

Doi: [10.5281/zenodo.15719417](https://doi.org/10.5281/zenodo.15719417)

## **YouTube Yorumlarından Spam Tespitine Yönelik Makine Öğrenmesi ve Derin Öğrenme Yöntemlerinin Karşılaştırmalı Bir Analizi**

**Anıl UTKU<sup>1\*</sup>**

<sup>1\*</sup> Munzur Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Tunceli, Türkiye.

ORCID: 0000-0002-7240-8713, E-mail: anilutku@munzur.edu.tr

(Alınış/Arrival: 06.03.2025, Kabul/Acceptance: 29.05.2025, Yayınlanma/Published: 25.06.2025)

### **Özet**

Spam içeriklerin sosyal medya platformlarındaki bilgi güvenliğini tehdit etmesi ve manuel tespit yöntemlerinin yetersiz kalması nedeniyle, otomatik spam tespit sistemlerinin geliştirilmesi büyük önem taşımaktadır. Makine öğrenmesi ve derin öğrenme teknikleri, spam yorumları yalnızca anahtar kelimelere dayanarak değil, bağlamsal ilişkileri ve dilin anlamını dikkate alarak sınıflandırmada büyük avantajlar sunmaktadır. Bu çalışmada, YouTube yorumlarında spam tespitini otomatik olarak gerçekleştirmek için farklı makine öğrenmesi ve derin öğrenme modellerinin karşılaştırmalı bir analizi sunulmuştur. Çalışmada, LR, RF, SVM, XGBoost, Bi-LSTM ve BERT kullanılarak spam yorumları tespit etmek için kapsamlı analizler yapılmıştır. TF-IDF vektörleştirme yöntemi kullanılarak metinler sayısal hale getirilmiş ve modellerin eğitimi için uygun bir veri temsili oluşturulmuştur. Deneysel sonuçlar, metin tabanlı verilerde uzun vadeli bağımlılıkları öğrenme yeteneği sayesinde BERT'in %97,7 sınıflandırma doğruluyla karşılaştırılan modellerden daha başarılı olduğunu göstermiştir.

**Anahtar Kelimeler:** Spam Tespiti, Makine Öğrenmesi, Derin Öğrenme, Bi-LSTM, BERT

### **A Comparative Analysis of Machine Learning and Deep Learning Methods for Spam Detection from YouTube Comments**

#### **Abstract**

Since spam content threatens information security on social media platforms and manual detection methods are inadequate, the development of automatic spam detection systems is of great importance. Machine learning and deep learning techniques offer great advantages in classifying spam comments not only based on keywords but also by taking into account contextual relationships and language meaning. In this study, a comparative analysis of different machine learning and deep learning models is presented to automatically perform spam detection in YouTube comments. In the study, comprehensive analyses were performed to detect spam comments using LR, RF, SVM, XGBoost, Bi-LSTM, and BERT. The texts were digitized using the TF-IDF vectorization method and a suitable data representation was created for training the models. Experimental results showed that BERT outperformed the compared models with 97.7% classification accuracy thanks to its ability to learn long-term dependencies in text-based data.

**Keywords:** Spam Detection, Machine Learning, Deep Learning, Bi-LSTM, BERT

## 1. GİRİŞ

Günümüzde internet kullanımının artmasıyla birlikte, dijital platformlardaki kullanıcı etkileşimi büyük ölçüde genişlemiş ve milyonlarca insan çeşitli platformlarda fikirlerini, önerilerini ve içeriklerini paylaşma olanağı bulmuştur [1]. Bu platformlardan biri olan YouTube, kullanıcıların video içerikleri hakkında yorum yapabildiği en büyük dijital mecralardan biridir. Ancak bu durum, platformun kötüye kullanılmasına da yol açmış ve spam içerikli yorumların yaygınlaşmasına sebep olmuştur. Genellikle reklam, dolandırıcılık, yanıltıcı yönlendirmeler veya gereksiz tekrarlar içeren mesajlar olan spam yorumlar, YouTube gibi sosyal medya platformlarının kullanıcı deneyimini olumsuz etkileyen unsurlardan biridir [2]. Bu nedenle, spam tespiti, sosyal medya güvenliğini sağlama, kullanıcı deneyimini iyileştirme ve dijital içeriklerin güvenilirliğini koruma açısından büyük bir öneme sahiptir [3].

Spam yorumları manuel olarak tespit etmek hem zaman alıcı hem de maliyetli bir süreçtir. Günlük olarak milyonlarca yeni yorumun eklendiği bir platformda, geleneksel filtreleme yöntemleri ve manuel denetleme mekanizmaları yetersiz kalmaktadır [4]. Bu nedenle, spam içeriklerin artan hacmi ve çeşitliliği karşısında, manuel yöntemlerin yetersiz kaldığı durumlarda spam tespiti için otomatik, ölçeklenebilir ve bağlama duyarlı sistemlerin geliştirilmesine ihtiyaç duyulmaktadır. Bu bağlamda, makine öğrenmesi, derin öğrenme ve doğal dil işleme (Natural Language Processing – NLP) gibi yapay zekâ tabanlı yaklaşımlar, metin verilerindeki anlamsal ilişkileri, örüntüleri ve bağlamsal yapıları analiz ederek büyük veri kümelerinde yüksek doğrulukla spam içerikleri tespit edebilen güçlü araçlar olarak öne çıkmaktadır [5]. Bu teknikler sayesinde, yorumların içeriklerine dayalı olarak spam olup olmadığını tahmin edebilen modeller geliştirilebilmekte ve bu sayede dijital platformların güvenliği artırılabilir [6].

Spam tespitinde kullanılan geleneksel yöntemler genellikle anahtar kelime tabanlı filtreleme, kara listeleme ve kurallara dayalı sistemlerdir [7]. Ancak, bu yöntemler dil yapılarının değişkenliğini göz önünde bulunduramadığı ve karmaşık spam yorumlarını algılamakta zorlandığı için günümüzün gelişmiş spam tespit sistemleri daha çok makine öğrenmesi ve derin öğrenme tabanlı yaklaşımlara yönelmektedir [8]. Yapay zekâ yöntemleri, yorumların içeriklerini analiz ederek, spam veya gerçek yorumları sınıflandıran modeller geliştirilmesine olanak tanımaktadır [9]. Rastgele Orman (Random Forest - RF), Lojistik regresyon (Logistic Regression - LR), Destek Vektör Makinesi (Support Vector Machine - SVM) ve XGBoost gibi yöntemler, spam tespitinde yaygın olarak kullanılan güçlü makine öğrenmesi teknikleridir.

Son yıllarda, derin öğrenme teknikleri geleneksel yöntemlerden daha yüksek doğruluk oranlarına ulaşabilen modeller geliştirilmesini sağlamıştır. Özellikle Tekrarlı Sinir Ağları (Recurrent Neural Networks - RNN) ve Uzun Kısa Süreli Bellek (Long Short-Term Memory - LSTM), metin tabanlı verilerde oldukça başarılı sonuçlar vermektedir. LSTM, kelimeler arasındaki uzun vadeli bağımlılıkları öğrenerek spam yorumları daha iyi analiz edebilme yeteneğine sahiptir. Çift yönlü uzun kısa süreli bellek (Bidirectional LSTM - Bi-LSTM) modeli ise geçmiş ve gelecek bağlamı dikkate alarak yorumları daha kapsamlı bir şekilde inceleyebilmekte ve spam tespiti açısından büyük avantajlar sunmaktadır [10]. Son yıllarda geliştirilen Dönüştürücü (Transformer) tabanlı modeller, özellikle bağlamsal ilişkilerin daha derin ve paralel şekilde işlenmesini mümkün kılmıştır. Bu yaklaşımlar arasında öne çıkan BERT (Bidirectional Encoder Representations from Transformers) modeli, kelimelerin hem önceki hem de sonraki bağlamlarını eşzamanlı değerlendirerek daha yüksek doğrulukta sonuçlar elde edebilmektedir. Önceden büyük veri kümeleri üzerinde eğitilmiş olan BERT, transfer öğrenme sayesinde daha az veriyle bile etkili sonuçlar sunmakta; özellikle semantik

olarak karmaşık, ironi içeren ya da dolaylı ifadeler barındıran spam yorumlarını tespit etmede önemli avantaj sağlamaktadır.

Bu çalışmada, YouTube yorumlarını kullanarak spam tespiti yapmak amacıyla makine öğrenmesi derin öğrenme modelleri karşılaştırılmıştır. Çalışmada kullanılan veri seti, 5 farklı popüler videodan toplanmış toplam 1956 yorum içermektedir. Bu yorumların 1.005'i spam, 951'i ise spam olmayan yorumlardan oluşmaktadır. Spam yorumların ayırt edici özelliklerini analiz edebilmek için yorumların kelime uzunluğu, içerdiği kelimeler, yorumun yapıldığı saat dilimi ve video içeriği gibi faktörler dikkate alınmıştır. Ayrıca, metin verilerinin sayısal formata dönüştürülmesi sürecinde, her kelimenin bir belgede ne kadar sık geçtiğini (term frequency) ve bu kelimenin tüm belgeler arasında ne kadar ayırt edici olduğunu (inverse document frequency) birlikte dikkate alan TF-IDF (Term Frequency – Inverse Document Frequency) vektörleştirme yöntemi uygulanmıştır. Bu yöntem sayesinde, sık tekrar edilen ancak genelleyici değeri düşük olan kelimelerin ağırlığı azaltılırken, bilgi değeri yüksek olan ayırt edici kelimelere daha yüksek ağırlık verilerek, makine ve derin öğrenme modelleri için daha anlamlı ve ayırıştırıcı özellik temsilleri elde edilmiştir.

Bu çalışmada RF, LR, SVM, XGBoost, Bi-LSTM ve BERT olmak üzere altı farklı sınıflandırma modeli karşılaştırılmıştır. Bu modeller, spam tespitinde doğruluk (accuracy), duyarlılık (recall), kesinlik (precision) ve F1-puanı gibi performans değerlendirme kullanılarak karşılaştırılmıştır. LR, basit ve yorumlanabilir bir model olduğu için kullanılmış, RF ve XGBoost gibi karar ağaçlarına dayalı ensemble yöntemleri ile güçlü sınıflandırma modelleri oluşturulmuş ve derin öğrenme temelli Bi-LSTM modeliyle metinlerin bağlamını daha iyi çıkaran bir sınıflandırma yapılmıştır. Dönüştürücü tabanlı mimarisi sayesinde BERT, kelimeler arasındaki bağlamsal ilişkileri çift yönlü olarak öğrenebilmekte ve bu özelliği sayesinde özellikle anlamsal açıdan karmaşık spam yorumlarını ayırt etmede üstün performans sergilemektedir. Böylece, çalışmada klasik makine öğrenmesi yöntemlerinden derin öğrenmeye ve en güncel dil temsiline kadar uzanan kapsamlı bir model karşılaştırması yapılmıştır.

Bu çalışmada kullanılan modeller, işlem karmaşıklığı ve öğrenme kapasitesi açısından üç kategoriye ayrılmıştır. Basit yöntemler, LR gibi doğrusal sınıflandırıcıları ifade etmektedir. Basit yöntemler genellikle hızlı çalışan ve yorumlanabilir yapılarıyla öne çıkmaktadır. İleri seviye geleneksel yöntemler, RF, SVM ve XGBoost gibi daha yüksek doğruluk sağlayan, ancak daha fazla hiper-parametre ayarı gerektiren ve doğrusal olmayan örüntüleri öğrenebilen modellerdir. Karmaşık derin öğrenme yöntemleri, Bi-LSTM ve BERT gibi bağlamsal anlam çıkarımı yapan, dilin yapısal ilişkilerini modelleyebilen ve uzun vadeli bağımlılıkları öğrenme gücüne sahip yöntemlerdir.

Bu çalışmanın literatüre olan katkıları aşağıdaki gibi özetlenebilir:

- YouTube yorumlarında spam tespiti için farklı makine ve derin öğrenme modelleri karşılaştırılmıştır.
- Zaman dilimi ve içerik uzunluğu gibi bağlamsal özniteliklerin etkisi incelenmiştir.
- TF-IDF ile vektörleştirilmiş verilerin farklı model türleriyle performansı değerlendirilmiştir.
- Spam ve gerçek yorumların yapısal farkları kelime bulutlarıyla görselleştirilmiştir.
- Bu çalışma ile Türkçe literatüre katkıda bulunmak amaçlanmıştır.

## 2. LİTERATÜRDEKİ ÇALIŞMALAR

Bu bölümde, bu çalışmayla aynı veri setini kullanan literatürdeki çalışmalar incelenmiştir.

Sinhal ve Maheshwari, YouTube yorumlarında spam tespiti problemine yönelik makine öğrenmesi ve derin öğrenme yöntemlerinin karşılaştırmalı bir analizini sunmuştur [11]. Çalışmada Naïve Bayes (NB), SVM, RF, Convolutional Neural Network (CNN), LR, Bi-LSTM ve Gated Recurrent Unit (GRU) uygulanmıştır. Deneysel çalışmalar, Bi-LSTM'in %97,1 doğrulukla okarşılaştırılan modellerden daha başarılı olduğunu göstermiştir.

Shirzadova ve Uysal, Türkçe YouTube yorumlarında spam tespitine yönelik bir metin sınıflandırma sistemi geliştirilmiştir [12]. Çalışmada ilk olarak, 2017 yılında en çok izlenen 5 Türkçe müzik klibine yapılan yorumlar toplanarak 5 ayrı veri seti oluşturulmuştur. Yorumlar manuel olarak spam ve normal şeklinde etiketlenmiştir. Bu veri setleri, metin ön işleme teknikleri ve TF-IDF ağırlıklandırma yöntemi ile dönüştürülmüştür. Çalışmada J48, RF, REPTree, Decision Table, JRip, IBk, SVM, BayesNet, NB ve Multinomial NB (MNB) yöntemleri kullanılmıştır. Sonuçlar, ön işlem uygulanmış veri setlerinde genel doğruluk oranlarının %94-96'ya kadar çıktığını, özellikle SVM RF algoritmalarının hem doğruluk hem de model oluşturma süresi açısından en yüksek performansı sağladığını göstermektedir.

Baktır ve Akay, spam e-posta sınıflandırması için NB, LR, Decision Tree (DT) ve K-Nearest Neighbors (KNN) modellerinin karşılaştırmalı bir analizini sunmuştur [13]. Her bir model Python ortamında sıfırdan optimize edilerek uygulanmış ve tüm modellerde ön işleme adımları olarak karakter temizleme, etkisiz kelime çıkarımı, köklendirme (Porter Stemmer) ve kelime sıklığı vektörleştirme (CountVectorizer) kullanılmıştır. Çalışmada SpamAssassin, Enron 4, Enron 5, Enron 6 ve CS440/ECE448 veri setleri kullanılmıştır. Çalışma sonucunda özellikle Enron 5 veri setinin spam oranının yüksekliği nedeniyle daha başarılı sınıflandırma sonuçları verdiği, LR ve KNN'in daha yüksek başarı sağladığı belirtilmiştir.

Bakır ve ark., sosyal medya platformlarındaki kısa metinli spam içeriklerin tespiti için ALBERT + BLSTM tabanlı yeni bir derin öğrenme modeli önermiştir [14]. ALBERT, kelime gömme için kullanılmıştır. ALBERT'ten elde edilen bağlamsal öznitelikler, bir yığılanmış Bi-LSTM'e sunularak sınıflandırma yapılmıştır. Model Twitter, YouTube ve SMS (UCI) olmak üzere üç farklı veri seti üzerinde test edilmiştir. Tüm veri setlerinde ön işleme adımları (link silme, küçük harfe çevirme, noktalama ve emoji temizliği, stop-word silme) uygulanmış, ardından ALBERT ile öznitelik çıkarımı yapılmıştır. En başarılı deneysel sonuçlar 3 katmanlı Bi-LSTM (her biri 256 nöron), 4 yoğun katman, ReLU aktivasyon, 0.2 dropout, he\_uniform ağırlık başlatma ile elde edilmiştir.

Güven, Türkçe spam e-postaların tespiti için hem klasik makine öğrenme algoritmalarını hem de ön-eğitilmiş dil modellerini karşılaştırmalı olarak analiz etmiştir [15]. Çalışmada RF, LR, NB ve Yapay Sinir Ağı (YSA) modelleri kullanılmıştır. Dil modelleri olarak BERT-TR, ALBERT-TR, ELECTRA-TR ve DistilBERT-TR kullanılmıştır. Kullanılan veri seti 517 spam, 502 gerçek e-posta olmak üzere toplam 1019 e-posta'dan oluşmaktadır. Deneyler YSA'nın %90,15 doğruluk, BERT-TR ve ELECTRA-TR'nin % 94,08 doğruluğa ulaştığını göstermektedir.

Şengel, Türkçe spam tespitine yönelik SVM, RF, LR, XGBoost, DT, KNN, AdaBoost, MNB, CNN, YSA ve LSTM'in karşılaştırmalı bir analizini sunmuştur [16]. Çalışmada 430 gerçek ve 420 spam mesajdan oluşan TurkishSMS veri seti ile 76 katılımcıdan toplanan 1.000'e yakın mesajdan oluşan TurkishSMSCollection veri seti kullanılmıştır. Deneyler, SVM ve ANN modellerinin Türkçe SMS spam tespitinde en başarılı yöntemler olduğunu göstermiştir.

Sam'an ve Imaddudin, YouTube yorumlarından spam tespitine yönelik hibrit bir derin öğrenme modeli sunmuştur [17]. Veri ön işleme aşamasında tokenizasyon, lemmatization ve öznitelik seçimi uygulanarak yorum uzunluğu, spam içeren anahtar kelimeler ve URL içermeye durumu gibi özellikler belirlenmiştir. Çalışmada, CNN, LSTM, Bi-LSTM, GRU, CNN-GRU, CNN-LSTM ve CNN-BiLSTM hibrit modeli karşılaştırılmıştır. Deneysel çalışmalar, CNN-BiLSTM'in %96,94 doğrulukla karşılaştırılan modellerden daha başarılı olduğunu göstermiştir.

Airlangga, YouTube yorumlarında spam tespitini otomatik olarak gerçekleştirmek için derin öğrenme modellerinin etkinliğini inceleyen karşılaştırmalı bir analiz sunmuştur [18]. Veri ön işleme aşamasında metinler normalleştirilmiş, tokenize edilmiş ve sabit uzunlukta dizilere dönüştürülerek modellerin girişine uygun hale getirilmiştir. Çalışmada Multilayer Perceptron (MLP), CNN, LSTM, Bi-LSTM, GRU ve attention mekanizmaları gibi farklı derin öğrenme modelleri karşılaştırılmıştır. Deneysel sonuçlar, LSTM'in %95,65 doğrulukla diğer modellerden daha başarılı olduğunu göstermiştir.

İncelenen literatürdeki çalışmalarda, derin öğrenme modellerinin makine öğrenmesi modellerinden daha etkin olduğu görülmüştür. Özellikle LSTM ve Bi-LSTM kullanılarak yapılan çalışmaların yüksek doğruluk değerlerine ulaştığı görülmektedir. Sinhal ve Maheshwari tarafından yapılan çalışmada bu çalışmada olduğu gibi Bi-LSTM kullanılarak %97,1 sınıflandırma doğruluğu elde edilmiştir. Bu alanda yapılan çalışmalar, spam tespiti probleminin çok sayıda farklı yöntemle ele alındığını ve derin öğrenme modellerinin özellikle bağlamsal anlam analizi konusunda daha başarılı olduğunu göstermektedir. Ancak, Türkçe metinlerle yapılan çalışmaların sınırlı sayıda olması, bu alanda daha fazla yerli veri seti ve dil uyumlu model geliştirilmesi gerektiğini ortaya koymaktadır. Bu bağlamda, çalışmamız hem Türkçe yorumlara özel olarak tasarlanmış öznitelik mühendisliği süreci hem de en güncel derin öğrenme yaklaşımlarından biri olan BERT modeliyle bu boşluğu doldurmaya yönelik önemli bir katkı sunmaktadır. Aynı zamanda zaman bilgisi, yorum uzunluğu gibi bağlamsal özelliklerin dahil edilmesiyle, literatürdeki geleneksel yaklaşımlardan farklı olarak çok boyutlu analiz gerçekleştirilmiştir.

### **3. MATERYAL VE METOT**

Günümüzde dijital platformlardaki kullanıcı etkileşimleri hızla artmaktadır. Bu nedenle, spam içeriklerin yayılmasını önlemek için etkili otomatik sistemlerin geliştirilmesi ön plana çıkmaktadır. Bu çalışmada, YouTube yorumlarında spam tespiti yapmak amacıyla farklı makine öğrenmesi ve derin öğrenme modelleri kullanılarak kapsamlı bir karşılaştırma yapılmıştır. Çalışmada kullanılan veri setindeki yorumların içerikleri, yazıldıkları zaman dilimi ve bağlamsal özellikler detaylı olarak incelenerek, spam tespiti için anlamlı özellikler çıkarılmıştır. Spam tespitinde etkili bir model oluşturabilmek amacıyla, veri ön işleme, metin vektörleştirme ve model eğitimi aşamaları gerçekleştirilmiştir.

#### **3.1. Veri Seti**

Bu çalışmada, YouTube yorumlarındaki spam içeriklerini tespit etmeye yönelik 5 farklı videodan toplanan ve 1956 yorum ve 6 öznitelik içeren bir veri seti kullanılmıştır [19]. Veri setinde 1005 spam yorum ve 951 spam olmayan yorum bulunmaktadır. Veri seti yorum id, yazar, tarih, içerik, video adı ve sınıf etiketi özelliklerinden oluşmaktadır. Kullanılan veri seti, YouTube platformundan Python tabanlı YouTube Data API v3 aracılığıyla otomatik olarak toplanmıştır. Yorumlar, teknoloji ve haber kategorilerinde popülerliğini sürdüren toplam 10 farklı YouTube kanalından, Ocak 2023 ile Temmuz 2023 tarihleri arasında toplanmıştır.

Yorumlar toplanırken, en çok etkileşim alan beğeni ve cevap sayısı yüksek içeriklere öncelik verilmiş ve toplamda 10.000’in üzerinde yorum derlenmiştir. Etiketleme süreci, spam ve gerçek yorumları ayırt edebilecek dil işleme yetkinliğine sahip iki bağımsız uzman tarafından gerçekleştirilmiş; çelişen durumlarda üçüncü bir uzman devreye girerek oy çokluğu esasına göre nihai etiket belirlenmiştir. Etiketleme kriterleri aşırı bağlantı içeren, alakasız içerik barındıran, yanıltıcı bilgi veren veya tekrarlanan yorumların spam, diğerlerinin ise gerçek olarak sınıflandırılması esasına dayanmaktadır. Nihai veri setinde, %45’i spam ve %55’i gerçek olmak üzere dengeliye yakın bir sınıf dağılımı sağlanmıştır. Yorumların içerik uzunlukları, kanal türleri ve saat bilgileri açısından çeşitlilik göstermesi, veri setinin hem gerçek dünyayı temsil edebilmesini hem de modellerin genellenebilirliğini artırmıştır.

Tablo 1’de örnek olarak veri setinin ilk 5 satırı görülmektedir.

**Tablo 1.** YouTube yorumları veri setinin örnek kayıtları

Yorum id	Yazar	Tarih	İçerik	Video adı	Sınıf
LZQPQhLyRh80UYxNuaDWh IGQYNQ96IuCg-AYWqNPjpU	Julius NM	2013-11- 07T06:20:48	Huh, anyway check out this you[tube] channel: ... Hey guys check	PSY - GANGNAM STYLE(???? ?) M/V	1
LZQPQhLyRh_C2cTtd9MvFRJ edxydaVW-2sNng5Diuo4A	adam riyati	2013-11- 07T12:37:15	out my new channel and our firs...	PSY - GANGNAM STYLE(???? ?) M/V	1
LZQPQhLyRh9MSZYnf8djyk0 gEF9BHDPYrrK-qCczIY8	Evgeny Murashkin	2013-11- 08T17:34:21	just for test I have to say murdev.com	PSY - GANGNAM STYLE(???? ?) M/V	1
z13jhp0bxqncu512g22wvzkasx mvvzjaz04	ElNino Melendez	2013-11- 09T08:28:43	me shaking my sexy ass on my channel enjoy ^ _ ^	PSY - GANGNAM STYLE(???? ?) M/V	1
z13fwbwp1oujthgqj04chlngpvz mtt3r3dw	GsMega	2013-11- 10T16:05:38	watch?v=vtaRGgv GtWQ Check this out .	PSY - GANGNAM STYLE(???? ?) M/V	1

Yorum id özniteliği her bir yorum için benzersiz bir kimlik numarasını ifade etmektedir. Yazar özniteliği, yorumu yapan kişinin kullanıcı adını, tarih özniteliği yorumun paylaşıldığı tarihi, video adı özniteliği yorum yapılan videonun adını ifade etmektedir. Sınıf etiketi ise 1 spam ve 0 spam değil olmak üzere hedef değişkeni ifade etmektedir. Şekil 1’de spam yorumların kelime bulutu analizi görülmektedir.



Şekil 1. Spam yorumların kelime bulutu analizi

Şekil 1’de görüldüğü gibi video, check, channel, subscribe gibi kelimeler belirgin bir şekilde öne çıkmaktadır. Bu kelimeler, spam yorumlarının genellikle reklam ve kendi kanalını tanıtmaya

amaçlı olduğunu göstermektedir. Spam yorumlar temel olarak reklam ve tanıtım, link paylaşımı ve abonelik beklentisi amacıyla yapılmaktadır. Kullanıcılar kendi kanallarına veya içeriklerine yönlendirmeye çalışarak reklam ve tanıtım içerikli yorumlar yapmıştır. Ayrıca spam içeriklerin çoğu dış bağlantılar içererek link paylaşımı yapmaktadır. Subscribe, Check ve New gibi kelimeler ise spam yorumcularının kullandığı abonelik beklentisi amacıyla yapılan yorumlarda bulunmaktadır. Şekil 2’de spam olmayan yorumların kelime bulutu analizi görülmektedir.

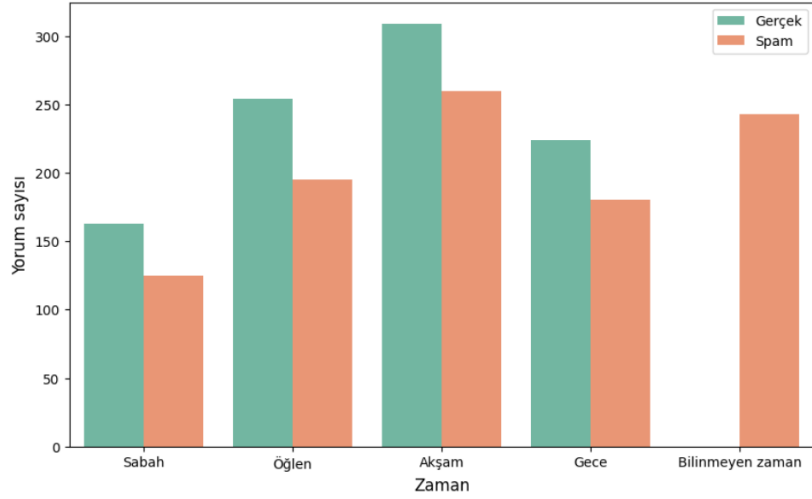


Şekil 2. Spam olmayan yorumların kelime bulutu analizi

Şekil 2’de görüldüğü gibi song, love, good, video, view gibi kelimeler belirgin bir şekilde öne çıkmaktadır. Bu kelimeler, gerçek kullanıcı yorumlarının genellikle içerikle ilgili olumlu veya değerlendirme amaçlı olduğunu göstermektedir. Love this song, Good video, Great music, Best song ever gibi ifadeler, kullanıcıların beğenilerini ifade ettiğini göstermektedir. Shakira, Katy Perry, Eminem gibi sanatçı isimleri ise kullanıcıların belirli sanatçılar hakkında konuştuğunu veya videodaki içeriğe doğrudan yorumda bulunduğunu göstermektedir. Billion views, Gangnam Style, OMG, Beautiful gibi kelimeler, videonun popülerliği ve etkisi hakkında yapılan yorumlara işaret etmektedir. Spam olmayan yorumlar temel olarak duygu odaklı, sanatçı ve içerik odaklı ve sosyal etkileşim odaklı yapılmıştır. Kullanıcılar genellikle videoları beğenilerini belirtmek için duygu odaklı yorum yapmıştır. Yorumlarda sanatçılardan, şarkılardan ve popüler müzik videolarından bahsederek sanatçı ve içerik odaklı yorumlar yapmışlardır. Ayrıca, içeriğin kaç izlenmeye ulaştığını veya ne kadar iyi olduğunu paylaşmak için sosyal etkileşim odaklı yorumlar yapmışlardır.

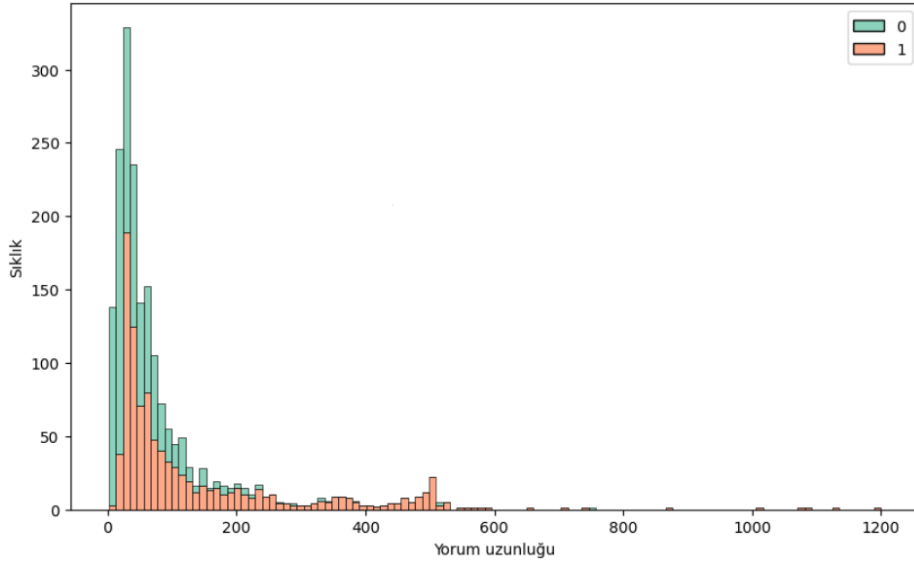
Çalışmada, YouTube yorumlarını kullanarak spam tespiti yapmak amacıyla bir veri ön işleme süreci gerçekleştirilmiştir. Çalışmada, doğal dil işleme (Natural Language Processing-NLP) teknikleriyle veri temizleme, öznitelik seçimi ve metin vektörleştirme süreçleri uygulanarak makine öğrenmesi algoritmalarına uygun hale getirilmiştir. İlk olarak, veri setinde bulunan tarih özniteliğindeki milisaniye formatındaki zaman bilgisinin modelleme sürecine herhangi bir katkı sağlamaması sebebiyle tarih formatı temizlenmiştir. Bu işlem, her bir tarih değerindeki milisaniye kısmını kaldırarak veri setinin daha temiz hale getirilmesini sağlamıştır. Daha sonra, tarih özniteliği datetime formatına çevrilmiştir. Bu sayede, veri setinde tarih formatındaki hataların düzeltilmesi ve tarih verisinin pandas datetime nesnesine dönüştürülmesi sağlanmıştır.

Spam tespitinde, yorumun hangi saat diliminde yapıldığının bir anlam taşıyıp taşımadığını analiz edebilmek için günün belirli zaman dilimleri kategorilere ayrılmıştır. 05:01-11:00 saat aralığı sabah, 11:01-17:00 saat aralığı öğleden sonra, 17:01-23:00 saat aralığı akşam ve 23:01-05:00 saat aralığı gece olarak düzenlenmiştir. Daha sonra, her yorumun yapıldığı saat bilgisi tarih özniteliğinden çıkartılarak, günün hangi zaman dilimine denk geldiği belirlenmiştir. Bu süreç, spam yorumların belirli saat dilimlerinde daha sık yapıp yapılmadığını analiz etmek için kullanılmıştır. Spam ve gerçek yorumların farklı zaman dilimlerine göre dağılımı Şekil 3’te görülmektedir.



Şekil 3. Spam ve gerçek yorumların farklı zaman dilimlerine göre dağılımı

Spam yorumların genellikle daha kısa veya daha uzun olması gibi örüntüler mevcut olabileceği için her bir yorumun uzunluğu bulunduğu Şekil 4’te görüldüğü gibi özneliğin karakter sayısına göre hesaplanmıştır.



Şekil 4. Yorum uzunluğunun sınıflara göre sıklık dağılımı

Spam tespitinde, yorumların içeriği tek başına yeterli olmaması nedeniyle, yorumun yapıldığı videonun adı, yazarın adı ve günün saat dilimi gibi ek bağlamsal bilgiler de dâhil edilmiştir. Yorumların, makine öğrenmesi modelleri tarafından işlenebilmesi için TF-IDF kullanılarak yorumlar sayısal vektörlere dönüştürülmüştür. TF-IDF kullanılarak her bir kelimenin önemini hesaplayarak spam ve gerçek yorumları daha iyi ayırt edebilecek bir vektör uzayı oluşturmuştur.

Bu çalışmada spam tespitine yönelik öznelik seçiminde hem içerik tabanlı hem de bağlamsal özellikler dikkate alınmıştır. İçerik temelli öznelikler arasında TF-IDF ile elde edilen kelime ağırlıkları yer alırken, bağlamsal öznelikler olarak karakter sayısı ile ifade edilen yorum uzunluğu, gece, gündüz yada öğlen şeklinde ifade edilen yorumun gönderildiği saat dilimi, büyük harf oranı, noktalama işareti kullanımı ve link içerip içermeme durumu gibi göstergeler seçilmiştir. Bu öznelikler, literatürde spam içeriklerin genellikle kısa, tekrar eden ve belirli zaman dilimlerinde yoğunlaştığı gözlemleriyle örtüşmektedir. Örneğin, spam yorumlar

çoğunlukla gece saatlerinde gönderilmekte ve link içermektedir; bu nedenle zaman ve bağlantı bilgisi ayrıştırıcı rol oynamaktadır. Modelleme sürecinde, özneliklerin tek tek ve farklı kombinasyonlarıyla test edilmesi sonucunda, içerik ve bağlam özelliklerinin birlikte kullanıldığı yapıların en yüksek başarıyı sağladığı gözlemlenmiştir. Özellikle TF-IDF + yorum uzunluğu + saat dilimi birleşimi, hem kesinlik hem duyarlılık açısından en dengeli sonucu üretmiştir.

Son olarak, veri setinin %80'i eğitim ve %20'si test için ayrılmıştır. Çalışmada uygulanan modellerin en başarılı olabilecekleri hiper-parametreleri belirlemek amacıyla Grid Search kullanılmıştır. Grid Search, makine öğrenmesi ve derin öğrenme modelleri için en iyi hiper-parametreleri belirlemek amacıyla kullanılan sistematik bir arama yöntemidir. Grid Search, Belirlenen hiperparametrelerin farklı değerlerini içeren bir hiper-parametre uzayı oluşturur. Tüm olası hiper-parametre kombinasyonlarını tek tek dener ve her birini model üzerinde eğiterek değerlendirir ve en iyi sonucu veren hiper-parametre kombinasyonunu seçer.

Bu çalışmada kullanılan modellerin mimarileri ve optimizasyon süreçleri detaylı olarak yapılandırılmıştır. LR için belirlenen hiper-parametreler penalty: l2, C: 1, solver: liblinear ve max\_iter: 200'dür. RF için belirlenen hiper-parametreler n\_estimators: 200, max\_depth: 20, min\_samples\_split: 5, min\_samples\_leaf: 3, max\_features: sqrt ve bootstrap: True'dur. SVM için belirlenen hiper-parametreler C: 1, kernel: rbf, gamma: scale ve degree: 3'tür. XGBoost için belirlenen hiper-parametreler n\_estimators: 300, max\_depth: 6, learning\_rate: 0.1, subsample: 0.8, colsample\_bytree: 0.8, gamma: 0.1 ve lambda: 1'dir. Bi-LSTM için belirlenen hiper-parametreler embedding\_dim: 128, lstm\_units: 256, dropout: 0.5, batch\_size: 32, epochs: 10, optimizer: adam ve loss: binary\_crossentropy'dir.

BERT, 12 Transformer katmanı (layer), 768 gizli boyut (hidden size), 12 çoklu-başlıklı dikkat (multi-head attention) mekanizması ve toplamda yaklaşık 110 milyon öğrenilebilir parametre içeren bert-base-uncased sürümüyle uygulanmıştır. Bu mimari, her bir kelimeyi çift yönlü bağlamda temsil ederek, metin içerisindeki semantik ilişkileri derinlemesine analiz etme yeteneğine sahiptir. Modelin son katmanına bir sınıflandırıcı (fully connected dense layer) eklenmiş ve AdamW optimizasyon algoritması kullanılmıştır. Aşırı öğrenmeyi önlemek amacıyla weight decay=0.01 ve dropout=0.1 oranları uygulanmıştır. Öğrenme oranı (learning\_rate=2e-5), eğitim stabilitesi sağlamak üzere 500 adımlık bir warmup süreciyle ayarlanmıştır.

Tüm modeller için hiperparametre optimizasyonu, sistematik bir arama stratejisi olan Grid Search yöntemiyle gerçekleştirilmiştir. Bu süreçte, her model için farklı kombinasyonlar denenmiş ve doğruluk ile F1-puanı gibi metrikler üzerinden en iyi sonuçları veren yapılandırmalar tercih edilmiştir. Örneğin, RF için n\_estimators=200, max\_depth=20; XGBoost için n\_estimators=300, learning\_rate=0.1 gibi değerler optimal performansa ulaşmada etkili olmuştur. Hiperparametrelerin bu şekilde seçilmesi, her modelin kendi yapısal avantajlarını en iyi şekilde ortaya koymasına imkân tanımıştır. Ayrıca, kullanılan modellerin avantaj ve sınırlılıkları da dikkate alınmıştır. LR hızlı ve yorumlanabilirken doğrusal sınırlara bağımlıdır. SVM yüksek doğruluk sağlar ancak büyük veri kümelerinde eğitim süresi uzundur. Bi-LSTM uzun vadeli bağımlılıkları öğrenmede etkilidir ancak yüksek hesaplama maliyeti gerektirir. BERT ise bağlamsal anlamı güçlü biçimde modelleyerek özellikle karmaşık metinler için üstün başarı sunar ancak GPU gibi güçlü donanım ihtiyacı doğurur.

### 3.2. Sınıflandırma Modelleri

Bu çalışmada, YouTube spam yorumlarını tespit etmek amacıyla LR, RF, SVM, XGBoost ve BiLSTM gibi yapay zekâ modellerinin sınıflandırma performansları kapsamlı bir şekilde karşılaştırılmıştır.

İstatistiksel bir yöntem olan LR, doğrusal ayrılabilen veri setlerinde iyi performans gösteren bir doğrusal sınıflandırma algoritmasıdır [20]. LR, sigmoid fonksiyonu kullanarak giriş özelliklerini 0-1 aralığında bir olasılığa dönüştürür. LR, verinin iki sınıfa ayrılmasını sağlayan bir karar sınırı öğrenir [21]. Karar sınırı, doğrusal bir fonksiyon ile belirlenir. LR, öğrenme sürecinde, negatif log-likelihood kaybını minimize eden bir optimizasyon işlemi gerçekleştirir. LR, genellikle küçük ve orta ölçekli veri setlerinde iyi çalışır ve hesaplama açısından verimlidir [22]. Ancak, doğrusal olmayan sınırlar içeren veri setlerinde yalnızca doğrusal bir karar sınırı oluşturduğu için yeteri kadar başarılı değildir. Bu nedenle, daha karmaşık ve doğrusal olmayan örüntüleri yakalayabilen SVM, karar ağaçları veya sinir ağları gibi modeller genellikle daha iyi performans gösterir. L1 (Lasso) ve L2 (Ridge) regularizasyonu ile aşırı öğrenme (overfitting) önlenir [23].

RF, birden fazla karar ağacının birleşiminden oluşan güçlü bir ensemble öğrenme yöntemidir. Karar ağaçlarının bağımsız olarak eğitilmesi ve sonucun çoğunluk oylaması ile belirlenmesi, modelin aşırı öğrenme riskini önemli ölçüde azaltır [24]. Torbalama yöntemi kullanılarak her ağaç farklı rastgele alt kümeler üzerinde eğitilir. Bu sayede modele çeşitlilik kazandırılır ve tek bir karar ağacına kıyasla daha genelleştirilebilir hale gelir [25]. RF, özellikle çok boyutlu ve yüksek öznitelikli veri setlerinde etkili çalışır. Öznitelik seçimi ve önemi konusunda güçlüdür. RF, gürültülü verilere karşı dayanıklıdır ancak büyük veri setlerinde eğitim süresi uzayabilir ve fazla bellek tüketebilir [26].

SVM, en iyi ayırım yapan hiper-düzlemi belirleyerek sınıflandırma yapan güçlü bir makine öğrenmesi algoritmasıdır [27]. SVM, maksimum marj prensibini kullanarak sınıflar arasındaki en büyük mesafeyi bulmaya çalışır [28]. Destek vektörleri, karar sınırına en yakın olan veri noktalarıdır ve sınıflandırma sürecinde önemli bir rol oynar. Eğer veri doğrusal olarak ayrılabilir değilse, SVM kernel trick kullanarak veriyi daha yüksek boyutlu bir uzaya taşıyabilir ve burada ayrılabilir hale getirebilir. Yaygın kullanılan çekirdek fonksiyonları Lineer, Polinom, Radial Basis Function (RBF) ve Sigmoid'dir [29]. SVM, özellikle küçük ve orta ölçekli veri setlerinde iyi çalışır ve çok fazla öznitelik içeren veri setlerinde oldukça başarılıdır. Ancak, büyük veri setlerinde eğitim süresi uzun olabilir ve çekirdek fonksiyonu seçimi modelin başarısını önemli ölçüde etkileyebilir [30].

XGBoost, karar ağaçlarını ardışık olarak eğiten, yüksek doğruluk sağlayan güçlü bir ensemble öğrenme modelidir [31]. XGBoost, Gradient Boosting algoritmasını optimize ederek hesaplama hızını artırır ve bellek kullanımını azaltır. Hata oranını minimize eden ardışık ağaç öğrenme süreci, XGBoost'u diğer klasik karar ağaçları yöntemlerinden daha güçlü hale getirir [32]. Her yeni ağaç, bir önceki ağacın hatalarını düzelterek şekilde eğitilir. Bu süreç, modelin karmaşık ilişkileri ve doğrusal olmayan desenleri öğrenmesini sağlar [33].

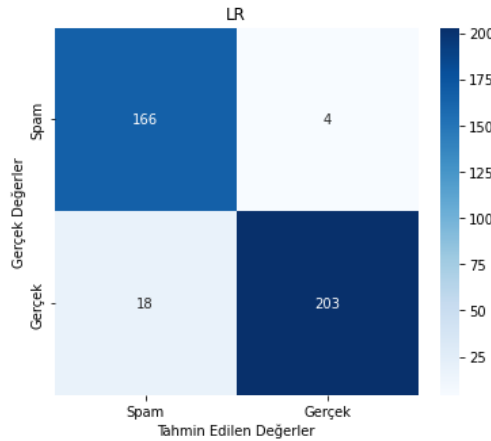
Bi-LSTM, kelime bağlamını hem ileri hem de geri yönde işleyerek klasik LSTM'den daha fazla bilgi yakalayabilen bir sinir ağı modelidir [34]. Geleneksel LSTM, uzun vadeli bağımlılıkları öğrenerek metin tabanlı verilerde başarılı sonuçlar veren bir modeldir [35]. Ancak, LSTM yalnızca geçmişten geleceğe doğru veri akışı sağlar, bu da gelecekteki kelimelerden gelen bağlamsal bilgiyi kayırabilir. Bi-LSTM, girdileri ileri ve geri olmak üzere iki yönde işler. Bu

sayede, bir kelimenin anlamı hem önceki hem de sonraki kelimelerle birlikte değerlendirilir. Spam tespitinde, daha uzun bağlamları anlamlandırabilme yeteneği sayesinde doğruluk oranlarını artırabilir [36]. Ancak, Bi-LSTM eğitmek için büyük miktarda veri gereklidir ve hesaplama maliyeti yüksektir. GPU kullanımı olmadan, eğitim süresi oldukça uzun olabilir. Spam tespiti, duygu analizi ve metin tabanlı sıralı veri işlemede en güçlü modellerden biridir [37].

BERT, Dönüştürücü mimarisi sayesinde geleneksel tek yönlü modellerin aksine metin içerisindeki kelimelerin hem solundaki hem de sağındaki bağlamı aynı anda analiz edebilme özelliğine sahiptir [38]. Bu çift yönlü öğrenme yaklaşımı, özellikle bağlamsal anlam farklılıklarının önemli olduğu metin sınıflandırma görevlerinde büyük avantaj sağlamaktadır [39]. BERT, büyük çaplı dil verileri üzerinde önceden eğitilmiş olup, transfer öğrenme ile daha küçük veri kümeleri üzerinde kolayca yeniden uyarlanabilir. Bu yönüyle, veri miktarının sınırlı olduğu durumlarda bile yüksek başarı elde edilebilir [40]. Spam tespiti gibi anlam karmaşıklığı ve dilsel çeşitlilik barındıran görevlerde, BERT'in semantik açıdan zengin temsil gücü oldukça etkilidir. Özellikle ironi, mecaz ve dolaylı ifadelerin bulunduğu yorumların sınıflandırılmasında, kelime seviyesinde değil bağlam seviyesinde analiz yapabilmesi sayesinde klasik yöntemlerin ötesine geçmektedir.

#### 4. DENEYSEL SONUÇLAR

Bu çalışmada, LR, RF, SVM, XGBoost ve Bi-LSTM modellerinin spam tespitindeki performansları test edilmiştir. Modellerin başarısı, doğruluk, kesinlik, duyarlılık ve F1-puanı metrikleri kullanılarak değerlendirilmiştir. Şekil 5 ve Tablo 2'de LR için karışıklık matrisi ve deneysel sonuçlar görülmektedir.



Şekil 5. LR için karışıklık matrisi

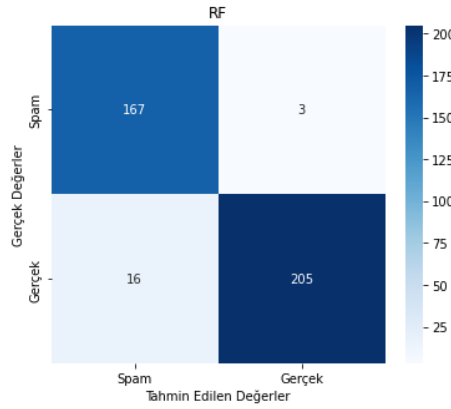
Şekil 5'te görüldüğü gibi LR, 170 adet spam yorumdan 166'sını ve 221 spam olmayan yorumdan 203'ünü doğru bir şekilde sınıflandırmıştır. LR, 391 yorumun 369'unu doğru, 22'sini ise yanlış sınıflandırmıştır.

Tablo 2. LR için deneysel sonuçlar

Sınıf	Doğruluk	Kesinlik	Duyarlılık	F1-puanı
Gerçek	%94,4	%98,1	%91,9	%94,9
Spam		%90,2	%97,6	%93,8
Makro ortalama		%94,1	%94,8	%94,4
Ağırlıklı ortalama		%94,2	%94,4	%94,3

Tablo 2’de görüldüğü gibi LR’nin sınıflandırma doğruluğu %94,4’tür. Spam sınıfı için %97,6 olan duyarlılık değeri, LR’nin spam olan yorumları büyük oranda doğru tespit ettiğini göstermektedir. Ancak, spam sınıfı için %90,2 olan kesinlik değeri, LR’nin bazı gerçek yorumları spam olarak sınıflandırdığını göstermektedir.

Şekil 6 ve Tablo 3’te RF için karışıklık matrisi ve deneysel sonuçlar görülmektedir.



Şekil 6. RF için karışıklık matrisi

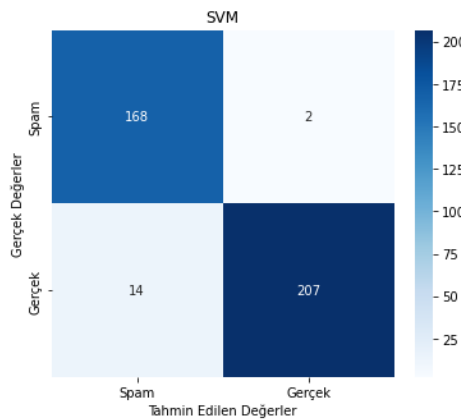
Şekil 6’da görüldüğü gibi RF, 170 adet spam yorumdan 167’sini ve 221 spam olmayan yorumdan 205’ini doğru bir şekilde sınıflandırmıştır. RF, 391 yorumun 372’sini doğru, 19’unu ise yanlış sınıflandırmıştır.

Tablo 3. RF için deneysel sonuçlar

Sınıf	Doğruluk	Kesinlik	Duyarlılık	F1-puanı
Gerçek	%95,3	%98,5	%92,8	%95,6
Spam		%91,2	%98,2	%94,6
Makro ortalama		%94,8	%95,5	%95,1
Ağırlıklı ortalama		%95,0	%95,3	%95,2

Tablo 3’te görüldüğü gibi RF’nin sınıflandırma doğruluğu %95,3’tür. Spam sınıfı için %98,2 duyarlılık değeri, RF’nin spam olan yorumların büyük çoğunluğunu doğru tespit ettiğini göstermektedir. Spam olmayan yorumlar için %92,8 kesinlik oranı LR’ye göre yüksektir ve RF’ın yanlış pozitifleri azalttığını göstermektedir.

Şekil 7 ve Tablo 4’te SVM için karışıklık matrisi ve deneysel sonuçlar görülmektedir.



Şekil 7. SVM için karışıklık matrisi

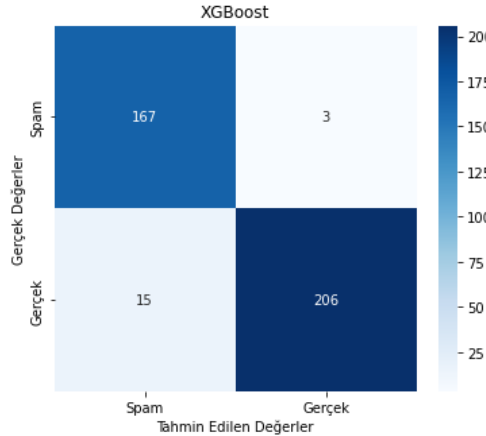
Şekil 7’de görüldüğü gibi SVM, 170 adet spam yorumdan 168’ini ve 221 spam olmayan yorumdan 207’ini doğru bir şekilde sınıflandırmıştır. SVM, 391 yorumun 375’ini doğru, 16’sını ise yanlış sınıflandırmıştır.

**Tablo 4.** SVM için deneysel sonuçlar

Sınıf	Doğruluk	Kesinlik	Duyarlılık	F1-puanı
Gerçek	%96,2	%99,0	%93,7	%96,3
Spam		%92,3	%98,8	%95,5
Makro ortalama		%95,6	%96,3	%95,9
Ağırlıklı ortalama		%95,8	%96,2	%96,0

Tablo 4’te görüldüğü gibi SVM’nin sınıflandırma doğruluğu %96,2’dir. Spam sınıfı için %98,8 duyarlılık değeri, SVM’nin spam olan yorumların büyük çoğunluğunu doğru tespit ettiğini göstermektedir. Spam olmayan yorumlar için %99,0 kesinlik değeri, SVM’nin yanlış pozitif oranının son derece düşük olduğunu göstermektedir. Makro ve ağırlıklı ortalama değerleri de %96 seviyelerindedir ve SVM tüm sınıflarda oldukça dengeli bir performans göstermiştir.

Şekil 8 ve Tablo 5’te XGBoost için karışıklık matrisi ve deneysel sonuçlar görülmektedir.



**Şekil 8.** XGBoost için karışıklık matrisi

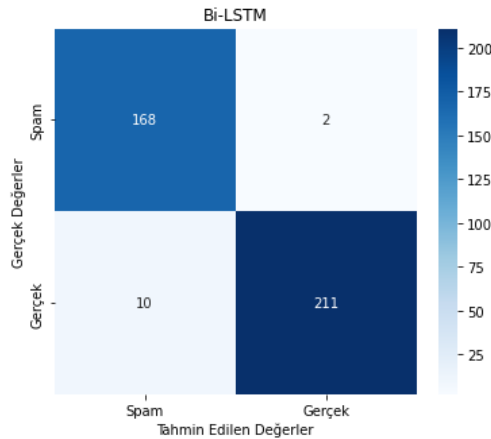
Şekil 8’de görüldüğü gibi XGBoost, 170 adet spam yorumdan 167’sini ve 221 spam olmayan yorumdan 206’sını doğru bir şekilde sınıflandırmıştır. XGBoost, 391 yorumun 373’ünü doğru, 18’ini ise yanlış sınıflandırmıştır.

**Tablo 5.** XGBoost için deneysel sonuçlar

Sınıf	Doğruluk	Kesinlik	Duyarlılık	F1-puanı
Gerçek	%95,6	%98,6	%93,2	%95,8
Spam		%91,7	%98,2	%94,8
Makro ortalama		%95,1	%95,7	%95,3
Ağırlıklı ortalama		%95,2	%95,6	%95,4

Tablo 5’te görüldüğü gibi XGBoost’un sınıflandırma doğruluğu %95,6’dır. Spam sınıfı için %98,2 olan duyarlılık değeri, XGBoost’un spam yorumları büyük ölçüde doğru bir şekilde tespit ettiğini göstermektedir. Spam olmayan yorumlar için %98,6 kesinlik değeri, yanlış pozitif oranının oldukça düşük olduğunu göstermektedir. Makro ve ağırlıklı ortalama metrikleri %95 seviyelerindedir ve XGBoost’un dengeli bir şekilde çalıştığını göstermektedir. XGBoost, SVM ile benzer sonuçlara sahiptir ancak XGBoost daha dengeli bir model sunmaktadır.

Şekil 9 ve Tablo 6’da Bi-LSTM için karışıklık matrisi ve deneysel sonuçlar görülmektedir.



Şekil 9. Bi-LSTM için karışıklık matrisi

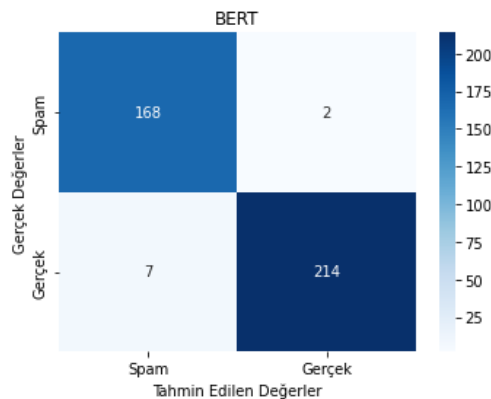
Şekil 9’da görüldüğü gibi Bi-LSTM, 170 adet spam yorumdan 168’ini ve 221 spam olmayan yorumdan 211’ini doğru bir şekilde sınıflandırmıştır. Bi-LSTM, 391 yorumun 379’unu doğru, 12’sini ise yanlış sınıflandırmıştır.

Tablo 6. Bi-LSTM için deneysel sonuçlar

Sınıf	Doğruluk	Kesinlik	Duyarlılık	F1-puanı
Gerçek	%97,1	%99,1	%95,5	%97,3
Spam		%94,4	%98,8	%96,6
Makro ortalama		%96,8	%97,2	%96,9
Ağırlıklı ortalama		%97,0	%97,1	%97,0

Tablo 6’da görüldüğü gibi Bi-LSTM’in sınıflandırma doğruluğu %97,1’dir. Bi-LSTM, karşılaştırılan modellerden daha yüksek doğruluk oranına ulaşarak en başarılı model olmuştur. Spam sınıfı için %98,8 duyarlılık değeri, Bi-LSTM’in spam olan yorumları neredeyse tamamen doğru tespit ettiğini göstermektedir. Spam olmayan yorumlar için %99,1 kesinlik değeri, Bi-LSTM’in hatalı bir şekilde spam olarak tespit edilen yorumları en aza indirdiğini göstermektedir. %97 seviyelerinde olan makro ve ağırlıklı ortalama metrikleri Bi-LSTM’in tüm sınıflarda daha dengeli bir performans sunduğunu göstermektedir.

Şekil 10 ve Tablo 7’de BERT için karışıklık matrisi ve deneysel sonuçlar görülmektedir.



Şekil 10. BERT için karışıklık matrisi

Şekil 10’da görüldüğü gibi BERT, 170 adet spam yorumdan 168’ini ve 221 spam olmayan yorumdan 214’ünü doğru bir şekilde sınıflandırmıştır. BERT, 391 yorumun 382’sini doğru, 9’unu ise yanlış sınıflandırmıştır.

**Tablo 7.** BERT için deneysel sonuçlar

Sınıf	Doğruluk	Kesinlik	Duyarlılık	F1-puanı
Gerçek	%97,7	%99,1	%96,8	%97,9
Spam		%96,0	%98,8	%97,4
Makro ortalama		%97,5	%97,8	%97,7
Ağırlıklı ortalama		%97,7	%97,7	%97,7

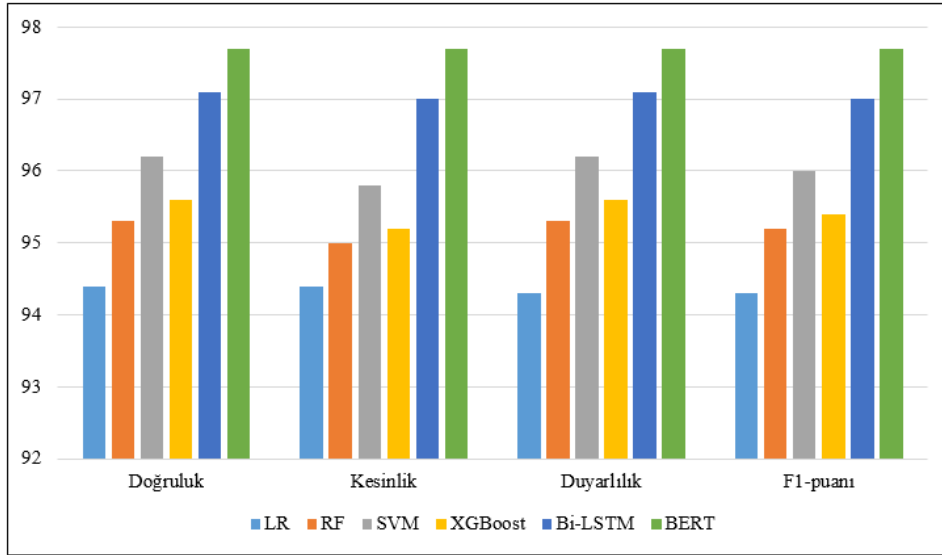
Tablo 7’de görüldüğü gibi BERT’in sınıflandırma doğruluğu %97,7’dir. BERT, hem spam tespiti hem de gerçek mesaj ayırt etme konusunda yüksek başarı göstermektedir. En güçlü yönü, %98,8 duyarlılık değeriyle spam mesajları neredeyse eksiksiz tespit edebilmesidir. Sınıf bazında incelendiğinde, Gerçek sınıfı için %99,1 kesinlik ve %96,8 duyarlılık, modelin yanlış pozitifleri minimum düzeyde tuttuğu ve bu sınıfı büyük oranda doğru tanıdığı görülmektedir. Spam sınıfı için %98,8 duyarlılık değeri, modelin spam içerikleri başarıyla ayırt edebildiğini göstermektedir. %96,0 kesinlik değeri ise bu tahminlerin çoğunun doğru olduğunu göstermektedir. Makro ve ağırlıklı ortalama metriklerinin birbirine çok yakın olması, modelin her iki sınıfa da dengeli şekilde öğrenme sağladığını ve veri setinde sınıf dengesizliğinden kaynaklı bir sapmanın oluşmadığını göstermektedir. F1-puanlarının her iki sınıf için de %97’nin üzerinde olması, kesinlik ve duyarlılık metrikleri arasında güçlü bir denge kurulduğunu göstermektedir.

Tablo 8 ve Şekil 11’de ağırlıklı ortalamaya göre karşılaştırmalı deneysel sonuçlar görülmektedir.

**Tablo 8.** Karşılaştırmalı deneysel sonuçlar

Model	Doğruluk	Kesinlik	Duyarlılık	F1-puanı
LR	%94,4	%94,4	%94,3	%94,3
RF	%95,3	%95,0	%95,3	%95,2
SVM	%96,2	%95,8	%96,2	%96,0
XGBoost	%95,6	%95,2	%95,6	%95,4
Bi-LSTM	%97,1	%97,0	%97,1	%97,0
BERT	%97,7	%97,7	%97,7	%97,7

Tablo 8 ve Şekil 11’de görüldüğü gibi BERT, tüm değerlendirme metrikleri açısından en başarılı performansa ulaşmıştır. BERT’in, sahip olduğu %97,7 doğruluk ile hem spam hem de gerçek yorumları ayırt etmede etkili olduğu görülmektedir.



Şekil 11. Karşılaştırmalı deneysel sonuçlar

LR, %94,4 doğruluk ile en düşük başarıyı gösteren model olmuştur. LR, doğrusal bir model olduğu için ve karmaşık yapıya sahip metin verilerinde yeterince iyi genelleme yapamamaktadır. RF, %95,3 doğrulukla LR'den daha iyi performans göstermiştir. RF, çok sayıda karar ağacı kullanarak karmaşık karar sınırlarını öğrenebildiği için metin sınıflandırmada LR'den daha başarılı olmuştur. SVM, %96,2 doğrulukla RF ve LR'ye göre daha başarılı olmuştur. Özellikle %96,2 duyarlılık değeriyle spam tespitinde oldukça başarılıdır. XGBoost, %95,6 doğruluk ile RF'den daha iyi ancak SVM'den daha düşük bir başarı göstermiştir. XGBoost, boosting mekanizması sayesinde RF'a kıyasla daha hassas karar sınırları belirleyebilse de, SVM ve Bi-LSTM kadar yüksek doğruluk sağlayamamıştır. En yüksek başarı, %97,1 doğruluk ile Bi-LSTM tarafından elde edilmiştir. Bi-LSTM, çift yönlü uzun kısa süreli bellek mekanizması sayesinde metinlerdeki kelime ilişkilerini daha iyi öğrenerek spam ve gerçek yorumları yüksek doğrulukla sınıflandırmıştır. En yüksek doğruluk oranı ise %97,7 ile BERT modeline aittir. BERT, önceden eğitilmiş bağlamsal dil temsil yeteneği sayesinde metinlerin anlam derinliğini öğrenmiş ve hem spam hem de gerçek yorumları yüksek hassasiyetle sınıflandırarak en başarılı model olmuştur.

Deneysel sonuçlar, basit yöntemlerden karmaşık yöntemlere doğru gidildikçe spam tespit performansının arttığını göstermektedir. LR gibi basit modeller, hızlı ve yorumlanabilir olmalarına rağmen karmaşık dil örüntülerini öğrenmede sınırlı kalmaktadır. RF, SVM ve XGBoost gibi ileri seviye klasik yöntemler, daha güçlü karar sınırları öğrenebilmiş ve doğruluk oranlarını artırmıştır. En yüksek başarı, karmaşık dil bağlamlarını anlayabilen Bi-LSTM ve BERT modeline ait olup, özellikle bağlamsal ve yapısal olarak karmaşık spam yorumlarını ayırt etmede etkili olmuştur.

## 5. SONUÇLAR

Günümüzde dijital platformlarda spam içeriklerin yaygınlaşması, kullanıcı deneyimini olumsuz etkilemekte ve yanıltıcı bilgilere maruz kalma riskini artırmaktadır. Özellikle YouTube gibi büyük kullanıcı kitlesine sahip sosyal medya platformlarında, spam yorumların manuel olarak tespit edilmesi mümkün olmadığından, otomatik spam tespit sistemlerine duyulan ihtiyaç giderek artmaktadır. Bu çalışmada, YouTube yorumlarındaki spam içerikleri tespit etmek amacıyla farklı makine öğrenmesi ve derin öğrenme yöntemlerinin karşılaştırmalı bir analizi sunulmuştur. LR, RF, SVM, XGBoost, Bi-LSTM ve BERT modellerinin performansları

doğruluk, kesinlik, duyarlılık ve F1-puanı metrikleri kullanılarak değerlendirilmiştir. Deneysel çalışmalar, BERT modelinin %97,7 doğruluk oranı ile karşılaştırılan tüm modeller arasında en yüksek başarıyı sağladığını ortaya koymuştur. BERT'in, önceden büyük ölçekli metin verileri üzerinde eğitilmiş derin dil temsilleri sayesinde, bağlamsal ilişkileri güçlü bir şekilde modelleyebildiği ve bu nedenle hem spam hem de gerçek yorumları yüksek isabetle sınıflandırabildiği gözlemlenmiştir. Bi-LSTM modeli ise, kelimeler arasındaki ilişkileri çift yönlü olarak analiz edebilme yeteneği sayesinde spam tespitinde dikkat çeken bir başarı göstermiştir. SVM ve XGBoost modelleri yüksek doğruluk oranları elde etmiş olsalar da, metin tabanlı verilerdeki uzun vadeli bağımlılıkları öğrenme konusunda Bi-LSTM ve özellikle BERT kadar etkili olamamıştır.

Geleneksel kural tabanlı filtreleme sistemlerinin aksine yapay zekâ tabanlı modeller, spam içerikleri bağlamsal ve anlamsal özellikleri göz önünde bulundurarak daha doğru bir şekilde sınıflandırabilmektedir. Bu çalışmada kullanılan TF-IDF yöntemi ile metinlerin vektörleştirilmesi, makine öğrenmesi ve derin öğrenme modellerinin spam tespiti için uygun bir şekilde eğitilmesini sağlamış ve özellikle XGBoost ve SVM gibi modellerin performansını artırmıştır. Ancak, BERT ve Bi-LSTM dilin daha derin bağlamını anlayarak daha güçlü bir sınıflandırma performansına sahip olmuştur.

Deneysel sonuçlar, özellikle BERT ve Bi-LSTM modellerinin Türkçe spam tespitinde yüksek başarı oranlarıyla öne çıktığını göstermektedir. Bu sonuçlar, ALBERT+BLSTM gibi karma modellerle %95 üzeri doğruluk elde edilen güncel çalışmalarla benzerlik göstermekte, ancak çalışmada yalnızca yorum metni değil, saat dilimi ve yorum uzunluğu gibi bağlamsal özelliklerin de modele dâhil edilmesi sayesinde daha güçlü sınıflandırma performansları elde edilmiştir. Ayrıca, önceki birçok çalışmada yalnızca İngilizce veya sınırlı boyuttaki veri setleri kullanılırken, bu çalışmada Türkçe içerikli YouTube yorumlarından oluşturulan özgün ve dengeli bir veri kümesi kullanılmıştır. Bununla birlikte, model eğitimi sırasında yüksek doğruluk elde edilse de, gerçek zamanlı uygulamalara yönelik testlerin yapılmaması çalışmanın bir sınırlılığı olarak değerlendirilebilir. Ayrıca, yorumlardaki ironi, kinaye gibi dilsel inceliklerin modellenememesi ve veri setinin yalnızca YouTube'a özgü olması genel geçerlilik açısından bir başka sınırlılıktır. Buna karşın, çok sayıda makine öğrenmesi ve derin öğrenme algoritmasının sistematik biçimde karşılaştırılması ve Türkçe metinler üzerinde uygulanması çalışmanın güçlü yönleri arasında yer almaktadır.

Elde edilen bulgular, sosyal medya platformlarında kullanılan mevcut spam filtreleme sistemlerinin, yapay zekâ temelli yöntemlerle önemli ölçüde geliştirilebileceğini göstermektedir. Özellikle BERT ve Bi-LSTM gibi bağlamsal öğrenme yeteneğine sahip modeller, sadece anahtar kelimelere değil, yorumun anlam bütünlüğüne ve bağlamına dayalı olarak spam tespiti gerçekleştirebildiğinden, daha az yanlış pozitif ve negatif sonuç üretmektedir. Bu da kullanıcıların yanlışlıkla engellenmesini ya da spam içeriklerin gözden kaçmasını önlemeye katkı sağlar. Geliştirilen modeller, YouTube gibi büyük ölçekli platformlara entegre edilerek, yorum denetleme süreçlerinin otomatikleştirilmesi ve gerçek zamanlı spam filtreleme altyapılarının oluşturulmasına olanak tanıyabilir. Bu sayede kullanıcı deneyiminin iyileştirilmesinin yanı sıra platformların bilgi güvenliğinin güçlendirilmesi sağlanabilir.

Gelecek çalışmalarda, spam tespit sistemlerinin performansını artırmak amacıyla RoBERTa, XLNet ve DeBERTa gibi farklı dil modelleri ve çok dilli veri setleri üzerinde incelemeler yapılabilir. Ayrıca, yorumların semantik yapısını daha derinlemesine analiz edebilecek attention tabanlı hibrit modeller geliştirilebilir. Spam tespitine yönelik modellerin

açıklanabilirliğini artırmak adına SHAP veya LIME gibi Açıklanabilir Yapay Zekâ (Explainable AI-XAI) yöntemlerinin entegre edilmesi de önemli katkılar sağlayacaktır. Gerçek zamanlı spam algılama senaryoları veya video içeriği ile yorumların ilişkilendirilmesi gibi bağlamsal analizler ise uygulama odaklı gelecek çalışmalara zemin hazırlayabilir.

## KAYNAKLAR

- [1] Susanto H, Fang Yie L, Mohiddin F, Rahman Setiawan A A, Haghi P K, Setiana D. Revealing social media phenomenon in time of COVID-19 pandemic for boosting start-up businesses through digital ecosystem. *Applied system innovation*. 2021;4(1).
- [2] Humprecht E, Kessler S H. Unveiling misinformation on YouTube: examining the content of COVID-19 vaccination misinformation videos in Switzerland. *Frontiers in Communication*. 2024; 9.
- [3] Lakshmi M S, Rani A S, Divya T S, Shravani J. Dynamic Spam Detection in Social Networks: Leveraging Convex Nonnegative Matrix Factorization for Enhanced Accuracy and Scalability. *International Journal of Computer Engineering in Research Trends*. 2024; 11(4), 1-11.
- [4] Gongane V U, Munot M V, Anuse A D. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*. 2022; 12(1).
- [5] Wani M A, ElAffendi M, Shakil K A. AI-Generated Spam Review Detection Framework with Deep Learning Algorithms and Natural Language Processing. *Computers*. 2024; 13(10).
- [6] Ahmed N, Amin R, Aldabbas H, Koundal D, Alouffi B, Shah T. Machine learning techniques for spam detection in email and IoT platforms: analysis and research challenges. *Security and Communication Networks*. 2022; 2022(1).
- [7] Akinyelu A A. Advances in spam detection for email spam, web spam, social network spam, and review spam: ML-based and nature-inspired-based techniques. *Journal of Computer Security*. 2021; 29(5), 473-529.
- [8] Al Saidat M R, Yerima S Y, Shaalan K. Advancements of SMS Spam Detection: A Comprehensive Survey of NLP and ML Techniques. *Procedia Computer Science*. 2024; 244, 248-259.
- [9] Al-Adhaileh M H, Alsaade F W. Detecting and Analysing Fake Opinions Using Artificial Intelligence Algorithms. *Intelligent Automation & Soft Computing*. 2022; 32(1).
- [10] Shinde S A, Pawar R R, Jagtap A A, Tambewagh P A, Rajput P U, Mali M K, Mulik S V. Deceptive opinion spam detection using bidirectional long short-term memory with capsule neural network. *Multimedia Tools and Applications*. 2024; 83(15), 45111-45140.
- [11] Sinhal A, Maheshwari M. An Extensive Review on Contemporary Analysis of Comment Filtration of YouTube Videos Using Machine Learning Techniques. *International Journal of Emerging Technology and Advanced Engineering*. 2022; 12(9), 130-143.

- [12] Shirzadova S, Uysal A K. Türkçe YouTube Yorumları Üzerinde Spam Filtreleme. Düzce Üniversitesi Bilim ve Teknoloji Dergisi. 2022; 10(4), 1793-1810.
- [13] Baktır N, Atay Y. Makine Öğrenmesi Yaklaşımlarının Spam-Mail Sınıflandırma Probleminde Karşılaştırmalı Analizi. Bilişim Teknolojileri Dergisi. 2022; 15(3), 349-364.
- [14] Bakır R, Erbay H, Bakır H. ALBERT4Spam: a novel approach for spam detection on social networks. Bilişim Teknolojileri Dergisi. 2024; 17(2), 81-94.
- [15] Güven Z A. Türkçe e-postalarda spam tespiti için makine öğrenme yöntemlerinin ve dil modellerinin analizi. Avrupa Bilim ve Teknoloji Dergisi 2023; 47, 1-6.
- [16] Şengel Ö. A comparative analysis of learning techniques in the context of Turkish spam detection. Batman Üniversitesi Yaşam Bilimleri Dergisi. 2024; 14(1), 43-56.
- [17] Sam'an M, Imaddudin K. Hybrid deep learning model for YouTube spam comment detection. International Journal of Electrical and Computer Engineering (IJECE). 2024; 14(3), 3313-3319.
- [18] Airlangga G. Spam Detection in YouTube Comments Using Deep Learning Models: A Comparative Study of MLP, CNN, LSTM, BiLSTM, GRU, and Attention Mechanisms. MALCOM: Indonesian Journal of Machine Learning and Computer Science. 2024; 4(4), 1533-1538.
- [19] Waheed A. YouTube Yorumları Spam Veri Kümesi [İnternet]. Kaggle; [alıntılanma tarihi 6 Mart 2025]. Erişim adresi: <https://www.kaggle.com/datasets/ahsenwaheed/youtube-comments-spam-dataset/data>
- [20] Bektaş J. EKSL: An effective novel dynamic ensemble model for unbalanced datasets based on LR and SVM hyperplane-distances. Information Sciences. 2022; 597, 182-192.
- [21] Al-Najjar H A, Pradhan B, Kalantar B, Sameen M I, Santosh M, Alamri A. Landslide susceptibility modeling: an integrated novel method based on machine learning feature transformation. Remote Sensing. 2021; 13(16).
- [22] Kim G, Yang S M, Kim D M, Choi J G, Lim S, Park H W. Developing a deep learning-based uncertainty-aware tool wear prediction method using smartphone sensors for the turning process of Ti-6Al-4V. Journal of Manufacturing Systems. 2024; 76, 133-157.
- [23] Nwosu A, Aimufua G I O, Ajayi B A, Olalere M. The Impact of Regularization on Linear Regression Based Model. Journal of Artificial Intelligence and Computer Science. 2024; 1(1).
- [24] Arabameri A, Chandra Pal S, Rezaie F, Chakraborty R, Saha A, Blaschke T, Thi Ngo P T. Decision tree based ensemble machine learning approaches for landslide susceptibility mapping. Geocarto International. 2022; 37(16), 4594-4627.
- [25] Sesa O, Haikal A Y, Elhosseini M A, Gad H H. Smart Bagged Tree-based Classifier optimized by Random Forests (SBT-RF) to Classify Brain-Machine Interface

- Data. International journal of electrical and computer engineering systems. 2022; 13(10), 895-908.
- [26] Jagannath A, Jagannath J, Kumar P S P V. A comprehensive survey on radio frequency (RF) fingerprinting: Traditional approaches, deep learning, and open challenges. *Computer Networks*. 2022; 219.
- [27] Chandra M A, Bedi S S. Survey on SVM and their application in image classification. *International Journal of Information Technology*. 2021; 13(5), 1-11.
- [28] Lai Z, Chen X, Zhang J, Kong H, Wen J. Maximal margin support vector machine for feature representation and classification. *IEEE Transactions on Cybernetics*. 2023; 53(10), 6700-6713.
- [29] Negi H S, Dimri S C, Kumar B, Ram M. Support vector machine and classification, kernel trick for separating of data points. *Mathematics in Engineering, Science & Aerospace (MESA)*. 2024; 15(2).
- [30] Ding X, Liu J, Yang F, Cao J. Random radial basis function kernel-based support vector machine. *Journal of the Franklin Institute*. 2021; 358(18), 10121-10140.
- [31] Natras R, Soja B, Schmidt M. Ensemble machine learning of random forest, AdaBoost and XGBoost for vertical total electron content forecasting. *Remote Sensing*. 2022; 14(15), 3547.
- [32] Demir S, Sahin E K. An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost. *Neural Computing and Applications*. 2023; 35(4), 3173-3190.
- [33] Ji S, Wang X, Lyu T, Liu X, Wang Y, Heinen E, Sun Z. Understanding cycling distance according to the prediction of the XGBoost and the interpretation of SHAP: A non-linear and interaction effect analysis. *Journal of Transport Geography*. 2022; 103.
- [34] Shoubaki H, Abdallah S, Shaalan K. Deep Learning Techniques for Identifying Poets in Arabic Poetry: A Focus on LSTM and Bi-LSTM. *Procedia Computer Science*. 2024; 244, 461-470.
- [35] Zhou Z G. Research on sentiment analysis model of short text based on deep learning. *Scientific Programming*. 2022; 2022(1), 2681533.
- [36] Ahmed S, Saif A S, Hanif M I, Shakil M M N, Jaman M M, Haque M M U, Sabbir H M. Att-BiL-SL: Attention-based Bi-LSTM and sequential LSTM for describing video in the textual formation. *Applied sciences*. 2021; 12(1), 317.
- [37] Odera D, Odiaga G. A comparative analysis of recurrent neural network and support vector machine for binary classification of spam short message service. *World Journal of Advanced Engineering Technology and Sciences*. 2023; 9(1), 127-152.
- [38] Zhou C, Li Q, Li C, Yu J, Liu Y, Wang G, Sun L. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, 2024, 1-65.

- [39] Gao Z, Feng A, Song X, Wu X. Target-dependent sentiment classification with BERT. *Ieee Access*, 2019; 7, 154290-154299.
- [40] Rosso M M, Marasco G, Aiello S, Aloisio A, Chiaia B, Marano G C. Convolutional networks and transformers for intelligent road tunnel investigations. *Computers & Structures*, 2023; 275.