<>Türkiye
Açık Kaynak
Platformu</>

Turkey Open
Source Platform

ACIKHACK

Açık Kaynak Hackathon Programı

<>Türkiye
Açık Kaynak
Platformu</>

Turkey Open
Source Platform
www.turkeyopensourcep latform.com

• ACIKHACK
Açık Kaynak Hackathon Programı

MSKU
NATURAL LANGUAGE PROCESSING
RESEARCH GROUP

# - MSKU-CENG-NLP-2 -

TEKNOFEST
HAVACILIK, UZAY VE TEKNOLOJİ FESTİVALİ

www.turkiyeacikkaynakplatformu.com

<>Türkiye Açık Kaynak Platformu</>

Turkey Open Source Platform
www.turkeyopensourceplatform.com

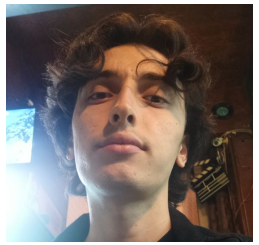
# EKİBİMİZ

• ACIKHACK
Açık Kaynak Hackathon Programı

**Prof. Dr. Bekir Taner Dinçer - Danışman**
**MSKÜ Bilgisayar Mühendisliği Öğretim Görevlisi**

**Şevval Özekinci - Takım Lideri**
**MSKÜ Bilgisayar Mühendisliği 3. sınıf öğrencisi**

**Hatice Nurel - Veri Önişleme**
**MSKÜ Bilgisayar Mühendisliği 3. sınıf öğrencisi**



**Enes Dertli - Model Geliştirme**
**MSKÜ Bilgisayar Mühendisliği 4. sınıf öğrencisi**

**Yasemin Demirkaya - Model Geliştirme**
**MSKÜ Bilgisayar Mühendisliği 4. sınıf öğrencisi**

TEKNOFEST
HAVACILIK, UZAY VE TEKNOLOJİ FESTİVALİ

*www.turkiyeacikkaynakplatformu.com*

# EKİP ÜYELERİNİN PROJEYE SUNDUĞU KATKI



**Şevval Özekinci (Takım lideri):**

- Takım üyelerinin iş paketlerini tamamlanmasını denetleme ve takım üyelerinin görevlendirilmesi.
- Gradio linkinin oluşturulması ve denetlenmesi.

**Hatice Nurel (Veri Ön işleme):**

- Verinin normalleştirilmesi ve modellemeye uygun hale getirilmesi.
- Bert modelinin oluşturulması.

**Enes Dertli (Model Geliştirme):**

- Naive Bayes modelinin oluşturulması.
- GitHub deposunun  oluşturulması ve düzenlenmesi.

**Yasemin Demirkaya (Model Geliştirme):**

- Modellerin geliştirilmesi ve denetlenmesi.
- Modellerin birleştirilmesi.

# Problemi Tanıyalım

Aşağılayıcı Söylem Tespiti Projesi, günümüzde giderek artan çevrimiçi iletişim ortamlarında karşılaşılan aşağılayıcı söylemlerin tespitinin ne kadar önemli olduğunu gösteren bir örnektir. Bu projede, doğal dil işleme yöntemleri kullanılarak aşağılayıcı söylemlerin tespiti amaçlanmıştır. Proje için oluşturulan veri kümesi, aşağılayıcı söylem içerip içermediğine ve içeriyorsa hangi alt kategoride (cinsiyetçi, ırkçı, küfür veya hakaret) olduğuna dair etiketlenmiştir. Bu sayede, projenin sonuçları hem çevrimiçi ortamların güvenliği hem de insanların psikolojik sağlığı açısından oldukça önemlidir.

# Problemin Çözümü

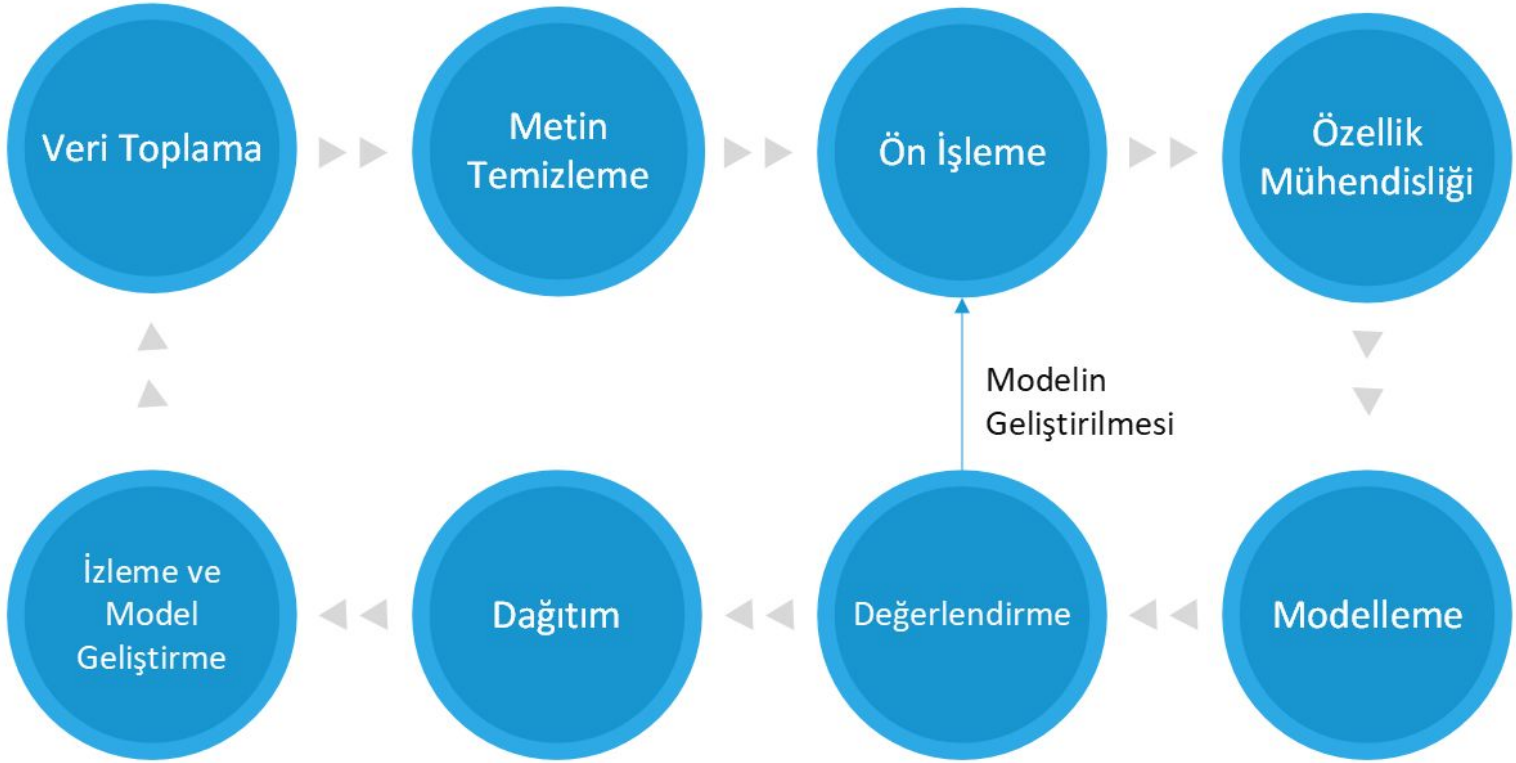
**Verilen probleme uygun bir duygu analizi (OFFENSİVE, NON-OFFENSIVE) ve OFFENSIVE altında 4 sınıf için sınıflandırma yaklaşımı.**

Bu görev için iki aşamalı yaklaşımımız mevcuttur.

**1)Duygu analizi:** Offensive, Non-offensive ayrımı için Naive Bayes Modeli kullanılmıştır.

**2)Offensive altında kategorizasyon:** Bu aşamada offensive ifadeleri 4 ayrı sınıfa ayırmak amacıyla BERT adlı derin öğrenme modeli, Türkçe metin sınıflandırması için kullanılmıştır.

# HANGİ YÖNTEMLE ÇÖZÜM GELİŞTİRİLDİ?



# Süreç

1) Yarışma tarafından sağlanan veri setine ön işleme uygulanması
- Tüm harflerin küçük harflere dönüştürülmesi
- Gereksiz karakterleri, noktalama işaretlerini, sayıları ve özel karakterlerin silinmesi
- Her kelimenin yerini kelime 5 harften fazla ise, ilk 5 harfinin alması

2) Veri setini dengeli hale getirme
- Non-offensive sınıfının offensive sınıfından az olması sonucu, 4 sınıftan eşit sayıda veri alınarak non-offensive sınıfındaki veri sayısına eşitlenmesi ile veri setinin dengeli hale getirilmesi

3) Yüksek F1 score elde etmek için modelleme çalışmaları
- Train veri setinin train-validation veri seti olarak ayrılması ve farklı modeller ile denenmesi ve sonuçların kaydedilmesi

4) Çalışmalar sonucu elde edilen verilerin karşılaştırılması ve en uygun modellerin seçilmesi

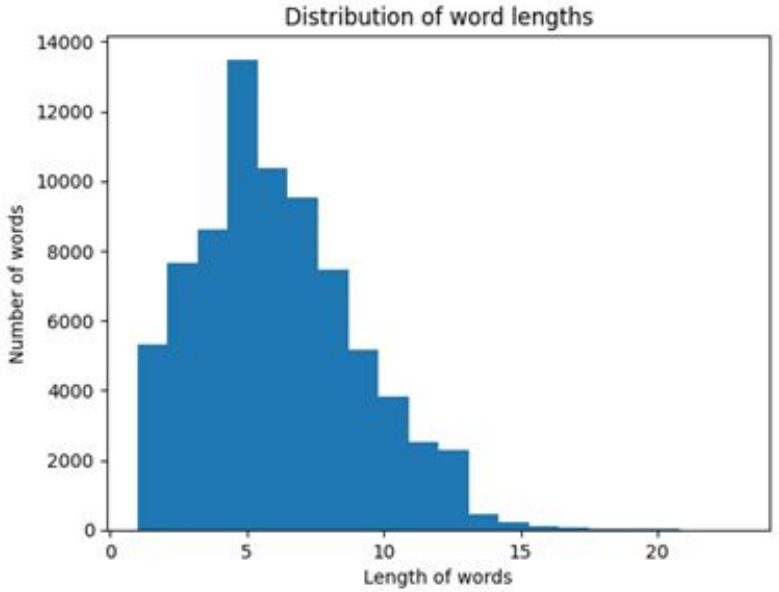
<>Türkiye Açık Kaynak Platformu</>

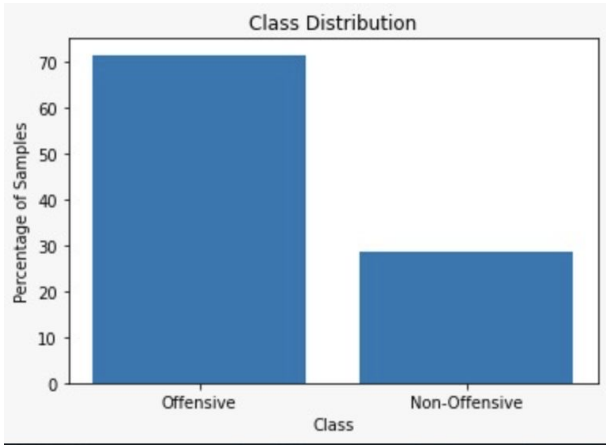
Turkey Open Source Platform
www.turkeyopensourcep latform.com

# YAPMIŞ OLDUĞUMUZ TEKNİK ÇALIŞMALAR

ACIKHACK
Açık Kaynak Hackathon Programı

Yarışma kapsamında verilen veri setinin talim derlemi,verideki gereksiz karakterlerin, noktalama işaretlerinin, sayıların ve özel karakterlerin kaldırılmasıyla metinler analiz edilebilir bir hale getirildi. Stemming methodu  agresif olduğundan, onun yerine  train veri setindeki kelimelerin uzunluk ortalamasına bakılarak, her kelimenin ilk beş harfi ile yeni text sütunu modellerde kullanılmak üzere oluşturuldu.




Veri seti dengesizliklerinin giderilmesi için scikit-learn kütüphanesinden faydalanıldı. is_offensive kolonunda değeri 0 olan veriler ile aynı sayıda, değeri 1 olan verilerden rastgele örneklem alınarak yeni bir dengeleştirilmiş veri seti oluşturuldu.

www.turkiyeacikkaynakplatformu.com

TEKNOFEST
HAVACILIK, UZAY VE TEKNOLOJİ FESTİVALİ

Projemizde iki türlü deney yaptık:

- Beş kategori olacak şekilde (Non-Offensive, Racist, Profanity, Sexist, Insult) sınıflandırma modelleri denedik.
- İlk aşama olarak Non-offensive ve Offensive sınıflandırması için bir model, ikinci aşama olarak diğer 4 sınıf için ayrı bir model denedik.

Modellemelerimizde SVM, Logistic Regression, Naive Bayes, Random Forest, LSTM, XGBoost, Bert vb. gibi çeşitli makine öğrenimi/derin öğrenme algoritmaları denedik.

Değerlendirmeler sonucunda oluşturulan modellerin performansları karşılaştırıldı ve problemin çözümüne ilişkin en uygun modellere karar verildi. Proje, github deposunda umuma açıldı.

<>Türkiye Açık Kaynak Platformu</>

Turkey Open Source Platform

www.turkeyopensourcep latform.com

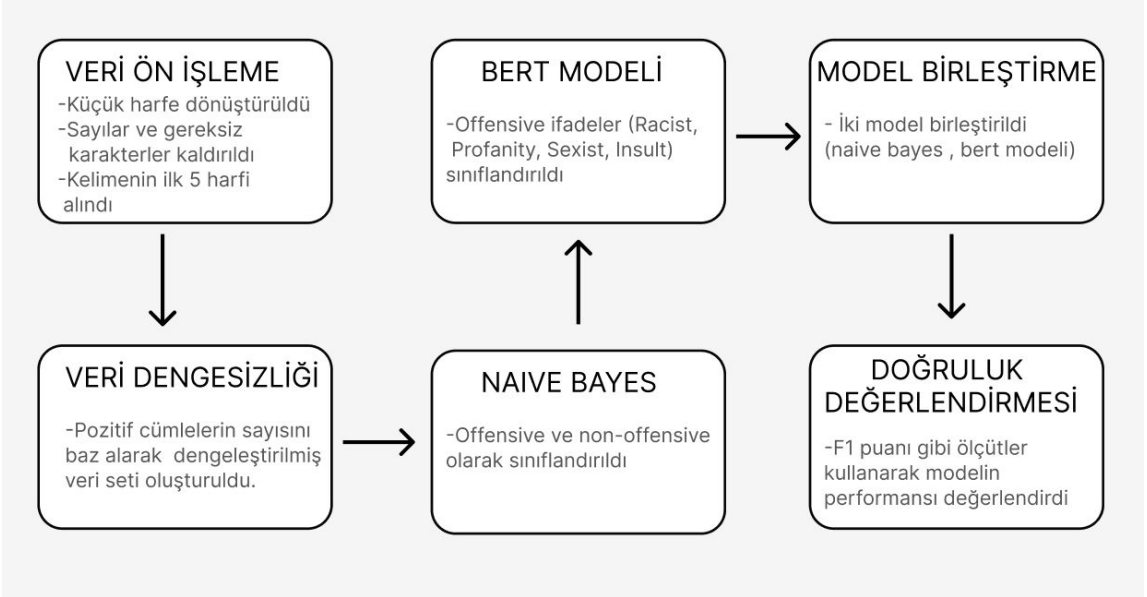
# PROJE İŞ AKIŞI VE YOL HARİTAMIZ

ACIKHACK

Açık Kaynak Hackathon Programı

**Yol haritamız birkaç makine öğrenmesi adımından oluşmaktadır:**

- Veri ön işleme
- Verinin dengeli hale getirilmesi için çeşitli yöntemlerin denenmesi
- Verinin sayısallaştırılması ile modele girdi olarak uygun hale getirilmesi
- Veri setinin train ve validation veri seti olarak ayrılması
- Makine öğrenmesi ve derin öğrenme modellerinin denenmesi
- Deney sonuçlarının karşılaştırılması




TEKNOFEST HAVACILIK, UZAY VE TEKNOLOJİ FESTİVALİ

# Planlarımız

Elde edilen çözüm gelecekte planlanan hedeflere ulaşmak için önemli bir adım olarak görülmektedir. Mevcut çözüme yenilikçi özellikler eklenmesi gelecekte kullanıcılara daha fazla değer sunabilir. Bu durumda, mevcut veri kaynaklarına ek olarak yeni veri kaynakları entegre edilmesi, veri çözümlemesi için daha kapsamlı bir veri seti sağlayabilir. Kullanıcı deneyimini iyileştirmek için kullanıcı arayüzünün geliştirilmesi, müşteri memnuniyetini artırabilir. Ayrıca bu çözüm, ekibimizin yeteneklerinin geliştirilmesine de katkı sağlayacak ve yenilikçi yaklaşımların benimsenmesiyle sürekli gelişim hedefine ulaştırılacaktır.

# Proje Linklerimiz

**Sunum Linki :**

https://www.youtube.com/watch?v=SUiZS_Xd3cY

**Github Linki :**

https://github.com/CCXXVII/MSKU-CENG-NLP-2-Final

**Gradio Linki :**

https://drive.google.com/file/d/1LMKMBpAwkXASGdZkwMhRUYj-yLGKQh8Q/view?usp=share_link

Not: Projeye ait Githup Linkini ekleyiniz.

**Dinlediğiniz için teşekkür ederiz.**
**Saygılarımızla,**

**MSKU-CENG-NLP-2**