MAKİNE ÖĞRENİMİ YARDIMIYLA HASTA DOKÜMANLARI İÇERİSİNDEKİ LABORATUVAR DEĞERLERİNİN YAKALANIP İLİŞKİ KURULARAK OTOMATİK BİR ŞEKİLDE EŞLEŞTİRİLMESİ

Mert İPAR
191402110

YÜKSEK LİSANS TEZİ
Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı
Danışman: Dr. Öğr. Üyesi Erdal GÜVENOĞLU

İstanbul
T.C. Maltepe Üniversitesi
Lisansüstü Eğitim Enstitüsü
Ağustos, 2022

MAKİNE ÖĞRENİMİ YARDIMIYLA HASTA DOKÜMANLARI İÇERİSİNDEKİ LABORATUVAR DEĞERLERİNİN YAKALANIP İLİŞKİ KURULARAK OTOMATİK BİR ŞEKİLDE EŞLEŞTİRİLMESİ

Mert İPAR
191402110
Orcid: 0000-0001-7724-1687

YÜKSEK LİSANS TEZİ
Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Yüksek Lisans Programı
Danışman: Dr. Öğr. Erdal GÜVENOĞLU

İstanbul
T.C. Maltepe Üniversitesi
Lisansüstü Eğitim Enstitüsü
Ağustos, 2022



JÜRİ VE ENSTİTÜ ONAYI

Bu belge, Yükseköğretim Kurulu tarafından 19.01.2021 tarihli *"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"* ile bildirilen 6689 Sayılı Kişisel Verilerin Korunması Kanunu kapsamında gizlenmiştir.

ETİK İLKE VE KURALLARA UYUM BEYANI

Bu belge, Yükseköğretim Kurulu tarafından 19.01.2021 tarihli *"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"* ile bildirilen 6689 Sayılı Kişisel Verilerin Korunması Kanunu kapsamında gizlenmiştir.

TEŞEKKÜR

Tez çalışmasında desteklerini esirgemeyen değerli akademisyen ve aynı zamanda tez danışmanım saygı değer Dr. Öğr. Üyesi Erdal Güvenoğlu'na, değerli ve saygı değer eşim ve aileme sonsuz minnetlerimi sunuyorum.

Mert İPAR

Ağustos 2022

ÖZ

MAKİNE ÖĞRENİMİ YARDIMIYLA HASTA DOKÜMANLARI İÇERİSİNDEKİ LABORATUVAR DEĞERLERİNİN YAKALANIP İLİŞKİ KURULARAK OTOMATİK BİR ŞEKİLDE EŞLEŞTİRİLMESİ

Mert İPAR

Yüksek Lisans Tezi

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı

Danışman: Dr. Öğr. Üyesi Erdal GÜVENOĞLU

Maltepe Üniversitesi Lisansüstü Enstitüsü, 2022

Makine öğrenimi ile laboratuvar parametreleri yakalayıp, arasındaki ilişkiyi çözerek ve yeni bir akış sağlaması hedeflenerek bir yapı kurulmuştur. Metin içerisinde bulunan laboratuvar değerlerini makine öğrenimi yardımıyla ve modeller aracılığıyla bulup birbiri ile alakalı olan değerlerini eşleştirip sonuçları elde ettiğimiz bir yapı oluşturulmuştur.

Kullanılan yöntemler ve modeller aracılığıyla laboratuvar parametreleri ve değerlerini yakalamada en iyi sonuç elde edilmek için hedef olan 100 cümle ile %80 oranında başarılı bir sonuç elde edilmiştir. Eğitilen cümleler arasında farklı kombinasyonlar yaparak da test edilebilen bu yapı sayesinde medikal datadan birden fazla değeri etiketleyip bize sonuç olarak sunabilmektedir. Karşılaşılan ve yeni olan her durum için yeniden regex yazmak yerine bunu makine öğrenimi yardımı ile yakalayıp istediğimiz veriyi sunabilmekteyiz.

Birkaç denemenin ardından medikal veri için en iyi model ve yöntemler birleştirilmiş ve eğitim ile artacak bir başarı oranı elde edilmiştir.

Anahtar Sözcükler: Yazılım, Makine Öğrenimi, Yapay Zeka, Yazılım Geliştirme, Doğal Dil İşleme, Data Analizi.

ABSTRACT

METHOD RECOMMENDATION ON CAPTURING AND LINKING LABORATORY VALUES IN PATIENT DOCUMENTS WITH MACHINE LEARNING HELP AND MATCHING THEM IN AN AUTOMATED MANNER

Mert İPAR
Master Thesis
Department of Computer Engineering
.Computer Engineering Programme
Thesis Advisor: Dr. Öğr. Üyesi Erdal GÜVENOĞLU
Maltepe University Graduate School, 2022

A structure has been established by capturing the laboratory parameters with machine learning, solving the relationship between them and aiming to provide a new flow. A structure was created in which we find the laboratory values in the text with the help of machine learning and models and match the values that are related to each other and obtain the results.

In order to achieve the best result in capturing laboratory parameters and values through the methods and models used, a successful result of 80% was achieved with 100 sentences, which is the target. Thanks to this structure, which can be tested by making different combinations between the trained sentences, it can label more than one value from the medical data and present it to us as a result. Instead of rewriting regex for every encountered and new situation, we can capture it with the help of machine learning and present the data we want.

After several trials, the best models and methods for medical data were combined and a success rate that would increase with training was achieved.

Keywords: The Software, Machine Learning, Artificial Intelligence, Software Development, Natural Language Processing, Data Analysis.

İÇİNDEKİLER

JÜRİ VE ENSTİTÜ ONAYI ..... **i**
ETİK İLKE VE KURALLARA UYUM BEYANI ..... **ii**
TEŞEKKÜR ..... **iii**
ÖZ ..... **iv**
ABSTRACT ..... **v**
İÇİNDEKİLER ..... **vi**
TABLOLAR LİSTESİ ..... **viii**
ŞEKİLLER LİSTESİ ..... **ix**
KISALTMALAR LİSTESİ ..... **x**
ÖZGEÇMİŞ ..... Hata! Yer işareti tanımlanmamış.
BÖLÜM 1. GİRİŞ **1**
BÖLÜM 2. LİTERATÜR ARAŞTIRMASI **3**
BÖLÜM 3. MATERYAL VE YÖNTEM **6**
3.1. Makine Öğrenimi 6
3.1.1. Gözetimli Makine Öğrenmesi 7
3.1.2. Gözetimsiz Makine Öğrenmesi 8
3.1.3. Yarı Gözetimli Öğrenme 8
3.2. Nlp (Doğal Dil İşleme) 8
3.3. Python 9
3.4. Nltk 10
Tokenizing 10
Word Tokenize 10
Sentence Tokenize 11
3.5. Pandas 11
3.6. Numpy 12
3.7. Mysql 13
3.8. Regex 13
3.9. Json 13

3.10. Rnn 14
İki yönlü Rnn'ler (Bidirectional Rnns): 15
Derin (İki Yönlü) Rnn'ler (Deep Rnns): 16
3.11. Lstm 17
3.12. Ner (Name Entity Recognition) 19
3.13. Bert 19
3.14. Flair 20
3.15. FastText 20
3.16. Gerçekleştirilen Uygulama 20
3.17. Elde Edilen Sonuçlar 30
3.18. Bulgular 33
BÖLÜM 4. SONUÇ 35
KAYNAKÇA 36

TABLOLAR LİSTESİ

Tablo 3.1.: *Laboratuvar Parametreleri Ve Eş Anlamları*..... 23

ŞEKİLLER LİSTESİ

Şekil 3.1. *Makine Öğrenimi Sınıflandırıcıları* 6

Şekil 3.2. *Makine Öğrenimi Algoritmaları* 7

Şekil 3.3. *Nlp İle Yapılabilecek Örnekler* 9

Şekil 3.4. *Nlp İle Yapılabilecek Örnekler* 12

Şekil 3.5. *Rnn'nin Bir Ağa Açılımı* 15

Şekil 3.6. *Tekrarlayan Sinir Ağı (Rnn)* 15

Şekil 3.7. *Çift Yönlü Rnn* 16

Şekil 3.8. *Derin Çift Yönlü Rnn* 16

Şekil 3.9. *Tekrarlayan Sinir Ağına Örnek* 17

Şekil 3.10. *Lstm Mimari Yapısı* 18

Şekil 3.11. *Ner Modelinin Çıktısı* 27

Şekil 3.12. *İlişki Modeli Çıktısı - 1* 28

Şekil 3.13. *İlişki Modeli Çıktısı - 2* 29

Şekil 3.14. *İlişki Modeli Çıktısı - 3* 30

Şekil 3.15. *Modellerin Testinde Kullanılan Örnek Cümleler* 30

Şekil 3.16. *Makine Öğrenimi Cümle girişi – 1* 31

Şekil 3.17. *Makine Öğrenimi Cümle Çıktısı – 1* 31

Şekil 3.18. *Makine Öğrenimi Cümle Girişi – 2* 33

Şekil 3.19. *Makine Öğrenimi Cümle Çıktısı – 2* 33

KISALTMALAR

AI	**:**	Artificial Intelligence
MR	**:**	Magnetic resonance
ML	**:**	Machine Learning
JSON	**:**	JavaScript Object Notation
NLTK	**:**	Natural Language Toolkit
LSTM	**:**	Long Short-Term Memory
RNN	**:**	Recurrent Neural Network
NER	**:**	Name Entity Recognition
POS	**:**	Part-of-Speech
PCA	**:**	Principal Component Analysis
SVD	**:**	Singular Value Decomposition
BERT	**:**	Bidirectional Encoder Representations from Transformers

BÖLÜM 1. GİRİŞ

Dünyada birden fazla teknoloji ile insan zamanının tasarrufu üzerine yoğunlaşıp en kolay şekilde sonuca ulaşmak için tasarlanmaktadır. Elimizde insan gücünün sınırlarını gördükten sonra birçok işi makinalar yardımıyla yapmayı planlamaktayız. Gelecekte insan gücünün belki de sadece yönetim kısmında var olacağı kaçınılmaz bir gerçektir. Bilgi teknolojileri alanında da bu aşamalı bir biçimde ilerleyen durumdur. Genel bir biçimde oluşturulan her işin hızı bizler için önem arz etmektedir. Bu aşamada dünyada hali hazırda kullanılan birçok teknoloji geliştirilmektedir.

Medikal anlamda hastanelerin, ilaç firmalarının, sigorta şirketlerinin birçok hastası ve elinde dokümanları bulunmaktadır. Bu dokümanlar içinde sayısız veri bulunmaktadır. Burada asıl dikkat edilmesi gereken kısımda hangi doktor için hangi bilgiye ihtiyaç duyulacağıdır. İnternet ortamında yaptığımız her işlem aslında arka planda bilmediğimiz bir amaca hizmet etmektedir. Bunun adı da veri işlemedir. Yaptığımız her işlem aslında bizim gözle göremediğimiz boş bir depoyu veri ile doldurmaktadır. Bu şekilde bir sonraki işlemlerde her defasında size zamandan kazanıp aynı anda da yeni veriler toplamayı hedeflemektedir. Medikal bilginin içerisinde var olan değerlerin en büyük kısmı sonuçlardır. Bu alanda daha hızlı hareketi sağlayacak bir ortam yaratmak hastanın ve bu dokümanları okuyanların yorumlayabilmesi için her zaman daha avantaj yaratan bir durum olabilmektedir. Aksi taktirde dakikalarca kontrol edip sonuçları bulmanın ardından yorumlaması ciddi zaman kaybına yol açabilmektedir.

Makine öğrenimi yeni teknolojiler içerisinde yaratılmış bir yapay beyin olarak düşünülmektedir. Aslında küçük bir bebek olarak da nitelendirebilirsiniz. Bir bebek doğumundan sonra sürekli olarak çevresini gözlemleyerek anlamaya çalışmaktadır. Yeri geldiğinde taklit eder, yeri geldiğinde hareketi kendi başında yapar. Bu durumu çok fazla çeşitlendirebilmektedir. Makine öğrenimi yardımıyla da metin olarak elimizde bulunan sayfalarca yazı arasında işimize en uygun olan 10 adet laboratuvar değerlerini yakalayıp en kısa sürede sonuç elde etmeyi hedeflenmektedir.

Hedeflenilen nokta iyi bir model yaratıp veya bulup her durum için yeni bir regex yazmak yerine, bir süre farklı cümle tiplerini eğitip sonrasında model aracılığıyla metin içerisinde otomatik olarak yakalanmasını sağlamaktır. Bu şekilde oluşacak karakter,

boşluk, rakam gibi farklı kombinasyon durumlarını en başarılı şekilde yakalayıp kullanıcıya sunmasıdır. Bu kısımda hedeflenen nokta hastanın medikal dokümanında metin olarak elimizde bulunmasıyla birlikte var olan laboratuvar değerlerinin otomatik olarak yakalanıp ilişkili bir şekilde eşleştirilmesini sağlanılmasıdır.

Bir hastanın sayfalarca hastalık ile alakalı dokümantasyonu olabilmektedir. Bunların içerisinde patoloji, mr (manyetik rezonans), laboratuvar, ameliyat, hastalık, yan etki gibi birçok sonucu bulunabilmektedir. Hedefimiz de doktor veya hemşirelerin bütün dokümantasyona bakıp laboratuvar değerlerini el ile sisteme girmek yerine, otomatik bir şekilde metni sisteme verip sonuçları elde etmelerini sağlamaktır. Bu alanda kullanacağımız yöntem ise makine öğrenimidir. Kullanılan diğer yöntemlerden bazıları sadece belirli kalıp cümleleri işleyerek herhangi bir karakter değişimi, yer değişimi gibi durumlarda işe yaramamaktadır. Biz ise yaptığımız model ve yöntemler daha yüksek başarı oranına sahip olmayı hedeflenmektedir.

Süreç içerisinde eğer yapılan model içerisinde verilen bilgiler hastaya ait değil ise veya yanlış bir ilişki kurup sonuç getirdiyse bu sorun çıkarabilecek bir ihtimal doğurabilmektedir. Burada eğitilen cümleler ile sorun giderilmektedir. Ancak uygulanan modeller arasında buna en uygun olan model seçileceği için ekstrem bir cümle üretilirse başarı oranı düşük çıkabilmektedir. Bu geliştirilebilir bir ihtimal olabilmektedir.

Tezin ikinci bölümünde literatür araştırılması yapılıp var olan bilgiler üzerinden çıkarımlar yapılarak kullanılmış teknolojiye değinilmektedir. Üçüncü bölümümüzde ise kullandığımız teknolojileri belirterek geliştirilmiş uygulamaların örnekli halleri yer almaktadır. Sonuç kısım olan dördüncü bölümümüzde hedeflenen durum üzerinden genel bir özetleme yapılmaktadır.

BÖLÜM 2. LİTERATÜR ARAŞTIRMASI

Makine öğrenimi yıllar önce hayatımıza adım atmış bir alt dal olarak gözlemlemekteyiz. İşlev olarak sürekli öğrenen bir yapıya sahiptir (Alkan, 2021). Bu alanda ilk çalışmaları IBM firmasında çalışan Arthur Samuel'in dama oyunu üzerinde olan araştırmasında kullanıldığı bilinmektedir. Profesyonel damacı Robert Nealey 1962 senesinde bir bilgisayarda dama oyununu oynamış ve kaybetmiştir (IBM, 2020). Burada yürütülen mantık düşünüldüğünde ise yapılan hamleye göre verilen tepki olarak düşünülebilinir.

Kullanılacak yöntem ihtiyaçtan doğmaktadır. Yani hasta verilerini hem zamandan hem de insan kaynağını minimal olarak kullanarak daha verimli bir yapı oluşturulabilmektedir. Şekil 3.1. makine öğreniminin türleri arasında kullanılan yöntemimiz ise öğrenim türleri arasında hem denetimli hem denetimsiz hem de yarı denetimli öğrenme şekillerini bir arada kapsamaktadır (Tamir, 2020).

Kullanılan hazır model ile denetimli öğrenme türüne girmektedir. Bunun sebebi ise denetimli öğrenme daha önce etiketlenmiş ve sınıflandırılmış veri kümesi kullanılmasındandır. Ancak yeni bir veri kümesi program içerisine verildiğinde bu cümle içerisinde hiçbir şey daha önce etiketlenmediği için denetimsiz öğrenme türüne ait olmaktadır. Etiketlenmemiş olan cümlelerin ve etiketlenmiş verinin de içinde bulunduğu veri ile programdan alınan çıktı bize öğrenme aşamasını göstermektedir. Bu da yarı denetimli öğrenme türünün uygulandığını ortaya çıkarmaktadır (Kaşgarlı, 2021:19).

Örneğin, tez incelemesi ve makaleler de intihal (bir fikri atıf yapmadan kullanma) skorlarını hesaplarken bir makine öğrenimi model geliştirilebilir. Bu alanda kullanılacak olan mantık ise benzer metin içeriğinin olup olmadığını kontrol edip sonrasında bir veri tabanı aracılığı ile arama yaptırıp sorgulamanın ardından içerik sahibine atıfta bulunulup bulunulmadığına göre bir skor belirlenebilir (Şafak, 2017).

Makine öğrenimi yaparken kullanılan diller arasında R, Python gibi birden fazla seçenek bulunmaktadır. Python ile yazımı kolay ve daha çok kütüphaneye ait olması öne çıkarmaktadır. İşlevsellik açısından birden fazla alanda kullanılmaktadır. Python, yazılımcıların da tercih ettiğini birden fazla zengin içeriğe sahip kütüphaneler ile anlaşılır ve kolay kod yapısı oluşturmaktadır (Eğitim 2021).

Çok boyutlu diziler için destek sağlayan Numpy adlı başka bir paketin üzerine inşa edilmiştir. En popüler veri tartışma paketlerinden biri olan Pandas, Python ekosistemindeki diğer birçok veri bilimi modülüyle iyi çalışır ve genellikle işletim sisteminizle birlikte gelenlerden ActiveState'in ActivePython'u gibi ticari satıcı dağıtımlarına kadar her Python dağıtımında bulunur (ActiveState, 2021).

Verinin temiz ve düzenli olması için Pandas kütüphanesini kullanarak veri kümeleri oluşturup içlerinde işlem yapabilmektesiniz. Bunun yanında Numpy kütüphanesi ile diziler arası işlemleri yapabilmekteyiz. Dizinlerde sayısal analizler gibi işlemlerin yanı sıra istatistik oranında operasyonel olarak da çalışabilmektedir (Doğan 2019).

Medikal alanda eğitimli bir model kullandığınız zaman bunu testlerini itinalı bir biçimde yapılması faydalı olmaktadır. Bunun için operasyonel olarak iyi bir işleyiş oluşturması gerekmektedir. Flair modeli ile daha önce 1000000+ veri ile eğitilmiş ve etiketlenmiş veri barındıran bir model aracılığıyla doğruluk oranı en yüksek etiketleme işlemini elde edebilmektedir (Schweter ve Golde, 2022).

Bu yöntem cümle içerisindeki bilgileri etiketleyerek ayıklama görevini yerine getirmektedir. Etiketlenmiş her kelime öncesinde belirlenmiş kategoride gruplanmıştır. Öncelikle var olarak kelimeyi bulmaktadır. Sonrasında bunu daha önceden etiketlenmiş veri üzerinden kategorize etmektedir. Bunu da İç-dış-başlangıç (etiketleme) yöntemi ile yapmaktadır (Marshall, 2019).

Kullandığımız verilerin düzenli bir yapıda oluşması için Mysql veri tabanı kullanılmaktadır. Burada basit ama işlevsel bağlantılarla verilerin güncellenmesi, eklenmesi, silinmesi ve sorgulanması gibi işlemle yazdığımız kod ve veri arasında bağlantı kurarak işlemler yapabilmekteyiz (W3schools, 2022).

Var olan veri üzerinde Nltk kütüphanesiyle cümle yapısında oynamalar yaparak kelimelere bölüp, modeller arasında işlemler yapıp aynı zamanda da temizlik görevini de bu alanda da düzenleyebilmekteyiz (NLTK, 2022).

Laboratuvar değerlerini araştırıp gerekli tüm bilgileri elde edilmesi gerekmektedir. Öncelikle etiketlenmiş veri içerisinde analiz yaparak eş anlamlı olan veriyi ayırıp bunları karşılaştırarak doğruluğu kanıtlandıktan sonra eğitmeye ve farklı örnekle bularak da bu veriyi çoğaltabilmekteyiz (IAPAC, 2021).

Bazı cümlelerde temizlik yaparken Nltk ile kelimelere böldükten sonra regex yardımıyla kolayca bunlar üzerinde değişim sağlayıp alfabe içerisinde olmayan yanlış karakterlerin modeli bozmasını engellenmesini sağlayabilmekteyiz (Pythex, 2021).

Temizlenmiş cümleler ile modele verilen girdi cümleler etiketlenmeye başlanmaktadır. Burada Flair Ner modeli kullanılmaktadır. Bu sayede tahmin yürüterek kelimenin ne olduğunu anlamaya çalışır ve çıktılar vermektedir. Daha önceden etiketlenmiş bir veriyi verilen cümle içerisinde algılar sonrasında kategoriye ayırır ve çıktısını sunabilmektedir (Schweter ve Golde, 2022).

Modellerin eğitim yapısında ise Lstm ve Rnn mimarileri kullanılabilmektedir. Geriye dönük öğrenme, karmaşık öğrenme gibi birden farklı yöntemleri içinde barındığı bu yapı sayesinde anlamlı tahminler çıkarması hedeflenmektedir (Kumaraswamy, 2021).

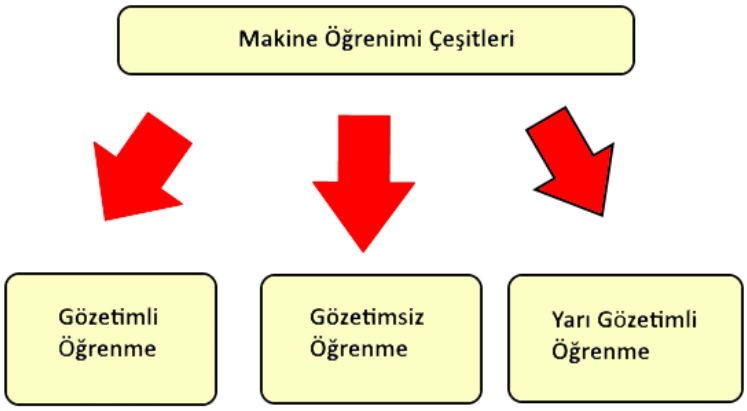
BÖLÜM 3. MATERYAL VE YÖNTEM

Makine öğrenimi çok farklı alanlarda yapılmaktadır. Görüntü işleme, ses işleme, metin işleme vb. gibi birçok alana hizmet etmektedir. Bu alanlar içerisinde her yapının farklı bir hiyerarşisi bulunabilmektedir. Kullanılan yöntemin amaca hizmet etmesi gerekmektedir. Bizim hedefimiz metin üzerinden makine öğrenimi yapıp tahmin edilebilirliği en iyi sonuçlar ile hedeflemektedir.

3.1. Makine Öğrenimi

Makine öğrenimi 1959 yılında bilgisayar biliminin yapay zekada sayısal öğrenme ve model tanıma çalışmalarından geliştirilmiş bir alt dalı olarak bilinmektedir. Makine öğrenmesi yapısal işlev olarak öğrenebilen ve veriler üzerinden tahmin yapabilen algoritmaların çalışma ve inşalarını araştıran bir sistem olarak düşünülmelidir. Bu tür algoritmalar statik program talimatlarını harfiyen takip etmek yerine örnek girişlerden veri tabanlı tahminleri ve kararları gerçekleştirebilmek amacıyla bir model inşa etme mantığıyla çalışmaktadırlar.

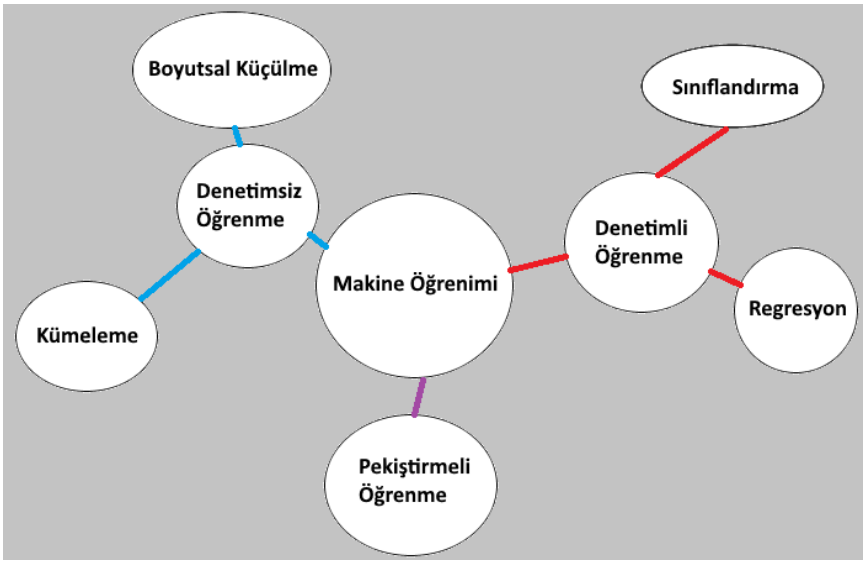
Kısaca belirtmek gerekirse makine öğrenmesi, insanların öğrenme şekillerini taklit etmek için veri ve algoritmaların kullanımına odaklanıp doğruluğunu kademeli olarak artıran bir yapay zeka (Ai) ve bilgisayar bilimi dalıdır. Makine öğrenmesi sınıflandırıcıları üç ana kategoriye ayrılır. Şekil 3.1. görüldüğü gibi 3 ana kategoriye ayrılmaktadır.



Şekil 3.1. *Makine Öğrenimi Sınıflandırıcıları*

Makine öğrenimi, matematiksel ve istatistiksel işlemler ile veriler üzerinden çıkarımlar yaparak tahminlerde bulunan sistemlerin bilgisayarlar ile modellenmesine anlamına gelen sisteme denilmektedir.

Şekil 3.2. de görüldüğü gibi makine öğreniminin belli başlı kullanım örnekleri yer almaktadır.



Şekil 3.2. *Makine Öğrenimi Algoritmaları*

3.1.1. Gözetimli Makine Öğrenmesi

Gözetimli öğrenme var olan verileri sınıflandırarak sonuçların doğru olarak tahmin edilmesini sağlayan algoritmaları eğitmesi için veri kümeleri kullanmasıyla tanınmaya denilmektedir. Model uygun hale gelene dek aktarılan girdi verilerinin ağırlıkları düzenlenmektedir. Eksik uyum veya aşırı uyumdan kaçınılmasını sağlayan çapraz doğrulamanın bir bölümü olarak gerçekleşmektedir. Gözetimli öğrenme, elektronik postalarda maillerin ayrı klasörlerde filtrelenmesi gibi, şirketlerin var olan sorunların ölçekli olarak çözümlerine yardımcı olmaktadır. Gözetimli öğrenmede kullanılan yöntemler arasında; neural networks, naïve bayes, linear regression, logistic regression, random forest, support vector machine (Svm) gibi yöntemler yer almaktadır.

3.1.2. Gözetimsiz Makine Öğrenmesi

Denetimsiz makine öğrenmesi olarak da adlandırılan, birçok makine öğrenmesi fonksiyonlarını çözümleyerek etiketlenmemiş veri kümeleri üzerinde kümeleme yapmak için kullanılmaktadır. İnsan eli değmeden veri grupları keşfeden algoritmalar olarak da düşünülmektedir. Bilgi içindeki benzerlikleri ve farklılıkları keşfetme yeteneği; bu algoritmayı, keşif amaçlı veri analizi, çapraz satış stratejileri, müşteri segmentasyonu, görüntü ve kalıp tanıma için ideal bir çözüm haline getirmektedir. Ek olarak modeldeki özellik sayısını, boyutluluk azaltma yoluyla düşürmek için de kullanılmaktadır. Pca(birincil bileşen analizi) ve Svd(tekil değer ayrıştırma) yaygın yaklaşım bulunmaktadır. Gözetimsiz öğrenmede kullanılan diğer algoritmalardan bazıları neural networks, k-means clustering, probabilistic clustering methods vb. şeklindedir (Tamir, 2020).

3.1.3. Yarı Gözetimli Öğrenme

Gözetimli ve gözetimsiz öğrenme arasında düşünülebilinir. Özellik ayrıştırmaya yol göstermek için daha küçük bir veri kümesi kullanmaktadır. Gözetimli öğrenme algoritmasını eğitmede yeterli veriyi etiketleyecek kapasitede olamama sorununu da çözmektedir (Tamir, 2020).

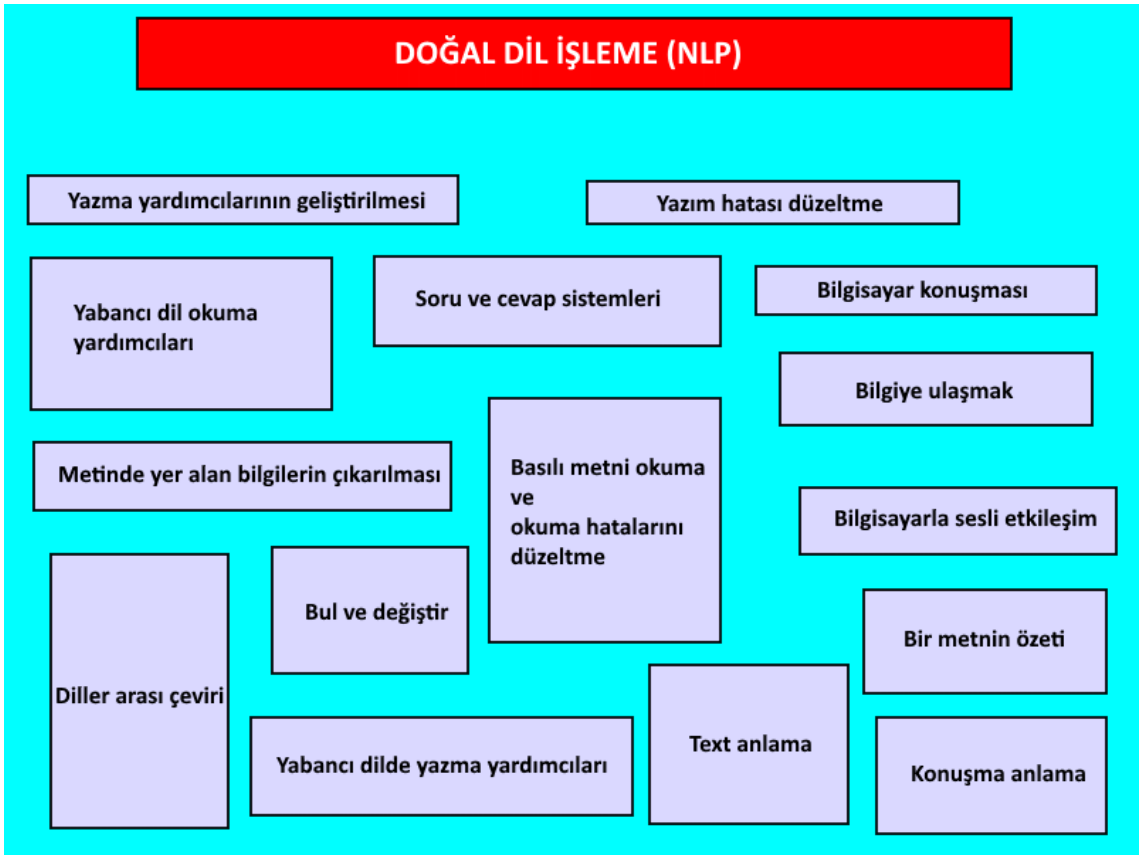
3.2. Nlp (Doğal Dil İşleme)

Bilgisayarların konuştuğumuz dili anlaması, işlemesi, yorum yapması ve hatta cümle üretebilmesi doğal dil işleme mantığına girmektedir (IBM, 2020). Doğal dil işleme, hesaplamalı bilimlerle (makine öğrenimi, yapay zeka) dil biliminin ortak kullanıldığı bir disiplindir olarak düşünülebilmektedir. Bugün banka sitesinde konuştuğumuz mesajlaşma robotu, telefonumuzdaki asistana verdiğimiz komutlar, google/microsoft çeviri de yaptığımız çeviriler, mesaj yazarken telefonumuzun bir sonraki kelimeyi tahmin etmesi gibi teknolojiler doğal dil işlemenin sonucu olarak düşünülmelidir. Bununla beraber yakın zamanlarda çokça kullanılan metin madenciliği (text mining) de doğal dil işlemeye dahildir. Metin madenciliği sayesinde internette yığınlar halinde biriken düşünceyi işleyip anlam çıkarabilmektedir. Örneğin, Twitter Api'yla iklim değişikliği içerikli tweetleri alıp tweeti yazanların yüzde kaçının iklim değişikliğini tehlikeli

bulduğunu öğrenip kamuoyu hakkında bilgi sahibi olabiliyoruz. Bunun dışında doğal dil işleme konuşma tanımada (speech recognition) kullanmaktadır.

Bunun gibi birçok alanda farklı yöntemler kullanılmaktadır. Diğer bir örnek de görüntü işleme yardımıyla trafikte bulunan kameralar gibi ciddi alanda işlevsel olarak hali hazırda da kullanılıp geliştirilmeye devam eden sistemlerdir.

Şekil 3.3. Nlp ile yapılabilecek uygulamalar yer almaktadır. Bu alanlar sağlık, teknoloji, doğal afetler gibi birçok alanda kullanılabilmektedir.



Şekil 3.3. *Nlp İle Yapılabilecek Örnekler*

3.3. Python

Python, nesne tabanlı, kolayca yorumlayabildiğimiz, modüler ve işlevsel olup kullanışlı bir programlama dilidir. Basit kod yapısı, dilin çözümlenmesinde ve basit pratiklerle akılda kalmasını kolaylaştırmaktadır. Bu da ona söz diziminin ayrıntıları ile

vakit yitirmeden programlama yapılmaya başlanabilen bir dil olma özelliği kazandırmaktadır.

Modüler yapısı sayesinde birçok veri yapısını desteklemektedir. Her türlü işletim sisteminde çalışabilmektedir. (Windows, Amiga, Symbian, Unix, Linux, Mac).

Python ile birçok alanda işlem yapabilir. Şu şekilde sıralayabiliriz:

- Web Geliştirme
- Robotik Kodlama
- Bilimsel ve Sayısal Hesaplama
- Gelişmiş İşlevsellik Sağlar
- Tarayıcı Web Uygulama Otomasyonu
- Makine Öğrenimi

3.4. Nltk

Doğal dil işleme (Nlp), insan etkileşimiyle bilgisayar arasında etkileşimi sağlayan, bilgisayarın insan dilini anlamasını sağlayan, dil incelemeleri için farklı birçok sınıfa sahip olan bir disiplin olarak düşünülmektedir. Örnek olarak, duygu analizi, metin özetleme sistemi, metin benzerliği hesabı, sanal asistanlar, çeviri sistemleri gibi birçok projeyi yapmamıza olanak sağlamaktadır (NLTK, 2022).

Tokenizing

Tokenize ederek, metni kelimeye veya cümleye göre kolayca bölebilmektedir. Bu, metnin geri kalanının bağlamı dışında bile nispeten tutarlı ve anlamlı olan daha küçük metin parçalarıyla çalışmamıza olanak tanımaktadır. Bu, yapılandırılmamış verileri, analiz edilmesi daha kolay yapılandırılmış verilere dönüştürmek için ilk adımımız olabilmektedir (NLTK, 2022).

Word Tokenize

Bir cümleyi kelimelere ayırmak istersek eğer bunun için word_tokenize() yöntemini kullanılmaktadır. Kelime tokenize çıktısı, makine öğrenimi uygulamalarında metnin daha iyi anlaşılması için veri çerçevesine dönüştürülebilmektedir. Noktalama işaretlerini kaldırma, sayısal karakterleri kaldırma veya kök alma gibi diğer metin temizleme adımları için girdi olarak da

sağlanabilmektedir. Makine öğrenimi modelleri eğitilmek ve bir tahminde bulunmak için sayısal verilere ihtiyaç duymaktadır. Kelime belirleme, metnin (dize) sayısal veriye dönüştürülmesinin çok arz eden durum haline bürünebilmektedir (NLTK, 2022).

Sentence Tokenize

Cümleleri tek tek bölmek için kullanılan bir yönteme denilmektedir. Cümle başına ortalama kelime saymanız gerektiğini düşündüğümüzde Sentence Tokenize yardımıyla hesaplaması kolaylaşmaktadır. Böyle bir görevi gerçekleştirmek için, oranı hesaplamak için hem Nltk cümle belirteci hem de Nltk kelime belirteci gerekmektedir. Bu tür çıktılar, cevap sayısal olacağından makine eğitimi için önemli özellik haline gelmektedir (NLTK, 2022).

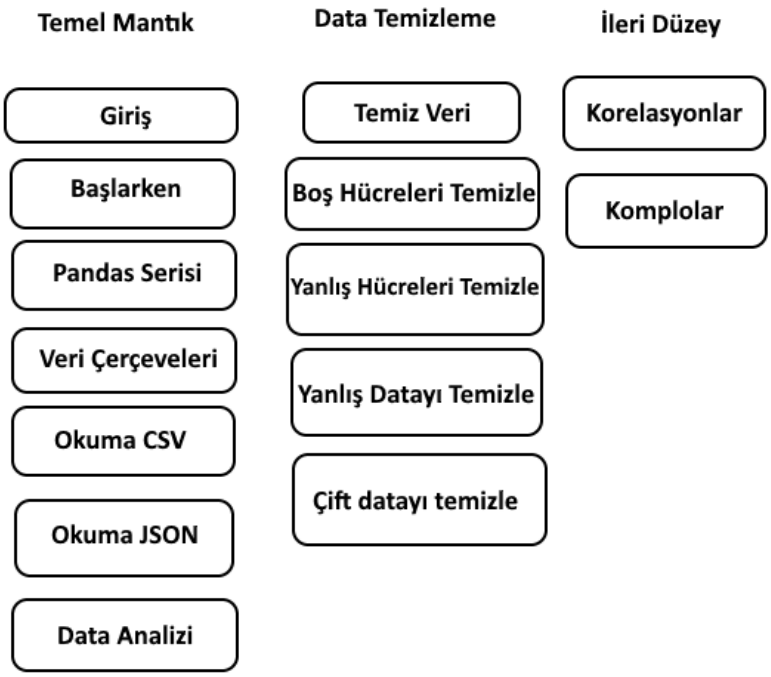
3.5. Pandas

Pandas, veri üzerinde birçok işlem yapmanızı sağlayan çok işlevselli bir kütüphanedir. Numpy üzerine inşa edilmiş çok boyutlu dizilere de destek vermektedir. Pandas birçok kütüphane ile birlikte çalışabilmekte ve işletim sisteminizle birlikte gelenlerden ActiveState'in ActivePython'u gibi ticari satıcı dağıtımlarına kadar her Python dağıtımına dahil edilebilmektedir.

Pandas, aşağıdakiler de dahil olmak üzere verilerle çalışmakla ilgili zaman alıcı, tekrarlayan görevlerin çoğunu yapmayı kolaylaştırmaktadır:

- Veri temizliği
- Veri incelemesi
- Veri doldurma
- Veri yükleme ve kaydetme
- Veri normalleştirme
- Veri görselleştirme
- İstatistiksel analiz
- Birleştirme ve düzenleme

Pandas ile yapılabilecek birçok örnek Şekil 3.4. içerisinde yer almaktadır.



Şekil 3.4. *Nlp İle Yapılabilecek Örnekler*

Pandas kullanırken bütün verileri son olarak bir yerde muhafaza etmemiz gerekebilmektedir. Bunu aslında anlık olarak veri tabanı veya ekranda çıktı olarak gösterebilmektesiniz. Ancak hız açısından tüm işlemler bitene kadar bütün işlemler sırasında işlenmiş verileri veri çerçevesinde(dataframe) tutulmaktadır. Pandas DataFrame, etiketli eksenlere (satırlar ve sütunlar) sahip iki boyutlu boyut-kesilebilmektedir. Potansiyel olarak heterojen tablolu veri yapısındadır. Veri çerçevesi iki boyutlu bir veri yapısıdır, yani veriler satır ve sütunlarda tablolu bir şekilde hizalanabilmektedir. Pandas veri kümeleri(dataframe), veriler, sütunlar ve satırlar olmak üzere üç temel bileşenden oluşmaktadır.

3.6. Numpy

NumPy, Python programlama diline ait çok boyutlu dizilerle ve matrislerle çalışmamıza yardım eden ileri düzey matematiksel işlemler yapabileceğimiz bir kütüphanedir. Veri bilimcilerinin özellikle sıklıkla kullandığı bu kütüphane Python dilinin önem arz eden bir kütüphanesi olmasını sağlamaktadır.

- Açık Kaynak koda sahip ve sayısal değerleri işleyen bir Python kütüphanesidir.
- Çok boyutlu dizin ve matris veri yapılarını içermektedir.

- Rastgele sayı üreteçlerine sahiptir.
- Dizinler de birçok sayısal işlem yapıp analiz çıktıları sağlayabilmek için kullanılmaktadır.

3.7. Mysql

Mysql, milyonlarca sistemin kullandığı, çok kullanıcılı, kullanışlı ve güvenli bir veri tabanı yönetim sistemidir. Kısacası tüm verilerin tablolar halinde bulunduğu veri tabanıdır.

3.8. Regex

Metin işleme yapılırken kullanılan en sık yöntemlerden bir tanesi de regex olarak bilinmektedir. Regex, metin içerisinde istenilen alanların kolay bir şekilde ayrıştırılmasını sağlayan ayıklayıcı olarak nitelendirilmektedir. Küçük veri içerisinde bu işlem başlangıç olarak kurtarıcı bir rol oynasa bile, veri büyüdükçe işin içinden çıkılmaz bir hale bürünebilmektedir.

Mantığı en kolay şekilde kavrayabilmek ve doğru bir kullanım için yapılması gereken öncelik ise veriyi iyice analiz edip neye ihtiyaç olduğunu belirlememiz gerekmektedir. Yazılan bir regex her mantıkta doğru bir şekilde çalışmayabilir. Gruplayarak çalışma mantığında hareket eden bu yöntem karakterlerin sıralanışı gibi ince detaylara dikkat edilip yazılması gerekmektedir. Bu projede belli adımlarda bu yöntem işe yarayabilmektedir. Medikal veri içerisinde bazı değer aralıkları bilindiği için ve bazı incelemelerin ardından birkaç data da farklı cümlelerde denk gelinen durumlarından kurtarmak adına kullanılmaktadır.

Pythex'e ait web sayfası içerisinde tek tek deneyerek aşamalı olarak regex'i de geliştirerek fonksiyonlar içerisinde konumlandırırken kullanıma uygun hale getirilebilmektedir.

3.9. Json

Json genel olarak veri depolamada düzenli ve tek tip veri ile basit işlemler gibi karışık işlemlere de düzen sağlayan bir veri yapısı olarak düşünülmektedir. Json yapısı itibariyle veri depolayan ve genellikle bir istemci ile sunucu arasında veri alışverişi için kullanılan bir format türüdür.

Xml tipinde biz dosyaya nazaran bir Json dosyası benzer fonksiyonlara sahip olduğunu varsayarak daha basit ve hafif bir alternatif olarak düşünülebilmektedir. Bu formatlar depolanan verinin eş zamansız yüklenebilmesi için uyumlu bir şekilde çalışabilmekte yani dinamik olarak işlevselliği bulunmaktadır.

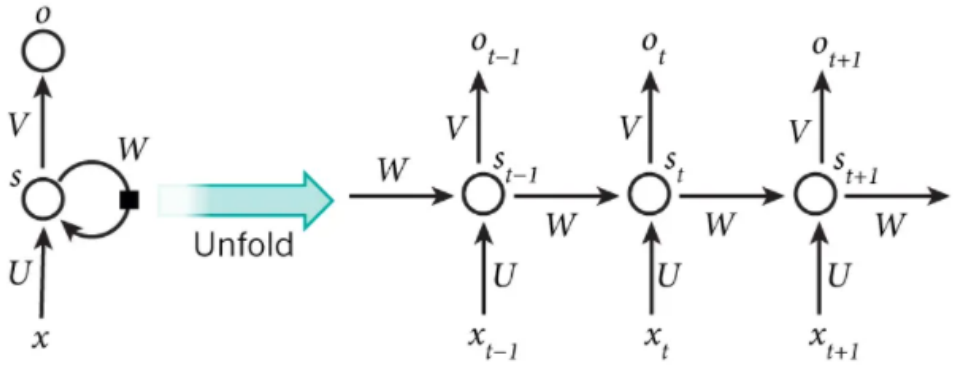
3.10. Rnn

Reccurent Neural Networks (Rnn), dilimizde ise tekrarlayan sinir ağları, düğümler arasındaki bağlantıların yönlendirilmiş bir döngü oluşturduğu sinir ağı sınıfı olarak tanımlanmaktadır. Rnn, zaman içinde dinamik bir biçimde davranmayı mümkün kılmaktadır.

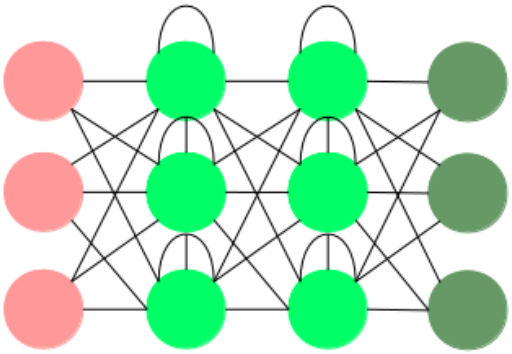
Rnn kullanım hedefi art arda olan bilgileri kullanmasıdır. Sinir ağlarının girdi ve çıktıları bağımsız olarak düşünülmektedir. Rnn'ler tekrarlayan olarak adlandırılmaktadır. Her bir dizinin, her öğesinin aynı görevi yerine getirip çıktı öncesi hesaplamalara bağlıdır. Rnn'ler, taşıdıkları bellek ile önceki hesaplamalarla bilgileri toplamaktadır. Rnn'ler teori de uzun diziler kullanabilmekte ancak pratikte sadece birkaç adım geriye dönebilmektedir ve takibini yapabilmektedir.

Bir Rnn mantığının formülsel olarak sinir ağlarına açılım mantığı Şekil 3.5. görselinde yer almaktadır. Tekrarlayan sinir ağları kendi içlerinde akışı ise Şekil 3.6. içerisinde görüntülenmektedir.

Tipik bir Rnn yapısı:

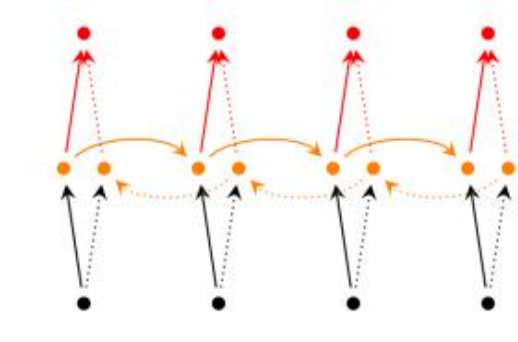


Şekil 3.5. *Rnn'nin Bir Ağa Açılımı*



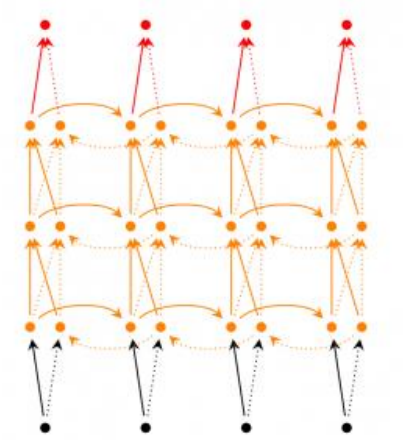
Şekil 3.6. *Tekrarlayan Sinir Ağı (Rnn)*

İki yönlü Rnn'ler (Bidirectional Rnns): t zamanındaki çıktıların sıralardaki önceki öğelere değil gelecekteki öğelere de bağlı olabileceği fikrine dayanmaktadır. Örneğin, bir dizideki eksik bir kelimeyi tahmin etmek için hem sol hem de sağ içeriğe bakmak istiyorsunuz. Çift yönlü Rnn'ler oldukça kolay olabilmektedir. Şekil 3.7. de gözlemlenebileceği gibi kolayca üst üste eklenmiş iki ayrı Rnn olduğunu düşünebilirsiniz. Çıktı verileri sonrasında iki ayrı Rnn'nin gizli halini temek alarak hesaplamaktadırlar.



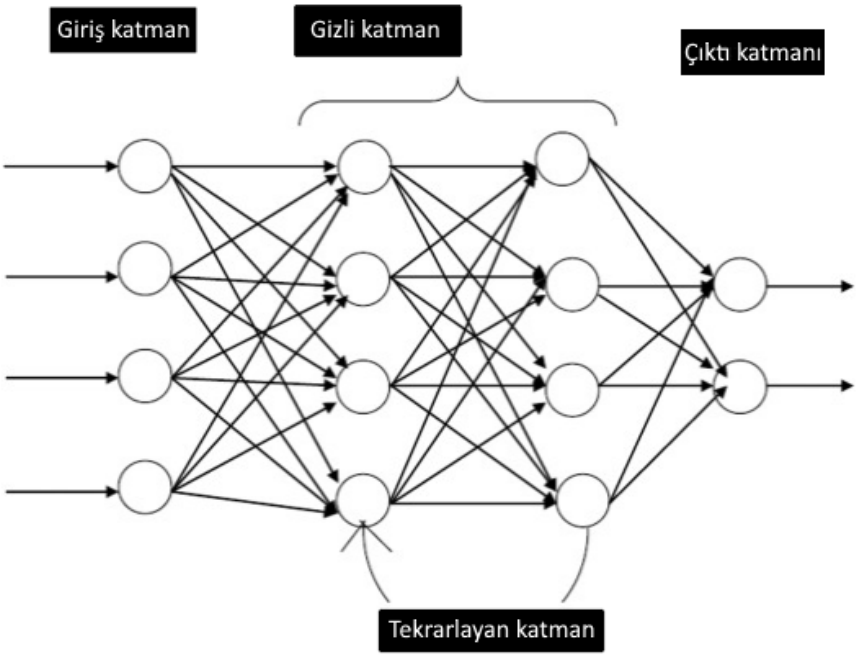
Şekil 3.7. *Çift Yönlü Rnn*

Derin (İki Yönlü) Rnn'ler (Deep Rnns): Zaman kısımlarında birden fazla alandan meydana gelmektedir. Şekil 3.8 de görüldüğü gibi iki yönlü Rnn'lere benzemektedir. Bu da yüksek öğrenme kapasitesi sağlamaktadır. Bunun başarılı olabilmesi için de daha çok veriye ihtiyaç duymaktadır.



Şekil 3.8. *Derin Çift Yönlü Rnn*

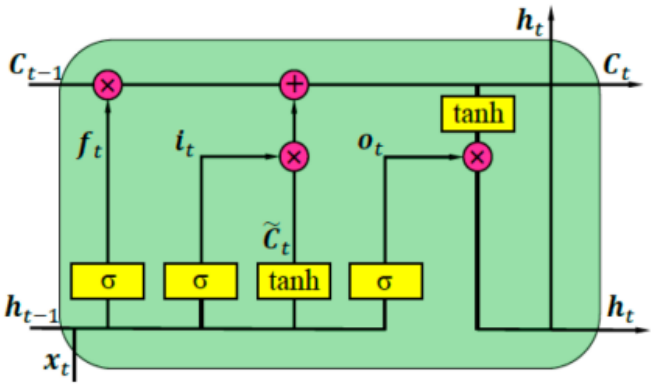
Şekil 3.9.'da gösterilen örnek aslında özet niteliği taşımaktadır. Girdilerin ardından kendi içerisinde bulunan tek yönlü tekrarlayan bir derin öğrenme göstermektedir.



Şekil 3.9. *Tekrarlayan Sinir Ağına Örnek*

3.11. Lstm

1990 yıllarının sonralarına doğru sıralı verilerin modellenmesi ihtiyacıyla Rnn mantığının özel bir türü olarak geliştirilmiştir. Rnn'nin mantığı her girdi bilgisinde bir önceki çıktı değerini de dikkate almakta ve incelemektedir. Bu mimaride ise önceki zaman aralıklarındaki bilgiler dikkate alan bir öğrenme olduğu belirtilmesine karşın gradyan patlaması/kaybolması problemi yüzünden bunun olasılık dışı olduğu belirtilmiştir. Bu sorunun ortadan kalkması ve uzun süreli bilgileri hatırlaması için Lstm mimarı yapısı geliştirilmiştir. Lstm mimarisi, Şekil 3.10.'da göründüğü gibi birbirini tekrar eden sıralı sistemden meydana gelmektedir.



Şekil 3.10. *Lstm Mimari Yapısı*

Lstm yapısı unut, girdi ve çıktı olmak üzere 3 farklı katmandan oluşmaktadır. Lstm mimarisinde öncelikle girdi olarak Xt ve $ht-1$ bilgileri kullanılarak hangi bilgilerin silinmesi gerektiğine karar verilmektedir. Bu işlemler unut katmanında (ft) Eşitlik (3.1) kullanılarak yapılmakta ve aktivasyon fonksiyonu olarak sigmoid kullanmaktadır.

$$ft = \sigma(Wf,x * Xt + Wf,h * ht-1 + bf) \quad (3.1)$$

İkinci adımda yeni bilgilerin belirleneceği girdi katmanı devreye girmekte ve öncelikle (it) Eşitlik (3.2) kullanılarak sigmoid fonksiyonu ile bilgiler yeniden güncellenmektedir. Ardından Eşitlik (3.3) ile yeni bilgiyi oluşturacak aday bilgiler tanh fonksiyonu tarafından belirlenmektedir.

$$it = \sigma(Wi,x * Xt + Wi,h * ht-1 + bi) \quad (3.2)$$

$$\tilde{C}t = tanh(Wc,x * Xt + Wc,h * ht-1 + bc) \quad (3.3)$$

Eşitlik (4) tarafından yeni bilgiler oluşturulmaktadır.

$$Ct = Ct-1 * ft + it * \tilde{C}t \quad (3.4)$$

Son kısımda ise çıktı katmanında Eşitlik (3.5) ve (3.6) kullanılarak çıktı verileri elde edilmektedir.

$$ot = \sigma(Wo,x * Xt + Wo,h * ht-1 + bo) \quad (3.5)$$

$$ht = ot * tanh(Ct) \quad (3.6)$$

Yukarıda belirtilen süreç sürekli olarak tekrarlanarak devam etmektedir. Ağırlık parametreleri (W) ve bias parametreleri (b) gerçek eğitim değerleri ile Lstm çıktı değerleri arasındaki farkı minimize edecek şekilde model tarafından öğrenilmektedir.

3.12. Ner (Name Entity Recognition)

Ner (Adlandırılmış varlık tanıma) metinler içerisindeki önemli bilgileri(varlıkları) belirleyip kategori olarak ayırma göreviyle çalışmaktadır. Daha açık olmak gerekirse cümle içerisinde bulunan kelimelerin ne olduğunu belirterek onları kategorize etmektedir. Daha önceden belirlenmiş kategoriler için verilen cümleler içinde o gruplar için araştırma yapar ve benzettiği mantıkta bir şeye rast gelir ise bunu belirtmektedir.

Örneğin önceden kategorilenmiş bir meyve adına cümle içerisinde denk geldiğinde bunu meyve olarak tahmin edebilmektedir. Bu başarı oranı da sizin datanızın büyüklüğüyle başarı oranı artmaktadır. Her metin yapısında öğrenim sağlanırsa bu yapının çalışması o kadar olumlu sonuçlar doğurmaktadır.

Herhangi bir Ner modelinin ana hattında iki adımlı bir süreç vardır:

- Adlandırılmış bir varlığı algılama
- Varlığı kategorilere ayırma

3.13. Bert

Bert (Transformers'dan Çift Yönlü Kodlayıcı Temsilleri) Google tarafından geliştirilmiş, doğal dil işleme (Nlp) ön eğitimi için transformatör tabanlı bir makine öğrenimi yöntemidir. Özünde değişken sayıda kodlayıcı katmanı ve kendi kendine dikkat eden kafaları olan bir dönüştürücü dil modelidir.

Bert iki görev üzerinde önceden eğitildi: dil modelleme (belirteçlerin %15'i maskelendi ve Bert bunları bağlamdan tahmin etmek için eğitildi) ve sonraki cümle tahmini (Bert, seçilen bir sonraki cümlenin olası olup olmadığını veya ilk cümlenin verilip verilmediğini tahmin etmek için eğitildi). Eğitim programının bir sonucu olarak, Bert kelimeler için bağlamsal yerleştirmeleri öğrenmektedir. Hesaplama açısından pahalı olan ön eğitimden sonra, Bert, belirli görevlerdeki performansını optimize etmek için daha küçük veri kümelerinde daha az kaynakla ince ayar yapılabilmektedir.

Bert ile çalışmak için yerel veya cloud alanda gpu kullanımı gerekmektedir. Bert diğer modellere nazaran cümleyi her iki yönden de değerlendirmektedir. Bu sayede anlamı ve kelimelerin birbiriyle olan ilişkileri daha iyi çıkarmayı planlayabilmektedir.

3.14. Flair

Güçlü bir Nlp kütüphanesi. Flair, adlandırılmış varlık tanıma (Ner), konuşma parçası etiketleme (Pos), biyomedikal veriler için özel destek, anlam gibi son teknoloji doğal dil işleme (Nlp) modellerimizi metninize uygulamanıza olanak tanımaktadır. Hızla artan sayıda dil desteği ile anlam ayrımı ve sınıflandırma.

Flair, önerilen Flair yerleştirmelerimiz, Bert yerleştirmelerimiz ve Elmo yerleştirmelerimiz dahil olmak üzere farklı kelime ve belge yerleştirmelerini kullanmanıza ve birleştirmenize izin veren basit arayüzlere sahiptir.

3.15. FastText

FastText, kelime temsillerinin ve cümle sınıflandırmasının verimli bir şekilde öğrenilmesi için bir kütüphanedir. Bu kitaplık, duygu analizi için denetimli metin sınıflandırıcıları eğitmek için de kullanılabilinmektedir.

3.16. Gerçekleştirilen Uygulama

Doğru sonuçlar elde edilene kadar birçok yöntem denenmiştir. İlk olarak her şeyden önce problem belirlenmelidir. Sonrasında da ihtiyaç listesi belirlenmektedir. Hedeflediğimiz sonuç ise laboratuvar değerlerini yakalayıp birbiri ile anlamlı bir sonuç elde etmektir. Öncelikli ilk olarak detaylı bir şekilde veriyi inceleyerek mantık oluşturmak gerekmektedir. Ancak hasta verileri gizli bilgi olduğu için örnek cümleleri oluşturulmalı veya türetilmelidir.

Örnek olabilecek cümleler:

- Laboratory work-up on November 16, 2017 was unremarkable aside from a CEA of 12.4, platelet count of 671 and a hemoglobin of 10.5.
- The patient, hemoglobin was 10.9.
- Laboratory work-up on January 11, 2018 revealed a WBC of 13.89, Hgb 9.4, Plt 415, ANC 8640 and CA125 388.1.
- Her ECOG is 2, and KPS 70%.

- Her laboratory workup showed Hb 9, WBC 8, platelets 132000, ANC 4000, Neutrophils 50%, AST 35, ALT 60, CA 15-3 10, bilirubin 1.8, PT and PTT normal, CEA 30, creatinine 1.3. TSH 2.7.
- Laboratory Results CBC : WBC = 4.3 , HGB 10.3 , HOT = 30.4 , Plat = 122 (6 / 22 / 20) Other studies : CEA 1.9 and tumor marker CA 199 is 19 Guardian 360 is unremarkable for any EGFR or any mutation of ALK .
- Performance Status : 1 - No physically strenuous activity , but ambulatory and able to carry out light or sedentary work (e.g. office work , light house work) .
- Laboratory work-up on September 22, 2017 showed a slightly elevated AST (42), but was otherwise stable.

Sonrasında yapay zeka gibi birçok alanda kullanılan Python yazılım dilini kullanmak hem yazımı kolay hem de kütüphane içeriği olarak zengin olduğundan kaynaklı tercih sebebi olmaktadır.

Var olan verileri bir alanda bulundurmak gerekmektedir. Bunun için Mysql kullanarak verileri düzenli bir halde bir arada toplanması sağlanmaktadır. Bir veri tabanı oluşturarak verilerimi düzenli bir yapıda bulunması için ve temiz kod yapısını korumak için her şeyi tablolar halinde tutmak tercih edilmektedir. Sonrasında Python kodu içerisinde bağlantı ayarlarını yapılarak veri üzerinde işlem yapılmaktadır. İstenilen tabloya istediğim verileri silinebilir, eklenebilir, sorgulanabilir ve güncellenebilmektedir.

Kod yapısını tasarlarken ihtiyaç olan kütüphaneleri indirerek ikinci aşamaya geçinilmelidir. Bu aşamada birkaç dikkat edinilmesi gereken konu bulunmaktadır. Öncelikle indireceğiniz kütüphaneler versiyondan kaynaklı çalışmamaktadır. Bu duruma özellikle Flair modeli kullanırken dikkat edilmelidir ve indirilen internet sitesindeki alanlarda çıkan hatalara dikkat edinilmelidir. Bütün kütüphanelerin tamamlanmasının ardından basit kod yazımı ile hepsinin doğru bir şekilde çalıştığından emin olmak gerekmektedir.

Sonraki aşama ise kullanacağımız kütüphaneleri dosyalara kodun içerisine ekleyerek deneme yanılma yoluyla bazı yöntemler denemeye başlanmalıdır. Bir yapı tümdengelim yöntemi değil de tümevarım şeklinde gidilirse daha sağlıklı bir yol izlemiş olabilirsiniz. Bu projedeki amaç manuel bir işlemler dizisi yerine makine öğrenimi

kullanılarak otomatik olarak sonuç elde etmektir. Her cümle için birden fazla regex yazılarak da bu işlem çözülebilmektedir. Ancak her yeni çeşit cümle de tekrar yenisini yazmak durumunda kalınabilmektedir. Bu projede bunu amaçlarımız arasına eklemeyerek model yardımıyla halletmeye çalışmaktayız.

Veri tabanı ile oluşturduğumuz bağlantıların ardından kütüphaneleri de kodumuzun içine eklemiş bulunmaktayız. Birkaç denemenin ardından türettiğimiz medikal cümleleri tabloya ekleyerek devam edebilmekteyiz. Toplam da 100'den fazla cümle oluşturmuş veya türetmiş olarak hazır bulundurmaktayız. Bunların içerisinde üzerinde çalışabileceğimiz 10 adet laboratuvar değeri mevcuttur.

- Ecog
- Neutrophil Count
- Hemoglobin
- Lymphocyte Count
- Platelet Count
- Ast
- Alt
- Alp
- Total Bilirubin
- Albumin

Sonrasında bu laboratuvar değerlerini ayrı bir tabloya ekleyerek düzenli bir yapı oluşturmaktayız. Medikal olarak bu parametreleri arama motorlarında araştırdığınız zaman birçok eş anlamlılarına rast gelinebilmektedir. Cümleleri oluştururken de bunlara dikkat ederek metin yapısında eş anlamlılarının olduğu ve array yapısında bir sütun eklenmesi faydalı olacaktır. Yani tablo olarak birinci sütun da her tabloda olduğu gibi id, ikinci sütun da parametrenin adı, üçüncü sütunda da eş anlamlılarının bulunduğu bir tablolama oluşturmaktayız. Tablo 3.1. de bulunan yapıya birebir benzemektedir. Bu düzendeki gibi oluşturulmaktadır.

Tablo 3.1.: *Laboratuvar Parametreleri Ve Eş Anlamları*

Laboratuvar Parametreleri	Eş Anlamları
ECOG	["Performance Status","Eastern cooperative oncology group performance status","ps","ecog","eastern cooperative oncology group","ecog ps","performance status score","performance status (ecog ps)","ecog level","ecog score","ecog performance status","Zubrod Performance Status","zubrod ps","Zubrod performance score","ECOG Status","ECOG Performance Scale"]
Neutrophil Count	["ANC","neutrophil count","neutrophil count anc","Absolute neutrophil count (ANC)","Absolute neutrophil count","neutno","Neut","Neutrophil Absolute","neutrophil abs"]
Hemoglobin	["hgb","hbg","hemoglobin","hemoglobin count","haemoglobin count","hb","hb level","hgb level","hemoglobin level","haemoglobin level","hemoglobin value"]
Lymphocyte Count	["lymphocytes","lymphocyte count","lymphocyte","absolute lymphocyte count ALC","absolute lymphocyte count","ALC","absolute lymphocyte count (ALC)","LYMPH","lymph","Lymphocytes Absolute","LY"]
Platelet Count	["plt","platelet","platelet count","platelet level","platelets","platelet counts","plat","plts"]

AST	["ast","aspartate aminotransferase","serum aspartate aminotransferase ast","serum ast","ast level","aspartate aminotransferase (AST) level","aspartate aminotransferase (AST)","AST(SGOT)","GOT"]
ALT	["alt","alt levels","alt sgpt","alt gpt","aspartate transaminase alt","ALT(SGPT)","alanine aminotransferase (ALT)","alanine aminotransferase (ALT) level","sgpt alt","ALT(SGPT)","GPT"]
ALP	["alp","ALP limit","alkaline phosphatase","alkaline phosphatase level (alp)","ALP Alkaline phosphatase","alkphos","alk.phos","alk. phos","alkp","Alk Phos"]
Total Bilirubin	["bilirubin","bilitotal","bilirubin, total","bilirubin,total","Bilirubin, total","tbili","TBILI","bilirubin total","total bilirubin","total serum bilirubin levels","total bilirubin levels","total bilirubin count","serum total bilirubin","total bilirubin tbl","Total serum bilirubin","TBIL","total bilirubin TBILI","total bilirubin level","Serum bilirubin","Bilirubin","total bili","tbilirubin","bilitotal","T Bili","bilirubin, total","bilirubin,total","Bilirubin, total","bili t","TBILIRUBIN","bilitotal","T bili","bilitot"]
Albumin	["albumin level","albumin levels","albumin","Alb"]

Elimizde olan türetilmiş cümlelerin içerisinde istenmeyen karakterlerden kaynaklı veya aslında olması gerekenin dışında kısaltmalar gibi durumlar içinde bir değiştirme tablosu oluşturmak kullanışlı olacaktır. Bunun yapılmasındaki amaç birkaç unsur için yapıyı değiştirmek yerine akılcı ve hızlı çözüme kavuşturmaktır. Bu tablo da id, değiştirmek istediğimiz karakter, kelime veya sayı değeri, son sütunda da olmasını gereken yapıdaki numerik veya metin şeklindeki durum olarak planlamaktadır. Bunun yapısını da girilen cümleyi öncelikle bir veri çerçevesi (dataframe) içerisinde saklı tutmak basite indirgeyecektir. Sonrasında da oluşturulan tabloyu çağırarak 2. kolondaki var olan değerin aratılması gerekmektedir. Eğer var ise 3. kolondaki ile yer değiştirilir. Bu şekilde bir nevi temizlik de yapmış olunmaktadır.

Temizlenmiş datalara örnek cümleler:

1- PET/CT in December 2018 demonstrated resolution of FDG avid of RP nodes, which decreased in size, and slight interval decrease in size of two persistently FDG avid splenic lesions. A autologous stem cell transplant was planned for early January 2019 after partial response from 2 cycles of RICE, but had to be delayed due to upper respiratory infection with positive RVP for rhinovirus. Neupogen was initiated on January 27, 2019, but stem cell collection was unsuccessful. Laboratory work-up on February 7, 2019 was unremarkable: Hgb 11.2, ANC 3.69, Total bilirubin 0.4, ALP 98, ALT 16, AST 21, Albumin 4.1. PET/CT on February 7, 2019 demonstrated multiple foci of intense FDG activity including within the spleen, soft tissue mass within the splenic hilum and left internal mammary lymph node in keeping with active diffuse large B-cell lymphoma.

2- CT soft tissue neck without contrast showed scattered lymph nodes throughout the neck, borderline right level 3 lymph node. Laboratory work-up on January 25, 2019 was significant for Hgb 7.5, Total bilirubin 0.5, ALP 90, AST 292, ALT 45, Albumin 2.4 VAS Duplex lower extremity veins with compression on January 26, 2019 showed no evidence of lower extremity deep venous thrombosis bilaterally.

Kullanacağınız verileri iyice incelemeniz gerekmektedir. Her şey veriye dayandığı için yazılan kod ile karşınıza çıkan cümleleri iyi bir şekilde sınıflandırmanız gerekmektedir. Kullanmış olduğumuz Flair Ner modeli de bu görevi üstlenmektedir. Cümleler içerisindeki kelimeleri sınıflayarak etiket mantığıyla hareket etmektedir. Medikal alanda iyi sonuçlar çıkardığını gözlemlediğim için özellikle bu modelleri tercih etmekteyiz.

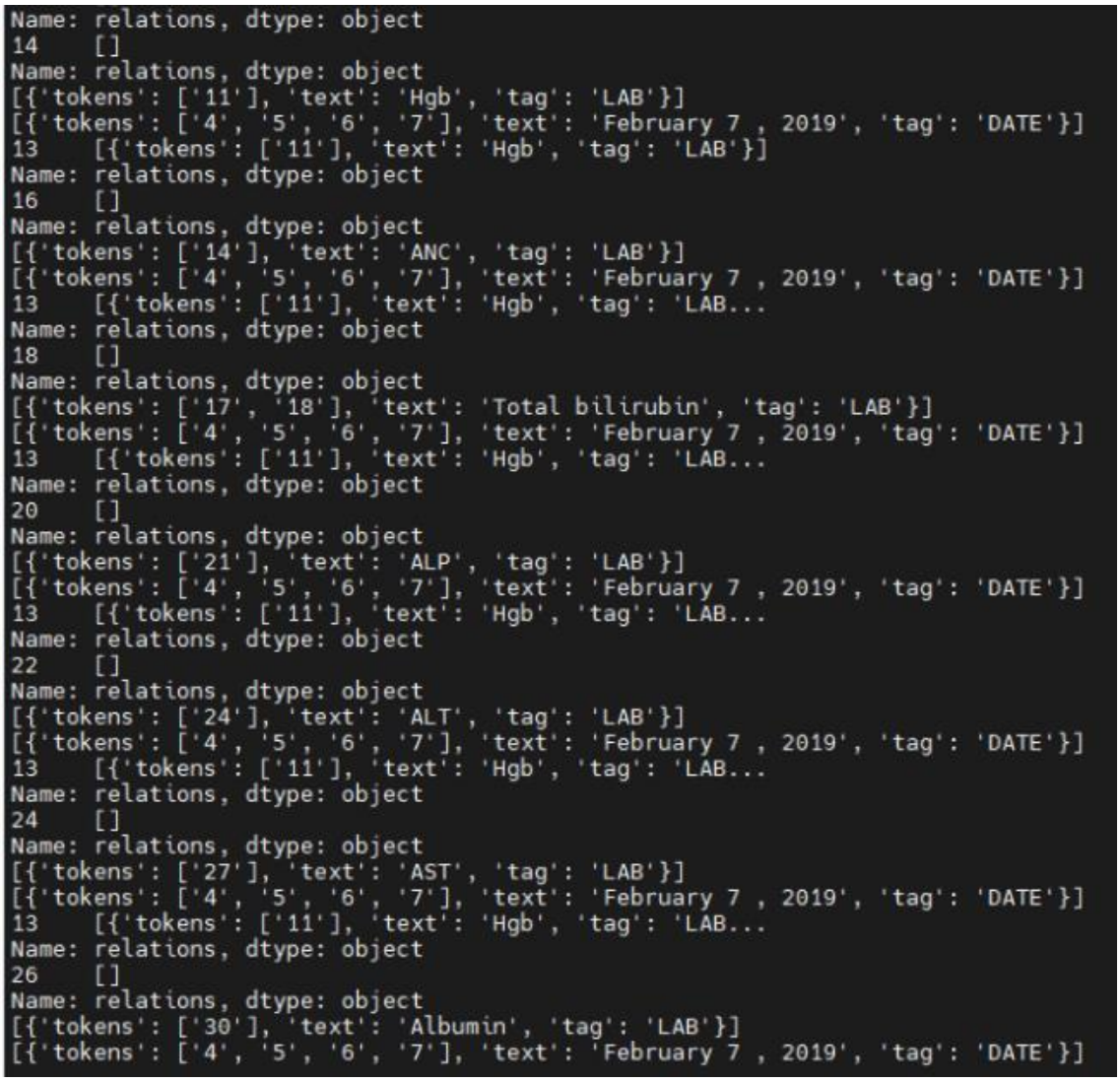
Temizlenmiş olan cümleleri Ner modelinden geçirerek vermiş olduğu çıktıyı analiz edip modelin analizi yapılması faydalı olabilmektedir. Şekil 3.11 de manuel veri girişi yapıktan sonra ilk Ner modelden geçmektedir.

	sentence	entity	entity_tag
0	PET/CT in December 2018 demonstrated resolutio...	PET/CT	PETSCAN
1	PET/CT in December 2018 demonstrated resolutio...	December 2018	DATE
2	PET/CT in December 2018 demonstrated resolutio...	nodes	MEASURABLEDISEASEPERRECISTv11
3	PET/CT in December 2018 demonstrated resolutio...	2	VALUE
4	PET/CT in December 2018 demonstrated resolutio...	lesions	MEASURABLEDISEASEPERRECISTv11
5	A autologous stem cell transplant was planned ...	autologous stem cell transplant	TRANSPLANT
6	A autologous stem cell transplant was planned ...	January 2019	DATE
7	A autologous stem cell transplant was planned ...	2 cycles	LINEOFTHERAPY
8	A autologous stem cell transplant was planned ...	RICE	DRUGMEDICATION
9	A autologous stem cell transplant was planned ...	upper respiratory infection	DISEASE
10	A autologous stem cell transplant was planned ...	positive	STATUS
11	Neupogen was initiated on January 27, 2019, bu...	Neupogen	DRUGMEDICATION
12	Neupogen was initiated on January 27, 2019, bu...	January 27 , 2019	DATE
13	Laboratory work-up on February 7, 2019 was unr...	February 7 , 2019	DATE
14	Laboratory work-up on February 7, 2019 was unr...	Hgb	LAB
15	Laboratory work-up on February 7, 2019 was unr...	11.2	VALUE
16	Laboratory work-up on February 7, 2019 was unr...	ANC	LAB
17	Laboratory work-up on February 7, 2019 was unr...	3.69	VALUE
18	Laboratory work-up on February 7, 2019 was unr...	Total bilirubin	LAB
19	Laboratory work-up on February 7, 2019 was unr...	0.4	VALUE
20	Laboratory work-up on February 7, 2019 was unr...	ALP	LAB
21	Laboratory work-up on February 7, 2019 was unr...	98	VALUE
22	Laboratory work-up on February 7, 2019 was unr...	ALT	LAB
23	Laboratory work-up on February 7, 2019 was unr...	16	VALUE
24	Laboratory work-up on February 7, 2019 was unr...	AST	LAB
25	Laboratory work-up on February 7, 2019 was unr...	21	VALUE
26	Laboratory work-up on February 7, 2019 was unr...	Albumin	LAB
27	Laboratory work-up on February 7, 2019 was unr...	4.1	VALUE
28	PET/CT on February 7, 2019 demonstrated multip...	PET/CT	PETSCAN
29	PET/CT on February 7, 2019 demonstrated multip...	February 7 , 2019	DATE
30	PET/CT on February 7, 2019 demonstrated multip...	mass	MEASURABLEDISEASEPERRECISTv11
31	PET/CT on February 7, 2019 demonstrated multip...	lymph node	MEASURABLEDISEASEPERRECISTv11
32	PET/CT on February 7, 2019 demonstrated multip...	diffuse large B-cell	HISTOLOGY
33	PET/CT on February 7, 2019 demonstrated multip...	lymphoma	CANCERTYPE

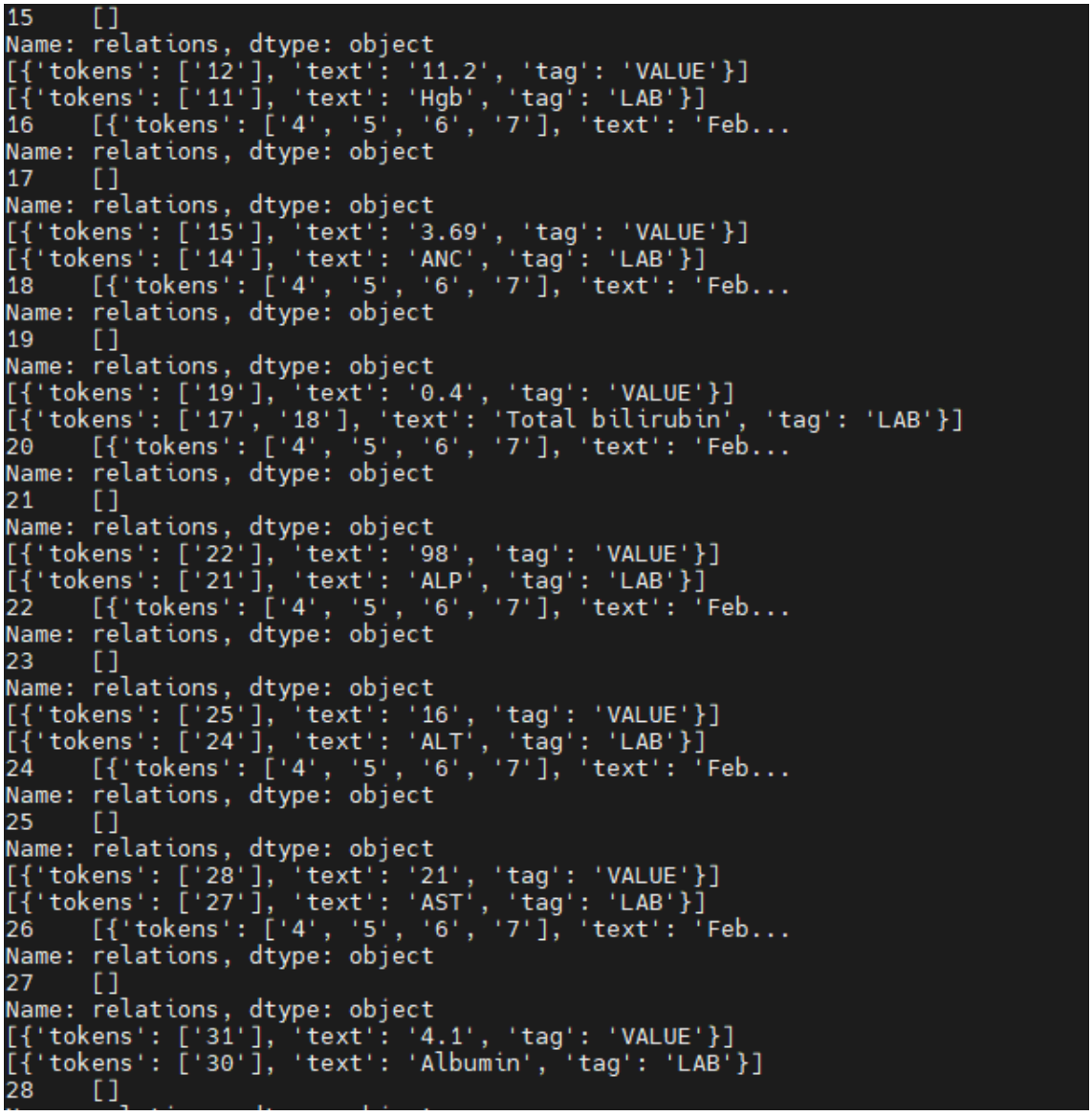
Şekil 3.11. *Ner Modelinin Çıktısı*

Deneme yanılma metodu ile birçok hatadan süzerek çıktı olarak elde edilen sonuçlar ile karşılaştırılmalıdır. Bu alanda 3 sütun halinde çıktıyı görüntülemekteyiz. İlk satırda Ner'den geçirdiğimiz cümleler yer almaktadır. İkinci alanda Flair modelimizin etiketlediği kelimeler yer almaktadır. 3. sütun da ise bunların model içerisinde karşılık olarak ne bulduğunu tahmin edip belirtmektedir. İstenilen kelimeleri doğru etiketlenip etiketlenmediğini sonuçları analiz ederek bir başarı oranı elde edinilmektedir. Bunu yaklaşık olarak 100 cümle üzerinden değerlendirmesini yapmaktayız. Bu şekilde kelime etiketleme işlemi istenilen doğruluk oranına ulaştığı taktirde ikinci aşama olan ilişki modelini kurgulamak gerekmektedir. Bu kısımda yapılması gerekli en doğru yöntemlerden biri de kullanılmış olan Flair modelinin testlerini doğru bir şekilde yapılmasıdır. Kısa cümleler ile sadece laboratuvar adı ile değeri arasındaki bağlantı da sonuçların gelip gelmediğini kontrol edilmesi gerekmektedir. Eğer belli bir oranda doğru sonuçlar elde edilebiliyorsa doğru bir yöntem izlenilmiş olunur. Aksi halde 100 adet cümle sayısı yetersiz geldiği için istenilen şekilde cümle eğitilmelidir.

Sonrasında değerler arasındaki bağlantı için de Flair'e ait ilişki modeli kullanılmaktadır. Bu şekilde benzer yapıda oluşturulmuş modeller bütünü sağlanmış olacaktır. Bu modelin başlangıç olarak tahmin oranı düşük olsa bile eğitilen cümle sayısı ne kadar artarsa bu oran da artacaktır. Ner modelinin bulduğu sonuçlardan sonra 3.12. de ve 3.13. görüldüğü gibi ilişki modeli devam etmektedir.



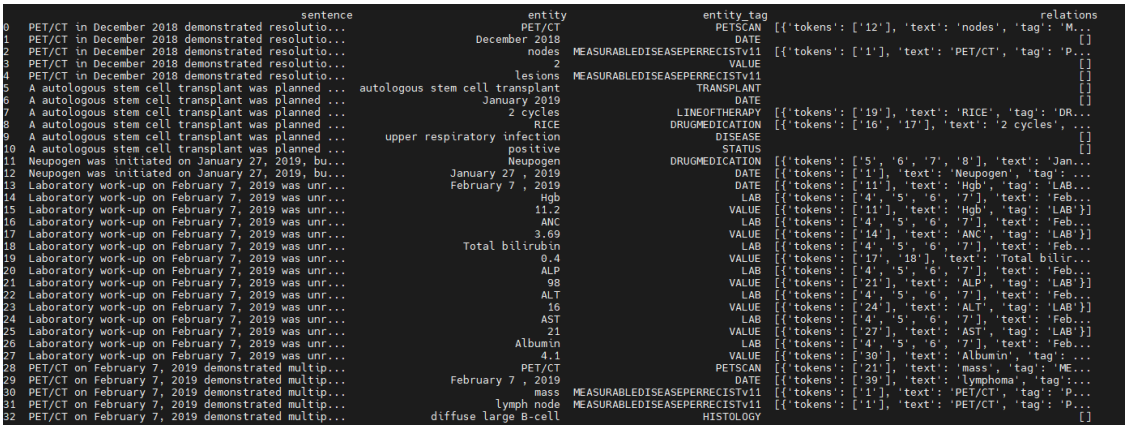
Şekil 3.12. *İlişki Modeli Çıktısı - 1*



Şekil 3.13. *İlişki Modeli Çıktısı - 2*

Model içeriğinde medikal olarak birden fazla içerik bulunabilmektedir. Bizim istediğimiz parametreler ise 10 adet ile sınırlandırılmış laboratuvar parametreleridir. Burada da ilişkisi bulunan laboratuvar değerlerini ve hangi tag ile etiketlendikleri görülmektedir. Bizim belirlemiş olduğumuz “LAB” ve “VALUE” etiketi ile önce Ner modelinden geçerek bu kelimeler etiketlenmektedir. Sonrasında ise ilişki modeliyle birlikte kelimeler arasındaki bağlantı ile veri kümesi içerisinde ilişkiler görünmektedir.

Modellerden geçen cümlelerinden ardından Ner etiketi verilen cümleler ilişki kurularak çıktıyı yazdırdıktan sonra Şekil 3.14. de göründüğü gibi ekranda göstermektedir. Bu şekilde oluşturulmuş veri tablosuna yazdırarak birden fazla cümleyi analiz edip doğruluk oranına göre cümle eğitimiyle bu oran arttırılabilir.



	sentence	entity	entity_tag	relations
0	PET/CT in December 2018 demonstrated resolutio...	PET/CT	PETSCAN	[{'tokens': ['12'], 'text': 'nodes', 'tag': 'M...
1	PET/CT in December 2018 demonstrated resolutio...	December 2018	DATE	[]
2	PET/CT in December 2018 demonstrated resolutio...	nodes	MEASURABLEDISEASEPERRECISTV11	[{'tokens': ['1'], 'text': 'PET/CT', 'tag': 'P...
3	PET/CT in December 2018 demonstrated resolutio...	2	VALUE	[]
4	PET/CT in December 2018 demonstrated resolutio...	lesions	MEASURABLEDISEASEPERRECISTV11	[]
5	A autologous stem cell transplant was planned ...	autologous stem cell transplant	TRANSPLANT	[]
6	A autologous stem cell transplant was planned ...	January 2019	DATE	[]
7	A autologous stem cell transplant was planned ...	2 cycles	LINEOFTHERAPY	[{'tokens': ['19'], 'text': 'RICE', 'tag': 'DR...
8	A autologous stem cell transplant was planned ...	RICE	DRUGMEDICATION	[{'tokens': ['16', '17'], 'text': '2 cycles', ...
9	A autologous stem cell transplant was planned ...	upper respiratory infection	DISEASE	[]
10	A autologous stem cell transplant was planned ...	positive	STATUS	[]
11	Neupogen was initiated on January 27, 2019, bu...	Neupogen	DRUGMEDICATION	[{'tokens': ['5', '6', '7', '8'], 'text': 'Jan...
12	Neupogen was initiated on January 27, 2019, bu...	January 27 , 2019	DATE	[{'tokens': ['1'], 'text': 'Neupogen', 'tag': ...
13	Laboratory work-up on February 7, 2019 was unr...	February 7 , 2019	DATE	[{'tokens': ['11'], 'text': 'Hgb', 'tag': 'LAB...
14	Laboratory work-up on February 7, 2019 was unr...	Hgb	LAB	[{'tokens': ['4', '5', '6', '7'], 'text': 'Feb...
15	Laboratory work-up on February 7, 2019 was unr...	11.2	VALUE	[{'tokens': ['11'], 'text': 'Hgb', 'tag': 'LAB'}]
16	Laboratory work-up on February 7, 2019 was unr...	ANC	LAB	[{'tokens': ['4', '5', '6', '7'], 'text': 'Feb...
17	Laboratory work-up on February 7, 2019 was unr...	3.69	VALUE	[{'tokens': ['14'], 'text': 'ANC', 'tag': 'LAB'}]
18	Laboratory work-up on February 7, 2019 was unr...	Total bilirubin	LAB	[{'tokens': ['4', '5', '6', '7'], 'text': 'Feb...
19	Laboratory work-up on February 7, 2019 was unr...	0.4	VALUE	[{'tokens': ['17', '18'], 'text': 'Total bilir...
20	Laboratory work-up on February 7, 2019 was unr...	ALP	LAB	[{'tokens': ['4', '5', '6', '7'], 'text': 'Feb...
21	Laboratory work-up on February 7, 2019 was unr...	98	VALUE	[{'tokens': ['21'], 'text': 'ALP', 'tag': 'LAB'}]
22	Laboratory work-up on February 7, 2019 was unr...	ALT	LAB	[{'tokens': ['4', '5', '6', '7'], 'text': 'Feb...
23	Laboratory work-up on February 7, 2019 was unr...	16	VALUE	[{'tokens': ['24'], 'text': 'ALT', 'tag': 'LAB'}]
24	Laboratory work-up on February 7, 2019 was unr...	AST	LAB	[{'tokens': ['4', '5', '6', '7'], 'text': 'Feb...
25	Laboratory work-up on February 7, 2019 was unr...	21	VALUE	[{'tokens': ['27'], 'text': 'AST', 'tag': 'LAB'}]
26	Laboratory work-up on February 7, 2019 was unr...	Albumin	LAB	[{'tokens': ['4', '5', '6', '7'], 'text': 'Feb...
27	Laboratory work-up on February 7, 2019 was unr...	4.1	VALUE	[{'tokens': ['30'], 'text': 'Albumin', 'tag': ...
28	PET/CT on February 7, 2019 demonstrated multip...	PET/CT	PETSCAN	[{'tokens': ['21'], 'text': 'mass', 'tag': 'ME...
29	PET/CT on February 7, 2019 demonstrated multip...	February 7 , 2019	DATE	[{'tokens': ['39'], 'text': 'lymphoma', 'tag':...
30	PET/CT on February 7, 2019 demonstrated multip...	mass	MEASURABLEDISEASEPERRECISTV11	[{'tokens': ['1'], 'text': 'PET/CT', 'tag': 'P...
31	PET/CT on February 7, 2019 demonstrated multip...	lymph node	MEASURABLEDISEASEPERRECISTV11	[{'tokens': ['1'], 'text': 'PET/CT', 'tag': 'P...
32	PET/CT on February 7, 2019 demonstrated multip...	diffuse large B-cell	HISTOLOGY	[]

Şekil 3.14. *İlişki Modeli Çıktısı - 3*

3.17. Elde Edilen Sonuçlar

Hazır olarak eğitilmiş bir modelden alınacak sonuçları analiz ederken dikkat edilmelidir. Model ne kadar büyük bir veri ile eğitilmiş olur ise sonuçları o oranda başarılı olacaktır. Seçilen Flair modeli medikal veri olarak başarı oranı yüksek bir modeldir. Oluşturulan Şekil 3.15. de örneklerde görebildiğimiz 100 cümleyi teker teker denendiğinde %80 oranında tahmin skoru elde edilmektedir.

sentence
The patient's total bilirubin was also 1.81 at this time.
On May 11, 2016, the patient's liver fuction testing demonstrated signficant elevation of AST (232), ALT (413), and Total Bilirubin (2.36).
She began to receive Neulasta injections after the second cycle due to neutropenia (WBC 1.53, ANC 310).
Laboratory work-up on January 11, 2018 revealed a WBC of 13.89, Hgb 9.4, Plt 415, ANC 8640 and CA125 388.1.
Upon admission to the hospital, the patient had a hemoglobin of 6.7.
This is the case of a 57-year-old male who presented in early April 2018 with an elevated PSA of 4.873.
Laboratory work-up on November 16, 2017 was unremarkable aside from a CEA of 12.4, platelet count of 671 and a hemoglobin of 10.5.
The patient, hemoglobin was 10.9.
Laboratory work-up on November 21, 2018 was significant for a WBC of 3.3, Hbg 10.1, ANC of 1.7 and Potassium of 3.1.
On November 29, 2018, laboratory work-up was significant for a Hgb of 8.8, Platelet 76, Total Protein 5.8 and Potassium 2.7.
Most recently, laboratory work-up on December 6, 2018 was significant for WBC of 2.1, Hgb of 9.4, ANC of 1.22 and Potassium was stable at 3.4.
Laboratory work-up on January 25, 2019 was significant for WBC 17.91, Hgb 7.5, Plt 52, ANC 6.01, Total bilirubin 0.5, ALP 90, AST 292, ALT 45, LDH >6650, Sodium 149, Potas:
Laboratory work-up on January 30, 2019 was significant for WBC 4.58, Hgb 6.8, Plt 29, ANC 6.01, Total bilirubin 0.7, Creatinine 1.09, ALP 49, ALT 42, AST 105, Sodium 140, Pc
Laboratory work-up on January 2, 2019 was unremarkable (Creatinine 0.94, WBC 6.1, Hgb 10.2, Plt 415, ANC 4.0, Total bilirubin 0.22, ALP 65, ALT 14, AST 28, CA125 17, Gluc
Laboratory work-up on February 7, 2019 was unremarkable: BUN 19, Creatinine 0.92, WBC 4.65, Hgb 11.2, ANC 3.69, Total bilirubin 0.4, ALP 98, ALT 16, AST 21, LDH 207, Cal
Laboratory work-up on March 1, 2019 was significant for BUN 17, Creatinine 0.9, WBC 10.16, Hgb 7.8, Platelets 221, ANC 9.73, Total bilirubin 0.2, Alk Phos 82, ALT 6, AST 18
Laboratory work-up on February 11, 2019 was unremarkable: BUN 17, Creatinine 0.7, WBC 9.9, Hgb 11.9, Plt 355, ANC 8.05, Total bilirubin 0.3, Alk Phos 138, ALT 29, AST 39,

Şekil 3.15. *Modellerin Testinde Kullanılan Örnek Cümleler*

Sunucu üzerinden input()(giriş) komutu kullanaraktan Şekil 3.16. da görüldüğü gibi Python ile cümle girdi alanı oluşturulabilir. Şekil 3.17. de sonuç görüntülenmektedir. İstenildiği taktirde bir adet sayfa oluşturulup buraya button(buton) ekleyip, butona da yazdığımız fonksiyonun çalışmasını çalıştırabilirsiniz. Sonrasında çıktı için result (sonuç) ekranı oluşturup kullanıcı açısından basit bir gözlem alanı oluşturulmaktadır.

Örnek cümlelerin modellerden çıktısı:

Cümle:

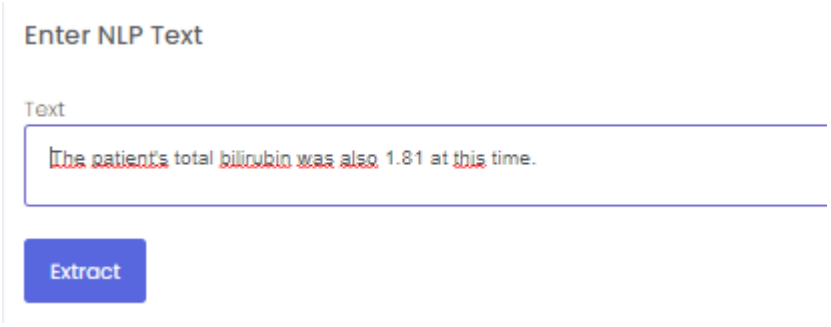
The patient's total bilirubin was also 1.81 at this time.

Ner Çıktısı:

[{"name":"The","value":"O"},{"name":"patien_s","value":"O"},{"name":"total","value":"LAB"},{"name":"bilirubin","value":"LAB"},{"name":"was","value":"O"},{"name":"also","value":"O"},{"name":"1.81","value":"VALUE"},{"name":"at","value":"O"},{"name":"this","value":"O"},{"name":"time","value":"O"},{"name":".","value":"O"}]

İlişki Çıktısı:

[{"reltag1":"LAB","reltag2":"VALUE","relstatus":"yes"}]



Şekil 3.16. *Makine Öğrenimi Cümle girişi – 1*



Şekil 3.17. *Makine Öğrenimi Cümle Çıktısı – 1*

Şekil 3.18. de istenilen veri textbox (metin kutusu) içerisine yazılıp butona tıklayıp sonucu görülmekte ve 3.19. bu sonuçlar gösterilmektedir.

Cümle:

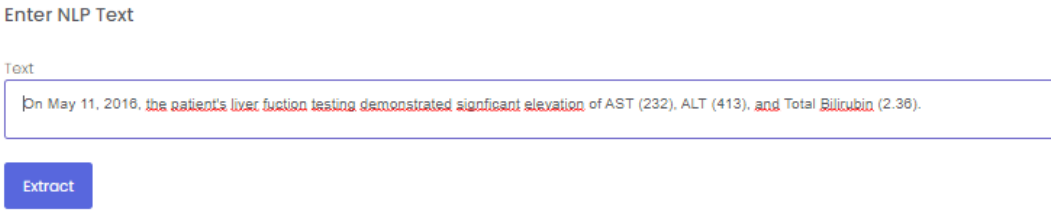
On May 11, 2016, the patient's liver fuction testing demonstrated signficant elevation of AST (232), ALT (413), and Total Bilirubin (2.36).

Ner Çıktısı:

[{"name":"On","value":"O"},{"name":"May","value":"DATE"},{"name":"11","value":"DATE"},{"name":",","value":"DATE"},{"name":"2016","value":"DATE"},{"name":",","value":"O"},{"name":"the","value":"O"},{"name":"patient_s","value":"O"},{"name":"liver","value":"LABTEST"},{"name":"fuction","value":"LABTEST"},{"name":"testing","value":"LABTEST"},{"name":"demonstrated","value":"O"},{"name":"signficant","value":"O"},{"name":"elevation","value":"O"},{"name":"of","value":"O"},{"name":"AST","value":"LAB"},{"name":"(","value":"O"},{"name":"232","value":"VALUE"},{"name":")","value":"O"},{"name":",","value":"O"},{"name":"ALT","value":"LAB1"},{"name":"(","value":"O"},{"name":"413","value":"VALUE1"},{"name":")","value":"O"},{"name":",","value":"O"},{"name":"and","value":"O"},{"name":"Total","value":"LAB2"},{"name":"Bilirubin","value":"LAB2"},{"name":"(","value":"O"},{"name":"2.36","value":"VALUE2"},{"name":")","value":"O"},{"name":".","value":"O"}]

İlişki Çıktısı:

[{"reltag1":"LABTEST","reltag2":"DATE","relstatus":"yes"},{"reltag1":"LAB","reltag2":"DATE","relstatus":"yes"},{"reltag1":"LAB","reltag2":"VALUE","relstatus":"yes"},{"reltag1":"LAB1","reltag2":"DATE","relstatus":"yes"},{"reltag1":"LAB1","reltag2":"VALUE1","relstatus":"yes"},{"reltag1":"LAB2","reltag2":"DATE","relstatus":"yes"},{"reltag1":"LAB2","reltag2":"VALUE2","relstatus":"yes"}]



Şekil 3.18. *Makine Öğrenimi Cümle Girişi – 2*



Şekil 3.19. *Makine Öğrenimi Cümle Çıktısı – 2*

3.18. Bulgular

Veri seti ile yapılan denemelerde alınan bazı sonuçlar sonrasında %80 oranında başarı oranı elde edildiğini kanıtlamaktadır. Bunlardan bazıları az veri olmasıyla beraber Bert yöntemi az sayıda veri üzerinde daha fazla ince ayar yapılmasını sağladığı için %20 oranında daha performanslı çalışmaktadır. Bert yöntemiyle ve Flair modelleriyle %100 başarı elde edemediğimizin birkaç sebebi bulunmaktadır. Bunlardan bazıları başka ek yöntemler kullanılarak çözümlenmiştir. Ancak bazıları için daha büyük verilere ihtiyaç duyulmaktadır.

Var olan ve bazı göz ardı kayıp olmasının sebebi de ince ayar yapılmasını sağlamasının yanı sıra bazı yanlış etiketlediği veya kelimeler arasında birden fazla boşluk olduğu için model olarak Ner mantığı daha düşük başarı oranı ile çalışmaktadır. Bu da %10'lik bir kayba sebep olmaktadır. Bu hatayı regex kullanarak ortadan kaldırılmaktadır.

Genel olarak %20'lik kayıp yanlış etiketlenmeden kaynaklı olmaktadır. Yani Ner modelinde olan eksiklikten kaynaklıdır. Cümleler içerisinde bulunan bazı ilaç isimleri, ameliyat isimleri, hastalıklar gibi parametreleri yanlış olarak laboratuvar değeri olarak etiketlemektedir. Bert içerisinde ince ayar yapılsa da bu hataların giderilmesi için daha büyük verilere ihtiyaç duyulmaktadır.

İlk olarak Bert ile %70 başarı oranı elde edilmiştir. Daha sonrasında %10'luk bir hata regex ile çözülmüştür. Ancak %20'lik hata payı içinde daha büyük verilere ihtiyaç duyulmaktadır.

BÖLÜM 4. SONUÇ

Yapılan girdi ve sonuç analizlerinin ardından, geliştirilen modeller ile 100 cümle arasından %80'lik bir başarı oranı ile laboratuvar sonuçlarının içinde bulunduğu cümlelere makine öğrenimi yardımıyla üretilen ve arka planında birden fazla yöntem geliştirilerek kullanılan uygulama oluşturulmuştur. Bu yöntem ile doktor bakış açısı gibi yüzdelik oranda tahmin derecesi yüksek bir yapay zeka oluşturulmuştur.

Kullanılan birçok yöntem arasından en uyumlu kütüphaneler kullanılarak öncelikli olarak veri üzerinden yapılan analize ve ihtiyaca göre deneme yanılma yöntemi kullanılmıştır. Verilerin düzgün bir şekilde temizlenmesi sağlanması için değişim tablosu oluşturulmuş ve temiz veri elde edilmiştir. Ardından nesne tabanlı kod yapısı oluşturulmuştur. Hasta verilerinin gizliliği açısında cümleleri oluşturma kısmı manuel olarak türetilmiştir. Bu şekilde belli bir düzenli yapıda cümleler elde ederek temiz veri üzerinde çalışmamız iş akışını kolaylaştırmıştır. Bu yapı genel olarak veri üzerine kurulu bir sistematik yapı olduğu için ne kadar çok veriye sahip olursanız karşılaşılacağımız cümleyi tahmin etme oranı o oranda artmaktadır. Eğitilmiş model üzerinde deneme yanılma yolu ile belirli 100 cümle arasından %80 başarılı tahmin oranına sahip bir yapay zeka üretmiş olunmaktadır.

Literatürde birden fazla makine öğrenimine dair oluşturulmuş olan dokümanları incelediğimizde sağlık alanında yöntem ve model uyumluluğunu da ekleyerek oluşturulan ve başarı oranı 4/5 oranında bir skora sahip mantık ve yöntem oluşturulmuştur. Daha iyi bir oran elde etmek içinde daha büyük veriye ihtiyaç duyulmaktadır.

KAYNAKÇA

ActiveState. (2022, 3 25). *Pandas:* https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/. Erişim tarihi:23/07/2021

ALKAN, M. A. (2022, 3 5). *Makine Öğrenimi Nedir? :* https://www.endustri40.com/makine-ogrenimi-nedir/. Erişim tarihi:01/07/2021

Beser, S. (2017, 9 26). *Uzun ve Kısa Süreli Bellek:* https://veribilimcisi.com/2017/09/26/uzun-kisa-sureli-bellek-long-short-term-memory/. Erişim tarihi: 3/12/2021

Breckbaldwin. (2009, 10 14). *Coding Chunkers as Taggers: IO, BIO, BMEWO, and BMEWO+* Erişim adresi https://lingpipe-blog.com/2009/10/14/coding-chunkers-as-taggers-io-bio-bmewo-and-bmewo/.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, 5 24). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *BERT*. Cornell University, Computer Science.

Doğan, Ö. (2020, 6 13). *Numpy Kütüphanesi Nedir? Nasıl Kullanılır?:* https://teknoloji.org/numpy-kutuphanesi-nedir-nasil-kullanilir/ Erişim Tarihi:13/08/2021.

Eğitim, U. (2021). *Python:* https://www.uzaktanegitim.com/blog/python-ile-neler-yapilabilir. Erişim tarihi:01/07/2020

IAPAC. (2021, 3). *Lab Values:* https://www.iapac.org/fact-sheet/normal-laboratory-values/. Erişim tarihi:01/12/2020

IBM. (2020, 7 15). *Makine Öğrenimi:* https://www.ibm.com/tr-tr/cloud/learn/machine-learning/. Erişim tarihi:01/06/2021

Kara, A. (2019). *Uzun Kısa Süreli Bellek Ağı Kullanılarak Küresel Güneş Işığı Zaman Serisi Tahmini:* Organizasyon. Fen Bilimleri Dergisi.

Kaşgarlı, K. M. (2021). Makine Öğrenim Sınıflandırıları *Makine Öğrenimi ile Twitter Duygu Analizi*. Kadir Has Üniversitesi: Erişim Mühendislik Ve Doğa Bilimleri Bölümü tarihi.

Kumaraswamy, B. (2021). Theories and Applications, Artificial Intelligence in Data Mining (109 - 131). *RNN:* https://www.sciencedirect.com/topics/engineering/recurrent-neural-network.

Lenzi, V. B., Speranza, M., & Sprugnoli, R. (2009). *Entity Recognition* https://www.evalita.it/evalita-2009/tasks/entity-recognition. Erişim tarihi:13/11/2021

Marshall, C. (2019, 12 18). *NER:* https://medium.com/mysuperai/what-is-named-entity-recognition-ner-and-how-can-i-use-it-2b68cf6f545d. Erişim tarihi: 04/12/2021

NLTK. (2022). *NLTK:*. https://www.nltk.org/api/nltk.tokenize.html/. Erişim tarihi:17/10/2021

Pedregosa, F. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning, (2825-2830).

Pythex. (2021). *Regex:* . https://pythex.org/. Erişim tarihi:01/08/2021

Ramshaw, M. (1995). Text Chunking using Transformation-Based Learning. ArXiv.

Schools, W. (2021). *Regex:* . https://www.w3schools.com/python/python_regex.asp/. Erişim tarihi:17/04/2021

Schweter, S., & Golde, J. (2022). *Github Repository:* . https://github.com/flairNLP/flair. Erişim tarihi:01/09/2021

ŞAFAK, H. İ. (2017). https://medium.com/t%C3%BCrkiye/makine-%C3%B6%C4%9Frenmesi-nedir-20dee450b56e. *Makine Öğrenmesi Nedir ?*

Tamir, D. M. (2020, 6 26). *Makine Öğrenimi Nedir? :* Berkeley School Of Information, Data Science Erişim tarihi:01/07/2021

W3schools. (2022). MySQL Database: https://www.w3schools.com/php/php_mysql_intro.asp. Erişim tarihi: 10/11/2021

Wikipedia. (2013, 9). *Inside–outside–beginning (tagging)*. https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging). Erişim tarihi: 11/10/2021

Zhaol, Z., Chen, W., Wul, X., Chen, P. C., & Liu, J. (2022). IET. *LSTM Networt* (1-8). Beihang University, School of Automation Science and Electrical Engineering.