T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

SOSYAL MEDYA İÇERİKLERİYLE HAYAT BİLGİSİ VERİ TABANI OLUŞTURMA

Deniz ALTINAL

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Danışman

Doç. Dr. Mehmet Fatih AMASYALİ

Aralık, 2021

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

SOSYAL MEDYA İÇERİKLERİYLE HAYAT BİLGİSİ VERİ TABANI OLUŞTURMA

Deniz ALTINAL tarafından hazırlanan tez çalışması 20.12.2021 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Doç. Dr. Mehmet Fatih AMASYALİ

Yıldız Teknik Üniversitesi

Danışman

Jüri Üyeleri

Doç. Dr. Mehmet Fatih AMASYALİ, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Banu DİRİ, Üye
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üye. Murat ORHUN, Üye
İstanbul Bilgi Üniversitesi _____

Danışmanım Doç. Dr. Mehmet Fatih AMASYALİ sorumluluğunda tarafımca hazırlanan Sosyal Medya İçerikleriyle Hayat Bilgisi Veri Tabanı Oluşturma başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Deniz ALTINAL

İmza

İÇİNDEKİLER

KISALTMA LİSTESİ iv
TABLO LİSTESİ v
ŞEKİL LİSTESİ vi
ÖZET vii
ABSTRACT viii
1 GİRİŞ 1
1.1 Tezin Amacı ... 1
1.2 Literatür Özeti ... 1
1.2.1 OMCS ... 2
1.2.2 WordNet ... 4
1.2.3 ConceptNet ... 5
1.2.4 WordNet'in Türkçeleştirilmesi ... 7
1.2.5 ConceptNet'in Türkçeleştirilmesi ... 8
1.2.6 Kelime Görev ve Eklerinin Kullanılarak Kavramlar Arası İlişki Bulma ... 9
1.2.7 Veri Kümesini Büyültme ve Verileri Doğrulamak İçin Oyun Ara yüzü ... 9
1.2.8 Hayat Bilgisi Verileriyle Benzer Anlamlı Cümle Oluşturma ... 9
1.2.9 Çizge Kullanılarak QA Veri Kümelerinde Doğru Cevap Bulma ... 9
1.2.10 QA Veri Kümeleri için Modellere Bağlamsal Bilgi Ekleme ... 12
1.2.11 Wikidata'dan Sağduyu Bilgileri İçeren Altküme Çıkarma ... 13
1.3 Hipotez ... 14
1.4 Kullanılan Mimarilere Genel Bakış ... 16
2 METOT 18
2.1 Çalışmanın Temel Adımları ... 18
2.1.1 Otomatik Çeviri ... 18
2.1.2 Veri Kümesi Oluşturma ... 20
2.1.3 Kavramlar Nasıl İlişkilendirildi ... 22
2.1.4 Mimariler ... 23
2.1.5 Deneysel Sonuçlar ... 23
3 SONUÇ VE ÖNERİLER 28
KAYNAKÇA 30
TEZDEN ÜRETİLMİŞ YAYINLAR 31

KISALTMA LİSTESİ

ALBERT	A Lite version of BERT
BERT	Bidirectional Encoder Representations from Transformers
BNC	British National Corpus
CSKG	Commonsense Knowledge Graph
DEKCOR	Descriptive Knowledge for Commonsense Reasoning
KCR	Knowledge Chosen by Relations
KG	Knowledge Graph
LDOCE	Longman Dictionary of Contemporary English
LSTM	Long Short-Term Memory
MLM	Masked Language Models
MLP	Multilayer Perceptron
NLP	Natural Language Processing
OMCS	Open Mind Common Sense
PWN	Princeton WordNet
RNN	Recurrent Neural Network
RoBERTa	A Robustly Optimized BERT Pretraining Approach

TABLO LİSTESİ

Tablo 1.1 Kavramları ve ilişkileri bulmada Varlık ismi tanıma etiketlemesinin kullanım örneği. 15

Tablo 2.1 ConceptNet'ten Türkçe'ye çevrilen kavram örnekleri 19

Tablo 2.2 ConceptNet'ten Türkçe'ye çevrilen bazı kavramların derlemelerde bulunma sıklıkları 19

Tablo 2.3 ConceptNet'ten Türkçe'ye çevrilen kavramları içeren ilişkilerden derlemelerdeki frekansı 0 ve üzerinde olan bazı örnekler 20

Tablo 2.4 Twitter API'den elde edilen orijinal tweetler, Türkçe karakter sorunu çözülmüş halleriyle birlikte 20

Tablo 2.5 Kavramlar, onların aralarındaki ilişki tipleri ve o kavramların bulunduğu tweetlerden örnekler 21

Tablo 2.6 Eğitim ve Doğrulama Kümelerindeki Tweet ve Etiket sayıları 22

Tablo 2.7 İlişkiler için doğrulama kümesindeki doğruluk skorları 23

Tablo 2.8 Antonym için BERT sonuçları 24

Tablo 2.9 Synonym için BERT sonuçları 24

Tablo 2.10 PartOf için BERT sonuçları 24

Tablo 2.11 IsA için BERT sonuçları 24

Tablo 2.12 UsedFor için BERT sonuçları 25

Tablo 2.13 İlişkiler için sadece doğrulama kümesinde olan ve doğrulama esnasında program tarafından bulunan bazı ilişkiler 26

Tablo 2.14 10 milyon tweetten elde edilen bazı yeni kavram ve ilişkiler 27

ŞEKİL LİSTESİ

Şekil 3.1 Mimarinin özet yapısı gösterilmiştir 28

ÖZET

Sosyal Medya İçerikleriyle Hayat Bilgisi Veri Tabanı Oluşturma

Deniz ALTINAL

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Doç. Dr. Mehmet Fatih AMASYALİ

,

Hayat bilgisi veri tabanları, hayata dair kavramlar ve aralarındaki ilişkileri bulunduran yapılardır. Bu yapılardaki kavramlar ve ilişkilere, hayata dair tahminler yapmak ve akıl yürütmek için ihtiyaç duyulur. Bu çalışmada sosyal medya içeriklerinden yararlanılarak birbirleriyle çeşitli ilişkiler içeren Türkçe kavramların bulunduğu bir veri kümesi oluşturmak amaçlanmıştır. Problem bir varlık ismi tanıma problemi olarak ele alınmıştır. Eğitim için kullanılan sosyal medya içeriklerinde aralarında ilişki olan kavramlar ve ilişki türleri işaretlenmiştir. Veriler LSTM ve BERT tabanlı yöntemlerle modellenmiştir. Bu sayede büyük miktarda etiketsiz içerik modellere verilerek eğitim kümesinde olmayan yeni kavramlar ve aralarındaki ilişkiler bulunabilmiştir.

Anahtar Kelimeler: Hayat bilgisi veri tabanları, varlık ismi tanıma, LSTM, BERT, dil modelleri

ABSTRACT

Building Common Sense Database By Using Social Media Content

Deniz ALTINAL

Department of Computer Engineering

Master of Science Thesis

Supervisor: Doç. Dr. Mehmet Fatih AMASYALİ

,

Commonsense databases are structures that contain concepts about life and the relationships between them. The concepts and relationships in these structures are needed to make predictions and reasoning about life. In this study, it is aimed to create a dataset of Turkish concepts that contain various relations with each other by using social media contents. The problem is treated as an entity name recognition problem. Concepts and types of relationships that are related to each other are marked in the social media content used for training. The data were modeled using LSTM and BERT based methods. In this way, a large amount of unlabeled content was given to the models, and new concepts that were not in the training set and the relationships between them were found.

Keywords: Commonsense databases, named entity recognition, LSTM, BERT, language models

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

1
GİRİŞ

1.1 Tezin Amacı

Bilgisayarlar birçok konuda çıkarım yapabilmektedir. Yüz tanıma gibi. Ancak insanın yaşamına dair bir çıkarım yapan sistemler de gerekmektedir. Bu konuda çıkarım yapacak bir sistem kurulmak istenmektedir. Hayat bilgisi veri tabanları, kelimeler ve kavramlar arasındaki yaşama dair gerçeklikleri çeşitli ilişki yapıları ile birbirlerine bağlar ve bu yolla bilgisayarların bu gerçeklikleri yorumlayabilmesini basit hale getirir. Hayat Bilgisi verileri neredeyse tüm insanlar tarafından paylaşılan ve neredeyse tüm insanlardan makul bir şekilde beklenebilecek şeyleri tartışmaya gerek kalmadan algılama, anlama ve yargılama temel yeteneği olarak tanımlanmıştır. Açlık bir insanın yemek yemesine, yağmurda gezmek ıslanmasına neden olur gibi örneklenmiştir. Bu tür sağduyulu bilgiye sahip olmak, iletişimdeki boşlukları doldurmak ve metinden varlık tanıma ve bağlantı kurma, soru cevaplama ve planlama gibi görevleri yerine getirmek için hem insanlar hem de makineler için önemlidir. Bu tanımdan, onun örneğe dayalı bilgiden çok kavramsal bilgiyle ilgili olması, öncelikle yaygın olarak bilinen gözlemlerle ilgili olması ve genel alan bilgilerini hedeflediği sonucuna varılabilir. Zamanla birçok sistem geliştikçe, insanın hayatına yönelik ve ona dokunacak sistemlerin gelişmesi gerektiği düşünülmüştür. ConceptNet [1], bir bilgi kavramlar çizgesidir. İçinde eş anlamlılar gibi ilişki türlerine karşılık gelen kavramlar içerir. Bu çizgenin eksikliği ise sadece İngilizce dilinde yeterli kavram ve ilişki bulundurmasıdır. Bu araştırmada, Twitter verilerinden yararlanılmıştır. Bu verilerden ConceptNet gibi ilişkiler içeren ancak Türkçe kavramlardan oluşan bir veri kümesi oluşturulabilir mi sorusu üzerine odaklanılmıştır. ConceptNet'te bulunan ilişkiler Türkçe'ye çevrilmiştir. Aralarında çeşitli ilişkiler bulunan Türkçe kavramlar elde edilmiştir. O kavramların içinde bulunduğu tweetler kullanılarak başka farklı kavramlar içeren ilişkiler bulunmuştur. Bu yolla mevcut veri setini genişletmek amaçlanmıştır.

1.2 Literatür Özeti

"Cyc" ile başlayan bilgisayarlara sağduyu / akıl yürütme yetisini verme çalışmaları ilk zamanlarında belli bir grup araştırmacı tarafından anlaşabilecek ölçekteydi, CycL diye bir dil kullanılıyordu ve bu dili bilen sayısı az ve öğrenmesi zahmetliydi [2]. OMCS'le

herkes tarafından anlaşılabilecek bir yapı kurarak çok sayıda kişinin bilgisini sisteme aktarabilmesini sağlamıştır [3]. WordNet üzerine yapılan çalışmalarda ise eşkümeler ve bu eşkümeler arasındaki ilişkileri elde etmek amaçlanmıştır [4]. ATOMIC üzerine yapılan çalışmalarda ise boşlukları doldurunuz şeklindeki görevlerle bilgiler alınmıştır. ConceptNet'le birlikte ise WordNet'teki genel bilgi daha hayata dair hale getirilmiştir [1]. Türkçe için yapılan çalışmalarda WordNet'in [5] ve ConceptNet'in [6] Türkçeleştirilmesi, kelime görev ve eklerinin kullanılarak kavramlar arası ilişki bulma [7], veri kümesini büyültme ve verileri doğrulama yöntemi olarak oyun ara yüzlerinin kullanılması [8], hayat bilgisi verileriyle benzer anlamlı cümle oluşturma [9], cümle öğeleri bulma [10] konularında çalışılmıştır. Son zamanlarda ise İngilizce için kapsamlı ConceptNet kullanılarak sorular ve seçeneklerin olduğu SocialIQA gibi sağduyuyla ilgili sorular içeren veri kümeleri üzerinde çalışılarak modelin doğru seçeneği bulmasını sağlamak için bazı çalışmalar yapılmıştır. [11]. Bunu yaparken ConceptNet'in verisinin yetersiz olduğu düşünülmüş ve o kavramların tanımlarını da dahil ederek modellerin başarısını artırmak üzerine çalışılmıştır [12]. Bir çalışmada çocuk masallarının içerikleriyle sözlük türlerine bakılarak ilişki bulunmaya çalışırken [13] bazılarında ise gelişmiş Wikidata gibi veri kaynaklarından sadece sağduyu bilgisi içeren alt kümeleri elde edilmeye çalışılmıştır [14]. Bu bölümde bu çalışmaların ayrıntıları verilmiştir.

1.2.1 OMCS

Bu yapı üzerine çalışmalarda, web üzerinden katılan halktan katılımcıların desteğiyle bir hayat bilgisi veri tabanı elde etmek ve sıradan insan meseleleriyle meşgul olan ve bu bilgiyi tutan vekiller oluşturmak amaçlanmıştır. OMCS, genel halkı hedefleyen sağduyulu bir bilgi edinme sistemidir. Çeşitli meydana gelen etkinlikleri kullanarak gerçekleri, kuralları, hikayeleri ve açıklamaları toplayan bir web sitesidir. Bu çalışmada yenilikçi olarak webin kullanımı binlerce kişinin bu verileri elde etmek için kullanıla bilinmesi sağlanmıştır. Yapay zekâ topluluğundaki daha küçük grupların, genel halkın katılımını sağlayarak büyük miktarlarda bilgi gerektiren sistemler inşa etmesi artık mümkün olmuştur. Katılımcılar bazı önyargılar nedeniyle vereceği bilgiler kısıtlanmıştır. Katılımcıdan herhangi bir bilgi girmesi yerine bir bilgiyi çıkarmak için çeşitli aktiviteler oluşturulmuştur. Aktiviteye sistem bir olay veya durum vererek başlar kullanıcı ise bu olay ve durumu destekleyecek olay ve durumlar eklemiştir. Yani sistem X kişisi grip oldu denildiğinde onun doktora gittiğini katılımcının vermesi beklenmiştir. Veri tabanından yüzlerce bilgi türünü basit çerçeve temsillerine dönüştürmek için çıkarma kalıpları

geliştirilmiş her hayat bilgisi cümlesi belli bir dil bilgisel formülüne göre düzenlenmiştir. Katılımcılar faydaları katılımcılara göre değişmektedir. O yüzden farklı katılımcılar için farklı katılım sağlanacak aşama alanları sunulmuştur. Bu katılımcının istediği yapım aşamasına katkı sağlaması, kendi deneyimini daha iyi kontrol etmesini sağlamıştır. OMCS-1'de şablon kullanılmak istenmemiştir. Çünkü büyük ve kapsamlı bir şablon kümesi tasarlanamamıştır. OMCS-2'de yüz binlerce veri elde edildiğinden, şablonlar üretilmeye başlanmıştır. OMCS-2'de şablonlu ve şablonsuz kullanım da sağlanmıştır. Kullanıcı ilgisini çeken şablon bulana kadar aratmıştır. Bulunca da o şablonla ilgili sonuçları elde eder. Çıkarımları kabul etmeme şansı da verilmiştir. Şablonsuz da kullanıcı bir cümle kurabilmiştir o cümleyle ilgili şablon bulunduktan sonra çıkarımlar ona göre yapılabilmiştir. Önceki çalışmalarda Cyc'yi kullanmak için cycl dili ve onun ontolojisindeki binlerce terim öğrenilmesi gerekmiştir ve katılımcıların çoğu bunu öğrenmek istememiştir. Sisteme kısıtlı bir doğal dil ara yüzü sunmanın bir yolunu bularak bu sorun çözülmüştür. OMCS'de; kullanıcının altta yatan temsil ile tutarlı İngilizce formları seçebileceği açılır menüler ve kısıtlanmış bir İngilizce alt kümesi geliştirilip kullanılmıştır. Yeni sistem, katı doğal dil şablonlarını kullanmıştır, katılımcıların bu şablonları tasarlamasına izin vermiştir. Kelime anlamındaki belirsizliği ve girilen bilgiyi açıklığa kavuşturmak, geri bildirim sağlamak için analojik benzeşmeye dayalı çıkarım yapılmıştır ve katılımcıların bilgiyi ve dolayısıyla birbirlerini doğrulamasına izin verilmiştir. Ancak bireylerin bu veri tabanın geliştirilmesi ve girdikleri bilginin tüm yönleri konusunda dikkatli olmayabileceği düşünülmüş, bunu engellemek için katılımcıların birbirleriyle ilişki içerisinde olması sağlanmıştır. Birbirlerinin bilgi parçalarını organize bir şekilde onarabilmeli ve doğrulayabilmelidir. Web üzerindeki katılımcılar sayesinde kullanıcı dürüst olup olmadığı sorgulanmıştır. Büyük sayıda katılımcının olduğundan, yanlış bilgi veren kullanıcıları elle elemek zor olmuştur. Buna 2. Versiyonda, Kullanıcıların diğer kullanıcıların verdiği bilgilerini değerlendirerek yargılamalarına izin verilerek çözüm getirilmiştir. Girilen bilginin doğruluğunu ölçmek için puanlama yaparken kullanıcının güvenirliğine göre katsayı çarpımı yapılmıştır. Bilginin doğruluğu hem bilgi değerlendirilmesine hem de kullanıcı değerlendirmesine göre ölçülmüştür. Değerlendirme OMCS-1'de ise elle yapılmıştır. "Round objects roll with greater ease than other shape" ("Yuvarlak nesneler diğer şekillerden daha kolay yuvarlanır") gibi ifadeler içeren 3245 tane eşsiz parça değerlendirmeye alınmıştır. Bu ifadeler veri tabanının (400,151 parça) yaklaşık yüzde 1'ine denk gelmiştir. Bunlardan, yüzde 7'si yok sayılmıştır. Çünkü bu ifadeleri anlamlandırmak için, resim, hikâye gibi ek

materyal gerekmiştir, tek başlarına anlamları yoktur. Kalan 3009 tanesinden "it has a meaning" ("anlamı vardır") gibi parça sayılmayacak cümleler (370 adet) elenmiştir. Kalan 2639 ifadenin, genelliğine, doğruluğuna, taraflılığına ve anlamlılığına bakılarak 1'den 5'e kadar puanlanmıştır ve her puanlama için yüzdeler bulunmuştur, puanlanamayanlar ise NA olarak düşünülmüştür. Eylül 2000'den, Ağustos 2002'ye kadar web üzerinde çalıştırılarak 9296 kişiden 456.195 parça hayat bilgisi verisi toplandığı belirtilmiştir. OMCS'nin genel olarak aktifliği Ağustos 2016'da sona ermiştir. OMCS kullanılarak uygulamalar üretilmiştir. Bunlardan birincisi, arama motoru kullanıcının ulaşmaya çalıştığı hedefe göre arama yapmaktadır. Mesela kedim hasta denildiğinde veterinerin kedileri iyileştirildiği bilgisini vermektedir. İkinci uygulama ise ARIA'dır [15]. İnsanların fotoğraflarını yönetmektedir. OMCS-1'deki 50000 cümleyi kullanır. Nesne sınıflandırmaları, uzamsal ilişkiler, nesne amaçları, olaylar arasındaki nedensel ilişkiler ve nesneler ve olayların deneyimlenmesinden yola çıkarak çıkarım yapar. Mesela "x kişisi gelinin kız kardeşidir" cümlesiyle "Nedime genelde gelinin kız kardeşidir" cümlesinden yola çıkarak "nedimenin x kişisi olduğu" ortaya çıkarmıştır ve onunla ilgili fotoğrafa ulaşım sağlanmıştır. Üçüncüsü ise belgesel yapımcılarına çekim sürecinde yorum sunmak için kullanılmıştır. Ne zaman bir çekim yapıldığında bu çekim doğal dile dönüştürülür ve veri tabanına sorgu atılmıştır. Mesela "bir sokak artisti resim çiziyor" denildiğinde, "resim çizmenin en son aşaması fırçaları temizlemektir", "kırmızı boya kullanmak pahalıdır" gibi bilgilere ulaşılmıştır [3].

1.2.2 WordNet

Bu yapıda, geleneksel sözlüklerin arama şekline yani alfabetik aramaya ek olarak kavramsal bir arama şekli bulunmaya çalışılmıştır. WordNet'in içinde bulunduğu ilişki tipleri açıklanmıştır ve Morfolojik ilişkilerden bahsedilmiştir. Sözdizimsel sınıflandırmayı WordNet'e dayatmanın bedeli, geleneksel sözlüklerin kaçındığı fazlalıkları da içermesine neden olmuştur. Yani bazı kelimelerin birden fazla yer almasına neden olmuştur. Örneğin, geri gibi kelimeler birden fazla kategoride ortaya çıkmaktadır. Bu WordNet'in zayıf yönüyken, bu sözdizimsel kategorilerin anlamsal organizasyonundaki temel farklılıkların açıkça görülebilmesi ve sistematik olarak kullanılabiliyor olması güçlü yönüdür. WordNet makalesinin yazıldığı zamanda yaklaşık 48.000 eşanlamlı grup halinde düzenlenmiş yaklaşık 54.000 farklı sözlük girdisi içermektedir. İçerdiği ilişkiler Synonym, Antonym, Hyponym, Meronym gibi ilişkilerdir. Synonym ise en önemli ilişkidir. Birbirine benzer olan kelimelerin birbirleriyle yer

değiştirmesi birbirine benzemeyen kelimelerin yer değiştirmesinden daha olasıdır. Bu yüzden kavramların benzemesi yeterli görülmüştür. Antonym, zıtlıktır. Hyponym, Antonym ve Synonym aksine semantik yani kelime anlam ilişkisine dayandırılmıştır. Tabi olma/üstünlük, altküme/üst küme veya IsA ilişkisi olarak adlandırılmıştır. Alt kavram, üst kavramın özelliklerinin hepsini almıştır ve en az bir tane farklı özellik ortaya koymuştur. Akçaağaç, ağacın tüm özelliklerini taşımış ancak sertliği, yaprakların şekliyle diğer ağaçlardan ayrılmıştır. Meronym, parça-bütün ilişkisini gösteren semantik ilişki olarak kabul edilmiştir. Kelimenin hangi ağda nerede bulunduğunu bilmek, kelimenin anlamını bilmenin önemli bir parçası olmuştur. Morfolojik ilişkiler ise kelimelerin formları arasındaki sözlüksel ilişkilerin önemli bir sınıfıdır [4].

1.2.3 ConceptNet

Bu yapıda yenilikçi bir yaklaşım olarak bileşik kavramlar arasındaki önceki çalışmalardan daha zengin bir semantik ilişkiler kümesine odaklanmıştır. WordNet ve Cyc bilgi mühendisleri tarafından elle yapılmış iken ConceptNet, halk tarafından oluşturulmuş OMCS'in içerdiği İngilizce cümlelerden otomatik olarak yararlanarak oluşturulmuştur. Sadece bilgili mühendisler değil çocuklar bile bu bilgiyi sağlamak için kullanılmıştır. 14000 Web katılımcısı bunu sağlamıştır. 700000 İngilizce hayat bilgisi cümlesi kullanılmıştır. Bu cümleleri üzerinden 300000 kavram elde edilmiştir. OMCS'nin İngilizce cümlelerinden ConceptNet'in ikili ilişki ifadeleri eşlenmiştir. MontyLingua araç takımına, OMCS'deki İngilizce cümleler verilmiştir. Bu sayede, cümlelerin içeresindeki fiiller("buy"), isim kalıpları ("laptop computer"), edat kalıpları ("in restaurant") ve sıfat kalıpları ("very sour") bulunmuştur. Bu bulunan yapılar kullanılarak düğümler elde edilmiştir ve sıra sınırlaması konulmuştur. Örneğin, fiiller, isim kalıpları ve sıfat kalıplarından önce yer alması gerektiği düşünülmüştür. Sonra normalize edilmiştir. Mesela "Is", "are", "were" ifadeleri "be" ye dönüştürülmüş, "apples" ifadesi "apple" a dönüştürülmüştür. Normalize edildikten sonra, tekrarlanan ifadeler birleştirilmiştir. Kaç defa bulunduğu not edilmiştir. IsA hiyerarşik ilişkisi, bilgiyi alt düğümlerden üst düğüme buluşsal olarak "kaldırmak" için kullanılır. Yani "elma, şeftali, muz tatlıysa" ve "hepsi meyveyse" o zaman "meyveler tatlıdır" sonucu çıkarılmıştır. WordNet'te bulunan ilişkilere ek olarak EffectOf (nedensellik), SubeventOf (olay hiyerarşisi), CapableOf (yetenek), PropertyOf, LocationOf ve MotivationOf (etki) gibi ilişkiler eklenmiştir. ConceptuallyRelatedTo, ThematicKLineSuperThematicKLine'dan 1.25 million, IsA, PropertyOf, PartOf,

MadeOf, DefinedAs'dan 52 000, CapableOf'dan 104 000, PrerequisiteEventOf, FirstSubeventOf, SubeventOf, LastSubeventOf'dan 38 000, LocationOf'dan 36 000, EffectOf, DesirousEffectOf'dan 17 000, UsedFor, CapableOfReceivingAction'dan 115000, MotivationOf, DesireOf'dan 34 000 tane ilişki elde edilmiştir. Bu veri tabanı kullanılarak çeşitli uygulamalar geliştirilmiştir. Commonsense ARIA, kullanıcının girdiği e-postaya göre kullanıcının hikayesine göre fotoğraf paylaşabilmiştir. Açıklamalar ve kullanıcının hikayesi (ör. "gelin" ve "düğün") arasındaki anlamsal boşlukları kapatmıştır. GOOSE, acemi kullanıcılar için hedef odaklı bir arama motorudur. MAKEBELIEVE, bir kişinin sistemle etkileşimli olarak bir hikâye icat etmesine izin veren hikâye üreticisidir. Hikâye çizgileri oluşturmak için nedensel projeksiyon(x->y ve y->z ise x->z gibi) zincirleri oluşturmak için ConceptNet öncülünü kullanmıştır. GloBuddy ise verdiği cümlenin konumuna göre diğer kavramları gönderebilmiştir. Mesela ben restorandayım denildiğinde, insanlar, garson, sandalye, yemek gibi kavramları göndermiştir. AAA, bir kişinin hedefleri ve arzuları hakkında akıl yürütmek için ConceptNet'i kullanarak Amazon.com'dan ürünler önermiş ve tahmin edilen zevklerinin bir profilini oluşturmuştur. OMAdventure, oyuncuların dinamik olarak oluşturulmuş grafiksel bir dünyada gezindikleri etkileşimli bir çöpçü avı oyunudur. Emotus Ponens, altı temel duygu kategorisi kullanarak metni sınıflandırmak için sağduyudan yararlanan metinsel bir duygulanım algılama sistemidir. EmpathyBuddy, yazara bir ifade yüzü aracılığıyla otomatik olarak duygusal geri bildirim veren bir e-posta istemcisidir. Overhear, gündelik konuşmaların konularını özetlemek için sağduyu kullanan konuşmaya dayalı bir konuşma anlama sistemidir. Bubble Lexicon, dinamik bir anlam açıklaması veren bağlam merkezli bir bilişsel sözlüktür. ConceptNet, sözlüğün bağlantıcı-anlamsal ağını dünya semantik bilgisi ile önyükler. LifeNet, günlük birinci şahıs insan deneyiminin olasılıksal bir grafik modelidir. ConceptNet'i benmerkezci önermeler şeklinde yeniden formüle etmiştir. Bunları geçiş olasılıkları ile birbirine bağlayarak oluşturulmuştur. SAM, çocuklarla bir bebek evinde oynarken birlikte hikayeler anlatan somutlaşmış bir hikâye anlatma ajanıdır. ConceptNet, SAM'ın söylem geçişleri seçimini yönlendirmiştir. What Would They Think, e-postalar, web günlükleri ve ana sayfalar gibi kişisel metinleri analiz ederek bir kişinin kişiliğini ve tutumlarını otomatik olarak modellemiştir. ConceptNet'in analoji oluşturma özelliği, tutum tahminini daha sağlam hale getirmek için kullanılmıştır. Commonsense Predictive Text Entry, bir kullanıcının cep telefonu metin mesajının içeriğini anlamak ve olası kelime tamamlamalarını önermek için ConceptNet'ten yararlanmıştır. Commonsense Investing, ConceptNet'in bir kişinin amaç ve isteklerini

temsilini bir uzmanın teknik terimleriyle eşleştirerek kişisel yatırımcılara finansal kararlarda yardımcı olmuştur. Metafor ise çocukların programları İngilizce kullanarak tanımlamalarına izin vererek programlama fikirlerini keşfetmelerini kolaylaştırmıştır. ConceptNet, doğal dil girdisinin programlı-anlamsal yorumu için kullanılan "sağduyu sınıfları" nın programatik bir kitaplığını sağlamıştır [1].

1.2.4 WordNet'in Türkçeleştirilmesi

Bu çalışmada, Euro WordNet projesindeki ana kavramlar Türkçe'ye çevrilmiştir. Bu temel kavramlar, hiyerarşide üst sıralarda yer alan ve birçok eş anlamlısı olan önemli kavramlardır. Bu kavramlar seçilirken, diğer dillerde de ortak olarak bulunması istenmiştir. 1310 tane ortak temel kavram elde edilmiş diğer 6 balkan WordNet'lerdeki çevirilerde bulunan kavramlar kümesiyle karşılaştırma yapılarak maksimum şekilde uyuşması sağlanmıştır. Bu yolla yapılan çevirinin tam ve hızlı olması sağlanmaya çalışılmıştır. Synsetler, içinde belli bir bağlamda, birbirleriyle yer değiştirildiğinde, doğruluk değerini değiştirmeyen eş anlamlıları içeren kümelerdir. İlk fazda oluşturulan kümede bulunan, 1310 ortak temel kavramın içinde bulunduğu 1310 synset Türkçe'ye çevrilmiştir. Makine tarafından okunabilen Türkçe sözlükten elde edilenler, perl'le parse edildikten sonra 10846 synonym ve 235 antonym elde edilmiştir. Son ek olarak gillerle bitenler de eklendikten sonra 889 tür ve çeşit anlamına gelen hyponym-hyperonym otomatik olarak çıkarılmıştır. Synset sayısını çoğaltmak ve 5000 taneye çıkartmak için çeşitli synset elde etme yöntemleri önerilmiştir. 240 gaps fikrinde, ilk fazda bulunan synsetlerin hyperonym karşılığı bu ilk fazda bulunan synsetlerin içinde bulunmadığı için, bunların karşılıklarını elde etmek amaçlanmıştır. Bunun için, boşlukları, karşılıklarıyla doldurmak için, yani hyperonym karşılığı olmayan ve ilk fazda bulunan synsetlerin karşılıklarını elde etmek için bu synsetlerle hyperonym olarak eşleşen PWN'deki en üstteki girişler kullanılmıştır. Bu girişlerin, boşlukları doldurulmasıyla bağlantısı olmayan serbest düğümler yok edilmiştir. 1228 ek synset elde etme fikrinde ise amaç elde edilen synsetlerin toplamının diğer dillerde yapılan çalışmalarda elde edilen verilerle maksimum şekilde uyuşmasını sağlamak ve bunlardan elde edilen ilişkilerin maksimum olması amaçlanmıştır. Bunun için, önemli kavramları elde etmek amaçlanmıştır. İlk kaynak olarak LDOCE sözlüğü kullanılmıştır. İkinci kaynak olarak sık kullanılan kelimeleri almak için ise BNC kullanılmıştır. Bu iki kaynaktan elde edilen kelimelerin kesişimi bulunmuştur. Üçüncü aşamada ise hedef 8000 synset oluşturmak olmuştur. Bazen bir dildeki morfolojik ilişki diğer dilde bazen bulunur bazen bulunmaz. Bulunduğu

kısmında mesela Türkçe'de morfolojik ilişkisi olan iki sözcüğün İngilizce karşılıkları da morfolojik ilişki içeresindeyse bunu İngilizce'deki morfolojiden çıkarmak mümkündür. Aksi bir durumda yani karşılıkları İngilizce'de morfolojik bir ilişki içerisinde değilse bu ilişkiyi bulmak normal şartlarda zor iken Türkçe 'deki morfolojik ilişkiye güvenerek o kelimeleri ilişkilendirmek işi daha hızlandırmış ve daha güvenilir kılmıştır. Bu da değişik dillerdeki WordNet'lerin birbirlerini geliştirebilir hale getirmiştir. Bazı become, cause gibi ilişkiler bu yolla elde edilmiştir. Her ne zaman yeni synset eklendiğinde boşlukları da ele alınmıştır. İki kavram birbirine benziyor ve aralarında bir ilişki varsa ilk kavram WordNet'te var ve diğeri yoksa sarkık ilişki oluşur bu yüzden o diğer kavramın da WordNet'e eklenmesi gerekmiştir. WordNet'ten alınan her ilişki manuel, yarı otomatik ve otomatik bir şekilde doğrulanmıştır. Sözlüksel üyelerin kapsamına bakılmıştır. Kapsamı ölçmek için 2 tane farklı eşsiz kelimelerden oluşan kelime listesi oluşturulmuş, ilk kelime listesi George Orwell'ın 1984 romanının Türkçe çevirisinden elde edilmiş ve ikinci kelime listesi ise genel amaçlı bir derlemi temel alınmıştır. İşlev yapısındaki kelimeler atılmıştır. Morfolojik çözümleme yapılmış, sözlük türleri etiketlenmiştir. Türkçe kelime ağının 1984 derlemi ve genel amaçlı derlem ile kesişimi sırasıyla %87,40 ve %86,45 olarak bulunmuştur. Oluşturulan uygulamalardan Meaning-to-Word, girilen tanıma ait ilgili en yakın kelimeleri vermiştir. Örneğin "Evlenmemiş kimse" girildiğinde "bekar" ifadesi elde edilebilmektedir. Bu sistem ölçülürken rastgele 50 tane anlama karşılık kelime verilmiş ve bunların doğruluğu hesaplanmıştır. Başarı yüzdesi Türkçe WordNet'in synonym bilgisi eklendiğinde % 60'tan %68'e yükselmiştir. Hyperonym için 11251, holo member için 946, holo part için 1423, holo portion için 176, causes için 100, be in state için 577, near antonym için 1400, Subevent için 127, also see için 270, verb group için 896, category domain için 384, ve toplamda 17550 semantik ilişki elde edilmiştir. İstatistikler de şu şekildedir: Synsets 11628, Synset Members 16095, Average Synset Size 1.38, Unlexicalized Concepts 987, Definitions 4913 [5].

1.2.5 ConceptNet'in Türkçeleştirilmesi

Bu çalışmada yenilikçi yaklaşım olarak verilerin Türkçe olarak oluşturma fikri ele alınmıştır. ConceptNet ve WordNet'teki veriler otomatik Türkçe'ye çevrildikten sonra kullanılmıştır. Çalışmada dört değişik kaynaktan alınan 475407 adet kavram ve aralarında kırk değişik ilişki türüne ait 1089230 adet ilişki elde edilmiştir. Çalışmanın zayıf yönü, mevcut çizgelerden elde edilen ve çevrilen bilgiyle sınırlı kalmasıdır [6].

1.2.6 Kelime Görev ve Eklerinin Kullanılarak Kavramlar Arası İlişki Bulma

Bu çalışmada, Türkçe tümcelerin öğelerini kullanarak Türkçe hayat bilgisi veri tabanı oluşturmak amaçlanmıştır. Cümle öğelerinden yola çıkarak Bilkent Üniversitesi Bilgiye Erişim Grubu tarafından derlenmiş metinleri kullanarak veri tabanı elde edilmiştir. İngilizce'den Türkçe'ye çevirerek kavramları oluşturmak yerine, tümcelerden yola çıkılmıştır. Zayıf yanı aynı kelime çözümlemelerinde olan tümcelerin farklı öğe dizilişlerine sahip olabilmesidir. Ayrım anlamsal olabilir. Ama bu ayrım yapılamamıştır. Bu durum, çalışmada öğeleri ayırma işleminde istatiksel bir modelin uygulanmayışı ve Zemberek kütüphanesinin kelimeleri tümce içinde değil kendi içinde değerlendirmesinden kaynaklanmıştır [7].

1.2.7 Veri Kümesini Büyültme ve Verileri Doğrulamak İçin Oyun Ara yüzü

Bu çalışmada, ConceptNet, WordNet gibi ilişkisi çok ama güvenirliği düşük olmasına rağmen Google Translate'le bu ilişkiler elde edilmiş ve bir oyun tasarlanarak bu güvenirliği yükseltmeye yani kullanıcının verdiği kararlarla hem yeni kavramlar elde etmek hem de var olanların güvenirliğini ölçmek için bir oyun tasarlanmıştır. Bildiklerim, Araştır, Tahminlerim gibi bölümler içermiştir. Tahminlerim kısmında var olandan başka bir kavram elde etmeye çalışma fikri sunması açısından önemlidir. Oyun sayesinde kullanıcılara güvenip, kendi kendine gelişebilen bir yapıya kavuşturulmuştur [8].

1.2.8 Hayat Bilgisi Verileriyle Benzer Anlamlı Cümle Oluşturma

Bu çalışmada, var olan bir hayat bilgisi veri tabanında sisteme girilen bir cümleden başka bir cümle üretilmeye çalışılmıştır. Veri tabanın içinde yer alan ilişkilere kullanıcılar tarafından belli bir puan verilmiştir. Bu puanların ortalaması alınarak ilişkinin doğruluğu hesaplanmıştır. Veri tabanını zemberek kelime çözümleyicisiyle bağlantısını kurarak yüksek puanlı olanları ele alarak bu veri tabanındaki karşılığını yollayarak cümleler elde edilmiştir. İlk defa Türkçe hayat bilgisi veri tabanı kullanılarak bir cümleden benzer ve yakın anlamlı yeni cümleler elde edilmesi sağlanmıştır [9].

1.2.9 Çizge Kullanılarak QA Veri Kümelerinde Doğru Cevap Bulma

Bu çalışmalarda, ConceptNet gibi bilgi çizgeleri kullanmadan, eğitilmiş dil modellerinin akıl yürütmede, yetersiz olduğu düşünülmüştür. Bu yetersizliği gidermek için çalışmada, bilgi çizgeleri hem dolaylı olarak hem de doğrudan, modellere sokulmuştur. Amaç, Bir olayın C bağlamı ve buna karşılık gelen bir Q sorusu verildiğinde, A = (A1, A2, A3) yanıt kümesinden doğru olanı seçmektir. Örneğin, bağlam cümlesinin, "Cameron decided to

have a barbecue and gathered her friends together" olduğu, "how would Others feel as a result" sorusuna, model doğru cevap olarak "like attending." cevabını vermiştir. Bunun için ConceptNet'te barbecue ve attend arasında bağlantı bulunmuştur ("barbecue–isa→dish–synonym→serve←synonym–attend"). SocialIQA veri setindeki bağlam, soru ve cevapların içindeki kelimeler kullanılarak onları bilgi çizgesi olan ConceptNet'te taratarak ikililer oluşturulmuştur. Her veri noktası için k tane en iyi ikililer seçilmiştir. Bu seçim, ilişkinin ağırlığı ve sorgunun sıklığına göre yapılmıştır. Dolaylı yaklaşımda ön eğitim yapılırken, doğrudan yaklaşımda ince ayarın durumuna göre dinamik bilgi oluşturulmuş ve ilgi mekanizması kullanılmıştır. Bu modellerin başarısı SocialIQA veri kümesi üzerinde hem eğitim kümesinin tümünü hem de bir kısmını kullanarak ölçülmüştür. Dolaylı yaklaşımda, bilgi çizgeleri, ön eğitimde kullanılmıştır. İkililer elde edildikten sonra, bunlar metin haline getirilmiştir. Örneğin (barbecue, hascontext, cooking) ikilisi, "barbecue is a word used in the context of cooking." haline getirilmiştir. ATOMIC kullanırken, "PersonX öğle yemeğinde buluşuyor" gibi bazı boşluklar ve bilinmeyen isimler olduğundan, cinsiyet önyargısını önlemek için PersonX ve PersonY'yi iki farklı ortak soyadıyla değiştirilmiştir ve ardından, önceden eğitilmiş BERT-large'ın MLM'si kullanılmıştır ve boşluklar doldurulmuştur. RoBERTa tabanlı modelle ve MLM ile maskeleme yaparak eğitilmiştir. Maskeleme yapılacak yer, düğümün başı veya kuyruk kısmı olarak seçilmiştir. ("barbecue" veya "cooking"). Bilgi çizgelerinden, dolaylı olarak öğrenilen sağduyu bilgileriyle önceden eğitilmiş modeller eğitilmiştir. Bilgi çizgelerini doğrudan kullanan yaklaşımda ise, elde edilen ikililer belleğe alınmıştır her girişte değişecek şekilde ve girişin duruma göre dinamik olarak değiştirilmiştir. Model bunların hangilerinin daha önemli olduğunu ve kullanması gerektiğine karar vermiştir. Bunun için, KG ilgi mekanizması kullanılmıştır. Bu yaklaşımda her ikili bir segment olarak kabul edilmiş ve her segmente ayrı bir ağırlık puanı verilmiştir. K tanesini birleştirip SocialIQA girdisine göre oluşturulmuş RoBERTa'nın son katmanına verilerek, KG ilgi cümle gösterimi elde edilmiştir. Bunu yaparken RoBERTa'nın verilen her adaya giriş dizisiyle ve k tane ikiliyi ifade eden gösterim kullanarak bir softmax hesabı yapılmıştır. Bu softmax hesabı çizge ilgi cümlesi gösterimini elde etmek için kullanılmıştır. İkilileri birleştirmek için ve RoBERTa'nın temsil alanı göndermek için çeşitli yöntemler kullanılmıştır. Bir yaklaşımda, TransE kullanarak, ConceptNet'e dayalı önceden eğitilmiş bilgi çizgesi temsilleri kullanılmıştır, başka yaklaşımda ConceptNet tarafından uyarlanan önceden eğitilmiş kelime temsilleri kullanılmıştır. Bu yaklaşımda, bir kelimenin temsili çizgedeki komşularınınkilere yakın olacak şekilde ayarlanmıştır. Bir başka yöntemde de universal

sentence encoder kullanılmıştır. Bu encoder, metinleri vektörlere dönüştürmüştür. Bu top k ikililerin gömülmeleri, lineer dönüştürücü kullanılarak dönüştürüldükten sonra birleştirilerek bilgi gösterimi elde edilmiştir. Kaynaştırma tabakasında ise, ikililerin k tanesini birleştirip SocialIQA girdisine göre oluşturulmuş RoBERTa'nın son katmanıyla, çizge ilgi cümlesi gösterimi birleştirilmiştir ve max pooling kullanarak en yoğun olan tabaka bulunmuştur, MLP kullanılarak her cevap için softmax hesabı yapılmıştır ve doğru seçenek bulunmak istenmiştir. Bu çalışma farklı olarak ConceptNet gibi zengin kaynakları kullanarak model oluşturmak hedeflenmiştir. Böylece cevaplarda ve mevcut bağlamda olmayan bilgilere daha derin ve geniş anlayış sağlanmıştır. SocialIQA'daki bilgilerden nasıl demetler elde edildiğinden bahsedilmiş ve bunların neye göre sıralandığı belirtilmiştir. ConceptNet'teki demetlerden cümle eldesi sağlanmıştır. ATOMIC deki veriler için maskeleme doldurma yapılmıştır. Bilgi çizge ilgi gösteriminin nasıl elde edildiği, bilgi demetlerinin ne şekilde kodlanıp ne tür temsillerin üretildiği, kaynaşma katmanlarının oluşumundan bahsedilmiştir. Güçlü yönü, SocialIQA'daki bilgileri ConceptNet'teki hoplardan elde etmek ve elde edilen bilgileri ilişkinin ağılığı ve sorgunun idf skorlarının çarpımıyla sıralamanın haricinde her girişte değişen cachedlarla ve RoBERTa'yla belli temsillerle denenmesiyle ve kaynaşma katmanı oluşturmasıyla, max-pooling'le yoğunlaştırmasıyla ve MLP gibi her düğümün farklı ağırlığı önceliği olan bir yöntem kullanmasıyla elde edilen SoftMax skorlarıyla demetler arasında duruma göre sıralama yapılabilmesi olmuştur. Zayıf yönü ise bilgi çizgelerinden bilgi alırken uygulanan algoritmaların yetersiz olması mesela x kavramına ulaşırken en kısa yoldan erişim sağlanmayışıdır. Tüm veride, ConceptNet-attention-USE yüzde 79.2'yle en iyi çıkmıştır. ConceptNet-pretrain ise yüzde 76.8'le en kötü çıkmıştır. Few-Shot Learning kullanılmıştır. Limitli veriye göre yöntemlerin doğruluğu ölçülmüştür. Modeller SocialIQA'nın yüzde 5,10,20 sine göre ince ayar edilmiştir. ConceptNet-attention-USE, ATOMIC-pretrain, RoBERTa yöntemleri denemiştir. RoBERTa ve MLM birlikte kullanımı temel olarak kabul edilmiştir. ATOMIC pretrain yüzde beşte daha iyi çıkmıştır. ConceptNet-attention-USE yüzde 5'te en kötü olmuştur ama diğer iki yüzdede temelden daha iyi çıkmıştır. Yüzde 20'de ise ATOMIC-pretrain en iyi, temel yaklaşım ise en kötü olmuştur. Bert Modelin, eğitim kümesinin tümüyle eğitildiğinde yüzde başarısı olarak 63.3 çıkması RoBERTa modelinin, bazı sağduyuları ön eğitilme sürecinde zaten öğrenmiş olmasına bağlanabilir [11].

1.2.10 QA Veri Kümeleri için Modellere Bağlamsal Bilgi Ekleme

Bu çalışmada, bilgisayar tarafından, QA veri kümelerinde bulunan sorulara karşılık olarak verilen seçeneklerden, doğru seçeneğin bulunması için ConceptNet gibi bilgi çizgelerinin kullanımının tek başına yetersiz olduğu düşünülmüştür. Bu yüzden çizgelerde taranmış ilgili kavramlar, Wikitionary'deki tanımlarıyla birlikte modele sokulmuştur. Bu yolla doğru seçeneği bulmaya çalışan bu modellerin başarısı artırılmak istenmiştir. Modellerin, kavramlar ve ilişkileri bu tanımlarla daha iyi anlaması mümkün olmuştur. Wikitionary'de, bir kavrama karşılık birden fazla tanım vardır. Çalışmada birincil olan tanımlar kabul edilmiştir. Soru ve seçeneklerde bulunan ilgili kavramların tanımları elde edilmiştir. Çalışma, DEKCOR modelini önermiştir. Bu modellemede, soru ve seçeneklerde bulunan kavramlar arasında bağlantı varsa bunlar kabul edilmiştir. Aralarındaki bağlantıyla birlikte kavramlar, üçlüleri(eq,r,ea) oluşturmuştur. Eğer bir seçenekteki bir kavramın sorudaki kavramlardan birine bağlantısı yoksa, bu kavramın geçtiği tüm üçlüler bulunmuştur. Aralarındaki alaka puanları hesaplanmıştır. Bu puanı hesaplarken üçlünün ConceptNet'teki ağırlığı ve ilgili ilişki tipinin ağırlığıyla çarpılmıştır. Bu puanlardan yüksek olanlara karşılık gelen üçlüler seçilmiştir. Üçlülerdeki kavramlar belirlendikten sonra, tanımlar elde edilmiştir ve bunlar ilgili soru seçeneklerle ALBERT modeliyle modellenmiştir. İlgi ve softmax katmanlarıyla skor elde edilmiş, doğru seçenek bulunmaya çalışılmıştır. Örneğin, soru "Where is Mona Lisa", bir seçenek "Louvre" ise, Sorudan "Mona Lisa" kavramı, seçenekten ise "Louvre" kavramı çıkartılır. Aralarındaki bağıntıya bakılıp "Mona Lisa,is in,Louvre" gibi bir üçlü oluşturulur. Bu üçlemede bulunan kavramlar Wiktionary'de aratılır ve "Mona Lisa, Leonardo da Vinci'nin tablosudur", "Louvre Fransa Paris'teki en ünlü müzedir" gibi bilgiler elde edildikten sonra birleştirilip modellenmiştir. Model CommonsenseQA veri kümesi üzerinde değerlendirilmiştir. 12102 örnek kullanılmıştır. 9741'i eğitim kümesi, 1221'i geliştirme kümesi ve 1140 tanesi test kümesi olarak kabul edilmiştir. Kullanılan bu DEKCOR modellemesinde, test kümesi üzerinde son yöntemlerin en iyisi Albert+KCR'den yüzdeyi 80.7 çıkararak, tek modellemede yüzde 1.2, çoklu modellemede yüzdeyi 83.3'e çıkararak yüzde 3.8 daha iyi çıkmıştır. Geliştirme kümesi üzerinde yapılan çalışmalarda, üçlemesiz ve tanımlarsız başarı yüzdesi 80.1 iken, Wiktionary'den kavramlarla ilgili tanımlar elde edilerek modele yollanması CommonsenseQA üzerinde doğruluğu yüzde 2.7 artırmıştır, ConceptNet'ten üçlü elde

edilip kullanılmasıyla yüzde 4.6 katkı sağlamıştır ve yüzde 84.7 elde edilmiştir. Bu sonuç, tanım bağlam bilgilerinin dil modellerine dahil edilmesinin, sağduyu muhakemesine yardımcı olabileceğini göstermiştir [12].

1.2.11 Wikidata'dan Sağduyu Bilgileri İçeren Altküme Çıkarma

Gelişmiş Wikidata gibi bilgi kaynaklarından, sadece sağduyu bilgisi içeren alt kümelerinin elde edilmesi amaçlanmıştır. Sağduyu bilgisini tümden ayırmak için üç ana ilke benimsenmiştir. İlk ilkeye göre sağduyu bilgisinin kavram hakkında olması gerektiğidir. Sağduyu bilgisi, kavram düzeyindeki bilgilerle ve adlandırılmış varlık düzeyindeki(örnek) bilgiler arasındaki ayrımdan yararlanılmıştır. "Evlerde odaların olması", yaygın olarak uygulanabilir olduğu için sağduyu bilgisidir ancak "Versailles Sarayı'nın 700 odası olduğu" gerçeği ise değildir. Çünkü bu bilgi belirli bir durumla ilgilidir ve çoğu insan için ortak değildir. Bu ayrımı yapmak için etiketlere bakılmıştır çünkü Wikidata'da adlandırılmış varlıkların etiketleri büyük harfle yazılmıştır ve kavramların etiketleri küçük harfle yazılmıştır. Her iki düğümün de küçük harfle başlayan etiketlere sahip olduğu kenarlar seçilmiştir. İkinci ilkeye göre sağduyu bilgisini ayıran bir başka özelliğin o bilginin çoğu insan arasında paylaşılan iyi bilinen kavramlar olmasıdır. "Saklamak için bir kap kullanıldığı" bilgisi yaygın bir bilgidir, oysa "noma'nın aftöz stomatitin bir alt sınıfı olduğu" bilgisi çok bilinmemektedir. Bunun için kelimelerin metinler içindeki kullanım sıklıklarına bakılmıştır. Çok kullanılan kelimeler ve ifadeler, çok bilinen kavramlar olarak kabul edilmiştir. Bu yaklaşıma göre "saklamak" gibi yaygın bir kelimenin sıklığı, "noma" gibi nispeten bilinmeyen bir kelimenin sıklığından çok daha yüksektir. Bu iki yaklaşımla 420,822 kenar, 414 kenar tipine sahip 194,595 düğüm elde edilmiştir. Üçüncü ilke ise sağduyu bilgisinin kimya veya biyoloji gibi belirli bir alan hakkında uzman bilgisinden ziyade genel alan bilgisi hakkında olmasıdır. "Tekerlek arabanın bir parçasıdır" genel alan bilgisiyken ve sağduyu olarak kabul edilirken, "kolesterol bileşen hücre zarına sahiptir" bilgisi biyolojiyle ilgilidir ve sağduyu bilgisi olarak kabul edilmemiştir. Kenar tipleri içerdikleri kenar sayılarıyla sıralanmış, 50. sırayı alan kenar tipinden sonra frekans grafiği nispeten düz hale gelmiştir. Bu yüzden, 50 en çok sık olan kenar tipi kabul edilmiştir. 3. İlke ayrımını yapmak için bu en çok kenar bulunan kenar tipleri ConceptNet v5.7.5'teki ilişkilerle eşleştirilmiştir. Bu ilişkilerden bazıları tek bir ilişkiyle bazıları ise birden fazla ilişkiyle eşleştirilmiştir. 6 tanesi, biyolojiyle ve fizikle alakalı olduğu için çıkarılmıştır. Wikidata'daki opposite of ConceptNet'teki Antonymle eşleştirilmişken, Wikidata'daki genre,sport gibi ilişkiler

ConceptNet'teki HasContext'le eşleştirilmiştir. Sonrasında ise uzmanlık bilgilerinin içinde bulunduğu bir kara liste oluşturup o listedeki kelimelerle bağlantılı olan kenarlar silinmiştir. Mesela proteinle alakalı ilişkiler silinmiştir. Bazı kenarlar birleştirilmiştir. Benzer ilişkiler gruplandırılmıştır. Mesela instance of ve subclass of inheritance grubuna koyulmuştur. Çift yön içeren has cause and has effect ilişkileri tek yönlü ilişkiye çevrilmiştir. En son 15 kenar tipi bırakılmıştır. Wikidata'dan elde edilen sadece sağduyu bilgilerini içeren kümenin son halinin kenar sayısı, Wikidata'nın 0.01%'inin altında kalmıştır. Wikidata-CS ile ConceptNet ve WordNet dahil olmak üzere CSKG'ye dahil olan diğer kaynaklar arasındaki çakışmalar incelenmiştir. Bu, Wikidata'nın sağduyu bilgisinin zaman içindeki evriminin ölçülmesi sağlanmıştır. Wikidata zamanla büyümüştür ancak amacın tüm bilgileri artırmak olduğu ve sadece sağduyu bilgilerini artırmak olmadığı görülmüştür ve RelatedTo ilişkisi dışındaki sağduyu bilgilerindeki artışın Wikidata'nın tümüne oranla yavaş olduğu görülmüştür. SimilarTo, HasPrerequisite ve DistinctFrom gibi ilişkilerinin diğer ilişkilere göre daha hızlı arttığı gözlemlenmiştir. Bu da bazı ilişkilerin sağduyu bakış açılarının Wikidata'nın geneliyle daha ilgili olduğunu göstermiştir. Wikidata'da bulunan different from, opposite of, said to be the same as, partially coincident with, named after, fictional analog of, instance of, subclass of, subproperty of, part of, has part, has parts of the class, material used, is a list of, has list, product or material produced, use, uses, used by, color, has quality, properties of this type, Wikidata property, sex or gender, has cause, has effect, symptoms, followed by, follows, facet of, sport, field of this occupation, health specialty, competition class, genre, studied by, field of work, afflicts, practiced by, depicts, main subject, see also, subject item of this property gibi ilişkilerle, ConceptNet'te bulunan IsA PartOf, HasContext,DistinctFrom,HasPrerequisite,UsedFor,Antonym,MadeOf,Synonym,HasProperty,Causes,DerivedFrom,SimilarTo,CreatedBy,RelatedTo gibi ilişkiler ile eşleştirilmiştir. 71,243 düğüm ve 106,103 kenar elde edilmiştir [14].

1.3 Hipotez

Bu çalışmanın 2 temel hipotezi bulunmaktadır:

1-Hayat bilgisi verileri sosyal medya içeriklerinden elde edilebilir.

2-Bu işlem için varlık ismi etiketleme mimarileri kullanılabilir.

Aşağıda bu 2 hipotezin dayanakları verilmiştir.

Sunulan bu çalışmada bahsedilen yaklaşımlardan farklı olarak Türkçe sosyal medya içeriklerinden otomatik olarak hayat bilgisi verileri elde edilmeye çalışılmıştır. Çözülmesi gereken problemde sosyal medya içeriklerinden kavramlar ve aralarındaki ilişkilerin çıkarılması gerekmektedir. Problem bir varlık ismi tanıma problemi olarak ele alınmıştır. Varlık ismi tanımadaki etiketleme şekli ve mimarileri kullanılmıştır. Varlık ismi tanıma veri kümelerinde tanınmak istenen varlıklar varlık türü ile etiketlenir. Burada ise ilişkinin ismi ve kavramların sırası ile etiketlenmiştir. Tablo 1.1'de bir sosyal medya içeriğinin etiketlenme örneği gösterilmiştir.

Tablo 1.1 Kavramları ve ilişkileri bulmada Varlık ismi tanıma etiketlemesinin kullanım örneği

Kelime	**Etiket**
normal	**Antonym1**
bir	other
insan	other
olmaya	other
zorladılar	other
bana	other
boş	other
yere	other
vakit	other
kaybettiler	other
olmayınca	other
da	other
anormal	**Antonym3**
dediler	other
oğuzatay	other
iyigeceler	other
geceyebirnot	other

Bulunan kavramlar birden fazla kelimeden oluşabilir. Çalışma kapsamında kavramların en fazla 2 kelimeden oluştuğu varsayılmıştır. İlişkilerdeki kavramların sırası da önemli olduğundan etiketlerde sıra numarası da kullanılmıştır. İlk kavram 1-2, ikinci kavram ise 3-4 ile işaretlenmiştir. Ayrıca her kelimenin hangi tweete ait olduğunu gösteren

tweet idler tutulmuştur. Bu kelimelerin karşılıklarını, modellere vermeden önce tweet olarak gruplamayı sağlamıştır.

1.4 Kullanılan Mimarilere Genel Bakış

Mevcut bilgiyle değerlendirme yaparken, önceki bilgileri de düşünerek değerlendirme yapılması gerekmektedir. Mesela, bir kitap okurken bir sayfayı değerlendirirken önceki sayfaları da değerlendirmek gerekir. Bu bakış açısıyla bir veriyi değerlendirirken önceki değerlendirilmiş verileri de hesaba katmak gerekir. Doğal dil işleme problemlerinde bu değerlendirme gerekmiştir. Mevcut fonksiyon değerlendirilirken bir önceki fonksiyonun çıktısı, giriş olarak verilmesi gerektiği düşünülmüştür. Böylece bazı bilgiler hafızada tutulmuştur. Bunun için, RNN çevrimleri önerilmiştir. Ancak RNN, tek katman olan tanh'yı kullanır. Bu kısa süreli hafıza olması sebebiyle yetersiz kalmıştır. Çok fazla geçmiş bilgi olduğunda hafızası yetersiz kalmıştır. Bu sorunu çözmek için LSTM önerilmiştir [16].

LSTM'de, tanh'a ek olarak birden fazla katman eklenmiştir, böylece daha fazla bilginin hafızada kalması sağlanmıştır. Bu katmanlar, forget gate, input gate, output gate, cell state'dir. Cell state kapıları kullanarak, ağdaki bilgileri hafızada tutar ve hangi bilginin tutulup tutulmayacağına karar verir. Sigmoid aktivasyon kullanılır ve bilgiler, 0 veya 1 olarak değerlendirilir. 0 olan bilgiler kapıdan giremezken diğer bilgiler bir sonraki fonksiyonda kullanılır. Forget gate'de önceki fonksiyondan gelen bilgiyle mevcut bilgiyi birlikte sigmoid fonksiyonuna sokarak mevcut bilginin, hatırlanıp hatırlanmayacağına karar verir. Input gate'de, sigmoid ile önceden gelen bilgiyle, mevcut olan bilgiyi güncellemenin gerekip gerekmediğine karar verilir. Düzenleme işlemi için ise, -1 ve 1 aralığı sonucu veren tanh kullanılır ve sigmoid fonksiyonundan elde edilmiş sonuçla çarpılarak karar verilir. Output gatede ise bir sonraki fonksiyonun girişi tespit edilir. Bunun için önceki bilgi ve mevcut bilgiyi sigmoid fonksiyonuna, cell statedeki bilgiyi ise tanh fonksiyonuna sokulduktan sonra birbirleriyle çarpılır ve hangi bilgilerin bir sonraki fonksiyonda yer alacağı belirlenir. Çift yönlü LSTM'de ise tek bir modeli eğitmek yerine iki model kullanılır. İlk model sağlanan girdinin sırasını öğrenir ve ikinci model bu sıranın tersini öğrenir. Eğitimden sonra bu iki mekanizma birleştirilir. Bunu yapmak için Toplama, çarpma, ortalama alma, birleştirme gibi yöntemler kullanılır. Çünkü, tek yönlü LSTM bilgisi geriye doğru akar. Bunun aksine, çift yönlü LSTM bilgisi sadece geriye doğru değil, aynı zamanda iki gizli durumu kullanarak ileriden geriye doğru akar. Bu nedenle çift yönlü LSTM'ler bağlamı daha iyi anlar.

Transformers fikrinin ortaya atılmasındaki amaç, kelimelerin birden fazla anlamı olması ve bu anlamları temsil eden sadece tek kelimeyle gösterilmiş olması sorunu çözmek ve kelimeleri, bağlamlara göre karar veren bir yapı kurmak olmuştur. Kullanılan ilgi mekanizması verinin belli parçalarını daha önemli kılınmıştır. Transformers yapılardan biri de BERT'tir [17]. BERT'in temel teknik yeniliği, Transformer'ın çift yönlü eğitimini uygulamaktır. Ayrı yönlerde eğitim yapılıp bunları birleştirmek yerine, tek bir seferde çift yönlü eğitim uygulanmıştır. Bu özellik, modelin bir kelimenin tüm çevresine dayalı bağlamını yani kelimenin hem sol tarafından hem sağ tarafından öğrenmesini sağlamıştır. Bu da metindeki kelimeler arasındaki ilişkileri öğrenmesini sağlamıştır. Bu yapının Türkçe için kullanımını sağlamak için BERTurk [18] modeli oluşturulmuştur. Çünkü BERT modeli ile oluşturulmak istenen yapı için önceden eğitilmiş BERTurk gibi bir model gereksinimi vardır. BERTurk ile Türkçe doğal dil işleme çalışmalarında da BERT yapısının kullanımı sağlanmıştır.

2
METOT

2.1 Çalışmanın Temel Adımları

Çalışmanın temel adımları aşağıda verilmiştir:

- ConceptNet'ten Türkçe'ye otomatik çeviri
- Frekansı düşük olan kavramların elenmesi
- Kavram ikilileri içeren tweetlerin elde edilmesi
- Bu tweetlerin etiketlenerek eğitim ve doğrulama kümelerinin oluşturulması
- Etiketli verilerle modellerin eğitilmesi
- Modellerin performanslarının ölçülmesi
- Etiketsiz ham verilerden yeni kavramların ve aralarındaki ilişkilerin bulunması

2.1.1 Otomatik Çeviri

ConceptNet'in İngilizce kısımları ayrıştırıldıktan sonra elde edilen kavramlar Google Translate API ile Türkçeleştirilirken bir kavramın birden fazla ilişkide bulunduğu göze alınarak, değişik her kavram için sadece bir kez çeviri yapılmıştır. Çeviri örnekleri Tablo 2.1'de gösterilmiştir. Kavramlar Türkçeleştirilip yerine yerleştirilmeden önce derlemeler içindeki frekanslarına bakılmıştır. Frekansı 0 olan kavramlar elenmiştir. Frekansı 0 olan ve olmayan kavramlara örnekler Tablo 2.2'de gösterilmiştir. Türkçeleştirilmiş kavramlar, içerdikleri ilişkilere yerleştirilmiştir. Bu sayede ilişkilerin çevrilme zamanı azalırken, bu aşamanın performansında iyileştirmeler yapılmıştır. Her kavram birden fazla ilişkide olsa da sadece bir kez çevrilmiştir. Frekansı 0 olmayan kavramlar ilişkilerin içine yerleştirildikten sonra bu ilişkilerin derlemeler içerisindeki bulunma sıklığına bakılmıştır. Frekansı 0 olmayan ilişkiler kabul edilmiştir. Derlemelerde bulunan ve bulunmayan ilişkiler ve birbirleriyle ilişkisi olan kavramlara örnekler Tablo 2.3'te gösterilmiştir. Bir kavramın metin içinde bulunması yani derlemelerde bulunması onun daha belirleyici olduğunu göstermektedir. Bu yüzden derlemelerde bulunan kavramların seçilmesi tweetler elde edilirken daha makul sonuçlar elde edilmesi sağlanmıştır. Ayrıca

Türkçeleştirilmiş kavramlardan kaç tane seçileceği yeni kavramlar bulmada etkili olacağı düşünülmüştür. Bazı Türkçe kavramların karşılığı birden fazla İngilizce kavrama denk gelmektedir. Bu yüzden çeviri yapıldığında bazı ilişkilerin her iki tarafının Türkçe anlamı farklı örneklerde aynı olduğu gözlemlenmiştir. Bu tür ilişkiler veri kümesinden çıkarılmıştır.

Tablo 2.1 ConceptNet'ten Türkçe'ye çevrilen kavram örnekleri

KavramId	**İngilizcesi**	**Türkçesi**
78977	basin	havza
504614	lake	göl
593572	mountain	dağ
641898	ocean	okyanus
700083	river	nehir
826493	sea	deniz
999612	valley	vadi
1018916	waterfall	şelale

Tablo 2.2. ConceptNet'ten Türkçe'ye çevrilen bazı kavramların derlemelerde bulunma sıklıkları

Frekans>0	**Frekans==0**
good,iyi(68267)	goober,şapşal
alive,canlı(6919)	brine fly,salamura sinek
fish,balık(5578)	brinell number,brinel numarası
computer,bilgisayar(2592)	brimstone flower,kükürt çiçeği
engine,motor(1836)	warweariness,uysallık

Tablo 2.3 ConceptNet'ten Türkçe'ye çevrilen kavramları içeren ilişkilerden derlemelerdeki frekansı 0 ve üzerinde olan bazı örnekler

Frekans>0			**Frekans==0**		
İlişki	**Kav1**	**Kav2**	**İlişki**	**Kav1**	**Kav2**
Antonym	anormal	normal	Antonym	ızdırap	coşku
Synonym	nadiren	seyrek	Synonym	gaddarca	habis
PartOf	fren	araba	PartOf	östaki borusu	orta kulak
IsA	dolandırıcılık	suç	IsA	aplikatör	cihaz
UsedFor	akordeon	müzik	UsedFor	amfitiyatro	eğlence

2.1.2 Veri Kümesi Oluşturma

Bir önceki adımda elde edilen kavram ikililerinin her biri için bu kavramların her ikisinin de içinde bulunduğu tweetler Twitter API kullanılarak elde edilmiştir. Twitter API'den alınan tweetlerdeki Türkçe karakter sorunu çözülmüştür. Ayrıca veri kümesinde tekrarlanan tweetler, emojiler, linkler, boşluklar, paragraf boşlukları temizlenmiştir. Tablo 2.4'te Türkçe karakter düzeltilmesine örnekler verilmiştir.

Tablo 2.4. Twitter API'den elde edilen orijinal tweetler, Türkçe karakter sorunu çözülmüş halleriyle birlikte

Orijinal Tweet	**Türkçe Karakter Sorunu Çözülmüş Tweet**
Dedikleriniz kağıt hesabinda dogru evet ama piyasada büyüyen hisseler genelde temettü hisselerinden daha cok prim verir… Buyuyen sirket temettu veriyorsa o en iyisi tabi… Yoksa yavas buyuyen temettu veren mi hizli buyuyen temettu vermeyen mi dersek 2.yi seçerim	Dedikleriniz kağıt hesabında doğru evet ama piyasada büyüyen hisseler genelde temettü hisselerinden daha çok prim verir… büyüyen şirket temettü veriyorsa o en iyisi tabii.. yoksa yavaş büyüyen temettü veren mi, hızlı büyüyen temettü vermeyen mi dersek 2. yi seçerim

Tweetlerde bulunan her kelime etiketlenmiştir. Kavramlar iki kelimeden oluşuyorsa İlişki1 ve İlişki2 olarak, diğer ilişki iki kelimeden oluşuyorsa İlişki3 ve İlişki4 şeklinde etiketlenmiştir. Eğer tek kelimeden oluşuyorlarsa İlişki1 veya İlişki3 olarak etiketlenmiştir. Buradaki İlişki yerine hangi ilişki türündeyse o şekilde etiketlenmiştir. Örneğin Antonym için Antonym1, Antonym2, Antonym3, Antonym4 kullanılmıştır. Bu sayılar o kavramların tweetlerdeki sıralarına göre düzenlenmiştir. İlk gelen kavram tweette diğer kavramdan sonrasında da geliyorsa yine aynı şekilde etiketlenmiştir. Eğer kelime ilişkilerden bağımsız bir kelimeyse o da "other" olarak etiketlenmiştir.

Tablo 2.5'te Tweetlerden bazıları içerdikleri ilişkiler ve kavramlarıyla gösterilmiştir.

Tablo 2.5 Kavramlar, onların aralarındaki ilişki tipleri ve o kavramların bulunduğu tweetlerden örnekler

İlişki Türü	Kavram1, Kavram2	Tweet
Antonym	anormal,normal	**normal** bir insan olmaya zorladılar bana boş yere vakit kaybettiler olmayınca da **anormal** dediler
Synonym	elde etmek,kazanmak	tüm amacın para **kazanmak** ve güç **elde etmek** olduğu bu düzende yolsuzluk hırsızlık sahtekarlık kaçınılmaz
PartOf	fren,araba	duramayarak değil, hiç **fren** yapmamış. Muhtemelen eli telefonda kamyonu görmemiş bile direksiyon başında başka şeylerle ilgilenenlerin maalesef sonu böyle, ama çarptığı kamyon olmayabilirdi. **araba** kullanırken gözünüz yolda olsun. bu kadar ucuz ölmeyin.
IsA	tanıdık,kişi	artık neye inanacağımızı şaşırmış durumdayız zira etrafımızda **tanıdık** bir çok **kişi** korana'dan hayatını kaybetti
UsedFor	klima,soğutma	merhaba, bosch termoteknoloji olarak kombi, **klima**, kazan gibi ısıtma ve **soğutma** ürünleriyle hizmet veriyoruz bosch ev aletleri ile ilgili taleplerinizi ise 4446333 numaralı hattan bosch ev aletleri ekibiyle görüşerek iletebilirsiniz sevgi ve saygılarmızla.

2.1.3 Kavramlar Nasıl İlişkilendirildi

Kavramları ilişkilendirirken, Antonym, Synonym, PartOf, IsA, UsedFor ilişki türleri kullanılmıştır. Kavramlar elde edildikten sonra her ilişki türünden yaklaşık 200 ilişki seçilmiş ve bu ilişkiler içindeki her iki kavramı da içeren tweetler elde edilmiştir.

Bu tweetlerdeki kelimeler etiketlendikten sonra, her ilişki türü için yaklaşık yüzde 95'i eğitim kümesi ve yüzde 5'i ise doğrulama kümesi olarak ayrılmıştır. Modele sokmadan önce, kelimelerin eşsiz sayılara dönüştürülmesi ve kelimelerin bir aradayken oluşturduğu tweetlerin uzunluklarının eşit olması gerekmiştir. Bu yüzden eğitimden önce kelimelerin, karşılık bulduğu tweet id'lere göre gruplandırılmıştır ve bu grup içerisindeki kelime sayılarını eşitlemek için dolgulama işlemi yapılmıştır. Eğitim ve doğrulama kümesinde bulunan Tweet ve etiket sayıları her ilişki için Tablo 2.6'da gösterilmiştir.

Tablo 2.6 Eğitim ve Doğrulama Kümelerindeki Tweet ve Etiket sayıları

İlişki Türü	**Eğitim Tweet Sayısı**	**Doğrulama Tweet Sayısı**	**Eğitim Etiket Sayısı**	**Doğrulama Etiket Sayısı**
Antonym	9336	511	212581	11394
Synonym	4055	268	98765	6634
PartOf	3061	160	70886	3800
IsA	4739	258	120755	6749
UsedFor	2547	132	62778	2672

10 milyon tweet ise yeni kavram ve aralarındaki ilişkileri bulmak için elde edilmiştir. Bu 10 milyon tweette, Türkçe karakter sorunlarının çözülmesi, linklerin silinmesi gibi düzenlemeler otomatik olarak yapılmıştır. 10 milyon tweet giriş olarak verilip çalıştırıldıktan sonra içinde en az iki tane ilişki etiketi bulunan tweetler kabul edilmiştir. Diğerleri alınmamıştır. Program içinde ilişki olduğunu tespit ettiği tweeti, onu oluşturan kelime grubunun etiket karşılıklarında İlişki1 ve İlişki2 içeriyorsa, ilk kavramın veya İlişki3 ve İlişki4 içeriyorsa ikinci kavramın iki kelimeden oluştuğu kabul edilmiş ve ilişkilendirilmiştir. Eğer İlişki1 varsa ve İlişki2 yoksa ilk kavram, aynı şekilde eğer İlişki3

varsa ve İlişki4 yoksa ikinci kavram tek kelimeden oluştuğu kabul edilmiştir. Elde edilen ilişkiler gruplandırılmıştır. Eğitim veri kümesinde bulunan ilişkiler çıkarılmıştır ve kalanlar yani eğitim kümesinde bulunmayan ilişkilerin ne kadar sık bulunduğunu gösteren sayılara göre bir sıralama yapılmıştır. Frekansı yüksek olan ilişkilerin daha güvenilir oldukları değerlendirilmiştir.

2.1.4 Mimariler

Tablo 2.6'da yer alan eğitim kümesinin modellemesi için LSTM ve BERT tabanlı iki mimari kullanılmıştır. LSTM mimarisini için çift yönlü 150 adet LSTM hücresi tek katmanlı olarak kullanılmıştır [19]. Optimizasyon yöntemi olarak Adam, bir defada alınan eğitim örnek sayısı (batch size) 32, çevrim(epoch) sayısı ise 10 alınmıştır. BERT tabanlı mimari içinse Türkçe için eğitilmiş BERTurk [18] kullanılmıştır. Uyarlama (finetuning) işleminde optimizasyon yöntemi olarak AdamW, bir defada alınan eğitim örnek sayısı (batch size) 8, çevrim(epoch) sayısı ise 3 alınmıştır.

2.1.5 Deneysel Sonuçlar

Her ilişki için eğitim ve doğrulama kümeleri ayrı ayrı oluşturulduğundan, her ilişki türü için ayrı modeller eğitilmiştir. BERTurk ve LSTM modellerinin doğrulama kümeleri üzerindeki ağırlıklandırılmış başarı yüzdeleri (W avg) Tablo 2.7'de gösterilmiştir.

Tablo 2.7 İlişkiler için doğrulama kümesindeki doğruluk skorları

İlişki Türü	Doğruluk Skoru (BERTurk)	Doğruluk Skoru (LSTM)
Antonym	0.9648	0.8598
Synonym	0.9490	0.8819
PartOf	0.9265	0.8794
IsA	0.9302	0.8959
UsedFor	0.9427	0.8933

Tablo 2.7'deki sonuçlar incelendiğinde BERT tabanlı mimarinin daha başarılı sonuçlar ürettiği görülmektedir. Bu sebeple çalışmanın devamında BERT tabanlı mimari ile elde edilen sonuçlara odaklanılmıştır. Sonuçları deha detaylı incelemek için her ilişki ve her etiket için kesinlik(precision), duyarlılık(recall), F1 skoru, bunların ortalama (M avg) ve ağırlıklandırılmış (W avg) değerleri ve destek(support) değeri hesaplanmıştır. BERTurk modeli için bu değerler Tablo 2.8-12'te gösterilmiştir.

Tablo 2.8. Antonym için BERT sonuçları

Etiket	Kesinlik	Duyarlılık	F1	Destek
other	0.97	0.99	0.98	10167
Antonym1	0.91	0.70	0.79	620
Antonym2	1.00	0.45	0.62	11
Antoynm3	0.84	0.80	0.82	584
Antonym4	0.83	0.83	0.83	12
M avg	0.91	0.76	0.81	11394
W avg	0.96	0.96	0.96	11394

Tablo 2.9 Synonym için BERT sonuçları

Etiket	Kesinlik	Duyarlılık	F1	Destek
other	0.97	0.98	0.98	5995
Synonym1	0.77	0.59	0.67	322
Synonym3	0.68	0.71	0.69	317
M avg	0.81	0.76	0.78	6634
W avg	0.95	0.95	0.95	6634

Tablo 2.10 PartOf için BERT sonuçları

Etiket	Kesinlik	Duyarlılık	F1	Destek
other	0.97	0.97	0.97	3354
PartOf1	0.72	0.46	0.56	223
PartOf3	0.55	0.75	0.64	223
M avg	0.45	0.44	0.43	3800
W avg	0.93	0.93	0.93	3800

Tablo 2.11 IsA için BERT sonuçları

Etiket	Kesinlik	Duyarlılık	F1	Destek
other	0.96	0.98	0.97	6023
IsA1	0.79	0.40	0.53	354
IsA2	0.20	0.25	0.22	4
IsA3	0.59	0.67	0.63	363
IsA4	0.40	0.40	0.40	5
M avg	0.59	0.54	0.55	6749
W avg	0.93	0.93	0.93	6749

Tablo 2.12 UsedFor için BERT sonuçları

Etiket	Kesinlik	Duyarlılık	F1	Destek
other	0.98	0.98	0.98	2337
UsedFor1	0.70	0.68	0.69	174
UsedFor2	1.00	1.00	1.00	1
UsedFor3	0.70	0.66	0.68	158
UsedFor4	1.00	1.00	1.00	2
M avg	0.88	0.86	0.87	2672
W avg	0.94	0.94	0.94	2672

Tablo 2.8-12 arasında her ilişki türü için hesaplanan sonuçlar incelendiğinde her ilişki türü için en yüksek frekansa sahip olan etiketin doğal olarak "other" olduğu görülmektedir. Bu etiket tüm ilişkiler için yüksek başarı ile tahmin edilmiştir. Ancak yeni kavramların ve aralarındaki ilişkilerin bulunması için önemli olan etiket "other" değildir. Önemli olan etiketler (Antonym, Synonym, PartOf, IsA, UsedFor) için ortalama F1 skorları sırasıyla 0.81, 0.68, 0.6, 0.58 ve 0.69 dur. Buna göre BERT modelinin en başarılı olduğu ilişki türü Antonym, en başarısız olduğu ise IsA'dir.

Doğrulama kümesinde olup modellerin bulduğu bazı kavramlar Tablo 2.13'te verilmiştir.

Eğitilen modellerin doğrulama kümelerinde görece başarıları görüldükten sonra 10 milyon tweet üzerinde yeni kavramları ve aralarındaki ilişkileri bulma denemeleri yapılmıştır. Tablo 2.14'te eğitim kümesinde bulunmayan ilişkiler ve bunların kaç tane tweette bulunduğuna dair örnekler verilmiştir.

Tablo 2.13 İlişkiler için sadece doğrulama kümesinde olan ve doğrulama esnasında program tarafından bulunan bazı ilişkiler

ANTONYM		**SYNONYM**	
Kavram1	**Kavram2**	**Kavram1**	**Kavram2**
olağan	sıradışı	mükemmel	kusursuz
sabit	değişken	fakir	yoksul
değersiz	değerli	hata	kusur
kazanan	kaybeden	nezaketsiz	kaba
değişmiş	değişmemiş	özetle	kısaca
açmak	kapatmak	heybetli	gösterişli
aşağı	yukarı	iyileştirmek	düzeltmek
PARTOF		**ISA**	
Kavram1	**Kavram2**	**Kavram1**	**Kavram2**
merdiven	bina	makarna	gıda
klozet	tuvalet	gürültü	ses
semptom	hastalık	anne	ebeveyn
metin	yayın	cami	ibadethane
boğaz	boyun	süt	içecek
retina	göz	migren	hastalık
raf	dolap	evlilik teklifi	teklif

USEDFOR	
Kavram1	**Kavram2**
sakız	çiğnemek
ayak	dans
şişe	şarap
kumaş	dikiş
disk	depolama
fabrika	imalat
helikopter	ulaşım

Tablo 2.14 10 milyon tweetten elde edilen bazı yeni kavram ve ilişkiler

ANTONYM			**SYNONYM**		
Kavram1	**Kavram2**	**Frekans**	**Kavram1**	**Kavram2**	**Frekans**
gündüz	gece	798	kader	kısmet	255
zengin	fakir	592	fakir	fukara	188
gelir	gider	424	salak	aptal	148
alt	üst	391	güç	kuvvet	131
zayıf	güçlü	378	sağlık	sıhhat	129
gidiş	dönüş	279	makam	mevki	97
dolu	boş	232	şeref	haysiyet	64
PARTOF			**ISA**		
Kavram1	**Kavram2**	**Frekans**	**Kavram1**	**Kavram2**	**Frekans**
kelime	cümle	175	futbol	spor	192
doktor	hastane	170	tüfek	silah	168
kapak	kitap	94	şiir	yazı	136
sayfa	defter	21	sarı	renk	103
yelkovan	saat	20	makarna	yemek	40
tuş	klavye	7	forvet	oyuncu	35
pota	basketbol	4	telefon	cihaz	13

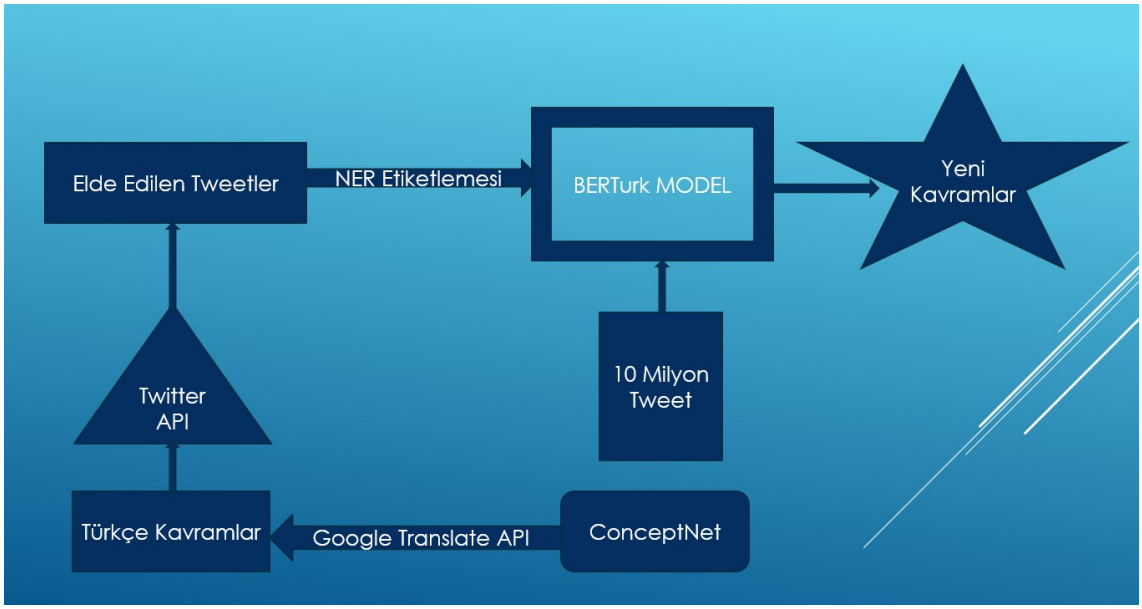
USEDFOR		
Kavram1	**Kavram2**	**Frekans**
demlik	çay	118
komedi	gülmek	99
top	gol	86
müzik	dinlemek	84
maç	izlemek	42
yol	gitmek	30
şort	giymek	28

Tablo 2.14 incelendiğinde bulunan yeni kavram ve ilişkilerin doğru olduğu görülmektedir. Bununla birlikte bulunan ilişkileri kaç tane farklı tweetten bulunduğunun doğruluk üzerinde etkili olduğu görülmüştür. Bir ilişki çok sayıda tweette bulundu ise doğru olma ihtimali yüksektir. Az sayıda tweetten elde edilen ilişkilerin doğru olma ihtimali düşük kalmaktadır. Ancak çalışmada kaynak olarak sosyal medya içeriklerini seçmemizin önemi burada ortaya çıkmaktadır. 10 milyon sayısı istenildiği kadar arttırılabilir ve bu sayede daha çok tweet ile daha çok sayıda ilişki güvenilir bir şekilde bulunabilir.

3
SONUÇ VE ÖNERİLER

Hayat bilgisi veri tabanları için kaynak olarak dijital ansiklopediler, insanlardan belli arayüzler aracılığı ile toplanan verilerin kullanımı yaygındır. Bu çalışmada bu yaklaşımlardan farklı olarak insanların günlük yaşama dair attıkları tweetlerin kullanımı önerilmiştir. Sosyal medya içeriklerinin hacimsel büyüklüğü ve çeşitliliği bu alanda büyük potansiyel taşımaktadır.

Bulunmak istenen ilişkiler için öncelikle çekirdek ilişkiler elde edilmiştir. Bunun için ConceptNet kullanılmıştır. İngilizce ilişkiler içeren bu bilgi tabanı otomatik makine çevirisi ile çevrilmiş ve her ilişki türü için belli sayıda ilişki seçilmiştir. Daha sonra seçilen ilişkilerdeki her iki kavramı da içeren tweetler elde edilmiştir. Bu tweetler etiketlenerek eğitim ve doğrulama kümesi oluşturulmuştur. Bu verileri modellemede varlık ismi tanıma problemi örnek alınmıştır. Etiketleme ve modelleme için varlık ismi tanıma süreçleri kullanılmıştır. Eğitilen modellerin başarıları doğrulama kümesi üzerinde ölçülmüş ve daha sonra yine topladığımız 10 milyon tweet üzerinde yeni kavram ve ilişkiler bulmaları istenmiştir. Bulunan ilişkilerin kaç farklı tweetten bulunduğu bulunan ilişkinin doğruluğunu gösteren bir işaretçi olarak kabul edilmiştir. Şekil 3.1'de mimarinin özet yapısı gösterilmiştir.



Şekil 3.1 Mimarinin özet yapısı gösterilmiştir

Sonuç olarak, bu çalışmada hayat bilgisi çıkarımı probleminin varlık ismi tanıma problemi olarak görülmesinin ve kaynak olarak sosyal medya içeriklerinin kullanılmasının mümkün olduğu ortaya konmuştur. Kullanılan modeller gelişime açıktır. Gelecek çalışmalar olarak mevcut mimarilerin hiper parametre optimizasyonu ve başka mimarilerin denenmesi düşünülmektedir.

KAYNAKÇA

[1] H. Lui and P. Singh, “ConceptNet — A Practical Commonsense Reasoning Tool-Kit”,2004.

[2] D.B. Lenat, R.V. Guha, K. Pittman, D.Pratt and M.Shepherd, “Cyc Toward Programs With Common Sense”, 1990.

[3] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins and W. L. Zhu, “Open Mind Common Sense: Knowledge Acquisition from the General Public”,2002.

[4] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. J. Miller, “Introduction to WordNet: An On-line Lexical Database”,1993.

[5] O. Bilgin, Ö. Çetinoğlu and K. Oflazer, “Building a wordnet for Turkish”,2004.

[6] M. F. Amasyalı, B.İnak and M.Z. Ersen, “Türkçe Hayat Bilgisi Veri Tabanının Oluşturulması”, 2010.

[7] C. Özköse and M. F. Amasyalı, “Tümce Öğelerinden Hayat Bilgisi Çıkarımı”, 2012.

[8] S. Özcan and M. F. Amasyalı, “TÜRKÇE HAYAT BİLGİSİ VERİTABANI VE CSOYUN”,2014.

[9] C.B. Özdemir and M. F. Amasyalı, “Hayat Bilgisi Veritabanı Kullanılarak Otomatik Cümle Üretimi”,2011.

[10] D. Jurafsky and J. H. Martin, “Speech and Language Processing”,2020.

[11] T. Y. Chang, Y. Liu, K. Gopalakrishnan, B. Hedayatnia, P. Zhou and D. H. Tür, “Incorporating Commonsense Knowledge Graph in Pretrained Models for Social Commonsense Tasks”,2020.

[12] Y. Xu, C. Zhu, R.Xu, Y.Liu, M.Zeng, and X. Huang, “Fusing Context Into Knowledge Graph for Commonsense Reasoning”,2020.

[13] J. Dy, C. Brito, V. Tan and J. Lola, “Acquiring Commonsense Knowledge during Collaborative Storytelling”,2021.

[14] F. Illievski, P. Szekely and D. Schwabe, “Commonsense Knowledge in Wikidata”,2020.

[15] Lieberman, H. and Liu. H. 2002. Adaptive Linking between Text and Photos Using Common Sense Reasoning. In Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems, (AH2002) Malaga, Spain

[16] A. A. Sharfuddin, M. N. Tihami and M. S. Islam, “A Deep Recurrent Neural Network with BiLSTM model for Sentiment Classification”, 2018.

[17] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 2019.

[18] S. Schweter, “Berturk - bert models for turkish,” Apr. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3770924.

[19] Graves, A., Fernández, S., & Schmidhuber, J. (2005, September). Bidirectional LSTM networks for improved phoneme classification and recognition. In International conference on artificial neural networks (pp. 799-804). Springer, Berlin, Heidelberg.

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirileri

1. D. Altınal and M. F. Amasyali, “Building Common Sense Database By Using Social Media Content”,2021.