderimi: Yok

Ek Dosya Sayısı: 0

İlgili Süreç Adı: Süreç Dışı

Tablo 4'te verilen örneklerde, modelleme sürecinde temel metinsel girdi olarak "Mail İçeriği" alanı kullanılmıştır. "İlgili Süreç Adı" alanı ise sınıflandırma modelinin tahmin etmeye çalıştığı hedef değişkeni temsil etmektedir. Diğer alanlar veri setinin yapısını açıklamak ve e-posta kayıtlarının kurumsal süreçlerle ilişkisini göstermek amacıyla sunulmuştur.

3.2.3 Veri hazırlama

Metin sınıflandırma çalışmalarında veri ön işleme aşaması, model performansını doğrudan etkileyen temel adımlardan biri olarak değerlendirilmektedir. Literatürde metin verilerinin temizlenmesi, gereksiz karakterlerin çıkarılması, sayısal temsillere dönüştürülmesi ve sınıflandırma modelleri için uygun hale getirilmesi yaygın olarak kullanılan ön hazırlık adımları arasında yer almaktadır (Kowsari ve arkadaşları, 2019). Benzer şekilde bir diğer çalışmada ise metin sınıflandırma çalışmalarında noktalama işaretlerinin kaldırılması, sayıların temizlenmesi, gereksiz kelimelerin çıkarılması, kök bulma ve tokenizasyon gibi ön işleme adımlarının model başarısı üzerinde etkili olduğunu belirtmiştir (Hendrawan ve arkadaşları, 2022).

Bu tez çalışmasında da literatürde yer alan temel veri ön işleme yaklaşımlarından yararlanılmıştır. Ancak kullanılan veri setinin kurumsal RPA e-posta içeriklerinden oluşması nedeniyle, klasik metin temizleme adımlarına ek olarak deneysel sürece özgü bazı işlemler uygulanmıştır. Bu kapsamda e-posta başlıklarında yer alan otomatik uyarı metinleri, footer alanları, e-posta meta verileri, süreçle ilişkili olmayan ifadeler ve kişisel veri içerebilecek kalıplar temizlenmiştir. Ayrıca benzer e-posta içeriklerinden kaynaklanabilecek veri sızıntısı riskini azaltmak amacıyla kosinüs benzerliğine dayalı grup kimliği oluşturma yaklaşımı uygulanmış ve değerlendirme süreci GroupKFold yapısı ile desteklenmiştir. Böylece veri ön işleme süreci, yalnızca metin temizleme adımı

olarak değil, model performansının daha güvenilir biçimde ölçülmesini sağlayan deneysel sürecin önemli bir bileşeni olarak ele alınmıştır.

Model geliştirme sürecinde kullanılacak verilerin analiz için uygun hale getirilmesi amacıyla kapsamlı bir veri ön işleme aşaması gerçekleştirilmiştir. Bu süreçte, e-posta içeriklerinin temizlenmesi, gereksiz metinlerin kaldırılması ve kişisel verilerin anonimleştirilmesi için çalışmalar yapılmıştır. Uygulanan veri ön işleme adımları aşağıda detaylandırılmaktadır:

3.2.3.1 Kullanılmayan kolonların çıkarılması

Ham veri setinde yer alan ancak sınıflandırma görevine katkı sağlamayan sütunlar belirlenerek veri setinden çıkarılmıştır. Böylece analiz yalnızca süreci temsil eden içerik üzerinde odaklanacak şekilde sadeleştirilmiştir.

3.2.3.2 E-posta başlıklarında yer alan uyarı metinlerinin temizlenmesi

E-postalarda otomatik olarak eklenen ve tüm kayıtlarda tekrar eden güvenlik uyarıları veri setinden kaldırılmıştır. Örneğin, şirket dışı gönderim uyarıları veya şüpheli e-posta bildirim açıklamaları metinlerden tamamen temizlenmiştir. Bu ifadeler modelin öğrenmesine katkı sağlamadığı gibi gereksiz tekrar niteliği taşımaktadır.

3.2.3.3 E-posta Footer bölümlerinin temizlenmesi

E-postaların sonunda yer alan gizlilik uyarıları, hukuki bildirimler ve kurumsal açıklamalar metinlerden çıkarılmıştır. Bu uzun ve tekrarlayıcı metinler süreçle ilişkili bilgi içermediğinden ön işleme sürecinde elimine edilmiştir.

3.2.3.4 E-posta meta verilerinin ön işleme sürecinde kaldırılması

E-postalara ait gönderici (From), alıcı (To), bilgi (CC) ve konu (Subject) alanları, sınıflandırma göreviyle doğrudan ilişkili bilgi içermemeleri ve kişisel veri niteliği taşıyabilmeleri nedeniyle veri ön işleme aşamasında veri setinden çıkarılmıştır. Bu işlem, modelin yalnızca e-posta içeriklerine odaklanmasını sağlamak ve kişisel verilerin korunmasına yönelik gereklilikleri karşılamak amacıyla gerçekleştirilmiştir.

3.2.3.5 Süreçle ilgili olmayan metinlerin temizlenmesi

E-posta gövdesinde yer alan imza, otomatik yanıt, tekrar eden bilgilendirme kalıpları ve sınıflandırma görevine katkı sağlamayan standart ifadeler temizlenmiştir. Ancak tanımlı RPA süreçlerinden herhangi biriyle ilişkilendirilemeyen ve "Süreç Dışı" sınıfını temsil eden kayıtlar veri setinden tamamen çıkarılmamış; bu kayıtlar modelin süreç dışı içerikleri ayırt edebilmesi amacıyla ilgili sınıf etiketi altında korunmuştur. Bu ayıklama işlemi, modelin yalnızca ilgili süreci tanımlayan içerikler üzerinde eğitilmesini sağlarken, süreç dışı örneklerin de ayrı bir sınıf olarak öğrenilmesine olanak tanımıştır.

3.2.3.6 Regex tabanlı temizlik işlemleri

Düzenli ifadeler (regex) kullanılarak, metinlerde yer alan T.C. kimlik numarası kalıpları, telefon numaraları, tarih ve saat ifadeleri, URL bağlantıları, sistem tarafından üretilen kodlar ve diğer özel biçimsel yapılar otomatik olarak tespit edilip metinlerden ayıklanmıştır. Bu işlem, veri setindeki yapısal formatta olan gereksiz verilerin azaltılmasına ve analiz sürecinde kullanılacak metinlerin daha tutarlı ve standartlaştırılmış bir formata dönüştürülmesine katkı sağlamıştır.

Bütün bu ön işleme çalışmaları sırasında, açık kaynaklı Python dilinde yaygın olarak kullanılan pandas, re (Regex Modülü), NumPy ve görselleştirme için matplotlib kütüphaneleri kullanılmıştır. Kullanılan kütüphaneler ve açıklamaları Tablo 5'teki gibidir.

Tablo 5. Ön İşleme Çalışmalarında Kullanılan Kütüphaneler

Kütüphane	**Açıklama**
pandas	Veri yönetimi, kolonların düzenlenmesi, eksik veya hatalı kayıtların ayıklanması ve metinsel alanların toplu biçimde işlenmesi için temel araç olarak kullanılmıştır.
re	Metinlerdeki kalıpsal yapıların (telefon numarası, T.C. kimlik numarası, tarih-saat ifadeleri, URL bağlantıları vb.) otomatik olarak tespit edilmesi ve temizlenmesi amacıyla kullanılmıştır.

NumPy	Veri işlemleri, dizisel işlemler ve metin temsillerinin sayısal yapılara dönüştürülmesi gibi performans gerektiren hesaplamalarda kullanılmıştır.
matplotlib	Çeşitli grafiksel görselleştirmeler oluşturulması için kullanılmıştır. Bu görselleştirmeler, veri setinin yapısının daha bütüncül biçimde değerlendirilmesini sağlamıştır.

Bu açık kaynak kütüphanelerin birlikte kullanılması sayesinde veri seti sistematik olarak işlenerek temizlenmiş ve ML modellerinin eğitimi için uygun bir yapıya dönüştürülmüştür.

Veri ön işleme işlemleri öncesi ve sonrasında veri sayıları Tablo 6'daki gibidir.

Tablo 6. Veri Ön İşleme Sonucu Veri Sayıları

Ön İşlemeden Önce Veri Sayısı	**Ön İşlemeden Sonra Veri Sayısı**
5243	2652

Tablo 6'da veri ön işleme süreci öncesinde ve sonrasında veri setinde meydana gelen niceliksel değişim sunulmaktadır. Ön işleme aşamasında sınıflandırma görevine katkı sağlamayan kolonların çıkarılması, tekrarlı ve süreçle ilişkisi bulunmayan metin parçalarının temizlenmesi, otomatik eklenen uyarı ifadeleri ile e-posta imza ve gizlilik bildirimlerinin metinlerden ayıklanması ve kişisel veri içerebilecek kalıpların düzenli ifadeler (regex) aracılığıyla giderilmesi işlemleri uygulanmıştır. Bu işlemler sonucunda, veri setinde yer alan gürültü niteliğindeki içerikler temizlenmiştir. Ancak süreçleri temsil eden anlamlı metinler korunmuştur. Ön işleme sürecinde tekrar eden, anlamsız ve süreçle ilişkili olmayan kayıtların çıkarılması nedeniyle veri seti yaklaşık %49,4 oranında azaltılmıştır. Tablo 6, ham veriden analiz edilebilir metin temsiline geçişte veri setinin daha tutarlı ve model eğitimi açısından daha uygun bir yapıya dönüştürüldüğü durumdaki sayılarını içermektedir. Bu veri ön işleme çalışmalarının ML modellerinin sınıflandırma performansını olumlu yönde etkilemesi beklenmektedir.

Bu çalışmada, e-posta içeriklerinin ML modelleri tarafından işlenebilir bir forma dönüştürülmesi amacıyla metinlerin sayısal temsilleri oluşturulmuştur. Bu kapsamda,

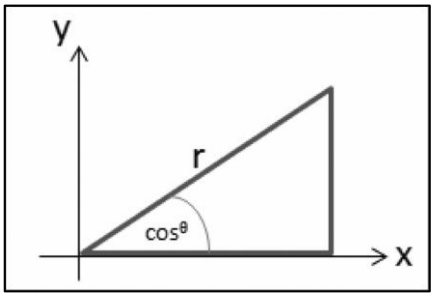
istatistiksel metin temsil yöntemleri kullanılarak e-posta içerikleri vektörel biçimde ifade edilmiştir.

İlk aşamada, e-posta metinlerindeki kelime ve kelime gruplarının önem derecelerini hesaplamak amacıyla TF-IDF yöntemi uygulanmıştır. Bu yaklaşımda her bir e-posta, unigram (tekli) ve bigram (ikili) düzeyindeki terim frekanslarına dayalı olarak sayısal bir vektörle temsil edilmiştir. TF-IDF yöntemi, tüm dokümanlarda sıkça geçen ancak ayırt ediciliği düşük olan kelimelerin etkisini azaltırken, belirli e-postalar için ayırt edici olan terimlerin ağırlığını artırarak sınıflandırma sürecine istatistiksel bir temel sağlamaktadır.

Kelime düzeyindeki temsile ek olarak, e-posta metinlerinin biçimsel ve karakter düzeyindeki benzerliklerini yakalayabilmek amacıyla karakter tabanlı TF-IDF temsilleri de kullanılmıştır. Bu yaklaşımda, kelimelerin iç yapısı ve yazım biçimleri dikkate alınarak belirlenen karakter n-gramları üzerinden vektörler oluşturulmuştur. Karakter tabanlı temsil, özellikle yazım farklılıkları, kısaltmalar ve küçük biçimsel varyasyonların bulunduğu kurumsal e-posta metinlerinde tamamlayıcı bir özellik seti sunmaktadır.

Metinlerin sayısal temsillerinin oluşturulmasının ardından, e-posta içerikleri arasındaki benzerlik ilişkilerinin belirlenmesi amacıyla kosinüs benzerlik analizi gerçekleştirilmiştir. Kosinüs benzerliği, iki metin vektörü arasındaki açının kosinüsünü hesaplayarak içerikler arasındaki göreli yakınlığı ölçmektedir. Bu analiz sayesinde, anlamsal ve yapısal açıdan yüksek benzerlik gösteren e-posta içerikleri nicel olarak tespit edilmiştir.

Kosinüs benzerliği, iki metin vektörü arasındaki açının kosinüsünü hesaplayarak benzerliği ölçmektedir. Şekil 5'te gösterildiği üzere, vektörler arasındaki açı küçüldükçe metinler arasındaki benzerlik artmaktadır (Lahitani ve arkadaşları, 2016).

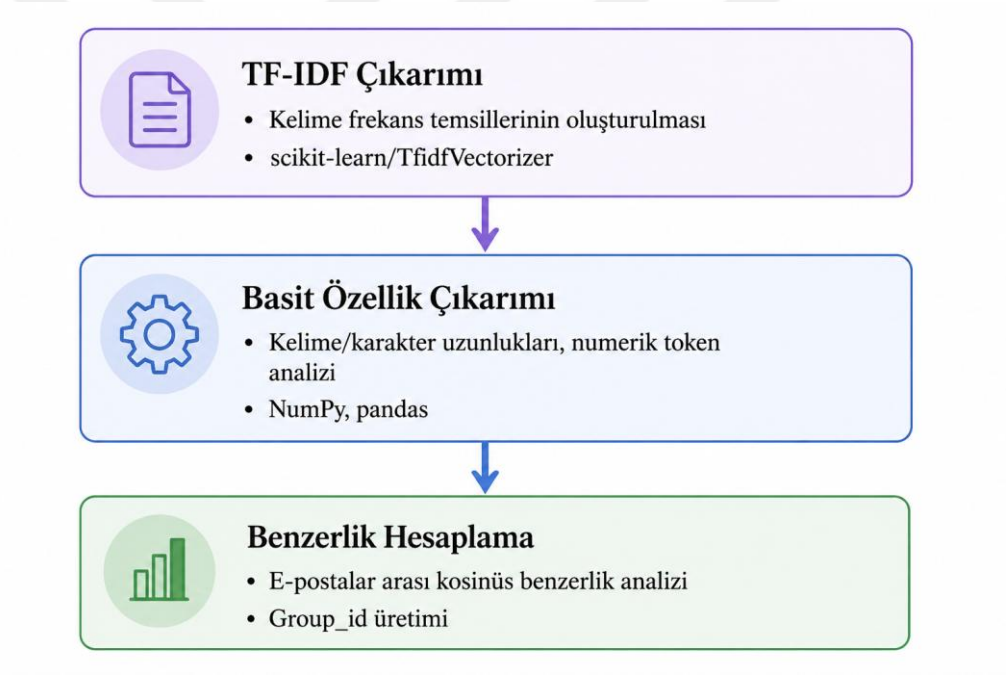


Şekil 5. Benzerlik Kavramı İçin Kosinüs Derecesi

Hesaplanan benzerlik değerleri kullanılarak, aynı süreci temsil eden ve yüksek derecede benzer içeriklere sahip e-postalar grup kimlikleri (group_id) altında birleştirilmiştir. Bu gruplama yaklaşımı, benzer metinlerin eğitim ve test kümelerine aynı anda düşmesini engelleyerek veri sızıntısı riskini azaltmayı amaçlamaktadır. Elde edilen grup yapısı, sonraki aşamalarda uygulanan GroupKFold çapraz doğrulama sürecinde kullanılmıştır.

Uygulanan bu sayısal temsil ve benzerlik temelli gruplama adımları, e-posta verilerinin daha tutarlı ve genellenebilir bir biçimde modellenmesine katkı sağlamıştır. Sınıflandırma modellerinin gerçekçi ve güvenilir biçimde değerlendirilmesine altyapı oluşturmuştur.

Şekil 6'da e-posta metinlerinin sınıflandırma sürecine hazırlanmasında izlenen NLP adımları ve bu adımlarda kullanılan temel kütüphaneler gösterilmektedir.



Şekil 6. E-posta Metinlerinin İşlenmesi Adımları ve İlgili Kütüphaneler

Şekil 6'da görüldüğü üzere, ilk aşamada TF-IDF çıkarımı ile metinlerin kelime frekans temsilleri oluşturulmuş, ardından basit özellik çıkarımı kapsamında kelime ve karakter uzunlukları ile numerik token analizi gerçekleştirilmiştir. Son aşamada ise e-postalar arasındaki içerik benzerlikleri kosinüs benzerliği yöntemi ile hesaplanmış ve benzer kayıtların gruplanabilmesi amacıyla grup kimlikleri üretilmiştir. Bu yapı, metinlerin

sayısal olarak temsil edilmesini ve benzer içeriklerin tespit edilerek modelleme sürecinin daha güvenilir biçimde yürütülmesini sağlamaktadır.

Veri hazırlama süreci kapsamında, e-posta metinleri ML modelleri tarafından işlenebilir sayısal temsillere dönüştürülmüştür. İstatistiksel metin temsilleri, temel biçimsel özellikler ve kosinüs benzerliğine dayalı ilişki çıkarımı adımları sistematik biçimde uygulanmıştır. Uygulanan TF-IDF tabanlı sayısallaştırma ve benzerlik temelli gruplama yaklaşımı, veri setinde yer alan benzer içeriklerin daha gerçekçi biçimde değerlendirilmesine olanak sağlamış ve modelleme aşamasında veri sızıntısı riskinin azaltılmasına katkıda bulunmuştur. Bu bütüncül veri hazırlama yaklaşımı, sonraki aşamalarda kullanılan sınıflandırma modellerinin daha genellenebilir ve güvenilir biçimde eğitilmesi için sağlam bir temel oluşturmuştur.

3.2.4 Modelleme

Çalışmada klasik ML modellerinin değerlendirilmesinde, benzer e-posta içeriklerinin eğitim ve test kümelerine aynı anda düşmesi ile oluşabilecek veri sızıntısını azaltmak amacıyla benzerlik temelli gruplama yaklaşımı uygulanmıştır. Böylece değerlendirme aşamasında veri bölme işlemi satır bazında değil, grup bazında gerçekleştirilmiştir. Deneylerde veri setindeki 2652 kayıt için yaklaşık 259 grup elde edilmiş ve performans raporlamasında grup tabanlı çapraz doğrulama tercih edilmiştir.

Bu kapsamda, metinlerin char-TF-IDF temsilleri üzerinden kosinüs benzerliği hesaplanmış ve belirlenen eşik değerinin (0.85) üzerinde benzer bulunan örnekler aynı group_id altında birleştirilmiştir. Kosinüs benzerliği, metin vektörleri arasındaki benzerliği ölçmek için yaygın olarak kullanılan bir yöntemdir. Viggiato ve arkadaşları (2023) çalışmalarında, sayısal temsil edilen test senaryoları için kosinüs benzerliği kullanmış ve 0.85 eşik değerinin en iyi performansı verdiğini belirtmiştir (Viggiato ve arkadaşları, 2023). Bu tez çalışmasında da 0.85 eşik değeri, birbirine yüksek derecede benzeyen e-posta içeriklerini aynı group_id altında toplarken farklı içeriklerin hatalı biçimde aynı gruba alınmasını azaltmak amacıyla literatürdeki ilgili çalışmalar ve deneysel gözlemler doğrultusunda tercih edilmiştir.

Model performanslarının daha gerçekçi biçimde ölçülebilmesi için çapraz doğrulama sürecinde GroupKFold yaklaşımı kullanılmıştır. Bu yaklaşım, metin verilerinde sık görülen veri sızıntısı problemini önlemek açısından kritik öneme sahiptir. Bu yöntemle önce her grup için baskın sınıf etiketi belirlenmiş, ardından ayrım grup düzeyinde yapılarak fold'larda sınıf dağılımının dengeli kalması hedeflenmiştir. Bu sayede hem grupların eğitim/test arasında bölünmesi engellenmiş hem de yalnızca tek sınıfın test setine düşmesi gibi olası dengesizliklerin önüne geçilmiştir. Her model, 5-katlı çapraz doğrulama ile değerlendirilmiş ve sonuçlar doğruluk, makro F1 ve ağırlıklı F1 metrikleri ile raporlanmıştır.

Bu çerçevede, klasik yöntemler olarak SVM, k-NN ve RF modelleri değerlendirilmiştir. Ek olarak, metinlerin bağlamsal olarak temsil edilmesini sağlayan transformer tabanlı bir yaklaşım olarak BERTurk modeli de modelleme sürecine dahil edilerek klasik yöntemlerle karşılaştırılmıştır.

Bu tez kapsamında gerçekleştirilen tüm deneysel çalışmalar, veri setinin kullanıldığı şirket bünyesinde yer alan 4 adet NVIDIA RTX A6000 GPU destekli kurumsal sunucu ortamında yürütülmüştür. Kullanılan donanım altyapısının yüksek işlem gücü ve bellek kapasitesi sayesinde veri hazırlama, model eğitimi, hiperparametre denemeleri ve performans değerlendirme süreçleri daha hızlı gerçekleştirilmiştir. Özellikle transformer tabanlı modellerin yüksek hesaplama maliyeti dikkate alındığında, 4 adet NVIDIA RTX A6000 GPU destekli ortam deneysel sürecin kısa sürede tamamlanmasına önemli katkı sağlamıştır. Bu altyapı sayesinde modelleme denemeleri daha etkin şekilde yürütülmüş ve farklı algoritmaların karşılaştırmalı olarak değerlendirilmesi mümkün olmuştur. Böylece çalışmanın deneysel aşamaları güçlü bir donanım ortamında gerçekleştirilerek elde edilen sonuçların sistematik biçimde analiz edilmesi sağlanmıştır.

Şekil 7'de yer alan kod parçasında, klasik ML modellerinin eğitim ve değerlendirme sürecine ilişkin genel yapı gösterilmektedir. Bu süreçte e-posta metinleri öncelikle TF-IDF yöntemi ile sayısal vektörlere dönüştürülmüş, ardından SVM, k-NN ve RF modelleri aynı değerlendirme yapısı içerisinde eğitilmiştir. Benzer e-posta içeriklerinin eğitim ve test kümelerine aynı anda düşmesini engellemek amacıyla GroupKFold çapraz doğrulama yöntemi kullanılmıştır. Model performansları doğruluk, makro F1 ve ağırlıklı F1 ölçütleri üzerinden hesaplanmıştır. Kod parçasında yer alan x değişkeni temizlenmiş

e-posta metinlerini, y değişkeni sınıf etiketlerini, group_id değişkeni ise benzerlik temelli oluşturulan grup bilgisini ifade etmektedir.

```
models = {
    "SVM": LinearSVC(class_weight="balanced", C=0.7),
    "k-NN": KNeighborsClassifier(
        n_neighbors=15,
        metric="cosine",
        weights="distance"
    ),
    "Random Forest": RandomForestClassifier(
        n_estimators=300,
        random_state=42,
        class_weight="balanced"
    )
}

for model_name, model in models.items():
    pipeline = Pipeline([
        ("tfidf", TfidfVectorizer(ngram_range=(1, 2))),
        ("classifier", model)
    ])

    scores = cross_validate(
        pipeline,
        X,
        y,
        groups=group_id,
        cv=GroupKFold(n_splits=5),
        scoring=["accuracy", "f1_macro", "f1_weighted"]
    )

    print(model_name)
    print("Accuracy:", scores["test_accuracy"].mean())
    print("F1-macro:", scores["test_f1_macro"].mean())
    print("F1-weighted:", scores["test_f1_weighted"].mean())
```

Şekil 7. Çalışmada Geliştirilen Model Eğitim ve Değerlendirme Kodları

Çalışmada kullanılan sınıflandırma modellerine ait temel kod tanımları Tablo 7'de sunulmaktadır.

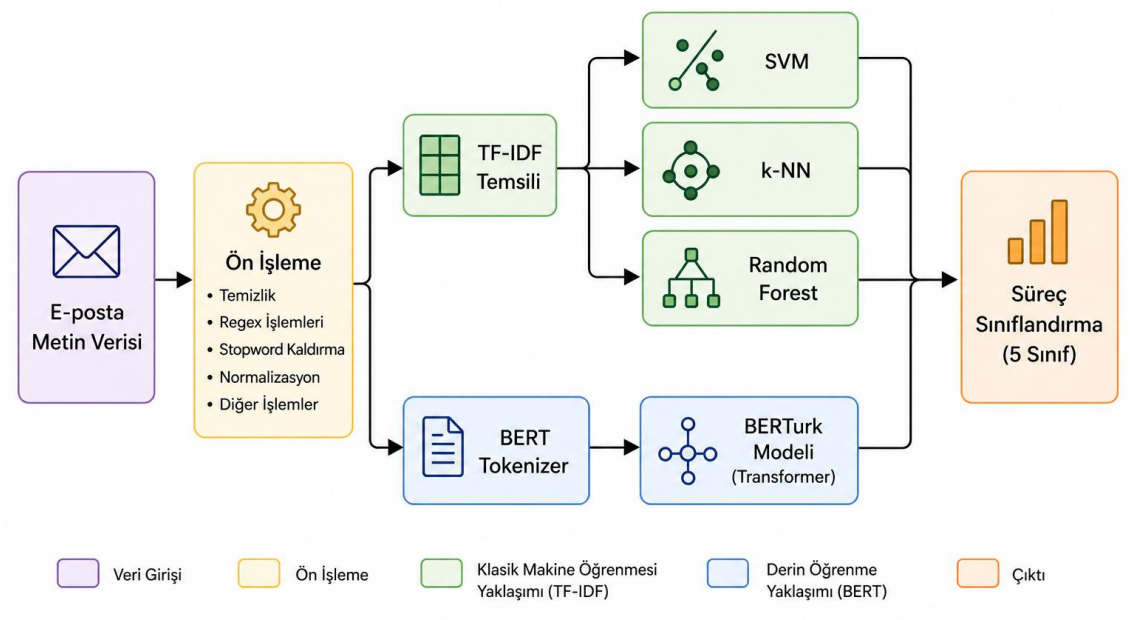
Tablo 7. Çalışmada Kullanılan Modellerin Kod Tanımları

Model	**Kod Tanımı**
SVM	LinearSVC(class_weight="balanced", C=0.7)
k-NN	KNeighborsClassifier(n_neighbors=15, metric="cosine", weights="distance")
RF	RandomForestClassifier(n_estimators=300, random_state=42, class_weight="balanced")
BERTurk	AutoModelForSequenceClassification.from_pretrained("dbmdz/bert-base-turkish-uncased")

Tablo 7'de verildiği gibi klasik ML modelleri uygun parametrelerle kullanılmıştır. SVM modelinde dengeli sınıf ağırlığı kullanılmış, k-NN modelinde metinler arası benzerliği daha uygun biçimde ölçebilmek amacıyla kosinüs metriği tercih edilmiştir. RF modelinde ise karar ağaçlarının topluluk yapısından yararlanılmıştır. BERTurk modeli, Türkçe metinler için önceden eğitilmiş transformer tabanlı bir model olarak sınıflandırma sürecine dahil edilmiştir.

BERTurk modeli transformer tabanlı bir derin öğrenme modeli olduğundan, klasik ML modellerinden farklı bir eğitim yapısına sahiptir. Bu nedenle BERTurk modelinin eğitim süreci, tokenizer kullanımı, maksimum token uzunluğu, epoch sayısı ve batch size gibi parametreler üzerinden ayrıca yapılandırılmıştır.

Şekil 8'de, tez kapsamında kullanılan modelleme yaklaşımlarının genel yapısı şematik olarak sunulmaktadır.



Şekil 8. Çalışmada Kullanılan Modelleme Yaklaşımlarının Şematik Gösterimi

Şekil 8'de görüldüğü üzere, metin verileri iki farklı modelleme yaklaşımı ile ele alınmıştır. İlk yaklaşımda TF-IDF temsili kullanılarak SVM, k-NN ve RF gibi klasik ML algoritmaları uygulanmıştır. İkinci yaklaşımda ise metin verileri BERT tabanlı bir model olan BERTurk aracılığıyla bağlamsal olarak temsil edilerek sınıflandırma gerçekleştirilmiştir.

3.2.4.1 Destek vektör makinesi

SVM, denetimli öğrenme yaklaşımına dayanan ve özellikle yüksek boyutlu veri uzaylarında etkili sonuçlar üretebilen güçlü bir sınıflandırma algoritmasıdır. SVM algoritmasının temel amacı, farklı sınıflara ait örnekleri en iyi şekilde ayıran ve sınıflar arasındaki marjini maksimize eden bir karar sınırı (hiper düzlem) bulmaktır. Bu özellik, metin sınıflandırma problemlerinde ortaya çıkan yüksek boyutlu ve seyrek TF-IDF temsilleri için önemli bir avantaj sağlamaktadır.

Bu tez çalışmasında SVM modeli, e-posta metinlerinden elde edilen kelime tabanlı TF-IDF özellikleri kullanılarak eğitilmiştir. Metinlerin TF-IDF ile temsil edilmesi, her bir e-postayı binlerce boyuttan oluşan seyrek vektörler halinde ifade ettiğinden, doğrusal sınıflandırıcıların etkinliğini artırmaktadır. Bu doğrultuda, doğrusal SVM tercih edilmiştir. Sınıf dengesizliğinin etkisini azaltmak amacıyla class_weight=balanced

parametresi kullanılmıştır. Ayrıca, aşırı öğrenmenin önüne geçebilmek için düzenlileştirme parametresi C, farklı değerler altında denenerek modelin genelleme yeteneği kontrol edilmiştir.

Elde edilen deneysel sonuçlar, SVM modelinin e-posta içeriklerini ilgili RPA süreçlerine göre ayırt edebilme konusunda güçlü bir performans sunduğunu göstermektedir. SVM, yüksek boyutlu metin temsilleri üzerinde istikrarlı bir ayırt edicilik sağlamaktadır. Özellikle gürültü içeren, kısa veya biçimsel farklılıklar barındıran kurumsal e-posta metinlerinde anlamlı sonuçlar üretebilmektedir. Bu yönüyle SVM, çalışmada kullanılan diğer algoritmalar için güçlü bir karşılaştırma temeli oluşturmuştur.

3.2.4.2 k-En yakın komşu

k-NN algoritması, denetimli öğrenme yaklaşımına dayanan ve sınıflandırma kararını eğitim verisindeki en benzer örneklerin sınıf bilgilerine göre veren örnek tabanlı bir yöntemdir. k-NN algoritması, klasik anlamda bir model öğrenme süreci içermemektedir. Yeni bir örneğin sınıfını belirlerken eğitim kümesindeki örneklerle doğrudan benzerlik hesaplamaktadır. Bu nedenle, özellikle metinler arası benzerlik ilişkilerinin belirgin olduğu problemlerde sezgisel ve anlaşılır bir yaklaşım sunmaktadır.

Bu çalışmada k-NN algoritması, e-posta içeriklerinin TF-IDF temsilleri üzerinden uygulanmış ve metinler arası benzerlik kosinüs benzerliği ölçütü ile hesaplanmıştır. Ayrıca, sınıflandırma kararında yakın komşuların etkisini artırmak amacıyla mesafeye bağlı ağırlıklandırma yöntemi kullanılmış ve farklı komşu sayıları (k) için deneyler gerçekleştirilmiştir. Bu yapı sayesinde, içerik açısından birbirine daha yakın e-postaların sınıflandırma sürecinde daha belirleyici olması hedeflenmiştir.

Deney sonuçları, k-NN algoritmasının bu problemde oldukça rekabetçi bir performans sergilediğini göstermektedir. Özellikle uygun k değeri seçildiğinde, k-NN modeli hem doğruluk hem de F1 skoru açısından SVM modelini aşan sonuçlar üretmiştir. Bununla birlikte, k-NN algoritmasının hesaplama maliyetinin veri boyutu arttıkça yükselmesi ve yüksek boyutlu uzaylarda mesafe hesaplamalarının zorlaşması, yöntemin pratik uygulamalardaki sınırlılıkları arasında değerlendirilmektedir.

3.2.4.3 Rastgele orman

RF algoritması, birden fazla karar ağacının bir araya getirilmesiyle oluşturulan ve topluluk öğrenmesine dayanan denetimli bir sınıflandırma yöntemidir. RF, her bir karar ağacını farklı örneklem alt kümeleri ve rastgele seçilmiş özellik alt kümeleri üzerinde eğiterek, tekil karar ağaçlarının aşırı öğrenme eğilimini azaltmayı ve daha dengeli bir genelleme performansı elde etmeyi amaçlamaktadır.

Bu tez çalışmasında RF modeli, e-posta metinlerinden elde edilen TF-IDF ve karakter tabanlı TF-IDF temsilleri kullanılarak değerlendirilmiştir. TF-IDF temsillerinin seyrek yapıda olması nedeniyle, RF algoritmasının eğitimi öncesinde bu temsilleri model uyumluluğu gereği uygun biçimde dönüştürülerek kullanılmıştır. Modelleme sürecinde ağaç sayısı ve ağaç derinliği gibi temel hiperparametreler farklı değerlerle denenmiş ve modelin performansı sistematik biçimde analiz edilmiştir.

Elde edilen bulgular, RF algoritmasının bu çalışma kapsamında klasik yöntemler arasında en yüksek ve en istikrarlı performansı sergileyen yöntem olduğunu ortaya koymaktadır. Özellikle ağaç sayısının artırılması ve uygun derinlik sınırlarının belirlenmesiyle, model hem doğruluk hem de F1 skoru açısından diğer yöntemleri geride bırakmıştır. Elde edilen sonuçlar, RF modelinin klasik yöntemler arasında dengeli ve istikrarlı bir performans sergilediğini göstermektedir. Bu durum, RF algoritmasının metin sınıflandırma probleminde güçlü ve güvenilir bir alternatif olduğunu ortaya koymaktadır.

3.2.4.4 Transformer tabanlı derin öğrenme modelleri

BERT, Google tarafından geliştirilen ve transformer mimarisine dayanan derin öğrenme tabanlı bir NLP modelidir (Devlin ve arkadaşları, 2019). BERT modeli, metinleri çift yönlü (bidirectional) olarak işleyerek bir kelimenin anlamını hem önceki hem de sonraki bağlamı dikkate alarak temsil edebilmektedir. Bu özellik, geleneksel kelime temsili yöntemlerine ve tek yönlü dil modellerine kıyasla metinlerin anlamsal yapısının daha doğru biçimde yakalanmasını sağlamaktadır.

Klasik ML yöntemleri metni ağırlıklı olarak istatistiksel özellikler üzerinden temsil ederken, transformer tabanlı modeller metinlerdeki bağlamsal ilişkileri dikkate alarak daha zengin temsiller oluşturabilmektedir. Bu kapsamda çalışmada, Türkçe metinler için

önceden eğitilmiş bir transformer modeli olan BERTurk (dbmdz/bert-base-turkish-uncased) kullanılmıştır. BERTurk, metni çift yönlü bağlamda işleyerek bir kelimenin anlamını hem önceki hem de sonraki kelimelerle ilişkisi üzerinden temsil edebilmektedir.

BERTurk modelinin değerlendirmesinde, veri sızıntısı riskini azaltmak ve fold'larda sınıf dağılımını korumak amacıyla label-içi benzerlik gruplama (group_id) yaklaşımı korunmuştur. Ardından grup düzeyinde ayrım yapabilmek için GroupKFold uygulanmıştır.

Elde edilen sonuçlarda BERTurk modeli yüksek başarı değerleri üretmiştir. Bunun temel nedeni, modelin bağlamsal temsilleri öğrenebilmesidir. Bu nedenle BERTurk sonuçları yorumlanırken değerlendirme stratejisi ve veri çeşitliliği dikkate alınmıştır.

3.2.5 Değerlendirme

Bu çalışmada, e-posta içeriklerinin ilgili RPA süreçlerine göre doğru biçimde sınıflandırılabilmesi amacıyla geliştirilen ML modellerinin eğitimi, doğrulanması ve seçimi sistematik şekilde yürütülmüştür. Öncelikle, veri seti denetimli öğrenme yaklaşımına uygun olacak şekilde eğitim ve test kümelerine ayrılmıştır. Eğitim–test ayrımı, modelin öğrenme sürecinde kullanılan veriler ile performansının değerlendirildiği verilerin birbirinden bağımsız olmasını sağlayarak elde edilen sonuçların genellenebilirliğini artırmayı amaçlamaktadır.

Tablo 8'de ön deneylerde kullanılan eğitim-test veri seti ayrım oranları yer almaktadır.

Tablo 8. Ön Deneylerde Kullanılan Eğitim-Test Veri Seti Ayrım Oranı

Eğitim Verisi Boyutu	Test Verisi Boyutu
%80	%20

İlk aşamada veri seti satır bazlı olarak %80 eğitim ve %20 test olacak şekilde ayrılarak ön deneyler gerçekleştirilmiştir. Tablo 8'de ön deneylerde kullanılan eğitim-test veri seti ayrım oranı sunulmaktadır. %80 eğitim ve %20 test ayrımı, modelin büyük bir veri bölümü üzerinden öğrenmesini sağlarken, ayrılan test kümesi aracılığıyla daha önce görmediği veriler üzerindeki genel başarımının gözlemlenmesine olanak tanıdığı için

makine öğrenmesi çalışmalarında yaygın olarak kullanılan bir yaklaşımdır. Shepard ve Naheed (2021), farklı eğitim-test ayrımlarını karşılaştırdıkları çalışmalarında standartlaştırılmış veri üzerinde %80 eğitim ve %20 test ayrımı kullanıldığında modelin %99'un üzerinde doğruluk elde ettiğini belirtmiştir (Shepard ve Naheed, 2021).

Bu ön deney aşamasında ayrıca ayrı bir doğrulama (validation) kümesi oluşturulmamıştır. Model doğrulama ve hiperparametre değerlendirme süreci, ön deneylerde elde edilen bulgular doğrultusunda daha sonra 5-katlı GroupKFold çapraz doğrulama (cross-validation) yaklaşımı içerisinde yürütülmüştür.

Satır bazlı %80-%20 eğitim-test ayrımı ile gerçekleştirilen ön deneylerde bazı modellerin 1.00'a yakın doğruluk değerleri üretmesi, veri setinde yer alan aynı veya yüksek derecede benzer e-posta içeriklerinin eğitim ve test kümelerine birlikte düşmesinden kaynaklanabilecek veri sızıntısı (data leakage) riski olarak değerlendirilmiştir. Benzer içeriklerin eğitim ve test kümelerine aynı anda düşmesi, modelin test verisini dolaylı olarak öğrenmesine ve performansın yapay biçimde yüksek ölçülmesine neden olabilmektedir.

Bu bulgu üzerine nihai değerlendirme stratejisi değiştirilmiştir. Metinler benzerlik temelli olarak gruplandırılmış ve her bir benzer içerik kümesine bir group_id atanmıştır. Böylece aynı veya yüksek derecede benzer e-posta içeriklerinin farklı değerlendirme kümelerine düşmesi engellenmiştir. Nihai model karşılaştırmaları ve hiperparametre seçimleri, satır bazlı %80-%20 ayrımı yerine 5-katlı GroupKFold çapraz doğrulama yaklaşımı ile grup bazında yürütülmüştür. Bu yaklaşımda veri seti grup düzeyinde beş farklı parçaya ayrılmış, her iterasyonda bir parça değerlendirme amacıyla kullanılırken kalan parçalar eğitim için kullanılmıştır. Böylece benzer e-posta içeriklerinin eğitim ve değerlendirme kümelerine aynı anda düşmesi engellenmiş ve model performanslarının daha gerçekçi biçimde ölçülmesi sağlanmıştır. Çalışmada raporlanan nihai model karşılaştırma sonuçları bu group-level çapraz doğrulama çıktılarından elde edilmiştir.

Model performansını en üst düzeye çıkarabilmek amacıyla hiperparametre optimizasyonu uygulanmıştır. k-NN, SVM ve RF algoritmaları için model karmaşıklığını ve genelleme kapasitesini doğrudan etkileyen temel hiperparametreler farklı değerler altında denenmiştir. Uygun hiperparametre kombinasyonları, eğitim verisi üzerinde

uygulanan GroupKFold tabanlı çapraz doğrulama sonuçlarına göre belirlenmiştir. Aşırı öğrenme riskini minimize eden ve dengeli performans sağlayan yapıların seçilmesine dikkat edilmiştir.

Sonuç olarak, model karşılaştırmaları ve hiperparametre ayarlamaları grup bazlı çapraz doğrulama ile gerçekleştirilmiştir. Bu tezde raporlanan nihai model karşılaştırma sonuçları 5-katlı GroupKFold çapraz doğrulama ortalamalarına dayanmaktadır. %80-%20 eğitim-test seti ayrımı yalnızca veri sızıntısı riskinin tespiti amacıyla ön analiz olarak kullanılmıştır ve nihai sonuçlara dahil edilmemiştir. Bu çok aşamalı değerlendirme yaklaşımı, elde edilen sonuçların hem metodolojik açıdan güvenilir hem de pratik uygulama açısından genellenebilir olmasını sağlamaktadır.

Bu çalışmada kullanılan performans ölçütleri, sınıflandırma modellerinin başarımını nesnel ve karşılaştırılabilir biçimde değerlendirebilmek amacıyla karışıklık matrisi (confusion matrix) bileşenlerine dayalı olarak hesaplanmıştır. Bu kapsamda doğru pozitif (TP), doğru negatif (TN), yanlış pozitif (FP) ve yanlış negatif (FN) değerleri temel alınmıştır.

Karışıklık matrisi bileşenleri, modelin tahmin ettiği sınıf ile gerçek sınıf arasındaki ilişkiyi ifade etmektedir. TP, modelin pozitif tahmin yaptığı ve gerçekte de pozitif olan durumları temsil etmektedir. FP, pozitif tahmin edilmesine rağmen gerçekte negatif olan durumları temsil etmektedir. FN, negatif tahmin edilmesine karşın gerçekte pozitif olan durumları temsil etmektedir. TN ise hem negatif tahmin edilen hem de gerçekte negatif olan durumları temsil etmektedir. Bu kavramlara ait açıklamalar Tablo 9'da yer almaktadır (Vujović, 2021).

Tablo 9. Karışıklık Matrisi Bileşenleri ve Açıklamaları

Kısaltma	Terim	Açıklama
TP	Doğru Pozitif	Model tarafından pozitif olarak tahmin edilen bir örneğin, gerçekte de pozitif olması durumudur. Modelin ilgili sınıfa ait bir veriyi doğru biçimde sınıflandırdığını ifade eder.
FP	Yanlış Pozitif	Model tarafından pozitif olarak tahmin edilen bir örneğin, gerçekte negatif olması durumudur. Bu durum Tip I Hata olarak adlandırılmakta olup, modelin olmayan bir durumu varmış gibi algılamasına neden olmaktadır.
FN	Yanlış Negatif	Model tarafından negatif olarak tahmin edilen bir örneğin, gerçekte pozitif olması durumudur. Bu senaryo Tip II Hata olarak bilinmekte ve özellikle kritik karar mekanizmalarında ciddi riskler doğurabilmektedir.
TN	Doğru Negatif	Model tarafından negatif olarak tahmin edilen bir örneğin, gerçekte de negatif olması durumudur. Modelin negatif sınıfa ait bir veriyi doğru şekilde tanımladığını göstermektedir.

Doğruluk (Accuracy), model tarafından doğru sınıflandırılan örneklerin toplam örnek sayısına oranını ifade etmektedir ve Formül 1'deki gibi hesaplanmaktadır.

(1)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Bu metrik, modelin genel sınıflandırma başarısını göstermektedir. Ancak sınıf dağılımının dengesiz olduğu veri setlerinde tek başına yeterli bir değerlendirme ölçütü olmayabilmektedir.

Hassasiyet (Precision), modelin belirli bir sınıfa ait olarak tahmin ettiği örnekler arasından gerçekten o sınıfa ait olanların oranını göstermekte olup, yanlış pozitif tahminlerin etkisini ölçmektedir. Hassasiyet değeri aşağıdaki Formül 2'deki gibi hesaplanmaktadır.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Geri çağırma (Recall), gerçek sınıfa ait örneklerin ne kadarının model tarafından doğru şekilde tespit edilebildiğini ifade etmektedir ve özellikle yanlış negatif tahminlerin kritik olduğu durumlarda önem taşımaktadır. Geri çağırma metriği Formül 3'teki gibi hesaplanmaktadır.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1 skoru, hassasiyet ve geri çağırma metriklerinin harmonik ortalaması olarak tanımlanmakta ve bu iki ölçüt arasında denge sağlamaktadır. F1 skorunun hesaplanma yöntemi Formül 4'teki gibidir.

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Bu çalışma kapsamında söz konusu metrikler, her bir sınıf için ayrı ayrı hesaplanmıştır. Ayrıca çok sınıflı yapı nedeniyle macro ortalama ve weighted ortalama değerler üzerinden genel model performansı da değerlendirilmiştir. Bu yaklaşım sayesinde,

geliştirilen sınıflandırma modellerinin hem sınıf bazlı hem de tüm veri seti genelindeki başarımı ayrıntılı ve dengeli bir biçimde analiz edilmiştir.

3.2.6 Uygulama

Bu tez çalışmasında geliştirilen sınıflandırma modelleri, deneysel bir araştırma kapsamında değerlendirilmiş olup, doğrudan bir RPA sistemi ile canlı entegrasyon gerçekleştirilmemiştir. Bununla birlikte, elde edilen sonuçlar doğrultusunda geliştirilen yaklaşımın kurumsal RPA süreçlerine entegre edilebilir bir yapı sunduğu öngörülmektedir.

Önerilen yaklaşımda, sınıflandırma modeli RPA altyapısından bağımsız bir bileşen olarak ele alınmaktadır. Bu sayede modelin yeniden eğitilmesi, farklı veri setleri ile güncellenmesi veya yeni süreçlerin eklenmesi durumunda, mevcut RPA akışlarının değiştirilmesine gerek kalmadan sistemin genişletilmesi mümkün olabilecektir. Bu mimari yaklaşım, ölçeklenebilirlik ve sürdürülebilirlik açısından önemli bir avantaj sunmaktadır.

Sonuç olarak, bu tez çalışmasında gerçekleştirilen deneysel analizler ve elde edilen bulgular, geliştirilen modelleme yaklaşımının kurumsal ortamlarda kullanılan RPA süreçlerine uygulanabilir nitelikte olduğunu göstermektedir. İleride gerçekleştirilecek çalışmalarda gerçek zamanlı entegrasyon ve canlı sistem uygulamaları için sağlam bir temel oluşturmaktadır.

3.2.7 Önerilen sistem mimarisi

Bu çalışma kapsamında canlı sistem entegrasyonu gerçekleştirilmemiştir. Geliştirilen sınıflandırma modelinin kurumsal RPA altyapısına nasıl entegre edilebileceğine yönelik önerilen bir mimari yapı tasarlanmıştır. Önerilen mimari, sınıflandırma modelinin RPA altyapısından bağımsız, servis tabanlı bir bileşen olarak kullanılmasını sağlamaktadır. Bu sayede model, mevcut RPA süreçlerinden ayrıştırılmış bir katmanda çalışabilmektedir ve modelde yeniden eğitim, güncelleme veya model değişikliği gibi durumlarda RPA akışlarının yeniden tasarlanmasına gerek kalmaması hedeflenmektedir.

Önerilen entegrasyon yapısında süreç akışı aşağıdaki şekildedir:

1. Kullanıcı tarafından gönderilen e-posta veya mesaj, ilgili RPA robotunun e-posta kutusuna düşer.

2. RPA robotu, mesaj içeriğini bir ön işleme servisine iletir.

3. Ön işleme katmanında gerekli metin temizleme, anonimleştirme ve format dönüşümleri gerçekleştirilir.

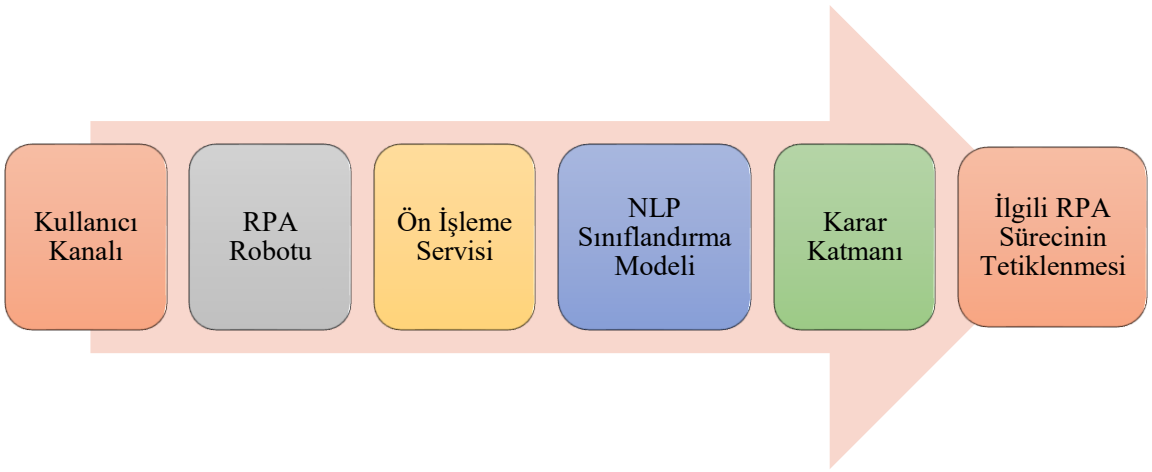
4. Temizlenen metin, NLP tabanlı sınıflandırma modeline gönderilir.

5. Model, ilgili süreç sınıfını ve güven skorunu üretir.

6. Karar katmanında belirlenen eşik değere göre ilgili RPA süreci tetiklenir veya "süreç dışı" olarak işaretlenir.

7. Sonuçlar loglama ve izleme sistemine kaydedilerek süreç sonlandırılır.

Şekil 9'da bu çalışmada önerilen RPA-NLP tabanlı sistem mimarisi gösterilmektedir. Mimari yapı, kullanıcı kanalı üzerinden e-posta veya mesajlaşma kanalları üzerinden gelen yapısal olmayan verilerin, ilgili RPA süreçlerine otomatik olarak yönlendirilmesini sağlayacak şekilde tasarlanmıştır.



Şekil 9. Önerilen Sistem Mimarisi

Süreç, kullanıcı tarafından gönderilen iletinin RPA robotu tarafından alınması ile başlamaktadır. RPA robotuna gelen mesaj öncelikle bir ön işleme servisine iletilmektedir. Ön işleme katmanında metin temizleme ve format dönüşümü gibi işlemler

gerçekleştirilerek veri modele iletilmek için uygun hale getirilmektedir. Temizlenen metin daha sonra NLP sınıflandırma modeline aktarılmaktadır. Model tarafından tahmin edilen ilgili süreç sınıfı ile bir güven skoru üretilmektedir.

Model çıktısı, karar katmanında değerlendirilmektedir. Belirlenen uygun eşik değer doğrultusunda mesajın hangi RPA sürecini tetikleyeceği tespit edilebilir. Eğer güven skoru beklenen düzeyde değilse, mesaj "süreç dışı" olarak işaretlenebilmekte ve manuel olarak incelenmesi üzere ilgili kişilere yönlendirilmektedir. Eşik değerini aşan durumlarda ise ilgili RPA süreci (queue veya workflow yapısı üzerinden) otomatik olarak tetiklenmesi sağlanmaktadır. Bu mimaride kullanılan güven skoru eşiği, sabit ve tüm kurumlar için geçerli bir değer olarak değerlendirilmemelidir. Eşik değer, kurumun operasyonel risk toleransı, süreçlerin kritiklik düzeyi ve yanlış sınıflandırma maliyeti dikkate alınarak belirlenebilir. Güven skorunun belirlenen eşik değerin altında kalması durumunda ilgili e-posta doğrudan otomatik sürece aktarılmak yerine manuel inceleme kuyruğuna yönlendirilerek hatalı süreç tetikleme riski azaltılabilir.

Önerilen mimari yapıda sınıflandırma bileşeni, RPA sisteminden bağımsız bir servis mimarisi olarak konumlandırılmıştır. RPA robotu, gelen e-posta içeriğini HTTP tabanlı bir REST API çağrısı ile sınıflandırma servisine iletmektedir. Bu çağrıda mesaj gövdesi ve gerekli meta veriler JSON formatında taşınmaktadır. Ön işleme işlemlerinin ardından veri eğitilmiş NLP modeline aktarılmakta ve model tarafından tahmin edilen süreç etiketi, güven skoru, model versiyon bilgisi ve işlem süresi gibi çıktılar üretilmektedir.

Karar katmanında, belirlenen eşik değere göre güven skoru değerlendirilmekte ve yeterli güven düzeyine sahip tahminler doğrudan ilgili RPA kuyruğuna yönlendirilmektedir. Eşik değerin altında kalan tahminler ise manuel inceleme kuyruğuna aktarılabilmektedir veya süreç dışı olarak etiketlenebilmektedir. Bu yapı, yanlış tetiklemelerin azaltılmasına ve operasyonel risklerin kontrol altında tutulmasına katkı sağlamaktadır.

Önerilen mimaride hata yönetimi süreci de sistemin önemli bir bileşeni olarak değerlendirilmektedir. Modelin düşük skor üretmesi, e-posta içeriğinin eksik veya belirsiz olması, ön işleme servisinde hata oluşması, sınıflandırma servisinden yanıt alınamaması ya da RPA platformunda süreç tetikleme hatası meydana gelmesi durumlarında ilgili kayıt doğrudan otomatik süreçlere aktarılmamalıdır. Bu tür

durumlarda hata türü sistem tarafından kayıt altına alınmalı, kayıt manuel inceleme kuyruğuna yönlendirilmeli ve gerekli durumlarda operasyon ekibine bildirim üretilmelidir. Böylece yanlış RPA sürecinin tetiklenmesi engellenerek operasyonel riskler azaltılabilir. Ayrıca hata kayıtlarının düzenli olarak izlenmesi, modelin yeniden eğitilmesi ve sistem performansının iyileştirilmesi için geri bildirim mekanizması sağlayacaktır.

Ayrıca sistemde loglama ve izleme mekanizması yer almaktadır. Modelin yaptığı tahminler, model performans metrikleri ve hata durumları merkezi bir kayıt altyapısında saklanmaktadır. Bu kayıtlar, modelin yeniden eğitilmesi ve performansının izlenmesi amacıyla geri bildirim olarak kullanılabilmektedir.

Bu servis tabanlı mimari yaklaşım, farklı RPA platformları ile entegre edilebilir nitelikte olup teknoloji bağımsız bir çözüm çerçevesi sunmaktadır. Önerilen yapı, bu tez kapsamında deneysel olarak geliştirilen modelleme yaklaşımının kurumsal ortamlarda uygulanabilirliğini göstermekte ve gelecekte gerçekleştirilecek gerçek zamanlı canlı entegrasyon çalışmaları için teknik bir temel oluşturmaktadır.

4. BULGULAR VE TARTIŞMA

Bu tez çalışmasında, RPA süreçlerinde kullanılan robotlara iletilen doğal dilde oluşturulmuş e-posta içeriklerinin anlamlandırılması, NLP teknikleri kullanılarak gerçekleştirilmiştir. E-postaların ilgili iş süreçlerine uygun biçimde sınıflandırılması ise yapay zekanın bir alt alanı olan ML yöntemleri aracılığıyla sağlanmıştır. Çalışma kapsamında, metin sınıflandırma problemi için hem klasik ML algoritmaları k-NN, RF ve SVM hem de ML'in bir alt alanı olan derin öğrenme temelli transformer tabanlı modeller (BERTurk) kullanılarak karşılaştırmalı bir analiz yapılmıştır. Bu yaklaşım sayesinde, metinsel verilerin otomatik olarak analiz edilmesi ve uygun RPA süreçlerine yönlendirilmesi sağlanmıştır.

Bu kapsamda, e-postaların içeriklerine bağlı olarak hangi iş sürecini tetiklemesi gerektiğinin otomatik olarak tahmin edilmesi amaçlanmıştır. Çalışmada kullanılan ML algoritmaları, metin sınıflandırma problemlerinde yaygın olarak tercih edilen ve farklı modelleme yaklaşımlarını temsil eden yöntemler arasından seçilmiştir. Modellerin eğitimi, doğrulanması ve performanslarının karşılaştırılması süreçleri sistematik bir şekilde yürütülmüştür. Elde edilen sonuçlar doğruluk, F1 skoru ve benzeri değerlendirme ölçütleri üzerinden analiz edilerek RPA süreçlerinde kullanılabilecek en uygun sınıflandırma yaklaşımının belirlenmesi hedeflenmiştir.

4.1 Bulgular

Bu bölümde, e-posta içeriklerinin RPA süreçlerine göre sınıflandırılması amacıyla uygulanan klasik ML algoritmaları ile transformer tabanlı derin öğrenme modelinin elde ettiği deneysel sonuçlar sunulmaktadır. Klasik ML modelleri TF-IDF temsilleri ile eğitilirken, BERT modeli bağlamsal gömme (embedding) yapısı ile eğitilmiştir. Benzer e-posta içeriklerinin eğitim ve test kümelerine aynı anda düşmesi sonucu oluşabilecek veri sızıntısını önlemek amacıyla, deneylerde benzerlik temelli grup kimlikleri (group_id) oluşturulmuş ve performans değerlendirmeleri GroupKFold yaklaşımı ile gerçekleştirilmiştir. Bu yöntem, veri sızıntısını engelleyerek model performansının daha gerçekçi değerlendirilmesini sağlamaktadır.

Model performansları doğruluk (Accuracy), makro F1 (F1-macro) ve ağırlıklı F1 (F1-weighted) metrikleri üzerinden raporlanmıştır. Her bir model için 5-katlı çapraz doğrulama sonucunda elde edilen ortalama değerler dikkate alınmıştır.

4.1.1 Klasik ML modellerinin sonuçları

Klasik ML yöntemleri kapsamında SVM, k-NN ve RF algoritmaları değerlendirilmiştir. Modeller, kelime düzeyinde TF-IDF temsilleri üzerinde farklı hiperparametre ayarlarıyla test edilmiştir.

Tablo 10'da TF-IDF tabanlı metin temsilleri kullanılarak eğitilen klasik ML modellerinin sonuçları sunulmaktadır.

Tablo 10. Klasik ML Modellerinin Sonuçlarının Karşılaştırılması

Model	Accuracy	F1-macro	F1-weighted
SVM (C=0.7)	0.84	0.85	0.82
k-NN (k=15, cosine)	0.88	0.85	0.87
RF	0.90	0.89	0.90

Sonuçlar, RF modelinin uygun hiperparametre ayarlarıyla en yüksek ortalama doğruluk ve F1 skorlarını elde ettiğini göstermektedir. k-NN modeli benzerlik temelli yapısı sayesinde rekabetçi sonuçlar üretirken, SVM modeli daha dengeli ancak görece daha düşük bir performans sergilemiştir.

4.1.2 Transformer tabanlı model (BERTurk) sonuçları

Klasik yöntemlere ek olarak, metinlerin bağlamsal temsillerini öğrenebilen transformer tabanlı bir yaklaşım olarak BERTurk modeli değerlendirilmiştir. Model, Türkçe metinler üzerinde önceden eğitilmiş dbmdz/bert-base-turkish-uncased kullanılarak ince ayar (fine-tuning) yapılmıştır. Değerlendirme sürecinde, klasik modellerle tutarlılığı korumak amacıyla label-içi benzerlik temelli group_id yapısı korunmuş ve GroupKFold uygulanmıştır.

BERTurk modeli sonuçları Tablo 11'de sunulmaktadır.

Tablo 11. BERTurk Modeli Sonuçları

Model	Accuracy	F1-macro	F1-weighted
BERTurk	0.94	0.93	0.96

Tablo 11'de BERTurk modelinin label-içi benzerlik temelli gruplama ve group-level çapraz doğrulama yaklaşımı ile elde edilen performans sonuçları yer almaktadır. Ortalama sonuçlar, BERTurk modelinin klasik ML yöntemlerine kıyasla daha yüksek doğruluk ve F1 skorları ürettiğini göstermektedir. Bu bulgu, transformer tabanlı modellerin e-posta içeriklerindeki bağlamsal ve anlamsal ilişkileri daha etkin biçimde yakalayabildiğini ve karmaşık doğal dil yapılarında üstün bir temsil gücü sunduğunu ortaya koymaktadır.

4.2 Tartışma

Bu çalışma kapsamında elde edilen deneysel sonuçlar, e-posta içeriklerinin RPA süreçlerine göre sınıflandırılmasında kullanılan model türünün ve uygulanan değerlendirme stratejisinin performans üzerinde belirleyici olduğunu göstermektedir. Klasik ML algoritmaları ile transformer tabanlı derin öğrenme modelleri arasında belirgin performans farkları gözlemlenmiştir.

Klasik yöntemler incelendiğinde, SVM modelinin yüksek boyutlu ve seyrek TF-IDF temsilleri üzerinde kararlı ve tutarlı sonuçlar ürettiği görülmüştür. Bu performans, SVM modelinin yüksek boyutlu ve seyrek yapıdaki TF-IDF vektörlerinde etkili bir şekilde genelleme yapabilmesinden kaynaklanmaktadır. SVM, özellikle sınıflar arası marjini maksimize eden yapısı sayesinde veri setindeki gürültüye karşı dayanıklılık göstermiştir. Ancak doğrusal karar sınırına dayalı olması, daha karmaşık bağlamsal ilişkilerin modellenmesini sınırlamıştır.

k-NN algoritması, benzerlik temelli yapısı sayesinde özellikle içerik açısından birbirine yakın e-postaların sınıflandırılmasında başarılı sonuçlar üretmiştir. Bu durum, veri setinde benzer içerikli e-postaların belirli kümeler oluşturması ve k-NN algoritmasının bu benzerlik yapısından doğrudan faydalanması ile açıklanabilir. Kosinüs benzerliği kullanımı, metinler arası göreli yakınlığın daha doğru biçimde ölçülmesine katkı

sağlamıştır. Bununla birlikte, yüksek boyutlu özellik uzayında hesaplama maliyetinin artması ve performansın seçilen k değerine duyarlı olması, k-NN'nin ölçeklenebilirliğini sınırlayan faktörler olarak değerlendirilmiştir.

RF modeli, uygun hiperparametre ayarlarıyla yüksek doğruluk ve F1 skorları elde etmiş olsa da metin verilerinde ortaya çıkan yüksek boyutluluk ve seyrek yapı nedeniyle performansının ayar seçimlerine oldukça duyarlı olduğu gözlemlenmiştir. Bu başarı, modelin farklı örneklem ve özellik alt kümeleri üzerinden öğrenme yaparak aşırı öğrenme riskini azaltmasından kaynaklanmaktadır. Bu durum, ağaç tabanlı yöntemlerin metin sınıflandırma problemlerinde dikkatli yapılandırılması gerektiğini ortaya koymaktadır.

Benzer şekilde e-posta sınıflandırma alanında yapılan bir çalışmada, spam ve spam olmayan e-postaların ayrıştırılması amacıyla NLP tabanlı özellik çıkarımı ve özellik seçimi yöntemleri kullanılmıştır. Çalışmada SVM algoritmasının %95,17 doğruluk oranına ulaştığı, RF algoritmasının ise %93,3 doğruluk sağladığı belirtilmiştir (Srinivasarao ve Sharaff, 2023). Bu tez çalışmasında da e-posta içeriklerinin sınıflandırılmasında SVM ve RF gibi klasik ML algoritmalarının değerlendirilmesi, söz konusu literatür bulgularıyla uyumludur. Ancak ilgili çalışmadan farklı olarak bu tezde spam tespiti yerine, e-posta içeriklerinin RPA süreçlerine göre çok sınıflı olarak yönlendirilmesi problemi ele alınmıştır.

Transformer tabanlı BERTurk modeli ise, bağlamsal ve anlamsal temsilleri doğrudan öğrenebilmesi sayesinde klasik yöntemlere kıyasla belirgin biçimde daha yüksek ortalama performans değerleri üretmiştir. Bu durum, modelin metin içerisindeki bağlamsal ilişkileri çift yönlü olarak öğrenebilme yeteneğinden kaynaklanmaktadır. BERTurk'un özellikle makro F1 skorlarında sağladığı artış, sınıflar arası dengesizliğin etkisini azaltabildiğini göstermektedir.

Bu tez çalışmasında transformer tabanlı BERTurk modeli değerlendirilmiştir. Daha büyük veri setleri üzerinde eğitilerek üretilen büyük dil modellerinin doğrudan kullanımı tercih edilmemiştir. Bunun temel nedeni, çalışmanın amacının e-posta içeriklerinden yeni metin üretmek veya serbest biçimli yanıt oluşturmak değil, gelen e-postaları önceden tanımlanmış RPA süreçlerine göre güvenilir biçimde sınıflandırmak olmasıdır. Büyük dil modelleri metin anlama, metin üretme ve bağlamsal çıkarım yapma gibi görevlerde güçlü

sonuçlar üretmektedir. Kurumsal RPA süreçlerinde sürekli ve yüksek hacimli kullanım söz konusu olduğunda, bu modeller yüksek işlem maliyetlerine neden olabilmektedir. Büyük dil modelleri çevrim içi servislere giderek daha fazla entegre edilmesine rağmen, pahalı GPU gereksinimi nedeniyle dağıtım maliyetleri yüksektir (Griggs ve arkadaşları, 2024). Bu nedenle büyük dil modellerinin eğitimi, ince ayarı veya kurum içi güvenli altyapıda çalıştırılması yüksek GPU kapasitesi, geniş bellek, model servis altyapısı, bakım süreçleri ve izleme mekanizmaları gerektirmektedir.

Bu çalışmada kullanılan veri seti, belirli bankacılık RPA süreçlerine ait sınırlı sayıda ve etiketli e-posta kaydından oluştuğu için, büyük dil modellerinin sağladığı üretken yetenekler problem için gerekli görülmemiştir. Ayrıca büyük dil modellerinin çıktı üretiminde zaman zaman kararsız veya açıklanması güç sonuçlar verebilmesi, hatalı süreç tetikleme riskinin yüksek olduğu kurumsal otomasyon senaryolarında ek kontrol ihtiyacına neden olabilir. RPA süreçlerinde yanlış sınıflandırma, hatalı iş akışı başlatma veya manuel düzeltme ihtiyacına yol açabileceğinden, modelin doğruluğu kadar çalıştırma maliyeti, gecikme süresi, açıklanabilirliği ve bakım kolaylığı da önem taşımaktadır.

Bu nedenle, çalışma kapsamında daha yönetilebilir, daha düşük maliyetli ve sınıflandırma amacıyla doğrudan kullanılabilir modellerin değerlendirilmesi tercih edilmiştir. BERTurk modeli, Türkçe metinlerde bağlamsal temsil gücü sağlaması nedeniyle karşılaştırmaya dahil edilmiştir. Ancak gerçek zamanlı kurumsal kullanım açısından RF gibi klasik ML modellerinin düşük donanım gereksinimi, hızlı tahmin üretme kapasitesi ve daha kolay bakım avantajı da dikkate alınmıştır. Bu bağlamda, büyük dil modellerinin bu problem için tamamen gereksiz olduğu değil, mevcut veri yapısı, altyapı gereksinimleri ve operasyonel maliyetler dikkate alındığında doğrudan önerilen ilk çözüm olmadığı değerlendirilmektedir.

Benzer şekilde, Türkçe oltalama e-postalarının tespitine yönelik gerçekleştirilen bir çalışmada TF-IDF temsilleri ile Naive Bayes, SVM ve BERT modelleri karşılaştırmalı olarak değerlendirilmiştir. Çalışmada Türkçe için önceden eğitilmiş BERTurk modelinin yaklaşık %97,5 doğruluk değeri ile diğer modellere göre daha başarılı sonuçlar ürettiği belirtilmiştir (Taş, 2025). Bu sonuç, Türkçe e-posta içeriklerinin sınıflandırılmasında bağlamsal temsillere dayalı BERTurk gibi transformer tabanlı modellerin klasik

yöntemlere kıyasla daha güçlü sonuçlar verebildiğini göstermektedir. Bu tez çalışmasında BERTurk modelinin klasik ML algoritmalarına göre daha yüksek performans üretmesi de söz konusu literatür bulgusuyla uyumludur. Ancak ilgili çalışmadan farklı olarak bu tezde oltalama e-postası tespiti yerine, e-posta içeriklerinin RPA süreçlerine göre çok sınıflı olarak sınıflandırılması problemi ele alınmıştır.

Bu tez çalışmasında bulgular, gerçekçi ve güvenilir bir değerlendirme için group-level çapraz doğrulama yaklaşımlarının kritik öneme sahip olduğunu ortaya koymaktadır. İlk aşamada satır bazlı eğitim-test ayrımı ile elde edilen yüksek başarı değerleri, benzer e-posta içeriklerinin eğitim ve test kümelerine aynı anda düşmesi nedeniyle veri sızıntısı riski taşıyabileceğini göstermiştir. Bu nedenle çalışmada benzerlik temelli grup kimliği oluşturularak GroupKFold çapraz doğrulama yaklaşımının kullanılması, model performansının daha güvenilir biçimde ölçülmesine katkı sağlamıştır.

Bu yaklaşım, benzerlik temelli sınıflandırma literatürüyle de uyumludur. Benzer başka bir çalışmada, benzerlik temelli sınıflandırma yaklaşımlarında bir örneğin sınıf etiketinin, eğitim örnekleriyle olan benzerlik ilişkilerine göre tahmin edilebileceğini belirtmektedir. Ayrıca metin madenciliği ve doküman sınıflandırma problemlerinde TF-IDF vektörleri arasındaki kosinüs benzerliğinin yaygın olarak kullanılan bir yöntem olduğunu vurgulamaktadır (Chen ve arkadaşları, 2009). Bu tez çalışmasında da e-posta içerikleri arasındaki benzerlik ilişkileri kosinüs benzerliği ile belirlenmiş ve yüksek benzerlik gösteren kayıtların aynı grup kimliği altında değerlendirilmesi sağlanmıştır. Böylece benzer içeriklerin eğitim ve test kümeleri arasında dağılması engellenerek değerlendirme sürecinin daha tutarlı ve genellenebilir olması hedeflenmiştir.

Sonuç olarak, bu tez çalışmasında elde edilen bulgular hem klasik ML yöntemlerinin e-posta metin sınıflandırmasında güçlü bir temel oluşturduğunu hem de BERTurk gibi transformer tabanlı modellerin Türkçe e-posta içeriklerinde bağlamsal ilişkileri daha başarılı biçimde temsil edebildiğini göstermektedir. Bununla birlikte, model başarısının yalnızca algoritma seçimine bağlı olmadığı ve veri setinin yapısı, sınıf dağılımı, metinlerin benzerlik düzeyi, ön işleme adımları ve değerlendirme stratejisi gibi birçok faktörden etkilendiği görülmektedir. Bu yönüyle çalışma, RPA süreçlerine gelen yapılandırılmamış e-posta içeriklerinin otomatik olarak sınıflandırılması ve doğru iş

süreçlerine yönlendirilmesi için ML ve NLP tabanlı yaklaşımların uygulanabilir olduğunu ortaya koymaktadır.

4.3 Deney Sürecinde Karşılaşılan Sorunlar ve Çözüm Yöntemleri

Bu tez çalışmasının deneysel süreci boyunca, model performanslarının değerlendirilmesi ve sonuçların güvenilirliği açısından dikkatle ele alınması gereken çeşitli teknik ve metodolojik sorunlarla karşılaşılmıştır. Bu sorunlar, özellikle doğal dilde üretilen metin verilerinin doğası ve kullanılan modellerin özellikleriyle doğrudan ilişkili olup, yapılan ek analizler ve iyileştirme adımları ile sistematik biçimde ele alınmıştır.

İlk deneysel aşamalarda, bazı klasik ML modellerinin %100 doğruluk (Accuracy = 1.00) değerleri ürettiği gözlemlenmiştir. İlk bakışta bu durum yüksek model başarımı olarak yorumlanabilecek olsa da metin sınıflandırma problemlerinde bu denli yüksek performans değerlerinin çoğu zaman veri sızıntısı (data leakage) veya aşırı öğrenme (overfitting) kaynaklı olabileceği bilinmektedir. Bu nedenle, elde edilen sonuçların geçerliliğini sorgulamak amacıyla veri seti ayrıntılı biçimde yeniden incelenmiştir. Yapılan analizler sonucunda, veri setinde aynı veya yüksek derecede benzer içeriklere sahip e-posta kayıtlarının bulunduğu ve satır bazlı eğitim–test veri seti ayrımının, bu benzer örneklerin farklı kümelere düşmesine yol açtığı tespit edilmiştir. Bu durumun, modelin test verisini önceden görmüş olmasına neden olarak yapay biçimde yüksek başarı değerleri ürettiği saptanmıştır.

Bu problemi gidermek amacıyla, benzerlik temelli gruplama (group_id) yaklaşımı uygulanmıştır. E-posta metinleri, karakter tabanlı TF-IDF temsilleri üzerinden hesaplanan kosinüs benzerliği kullanılarak gruplandırılarak yüksek benzerlik gösteren örnekler aynı grup altında birleştirilmiştir. Ardından, eğitim ve test ayrımı satır bazında değil, grup bazında gerçekleştirilmiş ve değerlendirme sürecinde group-level çapraz doğrulama yöntemi kullanılmıştır. Bu yaklaşım sayesinde, benzer içeriklerin eğitim ve test kümelerine aynı anda düşmesi engellenmiştir. Veri sızıntısı ve aşırı öğrenme riski önemli ölçüde azaltılmıştır. Elde edilen yeni sonuçlar, model performanslarının daha gerçekçi olduğunu göstermiştir.

Deney sürecinde karşılaşılan bir diğer önemli zorluk ise, transformer tabanlı BERTurk modelinin eğitimi sırasında ortaya çıkan donanımsal bellek ve zaman kısıtları olmuştur. BERT tabanlı modellerin yüksek parametre sayısına sahip olması ve uzun metinleri bağlamsal olarak işlemesi nedeniyle bellek kullanımı ve işlem süresi, klasik ML yöntemlerine kıyasla daha yüksek olabilmektedir. Bu çalışmada, eğitim ortamında karşılaşılan bellek aşımı (out-of-memory) riskini azaltmak amacıyla, batch size, maksimum token uzunluğu (sequence length) ve gradient accumulation gibi eğitim parametrelerinde optimizasyonlar yapılmıştır. Bu düzenlemeler, modelin kararlı biçimde eğitilmesini sağlamıştır. 4 adet NVIDIA RTX A6000 GPU destekli ortam sayesinde eğitim süreci verimli biçimde tamamlanmıştır.

Bununla birlikte, uygulanan bu önlemler sayesinde BERTurk modeli başarıyla eğitilmiş ve bağlamsal dil temsillerinin metin sınıflandırma performansına olan katkısı güvenilir biçimde gözlemlenebilmiştir. Elde edilen bulgular, transformer tabanlı modellerin yüksek hesaplama maliyetine rağmen, uygun değerlendirme stratejileri ve donanım farkındalığı ile gerçekçi ve güçlü sonuçlar üretebildiğini ortaya koymaktadır.

Sonuç olarak, deney sürecinde karşılaşılan bu sorunlar ve uygulanan çözüm yaklaşımları, yalnızca model performansını artırmakla kalmamış aynı zamanda elde edilen sonuçların geçerliliğini, güvenilirliğini ve bilimsel tutarlılığını güçlendiren önemli adımlar olmuştur. Bu deneyimler, özellikle yapısal olmayan metin verileri üzerinde çalışan benzer çalışmalar için yol gösterici niteliktedir.

5. SONUÇ ve ÖNERİLER

Bu bölümde, tez çalışması kapsamında elde edilen bulgular genel olarak değerlendirilmekte ve çalışmanın sonuçları ortaya konulmaktadır. Ayrıca, elde edilen sonuçlar doğrultusunda gelecekte yapılabilecek araştırmalar, kurumsal uygulamalar ve geliştiricilere yönelik öneriler sunulmaktadır.

Sonuç:

Bu tez çalışmasında, kurumsal ortamlarda RPA robotlarına iletilen serbest yapılı e-posta içeriklerinin, ilgili iş süreçlerine otomatik olarak atanması problemi ele alınmıştır. Çalışma kapsamında, NLP teknikleri kullanılarak e-posta metinleri sayısal temsillere dönüştürülmüştür. Klasik ML algoritmaları ve transformer tabanlı derin öğrenme modelleri karşılaştırmalı olarak değerlendirilmiştir. Bu çalışma, RPA süreçlerinde kullanılan yapısal olmayan e-posta verilerinin otomatik ve güvenilir biçimde sınıflandırılmasına katkı sağlamaktadır.

Elde edilen sonuçlar, klasik ML yöntemlerinin uygun veri hazırlama ve değerlendirme stratejileriyle tatmin edici performanslar sunabildiğini göstermektedir. Bununla birlikte, transformer tabanlı BERTurk modelinin bağlamsal temsil gücü sayesinde daha yüksek doğruluk ve F1 skorları ürettiği gözlemlenmiştir. Özellikle group-level değerlendirme yaklaşımı, model performanslarının daha gerçekçi biçimde ölçülmesine olanak sağlamış ve veri sızıntısı riskini önemli ölçüde azaltmıştır. Bununla birlikte, modelin yüksek hesaplama maliyeti ve donanım gereksinimleri, gerçek zamanlı sistemlerde kullanım planlanırken dikkate alınması gereken önemli faktörler arasında yer almaktadır.

Bu nedenle, daha büyük ölçekli üretken dil modellerinin kurumsal RPA süreçlerinde doğrudan kullanımı bu çalışma kapsamında öncelikli bir çözüm olarak değerlendirilmemiştir. Büyük dil modelleri metin anlama ve üretme görevlerinde güçlü yetenekler sunsa da bu modellerin kurum içi güvenli altyapıda çalıştırılması yüksek GPU kapasitesi, bellek, servis altyapısı, bakım ve izleme maliyeti gerektirmektedir. Ayrıca bu tez kapsamında ele alınan problem, açık uçlu metin üretimi veya serbest biçimli cevaplama değil, e-posta içeriklerinin önceden tanımlanmış RPA süreçlerine göre

sınıflandırılmasıdır. Bu nedenle, sınıflandırma görevi için daha düşük maliyetli, daha yönetilebilir ve kurumsal sistemlere entegrasyonu daha kolay olan klasik ML modelleri ile BERTurk modeli üzerinde durulmuştur. Büyük dil modellerinin kullanımı ise ancak yüksek doğruluk ihtiyacının işlem maliyeti, altyapı kapasitesi, veri güvenliği ve bakım gereksinimleriyle dengelenebildiği senaryolarda değerlendirilmelidir.

Bu çalışma kapsamında elde edilen bulgular, BERTurk modelinin sınıflandırma performansı açısından RF modeline kıyasla daha yüksek doğruluk ve makro F1 değerleri ürettiğini göstermektedir. Bununla birlikte model seçimi yalnızca performans metriklerine göre yapılmamalıdır. RF modeli, TF-IDF temsilleri üzerinde çalıştığından daha düşük hesaplama maliyeti gerektirmekte, CPU ortamında hızlı biçimde eğitilebilmekte ve tahmin aşamasında düşük gecikme süresi sunmaktadır. Bu durum, sınırlı donanım kaynaklarına sahip veya yüksek hacimli gerçek zamanlı işlem gerektiren kurumsal sistemler için önemli bir avantaj sağlamaktadır. Buna karşılık BERTurk modeli, milyonlarca parametre içeren transformer mimarisi nedeniyle eğitim sürecinde GPU desteği gerektirmekte, daha yüksek bellek ve işlem süresi kullanmakta ve yeniden eğitim süreçlerinde daha fazla operasyonel planlama ihtiyacı doğurmaktadır. Dolayısıyla BERTurk performans açısından üstünlük gösterse de donanım maliyeti, bakım gereksinimi ve ölçeklenebilirlik kriterleri dikkate alındığında RF modeli maliyet–performans dengesi açısından daha pragmatik bir alternatif olarak değerlendirilebilir. Model tercihi kurumun altyapı kapasitesi ve doğruluk gereksinimi doğrultusunda belirlenmelidir.

Bu çalışmanın temel katkılarından biri, benzerlik temelli grup oluşturma ile çapraz doğrulama sürecinin birlikte ele alınması ve RPA bağlamında gerçek operasyonel veriler üzerinde uygulanmasıdır. Elde edilen bulgular, e-posta tabanlı süreç tetikleme mekanizmalarının daha güvenilir ve otomatik hale getirilebileceğini göstermektedir.

Öneriler:

Bu tez çalışmasında elde edilen bulgular doğrultusunda, gelecekte yapılacak araştırmalar ve kurumsal uygulamalar için çeşitli öneriler sunulabilir. Öncelikle, çalışmada kullanılan veri seti bankacılık sektörüne ait belirli RPA süreçleriyle sınırlı olduğundan, gelecek çalışmalarda farklı sektörlerden ve daha fazla süreç türünden elde edilen e-posta

verileriyle modelin genelleme yeteneği değerlendirilebilir. Böylece geliştirilen sınıflandırma yaklaşımının farklı kurumsal süreçlerdeki başarısı daha kapsamlı biçimde analiz edilebilir.

Araştırmacılar için, benzer çalışmalarda yalnızca doğruluk metriğine bağlı kalınmaması ve makro F1, ağırlıklı F1 ve sınıf bazlı performans değerlerinin birlikte değerlendirilmesi önerilmektedir. Ayrıca metin verilerinde benzer içeriklerin eğitim ve test kümelerine aynı anda düşmesi veri sızıntısına neden olabileceğinden, grup temelli çapraz doğrulama yaklaşımlarının kullanılması önem taşımaktadır. Bu kapsamda, gelecekte yapılacak çalışmalarda farklı benzerlik eşik değerleri, farklı metin temsil yöntemleri ve farklı transformer tabanlı Türkçe dil modelleri karşılaştırmalı olarak incelenebilir.

Geliştiriciler için, sınıflandırma modellerinin RPA süreçlerine doğrudan gömülü şekilde değil, bağımsız ve servis tabanlı bir mimariyle entegre edilmesi önerilmektedir. Bu yaklaşım, modelin yeniden eğitilmesi, güncellenmesi veya farklı modellerle değiştirilmesi gerektiğinde mevcut RPA akışlarının yeniden tasarlanmasına gerek kalmadan sistemin sürdürülebilir biçimde yönetilmesine olanak sağlayacaktır. Ayrıca model çıktılarının güven skoru ile değerlendirilmesi, düşük güven düzeyine sahip tahminlerin manuel inceleme kuyruğuna yönlendirilmesi ve tüm tahminlerin loglanması, operasyonel risklerin azaltılması açısından önemlidir.

Kurumsal uygulamalar açısından, model seçiminin yalnızca en yüksek performans değerine göre yapılmaması önerilmektedir. BERTurk gibi transformer tabanlı modeller yüksek doğruluk ve F1 skorları sağlayabilse de donanım maliyeti, işlem süresi, bakım gereksinimi ve gerçek zamanlı kullanım koşulları dikkate alınmalıdır. Bu nedenle, yüksek doğruluk gerektiren süreçlerde transformer tabanlı modeller tercih edilebilirken, düşük gecikme süresi ve daha düşük altyapı maliyeti gerektiren süreçlerde RF gibi klasik makine öğrenmesi modelleri maliyet-performans dengesi açısından değerlendirilebilir.

Gelecek çalışmalarda, teknolojik gelişmelerle birlikte büyük dil modellerinin çalıştırılması için gerekli donanım, işlem gücü ve servis maliyetlerinin azalması durumunda, bu modellerin RPA süreçlerine gelen yapısal olmayan e-posta içeriklerinin sınıflandırılmasındaki etkinliği de incelenebilir. Ayrıca önerilen sistem mimarisinin gerçek zamanlı kurumsal RPA ortamına entegre edilerek canlı sistem performansının

ölçülmesi, kullanıcı geri bildirimleriyle modelin düzenli olarak güncellenmesi ve zamanla değişen süreç tanımlarına uyum sağlayabilen öğrenme yaklaşımlarının geliştirilmesi önemli araştırma yönleri olarak değerlendirilmektedir.

KAYNAKÇA

Aguirre, S., ve Rodriguez, A. (2017). Automation of a Business Process Using Robotic Process Automation (RPA): A Case Study. *Communications in Computer and Information Science* (s. 65-71). Springer. doi:10.1007/978-3-319-66963-2_7

Aktan, E. (2018). Büyük Veri: Uygulama Alanları, Analitiği ve Güvenlik Boyutu. *Bilgi Yönetimi*(1), 1-22.

Arslankaya, S., ve Toprak, Ş. (2021). Makine Öğrenmesi ve Derin Öğrenme Algoritmalarını Kullanarak Hisse Senedi Fiyat Tahmini. *International Journal of Engineering Research and Development*, 178-192. doi:https://doi.org/10.29137/umagd.771671

Axmann, B., ve Harmoko, H. (2020). Robotic Process Automation: An Overview and Comparison to Other Technology in Industry 4.0. *10th International Conference on Advanced Computer Information Technologies (ACIT)* (s. 559-562). Deggendorf, Germany: IEEE. doi:10.1109/ACIT49673.2020.9208907

Aylak, B., Oral, O., ve Yazıcı, K. (2021). Yapay Zeka ve Makine Öğrenmesi Tekniklerinin Lojistik Sektöründe Kullanımı. *El-Cezeri*(8), 74-93.

Baviskar, D., Ahirrao, S., Potdar, V., ve Kotecha, K. (2021). Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions. *IEEE Access*. doi:10.1109/ACCESS.2021.3072900

Blumberg, R., ve Atre, S. (2003). The Problem with Unstructured Data. *DM Review*, 42-46.

Chen, Y., Garcia, E., Gupta, M., Rahimi, A., ve Cazzanti, L. (2009). Similarity-based Classification: Concepts and Algorithms. *Journal of Machine Learning Research*(10).

Cho, S., Moon, J., Bae, J., ve Kang, J. (2023). A Framework for Understanding Unstructured Financial Documents Using RPA and Multimodal Approach. *Electronics*, 939.

Devlin, J., Chang, M.-W., Lee, K., ve Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding . *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, 4171-4186.

Eberendu, A. C. (2016). Unstructured Data: an overview of the data of Big Data. *International Journal of Computer Trends and Technology (IJCTT)*, 46-50. doi:10.14445/22312803/IJCTT-V38P109

Gharehchopogh, F. S., ve Khalifehlou, Z. A. (2011). Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing. *2011 5th International Conference on Application of Information and Communication Technologies, AICT 2011* (s. 1-4). IEEE. doi:10.1109/ICAICT.2011.6111017

Göker, H., ve Tekedere, H. (2017). FATİH Projesine Yönelik Görüşlerin Metin Madenciliği Yöntemleri İle Otomatik Değerlendirilmesi. *Bilişim Teknolojileri Dergisi*, 291-299. doi:https://doi.org/10.17671/gazibtd.331041

Griggs, T., Liu, X., Yu, J., Kim, D., Chiang, W.-L., Cheung, A., ve Stoica, I. (2024). *Mélange: Cost Efficient Large Language Model Serving by Exploiting GPU Heterogeneity.* arXiv:2404.14527.

Günerkan, M., Şahinaslan, E., ve Şahinaslan, Ö. (2019). Gümrük Beyannamesi Sürecinde Öğrenmeye Dayalı Algoritmaların Etkinliğinin İncelenmesi. *Acta Infologica, 6*(2), 175-188. doi:doi.org/10.26650/acin.1057060

Hendrawan, I. R., Utami, E., ve Hartanto, A. D. (2022). Comparison of Naïve Bayes Algorithm and XGBoost on Local Product Review Text Classification. *Edumatic: Jurnal Pendidikan Informatika*, 143-149. doi:10.29408/edumatic.v6i1.5613

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., ve Brown, D. (2019). Text Classification Algorithms: A Survey. *Information*. doi:https://doi.org/10.3390/info10040150

Kumar, M. (2022). Robotic Process Automation (RPA) in Healthcare Insurance: Streamlining Administrative Tasks. *Progress in Medical Sciences*, 1-6. doi:10.47363/PMS/2022(6)E140

Lahitani, A., Permanasari, A., ve Setiawan, N. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment. *International Conference on Cyber and IT Service Management (CITSM)* (s. 1-6). IEEE.

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

Lu, H., Li, Y., Chen, M., Kim, H., ve Serikawa, S. (2017). Brain Intelligence: Go beyond Artificial Intelligence. *Mobile Networks and Applications*(23), 368–375.

Machireddy, J. R., ve Devapatla, H. (2022). Leveraging Robotic Process Automation (RPA) with AI and Machine Learning for Scalable Data Science Workflows in Cloud-Based Data Warehousing Environments. *Australian Journal of Machine Learning Research ve Applications*, 234-262.

Mohamed, S. A., Mahmoud, M. A., Mahdi, M. N., ve Mostafa, S. A. (2022). Improving Efficiency and Effectiveness of Robotic Process Automation in Human Resource Management. *Sustainability*, 3920. doi:https://doi.org/10.3390/su14073920

Örpek, Z., Örpek, A., ve Şahinaslan, Ö. (2026). Classification of Unstructured Text for RPA: A Comparative Machine Learning Analysis. *Boletim da Sociedade Paranaense de Matemática, 44*(5), 1-12.

Özdem, H., ve Bora, M. P. (2022). Türkiye'de Robotik Süreç Otomasyonu. *Bilgisayar Bilimleri ve Teknolojileri Dergisi*, 01-09. doi:https://doi.org/10.54047/bibted.1008340

Punia, S. K., Stephan, T., Deverajan, G. G., ve Patan, R. (2021). Performance Analysis of Machine Learning Algorithms for Big Data Classification: ML and AI-Based

Algorithms for Big Data Analysis. *International Journal of E-Health and Medical Communications*, 60-75. doi:10.4018/IJEHMC.20210701.oa4

Seker, S. E. (2015). Doğal Dil İşleme(Natural Language Processing). *YBS Ansiklopedi, 2*(4), 14-22.

Selvi, O. (2016). Yapısal Olmayan Verilerin Büyük Veri Analiz Yöntemleri İle İşlenmesi Ve Yapısal Olan Verilerle İlişkilendirilmesine Yönelik Bir Platform: Sosyal Medya Temelli Tavsiye Motoru Geliştirme. *Doktora Tezi, Süleyman Demirel Üniversitesi*, (Yayın No. 442737). YÖKTEZ: https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp adresinden alındı

Shepard, A., ve Naheed, N. (2021). Application of Data Transformation Techniques and Train-Test split. *International Conference on Computational Science and Computational Intelligence (CSCI)*, (s. 58-63).

Srinivasarao, U., ve Sharaff, A. (2023). Spam email classification and sentiment analysis based on semantic similarity methods. *International Journal of Computational Science and Engineering*(26), 65-77.

Sun, C., Qiu, X., Xu, Y., ve Huang, X. (2019). How to Fine-Tune BERT for Text Classification. *China national conference on Chinese computational linguistics*, 194-206.

Syed, R., Suriadi, S., Adams, M., Bandara, W., Leemans, S., Ouyang, C., . . . Reijers, H. (2020). Robotic Process Automation: Contemporary themes and challenges. *Computers in Industry*(115).

Şahinaslan, E. (2023). İş Süreci Optimizasyonu: Yöntem, Teknoloji, Riskler ve Fırsatlar. *Journal of Academic Projection (JAP)*, 570-604.

Şahinaslan, E., Günerkan, M., ve Şahinaslan, Ö. (2023). Makine Öğrenmesinde Kategorik Veri Kodlama Tekniğinin Kullanımına Alternatif Bir Çözüm Yöntemi. *Journal of Intelligent Systems: Theory and Applications, 6*(1), 1-11. doi:10.38016/jista.1140499

Taş, M. (2025). Türkçe Oltalama E-Postalarının Anlamsal Tespiti: Doğal Dil İşleme ve Derin Öğrenme Tabanlı Bir Yaklaşım. *Siber Güvenlik ve Dijital Ekonomi*, 29-42.

Toprak, A. (2024). Automatic document verification with natural language processing. 15.01.2025 tarihinde İEEE web: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10711295 adresinden alındı

Venigandla, K. (2022). Integrating RPA with AI and ML for Enhanced Diagnostic Accuracy in Healthcare. *Power System Technology*, 33-42. doi:10.52783/pst.251

Viggiato, M., Paas, D., Buzon, C., ve Bezeme, C.-P. (2023). Identifying Similar Test Cases That Are Specified in Natural Language. *IEEE Transactions on Software Engineering, 49*(3), 1027-1043.

Vujović, Ž. (2021). Classification Model Evaluation Metrics. *(IJACSA)International Journal of Advanced Computer Science and Applications, 12*(6), 599-606.

Wewerka, J., ve Reichert, M. (2021). Robotic Process Automation - A Systematic Mapping Study and Classification Framework. *Enterprise Information Systems*. doi:10.1080/17517575.2021.1986862

Wirth, R., ve Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, *1*, s. 29-39.

Yadav, M., Choudhury, T., Chandel, A., ve Quttainah, M. A. (2025). Natural Language Processing for Capability Enhancement of RPA. *Intelligent Robotic Process Automation: Development, Vulnerability and Applications*, 257-282. doi:10.4018/979-8-3693-4365-4.ch011

ÖZ GEÇMİŞ

Zeynep ÖRPEK

Eğitim

Yüksek Lisans	2024	Maltepe Üniversitesi, Bilgisayar Mühendisliği
Lisans	2019	Karabük Üniversitesi, Bilgisayar Mühendisliği

İş

2025 – Halen	KFT Bilişim Sistemleri A.Ş.	AR-GE Yazılım Mühendisi
2019 – 2025	Vakıf Katılım Bankası	AR-GE Yazılım Mühendisi

Yayınlar

Örpek, Z., Örpek, A., ve Şahinaslan, Ö. (2026). Classification of unstructured text for RPA: A comparative machine learning analysis. *Boletim da Sociedade Paranaense de Matemática*, 44(5).

Örpek, Z., Şahinaslan, Ö. (2026). Robotik Süreç Otomasyonu Süreçlerinde Yapılandırılmamış E-postaların Makine Öğrenmesi Yöntemleriyle Sınıflandırılması. *II. International Graduate Students Interdisciplinary Symposium (IGSIS)*

Örpek, Z., Esin, E. M., (2025). Banka ATM Sistemlerinin Simülasyonu. *Bursa 4. Uluslararası Matematik ve Mühendislik Kongresi.*

Tural, B., Örpek, Z., Destan, Z. ve Alp, E. (2026). Information Access Strategies in Large Language Models: Base Model, Fine-Tuning, and RAG Analysis. *2026 8th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)*.

Tural, B. ve Örpek, Z. (2026). Core Components of Enterprise AI Agents. *2026 8th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)*.

Tural, B., Örpek, Z., ve Destan, Z. (2025, Kasım). Future AI tools: LLM and frameworks. *2025 9th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*.

Tural, B., Örpek, Z., ve Destan, Z. (2025, Haziran). Ethics in the age of artificial intelligence. *2025 3rd Cognitive Models and Artificial Intelligence Conference (AICCONF)*.

Tural, B., Örpek, Z., Destan, Z., Çoşkuner, İ., Alkan, M. A., ve Kazova, F. (2025, Mayıs). The impact of new technologies on digital transformation with RPA. *2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)*.

Tural, B., Örpek, Z., ve Destan, Z. (2024, Aralık). Retrieval-Augmented Generation (RAG) and LLM integration. *2024 8th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*.

Örpek, Z., Tural, B., ve Destan, Z. (2024, Kasım). SWOT analysis of open-source language models. *2024 8th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 1–4.

Örpek, Z., Tural, B., ve Destan, Z. (2024, Eylül). The language model revolution: LLM and SLM analysis. *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*.

Kulaglic, A., Örpek, Z., Kayı, B., ve Ozmen, S. (2024, Ocak). Analysis of emotions and movements of Asian and European facial expressions. *Advances in Science, Technology and Engineering Systems Journal*, *9*(1), 42–48.

Karadağ, B., Erbaşı, Z., Karagöz, K. T., ve Özmen, S. (2022, Aralık). Digital Academy for Mentally Disabled. *Orclever Proceedings of Research and Development*, *1*(1), 58–74.