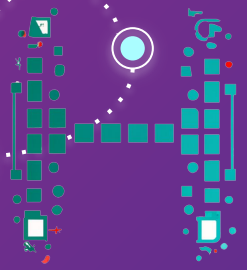


HERMES AI / Teknofest Türkçe Doğal Dil İşleme 2023

Özgür Doğan
Oğuzhan Şahin
Anıl Dursun İpek

Yarışma Konusu

Yarışmanın Konusu: Aşağılayıcı Söylemlerin Doğal Dil İşleme İle Tespiti

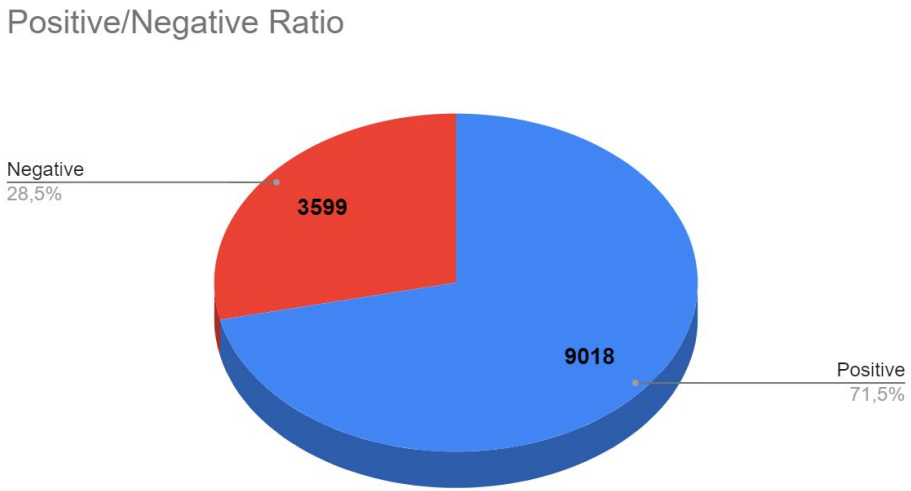
Yarışmanın Alt Başlıkları: Cinsiyetçi, ırkçı, küfür ve hakaret söylemleri gibi aşağılayıcı söylemler içeren cümlelerin doğal dil işleme yöntemleri ile tespit edilmesi ve ortaya çıkan teknik yetkinliğin sektörel kullanım alanları üzerine öneriler iletilmesi.

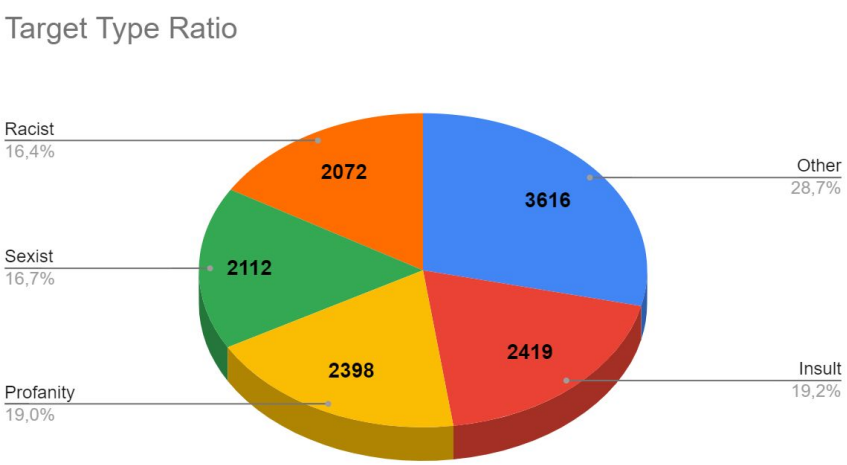
Yarışma Etiketleri:

- SEXIST
- RACIST
- PROFANITY
- INSULT
- OTHER

Keşifsel Veri Analizi - EDA

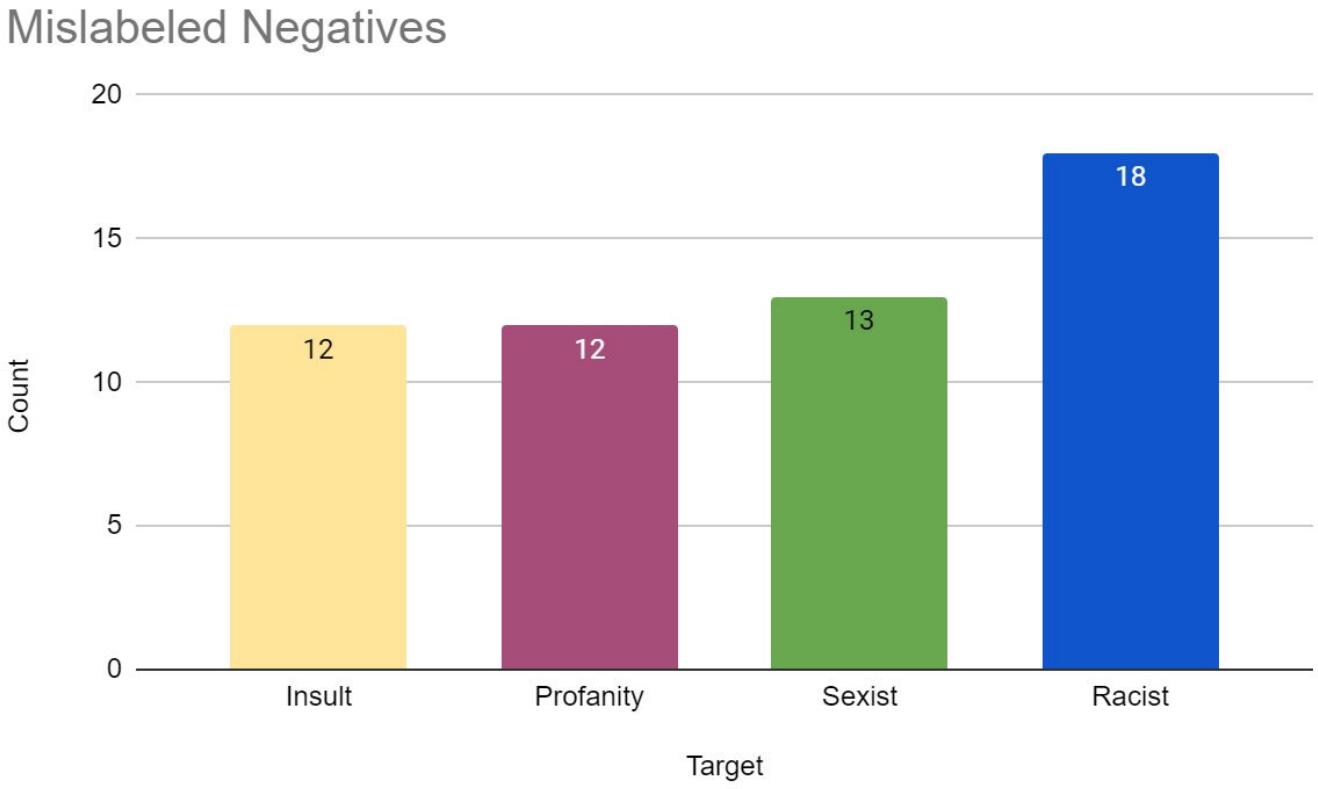
Veri Seti İçerisindeki Etiketlerin Dağılımı





Keşifsel Veri Analizi - EDA

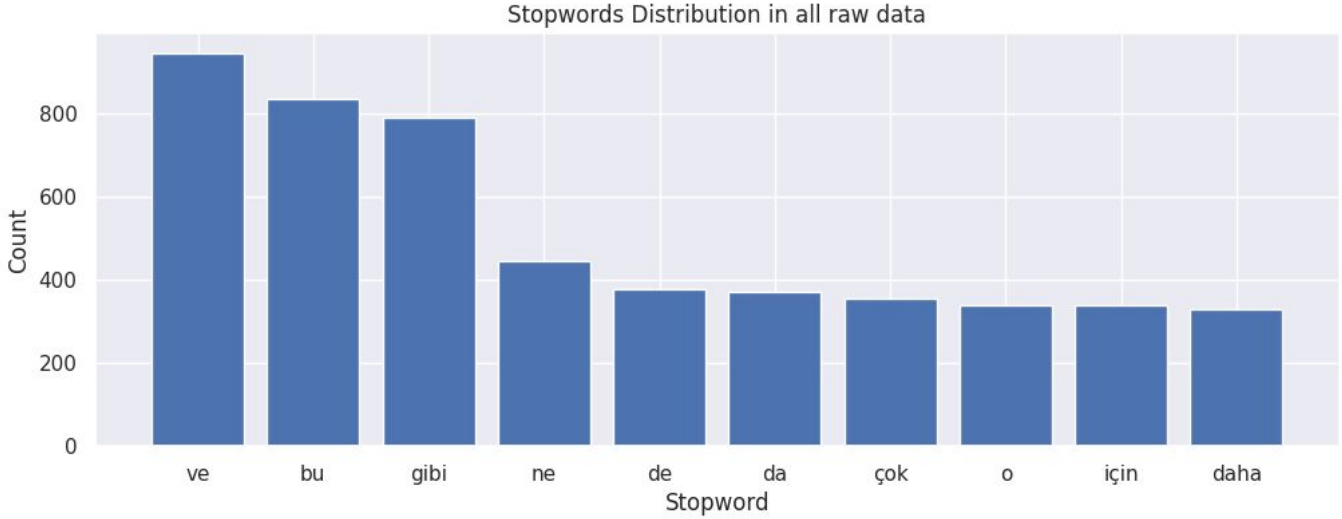
Veri Seti İçerisindeki Etiketlerin Dağılımı



Keşifsel Veri Analizi - EDA

Veriler İçerisindeki Stop Word lerin İncelenmesi

- Yarışma verisi içerisindeki stop words ler ile ilgili incelemeler yapıldı ve başarı skorunu etkilemediği gözlemlendi.



MODEL		HYPERPARAMETERS					METRICS		
Model Name	stopwords	max_length	train_batch_size	valid_batch_size	weight_decay	learning_rate	precision	recall	f1
bert-base-turkish-uncased	removed	32	32	64	0.01	1,00E-05	**0.93**	**0.93**	**0.93**
bert-base-turkish-uncased	kept	32	32	64	0.01	1,00E-05	**0.93**	**0.93**	**0.93**

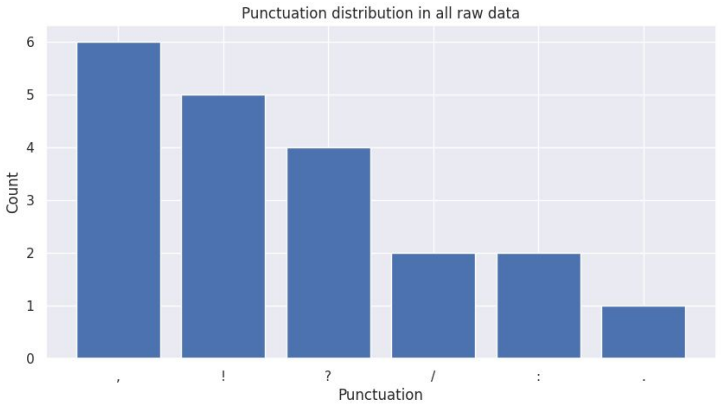
Aynı model kullanılarak 2 farklı eğitim gerçekleştirildi. İlk model stop word lerin çıkarılarak eğitildiği ikinci model ise stop word ler ile eğitilen modeldir.

Veri Ön İşleme Aşamaları

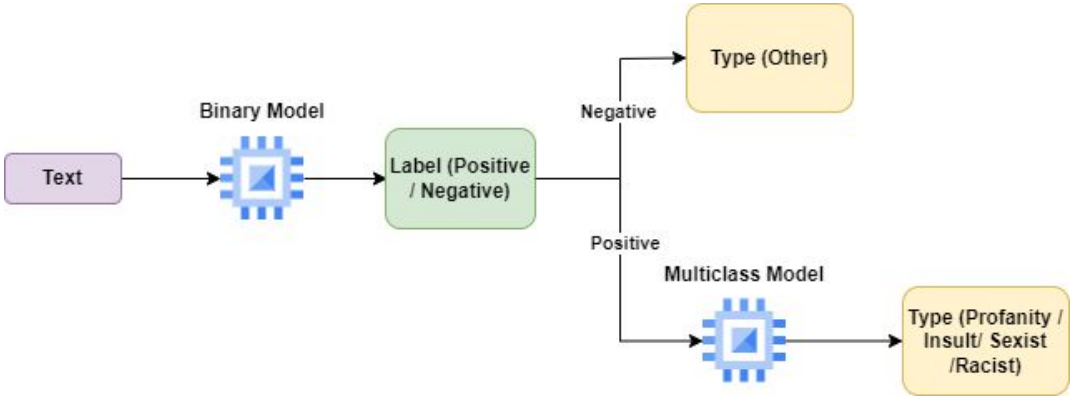
- Duplicate veriler tespit edildi ve çıkarıldı.
- Karakterler küçük harfe dönüştürüldü.
- Metinler içerisindeki stop word ler çıkarıldı. (nltk kütüphanesi içerisindeki türkçe stop word ler kullanıldı)
- Bozuk veriler tespit edildi ve çıkarıldı (Örneğin sadece tek bir harften oluşan 133 adet veri tespit edildi).
- Metinler noktalama işareti ve ifadelerden arındırıldı. (Noktalama işaretlerine ek olarak “<”, “>” gibi ifadeler tespit edildi ve çıkartıldı. Yapılan incelemeler ile çok fazla noktalama işaretinin bulunmadığı gözlemlendi).

	id	text	is_offensive	target
12340	e2d954b7-266d-43be-845e-015a8ecf1241	j	0	RACIST
12341	697c1629-d4f6-4e85-87f6-3fa5510f55cf	k	1	RACIST
12342	f760cf45-ad05-46e7-9971-2b515decae97	e	1	RACIST
12343	03307826-defb-4e34-aa5a-b74ca74c84c2	e	0	RACIST
12344	42bf5d9d-48ab-489f-a673-d6d792f97eb9	b	0	SEXIST
...	...	...	...	...
12485	65ca945d-15af-4d59-8d7d-b731578e45d8	e	1	PROFANITY
12486	593b9691-8287-4400-bb22-ca2add665b9a	j	0	OTHER
12487	6d93aea2-8130-4168-81ea-bd8557ce3272	b	1	OTHER
12488	b89720e0-fdf3-44c8-ae20-14e8fe2d94af	h	0	OTHER
12489	12694e03-fbdf-4c37-8183-daf1bbfb6b91	g	1	SEXIST

133 rows × 4 columns

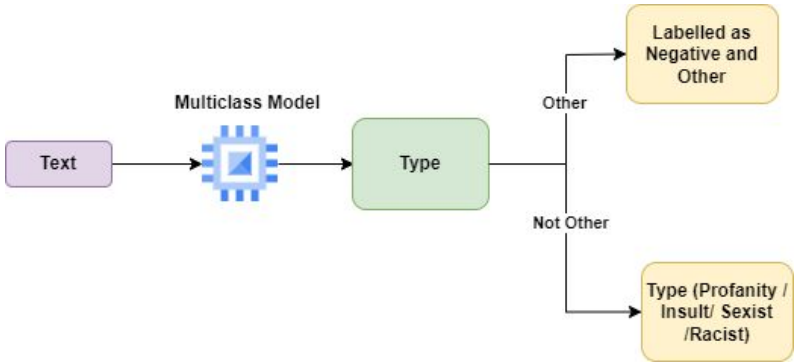


Binary Mimari vs. Multiclass Mimari



Binary Architecture

Multiclass Architecture



Model Denemeleri

Binary Sınıflandırma Denemeleri

MODEL	METRICS		
Model Name	precision	recall	f1
Tfldf + Naive Bayes	**0.86**	**0.83**	**0.80**
Countvectorizer + Naive Bayes	**0.90**	**0.89**	**0.89**
Tfidf + XgBoost	**0.87**	**0.87**	**0.87**

Multiclass Sınıflandırma Denemeleri

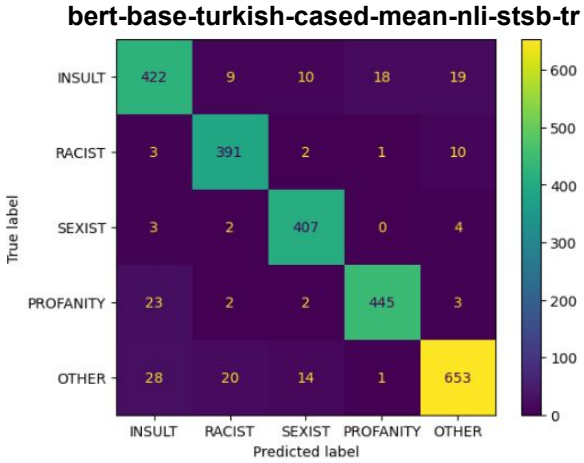
MODEL	HYPERPARAMETERS					METRICS		
Model Name	max_length	train_batch_size	valid_batch_size	weight_decay	learning_rate	precision	recall	f1
bert-base-turkish-uncased	32	32	64	0.01	1,00E-05	**0.93**	**0.93**	**0.93**
bert-base-turkish-128k-uncased	32	32	64	0.01	1,00E-05	**0.93**	**0.93**	**0.93**
convbert-base-turkish-mc4-uncased	32	32	64	0.01	1,00E-05	0.91	0.86	0.88
bert-base-turkish-cased-mean-nli-stsb-tr	32	32	64	0.01	1,00E-05	**0.93**	**0.93**	**0.93**
roberta-base-turkish-uncased	32	32	64	0.01	1,00E-05	0.89	0.89	0.89
electra-base-turkish-cased-discriminator	32	32	64	0.01	1,00E-05	**0.93**	**0.93**	**0.93**
bert-base-turkish-cased-mean-nli-stsb-tr	32	32	64	0.01	1,00E-05	**0.93**	**0.93**	**0.93**

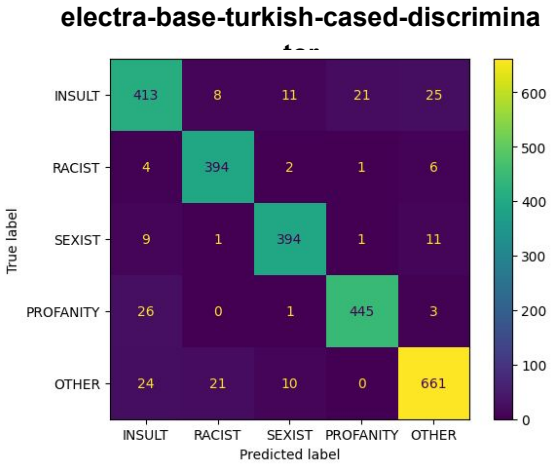
Gri arka planlı satırlar finalde kullanılan modellerdir.

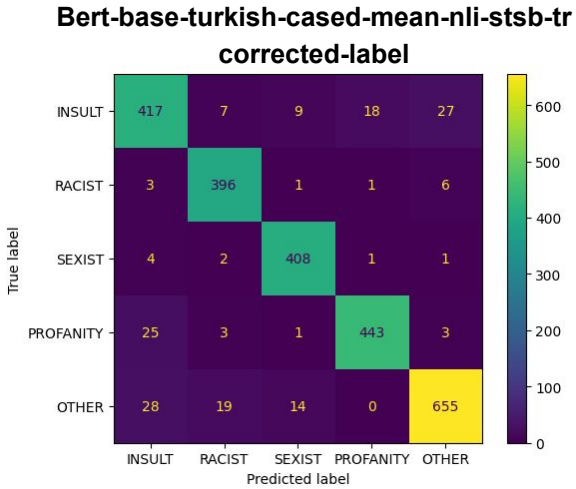
- Eldeki veriler ile farklı pre-trained model denemeleri yapıldı.
- Verideki yanlış etiketlemeler düzeltilip model denemesi yapıldı. 0.004'lük bir F1 skoru artışı gözlemlendi. (tablodaki son model)
- Binary model denemesi yapıldı. Sonuçların multi class a göre daha düşük skorlar verdiği gözlemlendi.
- Dış veri kullanılarak model denemesi yapıldı.

Model Denemeleri

Karmaşıklık Matrisleri (Confusion Matrix)







Model Tahminleri ve Hatalı Etiketlemeler

Text	Model Prediction	Actual Label
Defol burdan lan	Insult	Profanity
Yosma kadınlara benzedin	Sexist	Insult
Or*spu giyinme	Profanity	Insult
Bu ne derttir böyle anasını satayım		Profanity
Anasını satıyım bu ne böyle		Sexist

- **Bu tarz veriler, eğitim kümesinden çıkarıldığında elde edilen model sonucu;**

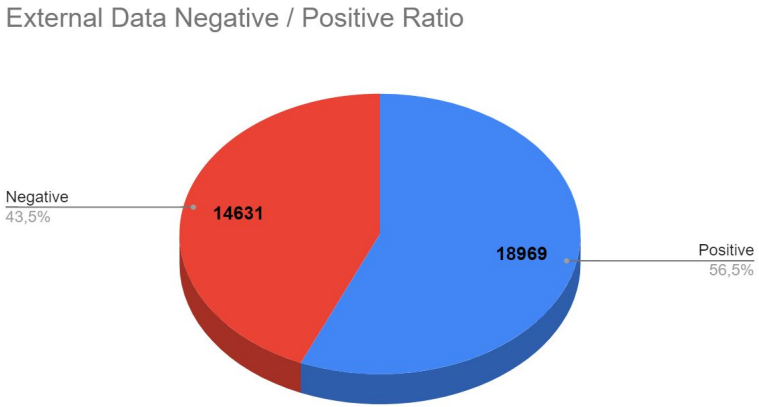
	precision	recall	f1-score	support
INSULT	0.87	0.87	0.87	478
RACIST	0.95	0.91	0.93	716
SEXIST	0.96	0.93	0.94	475
PROFANITY	0.93	0.97	0.95	407
OTHER	0.94	0.98	0.96	416
accuracy			0.93	2492
macro avg	0.93	0.93	0.93	2492
weighted avg	0.93	0.93	0.93	2492

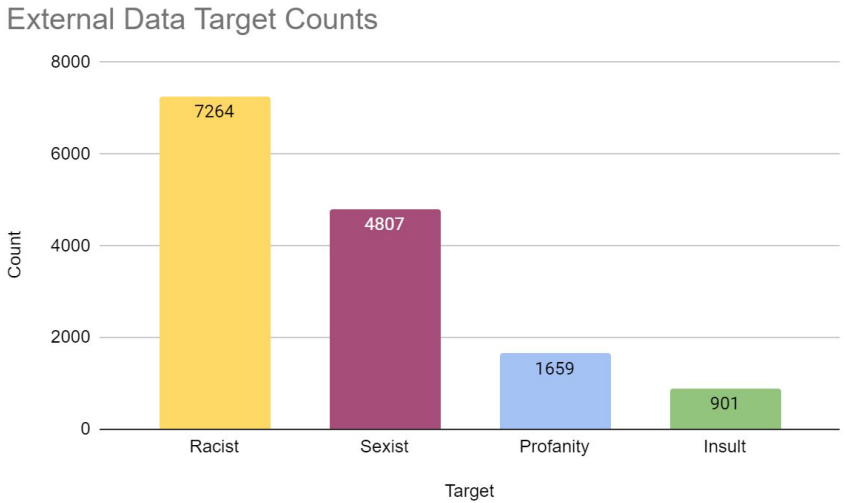
Dış Veri ile Modelleme

- Derogation*
- Animosity
- Threatening Language
- Support for Hateful Entities
- Dehumanization

→

- Racist
- Sexist
- Profanity
- Insult





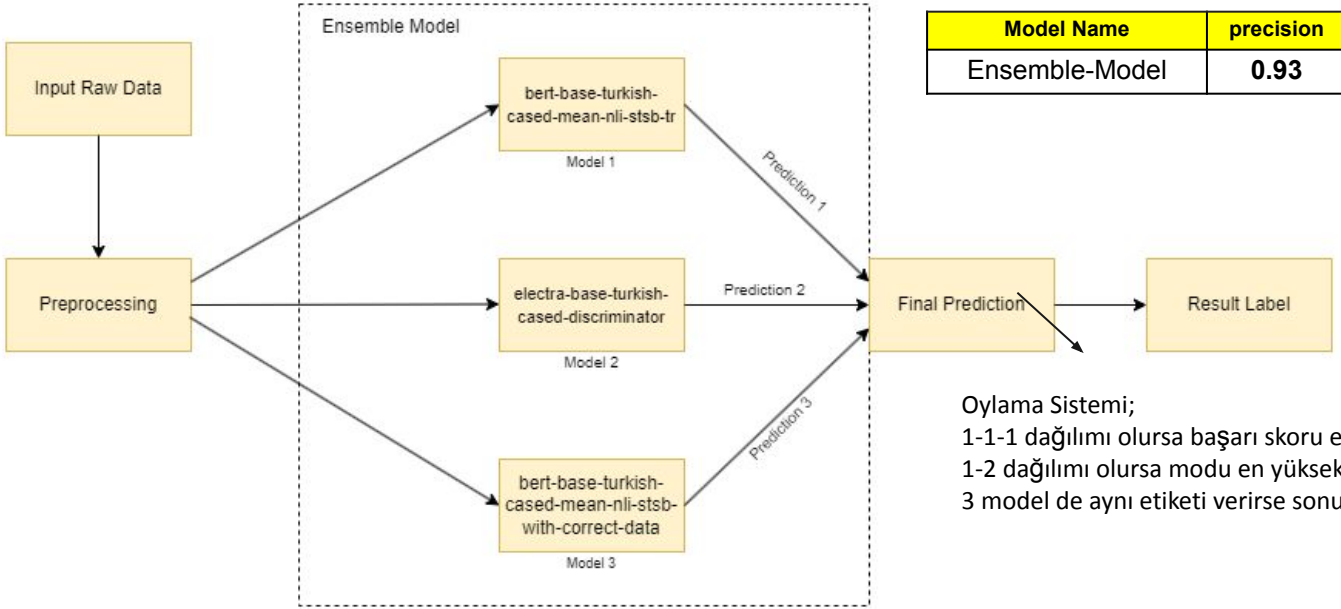
* Vidgen, B., Thrush, T., Waseem, Z., & Kiela, D. (2020). Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*.

Dış Veri ile Modelleme

Target	Precision	Recall	F1-score	Support
Racist	0,86	0,89	0,88	478
Sexist	0,96	0,94	0,95	475
Profanity	0,92	0,96	0,94	407
Insult	0,93	0,97	0,95	416
Other	0,95	0,90	0,92	716
Macro avg.	0,92	0,93	0,93	2492
Weighted avg.	0,93	0,93	0,93	2492

- Beklenen performans artışı yaşanmadı (+0,006)
- Bias riski
- Test verisinden uzaklaşma

Ensemble Model



Model Name	precision	recall	f1
Ensemble-Model	**0.93**	**0.93**	**0.93**

Oylama Sistemi;
1-1-1 dağılımı olursa başarı skoru en yüksek olan alınır,
1-2 dağılımı olursa modu en yüksek olan alınır,
3 model de aynı etiketi verirse sonuç etiket olarak alınır.

- En başarılı 3 model birleştirilerek ensemble bir model oluşturuldu.
- Model gelen etiketlere göre bir oylama sistemi kullanılarak(voting) final tahmini yapılmaktadır.
- Yapılan test sonuçlarına göre diğer modellerde 0.9311 bandında olan F1 skoru ensemble model ile 0.9391'e yükselmektedir.
- Bu modelin final modeli olarak kullanılması kararlaştırılmıştır.

TEŞEKKÜRLER!

Hermes AI