T.C.
GEBZE TEKNİK ÜNİVERSİTESİ

Bilgisayar Mühendisliği Bölümü

KANUN METİNLERİNDE VARLIK İSİMLERİNİN DERİN ÖĞRENME İLE TESPİTİ

Gökhan HAS

Danışman
Dr. Burcu YILMAZ

Ocak, 2021
Gebze, KOCAELİ

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ

Bilgisayar Mühendisliği Bölümü

KANUN METİNLERİNDE VARLIK İSİMLERİNİN DERİN ÖĞRENME İLE TESPİTİ

Gökhan HAS

Danışman
Dr. Burcu YILMAZ

Ocak, 2021
Gebze, KOCAELİ

Bu çalışma ..../..../2020 tarihinde aşağıdaki jüri tarafından Bilgisayar Mühendisliği Bölümü'nde Lisans Bitirme Projesi olarak kabul edilmiştir.

Bitirme Projesi Jürisi

Danışman Adı	Dr. Burcu YILMAZ	
Üniversite	Gebze Teknik Üniversitesi	
Fakülte	Mühendislik Fakültesi	

Jüri Adı	Doç. Dr. Mehmet GÖKTÜRK	
Üniversite	Gebze Teknik Üniversitesi	
Fakülte	Mühendislik Fakültesi	

Jüri Adı		
Üniversite		
Fakülte		

ÖNSÖZ

Bu raporun hazırlanmasında emeği geçenlere, raporun son halini almasında yol gösterici olan Sayın Dr. Burcu YILMAZ hocama ve bu çalışmayı destekleyen Gebze Teknik Üniversitesi'ne içten teşekkürlerimi sunarım. Ayrıca eğitimim süresince bana her konuda tam destek veren aileme ve bana hayatlarıyla örnek olan tüm hocalarıma saygı ve sevgilerimi sunarım.

Ocak, 2021 **Gökhan HAS**

İÇİNDEKİLER

ŞEKİL LİSTESİ

KISALTMA LİSTESİ

MLM : Masked Language Modeling
NSP : Next Sentence Prediction
BERT : Biderectional Encoder Representations from Transformers
MBS **:** Mevzuat Bilgi Sistemi
CRF **:** Koşullu Rastgele Alan (Conditional Random Field)
LSTM **:** Uzun-Kısa Süreli Bellek (Long Short-Term Memory)
CNN **:** Evrişimli Sinir Ağları (Convolutional Neural Network)
GRU **:** Kapılı Tekrarlayan Hücre (Gated Recurrent Unit)

ÖZET

Bu projede, kanun metinlerinde varlık isimlerinin derin öğrenme ile tespiti yapılması amaçlanmıştır. Varlık isimlerinin tespiti; doğal dil işleme ve metin madenciliği alanlarının kapsamında yer alan bir bilgi çıkarımı görevidir [1]. Kapsam ve kullanılan derin öğrenme metotları açısından, çalışmalar arasında farklılıklar görülmüş olsa da, temel olarak, bir metin içerisinden yer, kişi, kurum ve kuruluş vb. belirten ifadelerin tespit edilmesi hedeflenir.

Bu alanda yapılan eski çalışmalarda istatiksel bazlı sistemlerin analizi kullanıldığı görülmektedir. Yakın zamanda yapılan çalışmalarda ise bu alandaki en başarılı sonuçların derin öğrenme metotları kullanılarak yapıldığı görülmektedir. Derin öğrenme metotlarında kelimelerin vektörel olarak temsil edilmesinden faydalanılır.

Bu projede Türkçe kanuın metinleri için bir varlık ismi tanıma modeli geliştirilmiştir. Projenin başlangıcında kanun metinlerini içeren veri kümesi oluşturulmuştur. Derin öğrenme yöntemleri kullanılarak bu veri kümesi üzerinde model eğitimi amaçlanmıştır.

Kanun metinleri dışında kalan veriler temizlenerek veri kümesi üzerinde önişleme işlemleri yapılmıştır. Ön işleme yapıldıktan sonra teker teker ilgili kelimenin hangi sınıfa girdiği veri kümesine kaydedilmiştir. Daha sonra derin öğrenme modeli olarak BERT modeli belirlenmiş ve bu model eğitilmiştir. Derin öğrenme modelinin aldığı parametreler değiştirilerek en optimum sonuç çıkarılmıştır.

Model eğitildikten sonra, başarı kriterleri hesaplanmış ve veri kümesinde bulunmayan kelimeler verilerek modelin bu kelimeler hakkından tahminleri alınarak sonuca ulaşılmıştır.

SUMMARY

In this project, it is aimed to determine the names of entities in the texts of the law through deep learning. Determination of asset names; It is an information extraction task within the scope of natural language processing and text mining [1]. Although there are differences between the studies in terms of the scope and the deep learning methods used, basically, the place, person, institution and organization etc. within a text. It is aimed to determine the expressions that indicate.

It is seen that the analysis of statistical based systems was used in previous studies in this field. In recent studies, it is seen that the most successful results in this field are achieved by using deep learning methods. In deep learning methods, it is used to represent words as vector.

In this project, an entity name recognition model has been developed for Turkish law texts. At the beginning of the project, a data set containing the texts of the law was created. Model training is aimed on this dataset using deep learning methods.

Data excluding legal texts were cleaned and preprocessed on the data set. After preprocessing, the class of the relevant word one by one was recorded in the data set. Later, the BERT model was determined as a deep learning model and this model was trained. By changing the parameters of the deep learning model, the optimum result was obtained.

After the model was trained, success criteria were calculated and the results were obtained by giving the words that are not in the data set, and the estimates of the model about these words.

1. GİRİŞ

Günümüzde yapay zeka kavramı artık günlük yaşantıdaki birçok farklı kategoride kullanılmaktadır. Artık bu kategoriler içerisinde hukuk alanında da çalışmalar yapılmaya başlanılmıştır.

Kullanılan bu sistemlerin yararları düşünüldüğünde hukuk sistemlerinde fazlaca zaman veya iş gücü ihtiyacı duyulan süreçleri kolaylaştırması sağlanabilmektedir.

Bu çalışma hukuk alanında kanun metinlerindeki gerekli bilgilerin çıkarılmasını kolaylaştırmak amacıyla yapılmıştır. Bu çalışmada Türkçe için Kanun metinlerinde bulunan Kanun, kişi, hak, organizasyon, ceza, silah, konum, suç unsuru, dönem, para, tarih ve ekipman gibi varlık isimlerinin doküman içerisinden etiketlenerek çıkarılması için bir model geliştirilmiştir.

Varlık İsmi Tanıma kişi, yer, organizasyon gibi önceden tanımlanmış kategorilerin metin dokümanları üzerinden çıkarılma işlemidir. Bilgi çıkarımının bir alt dalı olup makine çevirilerinden duygu analizine kadar birçok Doğal Dil İşleme probleminde kullanılmaktadır. Tanımı ilk olarak 1995 yılında MUC-6 konferansında yapılmıştır. ENAMEX, TIMEX ve NUMEX olmak üzere 3 temel kategoride tanımlamalar yapılmaktadır. Enamex, kişi, yer, organizasyon gibi ifadeleri; Numex, parasal ve yüzdesel ifadeleri; Timex, gün ve tarih gibi zamansal ifadeleri tanımlamak için kullanılmaktadır. [2]

Çalışma ile ilgili literatür araştırılması yapılarak konu hakkında önceden yapılan çalışmalar ve sonuçları incelenmiştir. Yapılan bu çalışmalara bir sonraki bölümde değinilmiştir.

2. LİTERATÜR TARAMASI

Literatür taramasında varlık isimlerinin tespiti ile ilgili iki ana yaklaşımın bulunduğu görülmüştür. Yaklaşımlardan biri makine öğrenmesi yöntemlerini vurgularken diğeri derin öğrenme yöntemlerini vurgular.

Makine öğrenmesini vurgulayan yaklaşım ontoloji tabanlıdır. Yapılandırılmamış veya yarı yapılandırılmış metinlerdeki bilinen terim ve kavramları tanımada mükemmeldir, ancak büyük ölçüde güncellemelere dayanır. Aksi takdirde, kamuya açık olarak sürekli artan bilgiye ayak uyduramaz. Varlık ismi tanıma görevini çözmek için geniş bir yelpazeye yerleştirilebilecek birçok makine ögrenimi tabanlı yöntem önerilmiştir. İyi bilinen yaklaşımlardan bazıları koşullu rassal alanlar (CRF), en büyük entropi, saklı anlamsal ilişkilendirme ve karar ağaçları yöntemlerini temel alır. [3]

Derin öğrenme yöntemini vurgulayan yaklaşım makine öğrenme yaklaşımına göre daha yüksek doğrulukta sonuçlar verir. Bunun nedeni selefine göre kelimeleri kümelendirebilme özelliğidir. Sözcükler arasındaki anlamsal ve sözdizimsel ilişkiyi anlayabilen teknikler kullanılır. Ontoloji'de bulunmayan terimleri ve kavramları tanıyabilir. Çünkü yazılı dillerde çeşitli kavramların nasıl kullanıldıkları konusunda eğitilmişlerdir. Otomatik olarak öğrenebilir ve analiz yapabilir. Projede bu yüzden derin öğrenme kullanılmıştır. Bu alanda yapılan Türkçe çalışmalar ne yazık ki çok çok az sayıdadır. Türkçe Kanunlarda herhangi bir varlık isimlerini tanıma projesi görülmemiştir.

Derin öğrenme yöntemlerinin kullanılmasından önce, dizi etiketleme problemlerinde Gizli Markov Modeli (HMM), Maksimum Entropi Markov Modeli (MEMM), Koşullu Rastgele Alan (CRF) gibi istatistiksel yöntemler yaygın olarak kullanılıyordu. İstatistiksel yöntemlerden sonra yapay sinir ağları kullanılarak bazı çalışmalar yapılmıştır. Bu çalışmalardan birinde Collobert (2011), kelime vektörlerini kullanan basit ama etkili bir yapay sinir ağı önermiştir. Bu modeldeki bazı özellikler manuel olarak çıkarılabilse de, çoğu özellik kelime vektörleriyle

öğrenilecek şekilde tasarlanmıştır. Son zamanlarda, özyinelemeli sinir ağları (RNN) ve uzun kısa süreli bellek ağı (LSTM), geçitli tekrarlayan hücre (GRU) kullanan mimariler, seri verileri modellemede büyük başarı göstermiştir. [4]

Uzun Kısa Süreli Bellek Ağları (LSTM) (Hochreiter ve Schmidhuber, 1997), uzun vadeli bağımlılıkları öğrenebilen özel bir Yinelenen Sinir Ağı türüdür. LSTM, kapılar adı verilen yapılar tarafından dikkatle düzenlenen hücre durumuna bilgileri kaldırma veya ekleme yeteneğine sahiptir.

Bir dizi kelimeyi işlerken, belirli bir süre için hem geçmiş hem de gelecek girdiler bilinir, bu nedenle özelliklerin hem sağ hem de sol yönlerde etkili bir şekilde kullanılmasına izin verir. LSTM'nin bu varyasyonu çift yönlü LSTM (BI-LSTM) olarak adlandırılır (Graves ve Schmidhuber, 2005). Burada giriş, kelimenin hem sol hem de sağ bağlamını yakalamak için ileri ve geri LSTM'lere verilir.

Bu konuda yapılan çalışmalar kısaca özetlenirse;
Huang (2015), [5] modelini çift yönlü uzun kısa süreli bellek kullanarak oluşturmuştur. Sadece kelime vektörü kullanılmış, karakter vektörü kullanılmamıştır. LSTM çıktısı koşullu rastgele alandan (CRF) geçirilmiştir.

Nichols ve Chiu (2016), [6] LSTM kullanan bir model tasarlamışlardır. Tasarladıkları bu model, girdi olarak hem kelime vektörü hem de karakter vektörü almaktadır.

Hovy ve Ma (2016), [7] bir üstte bulunan çalışmaya ek olarak CNN ile elde edilmiş karakter vektörlerini kullanmışlardır. Bu modelde ise yine çalışma sonuçları koşullu rastgele alandan geçirilmiştir.

Yang (2016), [8] hem kelime hem de karakter vektörleri kullanarak bir LSTM modeli oluşturmuştur. Ek olarak CRF katmanında da değişiklikler yapmıştır.

Yaptığı çalışmadaki farklılık, modelinde kapılı tekrarlayan hücre (GRU) kullanmış olmasıdır.

Lample (2016), [9] Hovy ve Ma'nın çalışmalarına ek olarak karakter vektörlerinin oluşturulma biçimini değiştirmiştir. Karakter vektörlerini uzun kısa süreli bellek modeli kullanarak üretmişlerdir.

Zhang (2018), [10] bir önceki yaptığı çalışmadan tamamen farklı bir çalışma yaparak girdilerini sadece koşullu rastgele alandan geçirmiştir. Modeline ek olarak bir katman daha eklemiştir.

Yang (2018), [11] o zamana dek yapılan çalışmaların karşılaştırmaları yapılmıştır. Yapılan karşılaştırmalar sonucunda uzun kısa süreli bellek (LSTM) kullanımının ve bu model çıktısının koşullu rastgele alandan geçirilmesinin en iyi sonucu verdiği izlenimi çıkarılmıştır.

Yapılan yukarıdaki çalışmalarda CoNLL-2003 veri kümesi kullanılmıştır. Ve bu çalışmaların F1 skorları aşağıdaki gibidir. 2018 yılında BERT modeli duyurulduktan sonra yapılan çalışmaların daha başarılı sonuçlar verdiği çıkarılmıştır. (BERT modeli ileriki bölümlerde açıklanmıştır.)

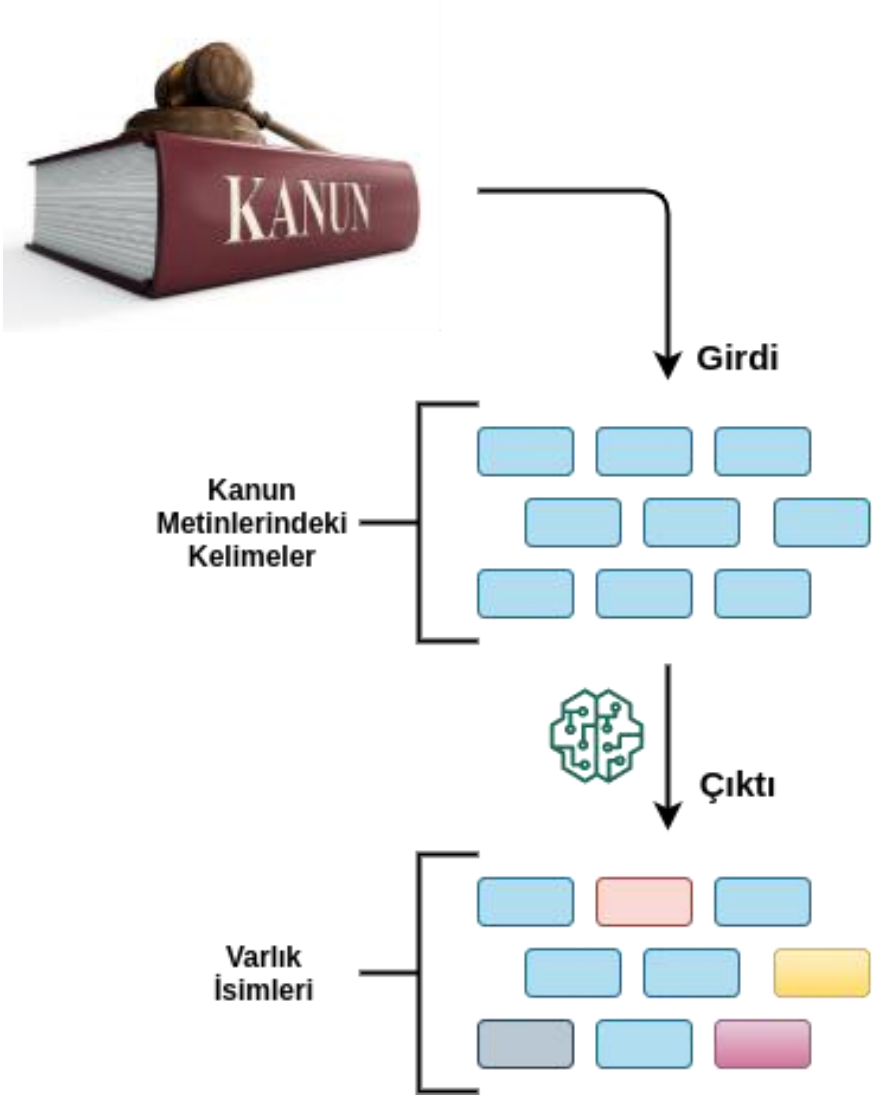
Yapılan Çalışma	F1 Skor (%)
Huang (2015)	90.1
Nichols ve Chiu (2016)	90.91
Hovy ve Ma (2016)	91.21
Yang (2016)	90.96
Lample (2016)	90.94
Zhang (2018)	91.22

Şekil 1 : Yapılan Çalışmaların F1 Skor Tablosu

3. PROJEDE KULLANILAN YÖNTEM VE MALZEME

3.1. PROJE ŞEMASI

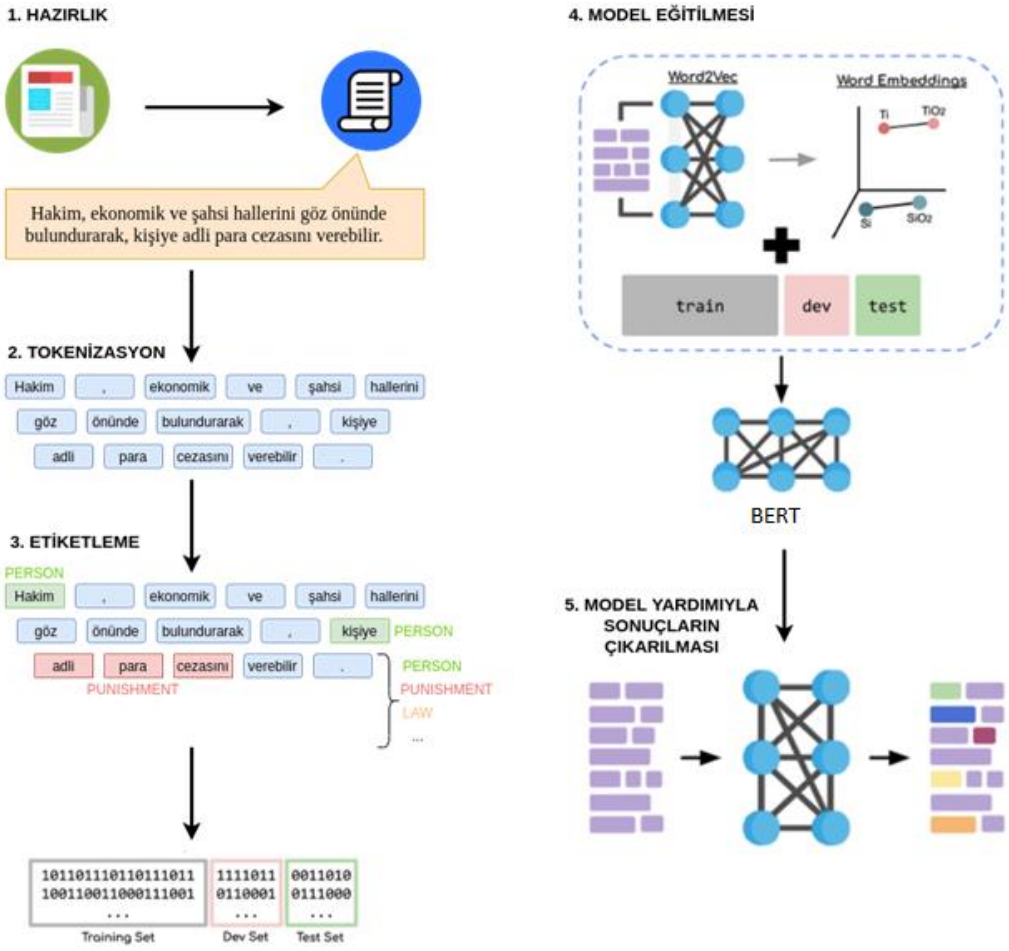
Proje şeması aşağıda görüldüğü gibidir. Kanun metinleri projeye girdi olarak verilir. BERT modeli kullanılarak bu metinlerdeki varlık isimleri tespit edilir.



Şekil 2 : Proje Şeması

3.2. PROJE TASARIM PLANI

Proje tasarım planı aşağıdaki gibidir.



Şekil 3 : Proje Tasarım Planı

3.3. KULLANILAN VERİ KÜMESİ

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 teker teker incelenmiş ve hangi varlık isimde oldukları belirlenmiştir. Veri kümesinde aşağıdaki etiketler kullanılmıştır:

ORGANIZATION : Kurum, kuruluş, topluluk, organizasyon isimleri

LAW : Kanun isimleri

PERSON : İnsan isimleri

RIGHT : Kanunlarda geçen hakların isimleri

PUNISHMENT : Ceza isimleri

TERM : Bir dönemi temsil eden isimler

TIME : Kanunlarda geçen saatler için kullanılan isimler

OFFENSIVE_WEAPON : Kanunlarda geçen silahların isimleri

GPE : Sıradağlar, deniz isimleri

LOC : Yer isimleri (ülke isimleri gibi)

NORP : Milliyet isimleri

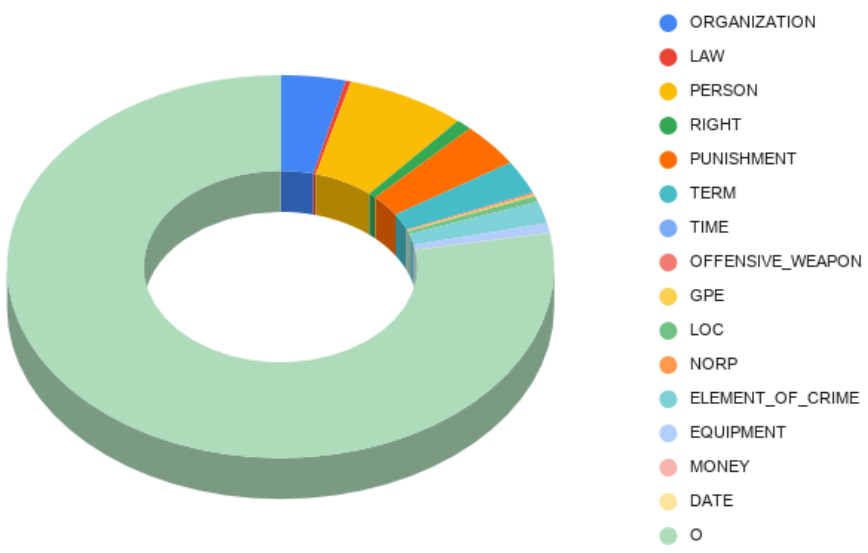
ELEMENT_OF_CRIME : Suç unsuru isimleri

EQUIPMENT : Kanunlarda geçen özel ekipmanlar

MONEY : Para birimi isimleri

DATE : Tarih isimleri

O : Hiçbir etikete girmeyen diğer isimler



Şekil 4 : Veri Kümesindeki Etiket Dağılımı

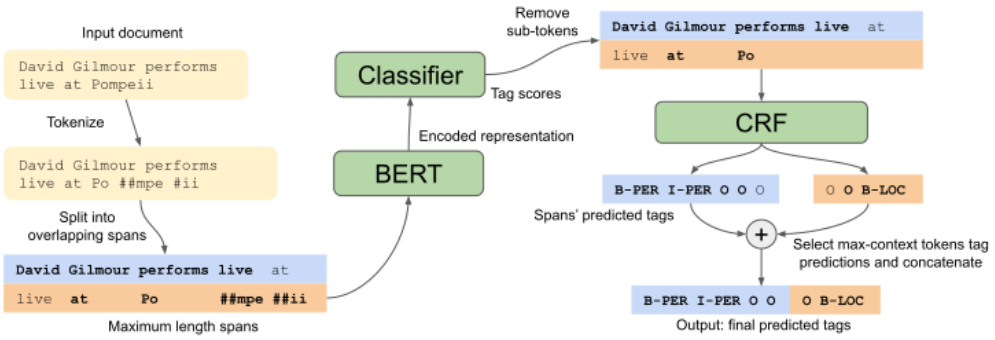
Kanun metinleri MBS'den indirildikten sonra eğik yazı ile yazılan kanunların bilgilendirici metinleri, madde numaraları, parantez içinde yazılan sayılar kaldırılır. Veri kümesi oluşturulurken kanun metinlerindeki sözcüklerin büyük harfle başlayıp, başlamadığı kontrol edilir. Büyük harfle başlayan sözcükler alınarak özel isim olup olmadıklarına bakılır. Daha sonra her sınıfa ait olan kurallar kelimelerin teker teker incelenmesiyle uygulanır.

3.4. BERT MODELİ

Google şirketi tarafından geliştirilmiş bir derin öğrenme yöntemidir. Büyük veri kümesi ile eğitilen genelleştirilmiş bir modeldir. Dildeki tüm dil bilimsel yapıyı bir çerçevede tutmuş, gerekli olduğunda diğer küçük dil işleme görevleri için transfer edilmesine olanak sağlamaktadır.

Transfer öğrenme, bir görev için model eğitildiğinde diğer görevler için bu önceden eğitilen model kullanılarak geliştirme becerisi olan derin öğrenme kavramını ifade eder. Projede önceden eğitilmiş olan bir BERT Modeli üzerinden transfer öğrenme tekniği kullanılmıştır.

BERT modeli, girdi metinlerini hem soldan hem sağdan değerlendirmektedir. Bu sayede kelimelerin anlamı ve kimelerin birbirleri ile olan ilişkilerinin daha iyi çıkarılması planlanmıştır. Girdi olarak verilen metinde ikinci cümlenin ilk cümlenin devamı olup olmadığı tahmin edilir.

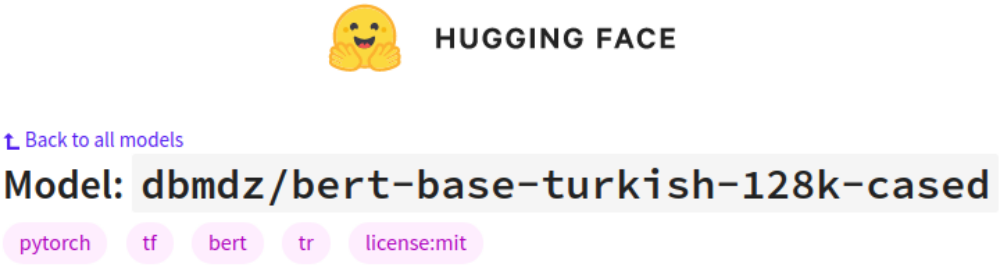


Şekil 5 : BERT Model Yapısı

BERT modeli Google tarafından ilk duyurulduğunda 800M kelime hazinesine sahip olan BookCorpus ve 2.5B kelime hazinesine sahip olan Wikipedia veriseti kullanılarak bert_large ve bert_base adı verilen 2 model şeklinde duyurulmuştur. Bert_large 16 adet TPU, bert_base ise 4 adet TPU ile 4 gün boyunca eğitilmiştir.

Projeye ilk başta veri kümesi 15.000 kelimeden oluştuğu sırada parametre analizi yapılmıştır. Bu analizde BERT modelleri için kullanılan optimum parametreler belirlenmeye çalışılmıştır. BERT modelleri Hugging Face açık kaynak kodlu kütüphaneleri barındıran siteden indirilmiştir.

BERT, çift-yönlü olması dışında MLM ve NSP adı verilen iki teknikle eğitilmektedir. Bir cümle modele girdiğinde, cümledeki kelimelerin %15'inde MLM tekniği kullanılır. Bu tekniğin kullanıldığı kelimelerin %80'i [MASK] etiketi ile, %10'u rastgele başka bir kelimeyle değiştirilmektedir. Geri kalan %10 da değiştirilmeden bırakılır. %15'lik değer, çok fazla kelimeyi maskelemenin eğitimi çok zorlaştırdığını, çok az kelimeyi maskelemenin de cümledeki içeriğin çok iyi kavranamama durumuna sebep olduğu için Google tarafından belirlenmiştir. MLM tekniğinde, maskelenen kelime, açık şekilde beslenen kelimelerle tahmin edilmeye çalışılır. (MLM'de sadece maskelenen kelimeler tahmin edilmeye çalışılır, açık olan veya üzerinde işlem uygulanmayan kelimelerle ilgili herhangi bir tahmin bulunmaz. bu sebeple kayıp değeri sadece işlem uygulanan kelimeler üzerinden değerlendirilir). İlk teknikte, cümle içerisindeki kelimeler arasındaki ilişki üzerinde durulurken, ikinci teknik olan NSP'de ise cümleler arasındaki ilişki kurulur. Eğitim esnasında ikili olarak gelen cümle çiftinde, ikinci cümlenin ilk cümlenin devamı olup olmadığı tahmin edilir. Bu teknikten önce ikinci cümlelerin %50'si rastgele değiştirilir, %50'si ise aynı şekilde bırakılır. Eğitim esnasındaki optimizasyon, bu iki tekniğin kullanılırken ortaya çıkan kaybın minimuma indirilmesidir. [12]



Şekil 6 : Projede Kullanılan Model İsmi

Yapılan literatür araştırmalarında genel anlamda BERT modelleri için kullanılan optimum parametreler çıkarılmıştır.

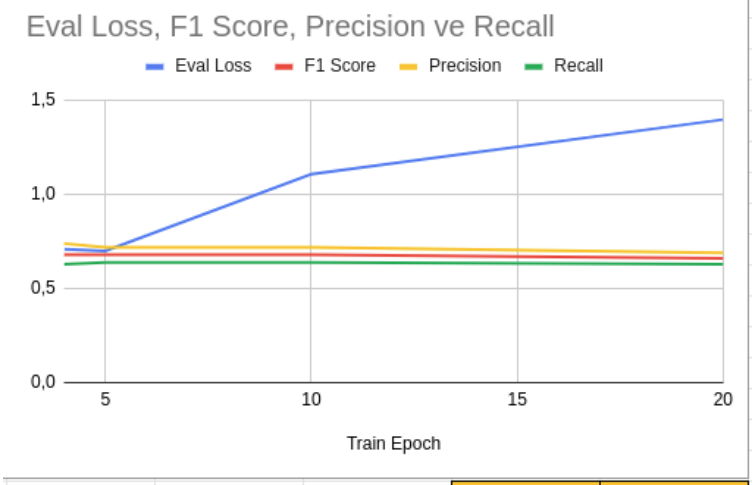
BERT MODELLERİ OPTİMUM PARAMETRE DEĞERLERİ			
Learing Rate	2,00E-05	3,00E-05	5,00E-05
Batch Size	8	16	32
Train Epoch	3	4	5

Şekil 7 : BERT Modelleri İçin Tavsiye Edilen Parametreler

Learing Rate	
1,00E-05	0,00001
2,00E-05	0,00002
3,00E-05	0,00003
5,00E-05	0,00005

Şekil 8 : Model Learning Rate Parametre Değerleri

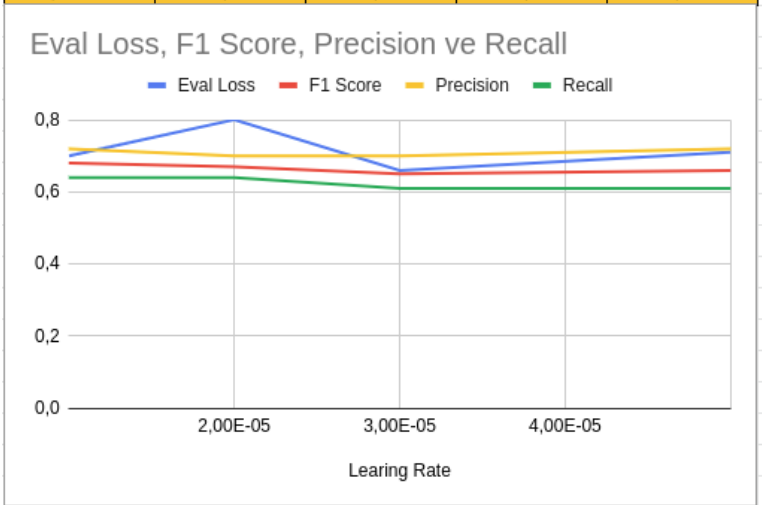
Train Epoch	Eval Loss	F1 Score	Precision	Recall
4	0,71	0,68	0,74	0,63
5	0,7	0,68	0,72	0,64
10	1,11	0,68	0,72	0,64
20	1,4	0,66	0,69	0,63



Learing Rate	1,00E-05
Train Batch Size	8
Eval Batch Size	32

Şekil 9 : Parametre Analizi 1

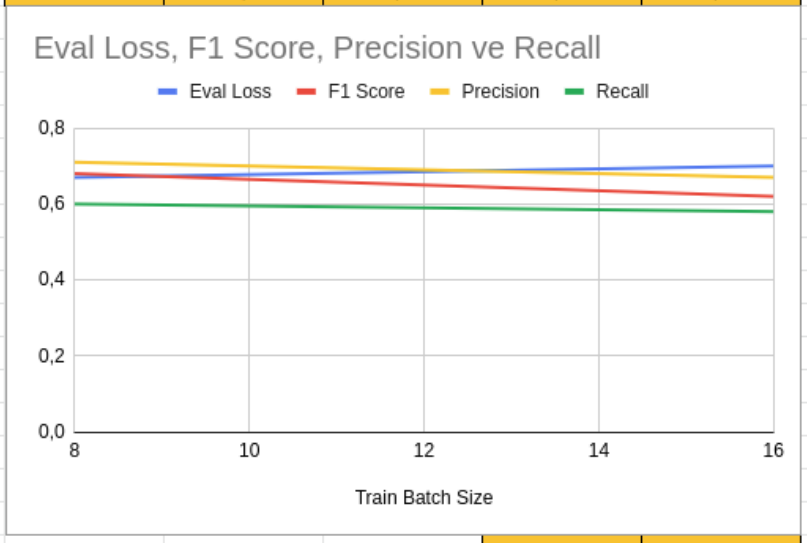
Learing Rate	Eval Loss	F1 Score	Precision	Recall
1,00E-05	0,7	0,68	0,72	0,64
2,00E-05	0,8	0,67	0,7	0,64
3,00E-05	0,66	0,65	0,7	0,61
5,00E-05	0,71	0,66	0,72	0,61



Train Epoch	5
Train Batch Size	8
Eval Batch Size	32

Şekil 10 : Parametre Analizi 2

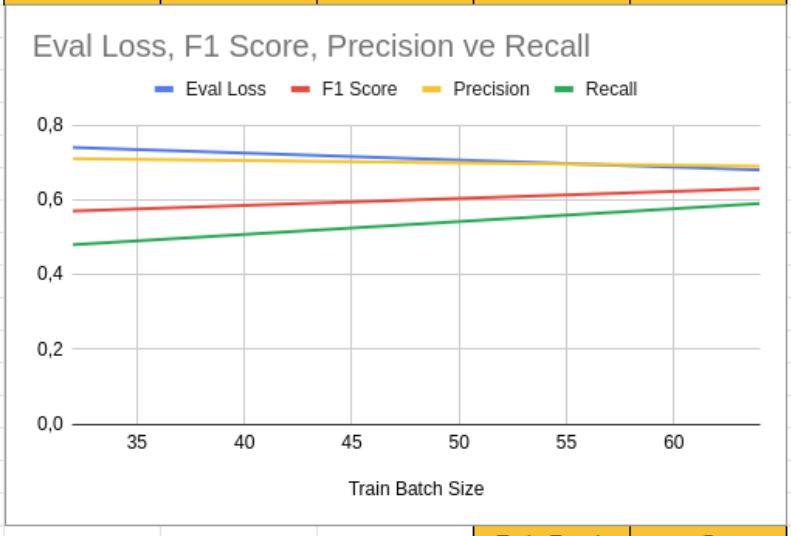
Train Batch Size	Eval Loss	F1 Score	Precision	Recall
8	0,67	0,68	0,71	0,6
16	0,7	0,62	0,67	0,58



Train Epoch	5
Learing Rate	3,00E-05
Eval Batch Size	32

Şekil 11 : Parametre Analizi 3

Eval Batch Size	Eval Loss	F1 Score	Precision	Recall
32	0,74	0,57	0,71	0,48
64	0,68	0,63	0,69	0,59



Train Epoch	5
Learing Rate	3,00E-05
Train Batch Size	8

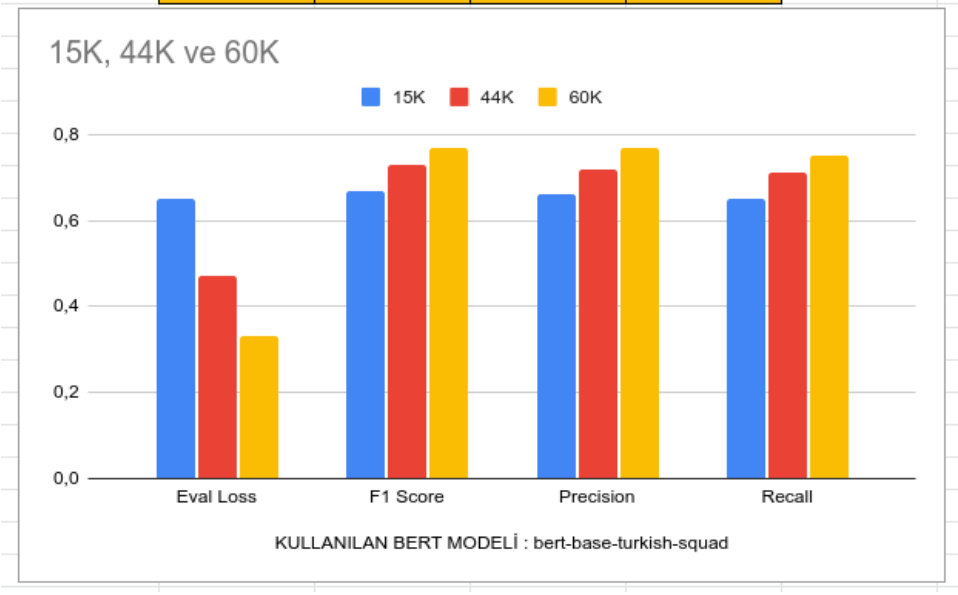
Şekil 12 : Parametre Analizi 4

Yapılan analizler sonucunda BERT modelleri için uygun bulunan parametreler kullanılan modelde daha doğru sonuçlar vermiştir ve çalışmalara bu parametrelerle devam edilmiştir.

Bu parametrelerden train_epoch 3, learning_rate 3e-5, train_batch_size 8 ve eval_batch_size 32 olarak belirlenmiştir. Bu analiz yapıldığında veri kümesi fazla büyük olmadığı için hata değeri görece yüksek, başarı oranları görece daha az çıkmıştır.

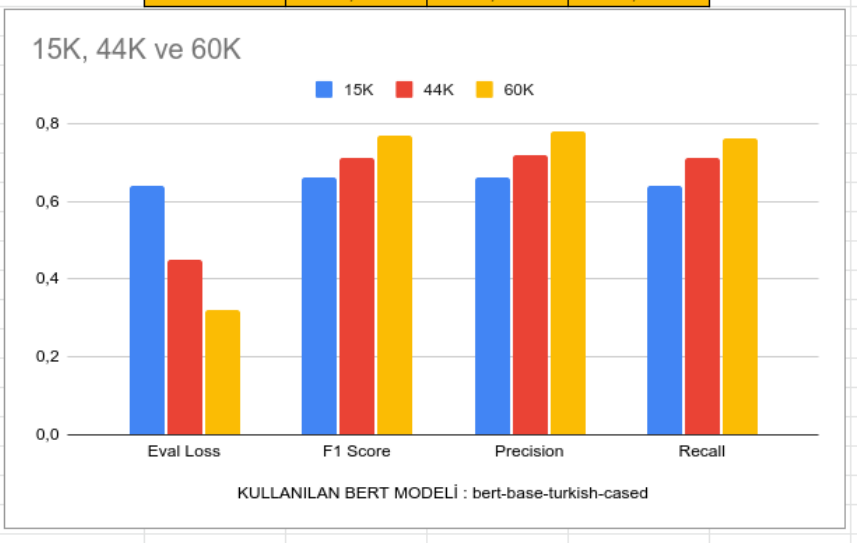
Veri kümesi genişletildikten sonra farklı BERT modelleri üzerinden tranfer öğrenmesi uygulanmıştır. Model parametreleri bir önceki analizdeki gibi seçilmiştir. Bu analiz sonucunda ise "bert-base-turkish-128k-cased" modelinin kullanılması uygun görülmüştür. [13]

KULLANILAN BERT MODELİ : bert-base-turkish-squad			
	Veri Kümesinin Büyüklüğü		
	15K	44K	60K
Eval Loss	0,65	0,47	0,33
F1 Score	0,67	0,73	0,77
Precision	0,66	0,72	0,77
Recall	0,65	0,71	0,75



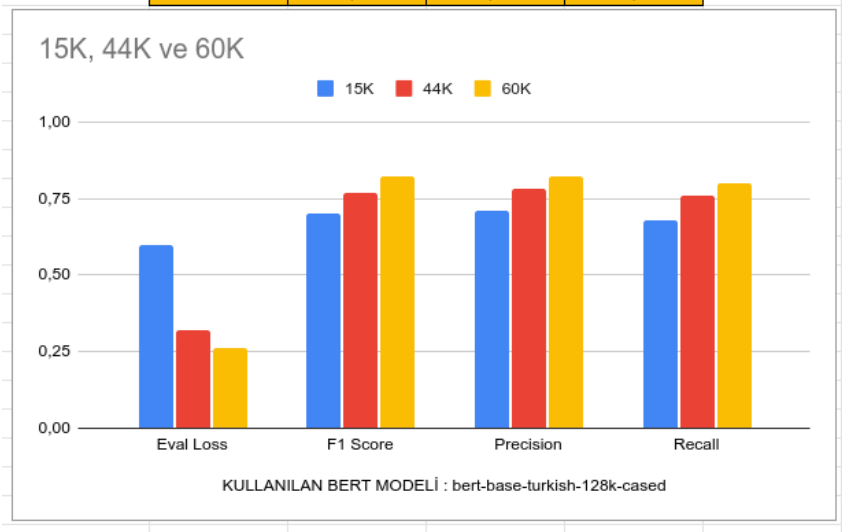
Şekil 13 : BERT Modelleri Analizi 1

KULLANILAN BERT MODELİ : bert-base-turkish-cased			
	Veri Kümesinin Büyüklüğü		
	15K	44K	60K
Eval Loss	0,64	0,45	0,32
F1 Score	0,66	0,71	0,77
Precision	0,66	0,72	0,78
Recall	0,64	0,71	0,76



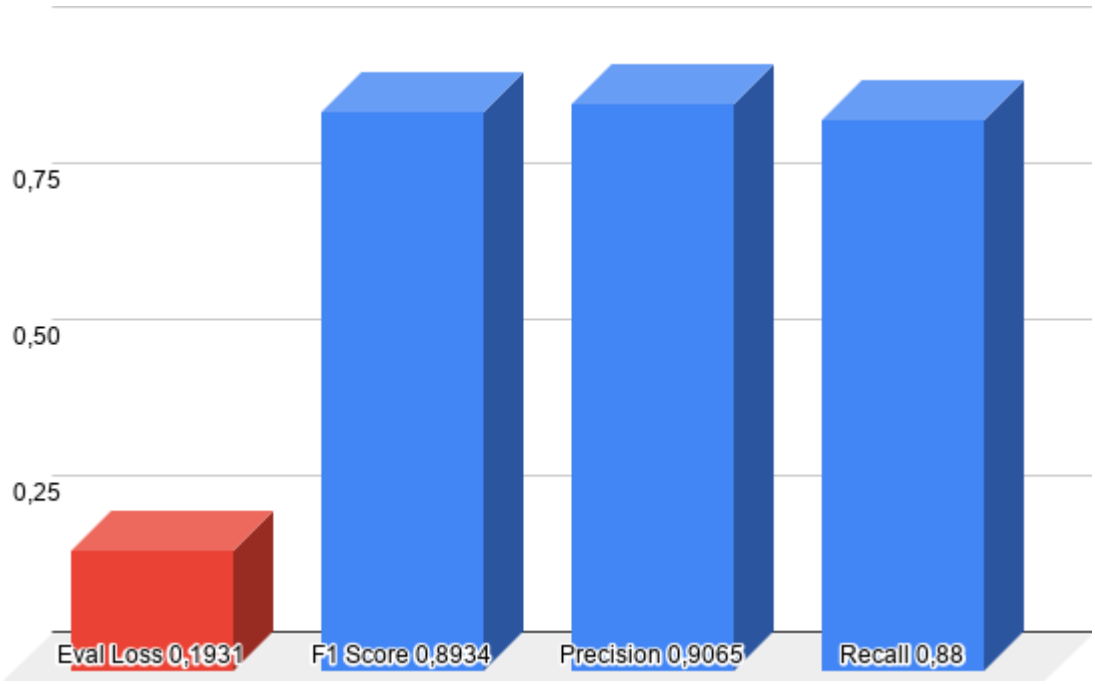
Şekil 14 : BERT Modelleri Analizi 2

KULLANILAN BERT MODELİ : bert-base-turkish-128k-cased			
	Veri Kümesinin Büyüklüğü		
	15K	44K	60K
Eval Loss	0,6	0,32	0,26
F1 Score	0,7	0,77	0,82
Precision	0,71	0,78	0,82
Recall	0,68	0,76	0,8



Şekil 15 : BERT Modelleri Analizi 3

Modelin eğitiminden sonra istenilen başarı oranları yakalanmadığı için veri kümesi tekrardan gözden geçirildi. Veri kümesinde bulunan yanlışlıklar giderildikten sonra istenilen başarı oranları yakalanmıştır. Modelin başarı oranları aşağıdaki grafikte gösterilmiştir.



Şekil 16 : Model Başarı Parametreleri

4. SONUÇ

Bu projede en yeni derin öğrenme metotlarından biri olan BERT ile Türkçe kanun metinlerindeki varlık isimlerini bulan bir model eğitilerek, geliştirilmiştir. Bu model için oluşturulan veri kümesinde 16 adet etiket kullanılmıştır. 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.

KAYNAKLAR

[1] SARI Ö. C., AKTAŞ Ö., *A NAMED ENTITY RECOGNITION MODEL FOR TURKISH LECTURE NOTES IN HISTORY AND GEOGRAPHY DOMAINS,* Research Article

[2] DALKILIÇ F., GELİŞLİ S. DİRİ B., *SIU2010 - 18.Sinyal İşleme ve İletişim Uygulamaları Kurultayı - Diyarbakır,* IEEE

[3] Yazarlar Gizlenmiştir, *Özyinelemeli sinir aglarıyla Türkçe varlık ismi tanıma*, 1993.

[4] CHENBIN Li, GUOHUA Z., ZHIHUA Li, News Text Classification Based on Improved, *2018 9th International Conference on Information Technology in Medicine and Education*, 7, IEEE, 2018

[5] Z. Huang, W. Xu, and K. Yu. *Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991*, 2015.

[6] J.P.C. Chiu and E. Nichols. *Named entity recognition with bidirectional lstm cnns. Transactions of the Association for Computational Linguistics, 4:357– 370,* 2016.

[7] X. Ma and E. Hovy. *End-to-end sequence labeling via bi-directional lstm cnns-crf. arXiv preprint arXiv:1603.01354*, 2016.

[8] Z. Yang, R. Salakhutdinov, and W. Cohen. *Multi-task cross-lingual sequence tagging from scratch. arXiv preprint arXiv:1603.06270*, 2016.

[9] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. *Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360*, 2016.

[10] Y. Zhang, H. Chen, Y. Zhao, Q. Liu, and D. Yin. *Learning tag dependencies for sequence tagging. In IJCAI, pages 4581–4587*. 2018.

[11] J. Yang, S. Liang, and Y. Zhang. *Design challenges and misconceptions in neural sequence labeling. arXiv preprint arXiv:1806.04470*, 2018.

[12] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L. KAISER L., POLOSUKHIN I., (Google Researchers), *Attention Is All You Need,* 6 December 2017

[13] HUGGING FACE [online],
https://huggingface.co/dbmdz/bert-base-turkish-128k-cased
[Ziyaret Tarihi: 29 Ekim 2020]