Haber Makalelerinin Sınıflandırılması, Özetlenmesi ve Genişletilmesi

Patika.dev & New Mind Bootcamp Final Case

Merve TUTAR

Proje Tanıtımı

- Bu proje, haber makalelerini kategorize etme, özetleme ve metin üretme konularında NLP ve makine öğrenmesi tekniklerini birleştiren entegre bir sistem amaçlamaktadır.

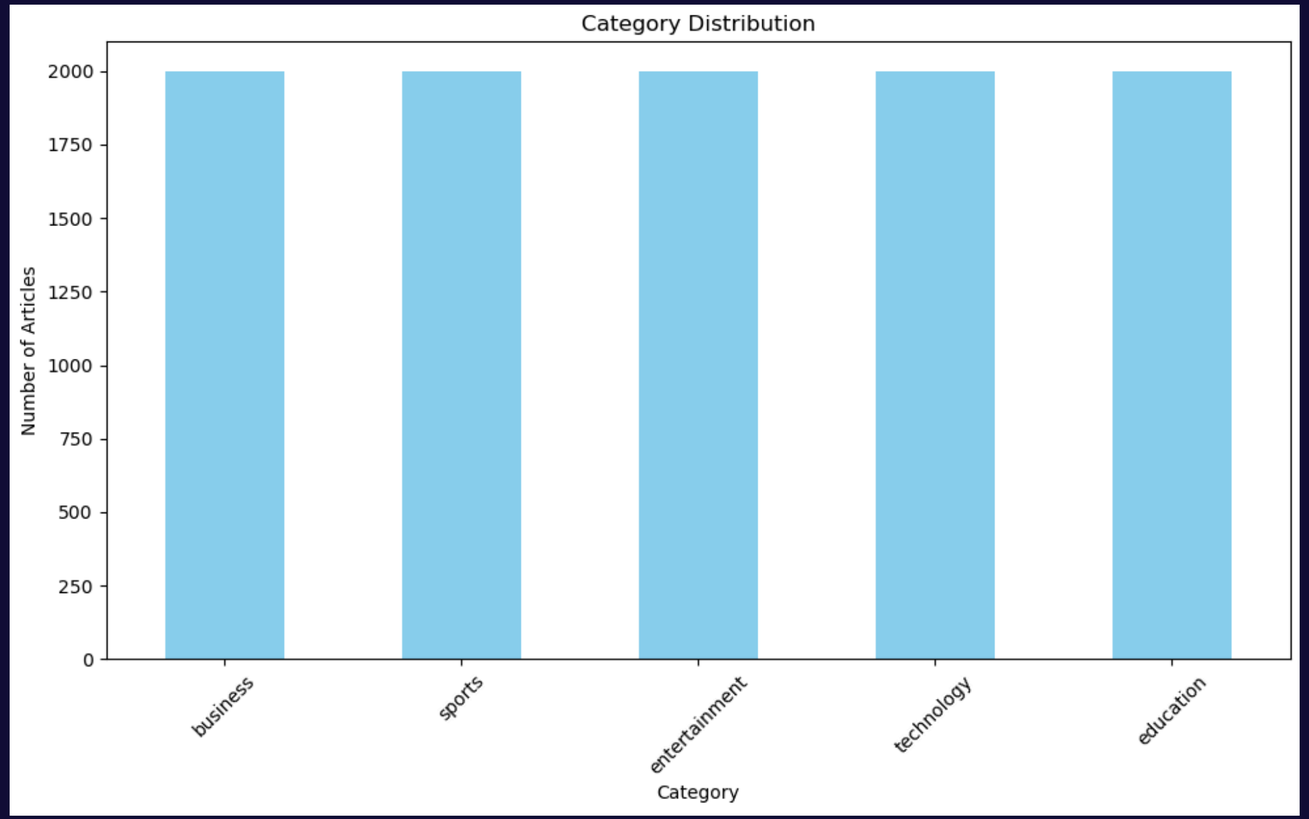
- Veri Kaynağı: Bu veri seti , İş, Teknoloji, Spor, Eğitim ve Eğlence gibi çeşitli alanlara yayılan kapsamlı bir haber makaleleri koleksiyonu sunmaktadır . Veriler, ünlü haber dergisi "The Indian Express" e ait olup Kaggle'dan alınmıştır.

- Veri Kategorileri:
- Business (İş Dünyası)
- Technology (Teknoloji)
- Sports (Spor)
- Education (Eğitim)
- Entertainment (Eğlence)

Veri Seti Özellikleri

Veriler, her haber kategorisi için başlık ve içerik bilgisi içermektedir. Toplamda 10.000 haber makalesi bulunmaktadır.

Manşetler: Haberin başlığı veya başlığı.
Tanım: Haber yazısının kısa özeti veya açıklaması.
İçerik: Haberin tam metin içeriği.
URL: Haberin orijinal kaynağına giden URL bağlantısı.
Kategori: Haber makalesinin kategorisi veya konusu (örneğin iş, eğitim, eğlence, spor, teknoloji).



Veri Analizi ve Ön İşleme

Veri, her kategori için ayrı dosyalardan yüklenerek tek bir veri çerçevesinde birleştirilmiştir. Eksik veri kontrolü ve veri dağılımı analizi yapılmıştır.

- Veri Ön İşleme:
  - Lowercasing: Tüm metinler küçük harfe dönüştürülmüştür.
  - Punctuation Removal: Noktalama işaretleri kaldırılmıştır.
  - Number Removal: Sayılar metinlerden çıkarılmıştır.
  - Stopwords Removal: Anlam taşımayan kelimeler (stopwords) temizlenmiştir.
  - Lemmatization: Kelimeler köklerine indirgenmiştir.
- Sonuç: Temizlenmiş metinler yeni bir sütunda (cleaned content) saklanmıştır.

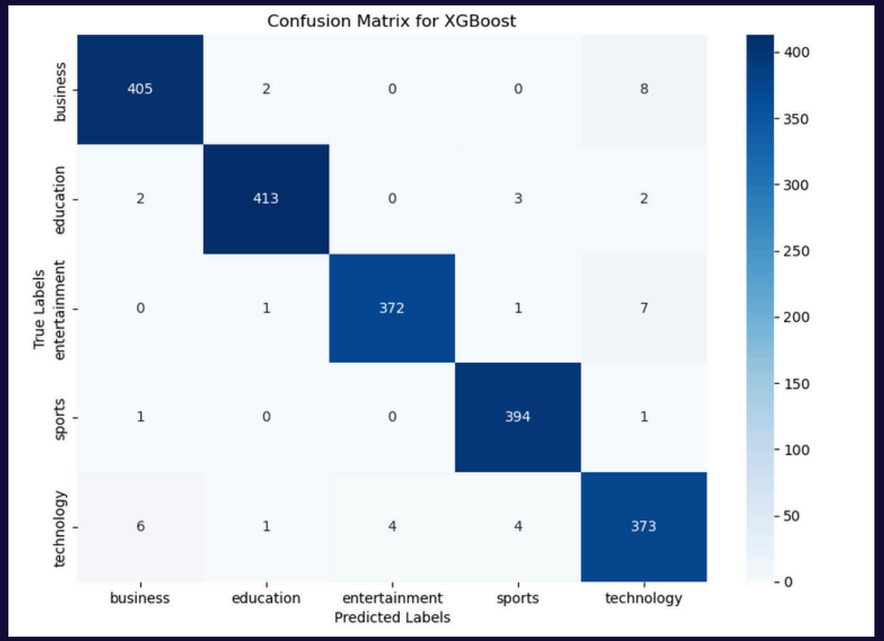
	content	cleaned content
0	Asus, Dell, HP and Foxconn are among 38 compan...	asus dell hp foxconn among company submitted a...
1	Salman Khan's Tiger 3 has worked majorly in fa...	salman khan tiger worked majorly favour bollyw...
2	Former Australian batsman Mike Hussey urged th...	former australian batsman mike hussey urged au...
3	It was fitting that the match schedule at the ...	fitting match schedule india open super h pran...
4	The National Medical Commission (NMC) has issu...	national medical commission nmc issued draft m...
5	A film on the Indian Army has to ensure that e...	film indian army ensure everything medal creas...
6	The Indian Institute of Management (IIM) Luckn...	indian institute management iim lucknow collab...
7	Shreyas Iyer is pleased with the way he has pr...	shreyas iyer pleased way prepared upcoming ser...
8	Top Universities in India 2023: The National I...	top university india national institutional ra...
9	The National Institute of Technology Rourkela ...	national institute technology rourkela nit rou...

Modelleme ve Değerlendirme

Modelleme

- Model Seçimi: Haber kategorilerini tahmin etmek için çeşitli modeller denenmiş olup XGBoost sınıflandırıcı modeli ile devam edilmiştir.
- Veri Bölme: Eğitim ve test verisi olarak %80 ve %20 oranında bölünmüştür.
- Modelin Eğitilmesi:
  - TF-IDF vektörizasyonu ile metin verileri sayısal hale getirilmiştir.
  - XGBoost sınıflandırıcı ile model eğitilmiş ve sonuçlar kaydedilmiştir.
- Model Performans Değerlendirmesi:
  - Accuracy, Precision, Recall, F1 Score gibi metriklerle model performansı ölçülmüştür.

	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9875	0.987568	0.9875	0.987492
Random Forest	0.9770	0.977149	0.9770	0.977018
XGBoost	0.9785	0.978577	0.9785	0.978514
SVM (Support Vector Machine)	0.9870	0.987027	0.9870	0.986992
Gradient Boosting	0.9735	0.973687	0.9735	0.973538



Metin Özetleme ve Genişletme

Özetleme için BART (Facebook) modeli, metin genişletme için DistilGPT-2 modeli kullanılmıştır. ROUGE ve BLEU metrikleri ile özetlerin kalitesi ölçülmüştür.

BART

Kategorilere göre özetler çıkarılmıştır.

DistilGPT-2

Metinleri daha detaylı bir şekilde genişletir.

Bu modelleri daha etkili hale getirmek için her kategoriye özgü prompt (yönlendirme) kullanılmıştır.

Business için prompt:

category_prompts = { 'business': "Summarize the key points from contents related to business, including key trends in trade, companies, and economics.", ... }

ROUGE ve BLEU Metrikleri:

ROUGE: Metnin referans metinle benzerliğini ölçer.
BLEU: Otomatik özetlerin doğruluğunu değerlendirmek için kullanılır.

Streamlit Uygulaması

Görselleştirme:
Kullanıcılar, metinlerini girerek haber kategorisini tahmin edebilir.

Tahmin edilen kategoriye ait özet bilgi ve kullanıcının girdiği habere ait genişletilmiş metin görüntülenir.

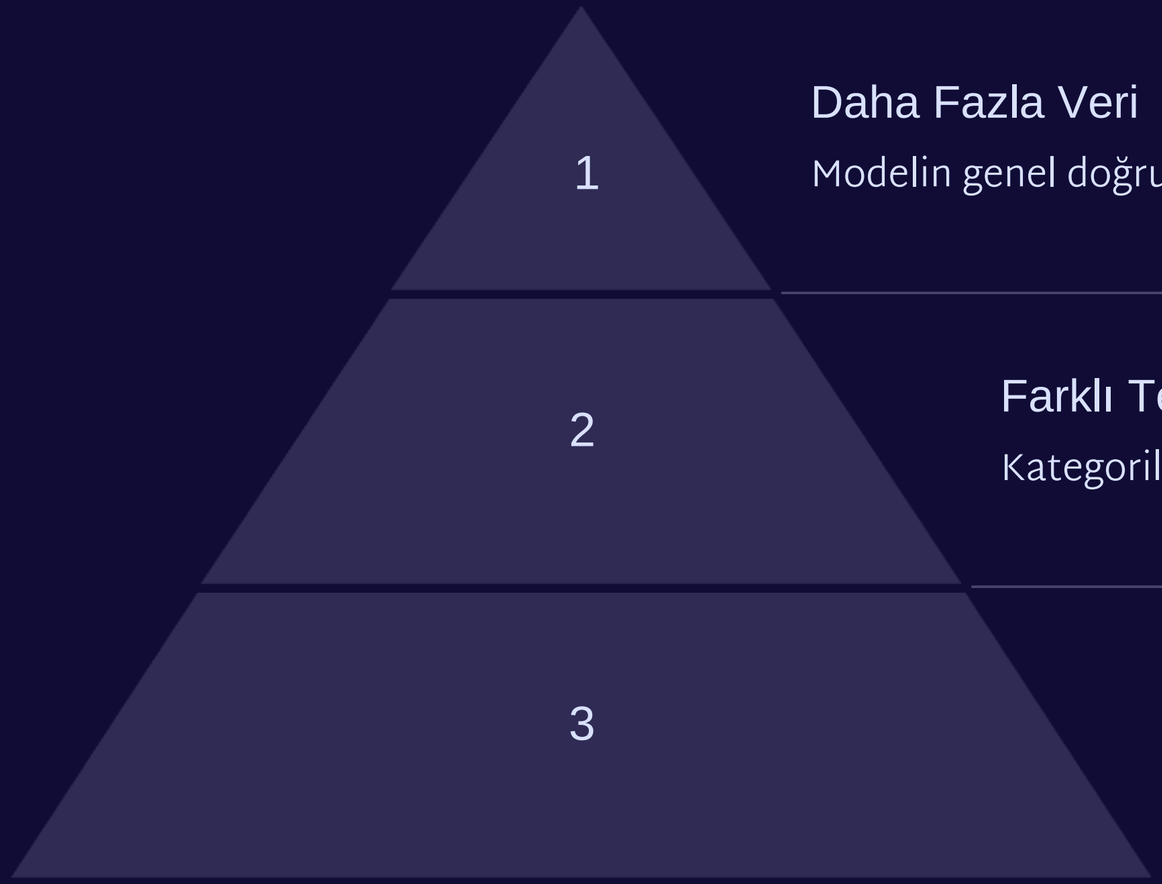
Modelin performans metrikleri (accuracy, precision, recall, vb.) ve ROUGE/ BLEU skorları da kullanıcıya sunulmaktadır.

Streamlit Uygulaması

Kullanıcılar, Streamlit uygulaması aracılığıyla kendi metinlerini girerek haber kategorisini tahmin edebilir, özetleyebilir ve haberi genişletebilirler.

Sonuçlar ve Gelecek Çalışmalar

Model başarıyla haber makalelerini 5 kategoriye ayırarak sınıflandırabilmektedir. Ancak sadece belirli bir bölgeden alınmış haberler olduğu için bazı konularda verinin yetersiz olduğu görülmüştür.



Daha Fazla Veri

Modelin genel doğruluğunu artırma

Farklı Teknikler

Kategorileri özetlemenin dinamik bir şekilde yapılması

Model Optimizasyonu

Farklı algoritmalar ve hiperparametre ayarları ile modelin performansını artırma