Haber Başlıkları Kullanılarak Sahte Haber Tespiti

Aleyna Er
Bilgisayar Mühendisliği
Yıldız Teknik Üniversitesi
Istanbul, Turkey
aleyynaer@gmail.com

***Özet*— Yapılan çalışmada, haber başlıkları analiz edilerek haberler gerçek yahut sahte olmalarına göre sınıflandırılmıştır. Haber başlıkları gerçek ve sahte olmalarına göre iki farklı dosyada bulunmaktadır ve dosyadaki her bir satır bir haber manşetine denk gelmektedir. Bu dosyalardan haber başlıkları çekilerek nltk kütüphanesiyle metindeki her kelime token'laştırılmış, stop word'ler temizlenmiştir, kalan kelimeler stemming ve lemmatizing işlemlerine tabi tutulmuştur. Bu işlemler sonucu cümleler kelimelerine ayrıştırılmış, tek başına anlamı olmayan (edat,bağlaç vs) kelimeler veriden silinmiş, kalan kelimeler köklerine ve ilk biçimine çevrilmiştir.**

Çalışmanın ilk aşamasında, veri seti cümle x tokenlar (kelimeler) matrisi ile temsil edilmek istenmiştir. Matrisin içi cümlelerin barındırdığı kelimelerin frekansı ile doldurulmuştur. Matrisin boyutunu azaltmak üzere, her kelime için IDF (inverse document frequency) hesaplanmış ve haberlerde fazlaca ortak kullanılan kelimeer veri setinden silinmiştir. Sonuç olarak 3266 x 3698 'lik haber başlığı x kelime matrisi oluşturulmuştur. Elde edilen matris üzerinde işlem ve modelleme yapmanın vakit olarak dezavantajlı olmasından dolayı bu yöntemden vazgeçilmiştir.

Projeye devam edilen yöntemde, gerçek ve sahte haberler için bu işlemler ayrı ayrı yapılmış ve elde edilen kelimeler iki farklı listede tutulmuştur, bunlar sahteKelime listesi ve gerçekKelime listesi olarak isimlendirilmiştir.

Haber başlıkları, oluştukları kelimelerin sahteKelime ve gerçekKelime listelerinde bulunma sıklığına göre işleme tabi tutulmuştur. Hesaplama sonucu mevcut haberin bulundurduğu sahte ve gerçek kelime sayısı o haberin uzunluğuna bölünerek normalize edilmiş, haber başlığının yüzde kaçının o listeye ait olduğu hesaplanmıştır.

Yapılan hesaplamala ile birlikte, veri seti modellemeye hazırlanmıştır. KNN, Decision Tree ve Naive Bayes algoritmalarıyla tahminleme işlemi gerçekleştirilmiştir. Kullanılan algoritmalar için, veri seti yüzde 70 train, 15 test ve 15 validation olmak üzere üç parçaya bölünmüştür. Algoritmaların parametreleri optimize edilerek ve k-fold cross-validation yöntemi kullanılarak model başarısı yükseltimek istenmiştir.

Keywords— news categorization; classification; machine learning

I. GİRİŞ

Sahte Haber Tespiti projesi ile, toplanan haber manşetlerinin (başlıklarının) analizi yapılarak haberlerin niteliğini (gerçek yahut sahte) tespit etmek amaçlanmıştır. Özellikle internet ortamında, sosyal medyada sahte haberlerin çok hızlı yayılabilmesi ve insanları paniğe sürükleyerek kötü sonuçlara sebep olabilmesi büyük tehlike taşımaktadır. Geliştirilecek proje ile, internet ortamında dolaşan haberlerin manşetleri analiz edilerek, haberin gerçek olup olmadığı hızlıca saptanabilecektir. Geliştirilen model ile, Teyit.org gibi oluşumlar desteklenebilir veya yeni bir oluşum kurulabilir. Ek olarak, sosyal medyada paylaşılan haberlerin analizini yapacak ve gönderinin güvenilirliğini ölçecek ve kullanıcılara gösterecek bir araç geliştirdiğimiz projeden faydalanılarak yazılabilir.

II. VERİ KÜMESİ

Sahte ve gerçek haber başlıkları olmak üzere iki ana txt dosyası bulunmaktadır. Txt dosyalarındaki her satır bir haber başlığına denk gelmektedir. İki dosya da Kaggle platformunda bulunan veri setleridir.
Sahte haber manşetleri, 244 farklı internet sitesinden toplanmıştır ve 1298 adet manşet içerir. Gerçek haber manşetleri ise ABC (Australian Broadcasting Corporation)'den kaynak alınmıştır, toplamda 1968 gerçek haber başlığı bulunmaktadır. Gerçek ve sahte haber oranlarına baktığımızda, iki sınıf için denge söz konusudur, aralarında çok fazla sample örneği bulunmamaktadır, dolayısıyla burada sentetik veri üretilmesine ihtiyaç yoktur.

Gerçek haber başlıklarına şöyle örnek verilebilir;

```
real_headers
```

```
['donald trump do you remember the year since he was elected',
 'trump defends son over emails as moscow hits back',
 'donald trump strategist says media wont easily give back america',
 'anthony scaramucci who is donald trumps new comms director',
 'donald trumps mobile phone use worries security experts',
```

Şekil 1. Gerçek Haber Başlıkları örneği

Sahte haber başlıkları örneği şekil 2'de görülebilir;

```
fake_headers
```

```
['trump warns of vote flipping on machines',
 'this election is not about trump its about a giant middle finger to washington dc',
 'more on trump populism and how it can be controlled by government',
 'trump bollywood ad meant to sway indian american voters is an hilarious fail',
 'dems could be up on charges for inciting trump rally violence',
```

Şekil 2. Sahte Haber Başlıkları örneği

III. SINIFLANDIRMA MODELLERİ

Projede gerçek ve sahte haberleri tahminleyebilmek için sınıflandırma algoritmalarından KNN, Decision Tree ve Naive Bayes kullanılmıştır.

K-Nearest Neighbor (KNN/ En Yakın Komşular) algoritması, yapı itibariyle lazy denilen bir öğrenim algoritmasıdır. Lazy denmesinin nedeni verilerin özelliklerini öğretmekten çok ezberletmesidir. Böylece yeni bir veri girildiğinde veya başka bir veri bütünüyle karşılaştığında, en yakınındaki komşuların özelliklerine bakarak en uygun bulduğu küme içerisine yerleştirir. Buradaki K değeri bakılacak elemanların (sample'ların) toplam sayısıdır. Yeni değer geldiğinde en yakınındaki k adet komşularına bakılarak uzaklık hesabı yapılır.

Decision Tree (Karar Ağaçları), diğer algoritmalardan farklı olarak hem regresyon hem de sınıflandırma problemlerinde kullanılabilir. Karar ağaçlarında veriler küçük karar kümelerine dönüşür. Bu kümelerden karar düğümleri oluşturulur. Her karar düğümünün birden fazla cevabı olabilir. Bu cevaplar karar veya yaprak düğümlere karşılık gelir. En üstteki düğüme kök düğüm denir. Ağacın oluşturulması, entropi hesaplanılarak yapılır. Entropi ile, veri setini en iyi ayrıştıracak öznitelik bulunur ve kök düğüme yerleştirilir. Her yerleşimden sonra, mevcut ağaç için tekrar entropi hesaplanır ve böylece diğer düğümler (öznitelikler) de yerleştirilir.

Naive Bayes, kurulum açısından diğer algoritmalara göre daha basittir ve çok büyük verilerin olduğu durumlarda kullanışlıdır. Koşullu olasılıktaki Bayes teorimine dayanır, özniteliklerin aldığı değerlerin olasılıkları hesaplanarak gelen sorgunun hangi sınıfa ait olma yatkınlığı olduğu saptanır.

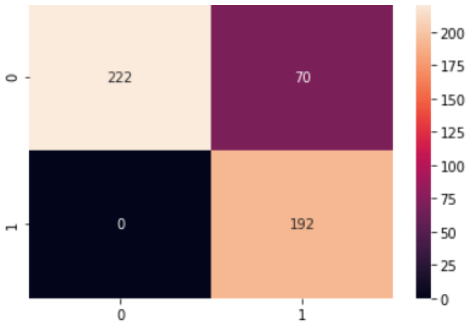
IV.DENEYSEL ANALİZ

KNN, Decision Tree ve Naive Bayes algoritmaları için, önce bu algoritmaların sklearndeki default parametreleriyle model kurulmuş, GridSearch fonksiyonu kullanılarak bu algoritmalar için optimum sonuçlar alındığı parametreler hesaplanmıştır. Sonra, validasyon aşamasında bu değerler ile parametre optimizasyonu yapılmış modeller kurularak tahminleme yapılmış ve iki durumdaki model performansları karşılaştırılmıştır. Validasyon aşamasında ek olarak cross validation yöntemi kullanılarak tahminleme yapılmış, burada alınan sonuçlar önceki sonuçlarla karşılaştırılmıştır.
Bu işlemleri sırasıyla algoritma bazında inceleyelim.

KNN algoritmasıyla kurulan modellerde, en iyi sonuç k komşu sayısının 3 olarak belirlendiği ve uzaklığın minkowski yöntemiyle hesaplandığı model ile alınmıştır. Model konfigürasyonunun başarısı şekil ve 4'te verilmiştir:

	precision	recall	f1-score	support
0.0	1.00	0.76	0.86	292
1.0	0.73	1.00	0.85	192
accuracy			0.86	484
macro avg	0.87	0.88	0.85	484
weighted avg	0.89	0.86	0.86	484

Şekil 3. KNN sınıflandırma raporu

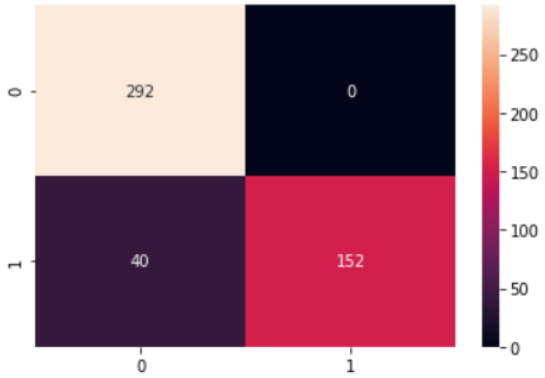


Şekil 4. KNN karmaşıklık matrisi

Decision tree ile kurulan modellerde en iyi sonuç, 'criterion': 'gini', 'max_depth': 2, 'min_samples_leaf': 5 parametreleri ile tahminleme yapıldığı model ile alınmıştır. Model konfigürasyonunun başarısı şekil 5 ve 6'da verilmiştir:

	precision	recall	f1-score	support
0.0	0.88	1.00	0.94	292
1.0	1.00	0.79	0.88	192
accuracy			0.92	484
macro avg	0.94	0.90	0.91	484
weighted avg	0.93	0.92	0.92	484

Şekil 5. Decision tree sınıflandırma raporu



Şekil 6. Decision tree karmaşıklık matrisi

Naive Bayes için; Bernoulli, Multinomial ve Gaussian olmak üzere üç versiyonu da deneysel analiz aşamasında impelente edilmişir. Kurulan modellerde en iyi sonuç, Gaussian NB versiyonunda 'var_smoothing': 0.533669923120631 parametresi ile tahminleme yapıldığı model ile alınmıştır. Model konfigürasyonunun başarısı aşağıda verilmiştir:

	precision	recall	f1-score	support
0.0	0.88	1.00	0.94	292
1.0	1.00	0.79	0.88	192
accuracy			0.92	484
macro avg	0.94	0.90	0.91	484
weighted avg	0.93	0.92	0.92	484

Şekil 7. Naïve Bayes sınıflandırma raporu

K-fold cross validation yöntemiyle doğrulama yapıldığı senaryoda ise,
KNN için 5 fold ile model başarısı: 0.925
KNN için 1 fold ile model başarısı: 0.897
Decision tree için 5 fold ile model başarısı: 0.925…252
Decision tree için 1 fold ile model başarısı: 0.925…109
Gaussian NB için 5 fold ile model başarısı: 0.925…252
Gaussian NB için 5 fold ile model başarısı: 0.925…109
ölçülmüştür.

V. SONUÇ

Yalnızca haber başlıklarını kullanarak haberin gerçek ya da sahte olduğunu tahminlediğimiz projede, validasyon veri setinde en iyi başarı Decision Tree ve Naive Bayes algoritmalarında alınmıştır. Sahte olmasına rağmen, gerçek olarak tahminlediğimiz haberler elimizdeki sahte haberlerin yaklaşık %20'sini oluşturur. Bu da, incelediğimiz her 5 sahte haberin 1'ini gerçek sınıfı olarak yanlış tahminlediğimiz anlamına gelir. Projeyi yapma amacımız göz önüne alndığında bu oran bizim için oldukça yüksektir.
Kurulan modellerin başarısının iyileştirilebilmesi için, başlangıçta düşünüldüğü gibi veri seti başlıklar x kelimeler matrisi ile tutulabilir ve bu matris üzerinde işlemler yapılabilir. Matris boyutunu küçültmek için IDF (inverse document frequency) hesaplanabilir ve fazlaca ortak kullanılan kelimeler veriden çıkartılabilir. Tam tersi şekilde, gerçek ve sahte haberleri ayrı matrislerde tutabilir ve IDF hesaplayarak bu iki sınıfın kendi içinde çokça ortak kullandığı kelimeler bulunarak model kurulabilir. Performansın iyileştirilebilmesi için haber başlığını farklı şekilde kullanmak tek başına yeterli değildir.
Haberlerin içeriği de benzer işlemlere sokularak kullanılabilir. Bununla birlikte ilerleyen aşamalarda; haberin kaynağı, haber başlığının ve içeriğinin uzunluğu da öznitelik olarak kullanılabilir.