TÜRKİYE CUMHURİYETİ
YILDIZ TEKNİK ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

YAPAY ZEKAYA GİRİŞ
DÖENM PROJESİ

Öğrenci Adı: Ali Albayrak
Öğrenci NO: 20011910
E-posta: ali.albayrak@std.yildiz.edu.tr

Ders Yürütücüsü
Prof.Dr. Mehmet Fatih AMASYALI
MAYIS, 2023

İçindekiler

● Ödev Anlatımı

Bu projede yapay zeka ile haberlerin doğruluğu kontrol eden bir (Logistic Regressor) model kuruldu. bu model sadece bir “Text” ve “Fake” labelleri üzerinde çalışır. yani haberin metini ve doğruluk sınıfı.

● Çözüm Genel Anlatımı

kullanılan programlama dili: Python (Jupyter Notebook)

çözüm 3 aşamalı

1- Verileri okuma ve önişleme

burda Text verileri okunur ve gereksiz sütunlar silinir. ardından oluşturduğumuz “DataPreProcessor” Sınıfını kullanarak linkler, html tags, özel karakterler, ve durma kelimeleri silinir. bu temizlenen veri yeni dosyaya kaydedilir(cleaned_data.csv).

2- Model Sınıfını oluşturma

Logistic Regression sınıfında 2 temel fonksiyonu vardır bunlar: fit ve predict. isimlerinden de anlaşılacak gibi fit fonksiyonu modelimizi veri kümesine uyarlıyor ve predict fonksiyonu ise görülmeyen verilerin sınıfını tahmin etmek için kullanılır.

fit fonksiyonu Gradient Descent mantığını ve sigmoid fonksiyonu kullanarak en başarılı katsayıları bulmaya çalışır bir sonraki bölümde biraz daha detaylı anlatılacak.

3- Modeli ve Sonuçlarını görüntüleme

burada eğitilen modelin performansı ve elde ettiği sonuçları izleyip görselleştirildi. fakat modeli kullanmak ile birlikte bizim verilerimiz string olduğu için Vectorizer kullanılmalı.

projenin sonunda 0.98 başarı oranına sahip bir model çıkıyor (puanlar sonraki kısımlarda gösterilecek)

● Sayısal Başarı

lojistik regression genel formülü şu şekilde oluyor

y: gerçek sınıf

h: tahmin edilen sınıf

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right)$$

$$\text{Cost}\left(h_\theta(x), y\right) = -\log\left(h_\theta(x)\right) \qquad \text{if y} = 1$$

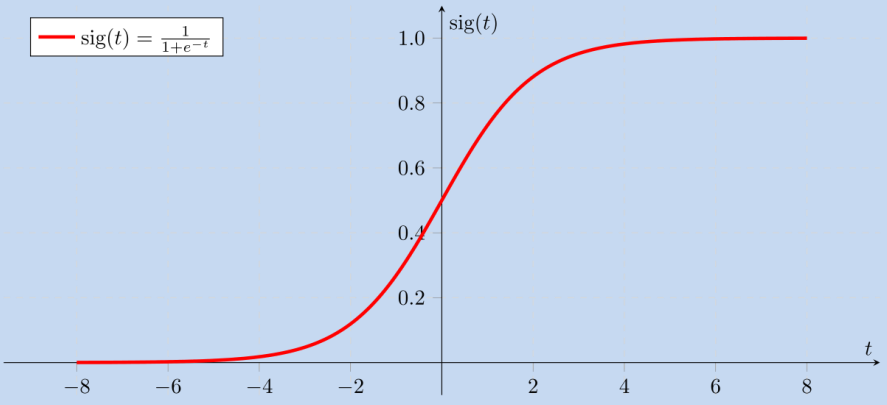
$$\text{Cost}\left(h_\theta(x), y\right) = -\log\left(1 - h_\theta(x)\right) \qquad \text{if y} = 0$$

bu iki cost fonksiyonu birleştiren bir formül var.

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log\left(1 - h_\theta(x^{(i)})\right)\right]$$

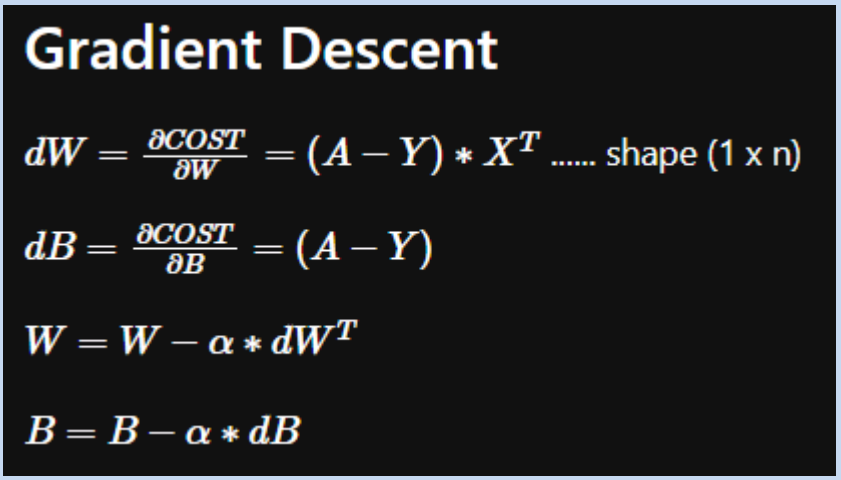
bu formul aynı sonuçlar veriri çünkü y=1 olunca artı işaretinden sonraki kısım sıfır olur ve y=0 olduğu zamanda da artı işaretinden önceki kısım sıfır olur

tahmin edilen sınıf sigmoid fonksiyonu kullanarak bulunur.



sigmoid fonksiyonun sonucu 0,5’ten küçük ise 0.sınıf, büyük ise 1.sınıf olarak tahmin edilir.

şimdi Gradient Descent algoritması uygulanır.



ve gratient descent’in temel amacı değişim hesaplayarak (Türev alarak) en az cost değeri yani en iyi W ve B değerleri bulmaya çalışır.

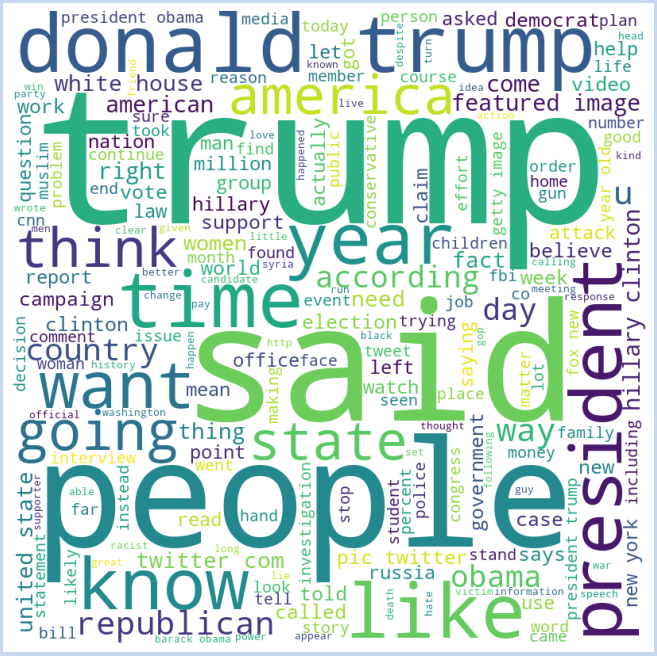
α: learning rate

W: weights

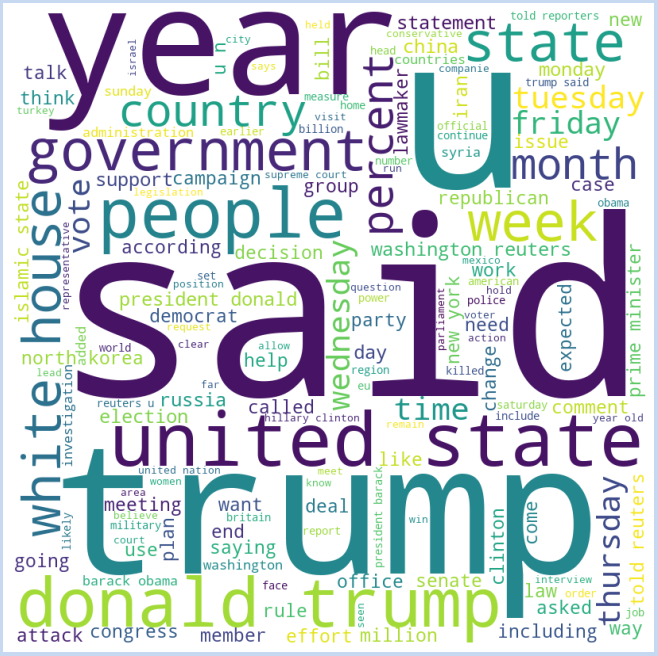
B: bias

- **Tablo ve Grafikler**

Fake ve True haberlerde en çok tekrarlanan kelimeler:



Fake



True

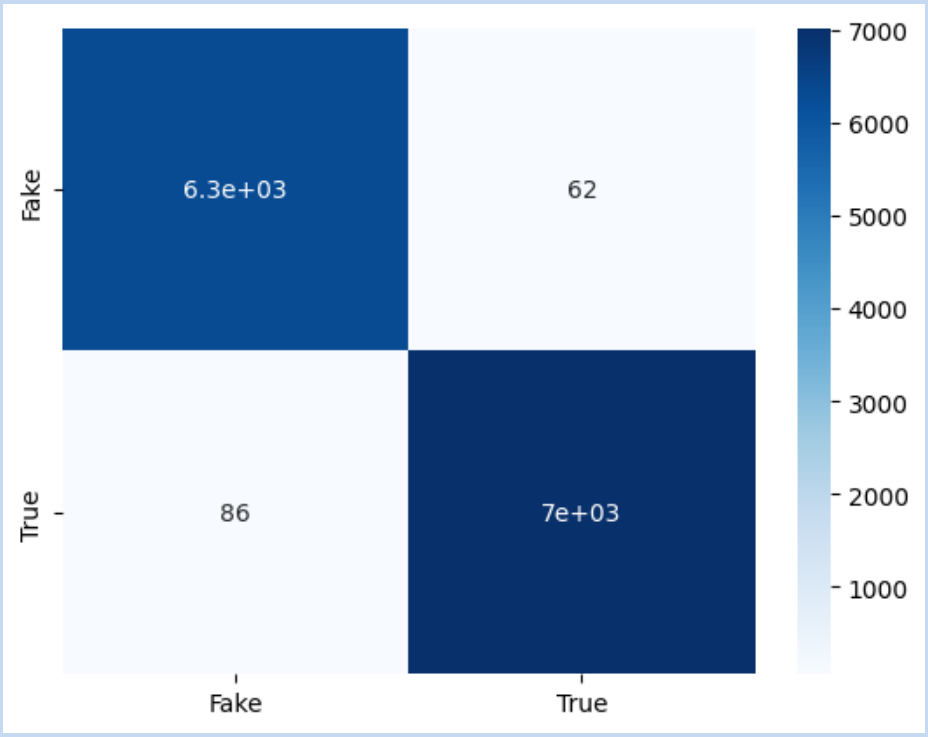
tahminleyicilerin performasını gösteren puan ve tablo:

accuracy_score: 0.9890126206384559

Report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	6368
1	0.99	0.99	0.99	7102
accuracy			0.99	13470
macro avg	0.99	0.99	0.99	13470
weighted avg	0.99	0.99	0.99	13470

tahminleyicilerin bir örneğini gösteren grafik:



● Bulgular ve Yorumlar

bu projenin sonunda TfidfVectorizer temel görevini anladık.stringi yazıya çevirir ve elde edilen değerler 0-1 arasında olur.

veri temizlemenin önemlerine bir önem daha katıldı o da “over fitting”den kaçınma. yani burda teamizlenmeyen veri genellikle daha yüksek puana sahip fakat model bu veriye over fit oluyor yani farklı verilerde aynı başarıyı göstermeyebilir.

● Video

Youtube: https://youtu.be/CgJPQvcmifw

Kaynakça

- Hesham Asem ML Course (Arabic)
  - https://www.youtube.com/@HeshamAsem
  - https://drive.google.com/drive/folders/1b8IaXG5KXDSunpOhWSBr4j8U2tlOSuBp
- Machine Learning Art page (Arabic)
  - https://www.facebook.com/Machine.Learning.Art/
- “Logistic Regression — Detailed Overview” Article on Meduim
  - https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
- AssemblyAI Youtube channel
  - https://www.youtube.com/@AssemblyAI
- StatQuest: Logistic Regression video
  - https://www.youtube.com/watch?v=yIYKR4sgzI8&ab_channel=StatQuestwithJoshStarmer