



# G4T0R2-NLP



BİLİŞİM  
VADİSİ



# İçindekiler

|                              |    |
|------------------------------|----|
| Ekip Tanıtımı .....          | 1  |
| Problem ve Amaç.....         | 2  |
| Literatür Taraması.....      | 3  |
| Kullanılan Veri Setleri..... | 6  |
| Modeller.....                | 10 |
| Gerçek Zamanlı Çalışma ..... | 16 |
| Gelecek Hedefleri .....      | 19 |
| Kaynakça .....               | 21 |

# Ekip Üyeleri



**Şengül BAYRAK**  
Danışman



**Ferhat TOSON**  
Veri Çekme  
Veri Ön İşleme



**Alper KARACA (K)**  
Model Geliştirme  
API Geliştirme



**Selçuk YAVAŞ**  
Veri Çekme  
Veri Ön İşleme



**Mehmet Emin Tayfur**  
API Geliştirme  
Web Tasarım

# Problem Ne?

Bu proje, "Entity Bazlı Duygu Analizi Yarışması" çerçevesinde müşteri geri bildirimlerini analiz ederek, hizmet veya ürün özellikleriyle ilgili duyguları sınıflandırmayı amaçlamaktadır. İlk aşamada yorumları doğru entity'lere atfetmek, ikinci aşamada ise bu entity'lerle ilgili duyguları (olumlu, olumsuz veya nötr) sınıflandırmak için yöntemler geliştirilmiştir. Metin madenciliği ve doğal dil işleme (NLP) teknikleri kullanılarak, müşteri memnuniyetini artırmak ve ürün/hizmet geliştirme süreçlerini optimize etmek hedeflenmektedir.





# Literatür Taraması

# Literatür

2016

Pontiki vd.

Pontiki vd. (2016), SemEval-2016 ABSA görevinde 8 dilde ve 7 farklı alanda duygusal analizi için 19 eğitim ve 20 test veri seti kullanmıştır. Kullanılan modeller arasında SVM, RNN, CNN ve Naive Bayes yer almış, SVM yüksek performans sergilerken, RNN ve CNN karmaşık dil yapılarında başarılı olmuştur. Ancak, modellerin spesifik doğruluk oranları belirtilmemiştir.

2018

Schüller vd.

Schüller vd. (2018) çalışmasında, Türkçe'de özne tespiti için kural tabanlı bir sistem ve coreference çözümlemesi için SVM kullanılarak METU-Sabancı Türkçe Ağaçbank ve Marmara Türkçe Özneleştirmeye veri setleri kullanılmıştır. Altın öznelerde SVC modeliyle %57.8 LEA skoru, tahmin edilen öznelerde SVR modelleriyle %53.6 LEA skoru elde edilmiştir. En iyi performans, SVC modelinin altın öznelerde sağladığı %57.8 LEA skoru ile elde edilmiştir.

2019

Erşahin vd.

Erşahin vd. (2019), Türkçe duygusal analizi için leksikon ve makine öğrenimi yöntemlerini birleştirerek doğruluğu artırmayı hedefleyen bir hibrit yaklaşım sunmaktadır. Naive Bayes (NB), SVM ve J48 modelleri kullanılarak Film, Otel ve Twitter veri setlerinde test edilmiştir. Hibrit yöntem, NB ile %88.93, %89.98 ve %83.37; SVM ile %86.31, %91.96 ve %81.83; J48 ile %77.92, %88.96 ve %72.72 doğruluk oranlarıyla mevcut yöntemlerden daha yüksek performans göstermiştir.

2020

Bardak vd.

Bardak vd. (2020), Türkçe sentiment analizi için iki yaklaşım önerilmiştir: BERT'in çok dilli modelini ince ayar yapma, Türkçe metinleri İngilizceye çevirip BERT'in ana modelini kullanma. Deneyler, olumlu veya olumsuz etiketlenmiş Türkçe film ve otel yorumları veri setlerinde gerçekleştirilmiş ve BERT modellerimiz mevcut yöntemleri aşarak yüksek doğruluk skorlarına ulaşmıştır.

2021

Salur vd.

Salur vd. (2021), Türkçe metinlerden konu terimlerini çıkarmak için çeşitli yöntemler kullanmıştır. SemEval ABSA 2016 Türk restoran veri seti üzerinde TF-IDF, LDA, NMF ve kural tabanlı yaklaşımları değerlendirderek, TF-IDF modelinin en iyi sonuçları verdiği belirlemiştir. Önerilen topluluk yöntemi Union-I stratejisiyle %61.08 F1 skoru elde etmiştir.

# Literatür

2022

Özçelik vd.

Özçelik vd. (2022), Türkçe adlandırılmış varlık tanıma (NER) çalışmasında, Transformer tabanlı dil modelleri en yüksek performansı göstererek tweetlerde %80.8 ve haber makalelerinde %96.1 F1 skoru elde etmiştir. Uzun varlıklarda ve sosyal medyada tüm modellerin performansı düşerken, esnek kelime sırası uygulandığında iyi yazılmış metinlerde %12, gürültülü metinlerde ise %7 performans kaybı gözlemlenmiştir.

2022

Dinç vd.

Dinç vd. (2022), Named Entity Recognition (NER) problemini ele alarak, derin öğrenme tabanlı modelleri farklı Türkçe veri setleriyle test etmektedir. Finans haberlerinden iki yeni anotasyonlu veri seti oluşturulmuş, BIO şeması ve ham etiketler kullanılarak anotasyon formatının performansı etkisi gözlemlenmiştir.

2023

Demir

Demir (2023), Marmara Türkçe Anafor Korpusu veri seti ile anafor çözümlemesinde iki modeli karşılaştırmıştır. Model\_1, klasik makine öğrenimi tekniklerini kullanırken, Model\_2 BERTurk, DistilBERT ve ELECTRA gibi bağlam bağımlı yöntemlere dayanmaktadır. BERTurk-128k-uncased ile Model\_2 en yüksek performansı göstererek MUC'da 0.697, B3'te 0.574 ve CEAFF-e'de 0.508 F1 skorlarına ulaşmıştır, bu da bağlam bağımlı yöntemlerin daha etkili olduğunu göstermektedir.

2024

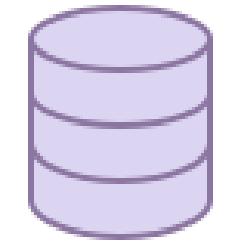
Arslan vd.

Arslan vd. (2024), BiLSTM-CRF modeli, FastText gömme vektörleri ile unstructured Türkçe metinlerde ürün isimlerini tespit ederken %57.40 F1 skoru, %55.78 hassasiyet ve %59.12 geri çağrıma oranı skoru elde etmiştir.

2024

Altınel vd.

Altınel vd. (2024), Türkçe nefret söylemini tespit etmek için BERT ve k-means+textGCN sınıflandırıcı gibi yenilikçi yöntemler geliştirilmiş ve Term Frequency, Word2Vec, Doc2Vec, and GloVe. Additionally, vektör temsili teknikleri ile öğrenme algoritmaları kullanılmıştır. Üç Türkçe nefret söylemi veri setinde %87.81 F1 skoru elde edilmiştir



# Kullanılan Veri Setleri

# Veri Çekilen Platformlar



**2.000 > 1.950**



**App Store**

**13.930 > 12.600**



**Google Play**

**841.240 > 132.465**

**kaggle**



**10.700 > 10.500**

\* Çekilen verilerden 157.515 veri seçilerek model eğitimi için kullanılmıştır.

# Veri Seti Düzeni

- RID Sütunu:** Çekilen metine ait eşsiz anahtarı (primary key) belirtir.
- SID Sütunu:** RID'e ait metindeki kaçinci cümle olduğunu belirtir.
- APP Sütunu:** Çekilen metinin kaynağını belirtir.
- Review Sütunu:** Yapılan yorumdaki cümleyi belirtir.
- Aspect Sütunu:** Cümledeki "Entity" belirtir.
- Sentiment Sütunu:** Cümledeki "Entity" ait duyguyu belirtir.

| RID | SID | APP     | REVIEW   | ASPECT       | SENTIMENT |
|-----|-----|---------|--|--------------|-----------|
| 1   | 1   | TV+     | Uygulamada kartlarda Troy kart geçmiyor.             | Troy         | Negatif   |
| 2   | 1   | BiP     | Tam 10 yıldır Turkcell kullanıyorum.                 | Turkcell     | Nötr      |
| 3   | 1   | Dijital | Türk Telekom çekim kalitesi çok iyi, tavsiye ederim. | Türk Telekom | Pozitif   |

\* Çekilen verilerdeki kelimelerin 2.085.057'i “O”, 155.754'ü “B-A” ve 6.466'sı “B-I” sınıfına aittir.  
85.690 adet “olumsuz”, 8.195 adet “nötr” ve 61.489 adet “olumlu” etikete sahiptir.

# Veri Ön İşleme Adımları

Ön işleme adımlarında sırasıyla :

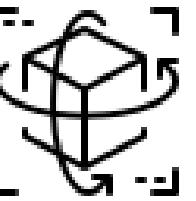
1. Metin içindeki linkler temizlenmiştir.
2. Metin içindeki mail adresleri temizlenmiştir.
3. Metin içindeki rakamlar temizlenmiştir.
4. Metin içerisindeki emojiler temizlenmiştir.
5. Metin içerisindeki HTML etiketleri temizlenmiştir.
6. Metindeki etkisiz kelimeler temizlenmiştir.
7. Ek olarak yazılan da/da temizlenmiştir.
8. Soru ekleri temizlenmiştir.
9. Tek harfden oluşan hatalı kelimeler temizlenmiştir.

**Örneğin;**

*'Uygulamada kartlarda Troy kart geçmiyor. Lütfen uygulamaya Troy kartı da ekleyin 😊😊 .*

**Ön işleme sonrası;**

*'Uygulamada kartlarda Troy kart geçmiyor. uygulamaya Troy kartı ekleyin.*

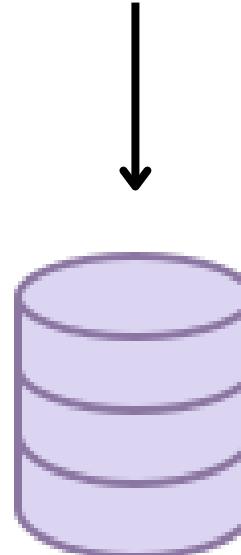


**Modeller**

# Model Eğitim Yöntemleri

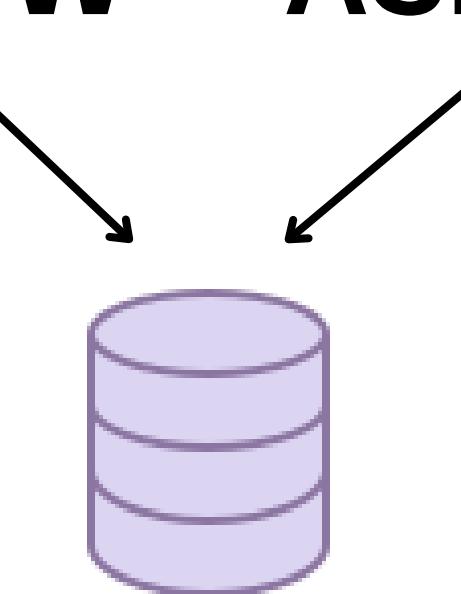
Cümle Tabanlı

**ASPECT + REVIEW**



Kelime Tabanlı

**REVIEW      ASPECT**



# Model Mimarileri

| Model Adı   | Model Tipi    | F1-Makro |
|---|---------------|----------|
| ConvBERT (dbmdz/convbert-base-turkish-cased)              | Cümle Tabanlı | %99.87   |
| BERT (dbmdz/bert-base-turkish-cased) (32k version)        | Cümle Tabanlı | %99.71   |
| BERT (dbmdz/bert-base-turkish-128k-cased)                 | Cümle Tabanlı | %99.57   |
| DeBERTa (microsoft/mdeberta-v3-base)                      | Cümle Tabanlı | %97.39   |
| ELECTRA (dbmdz/electra-base-turkish-cased-discriminator)  | Cümle Tabanlı | %97.28   |
| DistilBERT (dbmdz/distilbert-base-turkish-cased)          | Cümle Tabanlı | %97.22   |
| CRF   | Cümle Tabanlı | %96.36   |
| ELECTRA (dbmdz/electra-small-turkish-cased-discriminator) | Cümle Tabanlı | %94.54   |
| Bi-GRU + CRF  | Cümle Tabanlı | %91.21   |
| Bi-LSTM + CRF   | Cümle Tabanlı | %74.75   |

\* ‘Entity Extraction’ problemi içindir.

# ConvBERT Modeli

| CLASSIFICATION REPORT                        | O          | I-A    | B-A           |
|--|------------|--------|---------------|
| DEBERTA (MICROSOFT/MDEBERTA-V3-BASE)         | 1.0        | 0.93   | 1.0           |
| BERT (DBMDZ/BERT-BASE-TURKISH-CASED)         | 1.0        | 0.93   | 1.0           |
| CONVBERT (DBMDZ/CONVBERT-BASE-TURKISH-CASED) | 1.0        | 0.97   | 1.0           |
| MAX LENGTH                                   | BATCH SIZE | EPOCHS | LEARNING RATE |
| 128  | 32         | 2      | 2E-05         |

NVIDIA T4 (16 GB VRAM) ve CUDA 12.1

Eğitim Süresi = 39 dakika 18 saniye Sınama Süresi = 41 saniye

# Model Mimarileri

| Model Adı                                 | Model Tipi     | F1-Makro |
|---|----------------|----------|
| BERT (dbmdz/bert-base-turkish-128k-cased) | Kelime Tabanlı | %84.90   |
| MultinomialNaiveBayes + CountVectorizer   | Cümle Tabanlı  | %79.52   |
| MultinomialNaiveBayes + TFIDF             | Cümle Tabanlı  | %76.61   |
| BERT (dbmdz/bert-base-turkish-cased)      | Kelime Tabanlı | %76.39   |
| Bi-LSTM (Two Input)                       | Kelime Tabanlı | %76.25   |
| LSTM                                      | Cümle Tabanlı  | %75.02   |
| LSTM (Two Input)                          | Kelime Tabanlı | %74.84   |
| Attention                                 | Cümle Tabanlı  | %74.56   |
| GRU (Two Input)                           | Kelime Tabanlı | %74.56   |
| Bi-GRU (Two Input)                        | Kelime Tabanlı | %74.52   |

\* ‘Sentiment Analysis’ problemi içindir.

# BERT Modeli

| CLASSIFICATION REPORT                     | NEGATIVE   | NEUTRAL | POSITIVE      |
|---|------------|---------|---------------|
| Bi-LSTM (TWO INPUT)                       | 0.89       | 0.56    | 0.84          |
| MULTINOMİAL NAİVE BAYES + COUNTVECTORİZER | 0.84       | 0.33    | 0.75          |
| BERT (DBMDZ/BERT-BASE-TURKİSH-128K-CASED) | 0.94       | 0.70    | 0.91          |
| MAX LENGTH                                | BATCH SIZE | EPOCHS  | LEARNING RATE |
| 128                                       | 32         | 4       | 2E-5          |

NVIDIA T4 (16 GB VRAM) ve CUDA 12.1

Eğitim Süresi = 103 dakika 45 saniye Sınama Süresi = 1 dakika 12 saniye



# **Gerçek Zamanlı Çalışma**

# Arayüz

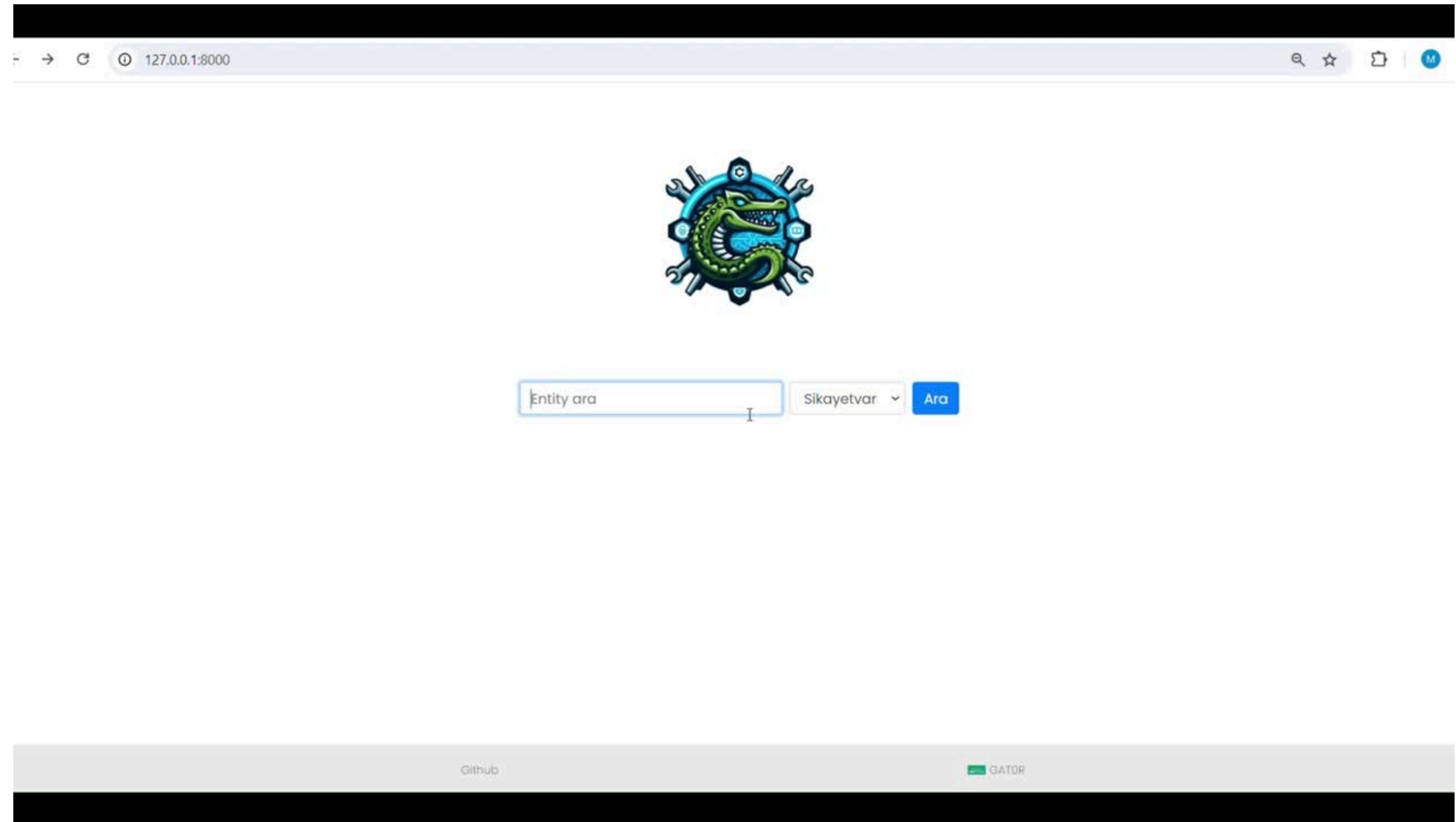
## Comment Analyzer



turkcell güzel bir iş çıkarmış ve spotifya çok benzeyen fizy uygulamasını yapmış. keşke paycell uygulamasına da gereken önemi verse ve uygulama sürekli çökmesse

Analyze

| Logo  | Varlık   | Duygu   |
|---|----------|---------|
|  | turkcell | olumlu  |
|  | spotifya | nötr    |
|  | fizy     | nötr    |
|  | paycell  | olumsuz |



Video Link



# **Gelecek Hedefleri**

# GELECEK HEDEFLERİ

1

Daha nitelikli ve temiz  
bir veri seti elde etmek.

2

Şirket istek ve  
önceliklerine göre bir  
model geliştirmek

3

Öneri ve şikayetleri ilgili  
departmana yönlendiren  
bir uygulama yapılması



# Kaynakça

# Kaynakça

- [1] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos & Eryiğit, G. (2016). Semeval-2016 task 5: Aspect based sentiment analysis.
- [2] S. ÖZEKES and E. N. KARAKOÇ, “Makine öğrenmesi yöntemleriyle anormal ağ trafiğinin tespit edilmesi,” Düzce Üniversitesi Bilim ve Teknoloji Dergisi, vol. 7, no. 1, pp. 566–576, 2019.
- [3] Erşahin, B., Aktaş, Ö., Kılıç, D., & Erşahin, M. (2019). A hybrid sentiment analysis method for Turkish. Turkish Journal of Electrical Engineering and Computer Sciences, 27(3), 1780-1793.
- [4] Acikalin, U. U., Bardak, B., & Kutlu, M. (2020, October). Turkish sentiment analysis using bert. In 2020 28th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.
- [5] Salur, M. U., Aydın, İ., & Jamous, M. (2022). An ensemble approach for aspect term extraction in Turkish texts. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 28(5), 769-776.
- [6] Ozcelik, O., & Toraman, C. (2022). Named entity recognition in Turkish: A comparative study with detailed error analysis. Information Processing & Management, 59(6), 103065.
- [7] Dinç, D. (2022). Financial named entity recognition for Turkish news texts (Master's thesis, Middle East Technical University).
- [8] Demir, Ş. (2023). Neural coreference resolution for Turkish. Journal of Intelligent Systems: Theory and Applications, 6(1), 85-95.  
ISO 690
- [9] Arslan, S. (2024). Application of BiLSTM-CRF model with different embeddings for product name extraction in unstructured Turkish text. Neural Computing and Applications, 36(15), 8371-8382.
- [10] Altinel, A. B., Sahin, S., Gurbuz, M. Z., & Baydogmus, G. K. (2024). So-haTRed: A novel hybrid system for Turkish Hate Speech Detection in Social Media with Ensemble Deep Learning improved by BERT and clustered-Graph Networks. IEEE Access.

Dinlediğiniz  
için  
Teşekkürler !

