TEKNOFEST
HAVACILIK, UZAY VE TEKNOLOJİ FESTİVALİ

DATA HACKERS

OĞUZHAN KIR

He graduated from Anadolu University, Open Education Faculty, Occupational Health and Safety with a grade of 3.10/4 in 2021, and from Kırklareli University, Faculty of Engineering, Department of Civil Engineering in 2022 with a grade of 3.68/4. He did his university internships at Kılıç İnşaat and Kırklareli OSB. He has been working as a Machine Learning Engineer at B2Metric since 27.06.2022.

ALİ OSMAN KAYA

He graduated from Marmara University, Faculty of Technology, Mechatronics Engineering with 2.73/4 in 2022. He did his internships at IQB Solutions for 2 months and then B2Metric for 5 months. He has been working as a Data Scientist at B2Metric since 05/2022.

ALİ OSMAN KAYA

- Araştırma Süreçleri
  - *Text mining süreçlerinin araştırılması*
  - *NLP literatürünün araştırılması*
  - *NLP State-of-art modellerin araştırılması*
- Ön İşlem Adımları
  - *Veri setinin temizlenmesi*
  - *Tokenization işlemi*
- Model Eğitim Mimarisi Kurulumu
  - *Model eğitim döngüsünün hazırlanması*
  - *Model eğitim süreci*

OĞUZHAN KIR

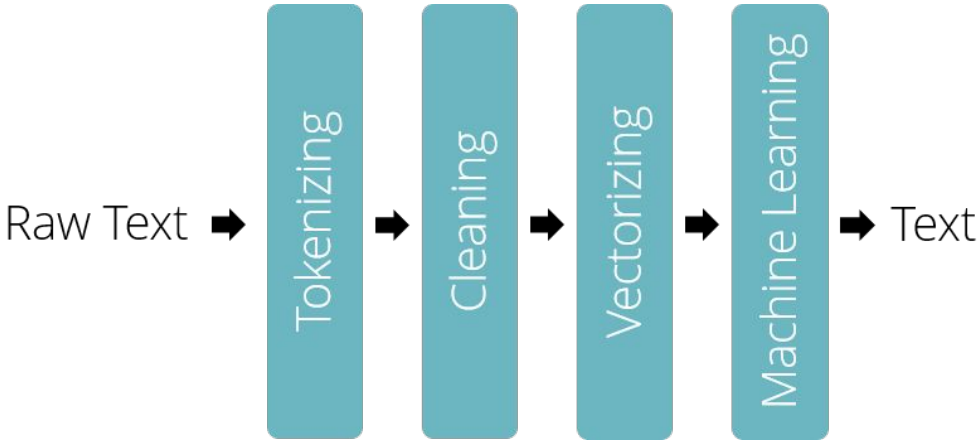
- Araştırma Süreçleri
  - *NLP literatürünün araştırılması*
  - *Açık kaynak model mimari ve kodların araştırılması*
- Model Geliştirme Yöntemleri
  - *Warm-up scheduler yöntemi*
  - *Stochastic Weight Averaging (SWA) scheduler yöntemi*
  - *Frequent Evaluation*
- Inference Kurulumu
  - *Tahmin sonuçlarının efektif bir şekilde istenilen formata getirilmesi*

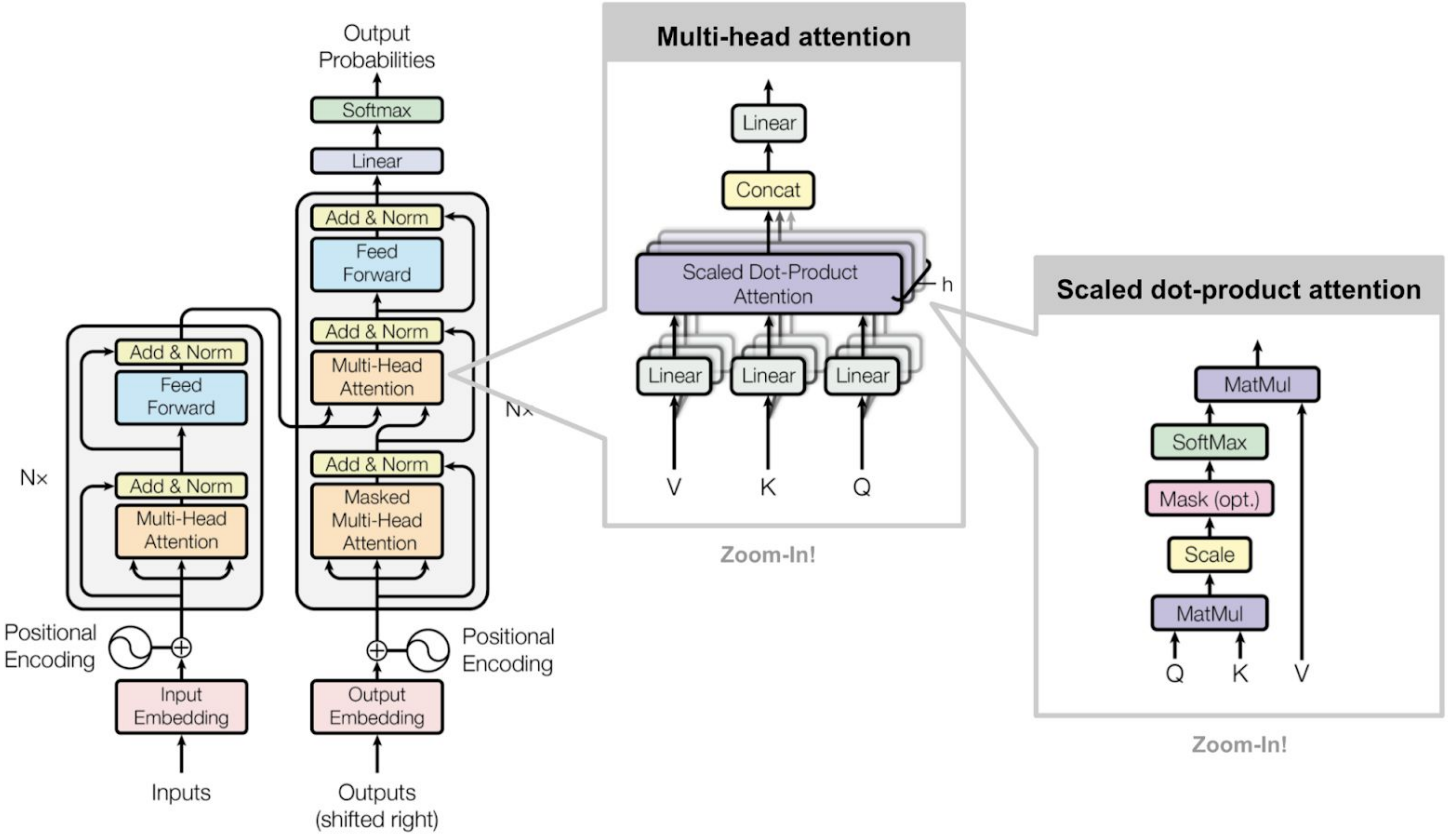
Problemimiz;

Dijital ortamlarda, metinlerde yer alan aşağılayıcı söylemleri tespit etmekte insan gücünün yetersiz kalması.

Buna bağlı olarak, aşağılayıcı söylemin türünün (Cinsiyetçi, ırkçı, küfür, hakaret) tespit edilmemesi üzerine uygun tedbirlerin alınamaması.

Yapay zekanın bir alt kategorisi olan Natural Language Processing (NLP) yöntemlerinin kullanılarak, metinlerde yer alan aşağılayıcı söylemlerin tespit edilmesi ve tespit edilen söylemlerin türlerinin (Cinsiyetçi, ırkçı, küfür, hakaret) belirlenmesi.

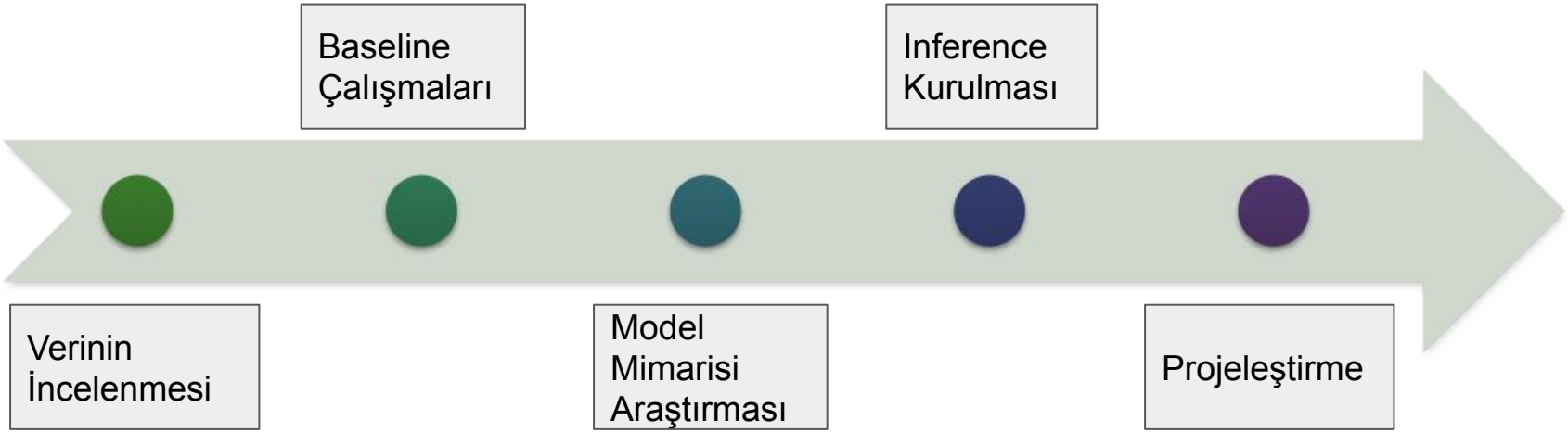



Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
Multi-head attention
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q
Zoom-In!
Scaled dot-product attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V
Zoom-In!

Teknik Çalışmalar ve Denemeler

- Baseline Modelleme Çalışmaları
  - Logistic Regression
  - Transformer
- Transformer + CatBoost
- Transformer Model Mimarileri
  - Bert
  - ConvBert
  - DistilBert
  - Electra
- Parametre Optimizasyonu
  - max_length
  - epochs
  - batch_size
  - learning_rate
- Stratified KFold Validation
- Modelleme Teknikleri
  - Layer-wise Learning Rate Decay (LLRD)
  - Warm-up Steps
  - Re-initializing Pre-trained Layers
  - Stochastic Weight Averaging (SWA)

Proje İş Akışı



Proje Yol Haritası

- Proje eksiklerinin kapatılması
- Skor arttırmak için diğer çözümlerden eklentiler yapılması
- Başka Türkçe doğal dil işleme probleminde denenmesi

TEŞEKKÜRLER

DATA HACKERS

- Ali Osman Kaya
- Oğuzhan Kır

https://github.com/Data-Hackers-Team/Teknofest2023