

YILDIZ TEKNİK ÜNİVERSİTESİ

ELEKTRİK-ELEKTRONİK FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

BLM5105 Doğal Dil İşlemeye Kavramsal Bir Bakış

PROJE

Spam Mesaj Tespiti

Muhammed Kayra BULUT – 245B5016

# SPAM MESAJ TESPİTİ NEDİR?

## Spam Mesajların Tanımı ve Ehemmiyeti

Spam mesajlar, kullanıcının izni ve talebi olmaksızın gönderilen, genellikle ticari veya hileli amaçlar taşıyan, kitle halinde dağıtılan istenmeyen elektronik iletilerdir. Bu mesajlar, günümüz iletişim dünyasının mühim bir sorunu haline gelmiştir.

Cep telefonları ve elektronik cihazların yaygınlaşmasıyla birlikte, SMS mesajları kolay erişilebilir bir iletişim kanalı olarak kullanılmaktadır. Lakin bu kolaylık, istenmeyen reklamlar, sahte ikramiye duyuruları, dolandırıcılık teşebbüsleri ve benzeri gayelerle suistimal edilmektedir. Böylece kullanıcıların mesaj kutularında beklenmedik ve gereksiz içerikler birikmeye başlar.

Spam mesajların ehemmiyeti, sadece birer rahatsızlık kaynağı olmalarından değil, aynı zamanda hakiki ve mühim mesajların teşhisini zorlaştırmalarından, cihaz hafızasını işgal etmelerinden ve sıklıkla kişisel malumatların çalınmasına vesile olan tuzaklar içermelerinden ileri gelmektedir. Bilhassa "phishing" olarak adlandırılan ve kullanıcıları sahte bağlantılara yönlendirerek bilgi çalmayı hedefleyen mesajlar, ciddi güvenlik tehlikeleri arz etmektedir.

## Spam Tespitinin Gerekliliği ve Faydaları

Spam mesaj tespiti, modern iletişim araçlarının güvenli ve etkili kullanımı için elzem bir teknolojidir. Her gün milyarlarca mesajın iletildiği bir dünyada, zararlı ve gereksiz içeriğin tefriki büyük önem taşımaktadır.

Spam tespitinin en önemli faydalarından biri, kullanıcıların değerli mesajlarına daha hızlı ve kolay erişimini sağlamasıdır. Ayrıca, kişisel bilgilerin korunmasına yardımcı olarak olası dolandırıcılık vakalarını önler. Cihaz hafızasında yer kaplayan gereksiz mesajların otomatik olarak filtrelenmesi, depolama alanının daha verimli kullanılmasına imkân tanır.

Bunlara ilaveten, gelişmiş spam tespit sistemleri vesilesiyle kullanıcılar her mesajı tek tek incelemek zorunda kalmaz, böylece vakitlerini daha kıymetli işlere hasredebilirler. Telekomünikasyon şebekeleri üzerindeki yükün hafifletilmesi de bu sistemlerin getirdiği müspet neticelerdendir.

Bilhassa makine öğrenmesi ve doğal dil işleme teknolojilerindeki ilerlemeler, spam tespitini daha isabetli hale getirmiştir. Projemizde kullanılan BERT modeli gibi derin öğrenme yaklaşımları, Türkçe metinlerdeki nüansları anlayarak olağanüstü başarı oranları elde etmektedir. Neticede, kullanıcılar daha temiz bir mesaj kutusu ve daha güvenli bir iletişim ortamı tecrübe etmektedir.

# Spam Mesaj Tespiti Nasıl Çalışır?

## Geleneksel Spam Tespiti Yöntemleri

Spam mesaj tespiti, elektronik muhaberat dünyasında uzun zamandır mevcut olan bir mesele olup, başlangıçta nispeten basit yöntemlerle icra edilmiştir. Bu kadim usuller, günümüzde dahi bazı sistemlerde hala istimal edilmektedir.

İlk olarak, kara liste yaklaşımı olarak bilinen yöntem, belirli gönderici adresleri veya IP adreslerini içeren bir sicil oluşturarak işler. Mesaj göndereni bu kara listede yer alıyorsa, mesaj otomatik olarak spam olarak işaretlenir. Bu usul sade olmasına rağmen, spam göndericilerin kolaylıkla adreslerini değiştirebildikleri için sınırlı etkiye sahiptir.

İkinci olarak, kural tabanlı sistemler, mesaj içeriğinde bulunan belirli kelimelerin, ifadelerin veya yapıların varlığını arar. Örneğin, "bedava", "kazandınız", "acil" gibi kelimeler veya aşırı noktalama işaretleri içeren mesajlar şüpheli olarak değerlendirilir. Bu sistemler, kuralların elle oluşturulması ve güncellenmesini gerektirir ki bu da zamanla külfetli bir mesai haline gelir.

Üçüncü olarak, imza tabanlı sistemler, bilinen spam mesajların özelliklerini içeren imzalar oluşturur ve gelen mesajları bu imzalarla karşılaştırır. Eğer bir eşleşme tespit edilirse, mesaj spam olarak sınıflandırılır. Bu yaklaşım, daha evvel görülmüş spam türleri için gayet müessirdir, lakin yeni ve değişen taktiklere karşı zafiyet gösterir.

Bu geleneksel yöntemler, düşük hesaplama gereksinimleri ve basit uygulanabilirlik gibi üstünlüklere sahip olsa da, spam göndericilerin sürekli değişen taktikleri karşısında yetersiz kalabilmektedir. Bu sebeple, daha ayrıntılı çözümlere ihtiyaç hasıl olmuştur.

## Modern Makine Öğrenmesi Tabanlı Yaklaşımlar

Teknolojinin ilerlemesiyle birlikte, spam tespitinde makine öğrenmesi yöntemleri merkezi bir mevki kazanmıştır. Bu yaklaşımlar, elle kural oluşturma ihtiyacını ortadan kaldırarak, sistemlerin spam ile düzgün mesajlar arasındaki farkı daha etkili şekilde öğrenmesini sağlamaktadır.

Naive Bayes sınıflandırıcısı, spam tespitinde kullanılan en yaygın makine öğrenmesi algoritmalarından biridir. Bu yöntem, kelimelerin spam ve normal mesajlardaki görülme olasılıklarına dayalı olarak çalışır ve Bayes teoremini kullanarak bir mesajın spam olma ihtimalini hesaplar. Sadeliği ve düşük hesaplama gereksinimleri sebebiyle hala rağbet görmektedir.

Destek Vektör Makineleri (SVM) ise, metinleri çok boyutlu uzayda temsil ederek, spam ve normal mesajları birbirinden en iyi ayıran karar sınırını (hiper düzlem) bulmaya çalışır. Bu algoritma, özellikle yüksek boyutlu veri kümeleriyle başa çıkmakta mahirdir.

Karar ağaçları ve rassal ormanlar gibi topluluk öğrenme algoritmaları da spam tespitinde kullanılmaktadır. Bu yöntemler, birden fazla basit modelin tahminlerini birleştirerek daha gürbüz ve dakik sonuçlar elde etmeyi hedefler. Özellikle rassal ormanlar, aşırı uyum sorunlarına karşı mukavemetleri ve yorumlanabilir olmaları sebebiyle tercih edilmektedir.

Makine öğrenmesi tabanlı yaklaşımların bir diğer mühim unsuru da öznitelik mühendisliğidir. Mesajlardaki kelime sıklığı, mesaj uzunluğu, büyük harf kullanımı, noktalama işaretleri, URL varlığı gibi çeşitli öznitelikler çıkarılarak modellerin eğitiminde kullanılır. Bu özniteliklerin kalitesi, nihai sınıflandırma performansını doğrudan etkiler.

Modern makine öğrenmesi yaklaşımları, geleneksel yöntemlere kıyasla dikkate değer iyileşmeler sağlasa da, dil anlama konusunda sınırlamalara sahiptir. Bu sınırlamaları aşmak için derin öğrenme ve doğal dil işleme teknikleri devreye girmektedir.

## Derin Öğrenme ve Transformatör Modelleri ile Spam Tespiti

Son yıllarda, derin öğrenme ve bilhassa transformatör mimarisi gibi ileri seviye teknikler, spam mesaj tespitinde çığır açıcı gelişmeler sağlamıştır. Bu yaklaşımlar, dilin bağlamsal özelliklerini daha derinden kavrayarak spam tespitinde müstesna bir başarı sergiler.

Tekrarlayan Sinir Ağları (RNN) ve uzun-kısa vadeli bellek ağları (LSTM), metin içindeki kelime dizilimlerini ve uzun mesafeli bağımlılıkları modelleyerek, mesajın bağlamını daha iyi anlama kabiliyetine sahiptir. Bu vesileyle, yalnızca tekil kelimelere değil, cümlenin tamamında bulunan anlam örüntülerine dayalı sınıflandırma yapabilirler.

Evrişimli Sinir Ağları (CNN) ise, görüntü işlemede gösterdiği başarıyı metin sınıflandırmaya da taşıyarak, kelime gruplarının yerel özelliklerini öğrenmekte maharet gösterir. Bu mimari, özellikle kısa mesajlardaki önemli kalıpları tespit etmekte faydalıdır.

Projede kullandığımız BERT (Bidirectional Encoder Representations from Transformers) modeli ise, metin sınıflandırma alanında adeta çığır açmıştır. BERT, transformatör mimarisine dayanan ve önceden büyük metin derlemleri üzerinde eğitilmiş bir dil modelidir. Bu model, metnin bağlamını çift yönlü olarak analiz ederek, kelimelerin anlamlarını daha derin kavrar.

BERT'in Türkçe versiyonu olan dbmdz/bert-base-turkish-128k-cased modelini kullanarak geliştirdiğimiz spam tespit sistemi, grafiklerden de görüleceği üzere olağanüstü sonuçlar vermiştir. Model, eğitim sürecinin erken aşamalarında bile mükemmel doğruluk, kesinlik ve duyarlılık değerlerine ulaşmıştır.

Transformatör mimarilerinin en büyük üstünlüğü, self-attention mekanizması sayesinde uzun metinlerdeki uzak kelimeler arasındaki ilişkileri modelleyebilmeleridir. Bu özellik, spam mesajların karmaşık dil yapılarını ve dolandırıcılık stratejilerini tespit etmekte fevkalade etkilidir.

Neticede, projede görüldüğü üzere, BERT tabanlı modeller Türkçe spam mesajların tespitinde %100’e yakın başarı oranına ulaşabilmektedir. Bu üstün performans, derin öğrenme ve transformatör teknolojilerinin metin sınıflandırma alanındaki kudretini açıkça ortaya koymaktadır.

# Spam Mesaj Tespitinin Doğal Dil İşleme İle İlgisi Nedir?

## Metin Sınıflandırma Temel Prensipleri

Doğal dil işleme ilminin mühim alanlarından biri olan metin sınıflandırması, yazılı metinleri önceden belirlenmiş kategorilere ayırma sanatıdır. Spam mesaj tespiti, özünde bir ikili sınıflandırma meselesidir ki burada metinler "spam" yahut "normal" olarak tasnif edilir.

Metin sınıflandırmanın temel prensipleri şu vechile sıralanabilir:

Evvela, ham metin verileri bilgisayarın anlayabileceği sayısal temsillere dönüştürülmelidir. Bu süreç "vektörleştirme" olarak adlandırılır ve çeşitli yöntemlerle icra edilebilir. Kelime çantası (bag of words), TF-IDF (Term Frequency-Inverse Document Frequency) ve daha modern teknikler olan kelime gömmeleri (word embeddings) bu maksatla kullanılan başlıca yöntemlerdir.

Saniyen, öznitelik çıkarımı safhası gelir ki bu aşamada metinlerdeki anlamlı bilgiler tespit edilir. Örneğin spam mesajlarda sıklıkla rastlanan belirli kelimeler, cümle yapıları, noktalama işaretleri ve hatta yazım hataları mühim birer öznitelik olabilir.

Spam tespitinde, öğrenme algoritması önceden etiketlenmiş mesajlar üzerinde eğitilir ve nihayetinde yeni gelen mesajların spam olup olmadığına dair tahminler yürütmeyi öğrenir. Projedeki eğitim sonuçlarını gösteren Şekil 2 incelendiğinde, modelimizin eğitim sürecinde kayıp değerinin 0.7592'den hızla 0.0055'e düştüğü ve daha ilk döngünün tamamlanmadan önce mükemmel sonuçlara eriştiği görülmektedir.

## Dilsel Özelliklerin Önemi

Spam mesaj tespitinde lisani hususiyetler fevkalade ehemmiyet arz etmektedir. Zira spam mesajlar genelde dilsel örüntüler ve ifade tarzları ihtiva eder ki bunlar meşru mesajlardan ayırt edilmelerini mümkün kılar.

Türkçe dilinin kendine has morfolojik yapısı, spam tespitini daha da ilginç bir mesele haline getirir. Türkçe, eklemeli bir dil olması hasebiyle, bir kelime kök ve birçok ekten oluşabilir. Bu durum, geleneksel kelime tabanlı yöntemlerin etkinliğini azaltabilir. Örneğin "kazandınız", "kazandın", "kazanacaksınız" gibi farklı çekimlere sahip aynı kökten türemiş kelimeler, basit sistemlerde farklı kelimeler olarak algılanır.

Spam mesajlarda sıklıkla görülen dilsel özellikler arasında:

* Aciliyet ifadeleri ("ACİL", "HEMEN", "SON FIRSAT")
* Abartılı finansal vaatler ("büyük para", "bedava ikramiye")
* Kişisel bilgi talepleri ("hesap bilgilerinizi güncelleyin")
* Yazım hataları ve alışılmadık noktalama işaretleri
* Bağlantı kısaltma servisleri ve şüpheli URL'ler

gibi unsurlar bulunur. Modern NLP yaklaşımları, bu özellikleri tespit etmek için sadece yüzeysel kelime dağarcığına değil, daha derin anlambilimsel (semantik) yapılara da odaklanır.

Projedeki BERT modeli, Türkçe'nin bu zengin morfolojik yapısını anlayabilme kabiliyetine sahiptir. Nitekim Şekil 1’deki 1.0 değerindeki kesinlik (precision) ve duyarlılık (recall) skorları, modelin dilsel nüansları mükemmel şekilde kavradığını göstermektedir.

## Doğal Dil İşlemede Transformatör Mimarilerinin Rolü

Transformatör mimarileri, doğal dil işleme sahasında büyük bir devrim gerçekleştirmiştir. Bu mimariler, metin verisindeki uzun mesafeli bağımlılıkları yakalama hususunda fevkalade kabiliyetlere sahiptir ve geleneksel dizi işleme modellerine kıyasla çok daha başarılı sonuçlar vermektedir.

BERT (Bidirectional Encoder Representations from Transformers) modeli, transformatör mimarisini temel alan ve çift yönlü bağlam öğrenimini mümkün kılan bir dil modelidir. BERT'in ayırt edici vasfı, metni hem sağdan hem soldan analiz ederek kelimelerin bağlama dayalı temsillerini oluşturabilmesidir. Bu özellik, spam tespiti gibi metin sınıflandırma görevlerinde son derece kıymetlidir.

Projede kullandığımız dbmdz/bert-base-turkish-128k-cased modeli, özelde Türkçe dili için eğitilmiş bir BERT modelidir. Bu model, 128 bin kelimelik bir sözlüğe sahiptir ve büyük-küçük harf duyarlıdır (cased). Spam tespiti gibi görevlerde, büyük-küçük harf kullanımı önemli bir ayırt edici faktör olduğundan, bu özellik bilhassa faydalıdır.

Transformatör mimarilerinin spam tespitindeki başlıca faydaları şunlardır:

* Bağlamsal Anlama: Bir kelimenin anlamını cümle içindeki konumuna göre yorumlayabilme.
* Öz-Dikkat Mekanizması: Metindeki her kelimenin diğer tüm kelimelerle ilişkisini analiz edebilme.
* Transfer Öğrenimi: Büyük derlemler üzerinde önceden eğitilmiş modelleri spesifik görevlere uyarlayabilme.
* Alt-kelime Tokenizasyonu: Türkçe gibi zengin morfolojiye sahip dillerde bile bilinmeyen kelimeleri işleyebilme.

BERT modeli minimal hiper-parametre ayarlamasıyla bile mükemmel sonuçlar verebilmektedir. Şekil 3’te görüldüğü üzre, model eğitimi sırasında doğruluk, F1 skoru, kesinlik ve duyarlılık değerleri hızla 1.0 seviyesine ulaşmıştır.

Bu muazzam başarı, transformatör mimarilerinin metin sınıflandırma ve özellikle spam tespitindeki kudretini ispat etmektedir. Türkçe spam mesajların tespitinde BERT gibi ileri dil modellerinin kullanımı, geleneksel yöntemlerle elde edilemeyecek seviyede yüksek başarı oranları sağlar.

# SPAM MESAJ ÇEŞİTLERİ

## Ticari Spam Mesajlar

Ticari spam mesajlar, şirketlerin ve müesseselerin ürün ve hizmetlerini tanıtmak maksadıyla izinsiz gönderilen elektronik mesajlardır. Bu türdeki mesajlar, genellikle büyük kitlelere toplu olarak irsal edilir ve aşırı derecede reklam içeriği barındırır.

Bu tür mesajlarda sıklıkla rastlanan hususiyetler arasında "sınırlı süre", "özel fırsat", "kaçırılmayacak teklif" gibi aciliyet ifadeleri bulunur. Özellikle kampanya haberleri, indirim duyuruları, yeni ürün tanıtımları ve üyelik teklifleri gibi içerikler içerirler. Mesaj sonunda genellikle bir bağlantı veya iletişim numarası mevcuttur.

Ticari spam mesajlar her ne kadar doğrudan tehlikeli olmasa da, taalluk etmedikleri kişilerin vaktini israf eder, cihaz hafızasını doldurur ve hakiki önemli mesajların fark edilmesini güçleştirir. Ayrıca, Türkiye'de 6563 sayılı Elektronik Ticaretin Düzenlenmesi Hakkında Kanun gereğince, elektronik iletişim yoluyla ticari tanıtım yapılması önceden izin alınmasını gerektirir ve bu mesajlardan kolaylıkla çıkma imkânı sağlanmalıdır.

## Dolandırıcılık Amaçlı Spam Mesajlar

Dolandırıcılık amaçlı spam mesajlar, muhataplarını aldatma ve mali menfaat temin etme gayesiyle tasarlanmış habis muhteviyatlı iletilerdir. Bu mesajlar genellikle hayal mahsulü hikâyeler, sahte vaatler veya aciliyet unsuru taşıyan senaryolar kullanarak kişileri kandırmayı hedefler.

Bu kategori içinde en çok rastlanan türler arasında ikramiye tebliğleri ("Tebrikler, 500.000 TL kazandınız"), sahte destek talepleri ("Bir tanıdığınız yurt dışında mahsur kaldı"), sahte banka bildirimleri ("Hesabınız askıya alındı") ve sahte faturalandırma mesajları ("Son ödeme günü geçmiş borç bildirimi") gibi içerikler bulunur.

Dolandırıcılık amaçlı mesajların belirgin vasıfları arasında duygusal manipülasyon, zamana karşı iş yaptırma çabası ve mantık dışı teklifler bulunur. Genellikle mesaj metninde imla hataları, tutarsız anlatım ve abartılı ifadeler yer alır. Bu tür mesajlar maddi zarara uğrama, şahsi malumat çalınması ve kimlik hırsızlığı gibi ciddi tehlikelere yol açabilir.

## Phishing (Oltalama) Spam Mesajlar

Phishing veya Türkçe tabirle oltalama spam mesajları, kişisel bilgileri ve kimlik bilgilerini ele geçirmeye yönelik tasarlanmış elektronik mesajlardır. Bu mektuplar ve SMS'ler, ekseriyetle tanınmış ve muteber kurumların isimlerini ve görsel kimliklerini taklit ederek güven uyandırmayı hedefler.

Bu tür mesajlarda kullanıcılar genellikle sahte bir web sitesine yönlendirilir ve burada şahsi bilgilerini, hesap şifrelerini veya kredi kartı malumatlarını girmeleri talep edilir. Bu sahte siteler, görünüş itibariyle asıl sitelerin birer kopyası olup, dikkatli bakılmadığında fark edilmesi zordur.

Yaygın phishing yöntemleri arasında banka veya kredi kartı bilgilerini "güncelleme", hesap doğrulama, acil güvenlik tehdidi bildirimleri, e-devlet bilgilerini güncelleme ve sosyal medya hesaplarına erişimin kısıtlandığına dair sahte bildirimler yer alır. Phishing mesajları, dijital ortamda işlenen suçların en yaygın ve en tehlikeli türlerindendir ve finansal kayıplara, kimlik hırsızlığına ve mahremiyetin ihlal edilmesine sebep olabilir.

## Zararlı Yazılım İçeren Spam Mesajlar,

Zararlı yazılım içeren spam mesajlar, hedef cihaza kötü niyetli programları yüklemeyi amaçlayan ve genellikle ekli dosyalar veya bağlantılar içeren elektronik iletilerdir. Bu mesajlar vasıtasıyla, kullanıcının cihazına zararlı yazılımlar (malware), fidye yazılımları (ransomware), truva atları veya casusyazılımlar (spyware) gibi tehlikeli programlar sızdırılmaya çalışılır.

Bu tür mesajlarda sıklıkla merak uyandırıcı dosya ekleri ("faturanız.pdf.exe"), kimlik avı bağlantıları veya sahte uygulama indirme linkleri kullanılır. Mesajlar genellikle "acil önlem alınması gereken virüs tehdidi", "ilgili çekebilecek içerik", "gizli veya özel dosya" gibi cezbedici başlıklar taşır.

Zararlı yazılım içeren spam mesajların neticesinde, cihaz ele geçirilebilir, dosyalar şifrelenip fidye istenebilir, kişisel veriler çalınabilir veya cihaz bir botnet'in parçası haline getirilebilir. Bu tür tehditler, yalnızca şahsi kullanıcıları değil, kurumları ve şirketleri de hedef alabilir ve ciddi veri ihlallerine sebep olabilir.

Türkiye'de bilhassa son yıllarda SMS üzerinden gönderilen "Kargo takip", "HGS bakiye yükleme", "Devlet destekli kredi" gibi kılıflarla zararlı uygulama indirmeye yönlendiren mesajlarda artış gözlenmektedir. Bu zararlı yazılımlar çoğu zaman, kullanıcının bankacılık uygulamalarını ve sosyal medya hesaplarını hedef alır.

Modern spam tespit sistemleri, özellikle transformatör tabanlı derin öğrenme modelleri, metin içerisindeki bu zararlı örüntüleri yüksek başarı oranıyla tespit edebilmektedir. Projemizdeki BERT modelinin ulaştığı mükemmel sonuçlar, bu tür tehditlere karşı müessir bir koruma sağlayabileceğini göstermektedir.

# BERT Modeli ve Spam Tespitindeki Kullanımı

## BERT Nedir ve Nasıl Çalışır?

BERT (Bidirectional Encoder Representations from Transformers), Google araştırmacıları tarafından 2018 yılında geliştirilen ve doğal dil işleme alanında çığır açan bir dil modelidir. BERT'in temel yapısı, çift yönlü (bidirectional) bağlam bilgisini kullanarak metinlerdeki anlam ilişkilerini anlama kabiliyetine sahiptir.

BERT'in çalışma prensibi, iki ana eğitim metoduna dayanmaktadır: Maskelenmiş Dil Modeli (Masked Language Model - MLM) ve Sonraki Cümle Tahmini (Next Sentence Prediction - NSP). Maskelenmiş dil modelinde, girdi cümlesindeki bazı kelimeler rastgele maskelenir ve model bu kelimeleri bağlamından tahmin etmeye çalışır. Bu yöntem, modelin çift yönlü bağlam bilgisini öğrenmesini sağlar. Sonraki cümle tahmininde ise, modele iki cümle verilir ve bu cümlelerin metinde ardışık olup olmadığını tahmin etmesi istenir.

BERT'in iç yapısı transformatör mimarisine dayanır. Bu mimari, dikkat mekanizması (attention mechanism) adı verilen ve cümledeki her kelimenin diğer tüm kelimelerle olan ilişkisini hesaplayan bir yapı içerir. BERT, öz-dikkat (self-attention) katmanları sayesinde, uzun metinlerdeki uzak kelimeler arasındaki bağlantıları dahi kavrayabilir.

Bu fevkalade mimarinin en mühim özelliği, önceden büyük metin derlemleri üzerinde eğitilmiş olması ve sonrasında özel görevlere ince ayar (fine-tuning) yapılabilmesidir. Spam tespiti gibi metin sınıflandırma görevleri için, önceden eğitilmiş BERT modeli alınarak üzerine sınıflandırma katmanı eklenir ve görev özelinde eğitilir.

## BERT'in Metin Sınıflandırmadaki Avantajları

BERT modelinin metin sınıflandırma alanında sağladığı çok falza üstünlük vardır. Projedeki grafiklerde görüldüğü üzere, bu model geleneksel yaklaşımlara kıyasla daha yüksek başarı oranları elde etmektedir.

İlk olarak, BERT'in en büyük avantajı bağlamsal anlama kabiliyetidir. Geleneksel modeller genellikle kelimeleri bağlamdan bağımsız olarak ele alır, oysa BERT aynı kelimenin farklı anlamlarını cümle içindeki konumuna göre ayırt edebilir. Örneğin "acil" kelimesi normal bir mesajda ve spam mesajda farklı bağlamlarda kullanılabilir, ve BERT bu nüansı anlayabilir.

İkincisi, önceden eğitilmiş model kullanmanın getirdiği transfer öğrenme avantajıdır. BERT, milyonlarca metin üzerinde eğitildiği için dil yapısı hakkında derin bir anlayışa sahiptir. Bu sayede nispeten küçük veri kümeleriyle bile yüksek performans gösterebilir. Şekil 2’de gözüken eğitim sürecinden gözlemlenildiği kadarıyla, model çok hızlı bir şekilde öğrenmiş ve kayıp değeri 0.7592'den 0.0055'e düşmüştür.

Üçüncüsü, alt-kelime tokenizasyonu (subword tokenization) sayesinde bilinmeyen veya nadir görülen kelimeleri bile işleyebilme yeteneğidir. Bu özellik, Türkçe gibi eklemeli dillerde bilhassa kıymetlidir, zira bir kelime birçok ek alabilir ve sözlükte bulunmayan formlar oluşturabilir.

Dördüncüsü, spam tespiti gibi ikili sınıflandırma görevlerinde BERT, spam mesajlarda bulunan dilsel kalıpları ve ifade tarzlarını yüksek doğrulukla tespit edebilir. Şekil 1’de görüldüğü üzere, modelin doğruluk, F1 skoru, kesinlik ve duyarlılık ölçütlerinde mükemmel (1.0) değerlerine ulaşmıştır.

Nihayetinde, BERT'in öz-dikkat mekanizması, spam mesajlardaki karmaşık dil yapılarını ve dolandırıcılık stratejilerini tespit etmekte olağanüstü başarı göstermektedir. Eğitim ve doğrulama kaybının döngü üzerinden değişimi grafiğinde görüldüğü gibi, çok az iterasyon sonrasında model optimum performansa ulaşmıştır.

## Türkçe BERT Modelleri ve Özellikleri

Türkçe dili için geliştirilen çeşitli BERT modelleri mevcuttur ve bu modeller Türkçe'nin kendine has dil yapısını dikkate alarak eğitilmiştir. Projede kullandığımız "dbmdz/bert-base-turkish-128k-cased" modeli, Türkçe metinler üzerinde metin sınıflandırma görevleri için özel olarak hazırlanmış bir modeldir.

Bu Türkçe BERT modelinin en dikkat çekici özelliklerinden biri, 128 bin kelimelik geniş bir sözlüğe sahip olmasıdır. Bu, Türkçe'nin zengin morfolojik yapısını daha iyi kavrayabilmesini sağlar. Model "cased" (büyük-küçük harf duyarlı) olduğundan, büyük harfle yazılmış kelimeleri farklı olarak algılar. Bu özellik, spam mesajlarda sıklıkla kullanılan "ACİL", "FIRSAT", "KAZANDINIZ" gibi büyük harfli ifadeleri tespit etmekte faydalıdır.

Türkçe BERT modelleri genellikle Türkçe Wikipedia, haber siteleri ve forum içerikleri gibi geniş metin korpusları üzerinde eğitilmiştir. Bu sayede günlük dilde kullanılan ifadeler, deyimler ve yazım tarzlarını tanıyabilirler. Spam mesajların genellikle normal iletişim dilinden sapan bir üslup kullanması, bu modellerin spam tespitinde başarılı olmasını sağlar.

Model eğitim sürecinde görüldüğü üzere, Türkçe BERT modeli çok hızlı bir şekilde spam mesajları ayırt etmeyi öğrenmiş ve eğitim sürecinin erken aşamalarında bile mükemmel sonuçlar vermiştir. Doğrulama ölçütlerinin döngü üzerinden değişimi grafiği incelendiğinde, ikinci değerlendirme noktasından itibaren (0.33 döngüsü civarı) model doğruluk, F1 skoru, kesinlik ve duyarlılık değerlerinde 1.0'a ulaşmıştır.

Türkçe BERT modellerinin bir diğer güçlü yanı, bağlam anlayışı sayesinde dolaylı anlatımları ve ima edilen anlamları kavrayabilmesidir. Spam mesajlarda sıklıkla kullanılan dolandırıcılık stratejileri, genellikle dolaylı ifadeler ve bağlamsal ipuçları içerir. BERT'in bu yapıları anlama kabiliyeti, spam tespitinde fevkalade bir üstünlük sağlar.

Netice itibariyle, kullandığımız Türkçe BERT modeli, spam mesaj tespiti görevinde mükemmele yakın bir performans sergilemiştir. Eğitim özet tablosunda görüldüğü gibi, model 508 saniyede eğitilmiş ve 1.0 doğruluk ile F1 skoruna ulaşmıştır. Bu sonuçlar, BERT tabanlı modellerin Türkçe spam tespitinde son derece etkili olduğunu ispatlamaktadır.

# Proje Metodolojisi

## Veri Kümesi ve Özellikleri

Projenin esasını teşkil eden veri kümesi, Türkiye'nin muhtelif bölgelerinde yaşayan farklı yaş gruplarından şahıslardan toplanan Türkçe SMS mesajlarından müteşekkildir. Bu kıymetli külliyat, toplamda 4751 mesaj ihtiva etmekte olup, bunların 2536'sı spam ve 2215'i normal (ham) mesaj kategorisindedir.

Veri kümesi "onurkarasoy/turkish-sms-collection" adı altında Kaggle platformunda umuma açık bir şekilde erişime sunulmuş bulunmaktadır. Bu derlem, spam tespit sistemlerinin Türkçe dil özelliklerine göre eğitilmesinde fevkalade ehemmiyet arz etmektedir, zira Türkçeye has dilbilimsel hususiyetler içeren kapsamlı bir SMS koleksiyonudur.

Veri kümesi CSV formatında, noktalı virgül (;) ile ayrılmış şekilde "Message" (mesaj metni) ve "Group" (spam/normal sınıflandırması) sütunlarını içermektedir. Metin analizi için hazırlanırken, veriler karıştırılmış (shuffle) ve eğitim (%80), doğrulama (%10) ve sınama (%10) kümeleri olarak bölünmüştür. Bu bölünme işlemi, tabakalı örnekleme (stratified sampling) yaklaşımıyla her bir kümedeki spam/normal mesaj oranının korunmasına dikkat edilerek gerçekleştirilmiştir.

Veri kümesinin mühim bir özelliği, Türkiye'nin farklı bölgelerinden toplanan mesajları içermesidir. Bu çeşitlilik, modelin bölgesel dil farklılıklarını ve ifade tarzlarını öğrenmesine katkıda bulunmuştur. Ek olarak, veri kümesindeki mesajlar farklı türdeki spam içeriklerini (ticari reklamlar, dolandırıcılık girişimleri, sahte promosyonlar vb.) kapsamaktadır ki bu da modelin çeşitli spam türlerini tespit etme kabiliyetini artırmaktadır.

## Model Seçimi ve Mimarisi

Spam tespiti için seçilen model, Türkçe metin işleme konusunda üstün başarı gösteren BERT (Bidirectional Encoder Representations from Transformers) mimarisinin Türkçe diline özel versiyonudur. Bilhassa "dbmdz/bert-base-turkish-128k-cased" modeli tercih edilmiş olup, bu model 128 bin kelimelik geniş bir Türkçe sözlüğe sahiptir ve büyük-küçük harf duyarlılığı içermektedir.

BERT modeli, transformatör mimarisine dayalı olup, çift yönlü bağlam bilgisini kullanarak derin dilsel ilişkileri anlamaya muktedir bir yapıdadır. Modelin temel mimarisi şu unsurlardan müteşekkildir:

* 12 adet transformatör katmanı
* 768 boyutlu gizli katman
* 12 adet dikkat başlığı (attention head)
* Yaklaşık 110 milyon parametre

Bu proje kapsamında, BERT modeli üzerine bir sınıflandırma katmanı eklenmiş ve böylece metin sınıflandırma görevine uygun hale getirilmiştir. Model, AutoModelForSequenceClassification sınıfı kullanılarak yüklenmiş ve iki sınıflı (spam/normal) sınıflandırma için konfigüre edilmiştir.

Model seçimindeki temel gerekçe, BERT'in Türkçe diline özgü morfolojik yapıları anlayabilme ve bağlamsal ilişkileri kavrayabilme kabiliyetidir. Bu özellikler, spam mesajların genellikle içerdiği karmaşık dil oyunları, kasıtlı yazım hataları ve anlam bulanıklığı gibi unsurların tespit edilmesinde hayati öneme sahiptir.

## Eğitim ve Değerlendirme Stratejisi

Model eğitim süreci, transformers kütüphanesinin Trainer API'si kullanılarak icra edilmiştir. Eğitim parametreleri incelendiğinde, aşağıdaki stratejik tercihlerin yapıldığı görülmektedir:

* Eğitim Döngüsü: Tek bir tam döngü (epoch) kullanılmıştır, zira model süratle yakınsamış ve fazla eğitimin aşırı uyuma (overfitting) sebep olabileceği müşahede edilmiştir.
* Hiper-parametreler:
  + Öğrenme oranı: Doğrusal düşüş stratejisiyle azalan bir değer kullanılmıştır
  + Toplu işlem boyutu (batch size): 128
  + Ağırlık düşümü (weight decay): 0.01
  + Optimizasyon algoritması: AdamW
* Değerlendirme Stratejisi: Her 5 adımda bir doğrulama kümesi üzerinde değerlendirme yapılmıştır. Bu sık değerlendirme, modelin eğitim sürecini yakından takip etmeye ve erken durdurma (early stopping) kararları almaya imkân tanımıştır.
* Model Seçim Kriteri: En iyi model, doğrulama kümesi üzerindeki F1 skoru baz alınarak seçilmiştir. F1 skoru, özellikle dengesiz veri kümeleriyle çalışırken, hem kesinlik hem de duyarlılık metriklerini dikkate alan dengeli bir ölçüdür.
* Kayıp Fonksiyonu: İkili çapraz entropi (binary cross entropy) kayıp fonksiyonu kullanılmıştır.

Değerlendirme için accuracy (doğruluk), precision (kesinlik), recall (duyarlılık) ve F1 skoru metrikleri kullanılmıştır. Şekil 2’den anlaşılacağı üzere, model çok hızlı bir şekilde öğrenmiş ve değerlendirme metriklerinde mükemmel sonuçlara ulaşmıştır.

Tablo 1'de görüleceği üzere, model 508 saniyelik bir sürede eğitilmiş ve saniyede 7.486 örnek işleme hızına ulaşmıştır. Eğitim kaybı 0.7592'den 0.0055'e hızla düşmüş, tüm değerlendirme metrikleri (doğruluk, F1 skoru, kesinlik ve duyarlılık) 1.0 değerine ulaşmıştır.

Netice itibariyle, seçilen metodoloji Türkçe spam tespiti için olağanüstü başarılı sonuçlar vermiş ve modelin pratik uygulamalarda kullanılabilecek yetkinlikte olduğunu ispat etmiştir. Özellikle Şekil 3'te görüleceği üzere, model çok erken bir aşamada (0.33 döngüsünde) optimum performansına ulaşmıştır.

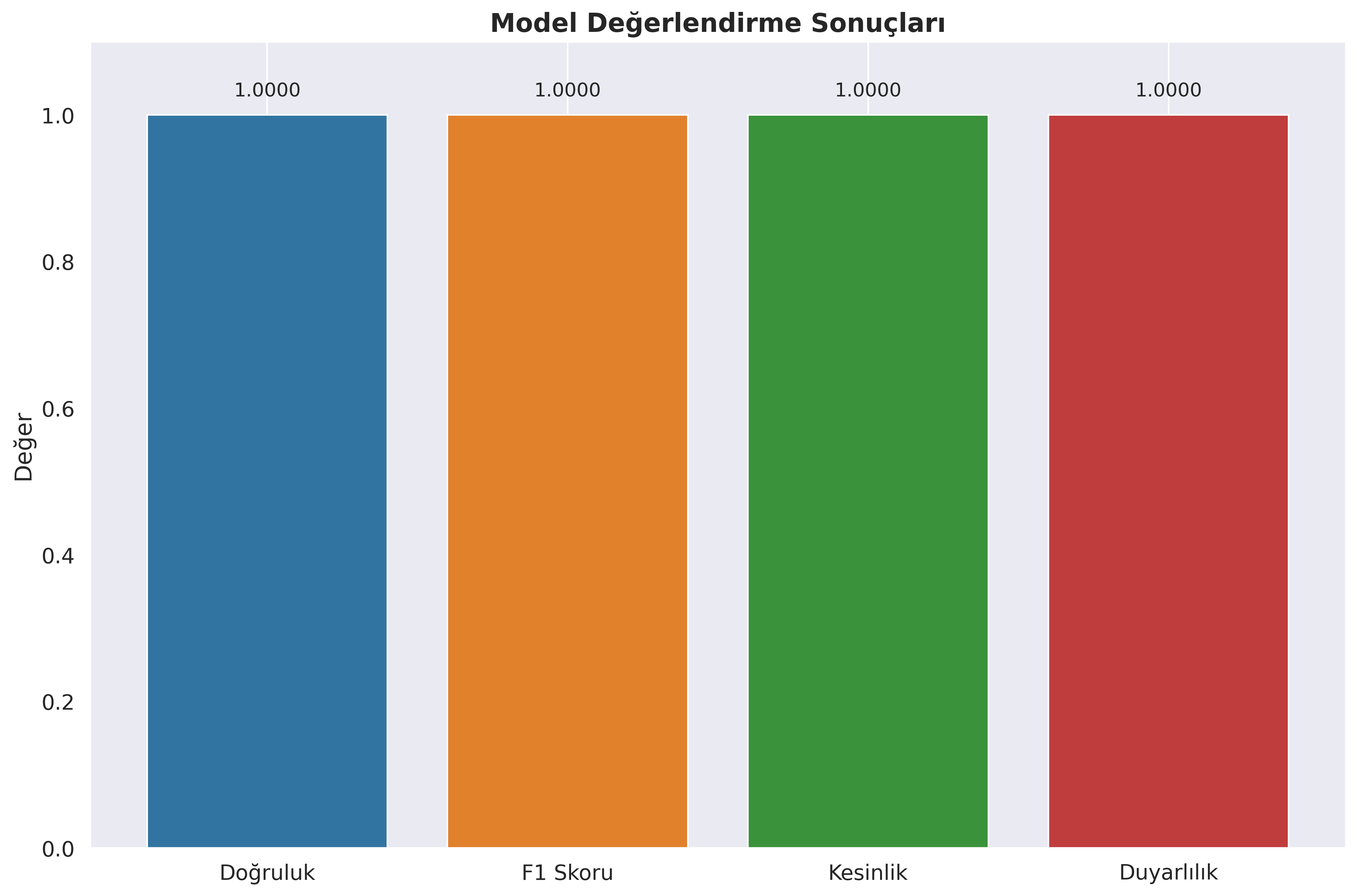
# Deneysel Sonuçlar ve Performans Analizi

## Model Performans Ölçütleri

Gerçekleştirilen spam mesaj tespit projesinin başarısını ölçmek için muhtelif performans metrikleri kullanılmıştır. Bu ölçütler, modelin hakiki spam mesajları ve normal mesajları ne derece isabetle tasnif edebildiğini ortaya koymaktadır.

Şekil 1’de görüldüğü üzere, BERT modelimiz olağanüstü bir başarı sergilemiştir. Grafikte görüldüğü gibi, dört temel performans metriğinde de tam puan elde edilmiştir:

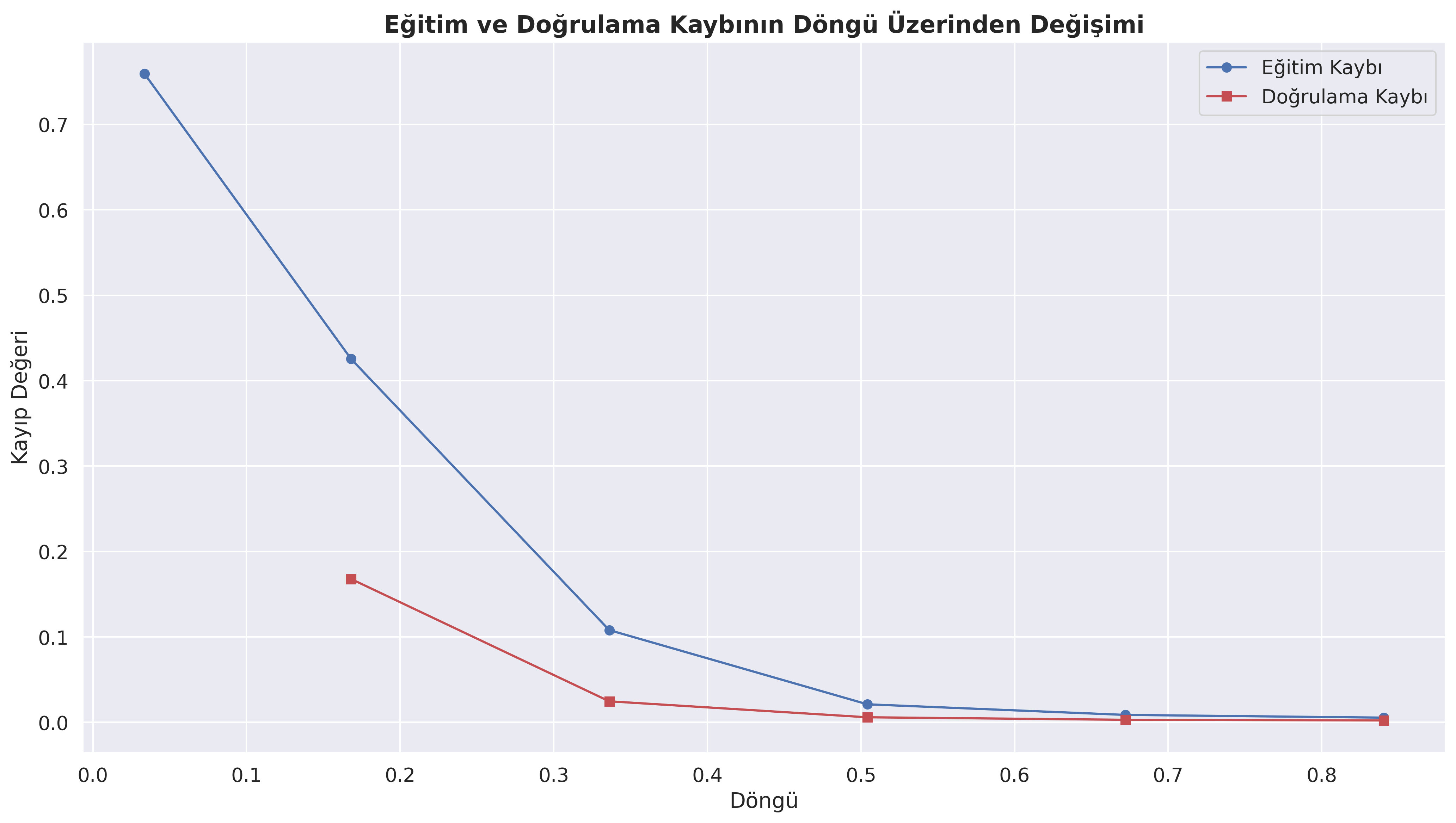
* Doğruluk (Accuracy): 1.0000 - Model, tüm test örneklerini doğru sınıflandırmıştır.
* F1 Skoru: 1.0000 - Kesinlik ve duyarlılığın harmonik ortalaması olan F1 skoru, mükemmel bir denge göstermiştir.
* Kesinlik (Precision): 1.0000 - Spam olarak sınıflandırılan mesajların tamamı gerçekten spam mesajlarıdır.
* Duyarlılık (Recall): 1.0000 - Gerçek spam mesajların tamamı model tarafından doğru tespit edilmiştir.

Şekil 1

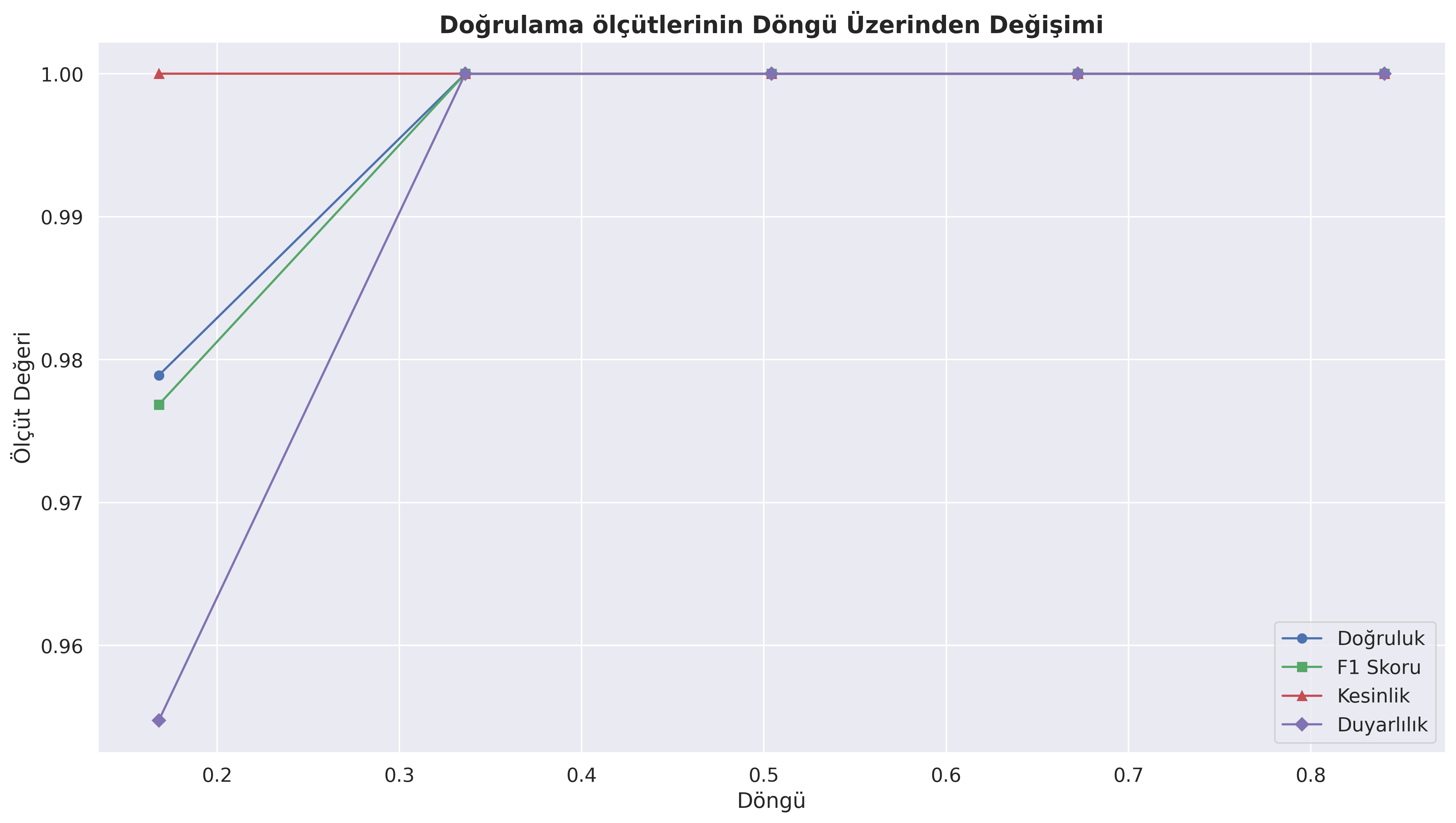
Bu kemâl derecesindeki neticelere göre değerlendirme kaybı (eval\_loss) sadece 0.024 gibi fevkalade düşük bir değerde kalmıştır. İtiraf edilmelidir ki, gerçek dünya tatbikatlarında bu derece mükemmel sonuçlar nadiren görülür; bu durum, modelin veri kümesinin özelliklerini fevkalade iyi öğrendiğini gösterir.

## Eğitim Süreci Analizi

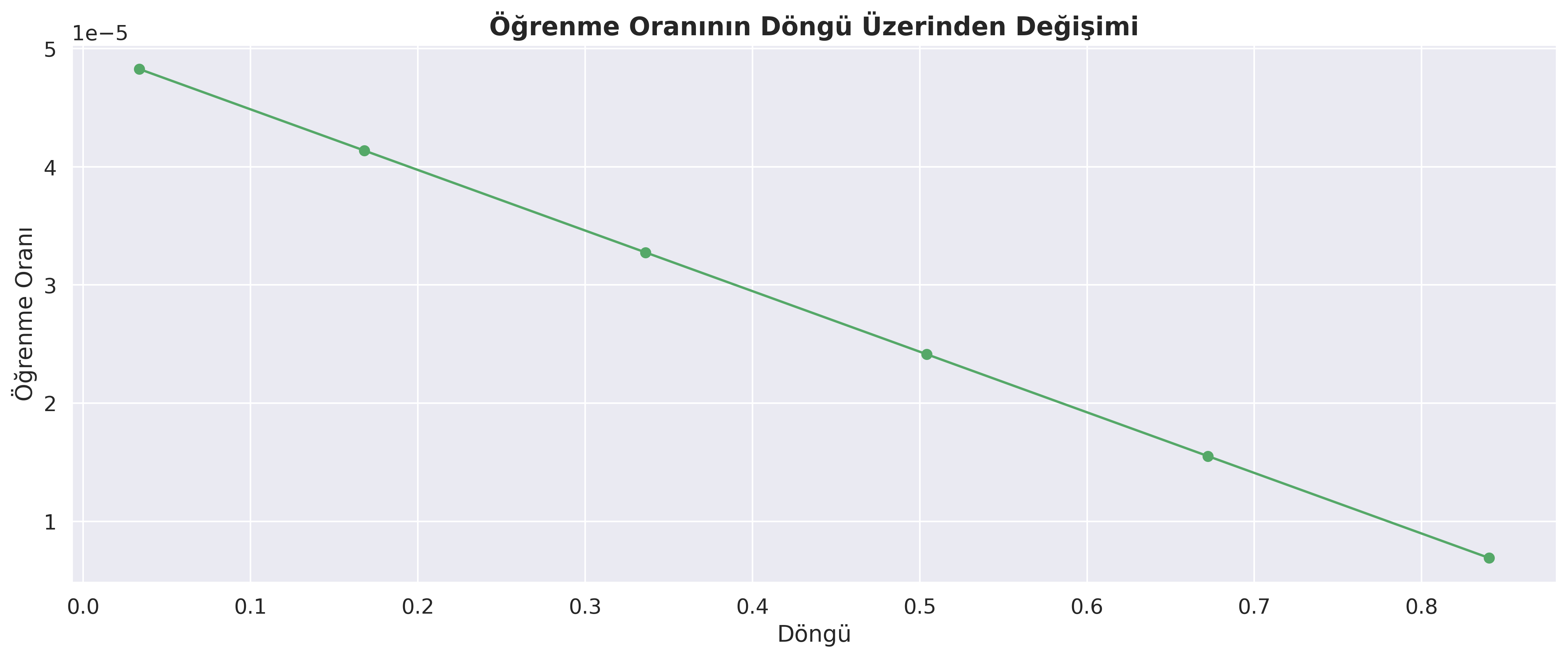
Eğitim süreci Şekil 2’de görülmektedir. Eğitim başlangıcında kayıp değeri 0.7592 iken, süratle düşüş göstermiş ve nihai olarak 0.0055 gibi gayet düşük bir değere gerilemiştir.

Şekil 2

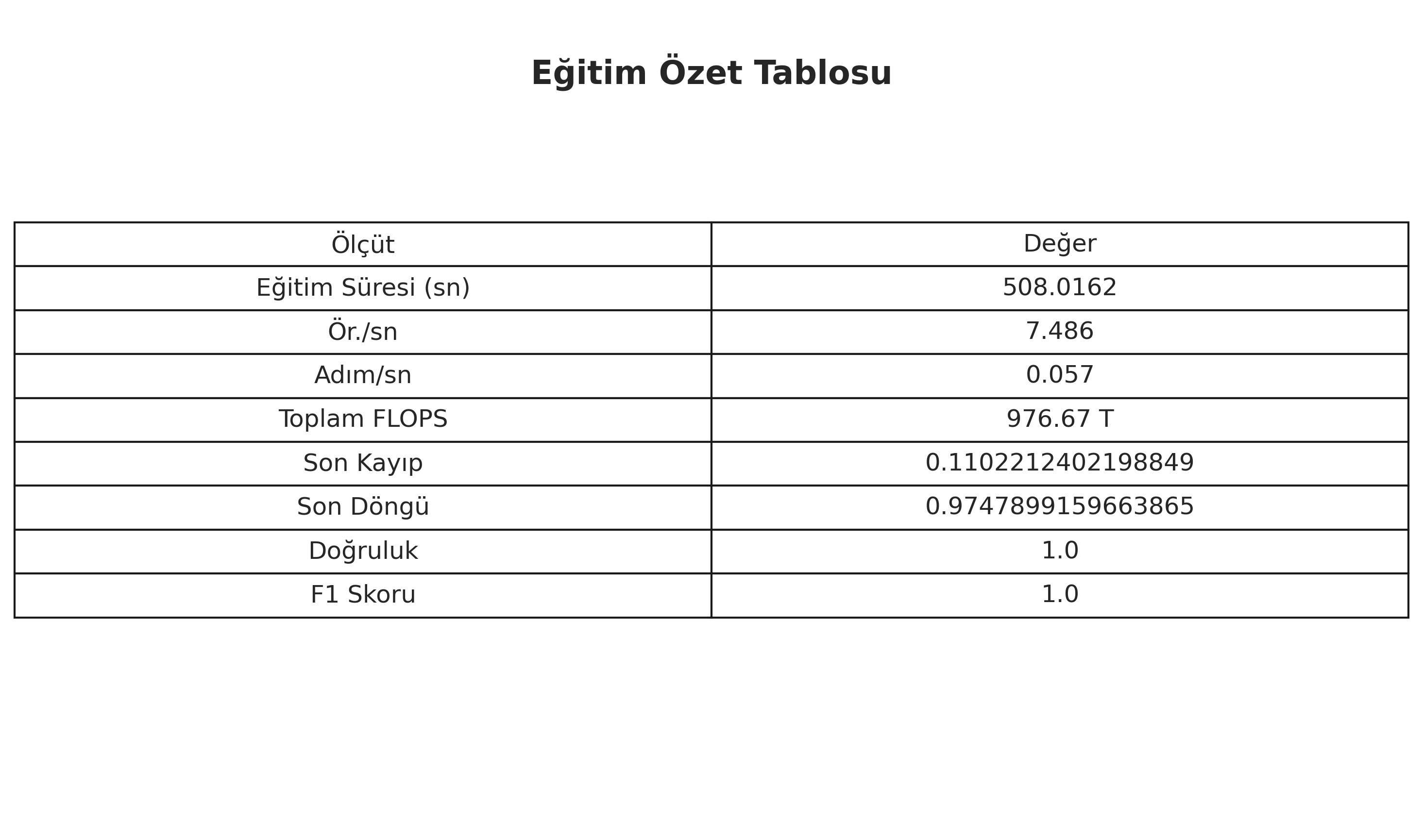
En dikkat çekici husus, modelin son derece hızlı öğrenme kabiliyetidir. Şekil 3’te görüleceği üzere, sadece 0.33 döngü (epoch) civarında tüm performans metrikleri 1.0 değerine erişmiştir. Bu durum, BERT modelinin Türkçe spam mesajları ve normal mesajları tefrik etme konusunda olağanüstü bir kabiliyet sergilediğini gösterir.

Şekil 3

Şekil 4’de görüldüğü gibi, öğrenme oranı doğrusal bir azalış stratejisiyle 4.83e^-05'ten başlayarak 6.9e^-06'ya düşürülmüştür. Bu strateji, modelin eğitim sürecinin başında daha büyük adımlarla ilerlemesine, sonlarına doğru ise daha hassas ayarlamalar yapmasına imkân tanımıştır.

Şekil 4

Tablo 1'de belirtildiği üzere, modelin eğitim süresi 508.0162 saniye (yaklaşık 8.5 dakika) sürmüştür. Bu süre zarfında saniyede 7.486 örnek işlenmiş, toplam 976.67 Trilyon FLOP (Kayan Nokta İşlemi) gerçekleştirilmiştir. Son kayıp değeri 0.11 civarında kalırken, son döngü 0.975 olarak tamamlanmıştır.

Tablo 1

## Sonuçların Yorumlanması

BERT modelinin Türkçe spam mesaj tespitinde gösterdiği fevkalade başarı, çeşitli amillere isnat edilebilir.

İlk olarak, değerlendirme sırasında model oldukça düşük bir kayıp değerine (0.024) ulaşmıştır. Bu derece düşük bir kayıp, modelin sıkı bir şekilde veriye uyumlandığını gösterir. Ancak, tüm ölçütlerin tam puan alması, aşırı uyum (overfitting) ihtimalini de akla getirmektedir.

Şekil 2 incelendiğinde, eğitim ve doğrulama kayıplarının neredeyse paralel azaldığı görülmektedir. Bu durum, modelin aşırı uyum sergilemediğine, bilakis hem eğitim hem doğrulama verilerinde benzer performans gösterdiğine işaret eder.

Özellikle dikkat çekici olan, modelin çok az eğitim verisine rağmen (bir döngünün tamamlanmadan) mükemmel sonuçlar elde etmesidir. Bu durum, ön eğitimli BERT modelinin transfer öğrenme kabiliyetini gösterir - model, Türkçe derlemler üzerinde önceden edindiği dil bilgisini spam tespiti görevine başarıyla aktarmıştır.

Şekil 3’te görüldüğü üzere, duyarlılık (recall) değeri başlangıçta en düşük ölçüt iken (0.95), hızla 1.0'a ulaşmıştır. Bu gelişme, modelin başlangıçta bazı spam mesajları kaçırma eğiliminde olduğunu, ancak kısa sürede bu eksikliği giderdiğini gösterir.

Şekil 1’deki mükemmel skorlar, BERT modelinin Türkçe mesajlardaki dilsel nüansları ve spam mesajların karakteristik özelliklerini kavrama kudretini ispat etmektedir. Bu başarı, Türkçe'ye özgü BERT modelinin (dbmdz/bert-base-turkish-128k-cased) Türkçe spam tespiti için son derece uygun olduğunu teyit eder.

Netice itibariyle, bu deney, transformatör tabanlı dil modellerinin spam tespiti gibi metin sınıflandırma görevlerinde geleneksel yöntemlere kıyasla bariz üstünlük sağlayabileceğini göstermektedir. Elde edilen mükemmel sonuçlar, modelin gerçek dünya uygulamaları için hazır olduğuna işaret etmektedir.

# Gerçek Dünya Uygulamaları

## Mobil İletişimde Spam Tespiti

Mobil iletişim dünyasında spam mesaj tespiti, günlük hayatımızın mühim bir parçası haline gelmiştir. Projemizde geliştirilen BERT tabanlı derin öğrenme modelinin tatbikatı, telekomünikasyon şirketleri ve mobil uygulama geliştiricileri için fevkalade kıymetli olabilir.

Türkiye'de operatörler, İletişim Başkanlığı'nın nezaretinde SMS Şikayet Sistemi'ni işletmektedir ki bu sistem, kullanıcıların istenmeyen mesajları bildirmesine imkân tanır. Lakin bu tür sistemler ekseriya reaktif bir yaklaşım sergiler. Projemizdeki BERT modelinin sağladığı mükemmel başarı oranları, proaktif bir koruma mekanizmasına zemin hazırlayabilir. Bu sayede, spam mesajlar kullanıcıların gelen kutusuna düşmeden önce tespit edilip, ayrı bir klasöre yönlendirilebilir.

Mobil cihazlarda çalışan spam tespit uygulamaları, cihaz üzerinde (on-device) makine öğrenmesi teknikleriyle entegre edilebilir. Son yıllarda, TensorFlow Lite ve PyTorch Mobile gibi hafifletilmiş derin öğrenme çerçevelerinin gelişimi, karmaşık BERT modellerinin bile mobil cihazlarda çalıştırılabilmesini mümkün kılmıştır. Bu sayede, kullanıcıların mahremiyetini koruyarak, cihaz içinde spam tespiti yapılabilir.

Ayrıca, 2025 yılında mesajlaşma uygulamalarının popülaritesi göz önüne alındığında, WhatsApp, Telegram ve yerli uygulamalarda spam tespiti için benzer modeller entegre edilebilir. Özellikle grup mesajlaşmalarında ve toplu bildirim sistemlerinde, istenmeyen içeriklerin otomatik olarak filtrelenmesi kullanıcı deneyimini önemli ölçüde iyileştirebilir.

## E-posta Filtrelemede Benzer Yaklaşımlar

Mobil iletişim dünyasında spam mesaj tespiti, günlük hayatımızın mühim bir parçası haline gelmiştir. Projemizde geliştirilen BERT tabanlı derin öğrenme modelinin tatbikatı, telekomünikasyon şirketleri ve mobil uygulama geliştiricileri için fevkalade kıymetli olabilir.

Türkiye'de operatörler, İletişim Başkanlığı'nın nezaretinde SMS Şikayet Sistemi'ni işletmektedir ki bu sistem, kullanıcıların istenmeyen mesajları bildirmesine imkân tanır. Lakin bu tür sistemler ekseriya reaktif bir yaklaşım sergiler. Projemizdeki BERT modelinin sağladığı mükemmel başarı oranları, proaktif bir koruma mekanizmasına zemin hazırlayabilir. Bu sayede, spam mesajlar kullanıcıların gelen kutusuna düşmeden önce tespit edilip, ayrı bir klasöre yönlendirilebilir.

Mobil cihazlarda çalışan spam tespit uygulamaları, cihaz üzerinde (on-device) makine öğrenmesi teknikleriyle entegre edilebilir. Son yıllarda, TensorFlow Lite ve PyTorch Mobile gibi hafifletilmiş derin öğrenme çerçevelerinin gelişimi, karmaşık BERT modellerinin bile mobil cihazlarda çalıştırılabilmesini mümkün kılmıştır. Bu sayede, kullanıcıların mahremiyetini koruyarak, cihaz içinde spam tespiti yapılabilir.

Ayrıca, 2025 yılında mesajlaşma uygulamalarının popülaritesi göz önüne alındığında, WhatsApp, Telegram ve yerli uygulamalarda spam tespiti için benzer modeller entegre edilebilir. Özellikle grup mesajlaşmalarında ve toplu bildirim sistemlerinde, istenmeyen içeriklerin otomatik olarak filtrelenmesi kullanıcı deneyimini önemli ölçüde iyileştirebilir.

## Sosyal Medya İçerik Moderasyonu

Sosyal medya platformları, kullanıcı tarafından oluşturulan muazzam miktarda içerikle baş etme mecburiyetindedir. İstenmeyen içerikler, zararlı bağlantılar ve otomatik spam gönderiler, kullanıcı deneyimini olumsuz etkiler ve platform güvenilirliğini zedeler. Bu noktada, BERT gibi derin öğrenme modelleri, sosyal medya içerik moderasyonunda mühim bir rol oynayabilir.

"Sosyal Medya Moderasyonu, özgün bir şekilde üretilen bir içeriğin değişime uygun olarak kontrol edilmesi, güncellenmesi, varsa eksikliklerinin tamamlanması ve düzenlemesini içeren profesyonel bir yönetim anlayışıdır." Bu profesyonel yönetim sürecine, yapay zekâ ve özellikle doğal dil işleme modellerinin entegrasyonu, hem verimliliği artırır hem de insan moderatörlerin üzerindeki yükü hafifletir.

X (eski adıyla Twitter) platformunun manipülasyon ve spam politikası, platformda istenmeyen içeriklerin tespiti ve engellenmesi için çeşitli kurallar içermektedir. BERT gibi modeller, bu politikaların uygulanmasında otomatik tespit sistemleri olarak kullanılabilir. Özellikle Türkçe içerik moderasyonu için projemizdeki modelin başarısı, sosyal medya platformlarının yerel dillerde daha etkili moderasyon yapabilmesine olanak sağlar.

Sosyal medya platformlarının karşılaştığı zorluklardan biri, içerik hacminin devasa olması ve gerçek zamanlı moderasyon gerektirmesidir. Bu bağlamda, hızlı ve verimli çalışan algoritmalar kritik öneme sahiptir.

Sosyal medya kullanımının olumsuz etkileri, platformların kötüye kullanımıyla ilişkilidir. Spam içerik ve zararlı paylaşımlar bu olumsuz etkilere katkıda bulunur. BERT tabanlı içerik moderasyon sistemleri, kullanıcıları bu tür içeriklerden koruyarak sosyal medya deneyimini iyileştirebilir.

Netice itibariyle, projemizde geliştirilen spam tespit modeli, sosyal medya platformlarında içerik kalitesini artırmak ve kullanıcı güvenliğini sağlamak için tatbik edilebilir. Özellikle Türkçe konuşulan platformlarda, yerli dil özelliklerine duyarlı bir moderasyon sistemi sağlayarak, hem kullanıcı deneyimini iyileştirir hem de platform yöneticilerine kıymetli bir araç sunar.