<SERBEST KATEGORİ>

TÜRKÇE DOĞAL DİL İŞLEME YARIŞMASI

Phorcys

8 - 9 Ağustos 2024

<EKİBİMİZ>

İbrahim Ünlü: Edebiyat Öğretmeni-Danışman

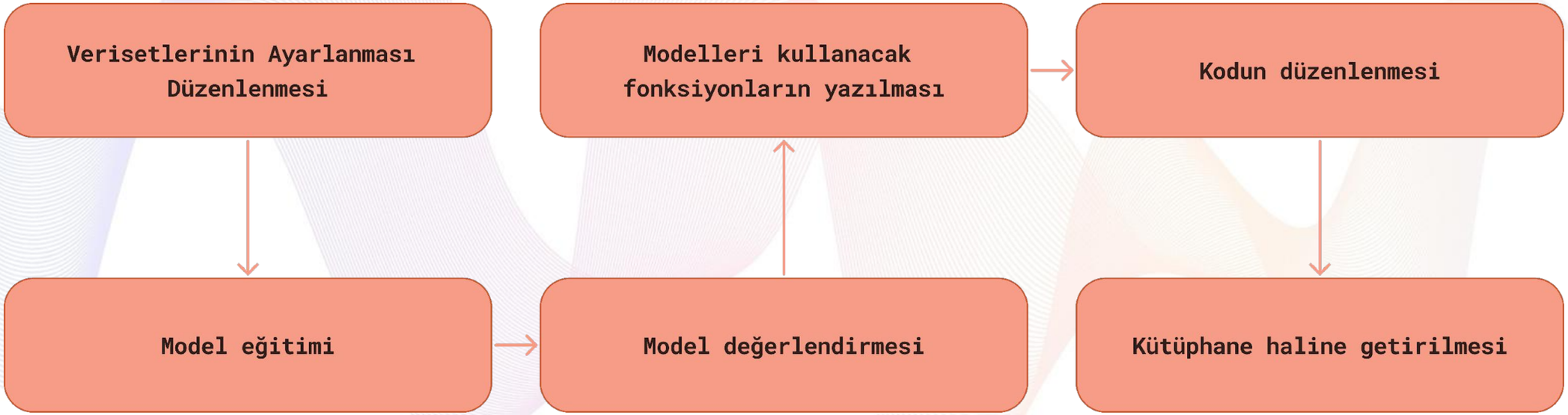
Mesut Aktaş: Proje Geliştirilmesi

<PROJENİN TANIMI>

Projemizin temel amacı herkesin kullanabilmesi için doğal dil işleme yöntemlerini kolaylaştırmaktır. Bunun için basit, anlaşılır ve kolay araçlar oluşturduk.

<PROJENİN SAĞLADIĞI ÇÖZÜM>

<PROJE İŞ AKIŞI>



<VERİ SETİ>

Opus Türkçe Veri Seti

TDD Milliyet Ner Veri Seti

Nane Limon Zorbalık veri seti

Kendi derlediğimiz veri setleri

Diğer veri setleri ve web scrapingle elde edilen veriler

<YÖNTEM VE TEKNİKLER>

Whisper ile text2speech uygulmamızı yazdık.

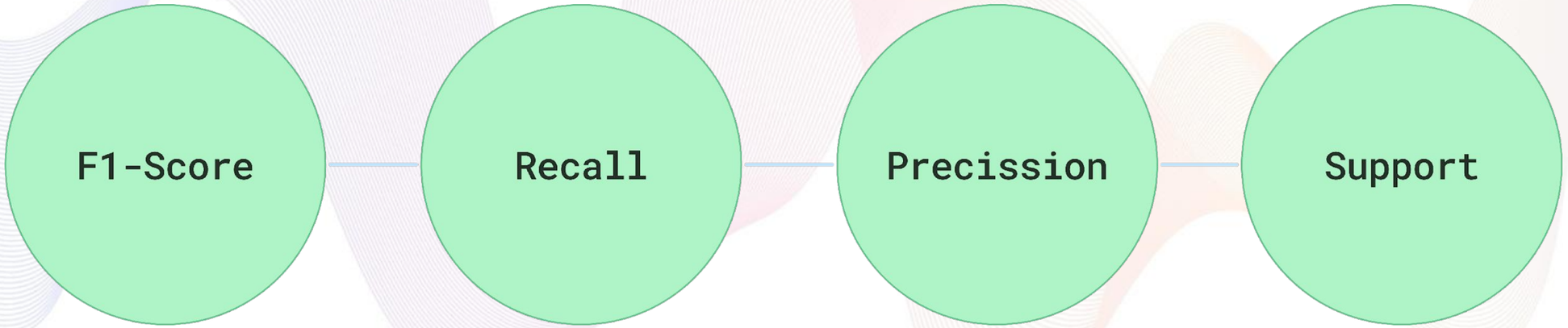
Veri Setlerinden anlamsız kelimeleri kaldırdık ve önemli kelimeleri öncelikli hale getirdik.

img2text modelleri ile Resimleri yorumlamayı sağladık,

Bert türkçe modeli ile modellerimizi eğittik.

Ollama ile LLM modellerini entegre ettik

<MODEL EĞİTİMİ VE DEĞERLENDİRME>



Class Name	Precision	1-Precision	Recall	1-Recall	f1-score
art	1.0000	0.0000	1.0000	0.0000	1.0000
economy	0.9947	0.0053	0.9947	0.0053	0.9947
health	1.0000	0.0000	1.0000	0.0000	1.0000
life	1.0000	0.0000	0.9930	0.0070	0.9965
magazine	0.9933	0.0067	1.0000	0.0000	0.9966
politics	0.9852	0.0148	0.9925	0.0075	0.9888
sport	1.0000	0.0000	1.0000	0.0000	1.0000
technology	1.0000	0.0000	0.9911	0.0089	0.9955
Accuracy	0.9965				
Misclassification Rate	0.0035				
Macro-F1	0.9965				
Weighted-F1	0.9965				

Training Set									
TARGET / OUTPUT	art	economy	health	life	magazine	politics	sport	technology	SUM
art	151 13.13%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	151 100.00% 0.00%
economy	0 0.00%	189 16.43%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	1 0.09%	190 99.47% 0.53%
health	0 0.00%	0 0.00%	132 11.48%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	132 100.00% 0.00%
life	0 0.00%	0 0.00%	0 0.00%	142 12.35%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	142 100.00% 0.00%
magazine	0 0.00%	0 0.00%	0 0.00%	0 0.00%	148 12.87%	1 0.09%	0 0.00%	0 0.00%	149 99.33% 0.67%
politics	0 0.00%	1 0.09%	0 0.00%	1 0.09%	0 0.00%	133 11.57%	0 0.00%	0 0.00%	135 98.52% 1.48%
sport	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	140 12.17%	0 0.00%	140 100.00% 0.00%
technology	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	111 9.65%	111 100.00% 0.00%
SUM	151 100.00% 0.00%	190 99.47% 0.53%	132 100.00% 0.00%	143 99.30% 0.70%	148 100.00% 0.00%	134 99.25% 0.75%	140 100.00% 0.00%	112 99.11% 0.89%	1146 / 1150 99.65% 0.35%

Class Name	Precision	1-Precision	Recall	1-Recall	f1-score
anger	0.9634	0.0366	0.9511	0.0489	0.9572
disgust	0.9623	0.0377	0.9640	0.0360	0.9632
fear	0.9614	0.0386	0.9803	0.0197	0.9708
joy	0.9612	0.0388	0.9758	0.0242	0.9685
neutral	0.9982	0.0018	0.9929	0.0071	0.9956
sadness	0.9813	0.0187	0.9481	0.0519	0.9644
surprise	0.9522	0.0478	0.9737	0.0263	0.9628
Accuracy	0.9681				
Misclassification Rate	0.0319				
Macro-F1	0.9689				
Weighted-F1	0.9681				

Training Set								
TARGET / OUTPUT	anger	disgust	fear	joy	neutral	sadness	surprise	SUM
anger	817 15.62%	7 0.13%	3 0.06%	4 0.08%	0 0.00%	15 0.29%	2 0.04%	848 96.34% 3.66%
disgust	19 0.36%	562 10.75%	1 0.02%	0 0.00%	0 0.00%	2 0.04%	0 0.00%	584 96.23% 3.77%
fear	10 0.19%	3 0.06%	798 15.26%	3 0.06%	0 0.00%	14 0.27%	2 0.04%	830 96.14% 3.86%
joy	13 0.25%	2 0.04%	5 0.10%	967 18.49%	4 0.08%	8 0.15%	7 0.13%	1006 96.12% 3.88%
neutral	0 0.00%	0 0.00%	0 0.00%	1 0.02%	560 10.71%	0 0.00%	0 0.00%	561 99.82% 0.18%
sadness	0 0.00%	0 0.00%	5 0.10%	8 0.15%	0 0.00%	840 16.06%	3 0.06%	856 98.13% 1.87%
surprise	0 0.00%	9 0.17%	2 0.04%	8 0.15%	0 0.00%	7 0.13%	518 9.91%	544 95.22% 4.78%
SUM	859 95.11% 4.89%	583 96.40% 3.60%	814 98.03% 1.97%	991 97.58% 2.42%	564 99.29% 0.71%	886 94.81% 5.19%	532 97.37% 2.63%	5062 / 5229 96.81% 3.19%

Class Name	Precision	1-Precision	Recall	1-Recall	f1-score
provocative	0.9015	0.0985	0.8655	0.1345	0.8831
provocative	0.7989	0.2011	0.8034	0.1966	0.8011
racism	0.8161	0.1839	0.8765	0.1235	0.8452
sexism	0.9048	0.0952	0.9301	0.0699	0.9172
Accuracy	0.8641				
Misclassification Rate	0.1359				
Macro-F1	0.8617				
Weighted-F1	0.8642				

Training Set					
TARGET / OUTPUT	provocative	provocative	racism	sexism	SUM
provocative	238 35.16%	15 2.22%	7 1.03%	4 0.59%	264 90.15% 9.85%
provocative	27 3.99%	143 21.12%	3 0.44%	6 0.89%	179 79.89% 20.11%
racism	2 0.30%	14 2.07%	71 10.49%	0 0.00%	87 81.61% 18.39%
sexism	8 1.18%	6 0.89%	0 0.00%	133 19.65%	147 90.48% 9.52%
SUM	275 86.55% 13.45%	178 80.34% 19.66%	81 87.65% 12.35%	143 93.01% 6.99%	585 / 677 86.41% 13.59%

Class Name	Precision	1-Precision	Recall	1-Recall	f1-score
neutral	0.9243	0.0757	0.9199	0.0801	0.9221
offensive	0.9081	0.0919	0.9130	0.0870	0.9105
Accuracy	0.9167				
Misclassification Rate	0.0833				
Macro-F1	0.9163				
Weighted-F1	0.9167				

Training Set			
TARGET / OUTPUT	neutral	offensive	SUM
neutral	5224 49.28%	428 4.04%	5652 92.43% 7.57%
offensive	455 4.29%	4494 42.39%	4949 90.81% 9.19%
SUM	5679 91.99% 8.01%	4922 91.30% 8.70%	9718 / 10601 91.67% 8.33%

Class Name	Precision	1-Precision	Recall	1-Recall	f1-score
negative	0.8290	0.1710	0.9229	0.0771	0.8735
neutral	0.9994	0.0006	0.9972	0.0028	0.9983
positive	0.9840	0.0160	0.9645	0.0355	0.9741
Accuracy	0.9716				
Misclassification Rate	0.0284				
Macro-F1	0.9486				
Weighted-F1	0.9722				

Training Set				
TARGET / OUTPUT	negative	neutral	positive	SUM
negative	8373 9.50%	24 0.03%	1703 1.93%	10100 82.90% 17.10%
neutral	2 0.00%	30640 34.76%	15 0.02%	30657 99.94% 0.06%
positive	697 0.79%	62 0.07%	46620 52.90%	47379 98.40% 1.60%
SUM	9072 92.29% 7.71%	30726 99.72% 0.28%	48338 96.45% 3.55%	85633 / 88136 97.16% 2.84%

<SONUÇLAR>

Proje boyunca çeşitli modellerle yüksek doğruluk değerine ulaştık hem kullanıcı hem geliştirici dostu bir kütüphane geliştirdik.

Sıkıştırma özelliği ile kullanımı kolay ve az boyutlu modeller geliştirdik

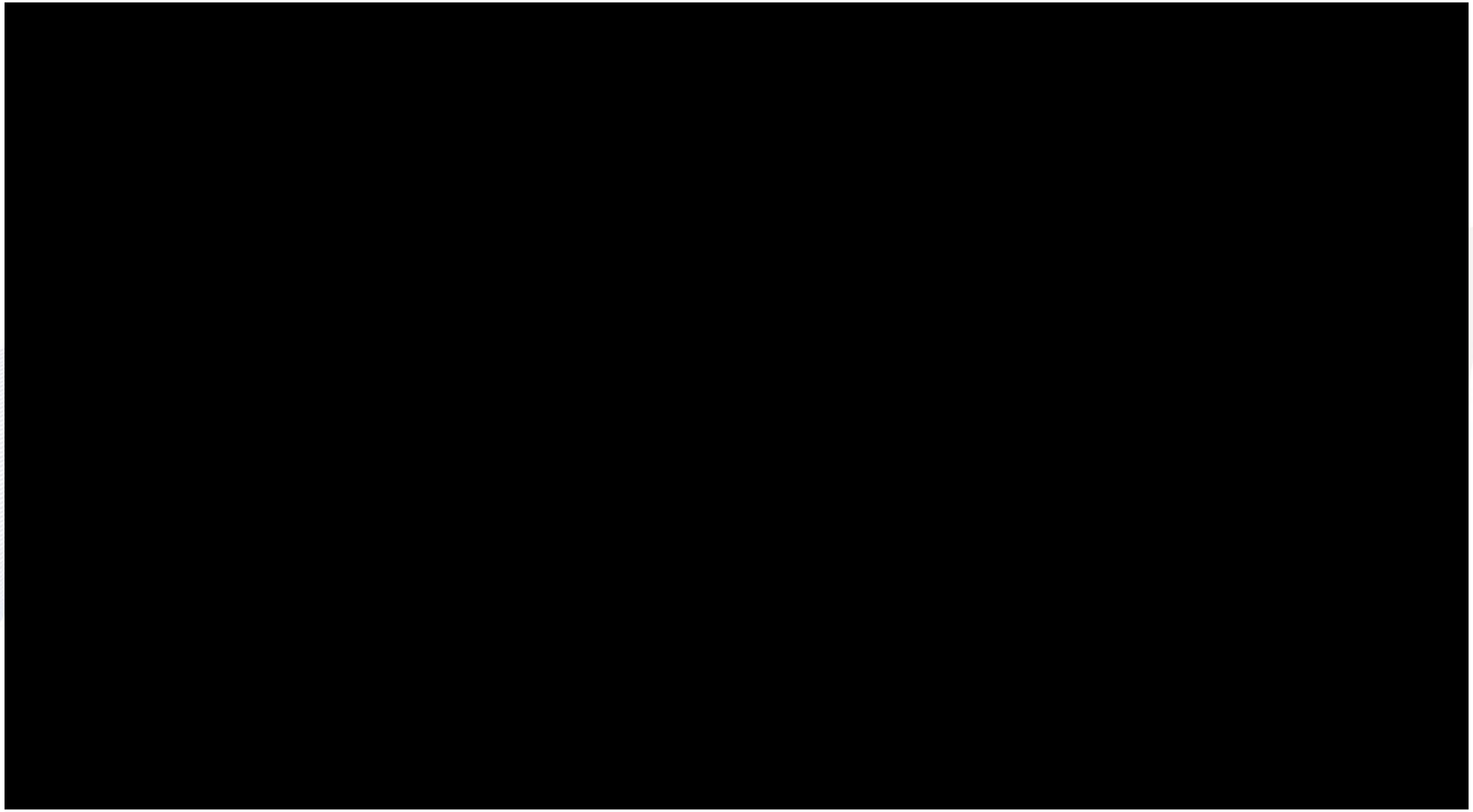
Kendi sınıflandırma modellerini üretmelerini kolaylaştırdık.

<PROJE YOL HARİTASI>

İleride geniş veri setleri oluşturarak kendi LLM modellerimizi oluşturmak istiyoruz

Öğretmen LLM’ler ile geliştiricilerin LLM modellerini fine tune işlemlerini kolaylaştırmak ve doğal dile en yakın sonuçları elde etmeyi hedefliyoruz.

Resim, Ses, Video gibi medya dosyalarını da işlemek ve Arşivcilik faaliyetleri için Osmanlı Türkçesiyle modeller oluşturmak istiyoruz.



phorcys.site/App

TEŞEKKÜRLER