KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ
ANABİLİM DALI

YÜKSEK LİSANS TEZİ

KİTLE KAYNAK KULLANIMI VE DOĞAL DİL İŞLEME İLE MARKET BAZINDA FİYAT ÖNERİ SİSTEMİ

FULYA ER

KOCAELİ 2023

KOCAELİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ

ANABİLİM DALI

YÜKSEK LİSANS TEZİ

KİTLE KAYNAK KULLANIMI VE DOĞAL DİL İŞLEME İLE MARKET BAZINDA FİYAT ÖNERİ SİSTEMİ

FULYA ER

Prof. Dr. Ahmet SAYAR		
Danışman,	**Kocaeli Üniversitesi**	**.....**
Prof. Dr. Sevinç İLHAN OMURCA		
Jüri Üyesi,	**Kocaeli Üniversitesi**	**.....**
Prof. Dr. Galip AYDIN		
Jüri Üyesi,	**Fırat Üniversitesi**	**.....**
Doç. Dr. Hidayet TAKÇI		
Eş Danışman,	**Cumhuriyet Üniversitesi**	

Enstitü Yönetim Kurulunun 08/12/2021 tarih ve 2021/42 nolu toplantısında II. tez danışmanı olarak atanmıştır.

Tezin Savunulduğu Tarih: 17.01.2023

ETİK BEYAN VE ARAŞTIRMA FONU DESTEĞİ

Kocaeli Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez/proje çalışmasında,

- Bu tezin/projenin bana ait, özgün bir çalışma olduğunu,
- Çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı,
- Bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi,
- Bu çalışmanın Kocaeli Üniversitesi'nin abone olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun olduğunu,
- Kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- Tezin/Projenin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez/proje çalışması olarak sunmadığımı,

beyan ederim.

■ Bu tez/proje çalışmasının herhangi bir aşaması hiçbir kurum/kuruluş tarafından maddi/alt yapı desteği ile desteklenmemiştir.

□ Bu tez/proje çalışması kapsamında üretilen veri ve bilgiler ..... tarafından ..... no'lu proje kapsamında maddi/alt yapı desteği alınarak gerçekleştirilmiştir.

Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

.....

Fulya ER

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI

Fen Bilimleri Enstitüsü tarafından onaylanan lisansüstü tezimin/projemin tamamını veya herhangi bir kısmını, basılı ve elektronik formatta arşivleme ve aşağıda belirtilen koşullarla kullanıma açma izninin Kocaeli Üniversitesi'ne verdiğimi beyan ederim. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet hakları bende kalacak, tezimin/projemin tamamının ya da bir bölümünün gelecekteki makale, kitap, tebliğ, lisans, patent gibi çalışmalarda kullanımı, danışmanımın isim hakkı saklı kalmak koşuluyla ve her iki tarafın bilgisi dâhilinde bana ait olacaktır. Tezin/projenin kendi özgün çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin/projenin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim kurulu tarafından yayınlanan "Lisanüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge" kapsamında tezim aşağıda belirtilen koşullar haricinde YÖK Ulusal Tez Merkezi/ Kocaeli Üniversitesi Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- ☐ Enstitü yönetim kurulu kararı ile tezimin/projemin erişime açılması mezuniyet tarihinden itibaren 2 yıl ertelenmiştir.
- ☐ Enstitü yönetim kurulu gerekçeli kararı ile tezimin/projemin erişime açılması mezuniyet tarihinden itibaren 6 ay ertelenmiştir.
- ■ Tezim/projem ile ilgili gizlilik kararı verilmemiştir.

.....

Fulya ER

ÖNSÖZ VE TEŞEKKÜR

Tez çalışmamda bana yardımcı olan danışman hocam Sayın Prof. Dr. Ahmet SAYAR'a ve eş danışman hocam Sayın Doç.Dr. Hidayet TAKÇI'ya teşekkürü borç bilirim. Eğitim hayatım boyunca yardımlarını ve desteklerini esirgemeyen annem ve babam başta olmak üzere tez çalışmam süresince motivasyonuma katkıda bulunan tüm aileme ve arkadaşlarıma teşekkür ederim.

Ocak – 2023 Fulya ER

İÇİNDEKİLER

ETİK BEYAN VE ARAŞTIRMA FONU DESTEĞİ i
YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI ii
ÖNSÖZ VE TEŞEKKÜR iii
İÇİNDEKİLER iv
ŞEKİLLER DİZİNİ v
TABLOLAR DİZİNİ vi
SİMGELER VE KISALTMALAR DİZİNİ vii
ÖZET viii
ABSTRACT ix
1. GİRİŞ 1
2. DOĞAL DİL İŞLEME 5
2.1. Doğal Dil İşleme Nedir? 5
2.2. Neden Doğal Dil İşleme? 8
2.3. Doğal Dil İşleme: Sorunlar ve Çözüm Önerileri 9
2.4. Doğal Dil İşleme Görevleri 10
2.5. Derlem Oluşturma 11
2.6. Dilbilimsel Çalışmalar 12
2.7. İstatistiksel Çalışmalar 13
2.8. Adlandırılmış Varlık Tanıma 13
2.8.1. Adlandırılmış Varlık Tanıma Nedir? 14
2.8.2. NER Nasıl Çalışır? 15
2.8.3. Kullanım Alanları Nelerdir? 15
2.8.4. Adlandırılmış Varlık Tanımına İlişkin Konular 16
2.8.5. Adlandırılmış Varlık Tanıma Yöntemleri 18
3. İLGİLİ ÇALIŞMALAR 19
4. ÖNERİLEN SİSTEM 25
4.1. Veri Temizleme 28
4.2. Normalizasyon 28
4.3. Kelimelere Ayrıştırma (Tokenizasyon) 29
4.4. Dil Bilimi İşlemleri (Morfoloji) 29
5. DENEYSEL ÇALIŞMALAR 31
6. SONUÇ VE DEĞERLENDİRME 38
KAYNAKLAR 40
KİŞİSEL YAYIN VE ESERLER 44
ÖZGEÇMİŞ 45

ŞEKİLLER DİZİNİ

Şekil 2.1. POS Tagging 12
Şekil 3.1. Kitle Kaynak Kullanım Süreci 20
Şekil 4.1. Önerilen Model 25
Şekil 4.2. Bilgi Çıkarımı Adımları 25
Şekil 4.3. Sözlük Tabanlı Adlandırılmış Varlık Tanıma İşlemi 27
Şekil 4.4. Veri Temizleme için Kod Parçası 28
Şekil 4.5. Normalizasyon Örneği 29
Şekil 5.1. İTÜ Normalizasyon Örneği 31
Şekil 5.2. İTÜ NER 32
Şekil 5.3. Zemberek Kütüphanesi ile Kök Bulma 32
Şekil 5.4. Kök Bulma ve NER İşlemi 33
Şekil 5.5. Stemming Örneği 34
Şekil 5.6. Kelimelerin Etiketleme İşlemi 34
Şekil 5.7. Birden Çok Kelimeli Ürün İsminin, Ürün Olarak Sözlüğe Eklenmesi 35
Şekil 5.8. Birden Çok Kelimeli Ürün İsminin Sözlüğe Eklenmesi 35
Şekil 5.9. Birden Çok Kelimeli Ürün İsminin Etiketlenmesi 36

TABLOLAR DİZİNİ

Tablo 4.1. Verilerin Toplanması.....26

SİMGELER VE KISALTMALAR DİZİNİ

Kısaltmalar

API	: Application Programming Interface (Uygulama Programlama Arabirimi)
BERT	: Bidirectional Encoder Representations for Transformer (Transformatörler için Çift Yönlü Enkoder Gösterimleri)
CONLL	: Conference on Computational Natural Language Learning (Hesaplamalı Doğal Dil Öğrenimi Konferansı)
CRF	: Conditional Random Fields (Koşullu Rastgele Alanlar)
DDİ	: Doğal Dil İşleme
ELMO	: Embeddings from Language Models (Dil Modellerinden Yerleştirmeler)
HMM	: Hidden Markov Model (Saklı Markov Modeli)
IREX	: International Research & Exchanges Board
MUC – 6	: 6. Message Understanding Conference (6. Mesaj Anlama Konferansı)
NER	: Named Entity Recognition (Adlandırılmış Varlık Tanıma)
NLP	: Natural Language Processing (Doğal Dil İşleme)
POS	: Part Of Speech
TÜFE	: Tüketici Fiyat Endeksi
TÜİK	: Türkiye İstatistik Kurumu
VİT	: Varlık İsmi Tanıma
WSJ	: Wall Street Journal

KİTLE KAYNAK KULLANIMI VE DOĞAL DİL İŞLEME İLE MARKET BAZINDA FİYAT ÖNERİ SİSTEMİ

ÖZET

Alışverişlerde insanların fiyat/performans oranı yüksek olan marketleri tercih ettiği bir gerçektir. İnsanlar almak istedikleri bir ürünü hangi mağazadan ne kadar fiyata alabileceklerini öğrenebilecekleri bir sisteme-yazılıma ihtiyaç duymuşlardır.
Bu çalışmada, bahsedilen probleme kısmen de olsa çözüm getirebilecek, kitle kaynak hesaplama modeli ile birlikte Doğal Dil İşleme yöntemlerini kullanarak ElasticSearch kelime tabanlı indexleme için gerekli olan datayı hazır hale getiren bir mimari geliştirilmiştir. Önerilen bu sistem, tüm bireylerin potansiyel bir müşteri olduğu ve marketlerdeki fiyatları düz mesajlarla sisteme besledikleri bir dünya üzerine kurulmuştur. Yapısız veri olan mesajların DDİ ile işlenerek mesajdaki ürün, marka, market, fiyat ve miktar bilgilerinin tespit edilebilmesi sağlanmıştır. Bu bilgiler tespit edilirken kelimelerin köklerinin bulunması için bir algoritma geliştirilmiştir. Geliştirilen bu algoritmayla birlikte, ürün isimleri iki veya daha fazla kelimeden oluşan kayıtlar için tutma oranı uygulanmıştır.
Mevcut durumun aksine yalnızca katalog veya fiyat/performans bilgileri üzerinden indirim yakalama değil aynı zamanda kişilerin birbirine gönderdiği mesajlar içerisinden indirim haberlerini yakalayacak doğal dil işleme prosedürlerinin uygulanması ile ürünlerin hangi markette en uygun olduğu bilgilerini listelemek mümkün olacaktır. Aynı zamanda, topluluklardan gücünü alan Kitle Kaynak Kullanımı ile birlikte, kullanıcı kitlesinin artması ve uygulamanın bu oranda kullanılabilirliğinin artması hedeflenmektedir.
Bu çalışmada market, marka, ürün, fiyat ve miktar bilgilerinin tespit edilmesi için Varlık İsmi Tanıma yöntemi kullanılmıştır. Literatüre katkı olarak, bu konudaki çalışmalara yardımcı olabilecek ve bu çalışmada da kullanılan etiketlerin bulunduğu bir sözlük oluşturulmuştur. Bu sözlük içerisinde yaklaşık olarak 5000 kayıt bulunmaktadır. Kullanıcının sistemi kullanma rolünün yanında besleme görevi de olduğu için, oluşturulan bu sözlüğe yeni bir veri eklenmesi gerektiğinde bunu kullanıcılar sağlayabilecektir. Böylece Kitle kaynak kullanımı ile marketteki ürünlerin fiyat bilgilerinin girildiği veriler üzerinde Adlandırılmış Varlık Tanıma yöntemi uygulanarak etiketsiz ve düzensiz veri etiketli ve düzenli hale getirilmiştir. Ayrıca çalışma bir arayüz de yazılarak sistemin çalışabildiği gösterilmiştir.

Anahtar Kelimeler: Adlandırılmış Varlık Tanıma, Doğal Dil İşleme, Kitle Kaynak Kullanımı.

PRODUCT BASED PRICE SUGGESTION SYSTEM IN THE MARKET BY USING CROWDSOURCING AND NATURAL LANGUAGE PROCESSING

ABSTRACT

It is a fact that people prefer markets with high price/performance ratios in shopping. People needed a system software, where they could learn from which store and at what price they could buy a product.
In this study, an architecture has been developed that can provide a partial solution to the aforementioned problem, and that makes the data necessary for ElasticSearch word-based indexing ready by using Natural Language Processing methods together with the crowdsourcing computing model. This proposed system is built on a world where all individuals are potential customers and feed the prices in the markets to the system with plain messages. Messages, which are unstructured data, are processed with NLP and it is ensured that the product, brand, market, price and quantity information in the message can be determined. While detecting this information, an algorithm has been developed to find the roots of the words. With this developed algorithm, the retention rate has been applied for records whose product names consist of two or more words.
Contrary to the current situation, it will be possible not only to catch discounts through catalogue or price/performance information but also to list information on which market the products are most suitable for, by applying natural language processing procedures that will capture discount news from messages sent by people to each other. At the same time, the use of Crowdsourcing, which is powered by communities, it is aimed to increase the user base and increase the usability of the application at this rate.
In this study, named entity recognition method was used to determine the market, brand, product, price and quantity information. As a contribution to the literature, a dictionary has been created that includes the labels used in this study, which can help studies on this subject, has been created. There are approximately 5000 entries in this dictionary. Since the user has the duty of feeding the system as well as the role of using the system, users will be able to provide this when new data needs to be added to this created dictionary. Thus, unlabeled and irregular data were made labelled and organized by applying the Named Entity Recognition method on the data in which the price information of the products in the market was entered with the use of crowdsourcing. In addition, it has been shown that the system can work by writing a working interface.

Keywords: Named Entity Recognition, Natural Language Processing, Crowdsourcing.

1. GİRİŞ

İnsanlar tarafından bilinçli bir çaba ile değil de görerek, duyarak ve doğuştan gelen özellikler ile öğrenilen, resmi olarak üretilmemiş dillere doğal dil adı verilir. Doğal diller, programlama dilleri gibi kısıtlı ve insan üretimi dillerden farklılık gösterirler. Onlar insan üretimi dillerden daha karmaşık ve daha canlıdır. Bunun bir sonucu olarak da doğal dillerin anlaşılması ve üretimi zorlu bir süreçtir.

Doğal dillerin anlaşılması ve üretimi zorlu görevler olmasına rağmen insanlarla bilgisayarların iletişimi ve insanlar arası iletişimin kolaylaştırılması için doğal dil işleme alanı ortaya çıkmıştır. Doğal dil işleme, ihtiyaca uygun yazılımlar geliştirmeyi hedefleyen, bilgisayar mühendisliği ve yapay zekâ alt alanlarından biridir. Konu sadece teorik değil aynı zamanda deneysel olarak ele alınmaktadır. İnsanoğlunun uzun yıllardır kendi doğal dilini kullanarak makinelerle etkileşim kurma isteği ve bu konuda yaptıkları araştırmalar en heyecan verici araştırmalar arasında yer almaktadır. Makineler ile insanlar arasındaki etkileşimin doğal dil üzerinden yapılmasını sağlamak, doğal dilin anlamlandırılması, diller arasında otomatik çevrimin yapılması, doğal dil ile yazılmış metinlerden bilgi çıkarımı gibi görevler DDI araştırma alanının çalışma konularıdır (Güneş, 2018).

Doğal dil işleme algoritmalarının çeşitli kullanımları vardır. Doğal dil işlemenin gücünden yararlanarak yapılan belge özetleme, başlık oluşturucu, altyazı oluşturucu, sahte evrak algılama, konuşma tanıma ile telefondaki asistana komut verme, Google / Yandex / Microsoft metin çevirileri, metin yazarken bir sonraki kelimenin tahmin edilmesi/önerilmesi, chatbot vb. çok çeşitli teknolojiler bulunmaktadır. Bunlar dünya problemlerini çözen teknolojilerdir.

Doğal dil işleme algoritmaları genel olarak makine öğrenimi algoritmalarına dayanır. Doğal dil işleme, büyük kural setlerini elle kodlamak yerine, bir dizi örneği analiz ederek (yani bir kitap gibi büyük bir bütünden, bir cümle koleksiyonuna kadar) ve istatistiksel bir çıkarım yaparak bu kuralları otomatik olarak öğrenmek için makine öğrenimine güvenebilir. Genellikle, ne kadar çok veri analiz edilirse, model o kadar doğru olur. Bununla birlikte son yılların önemli araştırma konularından olan metin madenciliği de doğal dil işleme dâhilindedir. Metin madenciliği ile milyarlarca internet kullanıcısının ürettiği veriyi işlenip anlamlandırılabilir. Doğal dil araç seti, Python programlama dili kullanılarak geliştirilmiş/geliştirilmekte olan konuşma dili verileriyle çalışan açık kaynaklı bir

kütüphanedir. Kütüphanenin içerdiği modüller veri ön işleme, makine öğrenmesi algoritmalarının uygulanması, Twitter API ile yapılan işlemler gibi birçok işlemde kullanılır.

Doğal dil işleme süreçlerinde, adlandırılmış varlık tanıma önemli bir yere sahiptir. Günümüzde bilgi erişim sistemlerinin çoğunda adlandırılmış varlık tanıma kullanılmaktadır. Denetimli makine öğrenimi sistemlerinin eğitimi için etiketlenmiş veri kümeleri gerekir. Bu kapsamda denetimli öğrenmenin bir örneği olan adlandırılmış varlık tanıma görevi için de bir derlemdeki her kelime etiketlenmelidir. İngilizce için veri etiketleme işlemi düzenli olarak yapılabilse de diğer diller için aynı durum geçerli değildir buna rağmen İngilizce için yapılan çalışmalarda bile belirli alanlarda etiketleme işlemi yapılmaktadır. Gazete makaleleri gibi alanlarda İngilizce için kelime etiketi yapılabiliyorken diğer alanlardaki görevler için etiketlenen kelime sayısı yok denecek kadar azdır. İnsanlar tarafından yapılan veri etiketleme işlemi ise oldukça az, maliyetli, zaman alan ve zor iştir.

Çalışmamız kapsamında diğer önemli konu bilgi çıkarımıdır. Yazılı metin içerisinden bilgi çıkarımı görevi ele alındığında ise metin içerisinde geçen tanımlı varlıkların tespit edilmesi önemli bir görev adımıdır. Metin içerisinde geçen kişi, organizasyon, yer adları, sayısal değerler, tarihsel ifadelerin tespit edilerek işaretlenmesi Adlandırılmış Varlık Tanıma olarak adlandırılmaktadır. NER görevinde tanınacak varlıklar ihtiyaçlara göre değişiklik gösterebileceği gibi NER çalışmalarında genel olarak ENAMEX (kişi, organizasyon ve yer isimleri), NUMEX (sayısal değerler, parasal ifadeler, yüzdeli ifadeler), TIMEX (tarih ve zaman ifadeleri) varlık kategorilerinin işaretlenmesi olarak değerlendirilmektedir. Bununla birlikte özelleşen sistemlerde e-posta adresleri, telefon numaraları, kitap başlıkları, proje isimleri gibi kategoriler de Varlık İsmi olarak ele alınabilmektedir. Adlandırılmış varlık tanıma makine çevirisi, bilgi elde etme, duygu analizi, multimedya indeksleme, soru-cevap sistemleri gibi diğer doğal dil işleme uygulamalarında kullanılabilmektedirler (Eken, 2015). Örnekleri çoğaltmak gerekirse biyoinformatik ve kimya alanlarındaki metinlerden protein isimleri, RNA, DNA, hücre bilgileri, ilaç adları, kimyasal adlarının tespit edilmesi de Adlandırılmış Varlık Tanıma görevi olarak değerlendirilebilmektedir.

Ülkelerin zaman zaman enflasyon değerlerinde artış yaşanması hiç şüphesiz insanların kaygılanmasına, varlığını en sağlıklı ve verimli harcama yapabilmek için nasıl değerlendirilebileceğinin yanıtını aramaya neden olmuştur. Mutluluk, sürdürülen yaşam biçiminden alınan hoşnutluk ve toplumun takdirini kazanacak değerlere sahip olup bunların

farklı oranlarda birleştirilmesi olarak tanımlanabilir. Mutluluk kavramı, insanın bir amaca ulaşıp gereksinimlerini karşılamasıyla gerçekleşir. Eğer bu amaç veya hedefe ulaşılamazsa veya gereksinimler beklentisini karşılamazsa kişi mutsuz olabilir. (Yıldırım & Demirel, 2015) Bu bilgiler de öngörülerek, projenin sağlamayı amaçladığı en büyük hedeflerden biri insan psikolojisini olumlu etkilemek, mutluluğuna katkıda bulunabilmek ve emeğinin karşılığını olabilecek en yüksek seviyede almasını sağlayan bir aracı olabilmektir. Bunların yanında gelişmeye devam etmekte olan Türkçe Doğal Dil İşleme alanına da yenilikler katabilmek, karşılaşabileceğimiz sorunlara farklı bakış açıları ile bir çözüm getirebilmek hedeflenmektedir.

Geçmişte yapılan birçok çalışmada veriler, ülke genelinde en sık kullanılan Migros, Bim, Carrefoursa gibi zincir marketlerin haftalık olarak fırsat ürünlerini sergilediği aktüel kataloglardan alınmıştır. Alışverişlerde tüm bunların bir arada kullanılması bir yere kadar fayda sağlayabilse de yeterli değildir. Aktüel Katalog yayınlayan birçok marketin yanında katalog yayınlamadan uygun fiyatlı ürün satışı yapan marketler de bulunmaktadır. Bir diğer hedef olan Türkçe Doğal Dil İşlemeye değer kazandırmak üzerine yapılan çalışmalarda ise market ve ürün verilerinin kullanıldığı bir çalışma bulunmamaktadır. Adlandırılmış varlık tanıma alanındaki mevcut çalışmalar geliştirmeyi düşündüğümüz çalışmaya katkı sunacak türden değildir. Bu nedenle çalışmamıza özgü bir sözlük oluşturulmuş olup çalışmalarımız bu özgün sözlüğe göre şekillendirilmiştir.

Günümüzde daha da değerli bir hal alan indirim yakalama, indirimi paylaşarak çevresini de faydalandırma işlemlerini kolaylaştırmak ve bu sayede toplum mutluluğunu artırmak, insanların bütçesine katkıda bulunmak, tekelciliği ve haksız rekabeti önlemek mümkündür. Alışverişlerde insanların fiyat/performans oranı yüksek olan marketleri tercih ettiği bir gerçektir. Bu çalışmada, mevcut durumun aksine yalnızca katalog veya fiyat/performans bilgileri üzerinden indirim yakalama değil aynı zamanda kişilerin birbirine gönderdiği mesajlar içerisinden indirim haberlerini yakalayacak doğal dil işleme prosedürleri uygulanacaktır. Böylece ürünlerin hangi markette en uygun olduğu bilgilerini listelemek mümkün olacaktır. Aynı zamanda, topluluklardan gücünü alan kitle kaynak kullanımı ile birlikte, kullanıcı kitlesinin artması ve uygulamanın bu oranda kullanılabilirliğinin artması hedeflenmektedir.

Bu çalışmada market, marka, ürün, fiyat ve miktar bilgilerinin tespit edilmesi için Adlandırılmış Varlık Tanıma yöntemi kullanılmıştır. İşlemlere yardımcı olması amacıyla market, marka, ürün, fiyat ve miktar etiketlerinin bulunduğu bir sözlük oluşturulmuştur. Bu sözlük içerisinde yaklaşık olarak 5000 kayıt bulunmaktadır. Kitle kaynak kullanımı ile marketteki ürünlerin fiyat bilgilerinin girildiği veriler üzerinde NER yöntemi uygulanarak etiketsiz ve düzensiz veri etiketli ve düzenli hale getirilmiştir. Çalışmamızda veri işleme adımına ilk olarak Normalleştirme ile başlanmıştır. Bu aşamada verilerdeki düzensizlik giderilip (hatalı sembol kullanımı, eksik sesli harf kullanımı, fazla harf kullanımı) bir sonraki aşama olan Tokenize Etme, Gövdeleme ve Baş Sözcük Çıkarma işlemlerine geçilmiştir. Tokenize işlemine tabi tutulan verilerin gerekli kütüphaneler kullanılarak gövde ve baş sözcük elde edilmesi sağlanmıştır. Veri işleme adımında son aşama olarak metinden bilgi çıkarımı yapma yöntemlerinden biri olan NER yapılmıştır. İşlenen veriler bu aşamada kendi oluşturduğumuz sözlükteki anahtar kelimeler ile etiketlenmiştir. Kullanıcının girmiş olduğu yorum cümlesindeki ürün ismi birden fazla kelimeden oluşuyorsa, buradaki ürün ismini içeren kelimelerin her biri kendi içinde de bir ürün ismini temsil ediyor mu diye kontrol edilir. Eğer her kelimesi sözlükte yer alırsa bu kelimelerin birleşimi de tek bir ürün ismi olabilir diye bir tutma oranı uygulanıyor. Bu tutma oranı, doğru tespit edilmiş varlık sayısının tespit edilmiş varlık sayısına oranı olarak hesaplanıyor. Böylece, üç kelimeden oluşan bir ürün isminin her kelimesi sözlükte ürün ismi olarak yer alıyorsa o üç kelime ayrı ayrı ürün olarak etiketlenmeyip, tek bir ürün ismi olarak etiketlenmiş olacaktır. Bu sayede kullanıcılar hangi ürün hangi markette en uygun fiyata satılıyor bilgisini görüntüleyebilmiştir. Böylece bir veritabanı oluşturularak elasticsearch ile en uygun fiyatlı ürün fiyat öneri sistemi geliştirilmesine olanak sağlanmıştır.

2. DOĞAL DİL İŞLEME

Doğal dil işleme, yapay zekânın bir alt dalıdır ve bu tez çalışmasının temelini oluşturmaktadır. Amacı, insanın konuşma ve anlama yeteneğini taklit etmektir. Bu kapsamda dillerin üretimi ve anlaşılmasına yönelik çalışmalar yapılmaktadır. Doğal dil işleme çalışmaları sadece yazılı değil aynı zamanda sözlü dille ilgili çalışmaları da kapsamaktadır.

2.1. Doğal Dil İşleme Nedir?

Doğal Dil İşleme, bir sistem aracılığıyla doğal diller üzerinden makine ve insanların birbiriyle haberleşmesini sağlayan bir araştırma alanıdır. Bu alanda araştırma yapanlar genelde, insanların dili nasıl kullandığı ve anladığıyla ilgili bilgi toplamayı amaçlamaktadır. Böylece istenen görevleri yerine getirip doğal dilleri anlamasını bilgisayar sistemleri sağlamış olur (Chowdhury, 2003).

Doğal dil işlemeyi başka bir şekilde ifade edecek olursak, insan dilini hesaplamalı olarak temsil etmek ve bu dilleri analiz etmek için kullanılan bir sistem olduğu ifade edilmektedir (Khurana ve diğ., 2017).

Doğal dil işleme çalışmaları yöntem açısından dilbilimsel ve istatistiksel dil işleme şeklinde iki alt alana ayrılır. Dilbilimsel çalışmaların hedefi dilin asırlar boyu gelişimini takip etmek, gramerini açığa çıkarmak, söz varlığını anlamak ve kullanım farklarını açıklamaktır. Dilbilimsel çalışma konu uzmanı olmayı gerektirir. Dünyada binlerce dil olduğunu düşünecek olursak bütün dillerle topluca mücadele etmek dilbilimsel çalışma yapanlar açısından mümkün değildir. Dilbilimciler bir veya en fazla birkaç dilde uzmanlık sahibi olabilir. Bununla birlikte başta Google olmak üzere dil uygulamalarında birçok dilin kapsandığı bilinen bir gerçektir. Dilin gramer ve morfolojiye dayalı analizine ek olarak istatistiksel yöntemlerle de çalışma yapılır. Dil anlamında başarılı uygulamalar işte bu çalışmaların ürünüdür. Bu nedenle özellikle yapay zekâ teknikleri yardımıyla yapılan çalışmalar istatistiksel doğal dil işleme olarak bilinir.

Dil çalışmalarının olduğu her yerde sözcüksel analiz, sözdizimsel analiz, anlamsal analiz gibi alt çalışma alanları bulunur. Sözcüksel analiz metnin atomlarına ayrılması bunları kökeniyle

ilgili bilgilerin elde edilmesiyle ilgilidir. Çalışma sözcük seviyesinde ele alındığı için bu adı alır. Bir metin bütünlüğü yerine sözcük seviyesinde bilgi çıkarımı hedeflenmiştir.
Sözdizimsel analiz kimi yerde gramer analizi olarak da bilinir. Sözdizimsel analizin amacı bir metnin dilin kurallarına uygun olup olmadığının tespit edilmesidir. Dilin matematiği ya da dilin formüle edilmesini ifade eder. Dilbilimsel çalışmalar kapsamında dillere ait sözdizim ağaçlarına rastlamak mümkündür.

Semantik analiz sözdizimsel analizden farklıdır. Amacı bir dilde üretilen içeriğin anlamlı olup olmamasını tespit etmektir. Örneğin bir programlama dili yardımıyla dil üretimi yapılırken üretilen içeriğin yazım kurallarına uygun olup olmaması sözdizimsel analizi ilgilendirir. Genellikle derleyiciler tarafından yakalanan hatalar sözdizimsel hatalardır. Komut sonunda noktalı virgül kullanılmaması veya blok oluşturma için ihtiyaç duyulan süslü ayraçların hatalı kullanımı sözdizimsel hata örnekleridir. Bununla birlikte bir koşul ifadesinde koşulun hatalı şekilde kontrol edilmesi veya bir döngü ifadesinde döngü komutlarının hatalı kullanımı nedeniyle istenen sonucun elde edilmemesi, ya da bir işlem yaparken işlem başında toplam değeri tutacak değişkenlere ilk değer verilmemesi semantik hatalar için örnek olarak verilebilir. Doğal dil açısından bakıldığında da düzgün şekilde yapılan konuşmaların sözdizim açısından sorunsuzluğu, içerik açısından da konuşulan sözlerin anlamlı olması semantik analiz ile ilişkilidir. Bir metinde yazım hatalarının tespiti dilin sözdizimsel analizi ile ilgiliyken metnin ana fikrinin ne olduğu semantik analiz ile ilgilidir.

Doğal dil işleme çalışmaları sadece yazılı diller için değil aynı zamanda sözlü diller için de yapılmaktadır. Sözlü dil ile çalışma yaparken sözlü dilin atomları olan fonemler karşımıza çıkar. Fonem, bir sesin çıkarılması ve o ses çıktığında meydana gelen sinyal ile ilgilidir. Dolayısıyla sözlü dillerin analiz edilmesinde çıkan seslerin sinyal işleme yöntemleriyle analizi önem kazanmaktadır. Yazı dilinde ise en temel bileşenler harflerdir. Yazılı dilin bileşeni olan harfler ve karakterlerin bir araya gelmesiyle heceler ve hecelerin bir araya gelmesiyle sözcükler meydana gelir. Sözcüksel analizin yapı birimi olan sözcükler bazen yalın halde bazen de ek almış halde bulunabilirler. Sözcüksel analiz yapabilmek için sözcüklerin eklerinin çıkarılmasına ihtiyaç vardır. O nedenle doğal dil işleme çalışmalarının önemli konularından birisi köke ulaşmak ile ilgilidir. Çünkü ancak köke ulaştıktan sonra doğru sözcüksel analiz yapma imkânı bulunabilecektir. Dilin temel ögeleri olan harfler,

heceler, kelimeler, onlardan meydana gelen kısa kelimeler, karakter veya kelime seviyeli n-gramlar özellikle istatistiksel doğal dil işleme için oldukça önemli bilgiler sunarlar. İstatistiksel doğal dil işlemede bu ögeler sayılır, sıraya dizilir, karşılaştırılır, işlem yapılır ve kelime çantası veya vektör uzayı modeli gibi sunum modelleri ile sunulurlar.

Doğal dil işleme çalışmaları bugün belki de en fazla ilerleme kaydeden yapay zekâ alanıdır. Çağrı merkezlerinde insanların yerini yavaş yavaş almaya başlayan chatbot isimli yapay zekâ uygulamaları yine doğal dil işlemenin kanıksadığımız uygulamalarından biridir. Hatta son dönemde hemen hepimizin kullanmaya başladığı chatgpt isimli chatbot bu alanda gelinen aşamayı göstermektedir. OpenAI firmasının geliştirdiği yapay zekâ robotu olan chatgpt sorulan hemen her türden soruya oldukça makul cevaplar vermekte ve insanı şaşırtan bir hızda bu işlemi yapmaktadır. Doğal dil işleme çalışmaları bu başarılarını derin öğrenme algoritmaları ve büyük miktarlardaki etiketli veriye borçludur.

Doğal dil işleme çalışmaları deyince akla gelen bir diğer uygulama makine çevirisi uygulamalarıdır. Google tarafından geliştirilen çeviri uygulaması yüksek doğruluklarda ve çok sayıda dil arasında çeviri hizmetini yerine getirebilmektedir. Sadece Google değil Microsoft tarafından geliştirilen uygulama da bu konuda oldukça başarılıdır.

Doğal dil işleme çalışmalarına başarıyı getiren şey insan dilinde yer alan modellerin anlaşılması ve taklit edilmesidir. İnsan dilindeki örüntülerin anlaşılmasında da bu alandaki yeterli miktarda verinin bulunması yer almaktadır. Özellikle Web 2.0 sonrası kullanıcıların kullandıkları sistemlere veri sunarak onları zenginleştiriyor olması ihtiyaç duyulan veri için en önemli kaynak olmuştur. Her ne kadar dillere has bazı problemler olsa da çalışmalar her geçen gün bir önceki günden daha iyi duruma gelmektedir.

Doğal dil işleme, insanlığa fayda sağlayacak birçok konu üzerinde çalışma sağlamıştır. Bu konular aşağıdaki gibi sıralanmaktadır (Adalı, 2016).

- Yazım yardımcı araçlarının geliştirilmesi
- Bir metnin özetini çıkarma
- Yabancı dilde yazma yardımcı araçları
- Bilgiye erişim

- Bilgisayarla sesli etkileşim
- Yabancı dil okuma yardımcı araçları
- Soru yanıt dizgeleri
- Metni anlama
- Doğal diller arası çeviri
- Bilgisayarın konuşması (metni seslendirme)
- Basılı bir metni okuma ve okuma yanlışlarını düzeltme
- Metnin içerdiği bilgiyi çıkarma
- Yazım yanlışlarının düzeltilmesi
- Bul ve değiştir
- Konuşmayı anlama (konuşmayı metne dönüştürme)

2.2. Neden Doğal Dil İşleme?

Doğal dil işlemenin en önemli iki görevi dil üretimi ve dilin anlaşılmasıdır. Dil üretimi metinden sese dönüşüm uygulamalarında sıklıkla ihtiyaç duyduğumuz bir çalışma alanıdır. Daha ileri bir uygulama olarak bir dilden diğerine çeviri de bir yönüyle dil üretiminin konusudur. Son dönemde özellikle mobil telefonlarda ve çağrı merkezlerinde sıklıkla kullanılan SIRI teknolojisi dil üretimi konusuyla ilgilidir. Dilin anlaşılması ise yazılı veya sözlü dilin ne anlattığını çözümlemek için ortaya konan çalışmalarla ilgilidir. Dolayısıyla doğal dil işleme çalışmasında öne çıkan dil üretimi ve dili anlamak için doğal dil işleme çalışmalarına ihtiyaç vardır. Her geçen gün artan kullanılabilir verinin büyük çoğunluğu metin formatındadır. Daha önceleri kütüphaneciler tarafından büyük bir özveri ile sınıflandırılan metinler artık insan gücüyle sınıflandırılamayacak kadar fazladır. Dokümanların çoğalması bir yandan dezavantaj gibi görünmekle birlikte diğer taraftan bir avantajı içinde barındırır. O avantaj da artan dokümanların işlenmesinin veriden daha fazla değer elde etmeye yardımcı olacak olmasıdır. Çalışmanın başında dokümanların otomatik yöntemlerle özetlenmesi ve özetler yardımıyla kategorilere atanmasına ihtiyaç vardır. Daha sonrasında dokümanlardan bilgi çıkarımı ve diğer işlerin yerine getirilmesi bir ihtiyaçtır. Bu ihtiyaçlar nedeniyle doğal dil işleme çalışmalarının önemi artmıştır. Yapay zekânın en önemli amaçlarından birisi insan gibi davranan sistemler geliştirmek olup bu sistemlerin önemli bir bileşeni de hiç şüphesiz insan gibi iletişime geçen sistemler oluşturmaktır. İnsanoğlunun dediğini anlayan ya da dediği insanlar tarafından anlaşılan sistemler oluşturabilmek için de doğal dil işleme çalışmalarına ihtiyaç vardır.

Farklı dillerde artan kaynakların ana dilde değerlendirilebilmesi için makine çevirisine ihtiyaç vardır. Aynı zamanda canlı çeviri için de gerçek zamanlı çeviri bir ihtiyaçtır. Bu türden ihtiyaçlar için de doğal dil işleme imdada koşacaktır.

Yukarıda sayabildiğimiz ve sayamadığımız birçok fayda için doğal işleme çalışmalarına ihtiyaç vardır. Doğal dil işleme çalışmaları olmasaydı bugün kullanmakta olduğumuz birçok hizmetten söz etmek mümkün olmayacaktı.

2.3. Doğal Dil İşleme: Sorunlar ve Çözüm Önerileri

Doğal dil işleme çalışmaları genellikle zorlu ve karmaşıktır. Bu durum doğal dillerin canlı, muğlak ve belirsiz olması ile ilgilidir. Detaylarına baktığımızda zorluğun sebepleri şöyledir:

- Doğal dil işleme, her dilin kendine has grameri olması nedeniyle genellikle zorlu bir problemdir. Gramer yardımıyla doğal dil işleme yapabilmek için dünyadaki bütün diller için ayrı yöntemlere ihtiyaç vardır. Birbirine benzeyen diller bile gramer açısından farklılıklar içermekte ve analiz için farklı yöntemlere ihtiyaç duymaktadır.
- Doğal dil işleme çalışmalarını zorlaştıran tek şey gramer kaynaklı değildir. Her dilde yer alan dile ait birimler de oldukça fazla boyutludur. Dillerin sözcükleri, dilin yardımcı fiilleri, edat, zamir, bağlaç ve diğerleri dili sözcük seviyesinde de karmaşık bir hale getirmiştir.
- Diller belirsizlik içerdiği için onları modellemek zordur. Programlama dilleri gibi diller sabit bir yapıya sahip olduğu için onları üretmek ve anlamak kolay iken doğal diller canlı organizmalar olduğu için onların üretimi ve anlaşılması zordur. Özellikle de Türkçe gibi sondan eklemeli dillerde yeni kelimelerin üretimi oldukça kolay fakat onların anlaşılması oldukça zor olabilmektedir.
- Dillerin kurallara uygun olarak kullanımı onların anlaşılmasını nispeten kolaylaştırmakla birlikte onların kuralsız kullanımı onun anlaşılmasını daha da zorlaştırmaktadır. Örneğin, devrik cümleler veya yarım bırakılmış cümleler konuyu zorlaştırmaktadır. Başta sosyal medya olmak üzere eksik ve gürültülü veri içeren çevrimiçi metinler bu problemi büyütmektedir.

Ortaya konulan problemlerin çözümü istatistiksel ve dilbilimsel özelliklerin bir arada kullanılmasıdır. Böylece dilbilimsel yöntemlerin ve istatistiksel yöntemlerin birbirini desteklemesi ile doğal dil işleme daha uygulanabilir bir çalışma alanına dönüşecektir. Ayrıca

doğal dil işleme çalışmalarına makine öğrenme yöntemlerin kullanımı bir çözüm sunacaktır. Bu sayede girdilerin ve çıktıların bilindiği bir model yardımıyla dildeki örüntülerin öğrenilmesi imkânı olacaktır. Örneğin, çeviri sistemlerinin son dönemlerde geldiği başarılı nokta makine öğrenme sistemleri yardımıyla bol miktarda eğitim verisinden öğrenme ile gerçekleştirilmiştir. Elimizde bulunan kaynak dildeki metin örnekleri ve onların hedef dildeki karşılıkları yer aldığı için kaynak dilde bir metin parçası görüldüğünde otomatik olarak bunun karşılığı olan metin parçası üretilebilmektedir. Bu konudaki bir başka çözüm önerisi ise veri madenciliği gibi yöntemler yardımıyla dilde yer alan sıralar öğrenilebilir ve bu sıraya uygun olarak çıkarım yapılabilir. Naive yaklaşımlar, gizli markov modelleri ve n-gram modelleri bu türden çalışma yaparken karşımıza çıkan yöntemler olacaktır.

2.4. Doğal Dil İşleme Görevleri

Doğal dil işleme kapsamında yerine getirilen çok sayıda göre bulunur. Özetleme, insan bilgisayar etkileşimi, makine çevirisi bunlardan sadece birkaç tanesidir. Benzer şekilde; dil tanıma, soru cevaplama, yazar tanıma, duygu analizi gibi çeşitli görevler de bulunmaktadır.

Özetleme görevi orijinal metnin anlamını kaybetmeden daha kısa şekilde ifade edilmesine dayanır. Bir metni özetlerken dile yardımcı ögeler, dili destekleyen yan cümleler ve diğerleri elimine edilebilir fakat metinde yer alan insanlar, yerler ve diğer önemli unsurlar tespit edilebilmelidir. Hata ile metindeki önemli kelimelerin metinden silinmemesi için insanların, yerlerin ve diğerlerinin bulunmasına ihtiyaç vardır. İsimlendirilmiş varlık tespiti bu noktada önem kazanan bir çalışma olarak ortaya çıkar. Ayrıca, sözcük seviyesinde kelimelerin türü tespit edildikten sonra kelimeler arasındaki ilişkilerin tespiti de bir ihtiyaç olarak karşımıza çıkacaktır.

Varlıklar ve bunların ilişkisinin bir sonucu olarak meydana gelen olayların tanımlanması da bu alandaki bir başka problemdir. İnsan bilgisayar etkileşimi de bir yönüyle doğal dil işleme görevidir. Doğal dil işleme açısından insan bilgisayar etkileşimi insanlar ile bilgisayarların doğal dilde etkileşimine izin veren bir çalışma alanıdır. İnsan bilgisayar etkileşiminin sorunsuz olabilmesi için anlamın modellenmesine ihtiyaç vardır. Ayrıca, iletişimin eksiksiz olması için alıcı ve gönderici arasında eksik bilgi kalmamalıdır. İnsan bilgisayar etkileşiminin tam anlamıyla gerçek olabilmesi için sesli dilin anlaşılması da yine önemli bir problemdir.

Makine çevrisi kaynak dilde yazılmış metnin hedef dile çevrimini ifade eder. Makine çevirisi yapabilmek için kelime haritalama ve gramer haritalama gibi görevlerin yerine getirilmesine ihtiyaç vardır. Bir makine çeviri sisteminin başarısı için en önemli ölçüt ise çıktı dilin anlaşılır ve okunabilir olmasıdır.

Doğal dil işleme ile ilgili önemli bir çalışma konusu ise duygu analizi ya da görüş madenciliğidir. İnsanlar ürünler, hizmetler, organizasyonlar veya olaylar hakkındaki görüşlerini genellikle sosyal medya veya diğer ortamlarda ifade ederler. Kişilerin bir konu hakkındaki düşüncelerini anlayabilmek için onların yazdıkları mesajları analiz etmek genellikle yeterlidir. Özellikle hizmet sunanlar sundukların hizmetin kalitesi hakkında geri bildirim almak için görüş madenciliğinden yararlanırlar. Görüş madenciliğinin tabanında bir metnin olumlu, olumsuz veya tarafsız şekilde sınıflandırılması ile ilgilidir. Dolayısıyla eğer bir ürün hakkında yazılan yorumun sınıfı olumsuz ise ürünün beğenilmediği ortaya çıkmaktadır. Bir mesajın çıktısı kadar mesajın ne ile ilgili olduğu da önemli bir konudur. Dolayısıyla görüş madenciliği alanında son dönemde hedef odaklı duygu analizi de yapılmaya başlanmıştır.

Bu alandaki önemli çalışmalardan birisi de dil tanımadır. Doğal dil işleme çalışmaları dilin özelliklerini bilmeye ihtiyaç duyduğu için analizin başında metnin hangi dilde olduğunu bilmeye ihtiyaç vardır. Üzerinde çalışma yapılacak dilin hangi dil olduğunu bulma işlemine dil tanıma adı verilir. Dil tanıma, özelliği nedeniyle doğal dil işlemenin giriş kapısı niteliğindedir. Dil tanıma çalışmalarında; kelime seviyeli, karakter seviyeli çok sayıda çalışma bulunmaktadır. Bu alanda özellikle n-gram analizine dayalı çalışmalar yoğun olarak tercih edilmiştir.

2.5. Derlem Oluşturma

Doğal dil işleme çalışmalarında makine öğrenmesi yöntemlerinin kullanılabilmesi için yeterli miktarda veriye ihtiyaç vardır. Bu veri metin verisidir ve sıklıkla dokümanlar şeklinde organize edilir. Doğal dil işleme için kullanılacak dokümanların organize edildiği veri topluluğuna ise derlem adı verilir. Derlem, dokümanlar yığını olmayıp dokümanların üst veri yardımıyla organize edildiği veri topluluğudur.

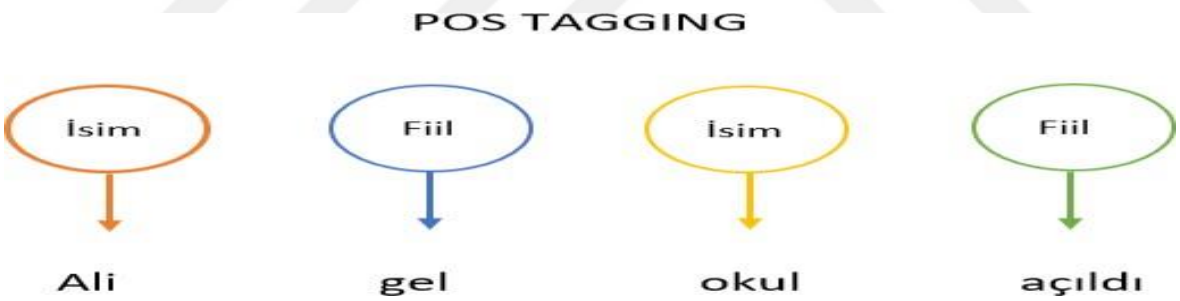
Her derlem bir amaca göre oluşturulur. Eğer yapmak istediğimiz çalışma Türkçe'den İngilizce'ye çeviri sistemi geliştirmek ise derlemde bulunan dokümanlar Türkçe dokümanlar

ve onların İngilizce karşılıkları şeklinde olmalıdır. Veya yazar tanıma çalışması yapıyorsak her bir doküman yazarlarına göre kategorilerine ayrılmalı ve her bir kategoride yazarların örnek dokümanları yer almalıdır. Yazarlara göre ayrılmış dokümanlar ileride yazar modeli oluşturmak ve yazarları test etmek için kullanılacaktır.

Derlem kimi zaman da ham dokümanların belli ön işlemden geçirilmiş sürümlerini de tutar. Bu sayede her doğal dil işleme çalışmasında yapılmakta olan bazı işlemler otomatik olarak yerine getirilmiş olacaktır. Ön işlemden geçirilmiş doküman bilgileri tutan derlemler üzerinde doğrudan istatistiksel ve makine öğrenme yöntemleri kullanılabilir.

2.6. Dilbilimsel Çalışmalar

Doğal dil işleme çalışmalarında dilbilimsel çalışmalar sözcüksel analiz, söz dizimsel analiz, anlamsal analiz, söylem entegrasyonu ve faydacı analiz olarak bilinir. Sözcük analizi metinlerin paragraflara, cümlelere ve kelimeler bölünmesi ile ilgilidir. Her kelime lexicon adı verilen özel bir sözlükte tutulur ve lexicon tabanlı çalışma yapılır. Sözcük analizi bir metinde yer alan kelimelerin türünün işaretlenmesi ile ilgilidir. Bir metinde geçen kelimelerin türü POS etiketleme yardımıyla yerine getirilir (Şekil 2.1).



Şekil 2.1. POS Tagging

Sözdizimsel analiz kelime seviyesinde değil cümle seviyesinde çalışır. Amacı kelimeler arası ilişkileri açığa çıkarmaktır. Sözdizimsel analiz konusunda öne çıkan konulardan birisi dili oluşturan ögelerin ayrıştırılmasıdır. Ayrıştırma bize dili anlama fırsatı sunar. Dilin üretimi aşamasında nasıl bir çalışma yapıldığı dilin ayrıştırılması ile ortaya konar. Bu bakımdan parsing önemlidir. Ayrıca ayrıştırmaların çıktısı ayrıştırma ağacı olarak sunulur. Bu konuda WSJ tree-bank gibi örnek çalışmalar bulunur. Ayrıca, söz dizimsel analiz gramer analizi olarak da literatürde yer almaktadır. Dilin formülleştirilmesi ile ilgili olup gramer için dilin matematiği demek mümkündür.

2.7. İstatistiksel Çalışmalar

Doğal dil işlemede istatistiksel ve makine öğrenmesi tabanlı çalışmalar istatistiksel doğal dil işleme başlığı altında ele alınır. İstatistiksel doğal dil işleme dilin içeriğini bilmeden dil üzerinde işleme imkânı verir. Konu dilin modellenmesi ile ilgilidir. İstatistiksel dil modelleme sıklıkla bir dilin neye benzediğine dair olasılıklar bir model oluşturmaya odaklanır. İstatistiksel dil modelleme ile ilgili belki en iyi örnek metin tahmini veya sıradaki kelimenin ne olacağının bulunmasıdır.

İstatistiksel dil modelleme konusunda karşımıza bayesian yaklaşımlar ve markov yaklaşımı çıkmaktadır. Bayesian yaklaşımları maksimum olabilirlik yöntemindeki sorunları ortadan kaldırmak için dilin bağımsızlığı varsayımı ile olasılık bulmaya odaklanmıştır. Markov yaklaşımı ise geçerli kelimenin tahmini için ondan önceki n-1 kelimenin bilgisi yerine k<n olmak üzere k kelimenin bilgisine dayalı hesaplama yapar. Bayesian yaklaşımının zorluğunu ortadan kaldırmak üzere markov yaklaşımına benzer bir mantık kullanan n-gram modelleri de bu kapsamda karşımıza çıkan bir başka modeldir.

Makine öğrenme algoritmaları da istatistiksel yöntemler başlığında karşımıza çıkmaktadır. Makine öğrenmesi yardımıyla yapılan çalışmalar sıklıkla metin madenciliği başlığı altında değerlendirilir. Metin madenciliği bir yönüyle doğal dil işleme diğer yönüyle veri madenciliği uygulamasıdır. Doğal dil işleme özellik mühendisliği tarafı ile ilgili işlere odaklanmışken veri madenciliği sayısal hale gelen verilere algoritma uygulanması ile ilgilidir.

Bugüne kadar; dil tanıma, metin sınıflandırma, metin özetleme, duygu analizi, spam tespiti gibi birçok çalışma yerine getirilmiştir. Makine öğrenme yöntemleri bayesian yaklaşımı ve benzer yaklaşımların tahmin amacıyla kullanımı olarak da karşımıza çıkmaktadır.

2.8. Adlandırılmış Varlık Tanıma

Adlandırılmış varlık tanıma veya diğer adıyla adlandırılmış varlık tanıma doğal dil işlemedeki Bilgi Çıkarma konusunun bir alt alanıdır. Bilgi çıkarma, yapısal olmayan bir metin içerisinde yer alan varlıkları bulmanın yanı sıra onlar arasında ilişki kurmak, onları sınıflandırma ve depolamak ilgili bir çalışma alanıdır. Anlamsal olarak geliştirilmiş bilgi çıkarma, yapısal olmayan metinlerden çıkarılan varlıkları anlamları ve bağlantılarıyla bulmaya yarar. Bilgi çıkarma, metinsel kaynaklardan belirli bilgilerin çıkarılması işlemidir.

Bu konudaki en basit örneklerden birisi elektronik postalardaki içeriğe dayalı olarak mesajları takvime eklemek olabilir. Benzer şekilde bir elektronik posta içeriğinin analiz edilmesi ile mesaj rotalama görevi yerine getirilebilir. Metinlerden çıkarılan kavramlara veya bilgilere üst veriler eklenerek içerik yönetimi ve bilgi keşfi gibi görevler yerine getirilir.

2.8.1. Adlandırılmış Varlık Tanıma Nedir?

Adlandırılmış varlık tanıma (NER), genel olarak varlık tanımlama veya varlık çıkarma olarak da bilinir. Metindeki belirli varlıkları bulmayı ve kategorilere ayırmayı amaçlayan bir bilgi çıkarma alt görevidir. "Adlandırılmış Varlık" ifadesi, 1996 yılında, yapılan MUC- 6 ortaya çıkmıştır. Dil alanında, Adlandırılmış Varlık Tanıma, terim normalizasyonu ve kategorilere göre sınıflandırma için "varlıkları" bulmak için yapılandırılmamış metin aracılığıyla otomatik taramayı içerir, örneğin kişi adları, kuruluşlar (şirketler, devlet kuruluşları, komiteler gibi), konumlar (şehirler, ülkeler, nehirler gibi) veya tarih ve saat ifadeleri.

NER, temel olarak belirli veri türlerinin (kişi, konum, kuruluş adı, tarih-saat ifadeleri) tanımlanması ve kategorize edilmesi olarak tanımlanabilir. Adlandırılmış varlık tanıma, makine çevirisi, duyarlılık analizi ve bilgi çıkarma gibi çeşitli doğal dil işleme görevleri için önemli bir aşamadır.

NER - bazen varlık öbekleme, ayıklama veya tanımlama olarak da adlandırılır - metindeki önemli bilgileri tanımlama ve kategorilere ayırma görevidir. Bir varlık, tutarlı bir şekilde aynı şeye atıfta bulunan herhangi bir kelime veya kelime dizisi olabilir. Tespit edilen her varlık önceden belirlenmiş bir kategoride sınıflandırılır. Örneğin, bir NER makine öğrenimi modeli, bir metindeki "NASA" kelimesini algılayabilir ve onu "Şirket" olarak sınıflandırabilir.

MUC (Sundheim, 1995) ve CoNLL (Tjong Kim Sang, 2002) konferansları, adlandırılmış varlıkların üç temel türünü tanımlar. Bunlar;

- ENAMEX (kişi, yer ve kuruluş adları)
- TIMEX (tarih ve saat varlıkları)
- NUMEX (para ve yüzde gibi sayısal ifadeler)

Ancak adlandırılmış varlık tanıma araştırması sadece bu türlerle sınırlı değildir. Farklı uygulama alanları, protein adları, ilaç adları, kitap adları gibi alternatif varlık türlerinin belirlenmesinde de önemli role sahiptir.

2.8.2. NER Nasıl Çalışır?

Herhangi bir NER modelinin kalbinde iki adımlı bir süreç vardır:

- Adlandırılmış bir varlığı algıla
- Varlığı kategorilere ayır

Birinci adım, bir varlığı oluşturan bir kelimeyi veya kelime dizisini tespit etmeyi içerir. Her kelime bir simgeyi temsil eder: "Büyük Göller" bir varlığı temsil eden iki simgeden oluşan bir dizidir. İç-dış-başlangıç etiketleme, varlıkların nerede başlayıp nerede bittiğini göstermenin yaygın bir yoludur. İkinci adım, varlık kategorilerinin oluşturulmasını gerektirir. İşte bazı yaygın varlık kategorileri:

Person (kişi)

- Örneğin, Elvis Presley, Audrey Hepburn, David Beckham Organization (organizasyon)
- Örneğin, Amazon, Mastercard, Sivas Cumhuriyet Üniversitesi, Kocaeli Üniversitesi

Time (zaman)

- Örneğin, 2006, 16:34, 2am

Location (yer)

- Örneğin, Kadıköy Meydanı, Ankara Work of art (sanat eseri)
- Hamlet, Kuğu Gölü Balesi

2.8.3. Kullanım Alanları Nelerdir?

Adlandırılmış Varlık Tanıma, bir metin yığınından adlandırılmış varlıkları (kişiler, yerler, kuruluşlar vb.) tespit etmeyi ve bunları önceden tanımlanmış bir kategoriler kümesinde sınıflandırmayı içeren standart bir doğal dil işleme problemidir. Örneğin elimizde büyükçe bir metin var. Bu metindeki kişi ve organizasyon isimlerini bulmak istiyorsunuz. Bunun için Adlandırılmış Varlık Tanıma kullanabiliriz.

NER’ nin pratik uygulamalarından bazıları şunlardır:

- Raporlanan kişi, kuruluş ve konumlar için haber makaleleri taranır.
- Arama optimizasyonu için özlü özellikler sağlamak: tüm içeriği aramak yerine, ilgili ana varlıkları basitçe arayabilirsiniz.
- Twitter gönderilerinde bahsedilen coğrafi konumları hızlı bir şekilde almak.

Bu çalışmalara ek olarak adlandırılmış varlık tanıma yardımıyla arama algoritmaları, metin sınıflandırmaları, makine çeviri sistemleri ve daha birçok alan örnek gösterilebilir. Makine çeviri sistemlerinde varlık isimlerinin tanınması ENAMEX varlık tiplerinde önemli gibi görünse de aynı önem TIMEX ve NUMEX veri tipleri için de geçerlidir.

2.8.4. Adlandırılmış Varlık Tanımına İlişkin Konular

Tokenizasyon: Hem doğal dil işlemenin hem de adlandırılmış varlık tanımanın ilk adımıdır. Bu aşamada metinler kelime ve cümle seviyesinde alt bölümlere ayrılır. Tokenizasyon işlemi için akla ilk gelen python komutu split() komutudur. Bununla birlikte daha doğru bir tokenizasyon için word_tokenize() ve sent_tokenize() komutları kullanılır. Adından da anlaşılacağı üzere word_tokenize() metni kelimelere dayalı olarak, sent_tokenize() ise metni cümlelere dayalı olarak parçalara ayırır.
POS (Parts of tags) etiketleme: Konuşma bölümü (POS) etiketleme, bir derlemdeki (korpus) kelimeleri, kelimenin tanımına ve bağlamına bağlı olarak konuşmanın belirli bir bölümüne karşılık gelecek şekilde kategorilere ayırmaya atıfta bulunan popüler bir Doğal Dil İşleme sürecidir.Konuşma bölümü etiketleri, bir cümle veya metin içindeki sözcüksel terimlerin karakteristik yapısını tanımlar, bu nedenle bunları anlambilim hakkında varsayımlarda bulunmak için kullanabiliriz.
POS etiketlemenin diğer uygulamaları şunları içerir:

- Adlandırılmış Varlık Tanıma
- Eş referans Çözünürlüğü
- Konuşma tanıma

POS etiketleme esas olarak, bir kelimenin bir cümlede nasıl kullanıldığını açıklar. Konuşmanın sekiz ana bölümü vardır - isimler, zamirler, sıfatlar, fiiller, zarflar, edatlar, bağlaçlar ve ünlemler.

Morfolojik analiz: bir kelimenin kök formunu elde etmek için kullanılan çalışmadır. Konu lemmatization olarak da bilinir. Bir kelimenin kökünü bulmada sıklıkla stemming işlemi kullanılsa da morfolojik analiz bağlamındaki kök bulma oldukça farklıdır. Stemming işleminde sadece ekler çıkarılıp kökler elde edilirken morfolojik analizde kelimenin atası elde edilir. Örneğin, stemming işlemine children ifadesi sokulduğunda sonuç children olurken morfolojik analiz çıktısı child olacaktır. Kelime ek almadığı için stemming algoritması children bulmuştur ama kelimenin atası child olduğu için morfolojik analiz ile child bulunmuştur.

Sözdizimsel ayrıştırma: Sözdizimsel ayrıştırma sözcük analizi (lexical analysis) gibi kelime seviyesinde değil cümle seviyesinde çalışır. Cümlede yer alan kelimelerin gramer açısından analizini yaparak kelimeler arası ilişkileri ortaya çıkarır. Adlandırılmış varlık tanıma çalışmalarında da sıklıkla kullanılan bir çalışmadır. Sözdizimsel analiz konusunda birçok algoritma olmasına rağmen sıklıkla Context-Free Grammar (bağlamdan bağımsız gramer) ve Top-Down Parser (yukardan aşağıya ayrıştırıcı) tercih edilir. Bağlam bağımsız gramer, bir cümlenin üretilmesinde kullanılan kuralların sunulduğu bir analiz türüdür. Dil üretim kuralları metnin anlaşılması esnasında çözümleme yaparken kullanılır. Ayrıştırma (Parsing) bize bir çözümleme yöntemi sunar. Çözümleme yönteminin girdisi bir metin, çıktısı ise bir ağaçtır. Bir derlemin çıktısı ise ağaç bankasıdır (tree bank). Bir ağaç bankası, sözdizimsel veya anlamsal cümle yapısını açıklayan ayrıştırılmış bir metin bütünüdür. Ağaç bankalarının kullanımı hesaplamalı dilbilime çok fazla katkı sunmuştur.

Semantik analiz: semantik analizin amacı metnin ne anlattığı ile ilgilidir. Bu konu metindeki özne, nesne, eylem gibi bileşenlerin ortaya çıkarılması ile ilgilidir. Özellikle özne ve nesne bileşenlerinin bulunması adlandırılmış varlık tanıma ile doğrudan ilişkilidir.

Söylem analizi (discource analysis): Bir cümlenin anlamı tek başına anlaşılamaz. O cümle kendinden önceki cümleden etkilenir ve kendinden sonra gelen cümleyi etkiler. Bu durumun bir neticesi olarak kayan pencere yaklaşımı ile kelimeler birlikte ele alınır. Word embedding tekniklerinden Skipgram ve CBOW bu konuyla ilgili örneklerdir.

Pragmatik analiz: Söylenenlerin gerçekten ne anlama geldiği ile ilgilidir. Bazen, görünen anlamın gerisinde başka anlamlar olabilir. Örneğin, hikâyedeki farklı bir kahraman gerçek hayatta başka bir kişiyi temsil ediyor olabilir. Konu bağlam (context) konusuyla ilgilidir.

2.8.5. Adlandırılmış Varlık Tanıma Yöntemleri

Adlandırılmış varlık tanıma yöntemleri için öne çıkan beş farklı yaklaşım bulunmaktadır. Bunlar sırayla; kural tabanlı sistemler, istatistiksel modeller, sinir ağları, hibrit sistemler ve anlamsal rol etiketlemedir.

Kural tabanlı sistemler: genellikle konu uzmanı kişiler tarafından yazılan kurallar yardımıyla çalışır. Uzmanlar tarafından çıkarılan bu kurallar metinde yer alan kalıplarla, sözlüksel bilgilerle veya sözdizimsel yapıyla ilgili olabilir. Kuralların üretimi ve güncellenmesi maliyetli işlemlerdir. Ayrıca genelleme yetenekleri zayıftır.

İstatistiksel modeller: istatistiksel modellere göre adlandırılmış varlıklar metindeki diğer ögelerden istatistiksel açıdan farklılaşarak ayırt edilebilirler. Bu kapsamda; Gizli Markov modelleri, maksimum entropi modelleri ve destek vektör makineleri (SVM'ler) gibi algoritmalar kullanılabilir. Modelin en önemli sorunu büyük miktarda veriye ihtiyaç duymasıdır.

Sinir ağı modelleri: adlandırılmış varlık tanıma için kullanılabilecek makine öğrenmesi yöntemlerinin bir alt kümesini oluşturur. Yapay sinir ağları veride yer alan kalıpları öğrenebilme yeteneği nedeniyle bu problem için de uygundur. İstatistiksel modellerde olduğu gibi büyük miktarda veriye ihtiyaç duyar.

Hibrit sistemler: bunlar, adlandırılmış varlık tanıma yaklaşımlarını bir araya getiren yöntemlerdir. Örneğin kural tabanlı sistem ve istatistiksel yaklaşım bir arada kullanılarak adlandırılmış varlık tanıma yapılabilir. Mesela kişi adları istatistiksel yaklaşım ile kuruluş adları kural tabanlı sistemle elde edilebilir.

Anlamsal rol etiketleme: Anlamsal rol etiketleme (SRL), bir cümledeki her kelimenin rolünü etiketlemeye dayalı olarak çalışır. Bu durum cümlenin sözdizimsel yapısını ve cümledeki varlığın bağlamını elde ederek yerine getirilir.

3. İLGİLİ ÇALIŞMALAR

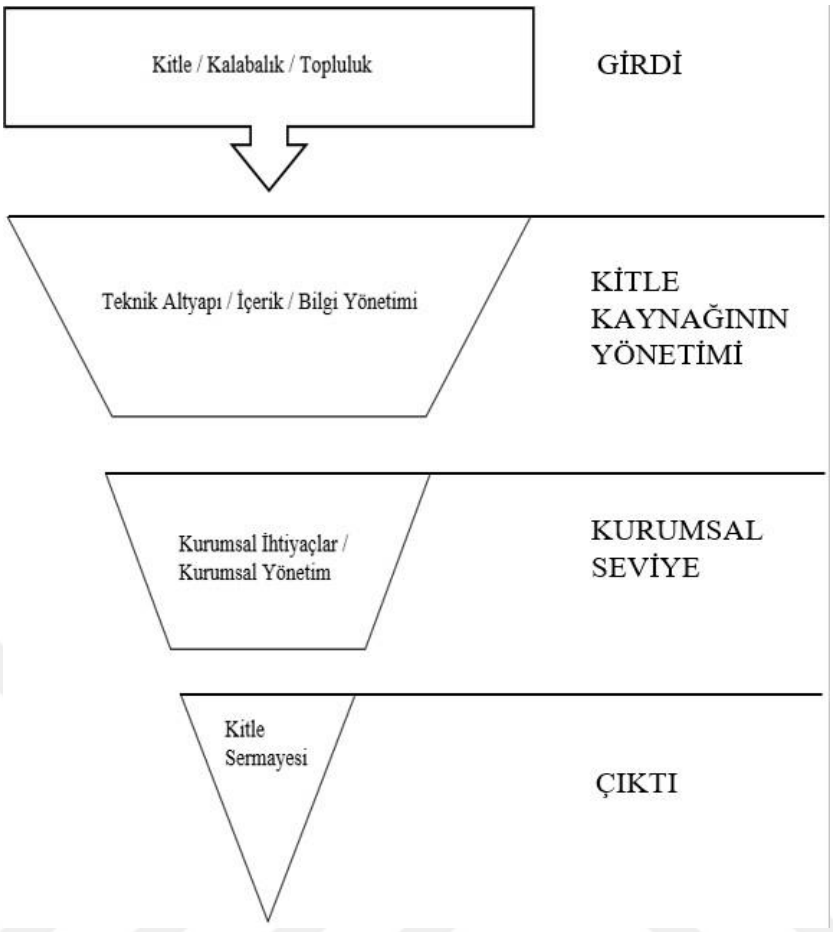
Kullanılan teknojilerle alakalı olarak literatürde birçok çalışmaktadır. Bu bölümde bu çalışmanın geçmiş çalışmalarla benzerlik ve farklılıkları ortaya konacaktır.

Kitle kaynak kullanımı, bir firmanın çevrimiçi bir topluluğa veya kalabalığa 'açık çağrı' biçiminde faaliyetlerin dış kaynaklardan sağlanması sürecini ifade eden yeni geliştirilmiş bir kavramdır. Kalabalığın herhangi bir üyesi atanmış bir görevi tamamlayabilir ve göstermiş olduğu çaba için ödüllendirilir. İşletmeler, verilen çeşitli görevlerin kendi çalışanları tarafından değil, bir kalabalığın üyeleri tarafından daha iyi tamamlanabildiklerini düşündükleri için 'kitle kaynak kullanımı' tercih etmeye başlamışlardır (Whitla, 2009)

Kitlesel kaynak, şirket dışı kurumlarla yapılan uzun süreli belli anlaşmalardan ziyade, kısa süreli bir işbirliği olup; bir iktisatçının, bir yönetim danışmanının ya da bir pazarlama gurusunun buluşuyla değil, sevdiği şeyleri özellikle başkalarıyla birlikte yapmaktan hoşlanan binlerce insanın plansız eylemleri sonucu ortaya çıkan bir kavramdır (Yavuz, 2014). Kitlesel kaynağı, yani bu gücü oluşturanlar, herhangi bir organizasyon olmadan organize olanlardır (Abrahamson, 2013).

Şekil 3.1’ deki Kitle Kaynak Süreci kapsamında değerlendirilebilecek birçok uygulama bulunur. Bu uygulamalara, genellikle çevrimiçi araştırmalarda kullanılan Survey Monkey ve MTurk uygulaması veri temizleme ve düzenleme işlemleri, wikipedia uygulamaları, anket formlarının ön testlerinin yapılması,belirli bir konuyla ilgili olarak kamuoyu eğimlimlerinin tespiti ve incelenmesi örnek gösterilebilir. Aynı zamanda yazılım programlarının beta sürümlerinin kullanıcılara ücretsiz olarak verilmesi ile birlikte yazılımdaki hataların tespiti ve çözümü, Kickstarter (yeni başlangıç yatırımcılar) fonları gibi uygulamalar en yaygın bilinenler arasındadır (Altunişik, 2017).

Ayrıca, sosyal bilimler alanındaki bilimsel araştırmalarla ilgili incelemelerde kitle kaynak kullanımı ile ilgili çeşitli uygulamalardan bahsedilmektedir. Örneğin, web-tabanlı bir tişört firması olan Threadless.com tasarımlarını çevrimiçi olarak düzenlenen yarışmalar şeklinde yürüterek kitle kaynak kullanımına dayandırmaktadır.



Şekil 3.1. Kitle Kaynak Kullanım Süreci (Şeker, 2015)

Benzer şekilde web-tabanlı bir işletme olan iStockphoto.com telifsiz resimler, animasyonlar ve videoları satan bir işletmedir. Bu işletme fotoğraflarını kitle kaynağı kullanarak temin etmektedir. iStockphoto fotoğrafçısı olabilmek için kişiler özgün fotoğraflarını işletmeye göndererek onaylanmayı beklerler. Onay alan fotoğrafçı artık iStockphoto'nun fotoğrafçısı olur ve sisteme fotoğraf yüklemeye başlar. Bir dolar gibi bir fiyatla satılan resimlerden elde edilen gelirin %20'si fotoğrafçıya gitmektedir. Benzer uygulamalara imza atan markalar arasında Converse, Sony, Chrysler, Doritos, Chevrolet, GoldCorp, vb. markalar da sayılabilir (Brabham, 2008). Ülkemizde de buna benzer uygulamalar strateji geliştirme bağlamında çeşitli markaların düzenlemiş olduğu oyunlar ve yarışmalar şeklinde ve kar amaçsız işletmeler için konsept geliştirme amaçlı çalışmalar olarak karşımıza çıkmaktadır.

Çalışmada kullanılan önemli yöntemlerden biri de DDİ tekniklerinden biri olan Adlandırılmış Varlık Tanıma yöntemidir. Varlık ismi tanımı konusunda birçok dilde çalışmalar yapılmıştır. İngilizce bu dillerin başında gelmektedir.

VİT sistemleri, ilk başta kural tabanlı olarak yapılsa da sonradan makine öğrenmesi teknikleri kullanılarak yaygınlaştırılmıştır (Nadeau ve Sekine, 2007). 1990'lı yıllarda VİT için ilk çalışmalar başlamıştır. Bu çalışmalardan biri İngilizce için yapılmıştır. Elle tanımlanmış kurallar ile metindeki şirket isimlerini bulmayı amaçlayan bir çalışmadır (Rau, 1991).

Varlık ismi kavramı ilk kez MUC-6 konferansında ortaya atılmıştır (Nadeau ve Sekine, 2007). MUC-6 konferansı ile birlikte VİT çalışmaları hızlanmıştır (Grishman ve Sundheim, 1996). Sonraki yıllarda adlandırılmış varlık tanıma alanında düzenlenen konferanslar artmıştır; CoNLL, IREX, HAREM, ACE bu konferanslardan bazılarıdır.

Ratinov ve Roth' un (Ratinov ve Roth, 2009) çalışması, CoNLL03 verisinde %90,8 başarım ile İngilizce için en yüksek başarı gösteren çalışma olarak gösterilebilir.

Tür ve diğ. (Tür ve Oflazer, 2003) tarafından yapılmış olan ve sadece Türkçeye özgü olan ilk çalışmada ise, %91,56'lık bir başarım elde edilmiştir. Yapılan bu çalışmada HMM haber metinleri üzerinde kullanılmıştır. İstatistiksel modelin oluşturulmasından itibaren kelimelerin sözlüksel, bağlamsal ve morfolojik özellikleri kullanılmıştır. İlgili çalışmada kişi, yer ve kuruluş adlarının tespiti söz konusudur. Bayraktar ve Temizel' in (Bayraktar ve Temizel, 2008) Türkçe finansal metinlerden kişi adlarını belirlemek için yaptıkları çalışmada, kelimelerin örüntü modelleri ve frekans analizleri kullanılmıştır. Türkçe için Küçük ve Yazıcı (Küçük ve Yazıcı, 2009) tarafından yapılan çalışma ise kural tabanlı geliştirilmiş bir sistemdir. Haber metinlerinde, tarihi metinlerde, çocuk hikayelerinde kişi, yer, kuruluş ve sayısal ifadelerin tespiti ile ilgilenir. Bu kural tabanlı sistem daha sonra istatistiksel öğrenme yöntemiyle geliştirilerek haber metinleri üzerindeki başarımı %87,96' dan %90,13' e çıkarılmıştır (Küçük ve Yazıcı, 2012). Özkaya ve Diri' nin (Özkaya ve Diri, 2011) Türkçe e-posta metinlerinde kişi, yer ve kurumu belirlemek için koşullu rastgele alanlar yöntemini kullanarak yaptıkları çalışmada %92,89 başarı oranı elde edilmiştir. Bu çalışmada sözlüklerden alınan kelimelerin uzunlukları, kısaltma olup olmadığı, noktalama işareti içerip içermediği, cümlenin ilk kelimesi olup olmadığı vb. bilgilerden yararlanılmıştır.

Şeker ve Eryiğit (Şeker ve Eryiğit, 2012) tarafından Türkçe için koşullu rastgele alanlar ile cümle başı bilgisi, sözlük bilgisi, büyük/küçük harf bilgisi ile kelimelerin morfolojik özelliklerinin kullanıldığı çalışmada ise haber metinlerinde CoNLL metriğine göre %92

oranında bir başarı elde edilmiştir. Bu zamana kadar genellikle formatı düzgün olan ve dilbilgisi kurallarına uygun olarak yazılan metinler üzerinde çalışmalar yapılmıştır. Bu sebeple web ortamındaki bilgiler her zaman uygun formatta olmadığından, bu gibi metinler ele alınarak yapılan çalışmalar son zamanlarda artış göstermiştir.

Ritter ve diğ. yaptıkları bir çalışmada İngilizce tweetleri ele almışlardır. Bu çalışmada (Ritter ve diğ., 2011) performansı arttırmak için hali hazırda bulunan DDİ araçlarını tweetlere uyarlamışlardır. Kelime sınıfı etiketleme, tamlama bulma aşamalarında CRF yöntemi kullanılmış, etiketlenmiş konu modelleme yöntemi ise varlık ismi sınıflandırılırken kullanılmıştır. Bu çalışmayla %51 oranında başarım elde edilmiştir. Bir başka çalışmada ise yine aynı şekilde İngilizce tweetler üzerinde çalışılmıştır. Yarı denetimli bir sistemi, koşullu rastgele alanlar yöntemi ve en yakın komşu algoritması ile oluşturmuşlardır (Liu ve diğ., 2011). Metinlerin imla, sözlük özellikleri ve ek sözlük sayesinde %80,2 oranında başarı sağlanmıştır.

Denetimsiz öğrenme yöntemiyle, İngilizce tweetlerdeki varlık isimlerini bulan bir sistem oluşturulmuştur (Li ve diğerleri, 2012). Filtre tabanlı adlandırılmış varlık tanıma sistemini (Oliveira ve diğ., 2013) İngilizce tweetleri baz alarak oluşturmuşlardır. Oluşturulan bu sistemin başarı oranı CRF'ye göre %3 daha fazladır.

Adlandırılmış varlık tanıma üzerine Türkçe tweetler için de iki çalışma bulunmaktadır. İlk çalışmada (Çelikkaya ve diğ.,2013) tweetler üzerinde normalizasyon yaptıktan sonra haber metinlerini eğitilmiş olan istatistiksel model ile test etmişlerdir ve bunun sonucunda %19,28 lik başarı oranını CoNLL metriğine göre elde edilmiştir. Test için kullanılan istatistiksel modeli oluşturmak için, Şeker ve Eryiğit' in (Liu ve diğ., 2011) çalışmasındaki aynı teknikler kullanılmıştır. Küçük ve Steinberger (Küçük ve Steinberger, 2014) ise Türkçe tweetler üzerinde başka bir çalışma yaparak Küçük ve Yazıcı tarafından geliştirilmiş kural tabanlı sistemi iyileştirmiştir. Kural tabanlı sistem ile tweetler test edildiğinde %53,23 başarı oranı elde edilmiştir. Sonrasında bu sistem geliştirilip tekrar test edildiğinde ise başarım oranı %61,17' ye yükselmiştir. Bu sistemin geliştirilmesinde normalizasyon, kelimelerin büyük/küçük harf bilgileri ve sözlük bilgisi kullanılmıştır.

ITU, Türkçe Doğal Dil İşleme alanında normalizasyon, adlandırılmış varlık tanıma gibi birçok konu üzerine çalışmalarda bulunmuştur. Bu çalışmalar açık kaynak kodlu değildir

fakat ücretsiz web servis hizmeti sunmaktadır. Bu ücretsiz hizmet projelere destek sağlamaktadır. Projelerde verilerden anlamlı sonuçlar çıkarılabilmesi için uygun formata çevrilmesi gerekir. Bu kapsamda ilk işlem olarak normalleştirmedir. Metin normalleştiricinin görevi metindeki standart olmayan kelimeleri tespit ederek bunların harf bazında açılımlarını yapmaktır (Bicil, 2016). Sosyal medyada kullanılan dil, çeşitli nedenlerle resmi olarak yazılmış metinlerden önemli derecede farklıdır. İnsanlar genellikle çok daha kişisel bir şekilde yazar ve konuşulan deyimleri yansıtırlar ve bunun bir parçası olarak deyimsel dilbilgisi kurallarını tartışmalı bir şekilde kullanırlar (Eryiğit ve Torunoğlu Selamet, 2017). Normalleştirme adımında veriler noktalama işaretlerinden, kısaltmalardan, anlamsız seslerden ayrıştırılarak düzenleme işlemi yapılmaktadır. Düzenli hale getirilen verilerde adlandırılmış varlık tanıma işlemi yapılmaktadır.

Küçük ve Yazıcı (Can ve Toraman, 2010) tarafından birçok örüntü ve sözlük kaynağından yararlanan kural tabanlı Türkçe adlandırılmış varlık tanıma yaklaşımı geliştirilmiştir. Adlandırılmış Varlık Tanıma için birçok hazır araç vardır fakat neredeyse hiçbiri Türkçeyi desteklememektedir. Bu durumun bir istisnası olarak Avrupa Birliği Birleşik Araştırma Merkezi tarafından geliştirilmiş olan ve Türkçe gibi birçok dili destekleyen JRC-Names programı bulunmaktadır. Ayrıca Türkçe varlık ismi tanımlama için İstanbul Teknik Üniversitesi Türkçe NLP Grubu'nun CRF algoritmasını kullanan bir çalışması yer almaktadır (Eryiğit, 2015). JRC-Names, 20 farklı dilde günde 100.000 haber makalesine kadar yeni kişi ve kuruluş adlarını tanıma, yeni bulunan adların otomatik olarak tanınması için çok yıllı büyük ölçekli bir çabanın sonucudur (Steinberger ve diğ., 2011).

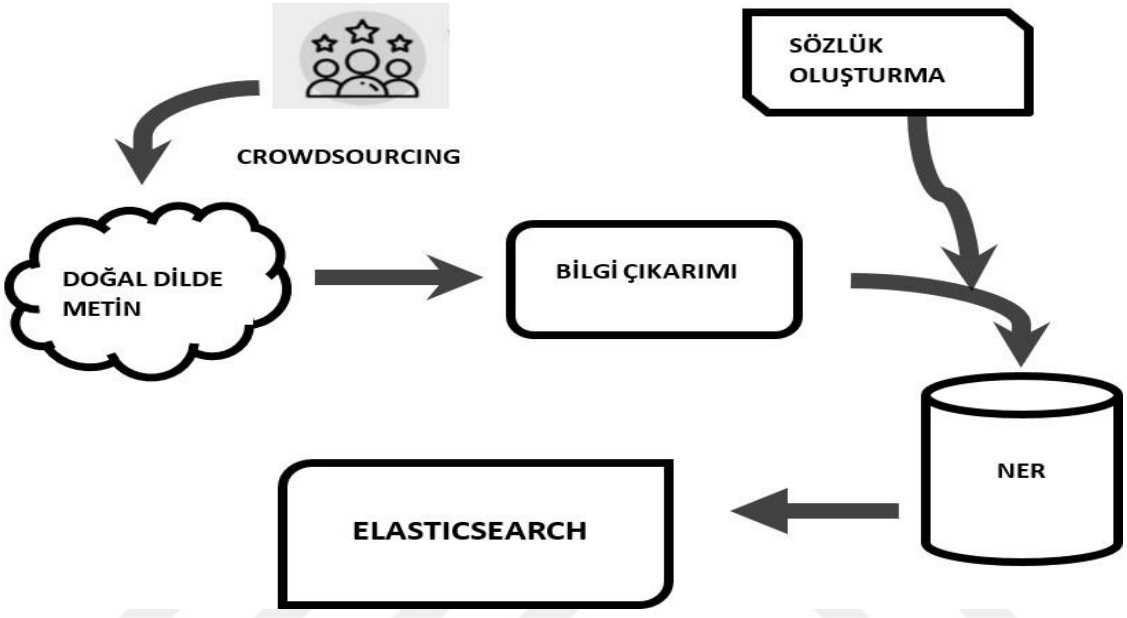
Ülkemizde kural tabanlı Adlandırılmış Varlık Tanıma çalışmalarına Yıldız Teknik Üniversitesindeki araştırmacıların çalışması örnek olarak verilebilir. Kural tabanlı tanıma için oluşturulan sistemin amacı, Türkçe dokümanlarda geçen kişi, yer ve kurum isimleri ile para, tarih ve saat varlıklarının bulunarak etiketlenmesidir (Özger, 1995). Yapılan çalışmada tüm varlıklar için o türe özgü tanımlayıcı özellikler kural haline getirilmiştir. Kişi, kurum, yer, tarih, saat, para varlıkları için ayrı ayrı kurallar oluşturulmuştur.

Çalışmada veri seti olarak güncel gazetelerden 6 farklı konudan 20’şer adet doküman olmak üzere toplam 120 adet doküman kullanılmıştır. Kurallar ve veri seti kullanılarak Adlandırılmış Varlık Tanıma işlemi gerçekleştirilmiştir.

NLP çalışmalarında, temel birimler olarak kelimeleri kabul ederken saf analiz veya üretime yoğunlaşmak gelenekseldir. Bununla birlikte, bu temel birimler açıkça ayrılmadan, herhangi bir analiz veya üretim yapmanın imkânsız olduğu bir gerçektir (Webster ve Kit, 1992). Bu işlem için cümle sonlarında kullanılan noktalama işaretlerinden yararlanılmaktadır. Her iki yöntemin de temel işlevi gövdeleme (Stemming) ve baş sözcük bulma (Lemmatization) benzeri çalışmalardır (Jivani ve Ganesh, 2012). Gövdeleme işleminde tüm yapım ve çekim eklerinden kurtulup kelimenin kökü bulunmaktadır. Baş sözcük bulma işlemi ise kelimenin çekim eklerini atıp, kelimenin sözlükteki anlamına ulaşmayı ifade etmektedir. Lemmatization işlemi sözlük, morfolojik analiz bilgisi ve bağlam bilgisi kullanır.

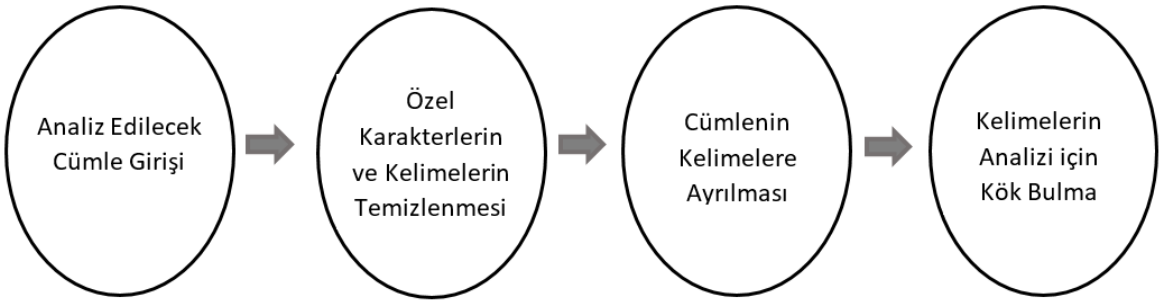
4. ÖNERİLEN SİSTEM

Bu çalışmada, marketlerdeki ürün bilgilerinin varlık isimlerini belirleyecek bir model oluşturulmuştur. Bu modelin oluşturulmasında sözlük tabanlı yaklaşım kullanılmıştır. Geliştirilen model market, marka, ürün, miktar ve fiyat gibi varlık tiplerinin tespitini yapabilmektedir.



Şekil 4.1. Önerilen Model

Kullanıcılardan market, marka, ürün, miktar ve fiyat bilgisini içeren yorum cümleleri toplanır. Bunlar düzensiz bir yapıda olduğu için DDİ bu cümleler üzerinde yapılır. Yapılan çalışma ile metinler içerisindeki anlamlı bilgiler çıkarılabilir.



Şekil 4.2. Bilgi Çıkarımı Adımları

Daha önce Türkiye'deki güncel market, ürün, marka, miktar ve fiyat etiketlemelerini toplu biçimde içeren bir veri seti üzerine herhangi bir çalışmaya rastlanmamıştır. Bu sebeple çalışmamızda kullanılan veri seti manuel olarak hazırlanmıştır. Sözlük oluşturmak için TÜİK ve TÜFE verileri ile marketlerin çevrimiçi alışveriş sitelerindeki veriler kullanılmıştır. Veri seti excel de oluşturulup, projede sözlük olarak kullanılmak üzere csv formata çevrilmiştir.

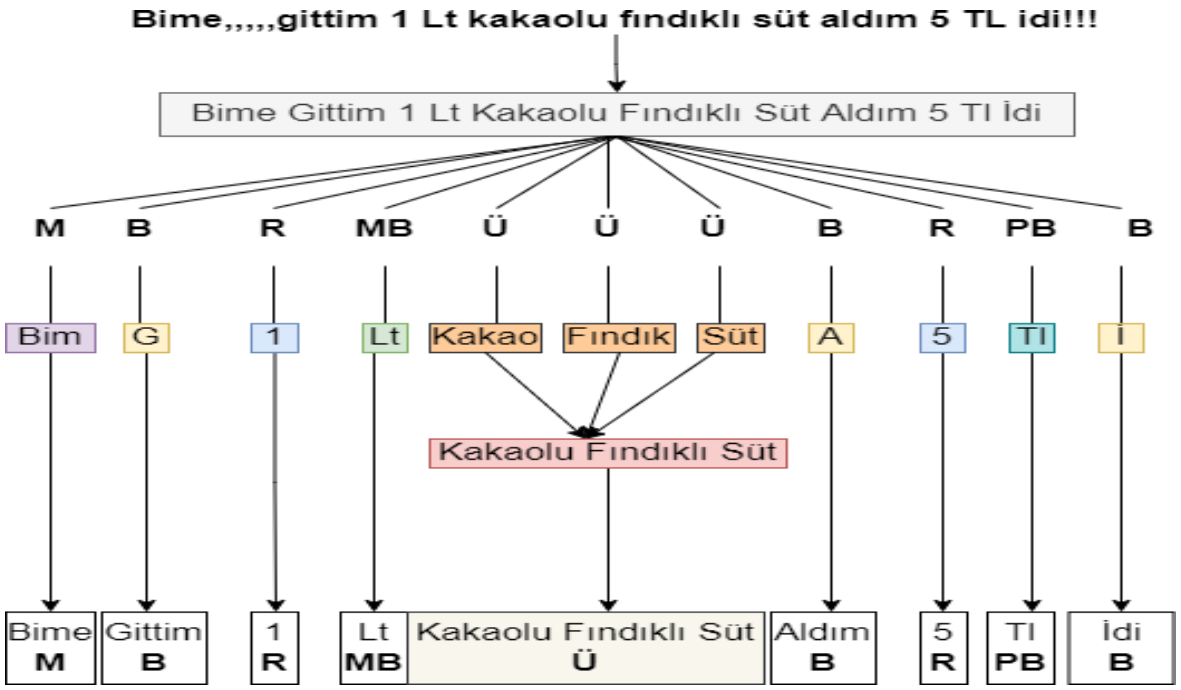
Tablo 4.1. Verilerin Toplanması

ID	Key	Tag	Category
106	Bim	Market	
107	Migros	Market	
109	Sütaş	Marka	
110	Tl	Para Birimi	
111	Dolar	Para Birimi	
112	Litre	Miktar Birimi	
113	Gram	Miktar Birimi	
114	Gofret	Ürün	Atıştırmalık
115	Domates	Ürün	Meyve-Sebze
116	Süt	Ürün	Süt-Süt Ürünleri
117	Ekmek	Ürün	Ekmek-Tahıl
118	Ezine Peyniri	Ürün	Süt-Süt Ürünleri

Veri seti zenginleştirmenin devamında algoritma geliştirme aşamaları yapılmıştır. Projenin en ucuz market ürününü bulma amacından yola çıkılarak kullanıcıların yorumları alınmıştır. Bu yorum cümlesi market, ürün, marka, fiyat ve birim bilgilerini içermektedir. Sistemin test edilmesi için ilk aşamada veriler veri setinden çekilmektedir. Kullanıcı tarafından gelen cümle üzerinde öncelikle özel karakterlerin temizlenmesi, olası yazım yanlışlarını en aza indirgemek amacıyla normalizasyon işlemi yapılmaktadır. Sonraki aşamada; cümleyi kelimelere ayırarak her bir kelimenin sözlükte karşılığı aranarak metindeki varlık isimleri tespit edilir ve önceden belirlenen kategorilere göre sınıflandırılırlar.

Metindeki varlık ismi tespiti yapılırken bazı ürün isimlerinde tutma oranı uygulanmaktadır. Bu oran ürün ismi birden fazla kelimeden oluşan ve ürün ismini içeren her bir kelimenin

sözlükte ürün ismi olarak tanımlandığı durumlarda her kelimeyi ayrı ayrı ürün şeklinde etiketlemek yerine doğru tespit edilmiş varlık sayısının, tespit edilmiş varlık sayısına oranıyla bir benzerlik oranı hesaplanır. Bu oran 1'e ne kadar yakınsa ürün isminin tespiti de o kadar doğru olmaktadır. Yani oran 1'e yaklaştıkça benzerlik artmaktadır. Tutma oranının örnek ile gösterilmiş hali Şekil 4.3'te gösterilmiştir.



Şekil 4.3. Sözlük Tabanlı Adlandırılmış Varlık Tanıma İşlemi

Araştırmamızda birleşik kelimeli ürünlerde her kelime sözlükte bulunduğu durumda etiketleme işlemi yapılabiliyorken (Örn: Çikolatalı Süt), kelimelerden biri veya daha fazlası sözlükte bulunmadığı durumda ise etiketleme işlemi yapılamamaktadır (Örn: Oda Spreyi).

Bunların yanı sıra bir de ELMO, BERT gibi bağlamsal olan gösterimler de mevcuttur (Devlin ve diğ., 2019). Bağlamsal duruma örnek vermek gerekirse, "5 litre sıvıyağı yüz liraya aldım." ve "Sabah uyandığımda ilk işim yüz yıkamak oluyor." cümleleri incelenebilir. İki cümlede de aynı kelime geçmesine rağmen cümle içinde kullanılma şekillerindeki (birlikte kullanıldığı kelimeler, cümledeki yeri …) farklılıktan dolayı farklı anlamlara gelmektedirler. Bu tarz bağlamsal durumlara çözüm olan teknikler ise sonraki çalışmalarda işlenebilir.

Ayrıca, sürekli sözlük güncelleme şeklindeki çözümler ve kural tabanlı çözümler sınırlı çözüm üretmektedir. Bunun için derin öğrenme tabanlı çalışmalar geliştirilebilir.

4.1. Veri Temizleme

Verinin söz konusu olduğu yerde yapılması gereken en önemli işlemlerden birisi veriyi analize hazır hale getirmektir. Bu işlemin ilk adımı da veri içerisinde yer alan fazlalıkların veriden ayıklanmasıdır. Veri temizleme aşamasında özellikle yapılacak işlemle alakası olmayan veri parçaları veriden temizlenir. Bu çalışmada veriler temizlenirken kelimeler arasındaki birden fazla boşluk kaldırılıp, içerisinde geçen özel karakterler temizlenmiştir. Aynı zamanda kelimelerin baş harf büyük devamı küçük harf olacak şekilde düzeltilmiştir. Böylelikle cümledeki kelimelerin etiketlenmesi için veri bir adım hazır hale getirilmiştir. Her kelimenin baş harfinin büyük diğer harflerin küçük olmasını python programlama dilinde capitalize() metodu ile yapabiliyorken c# toTitleCase() metodu ile yapılabilmektedir. Şekil 4.4' de kullanıcıdan alınan yorum cümlesinin özel karakterleri ve birden fazla boşluk olan cümleleri temizleyerek her kelimenin baş harfi büyük devamı küçük olacak şekilde düzenleyen kod parçasını görebiliriz.

```
string data = string.Empty;

Regex rgx = new Regex(@"[^\w\d]");

data = rgx.Replace(textBox1.Text, " ");

while (data.Contains("  "))
{
    data = data.Replace("  ", string.Empty);
}

textBox1.Text = data.ToLower();
textBox1.Text = CultureInfo.CurrentCulture.TextInfo.ToTitleCase(textBox1.Text);
```

Şekil 4.4. Veri Temizleme için Kod Parçası

4.2. Normalizasyon

Veriyi işleme hazırlamadaki ikinci adım normalizasyon adımıdır. Veri temizleme ile analizle ilişkili veriler bırakıldıktan sonra alakalı verilerde de sorunlar olabilir. Örneğin, bir hashtag ile alakalı twitter mesajında hatalı yazılmış kelimeler bulunabilir. Veri analiz edilecek veridir ama verinin bir kısmı gürültüler içermektedir. Verideki gürültülerin giderilmesi gibi ön işlemler normalizasyon olarak bilinir.



Şekil 4.5. Normalizasyon Örneği

Şekil 4.5'de görüldüğü üzere NGram kütüphanesi kullanılarak oluşturulan sözlükteki kelimelere göre normalizasyon işlemi yapılabilmektedir. Örnek olarak, "Migrsa gittim 1 litre dmes sütü 5 liraya aldım" şeklinde bir deneme yapıldığında veri setinde bulunan kelimelere göre yanlış yazılabilme ihtimali olan kelimeleri listeleyebilmekteyiz. Böylece cümledeki gürültüler temizlenerek, bu kelimelerin doğru yazılmış haliyle gerekli kök bulma işlemleri yapılabilir.

4.3. Kelimelere Ayrıştırma (Tokenizasyon)

Kelimelere ayrıştırma, belgelerin düz biçimlerinden hesaplama birimlerine, hesaplamalı dilbilimi uygulamalarına uygun hesaplama sembollerine dönüştürülmesi ve cümle sonlarının saptanmasıdır (Dinçer, 2004).

Kelimelere ayrıştırma işlemi gerçekleştirilirken metin içinden "Boşluk, yeni satır, noktalama işareti, roma rakamı, numara, yüzdelik ifade, zaman, tarih, URL, E-mail, HashTag, Mention, MetaTag, Emoji, Emoticon, Bilinmeyen" olan kelimeler temizlenmiştir.

4.4. Dil Bilimi İşlemleri (Morfoloji)

Kelime ayrıştırma ile ayrılmış kelimelerin her birinin kök halinin elde edilmesi dil bilimi işlemleri ile sağlanır.

Dil bilimi işlemleri Stemming ve Lemmatization olmak üzere iki kısma ayrılır.

- Stemming: Bir kelimede yer alan ön ekleri ve son ekleri keserek kelime kökünü bulmaya çalışır (Özker, 2019).
- Lemmatization: Kelimelerin morfolojik analizlerini esas alır. Böylelikle, algoritmanın kelimenin kökünü bulması için detaylandırılmış bir sözlüğe ihtiyacı vardır (Özker, 2019).

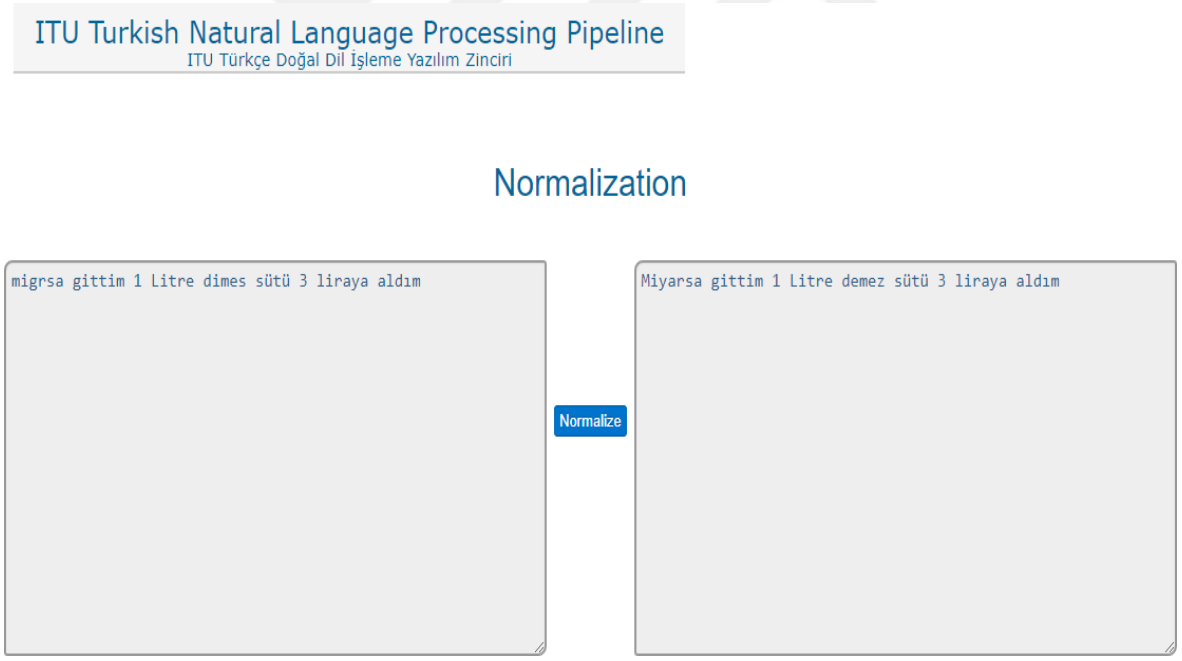
Örneğin "ekmeği" kelimesinin Stemming ile kökü "ekmeğ" olarak bulunurken Lemmatization ile "ekmek" olarak bulunur.

Dil bilimi işlemi sırasında her kelimenin baş harfi büyük devamı küçük harfe çevrilerek metin birimleri normalleştirilmiştir. Metin birimleri normalleştirildikten sonra proje kapsamında geliştirilen algoritma ile kök bulma tipi kullanılarak kelimelere ayrıştırma işleminde elde edilen kelimelerin kökü bulunur. Sözlük oluşturma işleminde sözlüğe eklenen kelimelerin hangi etikete ait olduğu "Tag" alanında belirtilmiştir.

5. DENEYSEL ÇALIŞMALAR

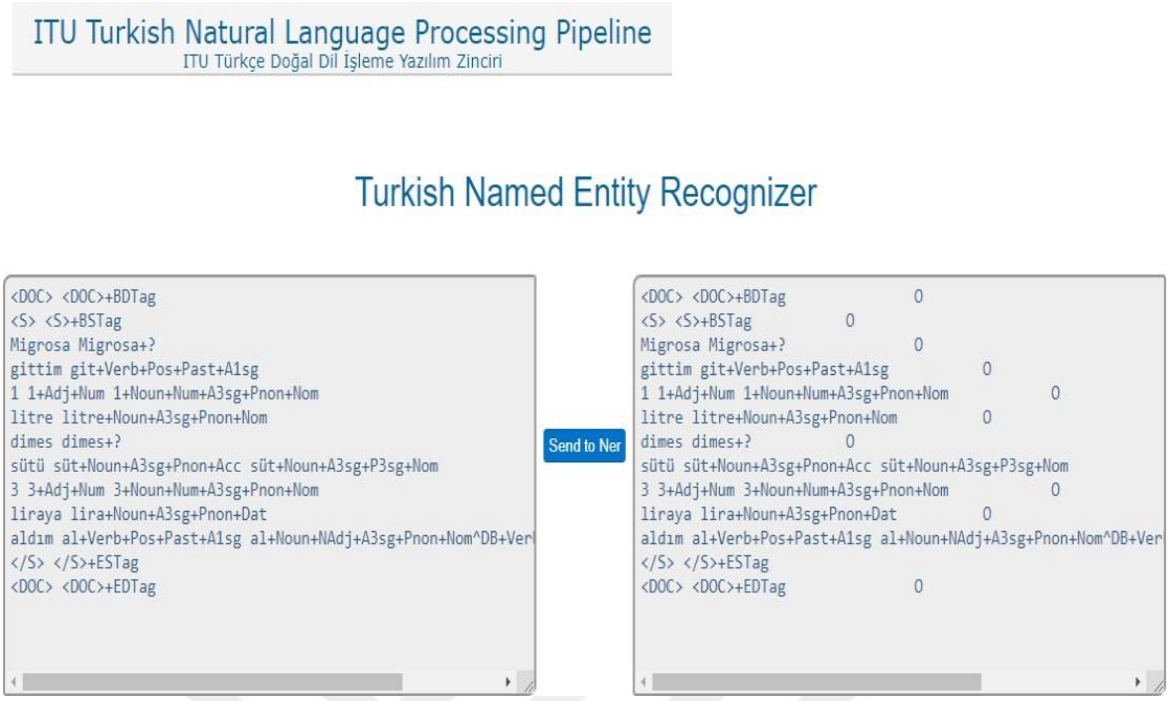
Çalışmada ilk olarak, kullanıcı tarafından gelen cümle üzerinde özel karakterlerin temizlenmesi ve olası yazım yanlışlarını en aza indirgemek amacıyla normalizasyon işlemi ve sonrasında tokenizasyon işlemi yapılmaktadır.

Adlandırılmış varlık tanıma için daha önce yapılan birkaç çalışma kullanılarak sisteme entegre edilmeye çalışıldı. Bu sebeple normalizasyon işleminde ilk olarak İstanbul Teknik Üniversitesi'nin hazırlamış olduğu Türkçe Nlp Tool'u kullanılması planlanmıştır. Türkçe Nlp Tool'u üzerinde birden çok bileşen bulunmaktadır. Bu bileşenlerden biri de Adlandırılmış Varlık Tanıma olarak ele alınmaktadır. Bu tool üzerinde önce normalizasyon sonrasında adlandırılmış varlık tanıma işlemleri denense de ilgili tool'un, bu çalışmada kullanılan veri seti için istenilen performansı göstermediği anlaşılmıştır. Günlük hayatta kullanılan kelimeler için normalizasyon bileşeni başarılı sonuç üretse de market, marka, ürün vs. etiketleri için başarılı olmamıştır.



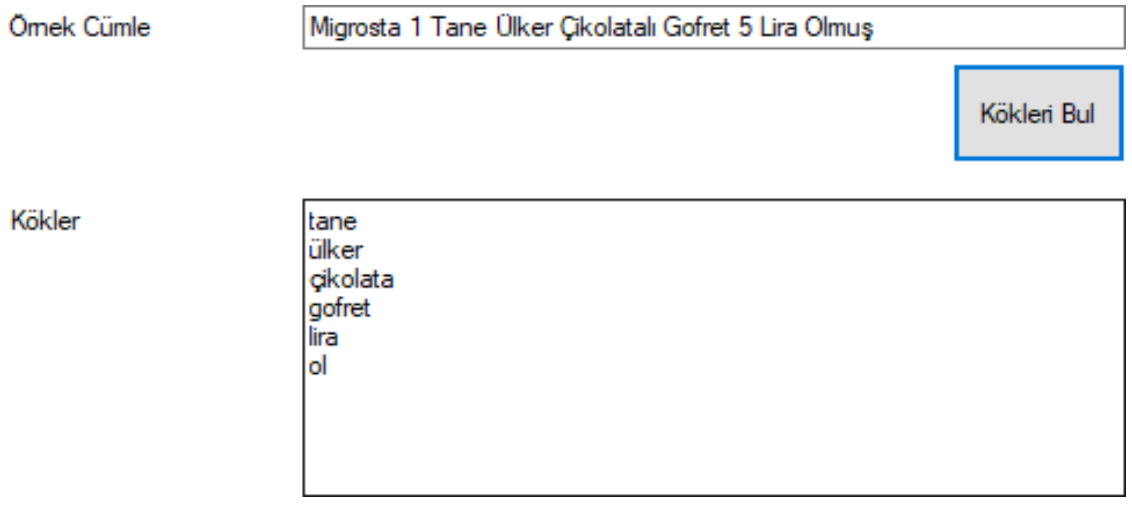
Şekil 5.1. İTÜ Normalizasyon Örneği

Şekil 5.1' de görüldüğü gibi "Migrs" kelimesi başarılı sonuç üretmemiştir. NER bileşeni için kontrol edildiğinde ise Şekil 5.2'deki gibi bir sonuç üretmiştir.



Şekil 5.2. İTÜ NER

Burada istenilen sonuç alınamadığı için kelimelerin köklerinin bulabilmek için Zemberek kütüphanesinin kullanılması planlanmıştır. Fakat yapılan denemeler sonucunda istenilen kökler elde edilememiştir. Çünkü Zemberek kütüphanesinin çalışma yapısına bakıldığında bu çalışma için kullanılan veri seti için istenilen sonuçları vermediği görülmüştür.



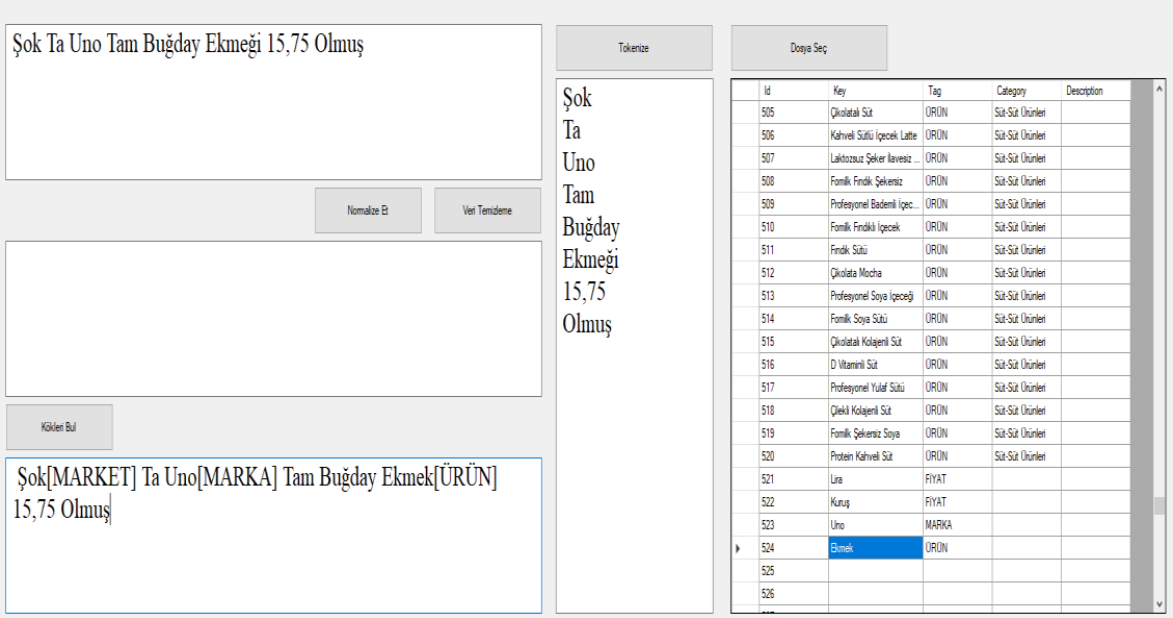
Şekil 5.3. Zemberek Kütüphanesi ile Kök Bulma

Şekil 5.4’ de görüldüğü üzere market ismini kök olarak algılayamıyor. Bu denemelerin sonucunda varlıkların etiketlenebilmesi için bir algoritma geliştirilmiştir. Geliştirilen bu algoritma için aynı örnek ile denediğimizde kökleri bularak, etiketleme işleminin başarılı şekilde yapıldığı görülmüştür.



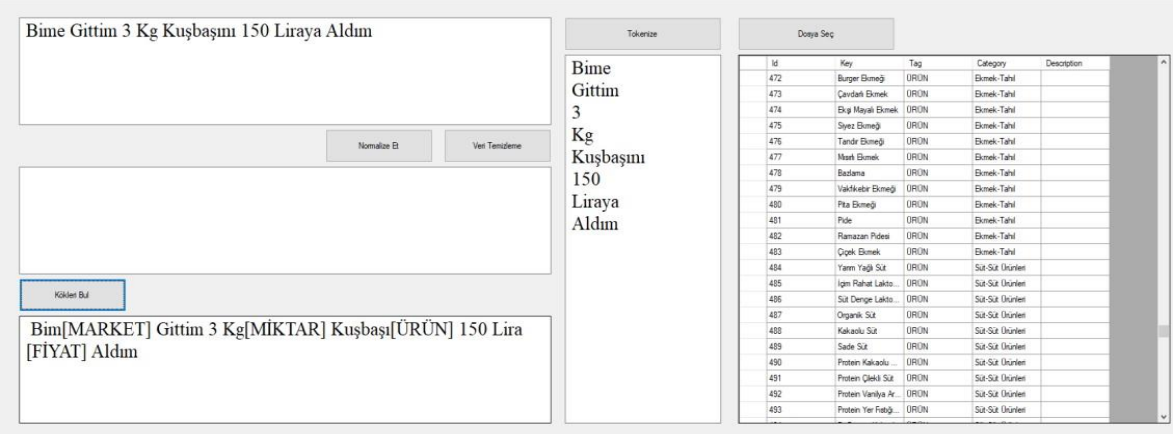
Şekil 5.4. Kök Bulma ve NER İşlemi

Bu algoritmayla birlikte, kullanıcıdan alınan cümledeki her bir kelime tek tek kontrol edilir. Eğer kelime oluşturduğumuz sözlükte yer alıyor ise direkt olarak kök listesine atılır. Fakat kelime herhangi bir ek aldıysa veya kelimede ses olayı var ise (ünsüz yumuşaması) ilgili kelime kök haline getirilene kadar sondan harf silme işlemi gerçekleştirilir. Kelime sonundan silinen harflerden sonra oluşabilecek ünsüz yumuşamalarını düzeltmek amaçlı pçtk kuralı uygulanmıştır. Geliştirilen algoritmada bu kural için gerekli harf ve dönüştürülecek harf, listede [['b', 'p'], ['c','ç'], ['d','t'], ['g','k'], ['ğ','k']] bu şekilde tutulmuştur. Bu işlem cümledeki tüm kelimeler için bu işlem yapılmıştır. Bu sayede Türkçe dil bilgisi kuralına uygun doğal dil işleme tekniği geliştirilmiştir. Örneğin Şekil 5.5’de ‘Şok’ta Uno tam buğday ekmeği 15,75 liraya aldım.’ Cümlesinde ekmeği kelimesinde sondan harf silinmesi ile ilk kelime ‘ekmeğ’ kelimesine dönüşecektir. Burada stemming işlemi yapılıyor ve pçtk kuralına göre ilgili kelime Ekmek’ e çevriliyor. Bu şekilde kelimenin kökü elde edilmiş olup, sözlükteki kategorisine göre gerekli sınıflandırma yapılmıştır.



Şekil 5.5. Stemming Örneği

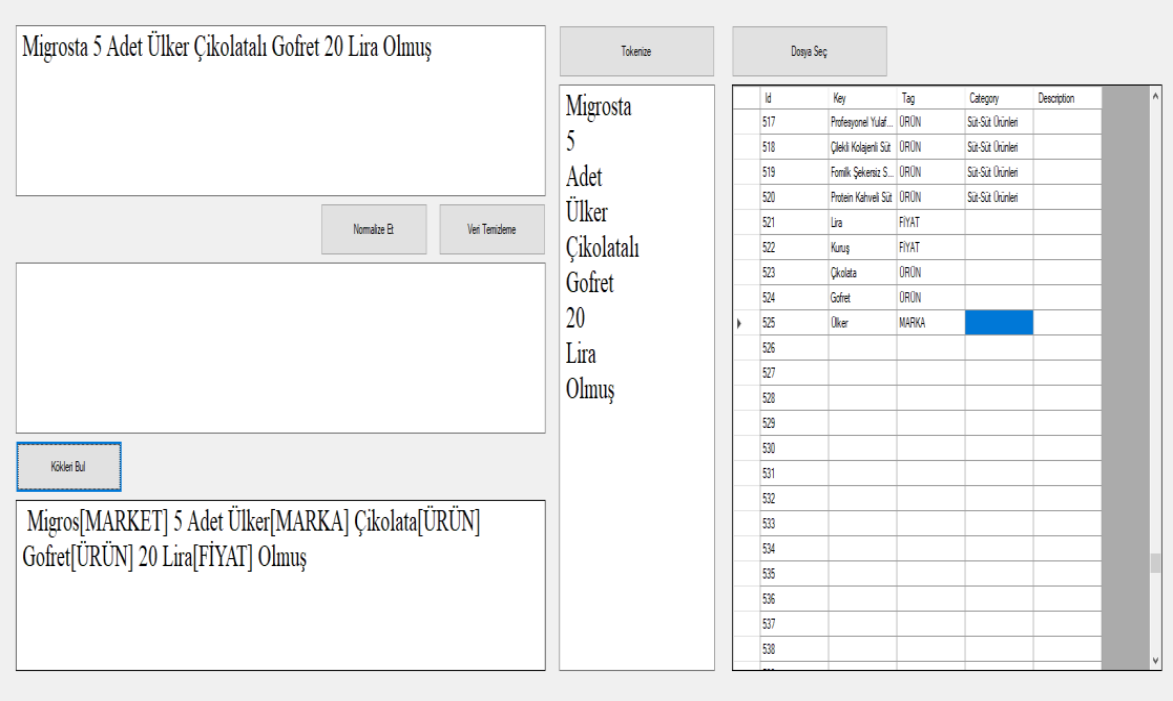
Cümle de market, marka, ürün, fiyat ve miktar etiketlerine ait kelimelerin hepsi sözlükte bulunuyorsa sınıflandırma işlemi başarılı şekilde yapılıyor ve ilgili kelimelerin etiketleme işlemi tamamlanıyor. Buna örnek olarak "Bime gittim 3 kg kuşbaşını 150 liraya aldım" cümlesinin normalizasyon, tokenize etme ve adlandırılmış varlık tanıma işlemlerine tabi tutarak etiketlemesi işlemi çıktısı Şekil 5.6'daki gibidir.



Şekil 5.6. Kelimelerin Etiketleme İşlemi

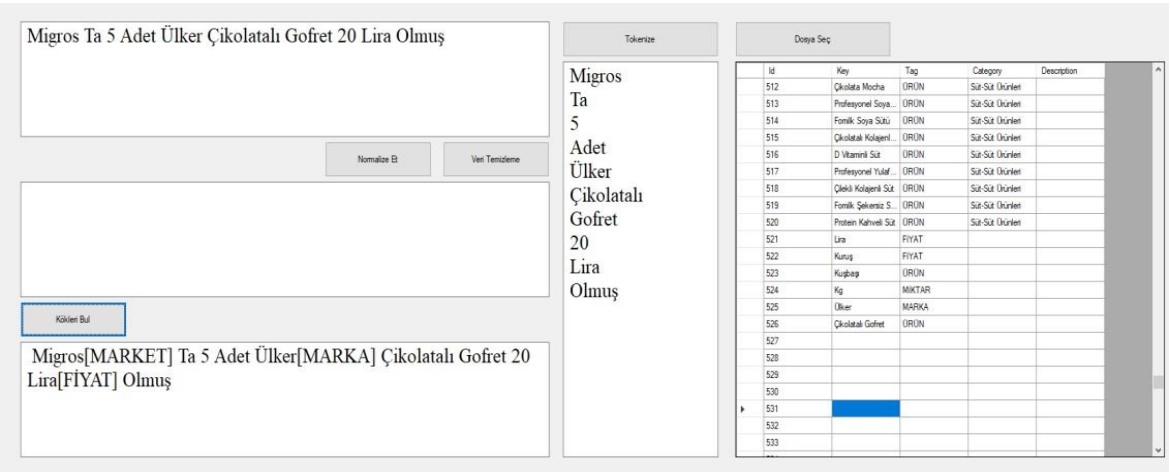
Bazı ürün isimleri iki veya daha fazla kelimeden oluşabiliyor. Eğer bu ürün ismindeki her kelime sözlükte ayrı bir ürün olarak tanımlı ise bir cümlede birden fazla ürün kategorisinde kelime olduğundan bunları kendi içinde de gruplayarak sözlükte tekrardan arama yapmaktadır. Örneğin; "Migros'ta 5 adet ülker çikolatalı gofret 20 lira olmuş." cümlesinde

“Çikolatalı Gofret” i incelediğimizde “Çikolata” ve “Gofret” sözlükte ayrı ayrı tanımlandıysa farklı sonuç, tanımlanmadıysa farklı sonuç üretmesi Şekil 5.7’ de gösterilmiştir.



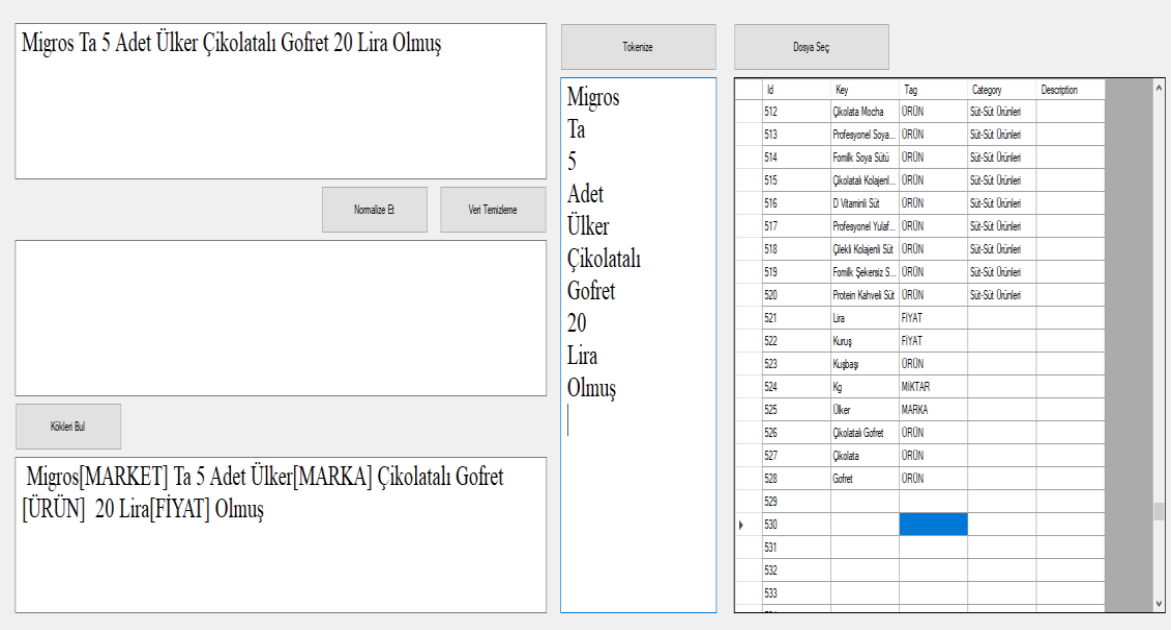
Şekil 5.7. Birden Çok Kelimeli Ürün İsminin, Ürün Olarak Sözlüğe Eklenmesi

Bu durumda “Çikolatalı Gofret” bir ürün olarak ifade edilmiş olsa da sistem bunu iki ayrı ürün olarak değerlendirmektedir. “Çikolata” ve “Gofret” kelimelerinin listede hiç olmayıp “Çikolatalı Gofret” tanımının yapıldığı örneğin ürettiği sonuç Şekil 5.8’deki gibidir.



Şekil 5.8. Birden Çok Kelimeli Ürün İsminin Sözlüğe Eklenmesi

Bu örnekte sistem ürün ismini bulamamaktadır. Çünkü cümlede her kelime sözlükte ayrı ayrı arandığından ürün isminin başlangıç ve bitiş noktası belirlenemiyor. Buna çözüm olarak eğer ürün ismi kendi içinde birden fazla ürün ismi içeriyorsa, o ürün isimleri sözlüğe ayrı ayrı eklendiğinde benzerlik oranlarına bakılarak ürün ismi belirlenebiliyor. Bu örnekte "çikolata", "gofret" ve "çikolatalı gofret" hepsi kendi başına bir ürün olduğundan bunların sözlüğe eklendiğinde üretilen sonuç Şekil 5.9' da gösterilmiştir.



Şekil 5.9. Birden Çok Kelimeli Ürün İsminin Etiketlenmesi

Kullanıcının yorum cümlesindeki ürün tek kelimeden oluşmadığı durumlarda başlangıç aşaması olarak cümle içinde geçen ürünlerin indexleri ve cümlede geçen ürün adetleri ilgili değişkenlerde tutulup, ürün adedi sayısı tespit edilmektedir. Bulunan ürün adedinin 2 ve üzeri olması durumunda ürün indexleri liste tipindeki değişkende tutulmaktadır. Ürünlerin indexleri bir değişkende tutulduktan sonra bir sonraki aşamaya geçilmektedir. Bu aşamada; kullanıcı cümlesinde geçen her bir ürün, daha önce hazırlanmış ortak kelimeler listesindeki tüm verilerde tek tek aratılmaktadır. Yapılacak aramalarda elbette kullanıcının bahsetmiş olduğu ürün ismine benzer ürünler bulunacaktır ve kontrol sağlanmadığı durumda kullanıcının bahsetmiş olduğu ürün yanlış tahmin edilecektir.

Adlandırılmış varlık tanıma sistemlerinin başarımı sistemin çıktısı olan metindeki tespit edilmiş varlık isimlerinin doğruluğu kontrol edilerek ölçülür. Çıktı, metindeki doğru tespit edilmiş varlık ismi sayısının tespit edilmiş toplam varlık ismi sayısına oranı tutma oranı (precision) olarak adlandırılır.

Burada "Çikolata" ve "Gofret" ayrı ayrı ürün olduğundan bunların bir arada bulunduğu ürün isimleri sözlükten bulunarak tutma oranı bulunmuştur. Tutma oranı için kullanılan denklem (5.1) gibidir;

$$tutma = \frac{doğru\ tespit\ edilmiş\ varlık\ ismi\ sayısı}{tespit\ edilmiş\ varlık\ ismi\ sayısı} \quad (5.1)$$

6. SONUÇ VE DEĞERLENDİRME

Makineler ile insanlar arasındaki etkileşimi doğal dil üzerinden yapılmasını sağlamak, doğal dilin anlamlandırılması, diller arasında otomatik çevrimin yapılması, doğal dil ile yazılmış metinlerden bilgi çıkarımı yapılması gibi görevler Doğal Dil İşleme (DDİ) araştırma alanının çalışma konularıdır.

Projenin sağlamayı amaçladığı en büyük hedeflerden biri insan psikolojisini olumlu etkilemek, mutluluğuna katkıda bulunabilmek ve emeğinin karşılığını olabilecek en yüksek seviyede almasını sağlayan bir aracı olabilmektir. Bunların yanında gelişmeye devam etmekte olan Türkçe Doğal Dil İşleme alanına da yenilikler katabilmek, karşılaşabileceğimiz sorunlara farklı bakış açıları ile bir çözüm getirebilmek hedeflenmektedir. Geçmişte yapılan birçok çalışmada veriler, ülke genelinde en sık kullanılan Migros, Bim, Carrefoursa gibi zincir marketlerin haftalık olarak fırsat ürünlerini sergilediği aktüel kataloglardan alınmıştır. Alışverişlerde tüm bunların bir arada kullanılması bir yere kadar fayda sağlayabilse de yeterli değildir. Aktüel Katalog yayınlayan birçok marketin yanında katalog yayınlamadan uygun fiyatlı ürün satışı yapan marketler de bulunmaktadır. Bir diğer hedef olan Türkçe Doğal Dil İşlemeye değer kazandırmak üzerine yapılan çalışmalarda ise market ve ürün verilerinin kullanıldığı bir çalışma bulunmamaktadır. Adlandırılmış varlık tanıma alanındaki mevcut çalışmalar geliştirmeyi düşündüğümüz çalışmaya katkı sunacak türden değildir. Bu nedenle çalışmamıza özgü bir sözlük oluşturulmuş olup çalışmalarımız bu özgün sözlüğe göre şekillendirilmiştir.

Bu çalışmada marketler ve marketlerdeki ürünler için adlandırılmış varlık tanıma sistemi geliştirilmiştir. Marketlerdeki ürün, marka ve fiyat bilgileri dilbilgisi, yazım ve imla kurallarına uymayan bir dille yazıldığından mevcuttaki DDİ yöntemlerinde yeterli performansa ulaşılamamaktadır. Literatürde bu içeriğe sahip metinle bir çalışma bulunmadığından bu tarz metinler normalizasyon yöntemleri ile mevcuttaki DDİ araçlarına uygun olacak formata dönüştürülmüştür ya da bu tarz metinler için yeni yöntemler geliştirilmiştir.

Veriler geliştirilen algoritma tasarımına uygun bir şekilde Key-Tag ikilisi biçiminde sözlük olarak tutulmuştur. Kullanıcıdan alınan cümle geliştirilen algoritma ile birçok aşamadan geçilerek DDİ teknikleri ile analiz edilmiştir. Uygulanan Nlp teknikleri ile kelime öbeği şeklindeki ürünlerin tespit ve analiz işlemini yapan algoritma geliştirilerek, ilgili cümlenin

anlamını kaybetmeden analiz işlemlerinin yapılmasına dikkat edilmiştir. Geliştirilen algoritma ile cümleye duygu katan ünlem, soru işareti gibi noktalama işaretleri ya da kelimelerin sonuna eklenen harf tekrarları göz ardı edilmiştir. Yani kullanıcının yorumda bahsettiği ürünü noktalama işaretleri ve eklerinden ayrıştırılarak anlamı korunmuştur.

Adlandırılmış varlık tanıma yapılırken kişi, yer, organizasyon gibi önceden tanımlanmış kategorilerin metin içerisinden çıkarılması hedeflenir. Bu çalışmada verilerde kullanılan etiketler market, ürün, marka, fiyat, ve miktar bilgileri olduğundan mevcut çalışmalardaki etiketlere uymamaktadır. Bu sebeple ilgili etiketlere uygun bir yapı kurmak amacıyla sözlük tabanlı adlandırılmış varlık tanıma işlemi için zengin bir sözlük oluşturulmuştur. TÜİK verileri de dahil edilerek hazırlanan sözlük, bu alanda daha sonra yapılabilecek çalışmalara katkısı olması ve alıntılanabilmesi adına kullanıma açılması, Türkçe Doğal Dil İşleme çalışmalarının yaygınlaşmasına katkı sağlayacaktır.

Sonuç olarak Crowdsourcing yöntemi kullanılarak gönüllü kullanıcıların ürün yorumları çalışma kapsamında oluşturulan sözlük ile ayrıştırılarak anlamlı ve düzenli hale getirilmiştir. Bu sayede kullanıcı dostu en ucuz ürünü bulan programın geliştirilmesi için gerekli olan verilerin hazır hale gelmesi sağlanabilecektir.

KAYNAKLAR

Abrahamson, S. R. (2013). *Crowd Storm: The Future of Innovation, Ideas and Problem Solving*, John Wiley, Sons, Inc., Hoboken, New Jersey.

Adalı, E. (2016). Doğal Dil İşleme. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 1-19.

Altunişik, R. (2017). Bir Veri Toplama Platformu Olarak Crowdsourcıng (Cs): Mechanıcal Turk Örneği Üzerinden Bir Literatür Taraması Ve Değerlendirme. *Uluslararası Yönetim İktisat ve İşletme Dergisi, 13*(5), 975-982.

Bayraktar, Ö. Temizel, T. T. (2008). Person Name Extraction From Turkish Financial News Text Using Local Grammar Based Approach. *In 23rd International Symposium on Computer and Information Sciences. ICSIS'08.* İstanbul.

Bicil, Y. (2016). Türkçe Metinden Konuşma Sentezleme, Yüksek Lisans Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Sakarya, 259140.

Brabham, D. (2008). *Crowdsourcing as a Model for Problem Solving: An Introduction and Cases*. Convergence: The International Journal of Research into New Media Technologies, 14(1), 75-90.

Çelikkaya, G., Torunoğlu, D., Eryiğit, G. (2013). Named Entity Recognition on Real Data. *Application of Information and Communication Technologies*, 1-5.

Chowdhury, G. G. (2003). Natural Language Processing. *Annual Review of Information Science and Technology. 37*(1), 51-89.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019, May 24). BERT: Pre- training of Deep Bidirectional Transformers for.

Dinçer, B. T. (2004). Türkçe için İstatistiksel bir Bilgi Geri Getirim Sistemi. İzmir: Ege Üniversitesi, Doktora Tezi, Ege Üniversitesi, Fen Bilimleri Enstitüsü, İzmir, 149708.

Eken, B. (2015). Kısa Metinlerde Varlık İsmi Tanıma., Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 389367.

Eryiğit, G. (2015). Itu Turkish Nlp Web Service., 1–4.

Eryiğit, G., Torunoğlu Selamet, D. (2017). Social media text normalization for Turkish. *Nat. Lang. Eng. 23*(6), 835–875.

Grishman, R., Sundheim, B. (1996). A Brief History. Coling'96. *Message Understanding Conference-6*, 466-471. Coling.

Güneş, A. (2018). Türkçe'de Varlık İsmi Tanıma.,Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 511546.

Jivani, M., Ganesh, A. (2012). A Comparative Study of Stemming Algorithms. *Oxford Handb. Cult. Psychol, 2*(6), 1930–1938.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2017). Natural Language Processing: State of The Art, Current Trends and Challenges.

Küçük, D., Arıcı, N. (2016). Türkçe için Wikipedia Tabanlı Varlık İsmi Tanıma. *Politeknik Dergisi*,, 19(3), 325-332.

Küçük, D., Steinberger, R. (2014). *Experiments to Improve Named Entity Recognition on Turkish Tweets*. Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM) ,71-78., Gothenburg, İsviçre, 2014.

Küçük, D., Yazıcı, A. (2009). Named Entity Recognition Experiments on Turkish Texts. *In Proceedings of the 8. International Conference on Flexible Query Answering Systems* ,524-535, Berlin, Almanya, 2009.

Küçük, D., Yazıcı, A. (2012). A hybrid named entity recognizer for Turkish. *Expert Systems With Applications*., 39(3), 2733-2742.

Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., & Lee, B. (2012, Ağustos). TwiNER: Named Entity Recognition in Targeted Twitter Stream.

Liu, X., Zhang, S., Wei, F., & Zhou, M. (2011). Recognizing Named Entities in Tweets. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 359- 367.

Nadeau, D., Sekine, S. (2007). *A Survey of Named Entity Recognition and Classification*. Lingvisticæ Investigationes, 30(1), 3-26, DOI: 10.1075/li.30.1.03nad

Oliveira, D., Laender, A., Veloso, A., Silva, A. (2013). FS-NER: A Lightweight Filter-Stream Approach to Named Entity Recognition on Twitter Data. *International World Wide Web Conference Committe*. May 2013, 597-604, DOI:10.1145/2487788.2488003

Özger, Z. B., Diri, B. (2016). Türkçe Dokümanlar İçin Kural Tabanlı Varlık İsmi Tanıma, *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5*(2).

Özkaya, S., Diri, B. (2011). Named Entity Recognition by Conditional Random Fields from Turkish informal texts. *2011 IEEE 19th Signal Processing and Communications Applications Conference (SIU)*. Antalya.

Özker, U. (2019). *Medium - Zemberek Doğal Dil İşleme*. https://medium.com/@ugrozkr_6539/zemberek-nlp-7add032881e9 (Ziyaret Tarihi : 25 Ekim 2022)

Ratinov, L., Roth, D. (2009). Proceedings of the Thirteenth Conference on Computational Natural Language Learning. *Proceedings of the Thirteenth Conference*., Boulder, Colorado, June, 2009

Rau, L. F. (1991). Extracting Company Names From Text. *In Proceedings Conference on Artificial Intelligence Applications of IEEE*., Florida, USA, February 24-28, 1991

Ritter, A., Clark, S., Etzioni, M., O., E. (2011). Named Entity Recognition in Tweets An Experimental Study. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, July 27-31, 2011

Sarı, Ö. C., Aktaş, Ö. (2018). Türkçe Ders Metinleri İçin Özelleştirilmiş Bir Varlık İsmi. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 11*(2), 2.

Şeker, G. A., Eryiğit, G. (2012). Initial Explorations on Using CRFs for Turkish Named Entity Recognition. *24th International Conference on Compuatational Linguistics*, Mumbai, Hindistan.

Şeker, S. E. (2015). Kitle Kaynak (Crowdsourcing). *YBS Ansiklopedi, 2*(2), 1-5.

Steinberger, R., Pouliquen, B., Kabadjov, M., Belyaeva, J., Goot, E. V. (2011). JRC-NAMES: A freely available, highly multilingual named entity resource. *Proceedings of Recent Advances in Natural Language Processing*, 104–110. Hissar, Bulgaria,

Sundheim, B. (1995). Overview of Results of the MUC-6 Evaluation., *In Sixth Message Understanding Conference (MUC-6): Processing of a Conference*, Colombia, Maryland, November 6-8, 1995, 13-31.

Tjong Kim Sang, E. F. (2002). CoNLL-2002, 155-158.

Toraman, Ç.,Can, F., (2010). Türkçe Haber Yazılarında Sosyal Ağların İncelenmesi.

Tür, G., Tür, D., Oflazer, K. (2003). *A Statistical Information Extraction System for Turkish*. Natural Language Engineering., 181-210.

Webster , J. J., Kit, C. (1992). Tokenization as the Initial Phase in Nlp.

Whitla, P. (2009). Crowdsourcing and its Application in Marketing Activities. *Contemporary Management Research, 5*(1), 15-28.

Yavuz, C. (2014). Tasarım ve Ürün Geliştirme Faaliyetlerinde Kitlesel Kaynak Kullanımı, Yükseklisans Tezi, Anadolu Üniversitesi Fen Bilimleri Enstitüsü, Eskişehir, 367768.

Yıldırım, A., Demirel, E. T. (2015). Ücret Tatmininin Yaşam Tatminini Belirleyici Etkisi Var Mı? Elazığ Banka Çalışanları Örneği. *Fırat Üniversitesi Sosyal Bilimler Dergisi*, 133-154.

Yılmaz, H., Yumuşak, S. (2021). Açık Kaynak Doğal Dil İşleme Kütüphaneleri. *İstanbul Sabahattin Zaim Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 3*(1).

KİŞİSEL YAYIN VE ESERLER

ER, F., Kul, S., Sayar, A., Product-Based Reasonable Price Suggestion System in the Market with Crowd-Source and Full-Text Search, *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Ankara, Turkey, 2021,1-6, DOI: 10.1109/HORA52670.2021.9461338.

ÖZGEÇMİŞ

İlk, orta ve lise öğrenimini Sivas'ta tamamladı. 2013 yılında girdiği Cumhuriyet Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü'nden 2017 yılında mezun oldu. 2019-2023 yılları arasında, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı'nda Yüksek Lisans öğrenimini tamamladı. Şu an da İstanbul'da özel bir şirkette kıdemli yazılım geliştirme uzmanı olarak görev yapmaktadır.