

Türkiye Açık Kaynak Platformu
Online Yarışma Programı

Türkçe Doğal Dil İşleme

MSKÜ-CENG-NLP-1

- EKİBİMİZ -

Alperen AYYILDIZ
Hasan ARCAS
Zehra RIDVANOĞULLARI
Murat Avzet ASAN

- EKİP ÜYELERİNİN PROJEYE SUNDUĞU KATKI -

Alperen AYYILDIZ

- Geliştirilecek modeller için gerekli literatür taraması ve dokümantasyon sağlanması.
- Modelin geliştirilmesinde Python programlama desteği.

Zehra RIDVANOĞULLARI

- Makine öğrenmesi modelleri geliştirilmesi ve model analizleri yapılması.

Murat Avzet ASAN

- Ek veri bulunması ve sınıflandırılması.

Hasan ARCAS

- Derin öğrenme modelleri geliştirilmesi ve model analizleri yapılması.

- PROBLEM -

Tarafımıza sağlanan etiketli veri kullanılarak cümleleri "ofansif" veya "ofansif değil" şeklinde metin sınıflandırması modeli geliştirilmesi. Ek olarak ofansif cümlelerin alt kırılmalarını saptamak için metin sınıflandırması yapılması.

- PROBLEMİN ÇÖZÜMÜ -

ÇÖZÜM 1: Multioutput Classification (Çoklu Çıktı Sınıflandırma)

- Her bir cümle için iki çıktı veren model geliştirilmesi.

ÇÖZÜM 2: İki Aşamalı Sınıflandırma

- Cümleler birinci modelde ofansif veya ofansif değil şeklinde sınıflandırılıyor.
- İkinci modelde, ilk aşamada ofansif şeklinde sınıflandırılan cümleler alt kırılımlarına göre sınıflandırılır.

- HANGİ YÖNTEMLE ÇÖZÜM GELİŞTİRİLDİ?-

- Ofansif olup olmamasına karar vermek ve ofansif olan cümlelerin alt kırımlarına karar verme iki ayrı sınıflandırma problemi olduğu için her bir probleme yönelik farklı birer model geliştirmeye karar verdik.
- BERT dil modelinin Türkçe kaynaklar üzerine eğitilmiş olan "dbmz/bert-base-turkish-cased" adlı ön eğitilmiş modelini kendi eğitim verimizle besleyerek bu sınıflandırma problemi için daha spesifik bir model elde etmiş olduk. Kullandığımız iki model de bu yaklaşıma dayanmaktadır.

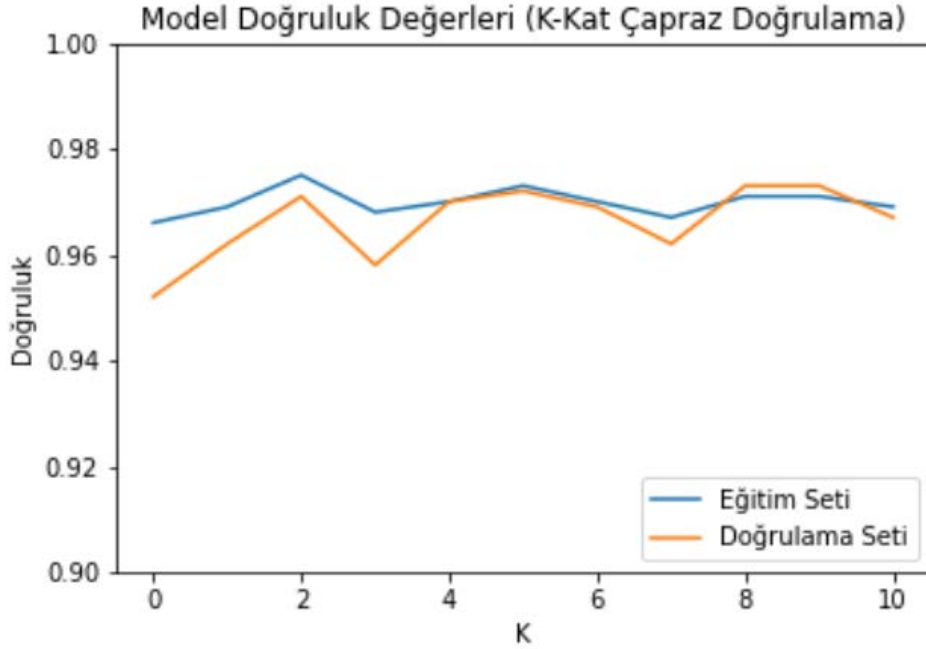
- TEKNİK ÇALIŞMALAR-

- Veri ön işleme (1. ve 2. model)
- Veri Arttırma (2. model)
- Ek Verileri Dahil etme (2. model)
- Model Hazırlığı (1. ve 2. model)
- Model Eğitimi (1. ve 2. model)
- Çapraz Doğrulama (1. ve 2. model)

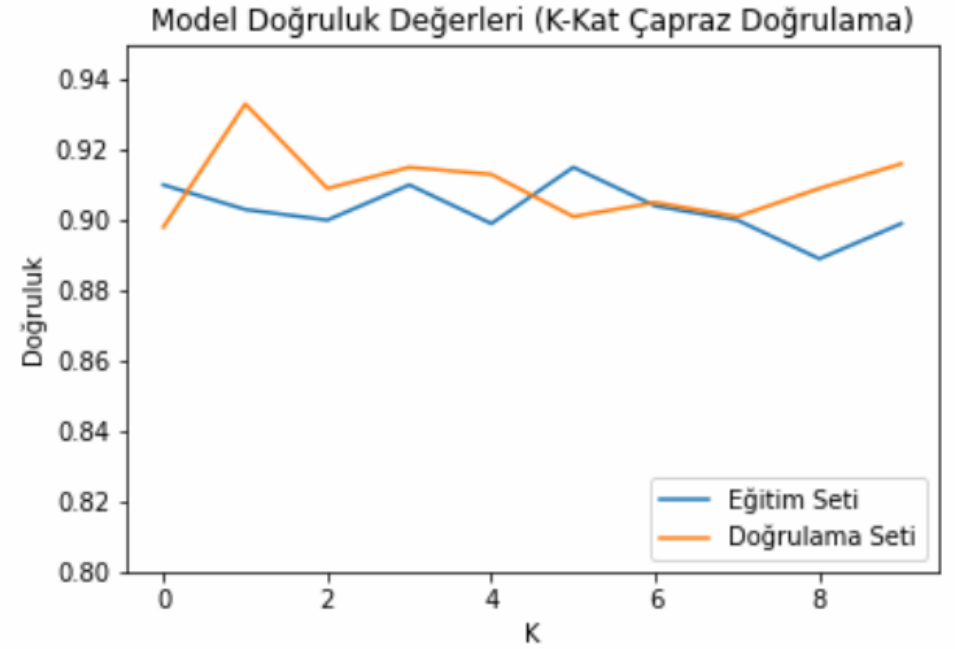
- PROJE İŞ AKIŞI -

1. Literatür Taraması ve benzer problem çözümleri tespiti
2. Veri ön işleme
3. Makine Öğrenmesi modelleri oluşturma
4. Dil Modeli oluşturma
5. Derin Öğrenme Modeli oluşturma
6. Ek veri bulma/yaratma
7. Model eğitimi
8. Çapraz doğrulama ile gerçek skor değerine ulaşma

SONUÇLARIMIZ

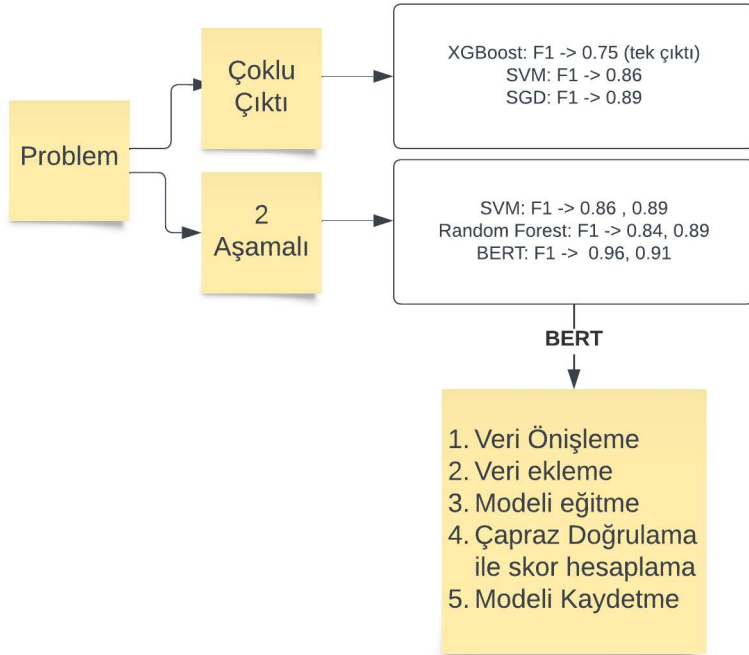


Eğitim Seti Ortalama Doğruluk Skoru ≈ 0.969
Doğrulama Seti Ortalama Doğruluk Skoru ≈ 0.966
Test Veri Seti F1-Skor ≈ 0.972

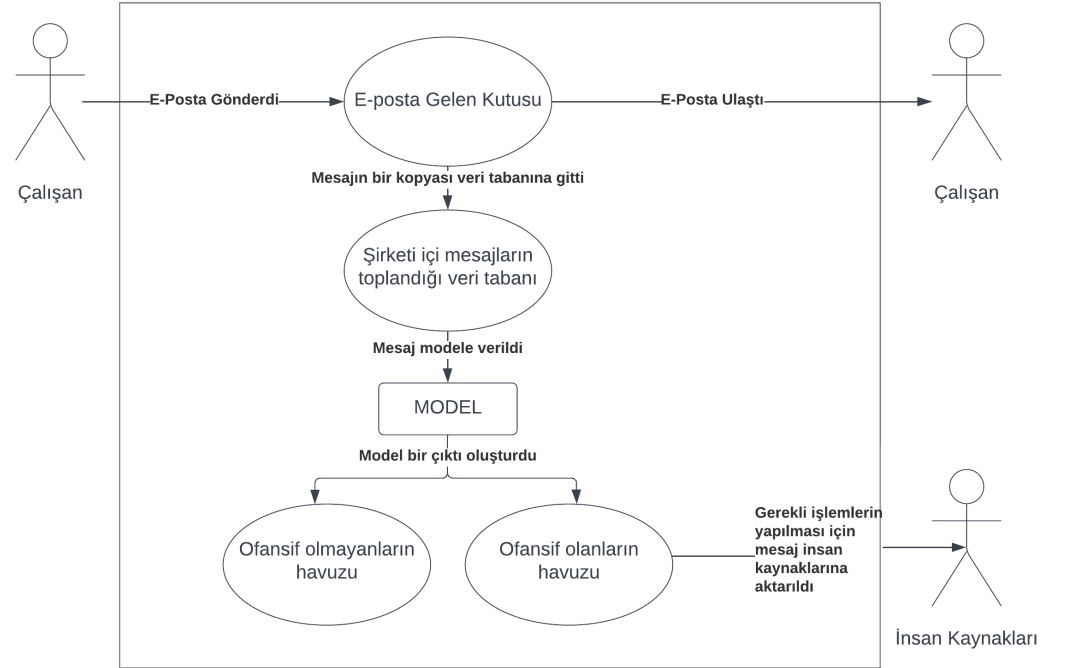


Eğitim Seti Ortalama Doğruluk Skoru ≈ 0.902
Doğrulama Seti Ortalama Doğruluk Skoru ≈ 0.910
Test Veri Seti F1-Skor ≈ 0.905

- PROJE İLE İLGİLİ YOL HARİTAMIZ-



USE-CASE DİYAGRAMI



LINKLERİMİZ

Demo Video Linki

Github Linki

Not: Projeye ait Github Linkini ekleyiniz.

www.turkiyeacikkaynakplatformu.com



- Teşekkürler -