SAKARYA
ÜNİVERSİTESİ

Doğal Dil İşleme (NLP) ile E-Posta Spam Tespiti ve Sınıflandırması

2025-2026

Younes Rahebi B221210588

Veri Madenciliği

Prof.Dr. NİLÜFER YURTAY
Arş.Gör. SEDA UÇAR

BİLGİSAYAR VE BİLİŞİM BİLİMLERİ FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

2. Özet

Bu projenin temel amacı, e-posta iletişiminde büyük bir sorun teşkil eden istenmeyen iletilerin (Spam) otomatik olarak tespit edilmesini sağlayan bir makine öğrenmesi modeli geliştirmektir. Proje kapsamında, literatürde yaygın olarak kullanılan Enron Spam veri seti ele alınmıştır. Çalışma sürecinde öncelikle veri temizleme ve ön işleme adımları uygulanmış, ardından metin verileri Doğal Dil İşleme (NLP) teknikleri kullanılarak sayısal vektörlere (TF-IDF) dönüştürülmüştür. Modelleme aşamasında Logistic Regression ve Multinomial Naive Bayes algoritmaları kullanılmıştır. Yapılan testler sonucunda Logistic Regression modelinin %98'in üzerinde bir doğruluk (accuracy) oranı ve yüksek AUC skoru ile Spam tespitinde başarılı bir performans sergilediği görülmüştür. Bu çalışma, metin madenciliği tekniklerinin siber güvenlik ve veri filtreleme alanındaki etkinliğini ortaya koymaktadır.

3. Giriş

Günümüzde dijital iletişimin en yaygın aracı olan e-postalar, "Spam" olarak adlandırılan, istenmeyen, reklam içerikli veya zararlı yazılım barındıran iletilerin hedefi haline gelmiştir. Spam e-postalar, sadece kullanıcıların zamanını almakla kalmayıp, güvenlik riskleri de oluşturmaktadır. Bu veri madenciliği probleminin tanımı; bir e-postanın içeriğine bakarak onun güvenli (Ham) veya zararlı (Spam) olduğuna karar verebilen bir sistem tasarlamaktır.

Projenin motivasyonu, yapısal olmayan (unstructured) metin verilerinin işlenerek anlamlı bilgiler çıkarılması ve yüksek başarımlı bir filtreleme sistemi oluşturulmasıdır. Projeden beklenen çıktılar; temizlenmiş bir veri seti, metin sınıflandırma yeteneği olan eğitilmiş modeller ve bu modellerin performansını gösteren karşılaştırmalı analiz raporlarıdır. Bu süreçte Python programlama dili ve Scikit-learn, Pandas, Seaborn gibi veri bilimi kütüphaneleri kullanılmıştır.

4. Literatür / Arka Plan

Metin sınıflandırma ve Spam tespiti, veri madenciliği ve yapay zeka alanında uzun yıllardır çalışılan bir konudur.

- **Naive Bayes:** Olasılık temelli yaklaşımı (Bayes Teoremi) sayesinde metin sınıflandırmada "baseline" (temel) model olarak kabul edilir. Özellikle kelime bağımsızlığı varsayımıyla hızlı çalışır.

- **TF-IDF (Term Frequency-Inverse Document Frequency):** Metinleri sayısal verilere dönüştürürken, bir kelimenin doküman içindeki sıklığı ile tüm veri setindeki nadirliğini dengeleyerek, o kelimenin ayırt ediciliğini ortaya çıkaran bir yöntemdir.
- **Kavramsal Çerçeve:** Proje, CRISP-DM (Cross Industry Standard Process for Data Mining) benzeri bir akış izleyerek; veriyi anlama, hazırlama, modelleme ve değerlendirme adımlarını takip eder.

5. Veri Seti

5.1 Veri Kaynağı

Projede kullanılan veri seti, Enron şirketinin kamuya açılan e-posta yazışmalarından derlenen ve literatürde "Enron Spam Dataset" olarak bilinen açık kaynaklı bir veri setidir. Veri seti, ham ve spam olarak etiketlenmiş binlerce e-posta içermektedir.

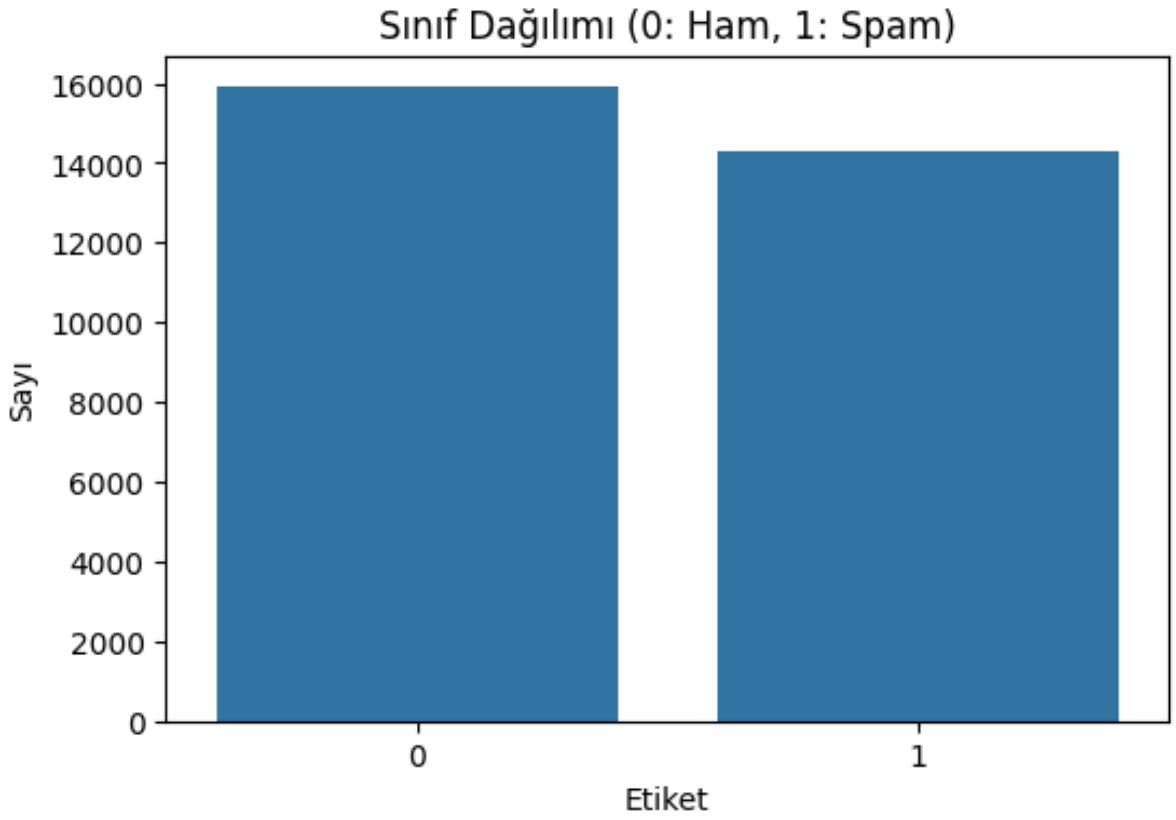
5.2 Değişkenlerin Açıklamaları

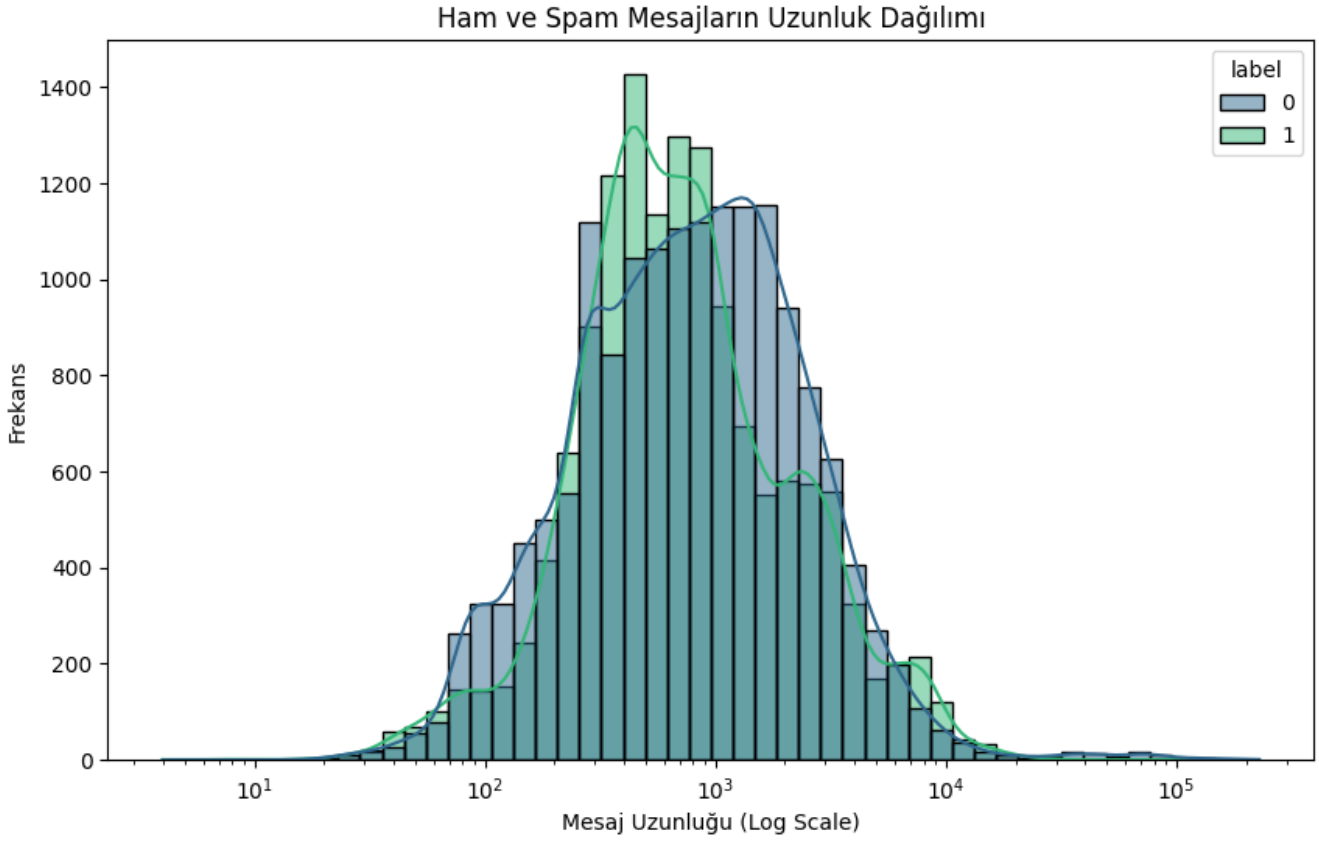
Veri seti ham haliyle şu temel değişkenleri içermektedir:

- **Subject (Kategorik/Metin):** E-postanın konu başlığı.
- **Message (Kategorik/Metin):** E-postanın gövde metni.
- **Spam/Ham (Kategorik/Hedef):** E-postanın sınıfı (Ham: Normal, Spam: İstenmeyen).
- **Date (Tarihsel):** E-postanın gönderim tarihi.

5.3 Veri Kalitesi Analizi

Veri setinde yapılan ilk incelemede, bazı satırlarda "Subject" veya "Message" alanlarının boş (Null) olduğu tespit edilmiştir. Ayrıca mükerrer (duplicate) kayıtlar gözlemlenmiştir. Sınıf dağılımı incelendiğinde, Ham ve Spam mesajlar arasında belirli bir dengesizlik olduğu, ancak modellemeye engel olacak seviyede olmadığı görülmüştür. Ayrıca Spam mesajların ve Ham mesajların karakter uzunlukları arasında belirgin farklar olduğu analiz edilmiştir.


Sınıf Dağılımı (0: Ham, 1: Spam)
Sayı
16000
14000
12000
10000
8000
6000
4000
2000
0
0
1
Etiket


Ham ve Spam Mesajların Uzunluk Dağılımı
label
0
1
Frekans
1400
1200
1000
800
600
400
200
0
10^1
10^2
10^3
10^4
10^5
Mesaj Uzunluğu (Log Scale)

6. Veri Ön İşleme

Ham metin verisini makine öğrenmesi algoritmalarına uygun hale getirmek için aşağıdaki adımlar uygulanmıştır:

1. **Veri Temizliği:** Eksik veriler (NaN) içeren satırlar silinmiş ve tekrar eden (duplicate) kayıtlar veri setinden çıkarılarak veri tekilleştirilmiştir.
2. **Metin Birleştirme:** Subject ve Message sütunları, daha zengin bir içerik elde etmek amacıyla birleştirilerek tek bir text sütunu oluşturulmuştur.
3. **Metin Temizleme (Text Cleaning):**
    - Tüm harfler küçük harfe (lowercase) çevrildi.
    - URL adresleri (http/www) Regex ile temizlendi.
    - Harf dışındaki karakterler (noktalama işaretleri, sayılar) kaldırıldı.
    - Fazla boşluklar silindi.
4. **Etiket Kodlama:** Hedef değişken olan Spam/Ham, sayısal değerlere dönüştürülmüştür (Ham: 0, Spam: 1).
5. **Veri Ayırma:** Veri seti, sınıf oranlarını koruyacak şekilde (stratified) **%80 Eğitim** ve **%20 Test** olarak ayrılmıştır.
6. **Özellik Çıkarımı (Vectorization):** Metin verisi, **TF-IDF Vectorizer** kullanılarak sayısal matrislere dönüştürülmüştür. Bu aşamada en sık geçen 5000 kelime özellik olarak seçilmiş ve İngilizce "stop words" (etkisiz kelimeler) filtrelenmiştir.

7. Yöntem ve Modelleme

7.1 Kullanılan Modeller

Bu bir "ikili sınıflandırma" (binary classification) problemi olduğu için denetimli öğrenme algoritmaları seçilmiştir:

1. **Logistic Regression:** Doğrusal ayrıştırılabilir problemlerde yüksek başarı gösterdiği ve yorumlanabilirliği yüksek olduğu için seçilmiştir.
2. **Multinomial Naive Bayes:** Metin verilerindeki kelime sıklıklarıyla (TF-IDF) çok iyi çalıştığı ve eğitim süresi kısa olduğu için tercih edilmiştir.

7.2 Model Kurulumu

Modeller, veri sızıntısını (data leakage) önlemek amacıyla Scikit-learn kütüphanesinin **Pipeline** yapısı içinde kurulmuştur. Logistic Regression modelinde ngram_range=(1,2) kullanılarak sadece kelimeler değil, kelime grupları da (bigram) analize dahil edilmiştir.

8. Sonuçlar ve Değerlendirme

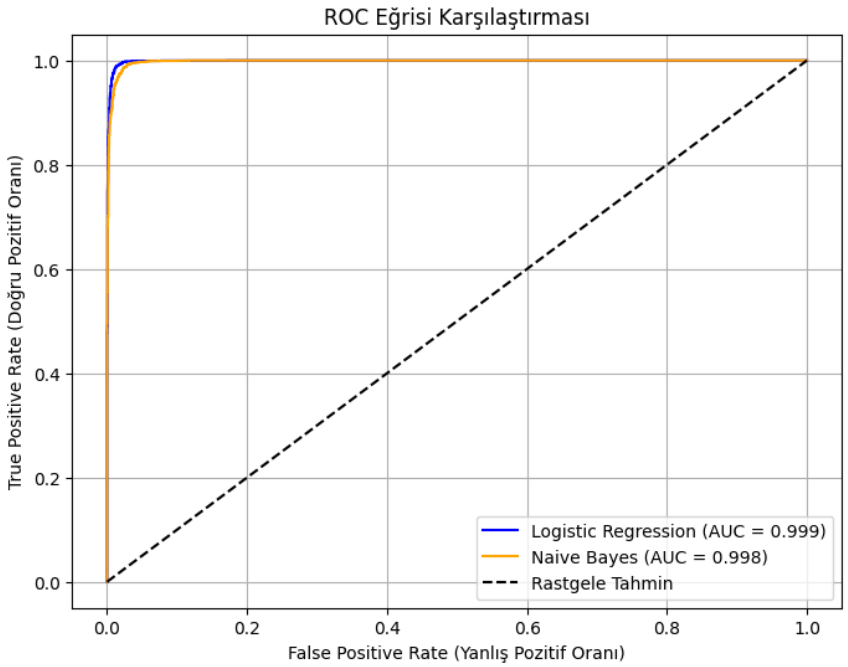
8.1 Performans Ölçütleri

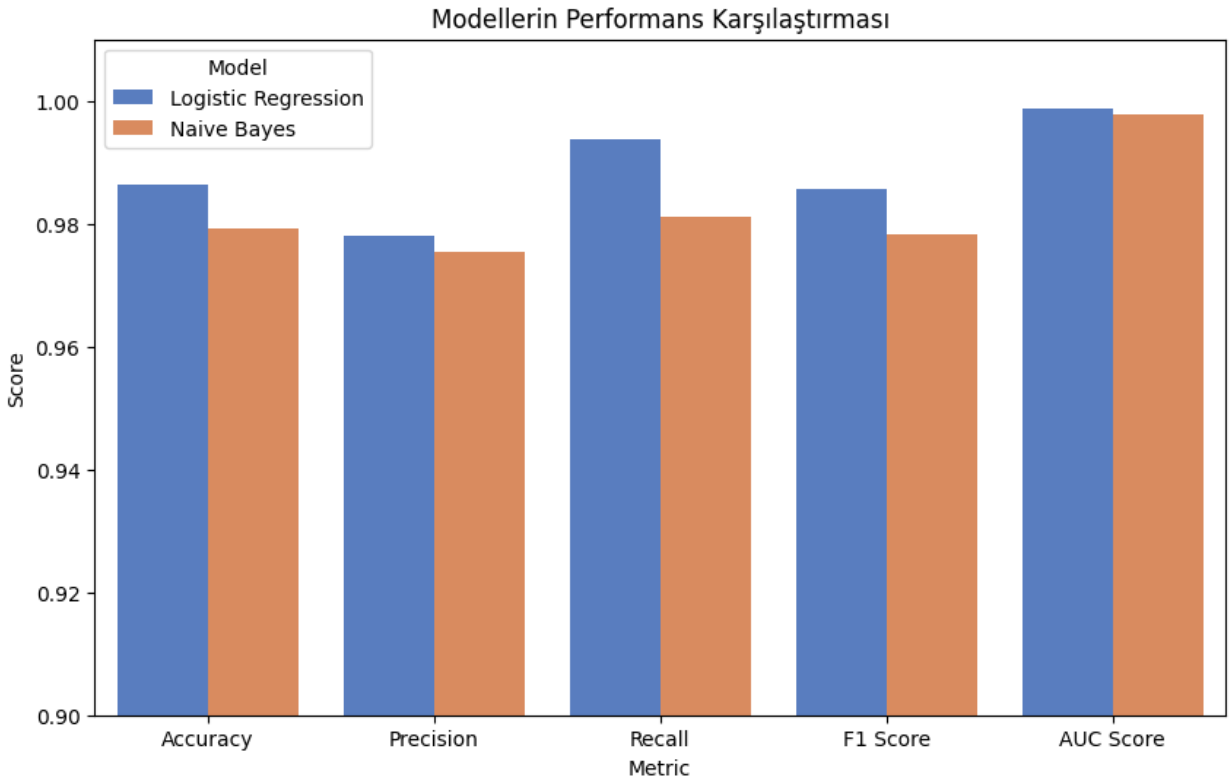
Modellerin başarısı Accuracy, Precision, Recall, F1-Score ve AUC metrikleri ile değerlendirilmiştir. Spam tespitinde "Yanlış Pozitif" (Normal bir maile spam denmesi) hatasını minimize etmek için Precision, spamleri kaçırmamak için Recall değerlerine özellikle dikkat edilmiştir.

8.2 Karşılaştırmalar

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.986	0.977	0.993	0.985	0.999
Naive Bayes	0.979	0.975	0.981	0.978	0.998

Elde edilen sonuçlara göre **Logistic Regression** modeli, hem genel doğruluk hem de Recall (Spam yakalama) oranında Naive Bayes modeline göre çok az farkla daha iyi performans göstermiştir.





8.3 Sonuçların Yorumlanması

Modellerimiz %98 bandında bir başarıya ulaşmıştır. Özellikle Logistic Regression modelinin yüksek Recall değeri, neredeyse tüm spam e-postaları yakaladığını göstermektedir. Yanlış sınıflandırmalar incelendiğinde (Confusion Matrix), modelin normal e-postaları spam olarak işaretleme (False Positive) oranının çok düşük olduğu görülmüştür. Bu, kullanıcı deneyimi açısından olumlu bir sonuçtur.

9. Tartışma

Projenin Sınırlılıkları: Model, sadece eğitim setindeki kelime dağarcığı ile sınırlıdır. Yeni türetilen spam kelimelerini veya karmaşık "sarcasm" (iğneleme) içeren metinleri anlamakta zorlanabilir.
Gelecek Çalışmalar: Proje, LSTM veya BERT gibi Derin Öğrenme (Deep Learning) tabanlı dil modelleri kullanılarak geliştirilebilir. Ayrıca modelin canlı bir e-posta akışına entegre edilmesi sağlanabilir.

10. Sonuç

Bu projede, Enron veri seti kullanılarak uçtan uca bir veri madenciliği süreci işletilmiştir. Ham metin verileri başarıyla işlenmiş ve yüksek performanslı bir Spam tespit sistemi geliştirilmiştir. Elde edilen %98 üzerindeki doğruluk oranı, TF-IDF ve Logistic Regression kombinasyonunun bu tür metin sınıflandırma problemleri için halen çok geçerli, hızlı ve etkili bir çözüm olduğunu kanıtlamıştır.

11. Kaynakça

1. Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). Spam filtering with Naive Bayes-which Naive Bayes?. *CEAS*, 17-28.
2. Scikit-learn Developers. (2024). *Scikit-learn: Machine Learning in Python*.
3. Enron Spam Dataset. (n.d.). [Dataset Source Link].