GTÜ BİL MUH BİL 495

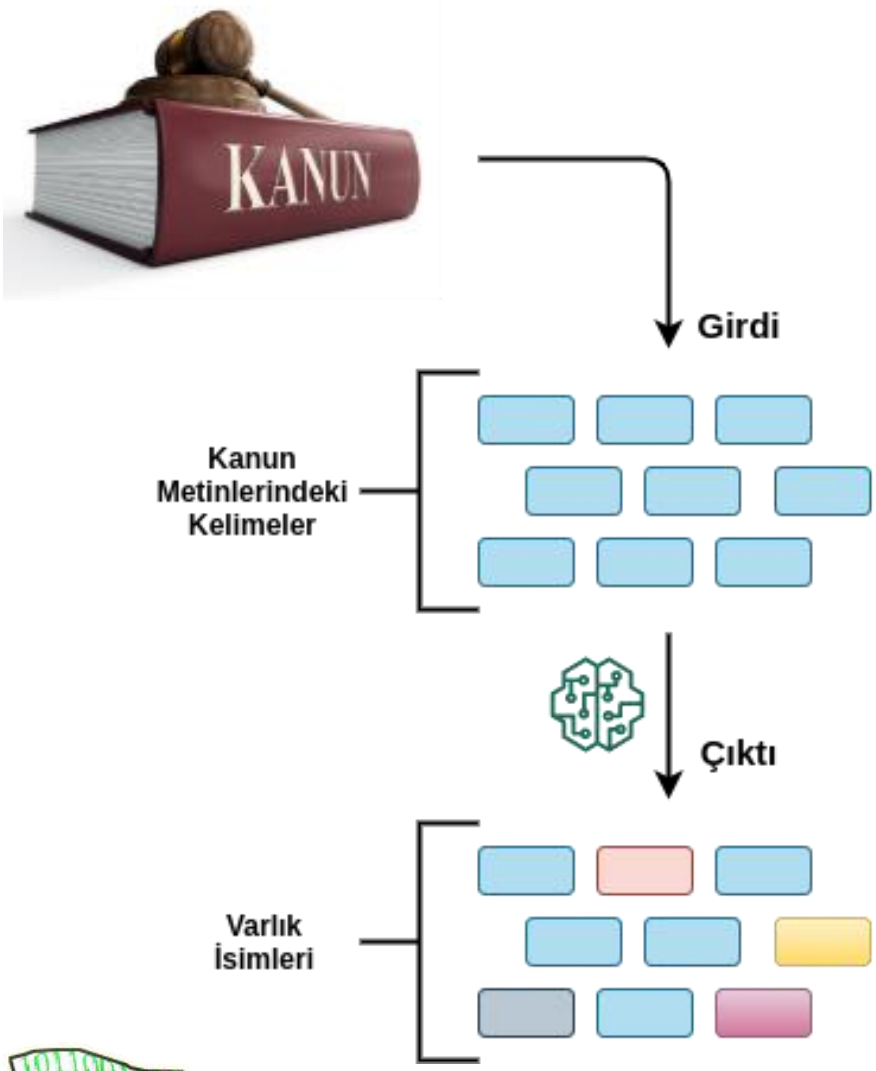
BIL 495
KANUN METİNLERİNDE VARLIK İSİMLERİNİN DERİN ÖĞRENME İLE TESPİTİ
Son Sunum

Gökhan HAS

Proje Danışmanı: Dr. Burcu YILMAZ
OCAK 2021

İçindekiler

Proje Şeması ve Tanımı



- Proje Nedir?

Türkçe kanun metinlerinde bulunan varlık isimlerinin çıkarılmasını sağlayan bir model ortaya çıkarmaktır.

- Varlık İsimleri Nelerdir?

Varlık isimleri; kişi, yer, organizasyon gibi önceden tanımlanmış kategorilerin metin dokümanları üzerinden çıkarılma işlemidir. Doğal Dil İşleme problemi olarak tanımlanmaktadır.

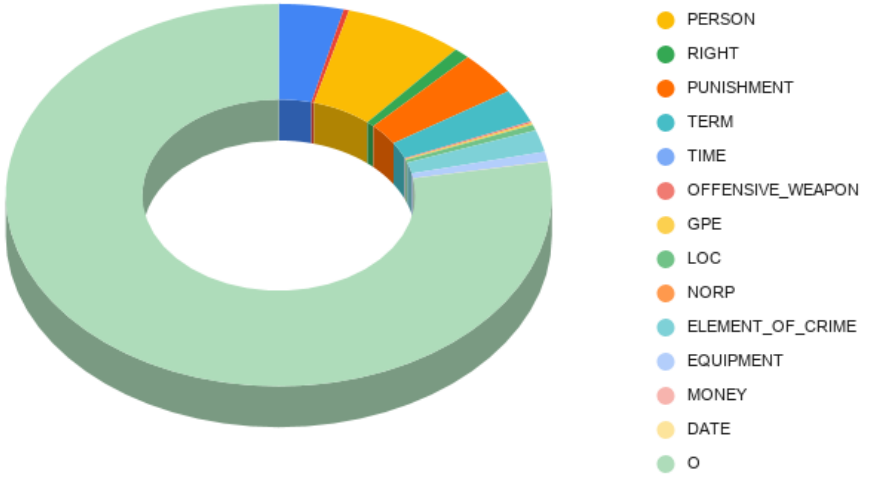
- Projedeki amaç, kanun metinlerine özgü etiketler oluşturularak varlık isimlerinin bulunmasını sağlamaktır.

Veri Kümesi - 1

Projede hazır bir veri kümesi kullanılmamış, sıfırdan veri kümesi oluşturulmuştur. Türkçe kanun metinleri pdf şeklinde indirilmiş, ön işleme uygulanarak başarıyı bozacak kelime ve kanun metninin yapıları temizlenmiştir.

Veri kümesinde 60.014 kelime bulunmaktadır. Bu kelimeler teker teker incelenmiş ve hangi varlık isimde oldukları belirlenmiştir. Veri kümesinde 16 adet etiket kullanılmıştır.

Örneğin kurum ve kuruluş isimleri için ORGANIZATION etiketi kullanılırken, haklar için RIGHT etiketi kullanılmıştır.

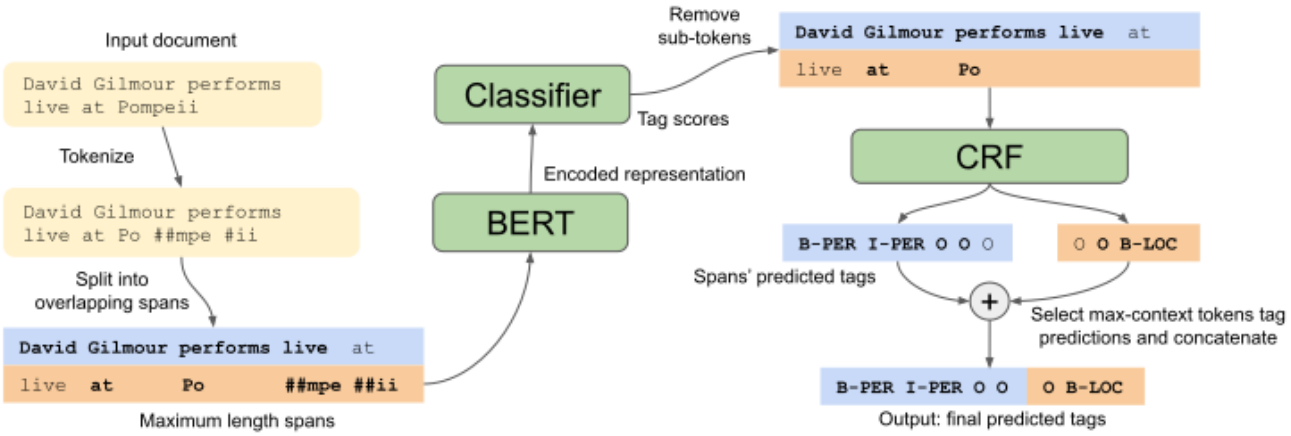


Veri Kümesi - 2

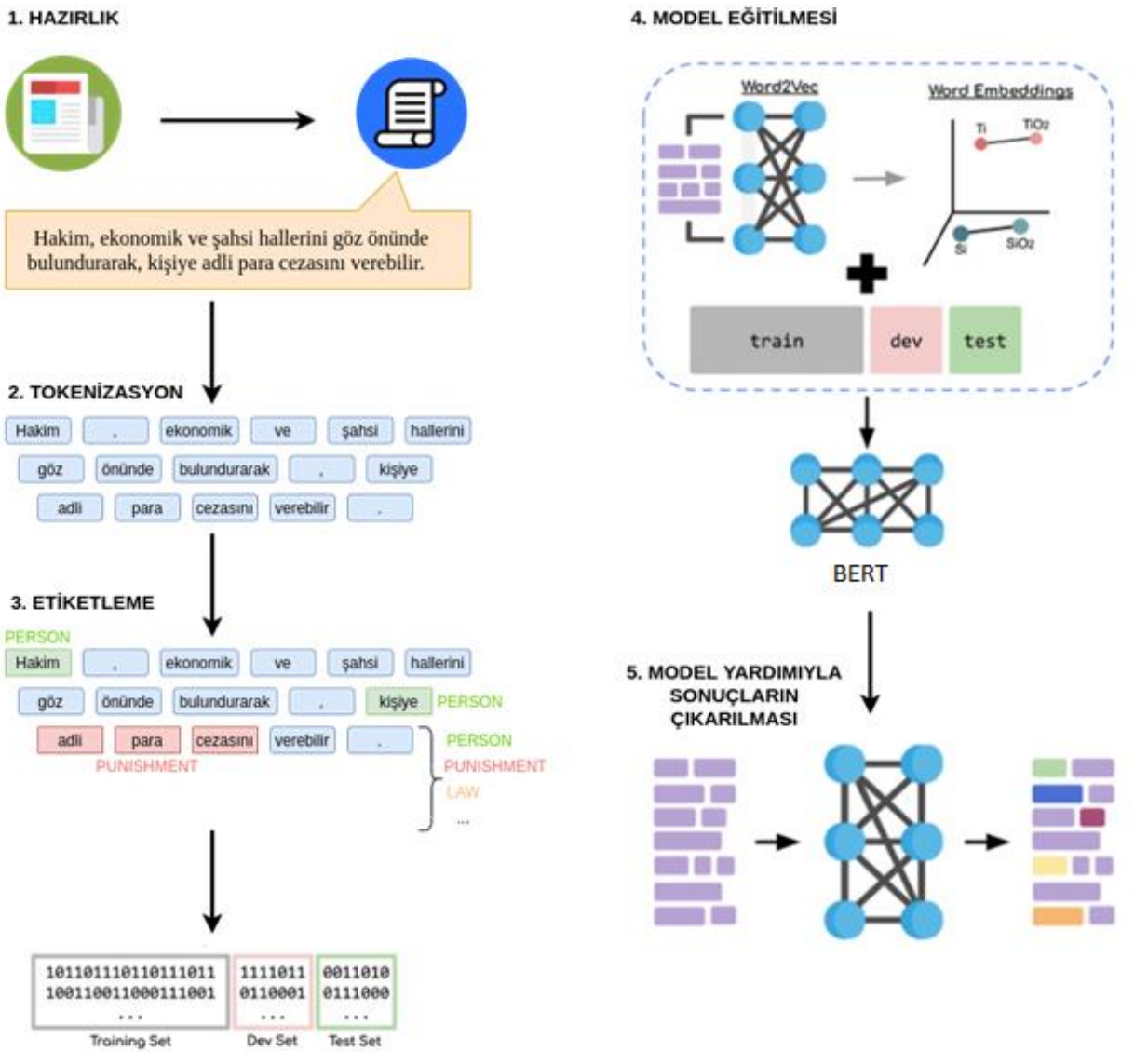
59998	Sentence : 2556	Alacakları	O
59999		,	O
60000		tereke	B-EQUIPMENT
60001		mallarıyla	O
60002		güvence	O
60003		altına	O
60004		alınmış	O
60005		olan	O
60006		alacaklılar	B-PERSON
60007		deftere	O
60008		geçirilmemiş	O
60009		olsa	O
60010		bile	O
60011		bu	O
60012		haklarını	B-RIGHT
60013		güvenceden	O
60014		alabilirler	O
60015		.	O
60016			

Yaklaşım

Yapılan literatür araştırmaları sonucunda bu problem için Bi-LSTM veya BERT derin öğrenme modellerinin yüksek doğrulukla sonuç verdiği çıkarılmıştır.
Yapılan çalışmaların az olması, daha yüksek sonuç vermesi ve daha yeni bir model olduğu için BERT derin öğrenme modelinin kullanılmasına karar verilmiştir.



Proje Tasarım Planı

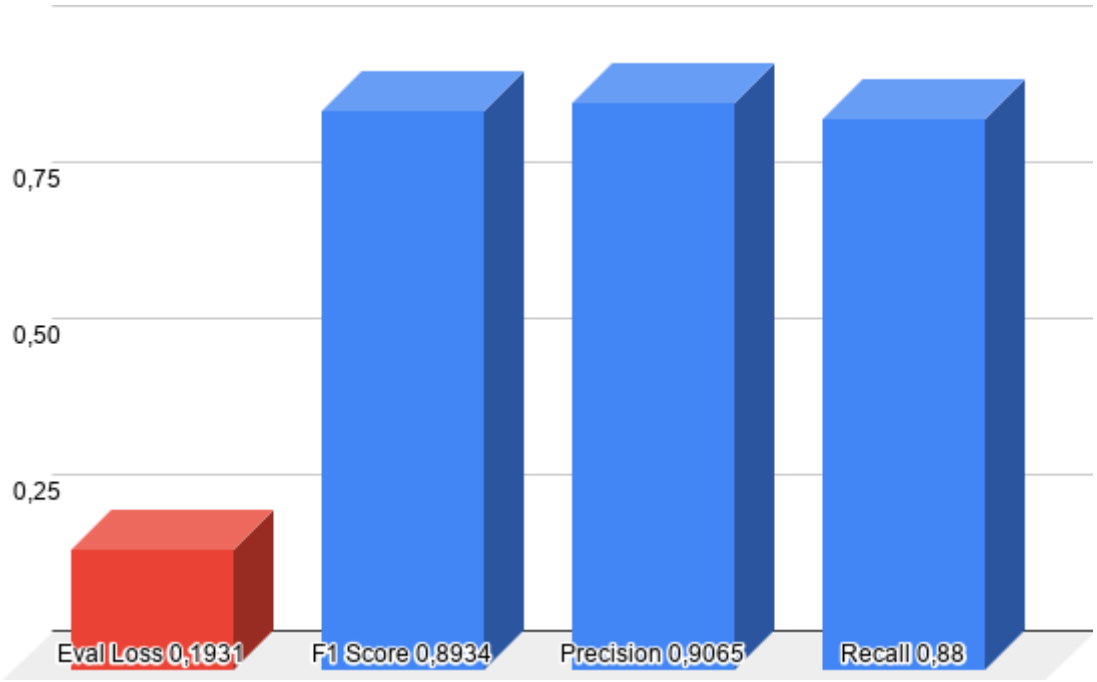


Yapılanlar

- Yapılan araştırmalar sonucunda 3 farklı BERT modeli kullanılmıştır. Bu modeller, aynı parametreler kullanılarak eğitilmiş ve en iyi sonucu veren model kullanılmıştır.
- BERT modelleri için önerilen parametre değerleri araştırılmış ve optimum parametreler belirlenerek modelin sonucunun iyileştirilmesi sağlanmıştır.
- Veri kümesi büyültme işlemleri yapılmıştır. Veri kümesinde önceden yapılan hatalı etiketlemeler giderilmiştir. Bazı etiketlerin yanlış yazılarak fazladan etiket oluşturulma problemi de çözülmüştür.

Başarı Kriterleri - 1

Veri kümesinde düzeltme ve hatalar giderildikten sonra başarı kriterleri aşağıdaki gibi görülmüştür.



Başarı Kriterleri - 2

Veri Kümesi boyutuna göre elde edilen başarı kriteri sonuçları aşağıdaki gibidir:

	EVAL LOSS	F1 SCORE	PRECISION	RECALL
15K KELIME	60%	70%	71%	68%
44K KELIME	33%	77%	78%	76%
60K KELIME	19%	89%	90%	88%
HEDEF	< 20%	> 50%	> 60%	...

Tahmin Sonuçları - 1

100% 1/1 [00:00<00:00, 7.24it/s]

Running Prediction: 100% 1/1 [00:00<00:00, 10.67it/s]

```
[[{'Fail': 'B-PERSON'},
  {'yararına': 'O'},
  {'cezayı': 'B-PUNISHMENT'},
  {'hafifletecek': 'O'},
  {'takdiri': 'O'},
  {'nedenlerin': 'O'},
  {'varlığı': 'O'},
  {'halinde,': 'O'},
  {'ağırlaştırılmış': 'B-PUNISHMENT'},
  {'müebbet': 'I-PUNISHMENT'},
  {'hapis': 'I-PUNISHMENT'},
  {'cezası': 'I-PUNISHMENT'},
  {'yerine,': 'O'},
  {'müebbet': 'I-PUNISHMENT'},
  {'hapis;': 'I-PUNISHMENT'},
  {'müebbet': 'I-PUNISHMENT'},
  {'hapis': 'I-PUNISHMENT'},
  {'cezası': 'I-PUNISHMENT'},
  {'yerine,': 'O'},
  {'yirmibeş': '#'},
  {'yıl': 'B-P_TERM'},
  {'hapis': 'I-PUNISHMENT'},
  {'cezası': 'I-PUNISHMENT'},
  {'verilir.': 'O'},
  {'Diğer': 'O'},
  {'cezaların': 'B-PUNISHMENT'},
  {'altıda': 'O'},
  {'birine': '#'},
  {'kadarı': 'B-P_TERM'},
  {'indirilir.': 'O'}]]
```

Tahmin Sonuçları - 2


```
[[{'Mahkeme,': 'B-ORGANIZATION'},
  {'hükümlünün': 'B-PERSON'},
  {'kişiliğini': 'O'},
  {'ve': 'O'},
  {'sosyal': 'O'},
  {'durumunu': 'O'},
  {'göz': 'O'},
  {'önünde': 'O'},
  {'bulundurarak,': 'O'},
  {'denetim': 'B-ORGANIZATION'},
  {'süresinin': 'I-ORGANIZATION'},
  {'herhangi': 'O'},
  {'bir': 'O'},
  {'yükümlülük': 'O'},
  {'belirlemeden': 'O'},
  {'veya': 'O'},
  {'uzman': 'B-PERSON'},
  {'kişi': 'I-PERSON'},
  {'görevlendirmeden': 'O'},
  {'geçirilmesine': 'O'},
  {'de': 'O'},
  {'karar': 'O'},
  {'verebilir.': 'O'}]]
```

Tahmin Sonuçları - 3

100% 1/1 [00:00<00:00, 4.54it/s]

Running Prediction: 100% 1/1 [00:00<00:00, 16.71it/s]

```
[[{'Bir': 'O'},
  {'kamu': 'B-ORGANIZATION'},
  {'kurumunun': 'I-ORGANIZATION'},
  {'veya': 'O'},
  {'kamu': 'B-ORGANIZATION'},
  {'kurumu': 'I-ORGANIZATION'},
  {'niteliğindeki': 'O'},
  {'meslek': 'B-ORGANIZATION'},
  {'kuruluşunun': 'I-ORGANIZATION'},
  {'iznine': 'O'},
  {'tabi': 'O'},
  {'bir': 'O'},
  {'meslek': 'B-ORGANIZATION'},
  {'veya': 'O'},
  {'sanatı,': 'O'},
  {'kendi': 'O'},
  {'sorumluluğu': 'O'},
  {'altında': 'O'},
  {'serbest': 'O'},
  {'meslek': 'B-ORGANIZATION'},
  {'erbabı': 'O'},
  {'veya': 'O'},
  {'tacir': 'O'},
  {'olarak': 'O'},
  {'icra': 'B-EQUIPMENT'},
  {'etmekten,': 'O'},
  {'yoksun': 'O'},
  {'bırakılır.': 'O'}]]
```

Sonuçlar

- Kanun metinlerine özel daha fazla kelimenin belirlenmesi için daha fazla etiket oluşturularak ve veri kümesi geliştirilerek daha başarılı modeller oluşturulabilir.

- Kanun metinlerinde varlık isimlerinin bulunması amaçlandığında bunu aramak için harcanan zaman eğitilen bu model kullanılarak daha verimli kullanılabilir.

Projede Öğrenilenler

- Derin öğrenme kullanılarak yapmış olduğum ilk projedir. Derin öğrenme modelleri konusunda aklımda oluşan sorulara cevap bulmamı sağladı.
- 2018 yılından itibaren yapılan çevirilerde son derece iyi sonuçlar veren Google Çeviri'nin arkasındaki teknolojiyi araştırma fırsatı olmuştur. (BERT)
- Derin öğrenme modellerinin sonuçlarının analiz edilip, en optimal parametrelerini belirlemek için nasıl bir süreçten geçirildiğini öğrenme fırsatım olmuştur.
- Bir derin öğrenme projesinde veri kümesi oluştururken nelere dikkat edilmesini öğretti.

Kaynaklar

1. https://huggingface.co/dbmdz/bert-base-turkish-128k-cased
2. https://app.diagrams.net/
3. https://data.mendeley.com/datasets/cdcztymf4k/1
4. http://www.madeinturkeydergisi.com/kanunlar/
5. https://www.mevzuat.gov.tr/#kanunlar
6. https://www.kaggle.com/abhishek/entity-extraction-model-using-bert-pytorchhttps://www.kaggle.com/shoumikgoswami/ner-using-random-forest-and-crf