

Türkiye Açık Kaynak Platformu  
*Online Yarışma Programı*

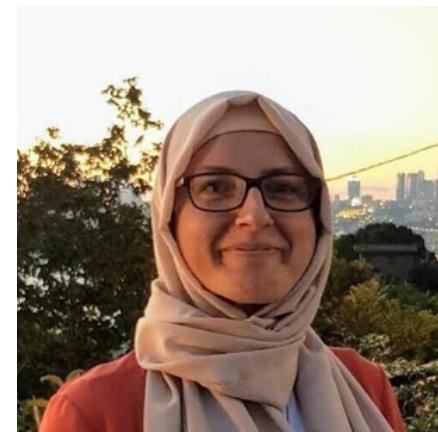
# Türkçe Doğal Dil İşleme

# Ekibimiz ve Sorumluluklarımız

Danışman



TRT DATA WARRIORS  
TACKLES HATE



Machine Learning Team Lead  
at TRT

Takım Lideri



Machine Learning Engineer  
at TRT

Üye



Machine Learning Engineer  
at TRT

Üye



Data Scientist  
at TRT

Üye



MLOPS Engineer  
at TRT

**M. Emir Koçak ve Burcu Şenol**



- Veri Analizi
- Metin Ön İşleme
- Veri toplama/büyütme
- Gradio Servisinin Hazırlanması
- Hugging Face ortamına model transferi
- Hugging Face aracılığıyla demo hazırlanması

**Mustafa Budak ve Nusret Özates**



- Sanal Tensorflow ve Torch ortamlarının kurulması
- Veri toplama/büyütme
- Modelleme
- Kod yapısının tasarılanması
- GitHub entegresi

# PROBLEM: Aşağılayıcı Söylemlerin Doğal Dil İşleme İle Tespiti



- **OTHER:** Cümplenin nefret söylemi içermemesi / nötr cümleler
  - Defnےyle artık konuşmuyoruz
- **INSULT:** Hakaret edici, aşağılayıcı, hor görülmeye gidi söylemlerin var olduğu cümleler
  - Çürüklü dişli
  - Haysiyet yoksunu herif
- **PROFANITY:** Küfür içeren cümleler
  - Ağzın kenef olsa s\*\*\*lmaz
- **RACIST:** Irkçı söylemlerin yer aldığı cümleler
  - Zenciler hayatı bir sıfır yenik başlar
- **SEXIST:** Cinsiyetçi, cinsiyet ayrıımı yapılan söylemlerin yer aldığı cümleler
  - Erkekler zora gelmez

# Literatür Taraması

- F. Beyhan, B. Çarık, İ. Arın, A. Terzioğlu, B. Yanıkoglu, R. Yeniterzi, "**A Turkish Hate Speech Dataset and Detection System**", Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 2022, 4177--4185
- Ç. Toraman, F. Şahinuç, E. Y. Yılmaz, "**Large-Scale Hate Speech Detection with Cross-Domain Transfer**", Proceedings of the Language Resources and Evaluation Conference, Marseille, France, 2022, 2215-2225
- İ. Mayda, Y. E. Demir, T. Dalyan and B. Diri, "**Hate Speech Dataset from Turkish Tweets,**" 2021 Innovations in Intelligent Systems and Applications Conference (ASYU), Elazig, Turkey, 2021, pp. 1-6.
- Ç. Çöltekin, "**A Corpus of Turkish Offensive Language on Social Media**", Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 2020, 6174--6184
- J. S. Malik and G. Pang and A. van den Hengel, "**Deep Learning for Hate Speech Detection: A Comparative Study**", 2022, 2202.09517, arXiv, cs.CL
- Bao, Tong et al. "**A BERT-Based Hybrid Short Text Classification Model Incorporating CNN and Attention-Based BiGRU.**" J. Organ. End User Comput. 33 (2021): 1-21.

# PROBLEMIN ÇÖZÜMÜ 1: VERİ ANALİZİ

## Karşılaşılan Sorunlara Bakış:

- Yanlış etiketli veriler

Örnekler:

- İspanyolların hepsinin, en& az 10, çocukları& var -- RACIST
- Türkler çok misafirperverdir -- RACIST

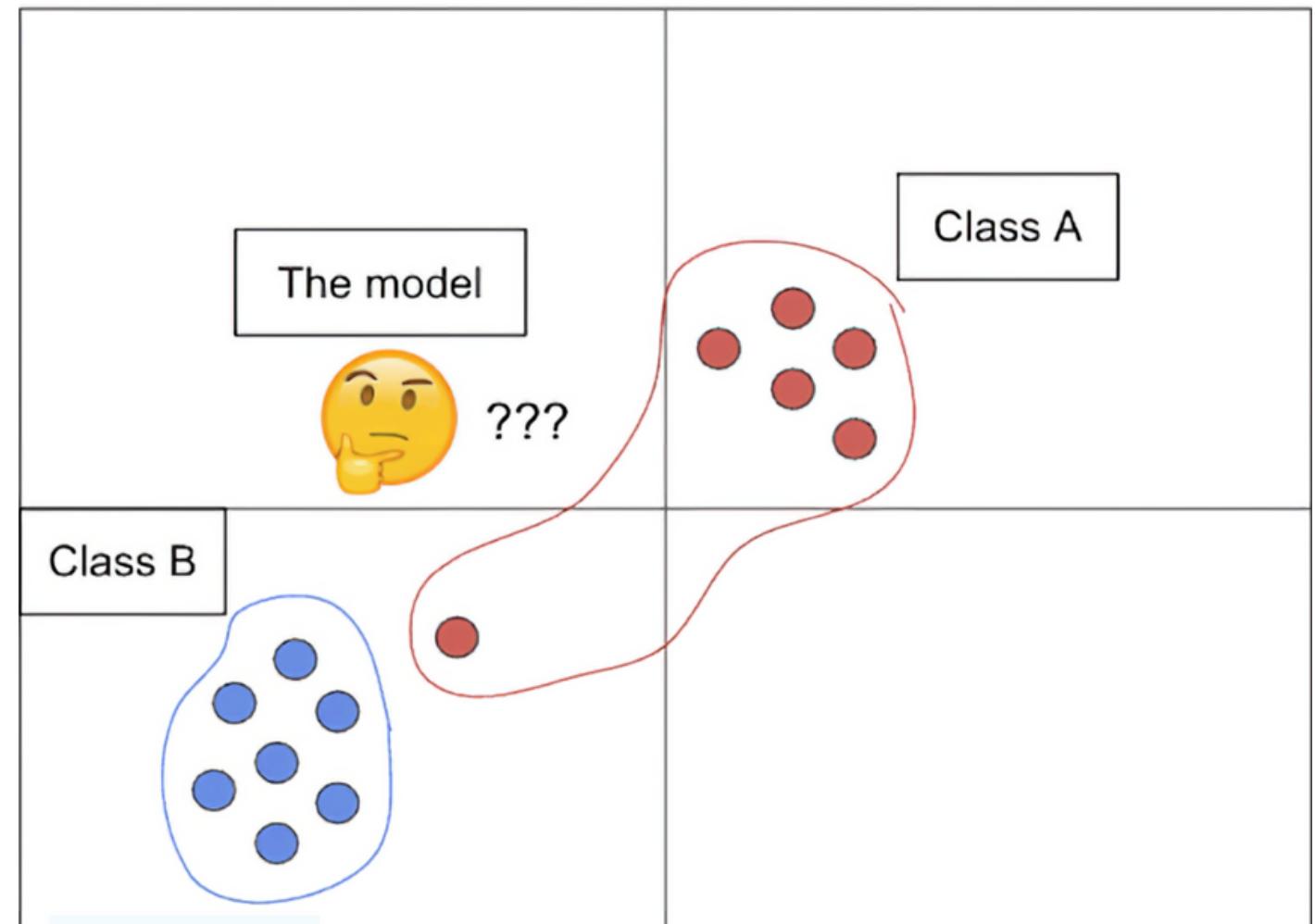
- Bir metnin birden fazla sınıfı ait olabileceği (multi-label problem)

- Türkiye'de yaşayan tüm ermenileri s\*\*\*yim -- RACIST? PROFANITY?

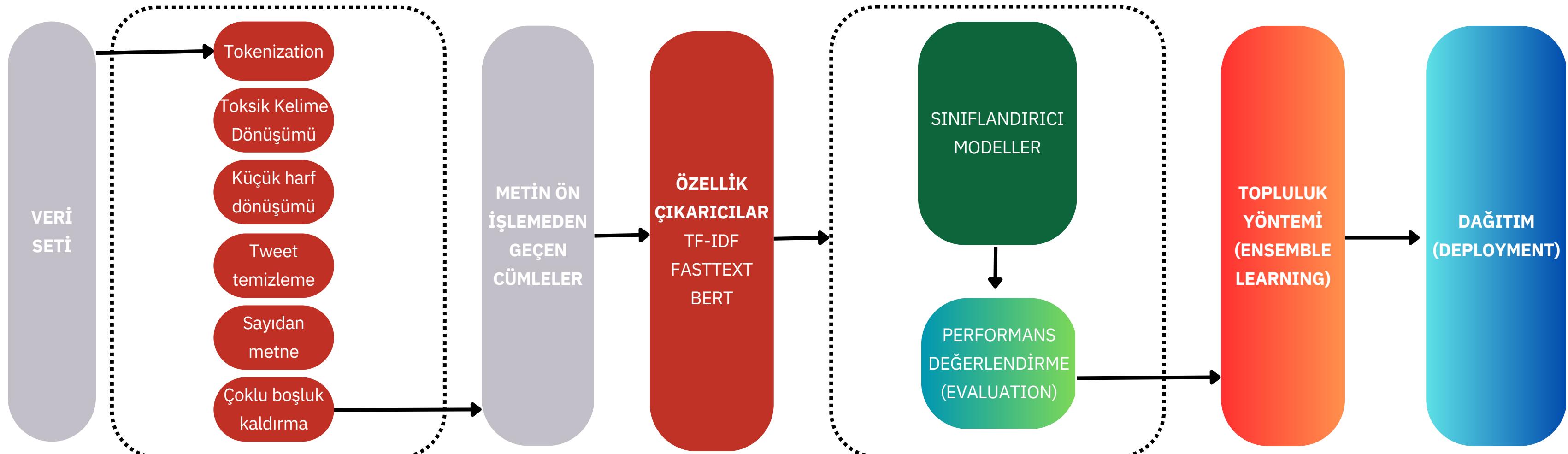
- Aynı anahtar kelimelere sahip fakat farklı sınıflara ait metinler

- lanet olsun -- PROFANITY
- lanet olsun -- INSULT
- O şerefsiz bir adamdı -- PROFANITY
- O bize şerefsizlik yaptı -- INSULT

- Verinin limitli olması

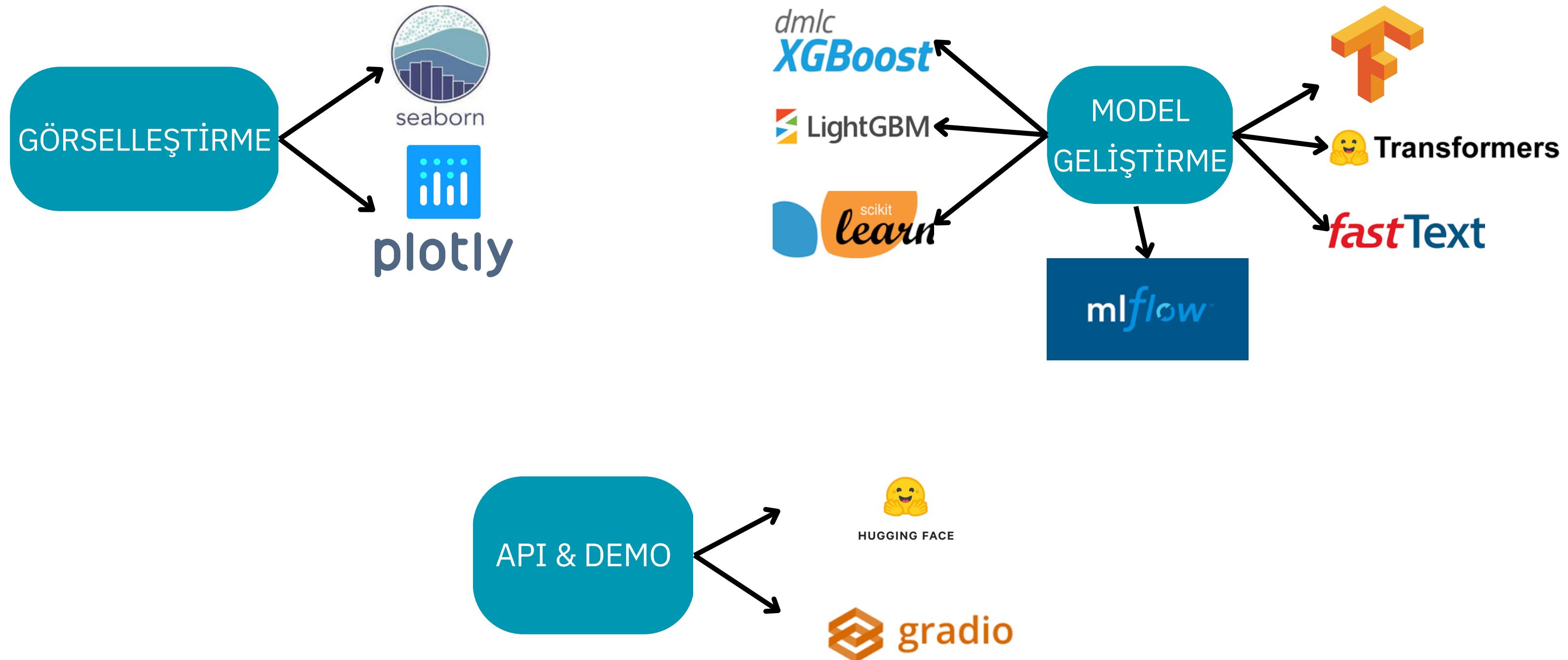


# PROBLEMIN ÇÖZÜMÜ 2





# Kullanılan Teknolojilere ve Kütüphanelere Bakış





# TEKNİK ÇALIŞMA 1: VERİ ÖN İŞLEME ADIMLARI

- **Toksik kelime düzeltme ve indirgeme:** Yanlış ve eksik yazılmış toksik kelimeleri düzeltme ve aynı anlama sahip fakat farklı yazılmış kelimeleri ortak bir kelimeye indirgeme, İngilizce toksik söylemleri Türkçe'ye çevirme

Cümlelerde Geçen Kelime	Düzeltilmiş Hali
amk, aq, awk, mq, a.w, a.q.	a**na k***yım
sg, sie, bsg, ybsg	s**tir git
oç, o.ç, oc, bitch, motherfucker	or**pu çocuğu

## Ön işleme öncesi

- 300 TANE şerefsiz toplanmış aq #isyanediyor <https://www.youtube.com/e1q...>

## Ön İşleme sonrası

- Üçüz tane şerefsiz toplanmış a\*\*na k\*\*\*yım isyanediyor

- Küçük harf dönüşümü
- Tweetlere ait olabilecek hashtags (#) ve mentions (@) kaldırıldı
- Metinlerde yer alan ve yer alabilecek tüm URL'leri kaldırma
- Çoklu boşlukları kaldırma:
  - Senin yapacağın işin amk --> Senin yapacağın işin amk
- Sayıları metne çevirme:
  - 300 tane adamvardı orada --> Üçüz tane adamvardı orada



Farklı kelime vektörleri, sınıflandırma sonuçlarını doğrudan etkiler. Bu problem özelinde kelime vektörlerini eğitmek için FastText ve BERT kullanılarak tüm temel modeller karşılaştırılmıştır.

## 1. TF-IDF + Boosting Algoritmaları

## 2. FastText embeddingleri kullanan modeller:

- a. *Bi-GRU + Bi-LSTM modeli*
- b. *Bi-LSTM + Attention modeli*
- c. *Bi-GRU + Bi-LSTM + CNN modeli*

## 3. Çeşitli transformers modelleri (*Electra, Bert, ConvBert ...*)

## 4. BERT embeddingleri kullanan modeller (BERT modelinin çıkışının son katmanından çıkarılan cümle vektörleri):

- a. *Bi-GRU + Bi-LSTM modeli*
- b. *Bi-LSTM + Attention modeli*
- c. *Bi-GRU + Bi-LSTM + CNN modeli*

FastText'in kelime vektörü boyutu **300**'e ve BERT'nin **768**'ine ayarlanmıştır. Batch size **256** ve cümle uzunluğu **32** olarak belirlenmiştir.



## ÇÖZÜM TEKNİKLERİ 2: MODELLERİN PERFORMANSLARI

Type	Model	PRECISON	RECALL	F1 MACRO	EĞİTİM SÜRESİ (20 epochs)
TF-IDF + Boosting	XGBoost / LightGBM / Catboost	71-73%	72-73%	75-77%	45s
FastText+Temel Yöntem	Bi-GRU + Bi-LSTM	89.45%	89.3%	89.4%	~271s (20 epochs)
	Bi-LSTM + Attention	86.3%	86.1%	86.75%	~265s (20 epochs)
	Bi-GRU + Bi-LSTM + CNN	89.4%	89.6%	89.6%	~282s (20 epochs)
ConvBertTurk mc4-based model	Bi-GRU + Bi-LSTM	94.35%	94.6%	94.47%	~475s (20 epochs)
	Bi-LSTM + Attention	94.12%	94.35%	94.23%	~451s (20 epochs)
	Bi-GRU + Bi-LSTM + CNN	94.46%	94.55%	94.5%	~492s (20 epochs)

# ÇÖZÜM TEKNİKLERİ 3: VERİ BÜYÜTME (DATA AUGMENTATION)

## Ek veriye neden ihtiyacımız var?

- Çok sınıfılı ama limit bir veri setine sahip olmamız
- Veri hataları
- Veriler arası yanılık (bias)

## Verileri nereden topladık?

1. <https://huggingface.co/datasets/Toygar/turkish-offensive-language-detection>
2. Çöltekin Troff
3. <https://github.com/avaapm/hatespeech> (Twitter scraping ile veri toplandı)
4. ChatGPT
5. Geri Çeviri Yöntemi (googletrans, translate kütüphaneleri)

bu kelimelerle sentetik cümle oluştur

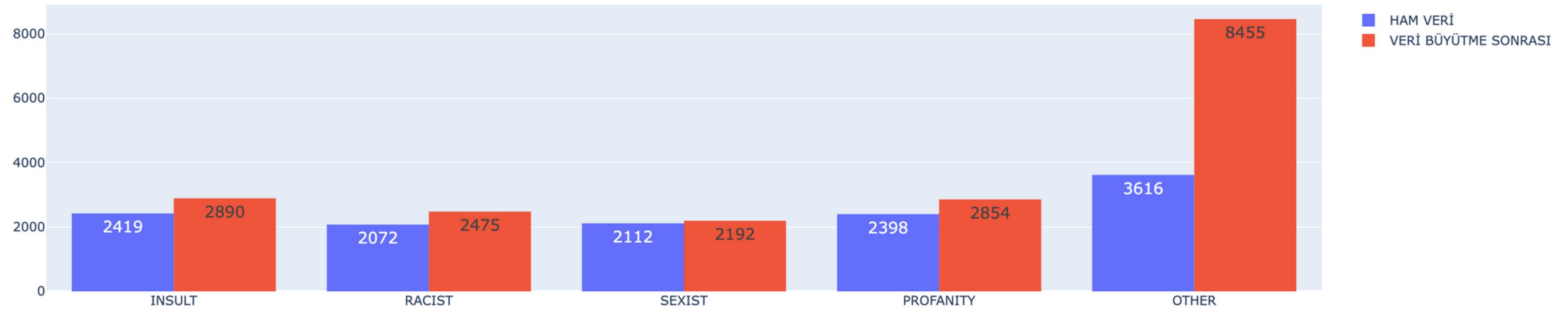
Tabii, örneğin:

- Vicdansız insanlar her zaman başkalarına zarar verir.
- Nankör davranışlarından dolayı kimse onunla iş yapmak istemez.
- Adın batsın dediği anda herkesin yüz ifadesi değişti.
- O, bir kukla gibi yönetiliyor gibi görünüyor.
- Adam değilsin sen, bir canavarsın!
- Sana diploma verenin yanlışlığını düşünüyorum.
- Eşkiya gibi davranışa devam edersen başın belaya girer.
- Kim oluyorsun da bana böyle konuşuyorsun?

DİLİ ALGILA	İNGİLİZCE	TÜRKÇE	NORVEÇÇE		TÜRKÇE	İNGİLİZCE	JAPONCA	
		<b>TÜRKÇE</b>				<b>İNGİLİZCE</b>		
		haysiyet yoksunu insanlar yüzünden belimizi doğrultamadık				we couldn't straighten our backs because of people with no dignity		
İLİ ALGILA	İNGİLİZCE	TÜRKÇE	NORVEÇÇE		TÜRKÇE	İNGİLİZCE	JAPONCA	
		we couldn't straighten our backs because of people with no dignity				onursuz insanlar yüzünden sırtımızı doğrultamadık		



# VERİ BÜYÜTME SONRASI SINIF DAĞILIMLARI



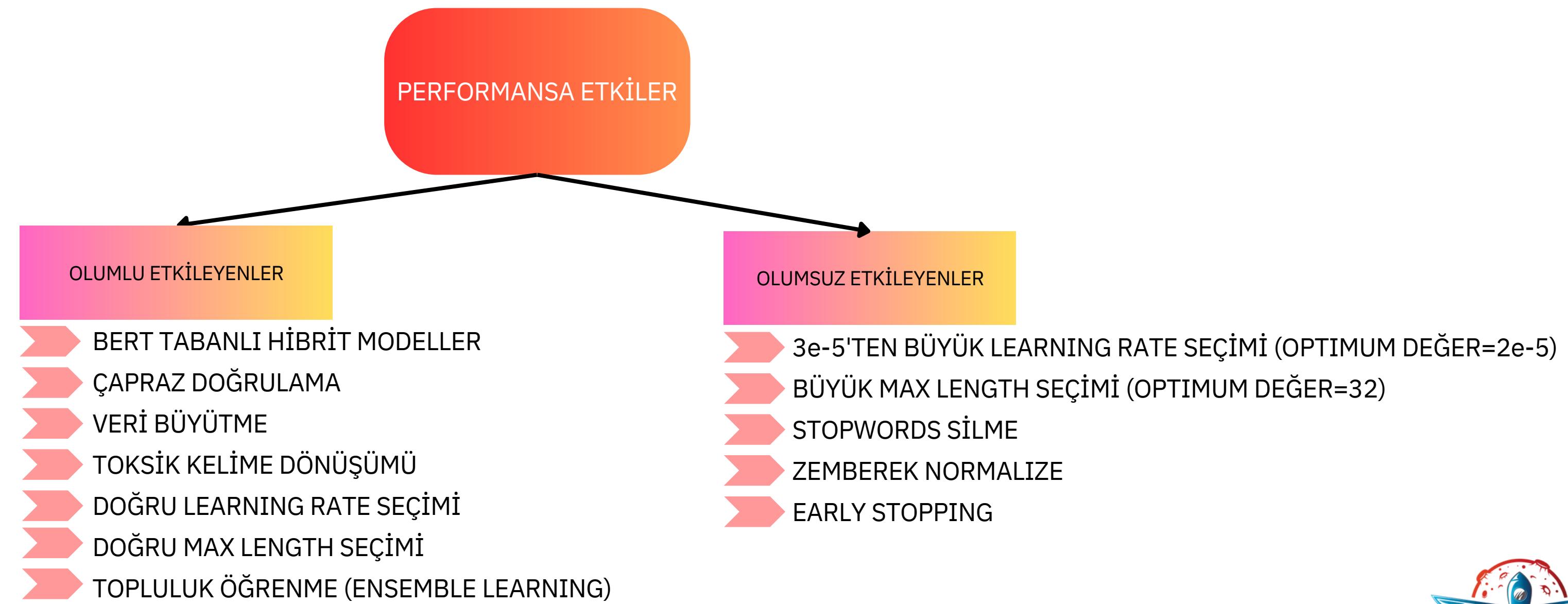


# VERİ BüYÜTME VE ÇAPRAZ DOĞRULAMA SONRASI MODELLERİN PERFORMANSLARI

Type	Model	PRECISON	RECALL	F1 MACRO	F1 MACRO / 5 CV
FastText+Temel Yöntem	Bi-GRU + Bi-LSTM	91.4% (+1.9)	91.4% (+2.1)	92.4% (+2)	-
	Bi-LSTM + Attention	88.2% (+1.9)	88% (1.9)	88.1% (+1.35)	-
	Bi-GRU + Bi-LSTM + CNN	91.8% (+2.4)	91.2% (1.6)	91.28% (1.68)	-
ConvBertTurk mc4+Temel Yöntem	Bi-GRU + Bi-LSTM	95.6% (+1.25)	95.3% (+0.7)	95.55% (+1.08)	96.74%
	Bi-LSTM + Attention	95.2% (+1.08)	95.25% (+0.9)	95.34% (+1.11)	96.64%
	Bi-GRU + Bi-LSTM + CNN	95.52% (+1.06)	95.4% (+0.85)	95.6% (+1.05)	96.72%

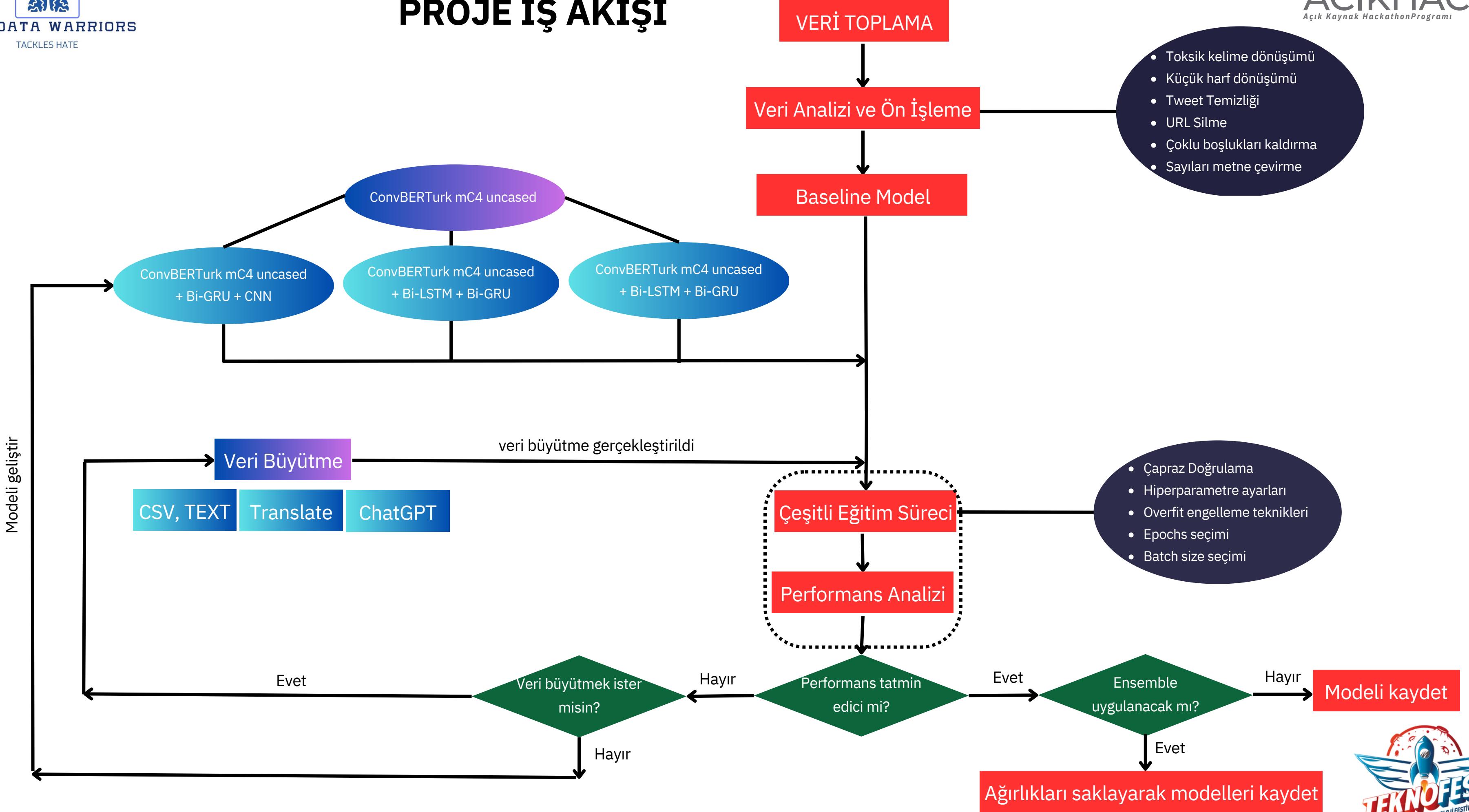
# Topluluk Öğrenme Yöntemi ve Performansa Etki Eden Unsurlar

Model	F1-MACRO	INSULT F1-MACRO	OTHER F1-MACRO	PROFANITY F1-MACRO	RACIST F1-MACRO	SEXIST F1-MACRO
Ensemble Model	<b>0.97 (+0.26)</b>	<b>0.931</b>	<b>0.982</b>	<b>0.981</b>	<b>0.982</b>	<b>0.97</b>





# PROJE İŞ AKIŞI





# DEMO VIDEO

The screenshot shows the Hugging Face Space interface for the project 'emirkocak/TRT\_Data\_Warriors\_tackling\_hate\_speech\_demo'. The interface includes a search bar, navigation links for Models, Datasets, Spaces, Docs, Solutions, Pricing, and user profile. Below the navigation is a banner indicating the space is 'Running'. The main area features two input fields: 'comment' on the left and 'output' on the right. Below these fields are two buttons: 'Clear' and 'Submit'. The 'Submit' button is highlighted in orange.

**GitHub Proje Sayfası:** <https://github.com/TRT-Data-Warriors/Tackling-Hate-Speech>

Hepsi bu  
kadardı!  
Teşekkürler

