E-Posta Spam Filtreleme

Gizem Aygün

Özet E-posta spam filtreleme projesinde, spam ve non-spam (istenmeyen ve istenen) e-posta mesajlarını sınıflandıran bir model geliştirilmiştir. Proje kapsamında 4000 mesajdan oluşan bir veri seti oluşturulmuş, bunlar 2000 spam ve 2000 non-spam e-posta olarak iki gruba ayrılmıştır. Veri seti, üç farklı kaynaktan toplanmıştır: Türkçe e-postalar oluşturulmuş ve ingilizceye çevrilmiştir, sentetik ingilizce veriler Hugging Face GPT2 kullanılarak oluşturulmuş ve bu birleşimden oluşan karma bir veri seti elde edilmiştir. Proje, spam tespiti için farklı veri temsillerinin (TF-IDF, Bag of Words, Word2Vec ve LSTM) ve sınıflandırma algoritmalarının (Naive Bayes, SVM, Random Forest, KNN ve LSTM) performanslarını karşılaştırmalı olarak incelemektedir. Her bir model, doğruluk, kesinlik, duyarlılık ve F1 skoru gibi önemli metrikler üzerinden değerlendirilenerek, hangi yöntemlerin daha yüksek başarı sağladığı belirlenmiştir. Ayrıca, sentetik veri setlerinin gerçekçi sonuçlar verme potansiyeli de analiz edilmiştir. Bu analizler, e-posta spam filtreleme alanında en etkili yöntemleri ortaya koyarak, gelecekteki uygulamalara katkı sağlamayı hedeflemektedir.

Anahtar Kelimeler: Spam filtreleme · Veri temsilleri · TF-IDF · LSTM · Naive Bayes · GPT-2 · Sentetik veri · Doğruluk · F1 skoru.

1. Giriş

E-posta spam filtreleme, dijital iletişimde önemli bir sorunu çözmeye yönelik geliştirilen sistemlerin temel bileşenlerinden biridir. Spam e-postalar, kullanıcıların istenmeyen mesajlarla karşılaşmasına ve verimli iletişimin engellenmesine yol açar. Bu nedenle, spam tespiti, e-posta hizmet sağlayıcıları ve kullanıcılar için büyük bir öneme sahiptir. Spam filtreleme sistemleri, makine öğrenimi ve doğal dil işleme tekniklerini kullanarak, e-posta mesajlarını spam ve non-spam (istenmeyen ve istenen) olarak sınıflandırır ve böylece kullanıcının gelen kutusuna yalnızca önemli mesajları iletir.

Bu projede, spam ve non-spam e-posta mesajlarını doğru bir şekilde sınıflandırabilen bir model geliştirilmesi amaçlanmıştır. Veri seti, 4000 mesajdan oluşmuş olup, bunlar 2000 spam ve 2000 non-spam mesaj olarak iki gruba ayrılmıştır. Veri kaynakları üç farklı biçimde toplanmıştır: Türkçe e-postalar oluşturulup İngilizceye çevrilmiş, GPT-2 tabanlı bir dil modeliyle sentetik İngilizce veriler üretilmiş ve ardından bu iki veri seti birleşerek karma bir veri seti elde edilmiştir. Bu veri seti, hem gerçek verilerin hem de sentetik verilerin birleşimini içermesi bakımından, daha geniş bir spam tespiti yelpazesinde test edilmiştir.

Projede, farklı veri temsillerinin (TF-IDF, Bag of Words, Word2Vec ve LSTM) ve sınıflandırma algoritmalarının (Naive Bayes, SVM, Random Forest, KNN ve LSTM) performansları karşılaştırılmıştır. Her bir model, doğruluk, kesinlik, duyarlılık ve F1 skoru gibi metriklerle değerlendirilmiş ve hangi yöntemlerin en yüksek başarıyı sağladığı belirlenmiştir. Ayrıca, sentetik veri setlerinin gerçekçi sonuçlar sağlama potansiyeli de incelenmiştir. Bu çalışma, spam filtreleme alanındaki en etkili yöntemlerin ortaya konulmasına ve gelecekteki e-posta güvenlik uygulamalarına katkı sağlamayı amaçlamaktadır.

2. Kullanılan Yöntem

Kullanılan yöntemler bu bölümde açıklanmıştır.

2.1. Veri Seti Hazırlığı

Orijinal veri seti, 2000 mesajdan oluşan bir koleksiyon olup, bunların 1000'i spam, 1000'i ise spam olmayan mesajlardan oluşmaktadır. Bu veri seti, gerçek kullanıcı davranışlarını yansıtacak şekilde oluşturulmuştur ve spam filtreleme algoritmalarının doğruluğunu test etmek amacıyla kullanılmaktadır. Başlangıçta, her iki sınıf (spam ve spam olmayan) Türkçe dilinde yazılmıştır. Türkçe veri seti, ardından dilsel çeşitliliği artırmak ve modelin daha geniş bir dil yelpazesinde çalışabilmesini sağlamak amacıyla İngilizceye

çevrilmiştir. Sentetik veri seti, Hugging Face platformu kullanılarak üretilmiştir. Hugging Face, doğal dil işleme alanında güçlü bir araç olup, mevcut spam mesajlarının özelliklerini öğrenen ve buna benzer sentetik mesajlar üretebilen modelleri barındırmaktadır. Bu sentetik mesajlar, orijinal veri setine benzer özellikler taşıyan, ancak tamamen yeni ve farklı içerikler sunan mesajlar olarak tasarlanmıştır. Sentetik veri seti üretimi, mevcut veri setinin çeşitlendirilmesi ve modelin daha geniş bir yelpazede test edilmesi amacıyla gerçekleştirilmiştir. Son olarak, orijinal veri seti ve sentetik veri seti birleştirilerek, toplamda 4000 mesajdan oluşan karma bir veri seti oluşturulmuştur. Bu birleşik veri seti, hem gerçek dünyadan alınan hem de yapay olarak üretilen verilerin harmanlanmasıyla oluşturulmuş olup, veri setinin çeşitliliğini artırmaktadır. Karma veri seti, modelin daha sağlam ve doğru sonuçlar verebilmesi için her iki veri türünü de içermekte ve böylece spam filtreleme modellerinin daha geniş bir veri yelpazesinde test edilmesini sağlamaktadır. Bu yaklaşım, modelin gerçek dünya senaryolarına daha yakın sonuçlar üretmesine olanak tanırken, aynı zamanda sentetik verinin ne kadar etkili olduğunu da değerlendirme imkânı sunmaktadır.

2.2. Veri Temsili

E-posta mesajları, makine öğrenimi modellerinde kullanılabilmesi amacıyla sayısal verilere dönüştürülmüştür. İlk olarak, TF-IDF (Term Frequency-Inverse Document Frequency) yöntemi kullanılmıştır. Bu yöntem, her bir kelimenin önemini hesaba katarak, kelimelerin metindeki sıklığının ve belgedeki yaygınlığının tespit edilmesini sağlar. Bu yöntemde, kelimelerin önem düzeyleri belirlenmiş ve veri seti, 5000 özellik ile sınırlanmıştır. Bu özelliklerin her biri, metindeki kelimelerin önem derecelerini yansıtmaktadır.

İkinci olarak, Bag of Words (BoW) yaklaşımı kullanılmıştır. Bu yöntem, kelimelerin frekanslarına dayalı bir temsil sunar ve her bir kelimenin belge içerisindeki tekrar sayısını hesaplar. BoW modelinde, İngilizce stop words (gereksiz kelimeler) filtrelenmiş, böylece sadece anlam taşıyan kelimeler dikkate alınmıştır.

Ayrıca, kelime bağlamını anlamaya yönelik olarak Word2Vec kullanılmıştır. Word2Vec, kelimeleri vektörler halinde temsil ederek, kelimeler arasındaki semantik ilişkileri ortaya koyar. Bu yöntemde, her bir kelimenin vektörü oluşturulmuş ve mesajların temsili, bu vektörlerin ortalaması alınarak sağlanmıştır. Bu sayede, kelimelerin anlamını daha derin bir düzeyde modellemek mümkün olmuştur.

Son olarak, LSTM (Long Short-Term Memory) modelleri için tokenization ve padding işlemleri uygulanmıştır. Tokenization, kelimelerin sayısal dizilere dönüştürülmesi sürecidir. Bu adımda, her kelime bir tamsayıya karşılık gelecek şekilde sayısallaştırılmıştır. Padding işlemi ise, mesajların uzunluklarını eşitlemek amacıyla yapılmış olup, metinlerin aynı uzunlukta olması sağlanmıştır. Bu adımlar, LSTM modelinin daha verimli çalışabilmesi için gerekli ön işleme adımlarıdır.

2.3. Kullanılan Modeller

Çalışmada, çeşitli sınıflandırma ve doğal dil işleme (NLP) modelleri kullanılarak e-posta mesajları sınıflandırılmıştır. İlk olarak, **Naive Bayes** algoritması kullanılmıştır. Naive Bayes, özellikle metin sınıflandırma görevlerinde yaygın olarak kullanılan, temel bir istatistiksel sınıflandırma yöntemidir. Bu model, özellikle TF-IDF gibi kelime temsilleriyle uyumlu çalışır ve her kelimenin sınıflandırma kararına katkısını bağımsız olarak değerlendirir. Bu yaklaşım, özellikle büyük veri setlerinde hızlı ve etkili sonuçlar verebilir.

Destek Vektör Makineleri (SVM), doğrusal ve doğrusal olmayan sınıflandırma görevlerinde etkili bir yöntem olarak öne çıkmaktadır. SVM, veriyi en iyi ayıran hiper düzlemi bulmayı amaçlar ve yüksek boyutlu veri setlerinde bile iyi performans gösterir. Bu özellik, metin verileri gibi çok sayıda özellik içeren veri setlerinde SVM'nin güçlü bir sınıflandırıcı olmasını sağlar. Bir diğer kullanılan model **Rastgele Orman (Random Forest)** algoritmasıdır. Random Forest, birçok karar ağacının birleştirilmesiyle oluşturulan bir ansamble yöntemidir. Bu model, bireysel karar ağaçlarının zayıf performanslarını dengeleyerek yüksek doğruluk oranları elde etmeyi amaçlar. Random Forest, overfitting (aşırı uyum) riskini minimize ederken, veri çeşitliliği sayesinde sağlam ve güvenilir sonuçlar üretir.

K-Nearest Neighbors (KNN), sınıflandırma işlemi için kullanılan basit ancak etkili bir yöntemdir. KNN, yeni bir örneği sınıflandırmak için, o örneğin en yakın komşularının sınıf etiketlerine dayanır. Bu model, özellikle küçük ve orta ölçekli veri setlerinde kullanılabilir ve sınıf ayırma için doğrudan mesafe ölçümlerine dayanır. Son olarak, **LSTM (Long Short-Term Memory)**, derin öğrenme tabanlı bir model olarak, doğal dil işleme alanında sıklıkla tercih edilmektedir. LSTM, uzun süreli bağımlılıkları öğrenebilen bir tür tekrarlayan sinir ağıdır (RNN). Bu model, kelimeler arasındaki bağlamı öğrenerek, metin verisinin anlamını daha derin bir seviyede kavrayabilir. LSTM, özellikle dil modelleme ve metin sınıflandırma gibi görevlerde üstün başarı göstermektedir. Bu derin öğrenme yöntemi, mesajların içerdiği dilsel yapıları anlamak ve sınıflandırma doğruluğunu artırmak için

kullanılmıştır.
Bu modellerin her biri, e-posta mesajlarının spam veya spam olmayan olarak sınıflandırılmasında farklı avantajlar sunmakta olup, genel doğruluk oranlarını artırmak amacıyla çeşitli kombinasyonlarla test edilmiştir.

2.4. Model Eğitimi ve Değerlendirme

Eğitim sürecinde, kullanılan modeller, TF-IDF, Bag of Words (BoW), Word2Vec ve LSTM yöntemleriyle temsil edilen veriler üzerinde ayrı ayrı eğitilmiştir. Her bir temsil yöntemi, metin verilerinin farklı yönlerini modellemek amacıyla seçilmiş olup, her modelin performansı, bu temsillere dayalı olarak değerlendirilmiştir. TF-IDF, kelimelerin belgelerdeki önemini vurgularken, BoW kelimelerin frekanslarına dayalı bir temsilde bulunur. Word2Vec, kelimeler arasındaki semantik ilişkileri yakalamaya odaklanırken, LSTM derin öğrenme tabanlı bir yöntem olarak metnin bağlamını anlamada üstün başarı göstermektedir. Her model, ilgili temsil yöntemleriyle eğitilerek, e-posta mesajlarını doğru şekilde sınıflandırmak için optimize edilmiştir.

Modellerin performansı, çeşitli değerlendirme metrikleri kullanılarak ölçülmüştür. Doğruluk (Accuracy), modelin tüm tahminlerinin doğruluğunu yansıtan temel bir metrik olup, doğru sınıflandırılan örneklerin tüm örneklere oranını ifade eder. Kesinlik (Precision), modelin pozitif sınıfa ait olarak tahmin ettiği örneklerin ne kadarının gerçekten doğru olduğunu ölçer ve yanlış pozitiflerin oranını azaltma başarısını yansıtır. Duyarlılık (Recall) ise, modelin gerçek pozitif örnekleri doğru şekilde tanımlama başarısını ifade eder ve yanlış negatiflerin oranını azaltmaya odaklanır. Son olarak, F1 Skoru, kesinlik ve duyarlılığın bir dengesini sağlayarak, her iki metriğin birleşik performansını sunar. F1 skoru, özellikle dengesiz veri setlerinde her iki metrik arasında bir denge kurarak modelin genel başarısını değerlendirmek için kullanılır. Bu metriklerin her biri, farklı model performanslarını karşılaştırmada önemli bir rol oynamaktadır.

3. Filtreleme Performansı

Bu bölümde oluşturulan veri setlerinin (Türkçe veri seti, İngilizce veri seti, Sentetik veri Seti, Karma veri seti) model performansına etkisi anlatılmaktadır.

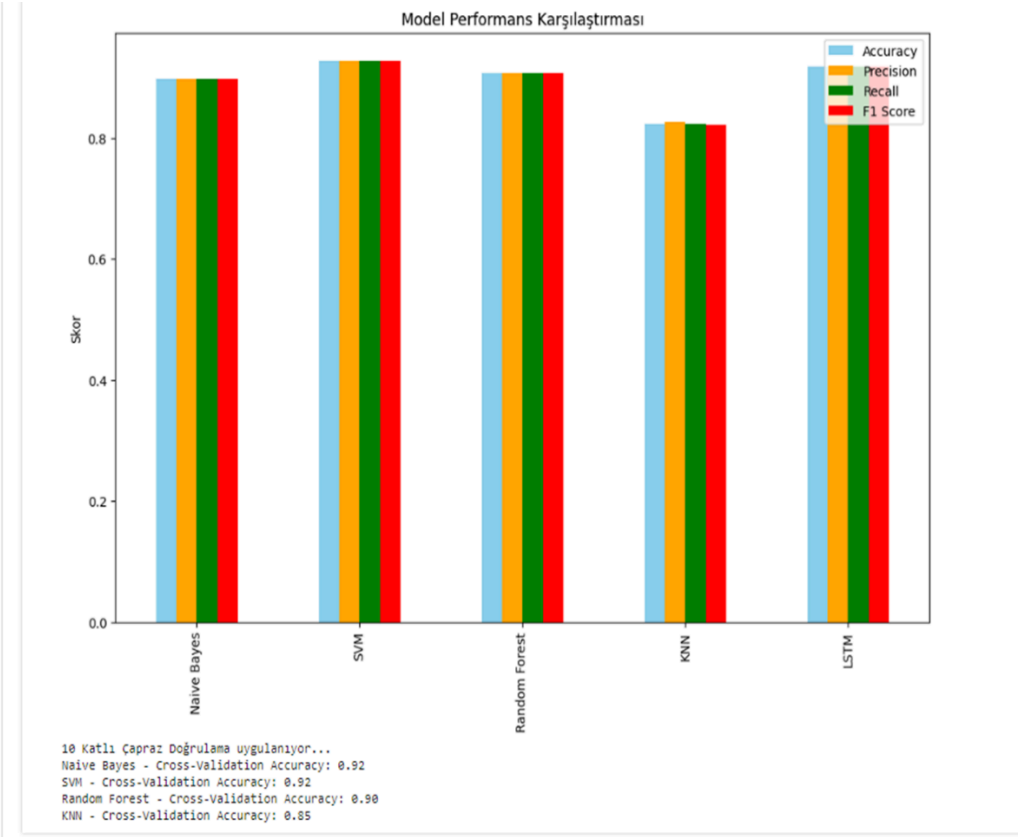
3.1. Türkçe veri seti

Deneysel sonuçlar incelendiğinde, Naive Bayes modeli, doğruluk, hassasiyet, geri çağırma ve F1 skoru gibi metriklerde %91 oranında bir performans sergilemiştir. Model, veri setinde spam ve non-spam sınıflarını başarılı bir şekilde ayırt etmesine rağmen, doğruluk oranı diğer modellere kıyasla daha düşük kalmıştır. Support Vector Machine (SVM) modeli ise %93 doğruluk oranı ile öne çıkmıştır. SVM, özellikle spam e-posta sınıflandırmasında dengeli sonuçlar vererek en iyi performans gösteren model olarak dikkat çekmiştir.

Random Forest modeli, %90 doğruluk oranı ile başarılı sonuçlar elde etmiştir, ancak bu modelin doğruluk oranı SVM'nin gerisinde kalmıştır. Random Forest'in avantajı, çoğu zaman daha iyi genelleme yapabilmesidir. K-Nearest Neighbors (KNN) modeli ise, doğruluk oranı %82 ile diğer modellere kıyasla daha düşük kalmış ve modelin sınıflandırma yaparken daha fazla hata yaptığı gözlemlenmiştir. Bu durum, KNN'nin spam e-posta sınıflandırmasında diğer modellere kıyasla daha az verimli olduğunu göstermektedir.

LSTM (Long Short-Term Memory) modeli, daha karmaşık bir yapıya sahip olmasına rağmen, %93 doğruluk oranı ile oldukça başarılı sonuçlar elde etmiştir. Model, eğitim sürecinde doğruluğun hızla arttığı gözlemlenmiş ve son aşamada %93 doğruluk oranına ulaşmıştır. LSTM, doğal dil işleme (NLP) için güçlü bir model olup, uzun bağıntıları öğrenme konusunda önemli bir avantaja sahiptir.

Son olarak, 10 katlı çapraz doğrulama uygulandığında, Naive Bayes ve SVM modelleri yine en yüksek doğruluk oranlarını (%93) elde etmiştir. Random Forest %90 doğruluk oranı ile takip ederken, KNN modeli %82 doğruluk oranı ile diğer modellere kıyasla daha düşük bir performans sergilemiştir. Çapraz doğrulama, modellerin genel performansını daha tutarlı bir şekilde değerlendirmek için kullanılmış ve sonuçlar, SVM ve Naive Bayes modellerinin en iyi performansı sunduğunu göstermiştir.



Grafik 1. Türkçe veri seti model performans karşılaştırması

3.2. İngilizce veri seti

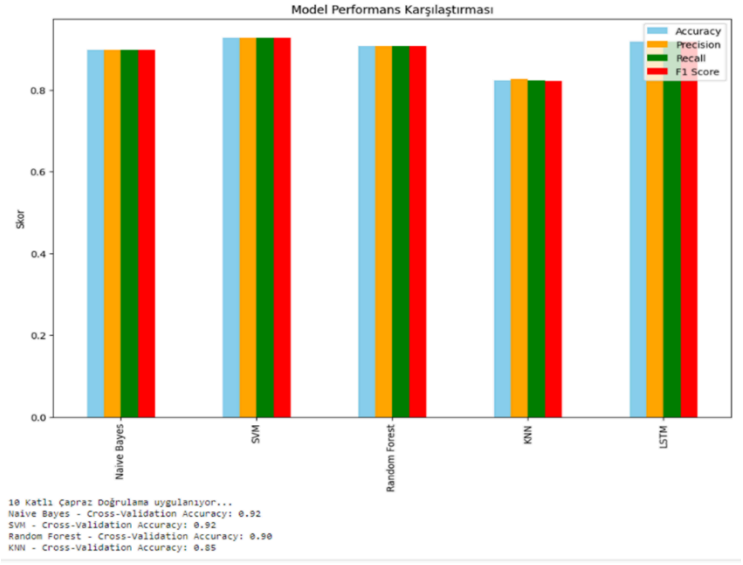
Çalışma kapsamında uygulanan modellerin performansları detaylı bir şekilde değerlendirildiğinde, en yüksek doğruluk oranı %93 ile Support Vector Machine (SVM) modelinde elde edilmiştir. SVM, verinin özellik uzayındaki ayrışabilirliğini başarıyla öğrenerek, sınıflandırma görevinde üstün bir performans sergilemiştir. Naive Bayes ve Random Forest modelleri ise %91 doğruluk oranı ile dengeli ve başarılı sonuçlar üretmiştir. Her iki model de hem spam hem de non-spam sınıflarını ayırt etme konusunda tatmin edici bir performans sergileyerek, veri setinin özelliklerini iyi bir şekilde öğrenebilmiştir. Random Forest, bir ensembıl yöntemi olarak farklı karar ağaçlarının birleşimiyle güçlü ve kararlı bir performans ortaya koymuş, ancak doğruluk açısından SVM'nin gerisinde kalmıştır.

K-Nearest Neighbors (KNN) modeli, %82 doğruluk oranıyla diğer modellere kıyasla daha düşük bir performans göstermiştir. KNN'nin, yüksek boyutlu verilerle başa çıkmada yaşadığı zayıflık bu düşük doğruluk oranına işaret etmektedir. Bu model, basit ve

açıklanabilir bir algoritma olmasına karşın, karmaşık sınıflandırma görevlerinde daha sınırlı kalmaktadır.

Long Short-Term Memory (LSTM) modeli, derin öğrenme tabanlı bir yaklaşım olarak, metin verilerinde güçlü bir performans sergilemiştir. %92 doğruluk oranına ulaşan LSTM, özellikle ardışık verilerde ve zaman serisi analizlerinde etkin olmasının yanı sıra, doğal dil işleme (NLP) görevlerinde de başarılı sonuçlar üretmiştir. Ancak, LSTM'nin daha yüksek hesaplama maliyetleri ve kaynak gereksinimleri dikkate alındığında, pratikte daha fazla işlem gücü gerektiren bir model olduğu söylenebilir.

Sonuç olarak, SVM ve LSTM modelleri metin sınıflandırma görevlerinde öne çıkmış ve en iyi performansı sergileyen modeller olarak değerlendirilmiştir. SVM'nin doğruluk oranı, verinin ayrışabilirliğini öğrenme konusunda güçlü bir yetenek gösterdiği için en yüksek başarıyı elde etmesini sağlamıştır. LSTM ise derin öğrenme tabanlı yaklaşımların güçlü yanlarını ortaya koyarak metin verilerinde yüksek doğruluk elde etmiştir. Naive Bayes ve Random Forest ise dengeli performanslarıyla iyi sonuçlar vermiştir. KNN modeli ise daha düşük bir doğruluk oranıyla diğer modellere kıyasla geride kalmıştır.



Grafik 2. İngilizce veri seti model performans karşılaştırması

3.3. Sentetik veri seti

Naive Bayes (NB) Modeli, metin sınıflandırma problemlerinde yaygın olarak kullanılan ve genellikle başarılı sonuçlar veren bir modeldir. Bu çalışma kapsamında, %91 doğruluk

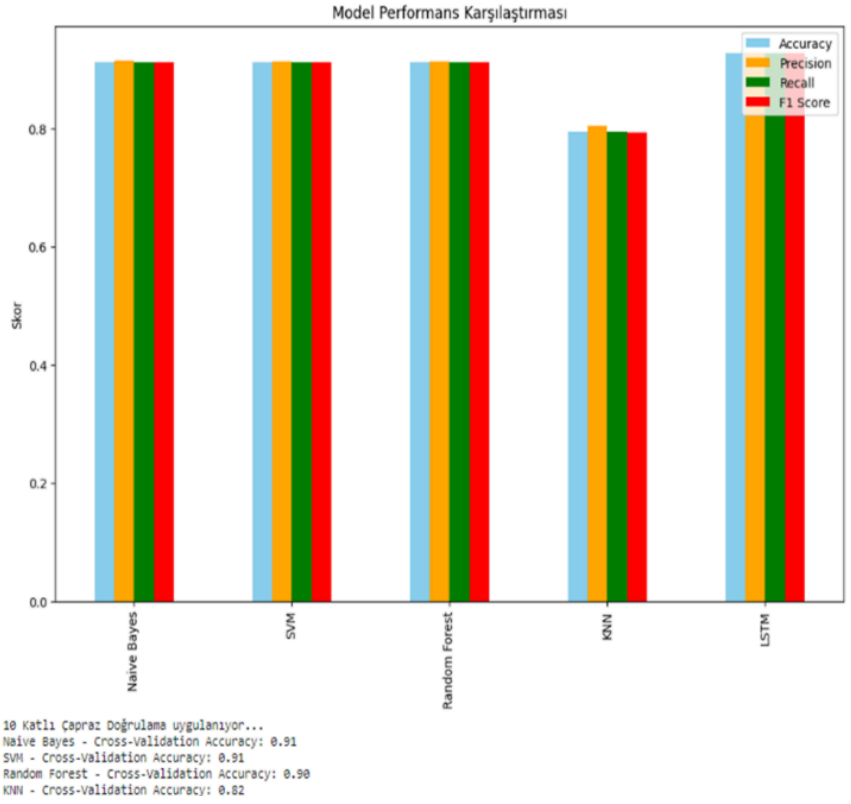
oranı elde eden Naive Bayes, spam ve ham mesajları ayırt etme görevinde oldukça dengeli bir performans sergilemiştir. Özellikle kesinlik (Precision) değeri %92, modelin spam mesajları doğru bir şekilde tahmin etme yeteneğini ortaya koymaktadır. Random Forest (RF) Modeli ise, benzer şekilde %91 doğruluk oranı elde etmiş, ancak çapraz doğrulama sonuçlarında Naive Bayes ile karşılaştırıldığında biraz daha düşük bir performans göstermiştir. Bu durum, Random Forest modelinin kararlılığının diğer modellere göre daha düşük olduğunu işaret etmektedir.

K-En Yakın Komşu (KNN) Modeli, doğruluk oranı bakımından diğer modellere kıyasla daha düşük bir performans sergilemiştir. Modelin doğruluk oranı %80 iken, F1 skoru ise %79'a gerilemiştir. Bu düşük performans, özellikle büyük veri setlerinde KNN'nin sınıflandırma görevlerinde yetersiz kaldığını ve hiperparametre optimizasyonunun önem taşıdığını göstermektedir. Destek Vektör Makineleri (SVM) Modeli ise, Naive Bayes ile benzer sonuçlar elde ederek %91 doğruluk oranı ile başarılı bir performans sergilemiştir. SVM'nin sınıflandırma başarısının yüksek doğruluk ve kararlılıkla sonuçlanması, modelin metin verisinin özellik uzayındaki ayrışabilirliğini öğrenme yeteneğinden kaynaklanmaktadır.

Long Short-Term Memory (LSTM) Modeli, özellikle zaman serisi verileri ve ardışık veri türlerinde güçlü performans gösteren derin öğrenme tabanlı bir yaklaşımdır. Bu model, projede %93 doğruluk oranına ulaşmış, eğitim sürecinde doğruluk hızla artmıştır. Eğitim doğruluğu ilk epoch'da %55 iken, ikinci epoch'da %88'e, üçüncü epoch'da ise %98'e çıkmıştır. Test seti sonuçlarında da doğruluk, kesinlik, duyarlılık ve F1 skoru %93 olarak ölçülmüştür. LSTM modelinin sürekli iyileşen doğruluk ve kayıp değerleri, metin verisini derinlemesine öğrenme yeteneğini doğrulamaktadır. Bu sonuç, LSTM'nin metin sınıflandırma görevlerinde diğer makine öğrenimi modellerine kıyasla daha yüksek doğruluk sağladığını ortaya koymaktadır.

Çapraz doğrulama sonuçları, modellerin genel performansını tutarlı bir şekilde değerlendirmeye olanak sağlamıştır. Naive Bayes ve SVM, her ikisi de %91 doğruluk oranı ile en kararlı performansı göstermiştir. Random Forest, %1'lik bir düşüşle orta seviyede kararlılığa sahipken, KNN modeli ise %82 doğruluk oranı ile daha düşük bir kararlılık göstermiştir. Bu, KNN'nin büyük veri setlerinde daha düşük performans sergileyebileceğini ve hiperparametre optimizasyonuna ihtiyaç duyduğunu göstermektedir. Genel olarak, SVM ve LSTM modelleri, metin sınıflandırma görevlerinde daha yüksek doğruluk ve kararlılık

sağlamış, bu modellerin özellikle doğal dil işleme ve metin verisi analizi için daha uygun olduğunu ortaya koymuştur.

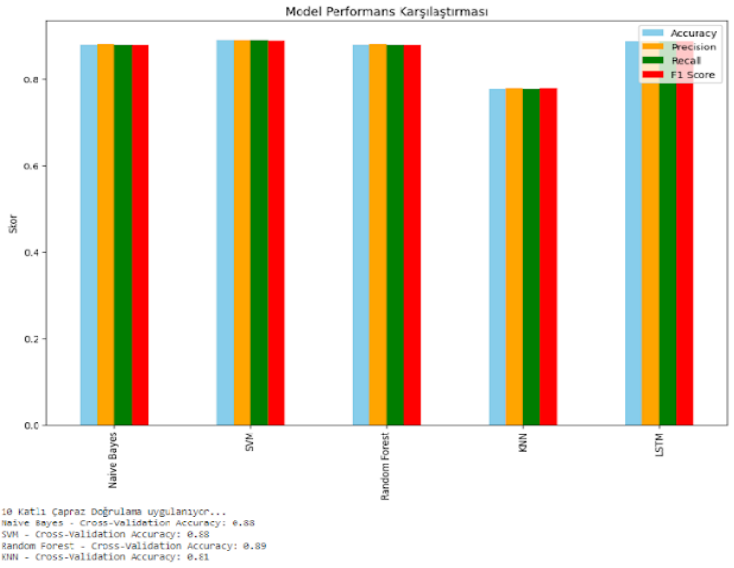


Grafik 3. Sentetik veri seti model performans karşılaştırması

3.4. Karma veri seti

Naive Bayes modeli, %88 doğruluk oranı ile başarılı bir performans sergilemiştir. Hem kesinlik hem de duyarlılık değerleri %88 seviyesindedir. Bu, modelin hem spam mesajları doğru tahmin etme oranının yüksek olduğunu hem de gerçek pozitif örnekleri doğru bir şekilde yakaladığını göstermektedir. Çapraz doğrulama sonuçlarında da %88 doğruluk oranı elde edilmiştir, bu da modelin genel kararlılığını doğrulamaktadır. Naive Bayes'in genellikle metin sınıflandırma görevlerinde hızlı ve etkili bir yöntem olarak kullanıldığını söylemek mümkündür. SVM modeli, %89 doğruluk ile bir miktar daha iyi performans sergilemiştir. Kesinlik, duyarlılık ve F1 skoru da %89 ile uyumlu bir şekilde yüksek çıkmıştır. SVM'nin yüksek doğruluk oranı, modelin veri kümesinin özellik uzayında ayrışabilirliğini öğrenme yeteneğinden kaynaklanmaktadır. Çapraz doğrulama sonuçlarında da %88 doğruluk oranı ile benzer bir performans elde edilmiştir, bu da modelin genelleme yeteneğini göstermektedir.

Random Forest modeli, %88 doğruluk ile Naive Bayes ile benzer bir performans sergilemiştir. Modelin kesinlik, duyarlılık ve F1 skoru da %88 civarındadır. Çapraz doğrulama sonuçları, %89 doğruluk ile küçük bir iyileşme göstermektedir, ancak çapraz doğrulama sırasında modelin kararlılığı biraz daha yüksek olmuştur. Random Forest, ensembıl yöntemleri kullanarak farklı karar ağaçlarının çıktılarının birleştirilmesiyle güçlü bir sınıflandırma yeteneği sağlar. Bu, modelin daha sağlam ve genelleştirilebilir sonuçlar ürettiğini göstermektedir. KNN modeli, %78 doğruluk ile diğer modellere kıyasla daha düşük performans göstermiştir. Kesinlik, duyarlılık ve F1 skoru da %78 ve %79 civarındadır. Bu sonuçlar, KNN'nin özellikle büyük veri setlerinde daha düşük performans gösterebileceğini ve hiperparametre optimizasyonunun gerekliliğini ortaya koymaktadır. Çapraz doğrulama sonuçları da %81 doğruluk ile daha düşük kararlılığı işaret etmektedir. KNN'nin, yüksek boyutlu veri setlerinde verimli çalışabilmesi için parametre ayarlarının dikkatle yapılması gerekmektedir. LSTM modeli, derin öğrenme tabanlı bir yöntem olarak özellikle ardışık ve zaman serisi verileri işlemek için güçlüdür. Modelin eğitim sürecinde doğruluk hızla artmış ve son epoch'ta %98'e kadar yükselmiştir. Test seti sonuçlarında doğruluk %89, kesinlik %89, duyarlılık %89 ve F1 skoru da %89 olarak elde edilmiştir. LSTM modelinin eğitim sürecindeki sürekli iyileşme, modelin derin öğrenme yöntemleriyle metin verisini anlamada güçlü bir kapasiteye sahip olduğunu göstermektedir. Ancak, çapraz doğrulama sırasında %89 doğruluk oranı ile benzer bir performans elde edilmiştir, bu da modelin kararlılığını desteklemektedir. Modellerin genel performansı incelendiğinde, SVM en yüksek doğruluk oranına (%89) sahip model olarak öne çıkmaktadır. LSTM de derin öğrenme tabanlı bir model olarak güçlü performans sergilemiş, ancak daha fazla hesaplama gücü ve eğitim süresi gerektirmiştir. Naive Bayes ve Random Forest modelleri de benzer doğruluk oranlarına (%88) ulaşmış, ancak LSTM ve SVM'ye göre biraz daha düşük performans göstermiştir. KNN modeli ise %78 doğrulukla diğer modellere kıyasla belirgin bir şekilde daha düşük bir başarıya ulaşmıştır. Çapraz doğrulama sonuçları, modellerin genelleme performanslarını ölçmek açısından önemlidir; burada SVM ve Naive Bayes modelleri, kararlılık açısından en iyi sonuçları verirken, KNN'nin kararlılığı düşük kalmıştır. SVM ve LSTM modelleri, metin sınıflandırma görevlerinde daha yüksek doğruluk ve kararlılık sağlarken, KNN modeli yüksek boyutlu veriyle başa çıkmada zorluklar yaşamakta ve hiperparametre ayarlamaları gerekmektedir. Bu bulgular, farklı modellerin metin sınıflandırma gibi görevlerdeki potansiyelini ve sınırlamalarını anlamak için önemli bir temel oluşturmaktadır.



Grafik 4. Karma veri seti model performans karşılaştırması

4. Sonuç

Bu çalışmada, dört farklı makine öğrenimi modeli (Naive Bayes, SVM, Random Forest, KNN) ve bir derin öğrenme modeli (LSTM), Türkçe, İngilizce, Sentetik ve Karma veri setleri üzerinde değerlendirilmiştir. Sonuçlar, her bir modelin doğruluk, kesinlik, duyarlılık ve F1 skoru gibi performans metrikleri üzerinden karşılaştırılmıştır.

Model	Çapraz Doğruluk (Accuracy)
Naive Bayes	%93
SVM	%93
Random Forest	%90
kNN	%82

Tablo 1. Türkçe veri seti 10 kat çapraz doğrulama sonuçları

Model	Çapraz Doğruluk (Accuracy)
Naive Bayes	%92
SVM	%92
Random Forest	%90
kNN	%85

Tablo 2. İngilizce veri seti 10 kat çapraz doğrulama sonuçları

Model	Çapraz Doğruluk (Accuracy)
Naive Bayes	%91
SVM	%91
Random Forest	%90
kNN	%82

Tablo 3. Sentetik veri seti 10 kat çapraz doğrulama sonuçları

Model	Çapraz Doğruluk (Accuracy)
Naive Bayes	%88
SVM	%88
Random Forest	%89
kNN	%81

Tablo 4. Karma veri seti 10 kat çapraz doğrulama sonuçları

4.1. SVM ve Random Forest

Bu iki model, çoğu veri setinde benzer şekilde yüksek doğruluk oranlarına ulaşmıştır (%89-93). Özellikle SVM, veri setlerinin farklı dil yapılarında ve özelliklerinde tutarlı bir performans göstermiştir. Random Forest ise özellikle veri setlerinde dengeli doğruluk ve hassasiyet oranları ile dikkat çekmiştir.

4.2. Naive Bayes

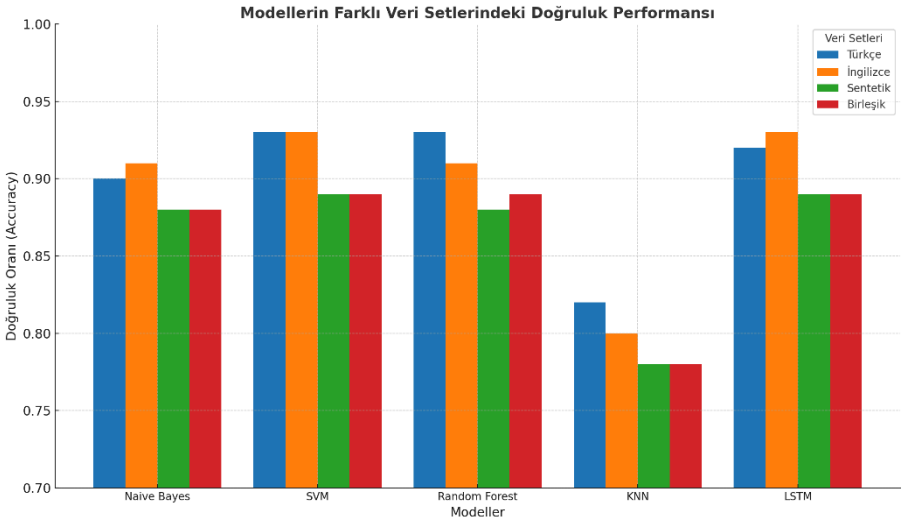
Naive Bayes, dilin temel özelliklerini yakalama konusundaki başarısı ve hızlı çalışması ile çoğu durumda etkili bir alternatif olarak öne çıkmıştır (%88-91 doğruluk). Ancak bağlamsal bilgiyi derinlemesine işlemede derin öğrenme modellerine kıyasla sınırlı kalmaktadır.

4.3. LSTM

LSTM modelleri, dilsel bağlamın öğrenilmesinde özellikle başarılıdır. Türkçe, İngilizce ve birleşik veri setlerinde %89-93 doğruluk oranlarına ulaşmıştır. Özellikle çok dilli veri setlerinde etkili bağlamsal çıkarımlar yaparak makine öğrenimi modellerine kıyasla üstünlük sağlamıştır. Bununla birlikte, sentetik veri setinde doğrulama kaybı, modelin genelleme yeteneğini geliştirme ihtiyacını ortaya koymaktadır.

4.4. KNN

KNN modeli tüm veri setlerinde en düşük doğruluk oranlarını sunarak (%78-82) sınırlı bir performans sergilemiştir. Özellikle veri setinin boyutu ve karmaşıklığı, KNN'nin düşük sonuçlar vermesine neden olmuştur.



Grafik 5. Farklı veri setlerinin performans karşılaştırması

Bu çalışma farklı modellerin veri setlerinin dil yapısı ve özelliklerine göre değişen performanslar sunduğunu ortaya koymaktadır. SVM, Random Forest ve LSTM modelleri, genellikle en yüksek doğruluk oranlarına ulaşarak diğer modellere kıyasla daha etkili sınıflandırıcılar olarak öne çıkmıştır. Özellikle LSTM modeli, hem Türkçe hem İngilizce veri setlerinde bağlamsal bilgiyi işleme yeteneği ile dikkat çekmiştir. KNN modeli ise genelde düşük doğruluk oranları ile bu çalışma kapsamında en az etkili yöntem olarak değerlendirilmiştir. Naive Bayes modeli ise basitliği ve hızlı çalışması ile diğer modellere kıyasla kabul edilebilir bir performans sergilemiştir. Bu sonuçlar, spam e-posta sınıflandırma problemi gibi karmaşık dilsel yapıları içeren problemlerde, veri setinin yapısına uygun model seçiminin ve dil özelliklerinin iyi anlaşılmasının önemini vurgulamaktadır. Ayrıca, derin öğrenme yaklaşımlarının, özellikle dilsel bağlamı öğrenmede makine öğrenimi modellerine göre daha etkili olduğu görülmüştür. Bu bağlamda, LSTM modeli özellikle çok dilli veri setlerinde ileri düzey performans sergileyerek dikkat çekici bir potansiyel sunmaktadır.

Sonuç olarak, bu çalışmada elde edilen bulgular, hem akademik çalışmalar hem de endüstriyel uygulamalar için rehberlik edebilecek niteliktedir. Spam sınıflandırma sistemlerinin dilsel karmaşıklığı ele almadaki başarısı, doğal dil işleme teknolojilerinin potansiyelini bir kez daha ortaya koymaktadır.

5. Öneriler

Gelecekteki çalışmalarda aşağıdaki kriterlerin uygulanması önerilmektedir.

1. LSTM modellerinin genelleme yeteneğini artırmak için daha geniş bir hiperparametre optimizasyonu yapılabilir.
2. Çok dilli veri setlerinde bağlamı daha iyi işleyebilecek hibrit modeller geliştirilebilir.
3. Sentetik veri setlerinde performansı artırmak için daha sofistike veri üretim yöntemleri uygulanabilir.

6. Referanslar

1.Google Colab: Python kodu yazıp çalıştırmak için interaktif bir not defteri ortamı sunan, makine öğrenmesi projeleri için bulut kaynaklarını kullanmayı sağlayan bir platformdur. https://colab.research.google.com/.

2.Hugging Face: Doğal dil işleme görevleri için makine öğrenmesi modellerinin paylaşılabildiği ve dağıtılabildiği bir platformdur. Sınıflandırma, metin üretimi ve daha fazlası için önceden eğitilmiş modellere erişim sunar. https://huggingface.co/.

3.Kaggle: Veri bilimi yarışmalarına ev sahipliği yapan ve makine öğrenmesi görevleri için veri setleri ve not defterleri sunan bir platformdur. Aynı zamanda büyük bir veri bilimi topluluğu bulunmaktadır. https://www.kaggle.com/.

4.ChatGPT (OpenAI): Yapay zeka tabanlı dil modelidir ve çeşitli konularda metin oluşturma, soruları yanıtlama ve önerilerde bulunma gibi görevlerde yardımcı olabilir. https://chat.openai.com/.

5.Veri Bilimi Türkiye (VeriBilimi.org): Türkiye'de veri bilimi, makine öğrenmesi ve yapay zeka üzerine bilgiler ve dersler sunan bir platformdur. https://veribilimi.org.

6. Dijital Akademi: Veri bilimi, yapay zeka ve makine öğrenmesi konularında Türkçe eğitimler sunan bir kaynaktır. https://dijitalakademi.com.

7. Medium - Veri Bilimi Türkiye (Makine Öğrenmesi Makaleleri): Veri bilimi ve makine öğrenmesi üzerine Türkçe yazılar ve rehberler içeren bir platformdur. https://medium.com/@veribilimiturkiye.