SPAM MAİL SINIFLANDIRMASI

Giriş

Spam analizi, e-posta iletişiminde kullanıcıları istenmeyen e-postalardan korumak adına oldukça kritik bir konudur.

Bu bağlamda, spam ve spam olmayan e-postaların sınıflandırılmasını içeren proje anlatılacaktır.

Problemin Tanımı

- E-Posta günümüzde çok yaygın olarak kullanılmaktadır
- Neredeyse her internet kullanıcısının bir e-posta hesabı vardır.
- E-posta kullanımı yaygınlaşması bazı problemleri de beraberinde getirmiştir.
- Bu sorunların en önemlilerinden bir tanesi istenmeyen (spam) e-postalardır.

Problemin Tanımı

- Şirketler için bu şekilde reklam yapmak daha az maliyetli olduğundan dolayı sık sık bu şekilde e-postalarla kullanıcıları rahatsız etmektedir.

- Reklamdan farklı olarak, kötü niyetli kişiler internet kullanıcılarının hassas bilgilerini ele geçirmek amaçlı spam e-postalar göndermektedir.

Proje Hakkında

- Spam e-postaların ve kısa mesajların tespit edilerek kullanıcılara sunulmadan engellenmesi önemlidir.

- Bu nedenlerden dolayı seçilen projede, Kaggle platformundan elde edilen bir e-posta veri kümesi üzerinde çalışılarak spam ve spam olmayan e-postaların sınıflandırılması amaçlanmaktadır.

Veri Setinin İncelenmesi

- Çalışmada, Kaggle platformunda erişime açık, dili İngilizce olan bir veri seti kullanılmıştır.
- Kullanılan veri setinde 4825 adet normal ve 747 adet spam elektronik posta örneği bulunmaktadır.

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will ü b going to esplanade fr home?
5569	ham	Pity, * was in mood for that. So...any other s...
5570	ham	The guy did some bitching but I acted like i'd...
5571	ham	Rofl. Its true to its name

[5572 rows x 2 columns]

Veri Setinin İncelenmesi

- Veri seti iki ana sütundan oluşmaktadır.
- "Category" sütunu e-postaların kategorisini belirtir. Bu kategoriler "ham" ve "spam" olmak üzere 2 çeşittir.
- İkinci sütun olan "Message" ise e-posta mesajlarının içeriğini saklar.

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will ü b going to esplanade fr home?
5569	ham	Pity, * was in mood for that. So...any other s...
5570	ham	The guy did some bitching but I acted like i'd...
5571	ham	Rofl. Its true to its name

[5572 rows x 2 columns]

Ön İşleme Adımı

Spam tespitinde başarılı bir model oluşturmanın ilk ve temel adımı elde bulunan verilerin ön işlemden geçirilmesidir.

Ön işlem bünyesinde, veri setini oluşturan tüm elektronik posta kayıtlarında aşağıdaki işlemler gerçekleştirilmiştir

- sayısal ifadelerin,özel karakterlerin kaldırılması
- Büyük harflerin küçük harfe dönüştürülmesi
- Etkisiz kelimelerin (Stopwords) çıkarılması
- Kelime köklerinin bulunması
- Label Encoding

1.Sayısal İfadelerin Kaldırılması

- Veri setindeki metinlerde bulunan sayısal ifadeler, genellikle spam analizi gibi metin madenciliği görevleri için önemli değildir.
- Bu nedenle, bu ifadeleri kaldırmak, modelin metni daha iyi anlamasına yardımcı olabilir.
- Bu amaçla sayısal ifadeler silinerek kelimeler üzerinde sınıflandırma işlemleri yapılmıştır.

2.Özel Karakterlerin Kaldırılması

- Metin verilerindeki özel karakterler, modelin kelime veya cümle yapılarını anlamasını zorlaştırabilir .
- Bu amaçla ‘@’, ‘|’, ‘<’, ‘>’, ‘#’, ‘$’, ‘%’ gibi özel karakterler çıkarılmıştır.

3.Büyük Harflerin Küçük Harfe Dönüştürülmesi

- Anlam bütünlüğünü sağlayabilmek ve aynı kelimenin farklı bir kelime gibi işlem yapılmasını engelleyebilmek amacıyla büyük-küçük harf uyumu sağlanmıştır.
- Büyük harfler küçük harflere dönüştürülerek standartlaştırma yapılmıştır.

4.Etkisiz Kelimelerin (Stopwords) Çıkarılması

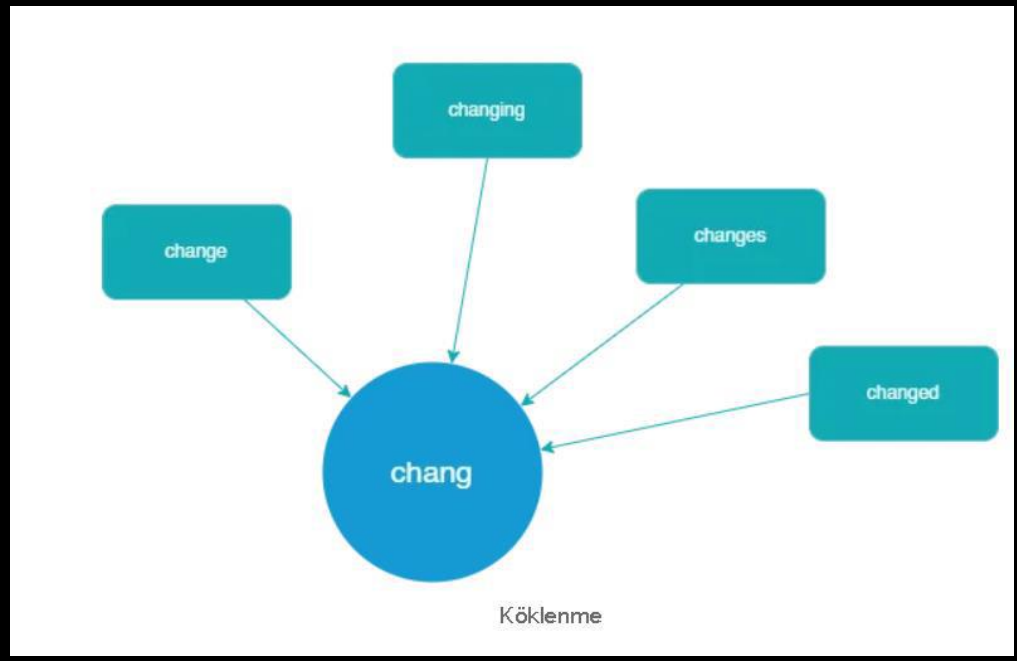
- Dil içerisinde sıkça kullanılan ancak genellikle anlam taşımayan kelimeler (stopwords), hem ham hem de spam maillerde yaygın olarak kullanıldığından ham ve spam ayrımı yaparken etkisi olmamaktadır.
- Bu tür kelimelerin kaldırılması, modelin daha önemli bilgileri öğrenmeye odaklanmasını sağlar.
- Metni daha anlamlı kılar. Bu amaçla etkisiz kelimeler çıkarılmıştır.

5.Kelime Köklerinin Bulunması

•Stemming, kelime köklerini bulma işlemidir ve metindeki kelimeleri temel köklerine indirgeme sürecidir.

•Spam analizi gibi görevlerde kelimelerin köklerini kullanmak, anlamı bozmadan metni daha basitleştirebilir.

- Bu ön işlemin amacı e-posta içerisinde geçen kelime sayısını azaltmaktır.



Özellik Çıkarımı (Feature Extraction)

Literatürde spam tespiti çalışmalarında kullanılan öznitelik çıkartma işlemleri incelendiğinde “TF-IDF” yaklaşımının ön plana çıkarıldığı görülmüştür. Bu çalışmada da TF-IDF kullanılmıştır.

TF-IDF, bir kelimenin bir belgedeki sıklığı ile o kelimenin genel belge koleksiyonundaki nadirliğini dikkate alarak bir terimin ağırlığını belirler.

Spam e-postalar genellikle belirli anahtar kelimeleri aşırı kullanma eğilimindedir. TF-IDF, bu tür anahtar kelimeleri belirlemede yardımcı olabilir.

Kullanılan Algoritmalar

- Veri setleri üzerinde yapılan ön işlemlerden sonra problemin algoritmalara uyarlanması gerçekleştirilmiştir.
- Veri seti eğitim ve test verileri olarak ikiye bölünmüştür. Bölme oranları %20 test ve %80 eğitim verisi şeklindedir.
- Probleme K-en yakın komşuluk algoritması ve lojistik regresyon algoritması uygulanmıştır.

1.KNN (K Nearest Neighbours)

KNN algoritmasında tahmin edilecek yeni bir veri noktası geldiğinde, bu noktanın en yakın k adet komşusu bulunur. "En yakın" komşular genellikle öklid mesafe ölçümü kullanılarak belirlenir.

K adet komşu veri noktasının sınıfları incelenir ve bu komşuların ait olduğu sınıflardan hangisi çoğunluğu belirliyorsa , yeni veri noktası bu çoğunluğa atanır.

2.Lojistik Regresyon

Lojistik Regresyon, sonuçları kategorik olarak değerlendiren (evet/hayır, geçti/kaldı vb.) makine öğrenmesi yaklaşımı sunmaktadır ve birden fazla sonuca dayalı sınıflandırma çıktısı üretmektedir.

Bu çalışma kapsamında LR yaklaşımı ikili sınıflandırma amacıyla kullanılmıştır.

Sonucun ham veya spam olarak sınıflandırılması şeklinde ele alınmıştır.

Modellerin Karşılaştırılması

2 model için de cross validation ve test verilerinin sonuçları değerlendirme metrikleri açısından karşılaştırıldığında sonuçlar hem birbirine yakın hem de yüksektir.

Değerler arasındaki benzerlik, modelin eğitim verisi üzerinde iyi performans sergileyip, bu başarısını bağımsız bir test seti üzerinde de sürdürdüğünü göstermektedir.

Modeller kendi arasında karşılaştırıldığında ise lojistik regresyon modeli knn'den daha iyi performans sergilemektedir.