T.C.
GEBZE TEKNİK ÜNİVERSİTESİ

Bilgisayar Mühendisliği Bölümü

Makine Öğrenmesi Yöntemleriyle Protein İsimlerinin Belirlenmesi

Eyyüp Aydın

Danışman

Doç. Dr. Fatih Erdoğan Sevilgen

Ocak, 2017

Gebze, KOCAELİ

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ

Bilgisayar Mühendisliği Bölümü

Makine Öğrenmesi Yöntemleriyle Protein İsimlerinin Belirlenmesi

Eyyüp Aydın

Danışman

Doç. Dr. Fatih Erdoğan Sevilgen

Ocak, 2017

Gebze, KOCAELİ

Bu çalışma ..../..../200.. tarihinde aşağıdaki jüri tarafından Bilgisayar Mühendisliği Bölümünde Lisans Bitirme Projesi olarak kabul edilmiştir.

Bitirme Projesi Jürisi

Danışman Adı	Doç. Dr. Fatih Erdoğan SEVİLGEN	
Üniversite	Gebze Teknik Üniversitesi	
Fakülte	Mühendislik Fakültesi	

Jüri Adı	Prof. Dr. Yusuf Sinan AKGÜL	
Üniversite	Gebze Teknik Üniversitesi	
Fakülte	Mühendislik Fakültesi	

Jüri Adı	Yrd. Doç. Dr. Yakup GENÇ	
Üniversite	Gebze Teknik Üniversitesi	
Fakülte	Mühendislik Fakültesi	

ÖNSÖZ

Bu projenin ve raporunun hazırlanmasında emeği geçenlere, projenin ve raporunun son halini almasında yol gösterici olan Sayın Y. Doç. Dr. F. Erdoğan SEVİLGEN 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, 2017 **Eyyüp Aydın**

İÇİNDEKİLER

ŞEKİL LİSTESİ

TABLO LİSTESİ

KISALTMA LİSTESİ

UML : Birleşik Modelleme Dili (Unified Modeling Language)

SVM : Destek Vektör Makinesi (Support Vector Machine)

CRF : Koşullu Rastgele Alanlar (Conditional Random Fields)

NER : Adlandırılmış Varlık Tanıma (Named Entity Recognition)

HHM : Gizli Markov Modelleri (Hidden Markov Models)

POS : Cümle Öğesi (Part Of Speech)

IOB : İçeri-Dışarı Başlangıç (Inside Outside Beginning)

ÖZET

Günümüzde bilimsel makalelerin sayısı ve yayımlanma hızı üstel zamanlı olarak artmaktadır. Bu ölçekte bir verinin işlenebilmesi için bilgisayar ortamında geliştirilen otomatik sistemler kullanılmalıdır.

Biyoloji literatüründe, patojen-konak, protein-protein, gen-protein ve türler arası etkileşimler önemli yer tutmaktadır. Bu tür etkileşimlerin ve daha fazlasının metin madenciliği teknikleriyle bulunabilmesi için gerekli ortak bir işlem mevcuttur: Metindeki biyolojik isimlerin belirlenmesi.

Bu çalışmada, biyolojik metin ve makalelerdeki protein isimlerinin otomatik bulunması için geliştirilen yöntemler incelenmiş ve bu alandaki bilimsel makalelerden seçilen yöntemler gerçekleştirilip, çalıştırılabilir hale getirilmiştir.

Geliştirilen üç farklı yöntemin en iyisi 64.28, en kötüsü 47.81 F-skoruna sahip olmakla birlikte ortalama F-skoru 54.86'dır. Kelime tanıma problemleri için koşullu rastgele alanlar metodunun daha iyi işlediği gözlenmiştir.

SUMMARY

Today, the number of scientific articles and the speed of publication are increasing exponentially. In order to process a data at this scale, automated systems developed in computer environment should be used.

Interactions between pathogen-host, protein-protein, gene-protein and species play an important role in biology literature. There is a common process for such interactions and more to be found by text mining techniques. This is the identification of biological names in the text.

In this study, methods developed for the finding of protein names automaticly in biological texts and articles were investigated and selected methods from the scientific articles in this area were carried out and made operable.

The average F-score of the three developed methods is 54.86, with the worst score being 47.81 and and the best score being 64.88.

1. GİRİŞ

Metin madenciliği, özellikle günümüzdeki bilimsel literatür kaynaklarının artmasıyla önemi günden güne artan bir alan olmuştur. Yeni araştırmaların yapılabilmesi için, var olan kaynaklardan çıkartılabilecek bilgilerin ışığında gidilmelidir.

Özellikle biyoinformatik gibi tıbbi alanlarda, başını PubMed'in [1] çektiği kaynaklara günde 2000'in üzerinde yeni makale girişi yapılmaktadır [2]. Bu ölçekteki verinin insanlar tarafından işlenebilmesi oldukça zordur. Bu nedenle birtakım metin madenciliği yöntemleri kullanılarak var olan bilgilerin elden geçirilmesi gerekmektedir.

Biyolojik literatür için pek çok metin madenciliği çalışmaları yapılmaktadır [3]. Bu çalışmalarda makalelerin içerisinden genler ve/veya proteinler arasındaki etkileşimler, veri tabanlarında henüz bulunmayan gen veya proteinler, mutasyonlu proteinler gibi bilgiler araştırılmıştır. Tüm bu işlemlerin yapılabilmesi için gerekli olan şey ise, metin içerisindeki gen ve protein isimlerinin belirlenebilmesidir.

Bu problemi çözmek için, gen ve protein veri tabanları kullanılarak bir isimler sözlüğü oluşturulup, metin içerisinde bunların eşleştirilmesi gibi bir yöntem kullanılabilmekle birlikte, makine öğrenmesi kullanarak metin içerisindeki bu isimler otomatik olarak belirlenebilmektedir.

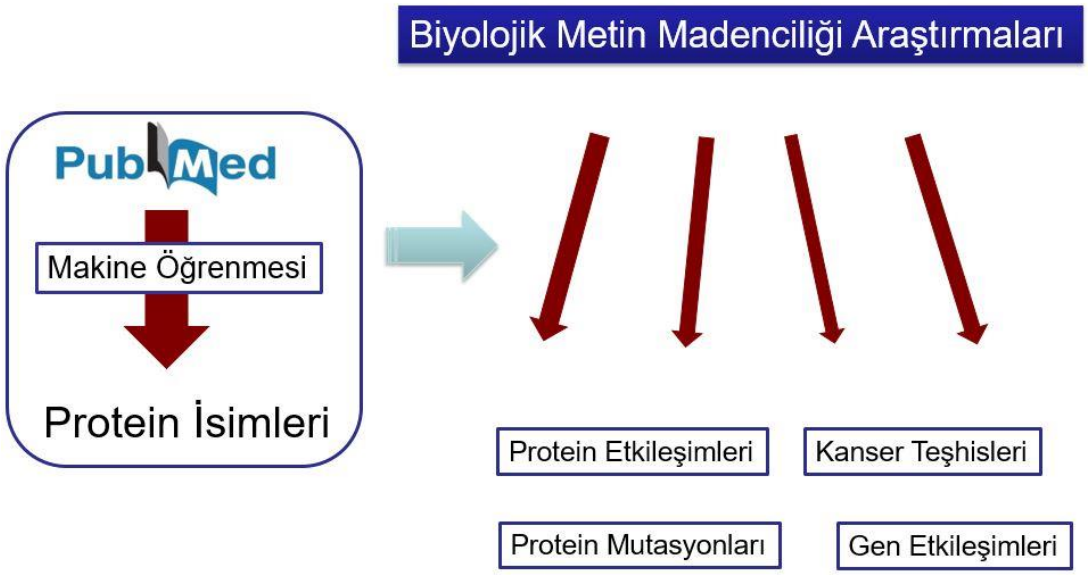
Bu çalışmada, biyolojik literatür içerisinden, ileri aşama metin madenciliği işlemleri için gerekli olan gen ve protein isimlerinin belirlenmesi işlemi makine öğrenmesi teknikleriyle incelenecektir.

Rapor boyunca projenin açıklanması, literatür araştırması, tasarlanan yapı, izlenilen iş planı ve sonuçlar izlenecektir.

1.1. PROJE TANIMI

Protein isimlerinin otomatik olarak belirlenmesi, biyoloji, biyoteknoloji ve biyoinformatik gibi alanların bilimsel makalelerin elde edilmesi, işlenmesi, makine öğrenmesi tekniklerinin uygulanması adımlarının izlenmesiyle sağlanacaktır. Protein isimlerinin

kolay bir şekilde bulunması, özellikle biyoinformatik alanındaki araştırmaların hızlanması ve daha etkili bir şekilde yapılabilmesi için önemlidir.



Şekil 1-1 Genel Yapı

1.2. PROJENİN NEDEN VE AMAÇLARI

Projenin ana amacı, metinler içerisinde İngilizce dili dahilinde bulunmayan protein isimlerinin belirlenerek, hangi makalelerin hangi proteinler ile ilgili bilgi içerdiğinin belirlenmesi sonucu, biyoinformatik alanında yapılacak diğer araştırmalar için literatürdeki fazla bilgi yığını içinden alakalı sonuçların bulunmasına yardımcı olmaktır.

Metin içinden bilgi çıkarımı alanında, alakalı isimlerin bulunması araştırılan bilgi ne olursa olsun çok etkili olmaktadır; eğer bir metin içindeki isimler silinirse, metin içerisindeki tüm bilgi ve değeri kaybedecektir.

Geliştirilmiş olan çoğu literatürden bilgi çıkarma araçları el ile hazırlanmış protein veri tabanlarını kullanarak geliştirilmiştir [4]. Ancak bu bir sınırlama teşkil etmektedir çünkü protein ve gen isimleri metin içerisinde çok farklı şekillerde yazılabilmektedir.

2. PROJE GEREKSİNİMLERİ

Bu projede başarılması gerekenler:

- Makine öğrenmesi için veri kümesi elde edilmesi/oluşturulması.
  PubMed'den makalelerin bulunup, indirilmesi.

- Makalelerin kullanılabilir hale getirilmesi.
  İndirilen makalelerin düz metin haline çevrilmesi ve temizlenmesi.

- Makine öğrenmesi yöntemlerinin geliştirilmesi (uygulanması).
  Verilerin makine öğrenmesi kütüphanelerinde çalışabilecek şekle getirilmesi.

- Geliştirilen sistemin, var olan sistemler ile farkına bakılıp, belirli bir doğruluk oranın sağlanması.

- Tüm bu sistemin bir kullanıcı için akıcı şekilde kullanılabilmesi için ara yüz geliştirimi.
  Veri kümesi girdisi, veri kümesinin hazırlanması, uygulanacak yöntemin seçilmesi ve sonuçların aktarılması adımlarının kullanıcı dostu bir ara yüz ile aktarılması.

Bunların sağlanması için gerekli ihtiyaçlar:

- Büyük veriler üstünde çalışabilecek şekilde özellikleri barındıran bir bilgisayar
- Üzerinde çalışılacak veri kümelerinin oluşturulabilmesi için makale kaynağı
- Geliştirilen sistemin karşılaştırılabilmesi için, başka bir protein ismi bulma aracı
- Intellij IDEA: Java programlarının geliştirilmesi için güzel bir tümleşik geliştirme ortamı.

3. LİTERATÜR

Varlık ismi belirlenmesi (Named Entity Recognition - NER) işlemi, Bilgi Çıkarımı (Information Extraction) işleminin bir alt dalıdır. Metin içerisindeki isimlerin (gerçek dünya objeleri: insanlar, yerler, ürünler…) nerede olduğunun bulunması ve sınıflandırılması işlemidir.

Biyoinformatik açısından, bir varlık ismi belirleyicisinin biyolojik isimlere karşı hassas olması gerekmektedir: Protein, gen, hastalık isimleri; varlık cinsleri, etkileşim kelimeleri.

Bu bölümde biyolojik isimlerin belirlenmesi için literatürde geliştirilmiş olan yöntemlerin incelemesi yapılacaktır.

3.1. SÖZLÜK TABANLI YAKLAŞIM

İsim belirleme işlemini gerçekleştirmenin temel bir yaklaşımı, metinde varlık deyişlerinin tanımlanmasına temel teşkil edebilecek sözlük gibi terimlerin açıklayıcı bir listesini kullanmaktır (terminolojik kaynaklar olarak da adlandırılır). Bu yaklaşım, sözlük tabanlı yaklaşım olarak bilinir. Metinden gelen kelime veya kelime grubu, listedeki terimle eşleşirse, isim eşleştirmesi sağlanmış sayılır. Bu yöntemin yüksek derecede hassas (*precision*) olduğu bulunmuştur ancak kötü bir geri çağırma özelliğine (*recall*) sahiptir.

3.2. KURAL TABANLI YAKLAŞIM

Kural tabanlı yaklaşım, isim belirlemenin bir diğer yaklaşımıdır. Burada kurallar, adlandırılmış varlıkların oluşum kalıplarını ve bağlamını tanımlayan varlıkları tanımaya çalışmak üzere tanımlanır. Bu yaklaşımda kurallar, sözcük-sözdizimsel özellikleri kullanarak veya mevcut bilgi listelerini kullanarak manuel olarak geliştirilir. Kural tabanlı yaklaşımların sözlük tabanlı yaklaşımlarla karşılaştırıldığında daha iyi bir performans elde ettiği belirtilir. Bununla birlikte, kurallar ağırlıklı olarak el işçiliği olduğu için zaman alıcı ve zorlayıcıdırlar. Ayrıca, kurallar, yüksek hassasiyet elde etmek için çok spesifik ve alan adına özgüdür.

3.3. MAKİNE ÖĞRENMESİ TABANLI YAKLAŞIM

Bu teknikte bir sistem, örneklerle bağlantılı özellikler yardımıyla, görev için olumlu ve olumlu eğitim örnekleri kullanarak otomatik olarak öğrenir. Seçilen makine öğrenme algoritmaları, negatif örnekleri pozitif örneklerden otomatik olarak ayırt eder ve yine de görülemeyen verilerden benzer bilgileri belirlemek için kullanılabilir. Makine Öğrenme algoritmaları genel olarak üç tipe ayrılır:

- Denetimli Öğrenme
- Yarı Denetimli Öğrenme
- Denetimsiz Öğrenme

3.4. MAKİNE ÖĞRENMESİYLE İLGİLİ ÖNEMLİ ÇALIŞMALAR

İsim belirleme alanında Gizli Markov Modelleri (Hidden Markov Models - HMM), Destek Vektör Makinesi (Support Vector Machine – SVM) ve Koşullu Rastgele Alanlar (Conditional Random Fields – CRF) teknikleri sıklıkla kullanılmaktadır.

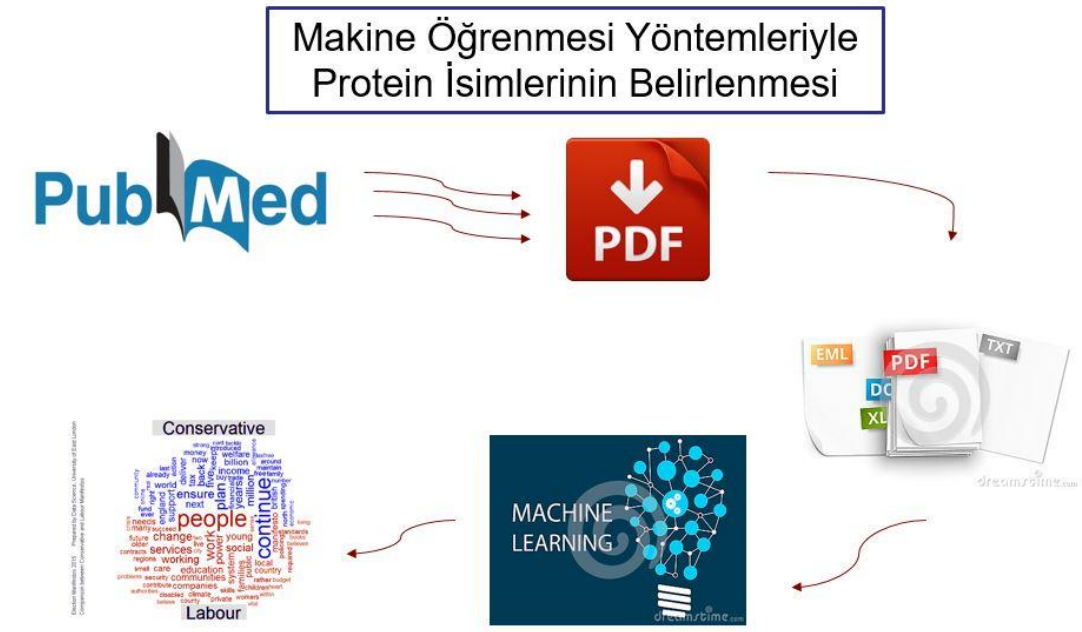
SVM'ler, otomatik metin sınıflandırmasında oldukça başarılı olmuştur. SVM'ler ikili sınıflandırıcılardır ve çoğunlukla "bire karşılık diğerleri" yaklaşımı kullanılarak eğitilirler. Bir SVM'nin eğitim süresi, eğitim setinin boyutuna göre süper doğrusaldır. Dolayısıyla, bir veride "bire karşılık diğerleri" yaklaşımıyla doğrudan eğitim mümkün değildir. Bu nedenle araştırmaların çoğu bu sorunları çözmeye yoğunlaşmıştır.

CRF’nin biyolojik isim tanımlama sistemleri için çok uygun olduğu yapılan çalışmalar sonucu gözlenmiştir.

4. SİSTEM MİMARİSİ

Öncelikle kullanıcının içerisinde protein bulunmasını istediği biyolojik metnin elde edilmiş olması gerekmektedir. Bu proje PubMed verileri üzerinden test edilmiştir çünkü bu hem alanındaki en kapsamlı portal hem de metin madenciliği projeleri için çok sık kullanılmaktadır.

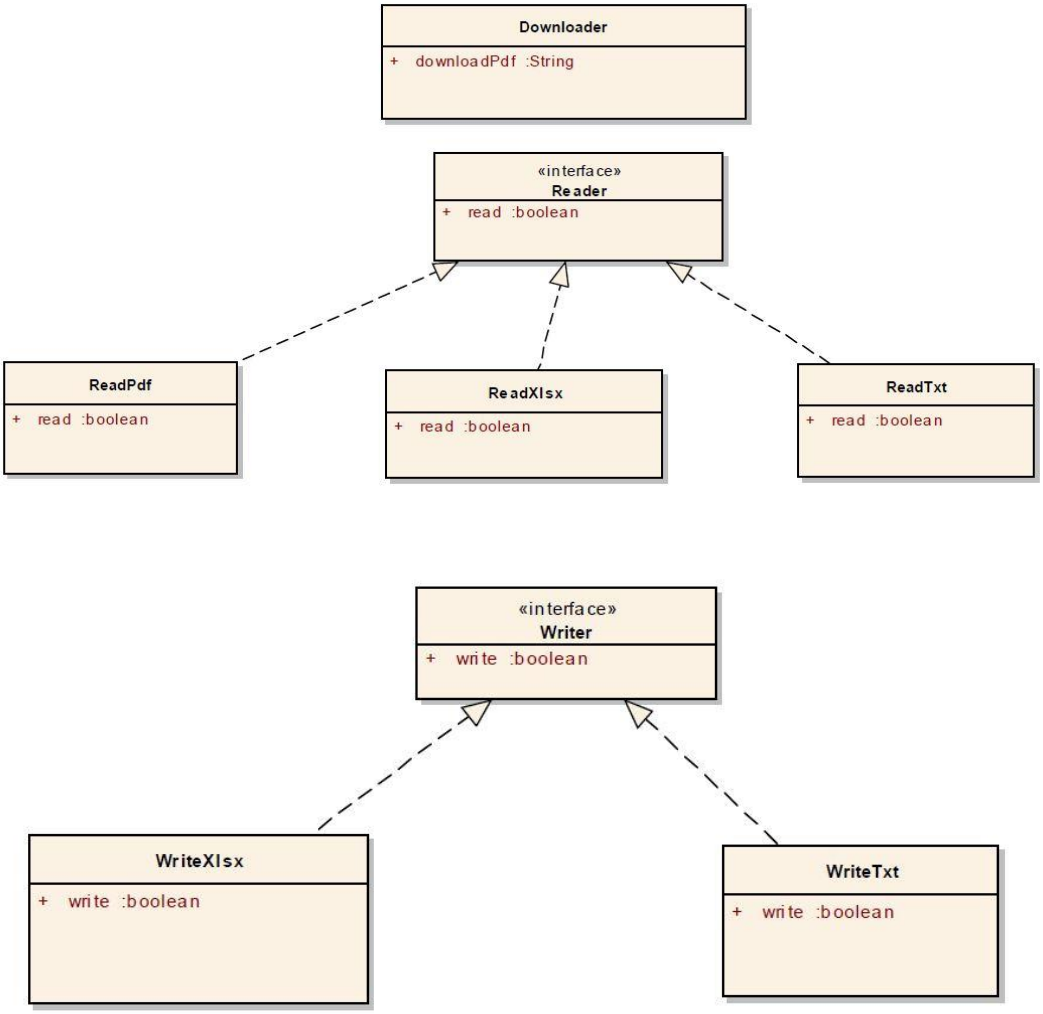
Şekil 4-1'de gösterildiği üzere makaleler düz metin haline getirildikten sonra programa girdi olarak verilmelidir ve programın, bu metni işleyerek seçilen yönteme göre bulduğu proteinleri sunması gerekmektedir.

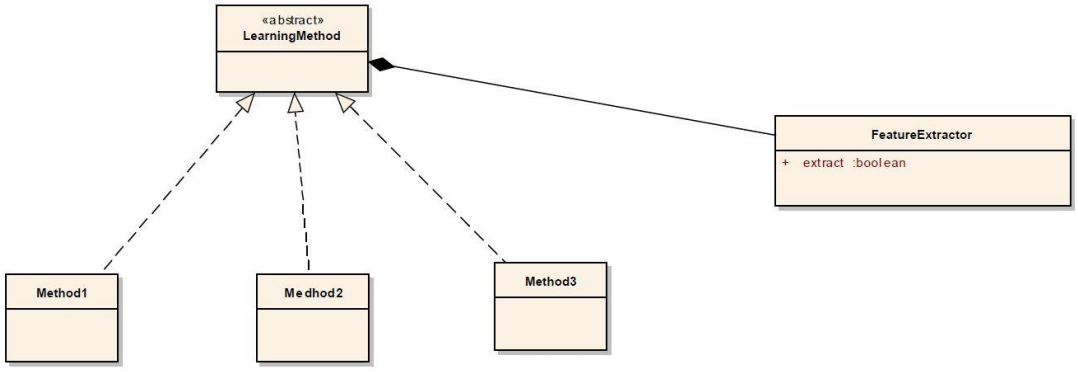


Şekil 4-1 Proje Mimarisi

4.1. UML DİYAGRAMLARI

Proje için tasarlanan sınıf diyagramları Şekil 4-2'de gösterilmiştir.


Downloader
+ downloadPdf :String
«interface»
Reader
+ read :boolean
ReadPdf
+ read :boolean
ReadXlsx
+ read :boolean
ReadTxt
+ read :boolean
«interface»
Writer
+ write :boolean
WriteXlsx
+ write :boolean
WriteTxt
+ write :boolean



Şekil 4-2 Sınıf Diyagramları

Çoğunlukla dosyalarla etkileşim halinde bulunulacağından, dosyaları okumak ve yazmak için özel sınıf hiyerarşileri oluşturulmuştur. Çevrimiçi olarak dosya indirilmesi tamamen ayrı bir işlem olduğu için buna özel sınıf oluşturulmuştur.

Bu projenin temel amacı birden çok metodun kullanılarak protein isimlerinin belirlenmesi olduğundan, metotların geliştirilebileceği bir hiyerarşi oluşturulmuştur.

5. PROJE KISIMLARI

Projede birden fazla makale gerçekleştirilmiştir. Bu kısımda, geliştirilen yöntemler ile birlikte veri kümesi anlatılacaktır.

5.1. KOŞULLU RASTELE ALANLAR

Koşullu rastgele alanlar (CRF), örüntü tanıma ve makine öğrenmede sıklıkla uygulanan istatistiksel modelleme yönteminin bir sınıfı olup, yapılandırılmış tahmin için kullanılırlar. Sıradan bir sınıflandırıcı, komşu numunelere bakılmaksızın tek bir numunenin bir etiketini öngörse de, CRF bağlamı hesaba katabilir. Örneğin, doğrusal zincirli koşullu rastgele alanlar girdi örneklerinin dizileri için etiket dizilerini öngörür.

CRF'lerin kelime etiketlemede (POS) ve isim belirlemede kullanışlı olduğu gösterilmiştir. [6][7]

Koşullu rastgele alanlar için çıkartılan özellikler şunlardır:

Büyük harf ile başlama, tümü büyük harf, içinde rakam var, doğal sayı, içinde tire var, tire ile başlıyor, tire ile bitiyor, kelimenin ön eki, kelimenin son eki, içerisinde Romen rakamı geçiyor, içerisinde Yunan harfi geçiyor ve içerisinde noktalama işaretleri var.

Bu proje kapsamında makine öğrenmesi desteği amacıyla Mallet kütüphanesi kullanılmaktadır.

5.2. DESTEK VEKTÖR MAKİNELERİ

Makine öğrenmesinde, destek vektör makineleri, sınıflandırma ve regresyon analizi için kullanılan, veriyi analiz eden ilişkili öğrenme algoritmaları ile denetlenen öğrenme modelleridir. Her biri, iki kategoriden birine ait olacak şekilde işaretlenmiş bir dizi eğitim örneği ele alındığında, bir SVM eğitim algoritması, yeni örnekleri iki kategoriden birine atayan olasılık dışı bir ikili doğrusal sınıflandırıcı halinde bir model oluşturur. Bir SVM modeli, örneklerin uzayda noktalar olarak gösterilmesidir. Ayrı kategorilerin örnekleri mümkün olduğunca açık bir boşluk ile bölünmüştür. Yeni örnekler daha sonra aynı uzaya eşlenir ve boşluğun hangi tarafına düştüğüne bakılır.

Proje kapsamında geliştirilen SVM tabanlı yöntemde, kelimeler öncelikle bir filtreleme işleminden geçmektedirler. İngilizce dilinde bulunan, gündelik hayatta kullanılan kelimeler protein ismi olmadığından, filtrelenerek çıkartılmaktadır. Böyle yapılarak yanlış sonuçların önüne geçilmeye çalışılır.

Girdilerden çıkartılan özellikler koşullu rastgele alanlar yönteminin özelliklerine ek olarak, aday kelimenin sağındaki ve solundaki kelimelerin POS (cümle öğesi) bilgisini de içerir.

5.3. KURAL TABANLI YÖNTEM

Projede, makine öğrenmesi yöntemleriyle birlikte kural tabanlı bir yöntem de geliştirilerek, iki metodun kıyaslanması amaçlanmıştır.

İşlenen kurallar, içinde harf bulunması, art arda beşten fazla rakam bulunamaması, ender karakterlerden içermemesi (#=;%°@$?!), bir DNA veya RNA dizisi olmaması, bir nükleotid ismi olmaması, içerisinde çok fazla noktalama işareti geçmemesi gibi kurallardır [8].

Bu yöntemde, öncelikle içerisinde durma kelimeleri (stop words) bulunmayan aday protein isimleri oluşturulur. Ardından, bahsedilen kurallar Regex (Regular Expression) [9] formatına getirilerek, aday protein isimlerine uygulanır.

5.4. VERİ KÜMESİ

Projede, GENIA [10] veri kümesi kullanılmaktadır. Bu veri kümesinde cümleler kelimeler halinde ayrıştırılmıştır ve her bir kelimenin yanında o kelime için etiketi bulunmaktadır.

“İçeri Dışarı Başlangıç” (Inside Outside Beginning – IOB [11]) formatına göre oluşturulmuştur.

GENIA, biyoloji tabanlı işler için oluşturulmuş olduğu için, kelimelerin yanındaki etiketler protein, gen ve hücre tipi şeklindedir.

```
In  O
137 O
cases   O
of  O
childbearing-aged   O
and O
pregnant    O
women   O
,   O
free    B-protein
form    I-protein
E   I-protein
receptor    I-protein
levels  O
(   O
sE  B-protein
)   O
in  O
serum   O
were    O
measured    O
by  O
ELISA   O
.   O
```

Şekil 5-1 GENIA veri kümesinden örnek bir cümle

Bu veri kümesinde 3856 adet cümle bulunmaktadır. Her bir cümlede en az bir etikete sahip kelime bulunmaktadır.

Proje kapsamında IOB formatında okunan eğitim dosyasından özellikler çıkartılarak hedef eğitim mekanizmasına verilir. Örneğin Şekil 5-1'deki cümle, *"In 137 cases of childbearing-aged and pregnant women, <protein> free form E receptor </protein> levels (<protein> sE </protein>) in serum were measured by ELISA."* şekline çevrilerek, hedef protein ismi çıkartıcısına yönlendirilir.

6. SONUÇLAR

Sistemin test edilmesi için, metin madenciliği alanında en sık kullanılan metrikler seçilmiştir: Precision (veya Accuracy (doğruluk)) ve Recall. Ayrıca bunların harmonik ortalaması olan F skoru da hesaplanmıştır.

Düzgün bir değerlendirme için, sonuçlar incelenerek bir karışıklık matrisi (confusion matrix) oluşturulmuştur. Bu tablodan okunan doğru-pozitif (TP), yanlış-pozitif (FP), yanlış-negatif (FN) ve doğru-negatif (TN) değerleri ile de bahsedilen metrikler hesaplanmıştır.

Precision (P),

$$P = \frac{TP}{TP + FP}$$

şeklinde tanımlanır ve bulunan sonuçların ne kadarının gerçeklerle alakalı olduğunu sorgular.

Recall (R),

$$R = \frac{TP}{TP + FP}$$

şeklinde tanımlanır ve semantik olarak, bulunması gereken sonuçların ne kadarının bulunduğunu temsil eder. Bu metriklerin birleşimi olan F-skoru ise aşağıdaki gibi tanımlanır:

$$F = \frac{2.P.R}{P + R}$$

PubMed'den rastgele seçilen on farklı makalenin en yoğun kısımlarının çıkartılması ve proteinlerin el ile belirlenmesiyle test verisi oluşturulmuştur (Tablo 1). Test verisinin her bir yönteme ve var olan sistemlere uygulanması sonucu Tablo 2'deki sonuçlara ulaşılmıştır.

Tablo 1 Test verisi bilgileri

Cümle Sayısı (On makale karışımı)	Protein Sayısı
103	45

Tablo 2 Yöntemlerin karşılaştırılması

Yöntem	**Precision**	**Recall**	**F**
SVM Tabanlı	50.3	55	52.5
CRF Tabanlı	65.2	63.4	64.28
Kural Tabanlı	45	51	47.81
ABNER	74.5	65.9	69.9
GENIA	75.20	66.56	70.61

Tablo 2'deki "SVM", "CRF" ve "Kural" tabanlı isim çıkartımı yöntemleri bu proje kapsamında geliştirilmiştir. ABNER ve GENIA ise CRF'yi kullanan diğer biyolojik isim belirleyicileridir.

Geliştirilen yöntemlerden en iyisi olan CRF, daha önce yapılmış olan kural tabanlı protein-protein etkileşimi çıkartıcı araç üzerine uygulanarak, aracın başarımının arttırılması hedeflenmiştir. Bu amaçla araç üç farklı küçük veri kümesinde CRF yöntemi varken ve yokken çalıştırılmıştır ve TP oranları ölçülmüştür. Veri kümeleri, üç farklı organizma için PubMed'de bulunan makale arşivinin indirilmesi ve içerisindeki protein-protein etkileşimlerinin belirlenmesi ile oluşturulmuştur.

Tablo 3 CRF'nin protein etkileşim çıkartıcı araca olan katkısı

Veri Kümesi	**CRF yok (%)**	**CRF var (%)**
SARS	40	45
Human Adenovirus 2	40	60
Parvoviridae	50	54
Ortalama	43.3	53

7. TARTIŞMA

Sonuçların gösterdiği üzere, biyolojik literatürden isimleri SVM tabanlı makine öğrenmesi ile çıkartımı pek yüksek başarıma sahip değilken, CRF tabanlı makine öğrenmesi bu iş için daha uygundur.

Makine öğrenmesi yöntemlerine karşılık olarak kural tabanlı yöntemin protein isimlerini bulmaktaki başarısı oldukça düşüktür. Pek çok farklı şekilde temsil edilebilen proteinler olduğu için, katı kuralların değişken girdiler karşısında başarısız olması doğaldır.

Ancak, kural tabanlı yöntemler ile makine öğrenmesi yöntemlerinin birleştirilmesi ile başarımı daha yüksek sistemler geliştirilebilir.

Geliştirilen yöntemler, literatürde yer etmiş diğer protein ismi bulucularıyla karşılaştırılacak olursa kabul edilebilir düzeyde oldukları görülmektedir. Bu yöntemler ileri düzey parametre ayarlamaları ve optimizasyonlar ile geliştirilerek, diğer protein ismi bulucularla aynı seviyeye veya daha iyi bir seviyeye getirilebilir.

Tablo 3'te görüldüğü üzere makine öğrenmesi teknikleri, var olan sistemlere katkı sağlayarak daha iyi bir araç haline getirmektedir. Biyoinformatik alanında yapılan metin madenciliği araştırmaları sıklıkla biyolojik terimlerin bulunmasına dayanır. Bu araştırmalar ilk olarak hedef terimleri belirleyip, bunun üzerinden hedef bilgilere giderler. Buradaki araç da, makale içerisindeki patojen proteini ve konak proteini arasında tanımlanmış olan etkileşimleri aramaktadır. Ancak öncelikle proteinlerin belirlenmesi gerekmektedir ki bunların üzerinden etkileşim aranabilsin.

Önceki yöntemde bir sözlükten karşılaştırma yapılarak proteinler bulunmaya çalışılmıştır. Ancak bu yöntem, pek çok proteini bulmakta başarısız kalmıştır çünkü proteinler çok farklı şekilde isimlendirilebilmektedir. Bu nedenle isimlerin belirlenmesi işlemi makine öğrenmesiyle yapılarak, başarım arttırılmaya çalışmıştır. Sonuçların gösterdiği üzere, makine öğrenmesiyle proteinlerin belirlenmesi, diğer metin madenciliği çalışmalarına büyük katkılar sağlamaktadır.

8. BAŞARI KRİTERLERİ

1. Herhangi bir makaleden en az %70 doğruluk oranı ile protein isimlerinin bulunması.
2. Var olan kural tabanlı protein etkileşimi çıkarma aracının doğruluk oranının %10 arttırılması.
   Daha önceden geliştirilmiş olan bir "protein-protein etkileşim çıkartımı" aracının başarımının artmasına katkı sağlamak.
3. En az 3 tane yöntemin (makalenin) geliştirilmesi.
   Protein isimlerinin belirlenmesi için 3 tane farklı yöntemin çalıştırılabilir hale getirilebilmesi.

Ortaya konan başarı kıstaslarından ilki, yakalanması zor bir hedef olmakla birlikte, ancak yaklaşılabilmiştir. Bu doğruluk oranının yakalanmasındaki zorluğun sebebi, biyoloji alanında değişken isimlendirme terminolojisidir.

Var olan protein etkileşim bulucu aracın başarım oranı %22 arttırılmıştır ve hedeflenenin üzerine çıkılmıştır.

Proje kapsamında, üç farklı yöntem ile protein isimleri çıkartılmıştır ve yöntemler değerlendirilmiştir.

KAYNAKLAR

[1] PubMed, National Center for Biotechnology Information, U.S. National Library of Medicine, https://www.ncbi.nlm.nih.gov/pubmed [Ziyaret Tarihi: 21.12.2016]

[2] Guan, J. (2016). A study of the use of keyword and keyphrase extraction techniques for answering biomedical questions.

[3] Int J Med Inform. 2002 Dec 4;67(1-3):49-61. Protein names and how to find them. Franzén K(1), Eriksson G, Olsson F, Asker L, Lidén P, Cöster J.

[4] Shi, L., & Campagne, F. (2005). Building a protein name dictionary from full text: a machine learning term extraction approach. BMC Bioinformatics, 6, 88.s

[5] He, Y., & Kayaalp, M. (2008). Biological Entity Recognition with Conditional Random Fields. AMIA Annual Symposium Proceedings, 2008, 293–297.

[6] Lafferty, J., McCallum, A., & Pereira, F. (2001, June). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the eighteenth international conference on machine learning, ICML (Vol. 1, pp. 282-289).

[7] McCallum, A., & Li, W. (2003, May). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4 (pp. 188-191). Association for Computational Linguistics.

[8] Mika, S., & Rost, B. (2004). NLProt: extracting protein names and sequences from papers. Nucleic Acids Research, 32(Web Server issue), W634–W637. http://doi.org/10.1093/nar/gkh427

[9] Regular expression, https://en.wikipedia.org/w/index.php?title=Regular_expression&oldid=756907682 [Ziyaret Tarihi: 25.12.2016]

[10] Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi, and Jun'ichi Tsujii. (2003). "GENIA Corpus - a semantically annotated corpus for bio-textmining." In: Bioinformatics. 19(suppl. 1).

[11] Inside Outside Beginning. (2016, January 23). In Wikipedia, The Free Encyclopedia. Retrieved 10:28, January 23, 2016, from https://en.wikipedia.org/w/index.php?title=Inside_Outside_Beginning&oldid=701241154