T.C.
GEBZE TEKNİK ÜNİVERSİTESİ

Bilgisayar Mühendisliği Bölümü

NLP ve Ontoloji Yapısı İle Tıbbi Metinlerde Arama

FERHAT ŞİRİN

Danışman
Yrd.Doç.Dr. Burcu Yılmaz

Ocak, 2020
Gebze, KOCAELİ

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ

Bilgisayar Mühendisliği Bölümü

NLP ve Ontoloji Yapısı İle Tıbbi Metinlerde Arama

FERHAT ŞİRİN

Danışman
Yrd.Doç.Dr. Burcu Yılmaz

OCAK, 2020
Gebze, KOCAELİ

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

Bitirme Projesi Jürisi

Danışman Adı		
Üniversite		
Fakülte		

Jüri Adı		
Üniversite		
Fakülte		

Jüri Adı		
Üniversite		
Fakülte		

ÖNSÖZ

Bu kılavuzun hazırlanmasında emeği geçenlere, kılavuzun son halini almasında yol gösterici olan Sayın Yrd.Doç. Dr. Burcu Yılmaz 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, 2020 **Ferhat ŞİRİN**

İÇİNDEKİLER

ŞEKİL LİSTESİ

TABLO LİSTESİ

KISALTMA LİSTESİ

NLP : Natural Language Processing (Doğal Dil İşleme)

CBOW : Continuous Bag of Words (Devamlı Kelime Kümesi)

SG : Skip-Gram

PV-DM : Paragraph Vector-Distributed Memory (Dağıtık Hafıza)

PV-DBOW : Paragraph Vector-Distributed Bag of Word (Dağıtık Kelime Kümesi)

RDF : Resource Description Framework (Kaynak Tanımlama Çerçevesi)

RFDS : Resource Description Framework Schema (Kaynak Tanımlama Şeması)

OWL : Ontology Web Language (Varlık bilimi Ağ Dili)

ÖZET

Günümüzde gelişen yeni nesil teknolojiler ile birlikte, her alanda yapılan araştırmalar artmış ve bilgi akışı hızlanmıştır. İnternet ile ulaşabileceğimiz milyonlarca makaleyi sınıflandırmak, insan eliyle yapılamayacak boyutlara ulaşmıştır. Hızla artan bu bilgiyi sınıflandırmak ve anlamlı hale getirmek için bilgisayarların hızından faydalanmak gerekmektedir.

Doğal dil işleme algoritmaları ve varlık bilimi (Ontoloji) kullanılarak, bilgiyi sınıflandırma ve anlam kazandırma işlemi başarılı bir şekilde gerçekleştirilebilir. Her iki yöntemde kendisine verilen veri kümesini birbiri içinde ilişkilendirip, sınıflandırarak, bilgiyi anlamlandırmamızı sağlar.

Bu projede, doğal dil algoritmaları ve varlık bilimi kullanılarak, tıp alanında yazılmış makalelerden alınan bilgiler sınıflandırılmıştır. Bu sayede herhangi bir hastalık belirtisinden yola çıkılarak, bu belirtilere sebep olabilecek hastalıkların bulunması sağlanmıştır. Hastalıkları anlatan makalelerden hastalığın nedenleri ve belirtileri tespit edilerek, aralarında ilişki kurulmuş ve hastalık tespitinin hızlandırılması amaçlanmıştır.

Projede, nörolojik hastalıklar incelenmiş, bu hastalıkların nedenlerini ve belirtilerini anlatan makalelerden veri tabanı oluşturulmuştur. Bu veri tabanı eğitilerek, hastalık ile belirtileri ve nedenleri arasındaki ilişki kurulmuştur. Bu sayede hastalığın belirtilerinden ve nedenlerinden yola çıkarak hastalık tespit edilmeye çalışılmıştır.

SUMMARY

With the next generation technology, researchs on any specific field are increased and flow of information is getting faster nowadays, We can reach millions of article via internet but classifying that many article is nearly impossible for the humans. We need help from the computers to classify that information and get something meaningful from it.

Natural language processing algorithms and ontology can help us to classify information and extract meaningful data for us. Both methods correlate given dataset and extract meaningful information from it.

In this project, natural language processing algorithms and ontology are used to classify medical articles. Thus, based on any syndrome, we can find the disease that causes it. Syndrome and cause of the disease are detected from the articles and revealed the relationship between them. So that we can determine the disease faster.

Neurologic diseases are examined in the project. A dataset is created that consist of articles studying on neurologic diseases. The dataset is trained to correlate the syndrome and cause of the diseases. So that we find the disease by looking its syndrome and causes.

1. GİRİŞ

1.1 PROJENİN TANIMI ve AMACI

Tıp alanındaki gelişmeler gün geçtikçe hız kazanmakta ve bu alanda yapılan yeni çalışmaları takip etmek giderek zorlaşmaktadır. İnternet üzerinden ulaşılabilecek milyonlarca makaleyi yorumlamak ve anlamlı bilgi çıkarmak, insan eliyle oldukça zordur. Sürekli yeni hastalıklar keşfedilmekte ve bu hastalıkların nedenlerini ve belirtilerini anlatan makaleler yayınlanmaktadır.
Nufüsun artmasıyla birlikte, sağlık hizmetine duyulan talepte artmıştır. Bu alanda bilgisayarlardan faydalanarak, hastalık tespiti ve nedenlerini ortaya çıkaran algoritmalara ihtiyaç duyulmuştur.

Projede, nörolojik hastalıklar üzerine yazılmış makalelerin doğal dil işleme ve varlık bilimi kullanılarak sınıflandırılması ve anlamlı bilgi elde edilmesi amaçlanmıştır. Hastalık belirtilerinden yola çıkarak, bu belirtilere sebeb olan hastalıkların bulunması ve bu hastalık üzerine yazılmış makalelerin incelenebilmesi sağlanmıştır.

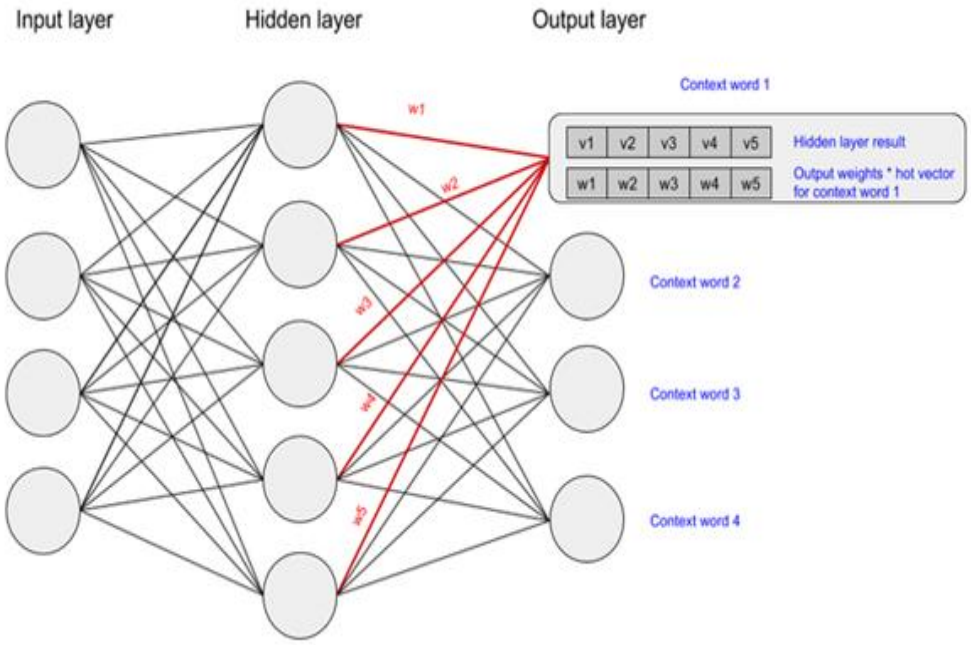
Bu iki yöntem birbilerinden bağımsız bir şekilde test edilip, birbirlerine göre avantajları ve dezavantajları ortaya çıkarılmıştır.

1.2 PROJEDE KULLANILAN YÖNTEMLER

1.2.1 DOĞAL DİL İŞLEME

Doğal dil işleme yapay zekanın gelişimi ve dil bilimle ortaklaşa geliştirilen çalışmalar sonucunda hayatımıza girmiş bir terimdir. En geniş kapsamıyla doğal dil işleme, Türkçe, İngilizce gibi doğal dillerdeki metinlerin, ses dalgalarının bilgisayar tarafından algılanarak yazılım programında çözümlenmesi ve bilgisayar ortamına aktarılmasıdır. Bilim insanları, doğal dil işleme üzerinde 50 yılı aşkın zamandır çalışmaktadırlar.

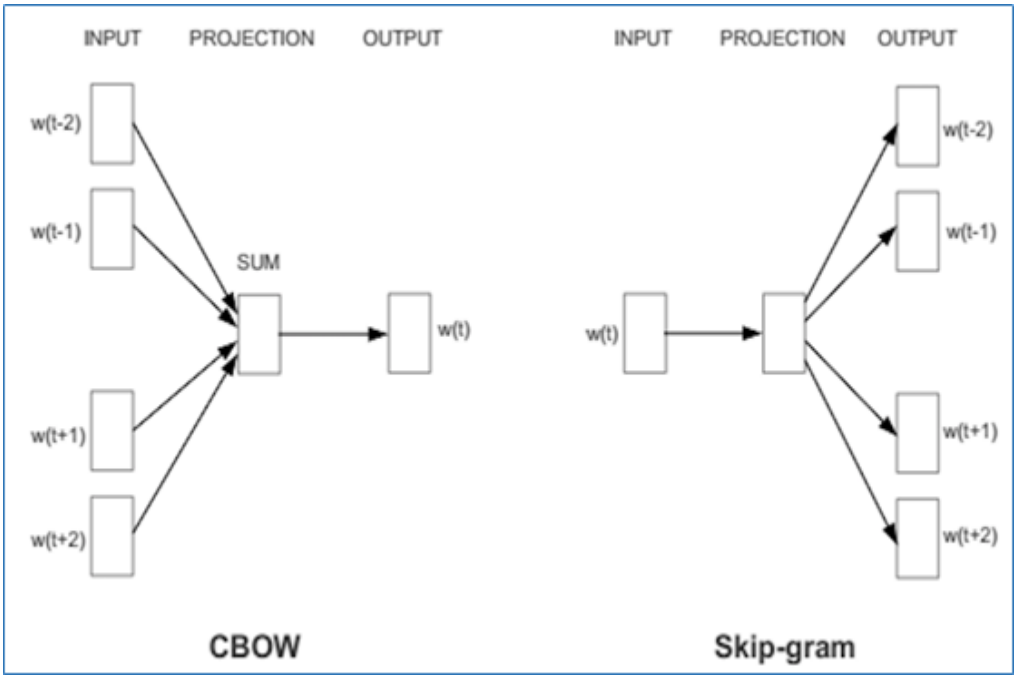
Doğal dil işleme algoritmalarıyla amaçlanan, verinin içersinden anlamlı ve istenilen bilginin çıkarılmasıdır. Bilgisayarın kelimeleri anlamlandırması için her kelimeye özgü bir vektör verilip, kelimenin kullanım yerine ve sıklığına göre vektörü düzenleyerek, birbirleriyle ilişkili vektörler haline getirilmeye çalışılır. Ortaya çıkarılan bu vektörler, bilgisayar için kelimeleri temsil eder.



Şekil 1-1 : Derin Ağ Örneği

Şekil 1-1 ‘de gösterilen derin ağ örneğinde olduğu gibi kelimeler için belirlenen vektörler bir ağırlık vektöründen geçerek birbirleri arasında ilişkilendirilir.

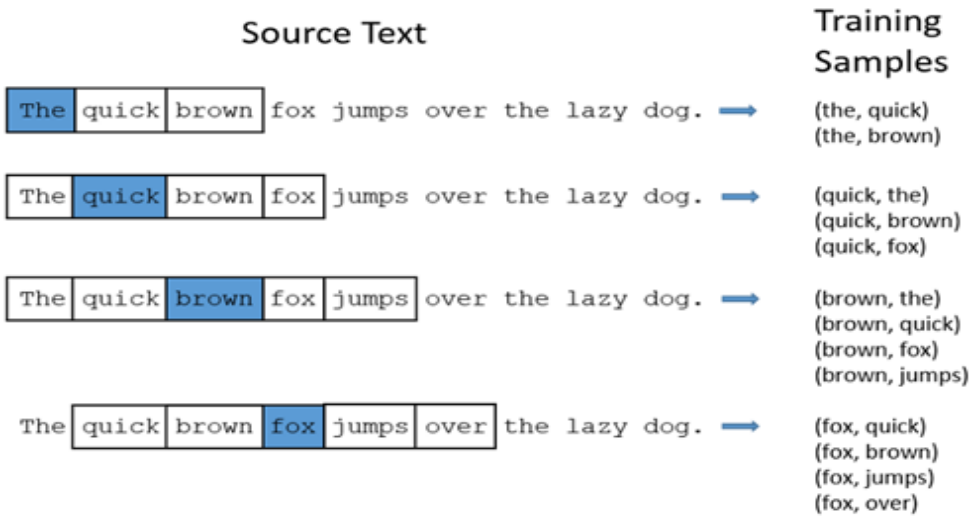
Doğal dil işlemenin çeşitli yöntemleri vardır. Son dönemlerde yaygın olarak kullanılan ve projede de yararlanılan Word2Vec ve Paragraph2Vec, bu yöntemlerden bazılarıdır.



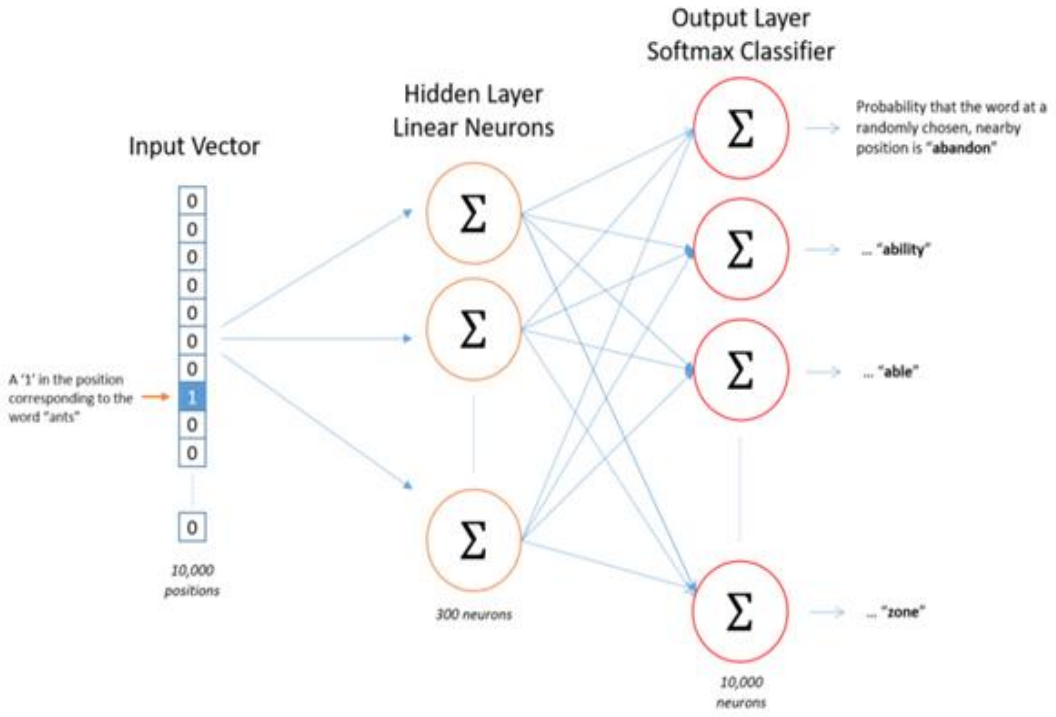
Şekil 1-2 : Word2Vec Algoritmasında Kullanılan 2 Farklı Yöntem

Word2Vec algoritması 2 farklı yöntemle gerçeklenebilir. CBOW (Devamlı Kelime Kümesi) veya Skip-Gram yöntemlerinden biri kullanılarak Word2Vec algoritması yazılabilir.

CBOW algoritmasında cümlenin anlamından, merkezdeki kelime türetilirken, Skip-Gram yönteminde merkezdeki kelimeden cümle anlamı çıkartılır. Pencere boyutu ise merkezdeki kelimenin çıkarılması için etrafında bulunan kaç kelimeye bakılması gerektiğini bildirir. Türkçe cümlelerde özne genellikle başta, fiil ise sondadır. Bu sebeble pencere boyutunun büyük olması Türkçe metinleri anlamlandırma da daha başarılı sonuç vermesini sağlar.



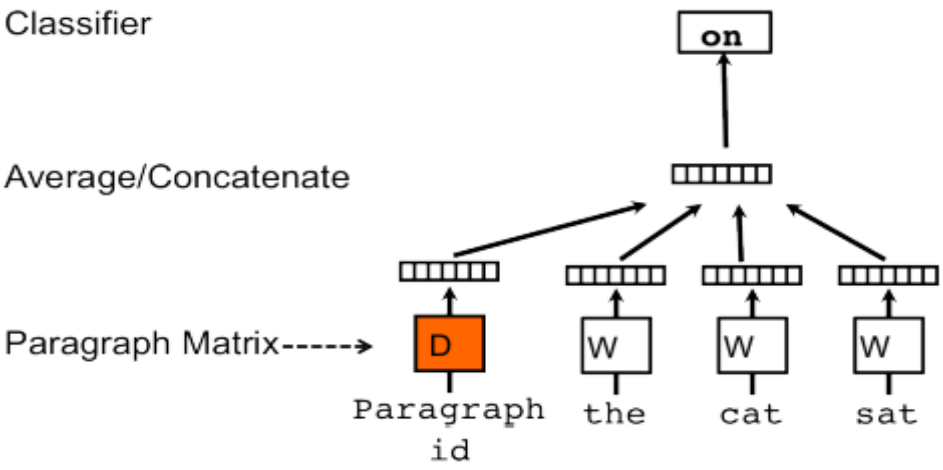
Şekil 1-3 : Pencere Boyutu 2 Olan CBOW Örneği



Şekil 1-4: Skip-Gram Örneği

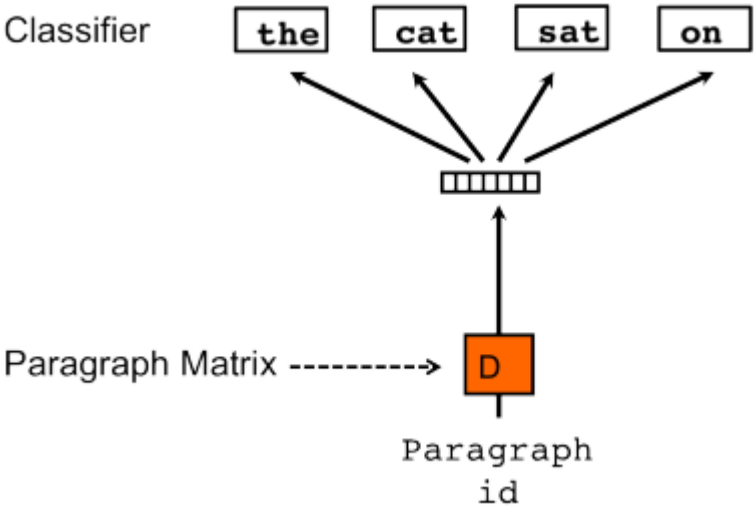
Word2Vec ve Paragraph2Vec modelleri 2 katmanlı ağ modelidir. Giriş ağırlık vektörü ve çıkış ağırlık vektörü olmak üzere 2 vektör oluşturulur. Bu vektörlere başlangıçta rastgele değerler atanır, her bir döngüde bu değerler yenilenerek algoritma sonunda son haline alırlar.

Paragraph2Vec yönteminde ise giriş vektörüne paragraf numarası eklenerek, ağ modeli oluşturulur. Böylelikle her bir paragraf ayrı bir vektör olarak temsil edilir. PV-DM veya PV-DBOW yöntemleri kullanılarak Paragraph2Vec algoritması gerçeklenebilir. PV-DM yöntemi Word2Vec'teki CBOW yöntemine benzer olup, kelime vektörüne paragraf numarası eklenmiş bir şekilde giriş vektörü oluşturulur. PV-DBOW yöntemi ise Word2Vec'teki Skip-Gram yöntemine benzer olup, bu sefer merkezdeki kelime yerine paragraf numarası kullanılarak cümle anlamı çıkarılmaya çalışılır.



Şekil 1-5 : PV-DM Örneği

Şekil 1-5 örneğinde görüldüğü gibi PV-DM yönteminde cümleye paragraf numarası eklenerek merkezdeki kelime çıkarılmaya çalışılır.



Şekil 1-6 : PV-DBOW Örneği

Şekil 1-6 örneğinde görüldüğü gibi PV-DBOW yönteminde paragraf numarasından cümle anlamı çıkarılmaya çalışılır.

Paragrph2Vec ve Word2Vec dışında da birçok doğal dil işleme yöntemi vardır. Fakat yapılan testler sonucunda Paragraph2Vec yöntemi doğru veri kümesi ile çalışıldığında, diğer yöntemlere göre daha az hata payına sahiptir.

Model	Error rate
Vector Averaging	10.25%
Bag-of-words	8.10 %
Bag-of-bigrams	7.28 %
Weighted Bag-of-bigrams	5.67%
Paragraph Vector	**3.82%**

Tablo 1-1 : Doğal Dil İşleme Yöntemleri Hata Payları

1.2.2 Varlık Bilimi

Varlık bilimi (Ontoloji) terim olarak olmak yada olmamak kavramı üzeride durur. Yani birşeyin var olup olmamasından, nasıl olduğuna kadar uzanan süreç varlık bilimidir. Varlık biliminin kökleri, felsefenin bir alt konusu olan dil bilimine dayanmaktadır. Buna göre kelimelerin anlamlarından yola çıkılarak, cümle ve hatta paradigmaların anlaşılması ve tam olarak olup olmadıkları, varsa nasıl oldukları ve hangi gruba ait oldukları ve hatta bu gruplar arası ilişkiler incelenmektedir.

Bilgisayar biliminde ise varlık bilimi, belirli bir tanım kümesi içindeki kavramların ve bu kavramlar arasındaki ilişkilerin kurallı bir şekilde temsili olarak tanımlanabilir. Ontolojiler yapay zeka, anlamsal web, ve yazılım mühendisliğinde, dünyanın tamamının veya bir kısmının temsil edilmesinde kullanılır ve tanımladığı kavramların en önemlileri şunlardır:

- Bireyler: temel nesneler
- Sınıflar: kümeler, nesne tipleri
- Özellikler: nesnenin sahip olduğu nitelikler, karakteristikler ve parametreler
- İlişkiler: nesnelerin birbirleri arasındaki etkileşim yollarının tanımları
- Kurallar

Bireyler

Bireyler ontolojinin temel seviyedeki bileşenleridir. Ontolojideki bireyler; insan, hayvan, araba, molekül, gezegen gibi somut nesneler olabileceği gibi sayılar ve kelimeler gibi soyut nesneler de olabilirler. Ancak ontolojilerin örneklendirilmesi zorunlu değildir, herhangi bir üyesi bulunmayan ontolojiler de olabilir.

Sınıflar

Ontolojide sınıflar, soyut gruplar ve kümelerdir, başka sınıfları içerebilirler ve birkaç sınıfın birleşiminden oluşabilirler. Örneğin;

İnsan: tüm insanların sınıfı
Araba: tüm arabaların sınıfı

Özellikler

Bir ontolojideki nesnelere özellik atayabiliriz. Her özelliğin bir adı ve en az bir değeri olmak zorundadır. Örneğin, Ferhat nesnesi şu özelliklere sahip olabilir:

İsim : Ferhat
Numarası : 161044080
Sınıfı: Öğrenci

İlişkiler

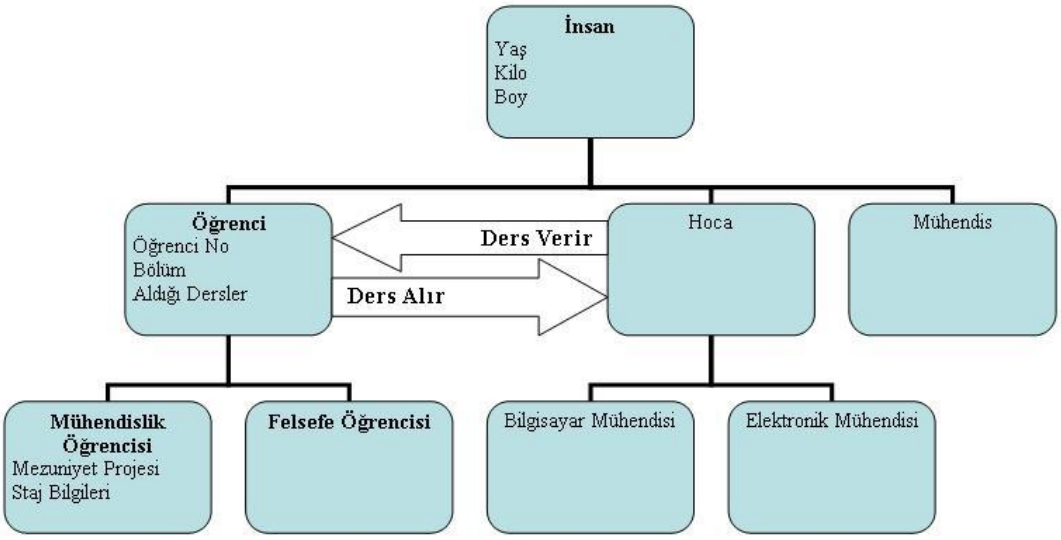
Yukarıda bahsettiğimiz özelliklerin verimli kullanımı nesneler arasında ilişki tanımlayarak kurulur. Burada ilişki adı altında bahsettiğimiz şey aslında, değeri başka bir nesne olan bir özellikten ibarettir. Örneğin :

Öğrenci sınıfı ile insan sınıfı arasında bir kapsama ilişkisi vardır. Öğrenci sınıfı, insan sınıfının bir alt kümesidir.

Ontolojilerde en önemli ilişki kapsama ilişkisidir. Bu ilişki, bize hangi nesnelerin hangi sınıflara ait olduğunu gösterir.

Kurallar

Ontoloji kuralları "eğer böyleyse, sonuçları bunlardır" yapılarından oluşur. Bu yapıların ontoloji ile birleşimi sağlandığında, ontoloji üzerinde sorgu yoluyla çıkarım yapabiliriz.



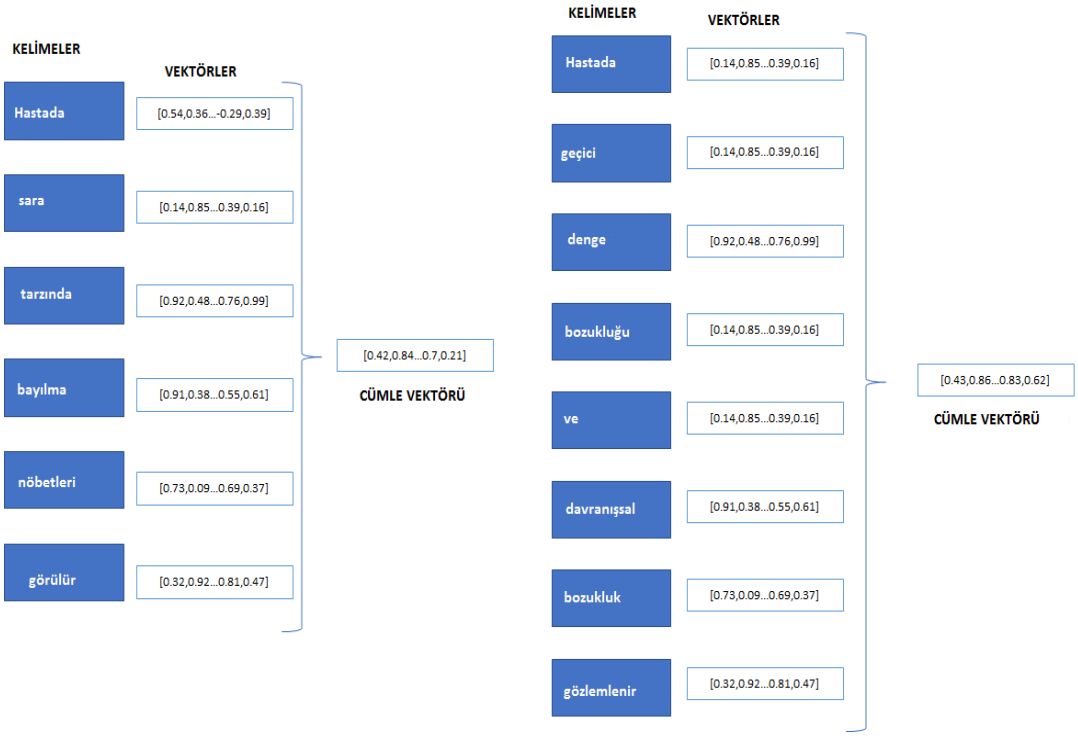
Şekil 1-7 : Ontoloji Örneği

1.3 PROJEDE KULLANILAN VERİ KÜMESİ

Projede nörolojik hastalıklara yer verilmiştir. Toplanan veriler büyük oranda, Wikipedia Türkçe sayfasında paylaşılan hastalıkları, nedenleri ve belirtileriyle anlatan makalelerden alınmıştır. Epilepsi dergisi ve diğer Türkçe tıp alanında yayın yapan kaynaklardan da yararlanılmıştır.

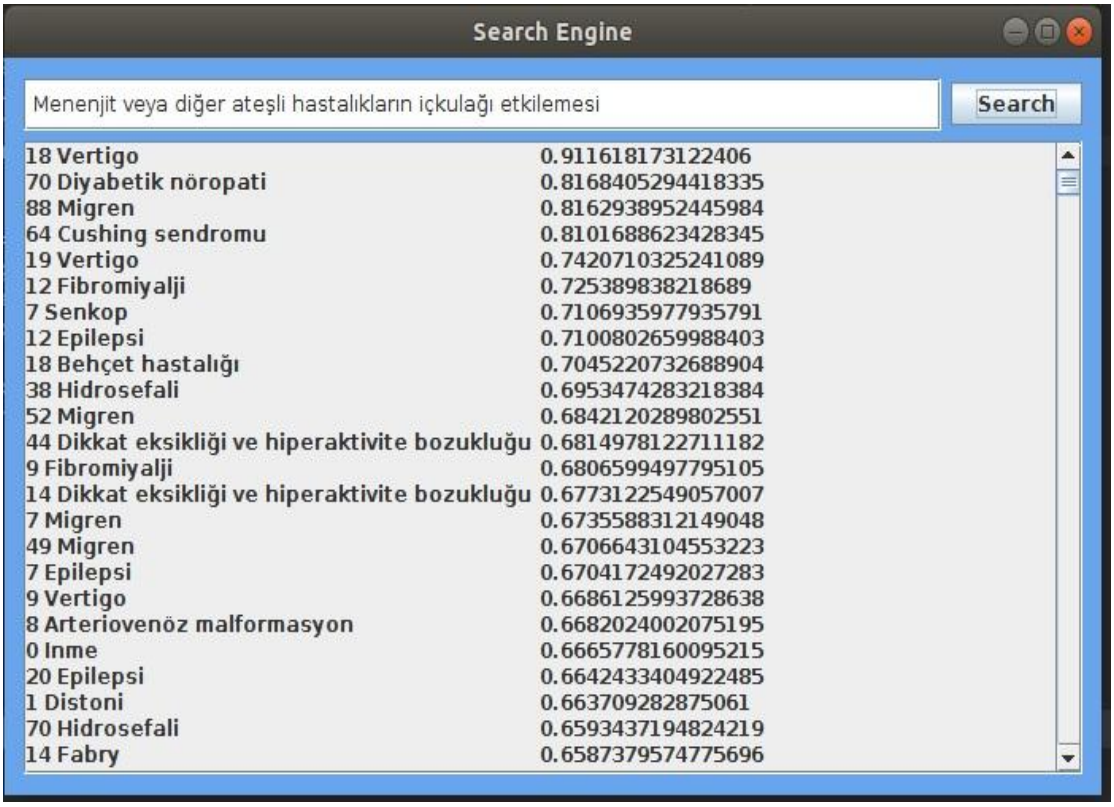
2. Projede Doğal Dil İşleme Kullanımı

Projede doğal dil işleme algoritması olarak Paragraph2Vec yöntemi kullanıldı. Algoritmanın gerçeklenmesinde Gensim kütüphanesinden yararlanılmıştır. İlk olarak bir öğrenim modeli oluşturulmuştur. Her bir hastalığın, paragrafları ayrı ayrı vektörler olarak ele alınıp PV-DBOW (Dağıtık Kelime Kümesi) yöntemi kullanılarak eğitim modeli oluşturulmuştur. Veri kümesini hastalık türlerine göre ayırmak yerine, paragraflarından ayırarak birbirleri arasında ilişkilendirmeyi sağlanma hedeflenmiştir. Oluşturulan bu eğitim modeli kaydedilmiştir ve bu eğitim modeli üzerinde arama yapılabilecek hale getirilmiştir.



Şekil 2-1 : Paragraph2Vec'in Projede Kullanımı

Aranmak istenilen cümle, eğitim modeli kullanılarak vektör haline getirilir. Oluşturulan bu vektör eğitim modelinde kaydedilmiş diğer vektörlerle karşılaştırılır. Kosinüs benzerliği yöntemi kullanılarak en yakın vektörler bulunur.



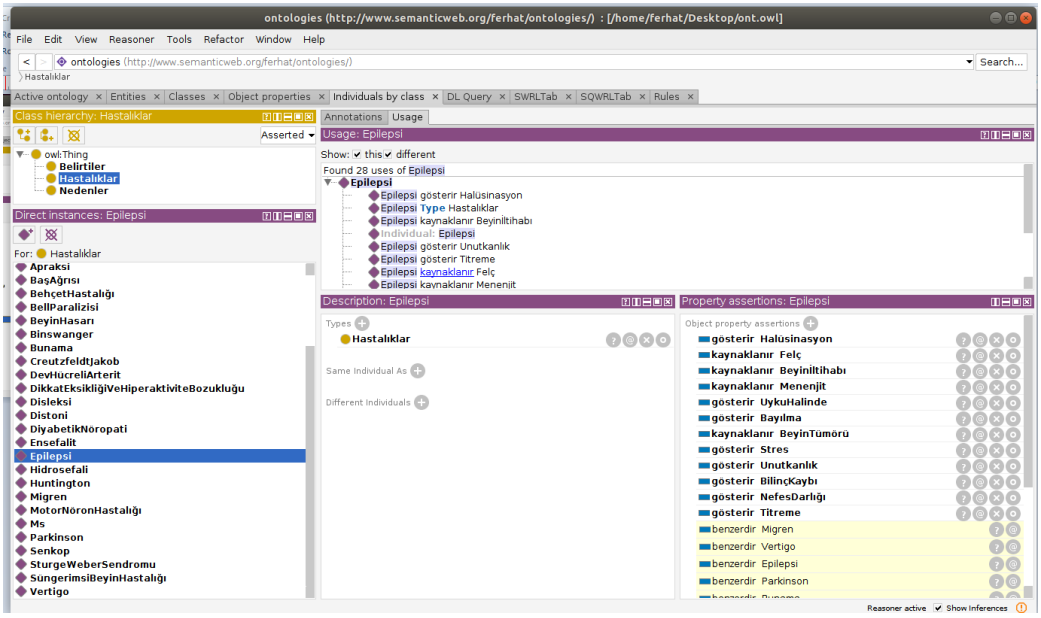
Şekil 2-2: Projede Doğal Dil İşleme Kullanım Örneği

Örnekte verilen “Menejit veya diğer ateşli hastalıkların iç kulağı etkilemesi” vertigo hastalığının nedenlerinden biridir. Vertigo hastalığının solundaki “18” değeri, hastalığın makalesindeki paragraf numarasını temsil eder, sağındaki değer ise çıkan vektörün o paragrafa olan kosinüs benzerliğini gösterir.

Paragraph2Vec yönteminde kullanılan giriş ve çıkış ağırlık vektörlerine, başlangıçta rastgele değerler atanır ve her bir döngüde bu değerler kelimelerin kullanımına göre asıl değerlerine yakınsarlar. Bu nedenle sonuçlar değişkendir, aranan hastalık bazen 2. veya 3. dereceden benzerlik gösterebilir. Eğer başlangıçta atanan değerler çok kötü ise çıkan sonuçlar beklenenden farklı olabilir. Bu nedenle eğitim algoritması defalarca test edilip, en uygun sonuçları veren algoritma kaydedildi.

3. Projede Varlık Biliminin Kullanımı

Projede doğal dil işleme algoritmalarının yanı sıra, varlık bilimi de kullanılarak hastalık tespit edilmeye çalışılmış ve bu iki yöntem arasındaki farklılıklar ortaya çıkarılmıştır. Varlık bilimini tanımlamak için "Hastalıklar", "Nedenler" ve "Belirtiler" olmak üzere üç sınıf ve bu sınıflara ait bireyler oluşturulmuştur. Sınıflar arasında "gösterir", "kaynaklanır" ve "benzerdir" ilişkileri kurulmuştur.



Şekil 3-1: Protege Editöründe Oluşturulan Varlık Bilimi

Varlık biliminin kurulumunda protege ontoloji editöründen yaralanılmıştır. Varlık bilimi OWL, RDF and RDFS dilleri kullanılarak inşa edilmiştir. RDF ile bireyler, RDFS ile sınıflar ve OWL ile sınıflar arasındaki ilişkiler kurulmuştur.

Varlık biliminde sorgular SPARQL dili ile yapılmıştır. SPARQL ile RDF kullanılarak tanımlanmış bireylerin aralarındaki ilişkilere ulaşılabilir.

SPARQL sorgu örneği :

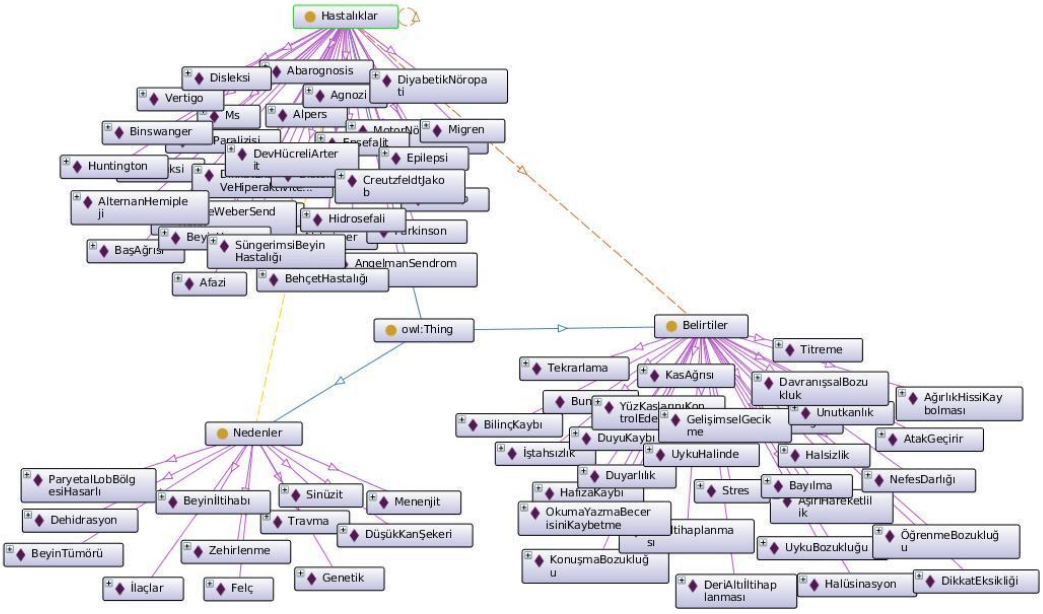
SELECT ?x WHERE { Migren rdfs:type ?x .}

SELECT ?x WHERE { ?x ont:gösterir ont:DikkatEksikliği . }

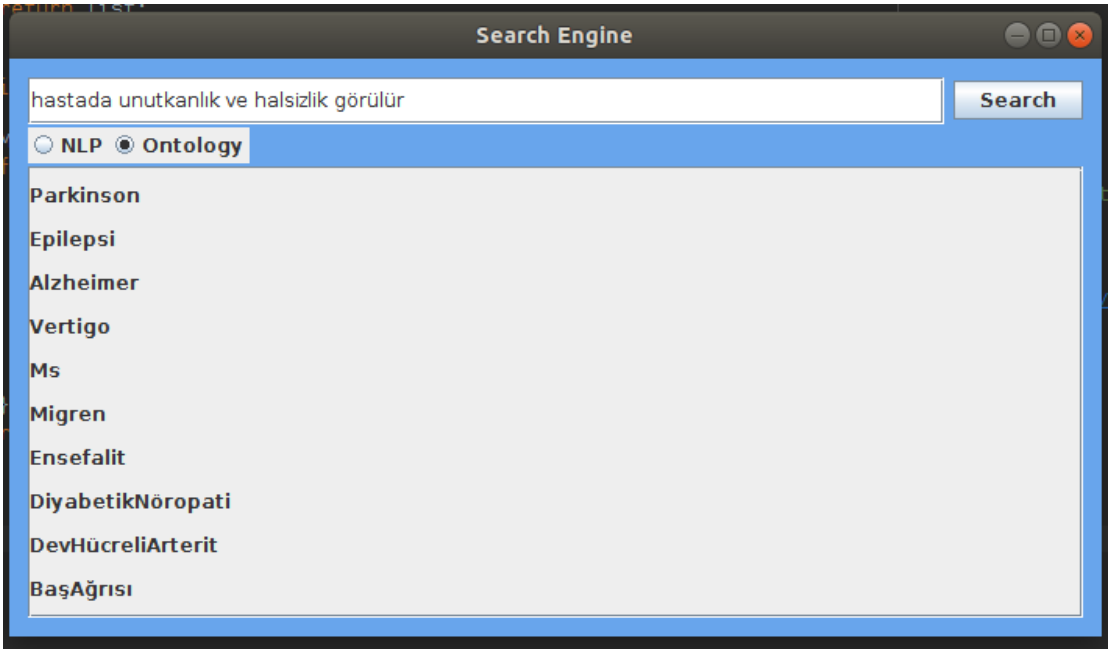
İlk örnekte migren hastalığı ve ona benzer hastalıklar elde edilirken diğer örnekte dikkat eksikliğine belirtisine sahip hastalıklar elde edilir.
Varlık biliminde, kurallar belirleyerek belli şartlar altında sınıflar arasında ilişki kurulması sağlanabilir.

Varlık Bilimi Kural Örneği :
Hastalık(?h1) ^ Hastalık(?h2) ^ differentFrom(?h1,?h2) ^ gösterir(?h1,?b1) ^ gösterir(^h2,^b1) -> benzerdir(?h1,?h2).

Kullanılan bu kural sayesinde iki farklı hastalık arasında eğer benzer belirtiler varsa "benzerdir" ilişkisi kurulması sağlanmıştır.



Şekil 3-2: Oluşturulan Varlık Bilimi Sonucunda Elde Edilen Grafik.



Şekil 3-3: Projede Varlık biliminin Kullanım Örneği

Aramak istenilen cümle, SPARQL sorgusuna çevrilerek varlık bilimi içinde aranır ve istenilen özellik ve ilişkilere sahip olan bireyler kullanıcıya verilir.

4. Literatür Taraması

Doğal dil işleme ve varlık bilimi üzerinde bir çok çalışma bulunmaktadır. Fırat Üniversitesinde, hastaya verilebilecek ilaçların tespiti için varlık biliminden faydalanılan bir proje geliştirilmiştir. Projede ilaçların kullanım yeri, sahip olduğu etken maddeler değerlendirilerek hastaya uygun olup olmadığına karar veren uygulama geliştirilmiştir.

Bu projeye benzer bir proje daha önce Japonya kökenli Fujitsu firması tarafından yapılmıştır. Tıbbi metinleri doğal dil işleme ve varlık bilimi yardımıyda kodlayarak hastalara uygulanan yöntemler incelenmiş ve benzer durumda ki hastalara ne gibi çözümler sunulabileceğine karar veren uygulama geliştirilmiştir.

5. Deneysel Sonuçlar

Doğal dil işleme yönteminin testi için, hastalığın adı, belirtileri ve nedenlerini gösteren bir tablo oluşturuldu. Yazılan test algoritmasında, hastalığın belirtileri verildiğinde, hangi hastalıkların sonuç olarak alındığına bakıldı.

HASTALIK	BELİRTİ_1	BELİRTİ_2	NEDEN_1	NEDEN_2
abarognosis	ağırlık hissi kaybolması		paryetal lob bölgesi hasarlı	
afazi	tekrarlama	konuşma bozukluğu	travma	felç
agnozi	duyu kaybı		felç	
alpers	denge bozukluğu		genetik	
alternan hemipleji	atak geçirir			
alzheimer	unutkanlık	konuşma bozukluğu	genetik	
angelman sendromu	davranışsal bozukluk	denge bozukluğu	genetik	
apraksi	konuşma bozukluğu	davranışsal bozukluk		
baş ağrısı	halsizlik		dehidrasyon	düşük kan şekeri
behçet hastalığı	göz iltihaplanması	deri altı iltihaplanması	genetik	
bell paralizisi	yüz kaslarını kontrol edememe		felç	beyin tümörü

Tablo 5-1: Testte Kullanılan Veri Kümesi

Toplam 30 hastalık üzerinde yapılan testte, bulunan her bir doğru sonuç 0.25 puan olarak kaydedildi. Eğer bir hastalığın tüm belirtileri ve nedenleri doğru sonuç verirse 1.0 puan elde edilir.

	Pencere Boyutu = 3								
Vektör Boyutu	100	250	500	750	1000	1500	2000	2500	3000
PV-DM	8.5	8.5	9.25	8.5	8.25	8.25	8.75	8	8.5
PV-DBOW	14.25	13.5	13	13.75	13.5	13.75	14	14.5	13.75

Tablo 5-2: Doğal Dil İşleme Test Sonuçları, Pencere Boyutu 3

İlk test, 3 boyutlu pencere ile yapıldı. Elde edilen sonuçlara bakıldığında PV-DBOW yöntemi 2500 vektör boyutunda en iyi sonuç olan 14.5 değerini vermektedir. Yüzdelik değer olarak %48.33 ‘lük başarı oranına sahiptir.

PV-DBOW yönteminin daha iyi sonuç vermesinin sebebi, veri kümesinin büyük olması ve bu kümede paragraftan cümle anlamı çıkarırken aynı zamanda kelime vektörlerininde eğitilmesidir. Büyük veri kümelerinde daha fazla örnekleme yapıldığı için daha iyi sonuçlar elde edilir.

	Pencere Boyutu = 6								
Vektör Boyutu	100	250	500	750	1000	1500	2000	2500	3000
PV-DM	9.75	9.5	9.25	9.5	8.75	8.75	9.25	9.5	9.25
PV-DBOW	14	13.75	15.25	13.25	13.75	12.25	14.5	14.5	13.5

Tablo 5-3: Doğal Dil İşleme Test Sonuçları, Pencere Boyutu 6

Pencere boyutu 6'ya çıkarıldığında sonuçlar da biraz daha iyileşme görülür. PV-DBOW yöntemi, PV-DM yöntemine göre daha iyi sonuç vermektedir. Vektör boyutu 500 iken en yüksek değer 15.25 elde edilir. Yüzdelik olarak %50.83 'lük başarı oranına sahiptir.

Türkçe cümlelerde kelime öznesi ve fiili birbirinden çok uzakta olabilir. Özne cümle başındayken, fiil cümle sonundadır. Bu nedenle pencere boyutu yükseltildiğinde, fiil ve cümle arasında ilişki kurulabildiği için sonuçlarda iyileşmeler görülür.

Varlık biliminde, aranılan bilgi eğer varlık bilimine eklenmişse sonuç kesin olarak çıkar. Varlık bilimi, doğal dil işlemenin aksine rastgelelik içermez ve çıkan sonuçlar kurulan varlık biliminin yapısına bağlıdır. Projede varlık bilimi "Hastalık" "Belirtiler" ve "Nedenler" olmak üzere 3 sınıfa ayrılmış ve aralarında "gösterir" "kaynaklanır" ve "benzerdir" ilişkileri kurulmuştur. Kurulan varlık bilimi, 35 hastalık üzerine kurulmuş, nedenleri ve belirtileri yazılmıştır. Bu hastalıklar üzerine yapılan aramalarda kesin sonuçlar elde edilir.

6. Tartışma

Bu çalışmanın sonucunda elde edilen çıkarımlar şu şekildedir.

- Doğal dil işlemede, veri kümesindeki hastalıkları ayrı ayrı işlemektense, hastalıklar ile ilgili tüm makaleleri bir araya getirip paragraflarına ayırarak öğrenim algoritmasında kullanmak birbirleri arasındaki ilişkiyi arttırdığı için daha iyi sonuçlar alındığı gözlemlenmiştir.

- Doğal dil işlemede, Paragraph2Vec yönteminin PV-DBOW yönteminin PV-DM yöntemine göre daha başarılı olduğu gözlemlenmiştir. Bunu sebebi olarak hastalıklar ile ilgili makaleler paragraflara ayrıldığında, ortaya daha geniş bir veri kümesi çıkmasıdır. Büyük veri kümelerinde PV-DBOW yönteminin daha iyi sonuç verdiği, yöntemi anlatan makalede de belirtilmiş olup, sebebinin ise paragraftan cümle anlamı çıkarma, geniş veri kümesinde paragraf vektörünü daha iyi yakınsatması olarak gösterilir. Yöntem aynı zamanda kelime vektörlerini de eğittiği için paragraf içindeki kelime vektörleride anlamlı vektörlere sahip olurlar.

- Varlık biliminde, tutarlı bir yapı oluşturmak için veri kümesinin iyi analiz edilmesi gerekmektedir. Konularını yeteri kadar anlamadan oluşturduğumuz sınıflar, tutarsız bir yapıya evrilir. Bu sebeble, varlık bilimini oluşturmadan önce, makaleler incelenmiş ve hastalıkların ne gibi belirtiler gösterdiği, ne tür nedenlerle meydana geldiği anlaşılmaya çalışılmıştır. Doğal dil işlemede veri kümesinin olabildiğince tutarlı ve istediğimiz bilgilerden oluşması isteriz fakat içeriği hakkında çok fazla bilgi edinmeden de eğitim algoritmasını yazabiliriz. Fakat varlık bilimi oluşturmak için, bu projede sağlık bilgisine de ihtiyaç duyulmuştur.

- Varlık biliminde, veri kümesini oluşturmak zor olsada, elde edilen sonuçlar kesin ve tutarlıdır. Varlık biliminde sınıfları oluşturmak en kritik aşamadır. Sınıfları, oluşturulan veri kümesine ve elde edilmek istenilen bilgiye göre seçersek, aralarındaki ilişkileri ve özellikleri de belirtebiliriz. Bu sayede, oluşturulan varlık biliminde elde edilen sonuçlar kesin yargı bildirir.

- Doğal dil işlemede, kelime vektörleri, ve sınıflandırma algoritmasındaki ağırlık vektörleri rastgele değerlerden seçilir ve eğitim boyunca bu değerler iyileştirimeye çalışılır. Kelimeye ve kelimenin kullanıldığı yere göre vektör son haline alır. Bu durum değişkenlik ortaya çıkarır. Her eğitim denemesi sonucunda farklı bir vektör ortaya çıkar, bu vektörlerde bazıları daha iyi sonuç verirken, diğerleri çok iyi yakınsamayabilir. Varlık biliminde ise bir kesinlik söz konusudur. Değişkenlik göstermez.

- Türkçe cümlelerde özne cümlenin başında fiil de cümlenin sonunda olduğu için, doğal dil işleme eğitim algoritmasında pencere boyutunu 6 gibi yüksek değerler verdiğimizde, daha anlamlı sonuçlar elde ettiğimiz gözlemlenmiştir. Pencere boyutunun yüksek olmasıyla, cümlenin öznesi ve fiili arasında ilişki kurulması sağlanır ve cümle anlamı ortaya çıkar.

KAYNAKLAR

[1] MİKOLOV, Tomas, *Efficient Estimation of Word Representations in Vector Space*, Google Inc, 2013

[2] MİKOLOV, Tomas, *Distributed Representations of Sentences and Documents*, Google Inc, 2014

[3] Lİ, Susan, *Multi-Class Text Classification with Doc2Vec & Logistic Regression*, https://towardsdatascience.com/multi-class-text-classification-with-doc2vec-logistic-regression-9da9947b43f4 , 2018

[4] ALİ, Zafar, *Word2Vec*, https://medium.com/@zafaralibagh6/a-simple-word2vec-tutorial-61e64e38a6a1 , 2019

[5] ÖZTÜRK, Övünç, *ANLAMSAL WEB İÇİN BİR ONTOLOJİ ORTAMI TASARIMI VE GERÇEKLEŞTİRİMİ*, Yüksek Lisans Tezi, Ege Üniversitesi Fen Bilimleri Enstitüsü, 2004

[6] BANSAL, Ritika, *Design and development of semantic web-based system for computer science domain-specific information retrieval*, Perspectives in Science, 2016

[7] Altay, Osman, *Anlamsal Web Kullanılarak İlaç Ontolojisi Çıkarılması*, Yüksek Lisans Tezi, Fırat Üniversitesi Teknoloji Fakültesi Yazılım Mühendisliği, 2018

[8] HORRİDGE, Matthew, *A Practical Guide To Building OWL Ontologies Using Protege 4*, The University Of Manchester, 2011