

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

**HABER ARŞİVLERİNDE OLAY ÇİZGE MODELİ İLE
YAPILANDIRILMIŞ BİLGİ KEŞFİ VE SORU CEVAPLAMA**

BAŞAK BULUZ KÖMEÇOĞLU

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN: DR. ÖĞR. ÜYESİ BURCU YILMAZ

ARALIK 2024

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

HABER ARŞİVLERİNDE OLAY ÇİZGE MODELİ İLE
YAPILANDIRILMIŞ BİLGİ KEŞFİ VE SORU
CEVAPLAMA

BAŞAK BULUZ KÖMEÇOĞLU

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN: DR. ÖĞR. ÜYESİ BURCU YILMAZ

ARALIK 2024

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

**STRUCTURED INFORMATION DISCOVERY AND
QUESTION ANSWERING WITH AN EVENT GRAPH
MODEL IN NEWS ARCHIVES**

BAŐAK BULUZ KÖMEÇOĐLU

A THESIS OF DOCTORATE
DEPARTMENT OF COMPUTER ENGINEERING

ADVISOR: ASST. PROF. DR. BURCU YILMAZ

DECEMBER 2024

DOKTORA JÜRİ ONAY FORMU

GTÜ Lisansüstü Eğitim Enstitüsü Yönetim Kurulunun 16/12/2024 tarih ve 24/66 sayılı kararıyla oluşturulan jüri tarafından 23/12/2024 tarihinde tez savunma sınavı yapılan Başak Buluz Kömeçoğlu'nun tez çalışması Bilgisayar Mühendisliği Anabilim Dalında DOKTORA tezi olarak kabul edilmiştir.

JÜRİ

ÜYE

(TEZ DANIŞMANI) : Dr. Öğr. Üyesi Burcu YILMAZ

ÜYE

: Doç. Dr. Yakup GENÇ

ÜYE

: Prof. Dr. Pınar KARAGÖZ

ÜYE

: Prof. Dr. Sevinç İLHAN OMURCA

ÜYE

: Doç. Dr. Habil KALKAN

ONAY

Gebze Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Yönetim Kurulunun
...../...../..... tarih ve/..... sayılı kararı.

İMZA/MÜHÜR



*Sonsuz sabır ve karşılıksız sevgi ile her daim yanımda olan
aileme, değerli eşime ve tüm fırsat eşitsizliklerine rağmen
geleceğe umut taşıma cesareti gösteren tüm çocuklara...*

ÖZET

Hızla artan dijital bilgi kaynakları, özellikle büyük ölçekli haber arşivleri gibi zengin bağlamsal verilerde, doğru ve etkili bilgiye erişim süreçlerini daha karmaşık hale getirmiştir. Bu tez, çevrimiçi haber metinlerinden olay tabanlı bilgi çıkarımı ve bu bilgilerin çok adımlı soru cevaplama sistemlerinde kullanılması için bir çerçeve sunmaktadır. Çalışmada geliştirilen olay çizge modeli, yalnızca bireysel bilgi birimlerini çıkarmakla kalmamış, aynı zamanda olaylar arasındaki bağlamsal ve zamansal ilişkileri de yapılandırarak olay örgüsü zincirleri oluşturmayı amaçlamıştır. Model, isimlendirilmiş varlık merkezli alt çizge madenciliği ve zamansal bilgi entegrasyonu gibi yöntemlerle desteklenmiş; bu yaklaşımlar özellikle çok adımlı soruların yanıtlanmasında dikkate değer sonuçlar elde edilmesini sağlamıştır.

Çalışmanın yeniliklerinden biri, Türkçe doğal dil işleme alanına uygun standart veri kümelerinin geliştirilmiş olmasıdır. Bu veri kümeleri, bilgi tabanlı soru cevaplama ve olay odaklı metin kümeleme gibi alanlarda kullanılmak üzere hazırlanmış ve literatürdeki büyük bir eksikliği gidermek üzere araştırmacılara sunulmuştur. Bunun yanı sıra, büyük dil modellerinin semantik ayrıştırma süreçleriyle zenginleştirilmesi ve komut istemi mühendisliği teknikleriyle performanslarının artırılması, çalışmayı yalnızca teorik bir katkı değil, aynı zamanda pratik bir dönüşüm aracı haline getirmiştir.

Bu çalışma, dijital haber arşivlerinden bilgi çıkarımı ve çok adımlı soru cevaplama sistemleri için yapısal bir temel sunmakta ve bilgi erişim süreçlerini optimize etmek için yeni yöntemler önermektedir. Bununla birlikte, elde edilen bulgular, özellikle olay çizge modelinin daha geniş veri kümelerinde uygulanabilirliği ve dil modellerinin maliyet-performans dengesi gibi konuların gelecekteki araştırmalar için potansiyel taşıdığını göstermektedir.

Anahtar Kelimeler: Büyük Dil Modeli, Çok Adımlı Soru Cevaplama, Doğal Dil İşleme, Metin Kümeleme, Sık Tekrarlı Alt Çizge Madenciliği

ABSTRACT

Rapidly growing digital information sources, especially rich contextual data such as large-scale news archives, have made the processes of accessing accurate and effective information increasingly complex. This dissertation presents a framework for event-based information extraction from online news texts and the application of this information in multi-step question answering systems. The event graph model developed in this study not only extracts individual information units but also structures contextual and temporal relationships between events to form coherent event chains. The model is supported by techniques such as named entity-centered subgraph mining and temporal information integration, which have demonstrated notable effectiveness, particularly in answering multi-step questions. One of the key innovations of this study is the development of standardized datasets tailored to the Turkish natural language processing domain. These datasets are designed for use in areas such as knowledge-based question answering and event-centric text clustering, addressing a significant gap in the literature and providing valuable resources for researchers. Furthermore, the enrichment of large language models with semantic parsing processes and the enhancement of their performance through prompt engineering techniques have transformed this study into not only a theoretical contribution but also a practical tool for implementation. This research establishes a structural foundation for information extraction from digital news archives and multi-step question answering systems, proposing novel methods to optimize information retrieval processes. However, the findings also highlight the potential for future research, particularly in the applicability of the event graph model to larger datasets and the exploration of cost-performance trade-offs in language models.

Keywords: Frequent Subgraph Mining, Large Language Model, Multi-hop Question Answering, Natural Language Processing, Text Clustering

TEŞEKKÜR

Bu doktora tezi, bilgiye duyduğum tutkunun ve etrafımda ışık saçan, bana destek ve ilham veren değerli insanların katkılarıyla şekillendi. Bu süreç boyunca yanımda olan, zorlukları birlikte göğüslediğim herkese şükran borçluyum.

Değerli danışmanım Dr. Burcu Yılmaz’a merakımı özgür bırakıp bana her zaman ilham verdiği ve bu süreç boyunca bana duyduğu güven ve sunduğu destek ile akademik yolculuğumda bana rehberlik ettiği için özel olarak teşekkür etmek isterim.

Bu tezin her adımında daha güçlü ve sağlam bir temele oturmasını sağlayan derin bilgi ve rehberlikleri ile tez izleme sürecimde değerli katkılarını esirgemeyen Prof. Dr. Pınar Karagöz ve Doç. Dr. Yakup Genç’e ve ayrıca, tez jürisinde yer alarak çalışmalarımı değerlendiren ve son haline gelmesine katkı sağlayan tüm hocalarıma teşekkür ederim.

En büyük teşekkürü ise sevgili aileme borçluyum. Bu zorlu süreçte en yakın arkadaşım, yol arkadaşım ve aynı zamanda meslektaşım olan sevgili eşime minnettarım. Tezin her aşamasında, hem teknik bilgisiyle hem de manevi desteğiyle yanımdaydı. Onun varlığı, her zor anı aşmamı sağladı ve bana güç verdi. Bu yolculuğu birlikte tamamlamanın mutluluğu bu başarının en değerli ödülü oldu. Bana en zor anlarda bile güç veren, sevgisi ve desteğiyle varlığını hissettiren tüm aileme sonsuz teşekkürlerimi sunarım. Onların inancı, sabrı ve fedakârlığı, yolda karşılaştığım tüm zorlukları aşmamda en büyük dayanağım oldu.

Bu süreçte bilgi paylaştığım, ilham aldığım ve ortak hedefler doğrultusunda çalıştığım araştırma arkadaşlarıma ve akademik topluluğa da şükranlarımı sunarım. Beraber paylaştığımız tartışmalar, fikir alışverişleri ve projeler, bu sürecin sadece bilimsel değil, insani anlamda da zenginleşmesine vesile oldu.

Son olarak, hayatıma değer katan tüm dostlarıma teşekkür ederim. Onların varlığı, en zorlu anları dahi aydınlatırken, bana daima motivasyon ve cesaret verdi. Bu süreçte gösterdikleri dostluk, başarılarımı daha anlamlı hale getirdi.

Bu tezi, tüm fırsat eşitsizliklerine rağmen bilime gönül vererek katkı sunan, bilgiye ve gerçeğe ulaşma yolunda mücadele eden herkese ithaf ediyorum. Bilimin ışığıyla nice zorlukların aşılmasını ve daha parlak bir geleceğe katkı sağlamasını diliyorum.

İÇİNDEKİLER

	<u>Sayfa</u>
ÖZET	vi
ABSTRACT	vii
TEŞEKKÜR	viii
İÇİNDEKİLER	ix
SİMGELER ve KISALTMALAR DİZİNİ	xi
ŞEKİLLER DİZİNİ	xiii
TABLolar DİZİNİ	xiv
1. GİRİŞ	1
1.1. Motivasyon	1
1.2. Tezin Amacı, İçeriği ve Katkısı	2
1.3. Hipotez	5
1.4. Tezin Organizasyonu	5
2. KAVRAMSAL ÇERÇEVE	7
2.1. Soru Cevaplama Problemi	7
2.1.1. Bilgiye Erişim Modülü	10
2.1.2. Okuduğunu Anlama Modülü	16
2.2. Bilgi Tabanlı Soru Cevaplama Problemi	19
2.2.1. Anlamsal Ayırıştırma Tabanlı Yöntemler	20
2.2.2. Bilgiye Erişim Tabanlı Yöntemler	21
2.2.3. Büyük Dil Modelleriyle Bilgi Tabanlı Soru Cevaplama	21
3. ÖN ÇALIŞMALAR	23
3.1. Gerçek Zamanlı Haber Akışında Aynı İçerikli Haberlerin Tespiti	23
3.1.1. Giriş ve Problem Tanımı	23
3.1.2. Metodoloji	25
3.1.3. Deneyler	29
3.1.4. Sonuçlar ve Tartışma	30
3.2. Adlandırılmış Varlıkların Olay Tabanlı Haber Kümelemeye Etkisi	32
3.2.1. Giriş ve Problem Tanımı	32
3.2.2. Metodoloji	35
3.2.3. Deneyler	38
3.2.4. Sonuçlar ve Tartışma	41
4. OLAY ÇİZGESİ TABANLI HABER KÜMELEME: İSİMLENDİRİLMİŞ VARLIK MERKEZLİ ALT ÇİZGELERİN ROLÜ	44

4.1. Giriş	44
4.2. Yöntem	45
4.3. Sonuçlar ve Tartışma	59
4.3.1. Deneysel Kurulum	59
4.3.2. Veri Kümeleri	60
4.3.3. Performans Değerlendirme	62
4.3.4. Sonuçlar	66
4.3.5. Kısıtlar	75
5. OLAY ÖRGÜSÜ BİLGİ GRAFİKLERİ VE BÜYÜK DİL MODELLERİ İLE ÇOK ADIMLI SORU CEVAPLAMA	77
5.1. Giriş	77
5.2. Metodoloji	79
5.3. Deneyler	88
5.3.1. Değerlendirme Veri Kümeleri	88
5.3.2. Değerlendirme Metrikleri	90
5.4. Sonuçlar ve Tartışma	91
5.5. Kısıtlar	103
6. SONUÇLAR VE GELECEK ÇALIŞMALAR	105
KAYNAKLAR	106
ÖZGEÇMİŞ	132
TEZ ÇALIŞMASI KAPSAMINDA YAPILAN YAYINLAR	133
EKLER	134

SİMGELER ve KISALTMALAR DİZİNİ

A	:	Dikkat Mekanizması Başlık Boyutu
B	:	Dallanma Faktörü Parametresi
Eps	:	Epsilon Yarıçapı
G_C	:	İçerik Çizgesi
G_E	:	Olay Çizgesi
G_{NerT}	:	İsimlendirilmiş Varlık ve Zaman Etiketleri Çizgesi
G_{SE}	:	Basit Olay Çizgesi
G_{SNer}	:	İsimlendirilmiş Varlık Tanıma ile Özetlenen İçerik Çizgesi
G_{SE+FS}	:	Sık Tekrarlı Alt Çizge Düğümlü Basit Olay Grafiği
$G_{SE+NEXTENTITY}$	:	“NEXT ENTITY” kenarları ile Basit Olay Grafiği
H	:	Gizli Katman Boyutu
L	:	Katman Sayısı
Min_heap	:	Öncelik Kuyruğu
$min_cluster_size$	:	Minimum Küme Boyutu Parametresi
$min_samples$	:	Epsilon Yarıçapı İçinde Bulunan Minimum Nesne Sayısı
$minsup$	:	Minimum Destek Parametresi
T	:	Eşik Değeri
$ev_{event}(G)$	:	Bir Olay Çizgesi Gömmesi
$ev(u)$	:	Bir Olay Çizgesi Düğümünün Gömme Vektörü
BERT	:	Transformatörlerden Çift Yönlü Kodlayıcı Temsilleri
BiLSTM	:	Çift yönlü LSTM
BIRCH	:	Dengeli Yinelemeli Azaltma ve Hiyerarşi Kullanarak Kümeleme
BoW	:	Kelime Torbası
CF Ağacı	:	Küme Özellik Ağacı
CNN	:	Evrişimli Sinir Ağları
CoT	:	Düşünce Zinciri

DBSCAN	:	Gürültülü Uygulamaların Yoğunluk Tabanlı Uzamsal Kümelenmesi
DFS	:	Önce Derinlik Araması
EM	:	Tam Eşleşme
GPT	:	Üretken Ön Eğitim Transformatörü
GRU	:	Geçitli Tekrarlayan Birim
HDBSCAN	:	Hiyerarşik DBSCAN
IR	:	Bilgiye Erişim
KBQA	:	Bilgi Tabanlı QA
KG	:	Bilgi Çizgesi
LLM	:	Büyük Dil Modeli
LSTM	:	Uzun Kısa Süreli Bellek
MRC	:	Okuduğunu Anlama
MST	:	Minimum Yayılma Ağacı
NC	:	News Clustering Veri Kümesi
NER	:	İsimlendirilmiş Varlık Tanıma
NLQ	:	Doğal Dil Sorusu
NLP	:	Doğal Dil İşleme
OCR	:	Optik Karakter Tanıma
OPTICS	:	Kümeleme Yapısını Belirlemek için Noktaları Sıralama
POS	:	Sözcük Türü Belirleme
QA	:	Soru Cevaplama
RDF	:	Kaynak Tanımlama Çerçevesi
RNN	:	Tekrarlayan Sinir Ağı
SBN	:	Hikaye Tabanlı Haber Veri Kümesi
SP	:	Anlamsal Ayrıştırma
SQUAD	:	Stanford Soru Cevaplama Veri Kümesi
TDT	:	Konu Algılama ve İzleme
TF-IDF	:	Terim Frekansı - Ters Belge Frekansı
VSM	:	Vektör Uzayı Modeli

ŞEKİLLER DİZİNİ

	Sayfa
Şekil 1.1: Büyük Dil Modelleri Kullanarak Semantik Ayırıştırma Tabanlı Bilgi Grafiği Soru Cevaplama Sistemi	3
Şekil 2.1: Geleneksel Soru Cevaplama Sistemi İşlem Hattı [1]	8
Şekil 3.1: İki Aşamalı Benzerlik Yaklaşımı ile Gerçek Zamanlı Haber Kümeleme Sistem Mimarisi	26
Şekil 3.2: Önerilen Olay Tabanlı Haber Kümeleme Yaklaşımı	34
Şekil 4.1: Olay Grafiği Tabanlı Haber Kümeleme Yaklaşımının Genel Yapısı [2]	46
Şekil 4.2: Haber Metninden Çizge Düğümlerinin Çıkarılması Örneği [2]	47
Şekil 4.3: Basit Olay Çizge Modeli	50
Şekil 4.4: Yıldız Çizge Yapısı Örneği	52
Şekil 4.5: Olay Çizgesi - Adlandırılmış Varlık Merkezli Sık Tekrarlı Alt Çizgelerle Genişletilmiş Olay Çizgesi	53
Şekil 4.6: Haber Metninden Olay Grafiği Oluşturma [2]	55
Şekil 5.1: Olay Örgüsü Örnek Grafik Gösterimi	80
Şekil 5.2: Olay Örgüsü Tabanlı Bilgi Grafiği İşlem Hattı	82
Şekil 5.3: Basit Neo4j Cypher Sorgu Oluşturma İstemi	85
Şekil 5.4: Neo4J Cypher Sorgu Oluşturma için Bilgi Eklentili LM İstemi	86
Şekil 5.5: Bilgi ile Güçlendirilmiş Dil Modeli İstem Hattı	88
Şekil 5.6: Ayrık Olay ve Olay Örgüsü Tabanlı Bilgi Grafiği Yapılarının Çok Adımlı QA Performansına Etkileri	94
Şekil 5.7: Hikaye Tabanlı KBQA Veri Kümesi üzerinde Zaman Etiketli Düğümlerinin Zamansal Soru Cevaplama Performansına Etkisi	98
Şekil 5.8: İstem Mühendisliğinin Soru Cevaplama Performansı Üzerindeki Etkisi	102
Şekil 5.9: İstem Mühendisliği ile Soru Cevaplama Performans İyileştirme Isı Haritası	102
Şekil 6.1: Olay Çizgesi Tabanlı Haber Kümeleme Geliştirme Prosedürü	139

TABLOLAR DİZİNİ

	<u>Sayfa</u>
Tablo 3.1: Farklı Gömme Seviyeleri için Kümeleme Performansının B-Cubed Değerlendirme Metriği ile Karşılaştırılması	31
Tablo 3.2: 2-adımlı Kümeleme Yaklaşımında Farklı Gömme Seviyelerinin Karşılaştırılması	31
Tablo 3.3: Türkçe Olay Tabanlı Metin Kümeleme Veri Kümesinden Örnek Haber Kümeleri	40
Tablo 3.4: Adlandırılmış Varlıkların Metin Kümeleme Performansı Üzerindeki Etkileri	42
Tablo 4.1: News Clustering Veri Kümesi'ne ait İstatistikler	60
Tablo 4.2: Hikaye Tabanlı Haber Veri Kümesi	63
Tablo 4.3: Kümeleme Algoritmalarının Parametre Ayarları	64
Tablo 4.4: Farklı Basit Çizge Modellerinin Kümeleme Performansı Üzerindeki Etkisi	68
Tablo 4.5: Adlandırılmış Varlık Merkezli Sık Tekrarlı Alt Çizge Düğümlerinin Kümeleme Performansı Üzerindeki Etkisi	70
Tablo 4.6: Veri Kümelerindeki Adlandırılmış Varlık Merkezli Sık Alt Çizge Sayıları	72
Tablo 4.7: Önerilen Yöntem ile SOTA Yöntemlerin Kümeleme Performansı Karşılaştırması	73
Tablo 5.1: Büyük Dil Modelleri Teknik Detay Tablosu	84
Tablo 5.2: Büyük Dil Modelleri Ortalama Girdi ve Çıktı Karakter Sayısına göre API Çağrı Maliyetleri	84
Tablo 5.3: Bilgi ile Güçlendirilmiş Komut İstemi ile İyileştirilmiş Sorgu Örneği	87
Tablo 5.4: Hikaye Tabanlı KBQA Veri Kümesi İstatistikleri	89
Tablo 5.5: MuSiQue-Ans Veri Kümesi İstatistikleri	90
Tablo 5.6: MuSiQue-Ans Veri Kümesindeki Muhakeme Grafikleri [3]	92
Tablo 5.7: Hikaye Tabanlı KGQA Veri Setinden Örnek Sorular	93
Tablo 5.8: MuSiQue-Ans (validation-mini) Veri Kümesi üzerinde Büyük Dil Modelleri ile Çok adımlı QA Performansı	95
Tablo 5.9: MuSiQue-Ans validation-mini Veri Kümesi Üzerinde Temel Soru Cevaplama Performansları	96
Tablo 5.10: Ayrık Olay ve Olay Örgüsü Tabanlı Bilgi Grafiği Yapılarının Çok Adımlı QA Performansı Karşılaştırması	97
Tablo 5.11: Hikaye Tabanlı KBQA Veri Kümesi Zaman Etiketli Düğümlerinin Zamansal Soru Cevaplama Performansına Etkisi	98
Tablo 5.12: MuSiQue-Ans (validation-mini) Veri Kümesi Zaman Etiketli Düğümlerinin Zamansal Soru Cevaplama Performansına Etkisi	99
Tablo 5.13: MuSiQue-Ans Doğrulama Veri Kümesi Zaman Etiketli Düğümlerinin Zamansal Soru Cevaplama Performansına Etkisi	100
Tablo 5.14: MuSiQue-Ans Doğrulama Kümesi Zamansal Soru İstatistikleri	100

Tablo 5.15: Hikaye Tabanlı KBQA Veri Kümesi Üzerinde Komut İstemi Mühendisliđinin Büyük Dil Modellerinin QA Performansına Etkisi	103
---	-----

1. GİRİŞ

Bu bölümde, tez çalışmasında ele alınan probleme yönelik motivasyonun yanı sıra, çalışmanın özgün değerini açıklayan amaçlar, tezin içeriği, literatüre sunulan katkılar ve son olarak tezin organizasyonu yer alacaktır.

1.1. Motivasyon

Dijitalleşmenin hız kazanması, bilgiye erişim alışkanlıklarında köklü dönüşümler yaratmıştır. Özellikle çevrimiçi haber kaynaklarının yaygınlaşmasıyla birlikte, kullanıcılar artık yalnızca statik ve basılı içeriklerle sınırlı kalmayıp, dinamik olarak güncellenen ve zengin çeşitlilikteki dijital bilgi kaynaklarına erişim sağlayabilmektedir [4]. Bu değişim, haberlerin hızla yayımlanması, düzenlenmesi ve güncellenmesi gibi kritik avantajlar sunarken, hızla genişleyen haber arşivlerinin sürekli artan veri hacmi, içeriklerin heterojen yapısı ve bağlamların çeşitliliği, kullanıcıların bilgiye etkin ve hızlı bir şekilde erişimini önemli ölçüde zorlaştırmaktadır [4, 5].

Geçmişteki bilgilerin doğrulanması, olayların zaman içerisindeki gelişiminin anlaşılması ya da tarihsel verilerin yalnızca incelenmesi gibi çeşitli bilgi ihtiyaçları, profesyonellerin ve bireysel kullanıcıların büyük ölçekli arşivlerde etkin arama yapabilmesi ve doğru bilgilere erişebilmesini gerektirir. Ancak, bu tür arşivlerde sıklıkla karşılaşılan düzensiz veri organizasyonu, olaylar arası açık bağlantıların eksikliği ve farklı haber olaylarını ele alan belgelerin karmaşık yapısı, anlamlı ve etkili sorgulama süreçlerini zor ya da imkansız hale getirmektedir. Bu tür zorlukları aşmak için doğal dilde ifade edilen sorulara doğru ve bağlamsal olarak uygun yanıtlar sağlamayı hedefleyen büyük ölçekli soru-cevaplama (QA) sistemleri, güçlü bir çözüm sunma potansiyeline sahiptir [6]. Bu sistemler, arşivlerdeki büyük ve düzensiz veri kümelerini yapılandırılmış bilgiye dönüştürerek, kullanıcıların ihtiyaç duyduğu bilgilere hızlı ve etkili bir şekilde erişim sağlamalarına olanak tanımaktadır.

Doğal dil işleme (Natural Language Processing - NLP) alanındaki soru cevaplama (Question Answering - QA) problemi, bilgi erişimi (Information Retrieval - IR) ve bilgi çıkarımı (Information Extraction - IE) gibi temel problemleri bütünleştiren bir araştırma alanı olarak önemli bir konuma sahiptir [7, 8]. Bilgi erişimi sistemleri, kul-

lanıcıların sorgularına uygun dokümanları tespit ederek bunları yanıt olarak sunarken; bilgi çıkarımı sistemleri yapılandırılmamış veya yarı yapılandırılmış veri kaynaklarından belirli bilgileri otomatik olarak çıkarmayı hedefler. QA ise bu iki temel bileşeni bir araya getirerek, kullanıcıların sorduğu soruların yanıtlarını, kaynak dokümanlardan anlamlı ve doğrudan bir çıktı olarak üretmeyi amaçlar [9]. QA sistemleri, yalnızca yüzeysel bilgi çıkarımını değil, olaylar, varlıklar ve bunlar arasındaki ilişkiler üzerinden derinlemesine bir analiz ve akıl yürütme yeteneği gerektirir [10].

Bu tez çalışması, bu zorlukların üstesinden gelmek amacıyla, çevrimiçi haber metinlerinden yapılandırılmış bilgi çıkarımı ve bu bilgilerin olaylar arası bağlamları ve zamansal ilişkileri kapsayan bir model çerçevesinde düzenlenmesine odaklanmaktadır. Önerilen yaklaşım, haber metinlerinden çıkarılan olayları, yalnızca bağımsız özellikleri ile ele almak yerine, bu olaylar arasındaki ilişkileri dikkate alarak bir "olay çizgesi modeli" geliştirmeyi hedeflemektedir. Olay çizgesi, isimlendirilmiş varlıklar ve zamansal bilgiler üzerinden yapılandırılarak, olayların bağlamını ve zamansal dizilimini koruyan bir "olay örgüsü zinciri" adı verilen bir anlatı yapısı sunmaktadır. Bu yapı, yalnızca basit soruların yanıtlanmasıyla sınırlı kalmayıp, karmaşık bilgi birleştirme ve akıl yürütme süreçlerini de desteklemektedir. Örneğin, aynı hikayeye ait farklı olayları ilişkilendirmek, çoklu kaynaklardan gelen bilgilerin anlamlı bir bağlamda birleştirilmesine olanak tanır. Böylece, QA sistemlerinin çok adımlı sorulara daha doğru, bağlamsal ve zamansal açıdan tutarlı yanıtlar verebilmesi sağlanmaktadır.

Sonuç olarak, bu çalışma, dijital haber kaynaklarının yoğun ve karmaşık bilgi yapısını düzenlemeyi, bu bilgiyi yapılandırılmış bir çerçevede temsil etmeyi ve QA sistemlerinin bilgi erişim süreçlerini iyileştirme motivasyonunu taşımaktadır.

1.2. Tezin Amacı, İçeriği ve Katkısı

Bu tez çalışmasında, haber arşivleri gibi büyük ölçekli ve bağlamsal açıdan zengin veri kaynaklarından olay tabanlı bilgi çıkarımının iyileştirilmesi ve bu bilgiyi kullanarak karmaşık sorulara uygun yanıtların otomatik olarak çıkarılması problemi ele alınmaktadır. Bu bağlamda, geleneksel yöntemlere kıyasla daha ileri bir seviyede bilgi çıkarımına olanak tanıyan isimlendirilmiş varlık merkezli olay çizge modeli geliştirilmesi ve bu modelin zaman bilgisi entegrasyonu ile çok adımlı soru yanıtlama performansının

bulunan doküman veya doküman parçalarını yapılandırılmış bilgiye dönüştürerek organize eder. Bu modül, özellikle ilgili haberlerin tespit edilmesi ve olayların kümeleme yoluyla yapılandırılması yoluyla sistemin soru cevaplama performansını doğrudan etkilemektedir.

2. Soru Cevaplama Modülü: Basit sorulardan karmaşık bilgi birleştirme ve akıl yürütme gerektiren sorulara kadar, büyük dil modellerinden faydalanarak anlamsal ayrıştırma yolu ile geniş bir yelpazede yanıt üretebilmektedir. Bu modül, olay örgüsü zinciri yapısını kullanarak, olayların bağlamsal ilişkilerinden yararlanır ve çok adımlı akıl yürütme gerektiren sorulara da doğru yanıtlar sunar.

Bu sistemin yenilikçi özelliklerinden biri, haber metinlerinde aynı hikayeye ait olayların kümeleme ve alt çizge madenciliği teknikleriyle ilişkilendirilerek bütüncül bir yapı oluşturulmasıdır. Olay örgüsü zinciri, olayların bağlamını ve ilişkilerini açık bir şekilde temsil ederken, sistemin çok adımlı soru yanıtlamadaki başarısını artırmaktadır.

Bu tez çalışmasının literatüre sağladığı katkılar aşağıdaki gibidir:

- Soru cevaplama problemi, haber içeriği barındıran metinlerin doğrudan metin analizi yerine çizge tabanlı bir modelleme ile ele alınmıştır. Çizge yapısında metin niteliklerine ek olarak, metinde geçen örtük/açık zaman bilgisi, adlandırılmış varlıklar veya haberin yayınlanma zamanı gibi ek bilgiler de entegre edilmiştir. Özellikle zaman bilgisinin çizge yapısında bulunması, zamanla değişen veya dönüşen bilgilerin takibini kolaylaştırmış; ayrıca, orijinal haber metninde yer almayan zaman ifadeleri ile ilgili sorulara da yanıt sağlanmasını mümkün kılmıştır.
- Soru cevaplama sisteminde arama uzayını daraltarak zaman ve depolama maliyetlerini azaltmak amacıyla, aynı haber içeriğine sahip metinler kümelenmiş ve her kümelenmiş metin grubu, ilgili içeriği temsil eden tek bir metine indirgenmiştir. Bu yaklaşım, gereksiz veri işleme yükünü azaltarak sistemin verimliliğini artırmıştır.
- Haber içeriğindeki olayları oluşturan örüntüleri belirlemek amacıyla sık tekrarlı alt çizge çıkarımı tekniklerinden faydalanılmıştır. Literatürdeki mevcut çalışmalardan farklı olarak, isimlendirilmiş varlık odaklı sık tekrarlı alt çizge keşfi ile olay örgüsünde yer alan diğer haber metinlerinin yüksek başarımla kümelenmesi sağlanmıştır.

- Haber içeriğinin zamanla gelişen olaylarla bağının kurulabilmesi amacıyla, belirli bir olayla ilgili farklı zaman dilimlerinde meydana gelen tüm ilintili olaylar tek bir çizge yapısı altında birleştirilmiştir. Bu sayede, olay örgüsü bütüncül bir yapı olarak öğrenilmiş ve birleştirilmiş olay zinciri yapısının çok adımlı soru cevaplama probleminin çözümüne katkıları değerlendirilmiştir.
- Türkçe doğal dil işleme alanında değerlendirilebilir standartlara uygun ilk Türkçe bilgi tabanlı soru cevaplama veri kümesi ve Türkçe olay odaklı metin kümeleme veri kümesi derlenmiştir. Ayrıca haber alanında gözetimsiz öğrenme yöntemleri için kullanılabilecek geniş ölçekli bir haber metin külliyatı literatüre kazandırılmıştır.

1.3. Hipotez

Bu tezi oluştururken ortaya konan hipotezler aşağıdaki gibidir:

- **H1:** İlgili haberlerin tespitinde, isimlendirilmiş varlık merkezli alt çizgelerle yapılandırılmış olay çizge modelinin kullanımı, geleneksel kümeleme algoritmalarının performansını artırmaktadır.
- **H2:** Aynı hikaye kümesine ait olay çizgelerinin birleştirilmesiyle oluşturulan "olay örgüsü" bilgi grafiği yapısı, çok adımlı soruların yanıtlanmasında "bağımsız olay" bilgi grafiklerinin kullanımına kıyasla daha yüksek bir başarımlı sunmaktadır.
- **H3:** Olay çizge modeline zaman bilgisinin entegre edilmesi, zamanla ilişkili değişimlerin izlenmesini sağlamakta ve örtük zaman bilgisi içeren metinlerden zaman odaklı sorulara yanıt verilmesini mümkün kılarak zaman ifadeli soruların yanıtlanabilirliğini artırmaktadır.

1.4. Tezin Organizasyonu

Tezi organizasyonu altı ana bölüm üzerine kurulmuştur. İkinci bölümde, soru cevaplama probleminin kapsamı tanımlanmakta ve geleneksel soru cevaplama sistemlerinde kullanılan bilgiye erişim modülü ile okuduğunu anlama modülüne dair mevcut literatür yaklaşımları incelenmektedir. Bu bölümde ayrıca, tezde ele alınan bilgi tabanlı soru cevaplama probleminin tanımı yapılmakta ve bu problem, anlamsal ayrıştırma tabanlı

yöntemler ve bilgiye erişim tabanlı yöntemler olmak üzere iki temel yaklaşıma ayrılarak değerlendirilmektedir. Her iki yöntemin çalışma prensipleri ve literatürdeki yeri bu bölümde detaylandırılmaktadır. Son olarak, tez çalışmasında da kullanılan son teknoloji büyük dil modelleri ile bilgi tabanlı soru cevaplama yaklaşımları incelenerek detaylı analizi sunulmuştur.

Üçüncü bölümde tez kapsamında geliştirilen, bilgi grafiği soru cevaplama sisteminin Bilgi Erişim Modülüne dair yapılan ön çalışmalar sunulmaktadır. Bu ön çalışmalar, semantik ayrıştırma için dil modellerinden faydalanan uçtan bilgi grafiği soru cevaplama sistemin altyapısını oluşturmakta ve metodolojik temelleri ortaya koymaktadır. Bölüm 3.1’de metodolojisi detayları ile sunulan aynı içerikli haberlerin tespiti yaklaşımı, Bilgi İşleme Modülünün iki temel bileşeninden ilkidir. Çalışma kapsamında yapılan deney sonuçları uluslararası bir kongrede sunulmuş olup, ayrıca kitap bölümü olarak literatüre kazandırılmıştır [11]. Bölüm 3.2 ise, aynı hikayenin parçası olan olay metinlerinin kümelenmesinde isimlendirilmiş varlıkların etkisini araştırmak üzere tasarlanmış ve sonuçları uluslararası kitap bölümü olarak sunulmuştur [5]. Bu çalışma, tez çalışmasının özgün değerlerinden biri olan Olay Çizge modeline isimlendirilmiş varlıkların entegre edilmesi konusunda bir temel oluşturmaktadır.

Dördüncü bölümde, adlandırılmış varlık merkezli sık tekrarlı alt çizgelerin entegrasyonu ile oluşturulan Olay Çizgesi ayrıntılı bir şekilde ele alınmaktadır. Önerilen çizge modelinin geleneksel metin kümeleme yaklaşımları üzerindeki etkileri analiz edilmekte ve elde edilen sonuçlar SCI-E indeksli uluslararası dergi makalesinde tartışılmaktadır [2].

Büyük dil modelleri kullanılarak geliştirilen anlamsal ayrıştırma tabanlı uçtan uca soru cevaplama sisteminin detayları ve olay çizgesi modelinin bu sistemin performansı üzerindeki etkileri beşinci bölümde sunulmaktadır. Bu bölümde, sistemin deneysel sonuçları ele alınmakta ve önerilen modelin soru cevaplama performansına olan katkıları verilerle desteklenmektedir. Soru cevaplama sistemine dair elde edilen temel bulgular 2024 yılı Ekim ayında uluslararası bir kongrede sunulmuştur [12].

Son bölümde ise doktora tez çalışması kapsamında elde edilen sonuçlar özetlenmiş ve gelecekteki çalışmalar için değerlendirmelerde bulunulmuştur.

2. KAVRAMSAL ÇERÇEVE

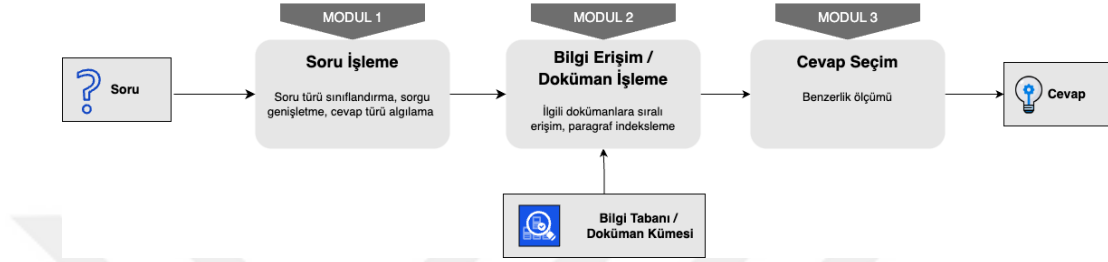
2.1. Soru Cevaplama Problemi

İnsanı bir ihtiyaç olan bilgi arayışı, bilgisayarlar ortaya çıkar çıkmaz insanların onlara sorular sormaya başlaması doğal sonucunu doğurmuştur [13]. Bu sebeple, doğal dil işleme araştırma alanında soru cevaplama probleminin ilk ele alındığı çalışmalar 1960'lı yıllara dayanmaktadır. Bu yıllarda yapılandırılmış veri tabanından, bir sorunun yanıtlanmasına dayalı ilk sistemler literatürde yerini almaya başlamıştır [14, 15]. Ancak bu sistemler uzmanlar aracılığı ile manuel olarak hazırlanmış kurallar ve oldukça dar sınırları olan kelime dağarcığına sahip ve alan uzmanları tarafından oluşturulmuş bilgi tabanlarına dayalı çalışmaktaydı. 1960'lı yıllardan günümüze kadar tasarlanan soru cevaplama sistemleri, farklı yapıdaki okuduğunu anlama modülleri, uygulama alanları, sorular, veri kaynakları, eşleştirme işlevleri veya cevaplar gibi farklı kriterlere göre birçok sınıf altında toplanabilmektedir ve diğer taraftan farklı bir çok alanda ve amaçta kullanım şekilleri de büyük oranda değişiklik gösterebilmektedir. [16, 17].

Soru Cevaplama, doğal dil işleme araştırma alanında bir sistemin, doğal dilde formüle edilmiş bir soru Q ve bir bilgi kaynağı C verildiğinde, doğru yanıt A 'yı üretecek bir f fonksiyonun oluşturulmasını amaçlayan temel bir problem olarak tanımlanır. Bu problem, yanıt üreten fonksiyonun $f : (Q, C) \rightarrow A$ şeklinde ifade edilmesini gerektirir. Burada Q doğal dilde ifade edilen soruyu; C soruya yönelik bilgi içeren bağlamı; A ise modelin döndürmesi beklenen doğru cevabı temsil eder [18]. QA problemi, modeli, soru ve bağlam arasındaki anlamsal ilişkiyi doğru bir şekilde kurmaya zorlayan ve doğru yanıtı seçmesini gerektiren bir optimizasyon problemi olarak ele alınır. Bu problem, yanıtın C bağlamında doğrudan bulunup bulunmadığına göre iki ana kategoriye ayrılır: çıkarımsal QA, cevabın doğrudan bağlamda yer aldığı durumları; üretken QA ise cevabın bağlam bilgisi kullanılarak oluşturulması gereken durumları kapsar. Her iki durumda da amaç, f fonksiyonunun, Q sorusu ile C bağlamı arasında anlamsal bir eşleme gerçekleştirerek en uygun A yanıtını döndürmesini sağlamaktır.

Temelde okuduğunu anlama (machine reading comprehension - MRC) yeteneğini gerektiren ve kullanıcının göndermiş olduğu bir sorguya doğrudan yanıt beklediği bir bilgi

alma görevi olarak da tanımlanan QA; Bilgi Erişim (Information Retrieval-IR) sistemlerinin büyük çoğunluğunda olduğu gibi, kullanıcıya sorgu sonrasında bir doküman dizisinin döndürülmesi yerine bir veya birden fazla kaynaktan toplanmış bilgilerden elde edilen damıtılmış bir yanıtın sağlanması amaçlar ve bir sistemin dili anlama ve muhakeme yeteneğini değerlendirmek için ölçülebilir bir yol sağlar [10, 19].



Şekil 2.1: Geleneksel Soru Cevaplama Sistemi İşlem Hattı [1]

Geleneksel soru cevaplama sistemleri Şekil 2.1’de görüldüğü gibi 3 temel bileşenden oluşur [1]:

1. Soru İşleme Modülü: Doğal dilde ifade edilmiş bir soru Q ’nun sınıflandırılması veya çeşitli örüntü eşleştirme yöntemleri ile soru odağını analiz eden ve tanımlayan modüldür. Probleme bakış açısına göre sorular sınıflandırılabilir, veritabanı sorgusuna dönüştürmek gibi yollarla yeniden formüle edilebilir veya soru içerisindeki öğeler analiz edilerek iç temsiller oluşturulabilir. Matematiksel olarak, Soru İşleme Modülü

$$g_1 : Q \rightarrow T_Q$$

şeklinde tanımlanır. Burada T_Q , sorunun probleme özgü bir bakış açısıyla yeniden yapılandırılmış ve işlemeye uygun bir biçimde temsil edilmiş halini ifade eder.

2. Bilgi Erişim / Doküman İşleme Modülü: Soru işleme modülünün T_Q çıktısını, girdi olarak kabul eder ve bu soru temsiline dayanarak bilgi kaynağı C üzerinde arama yapar, ardından yanıt içerme ihtimali bulunan bir veya bir dizi ilgili dokümanı D döndürür. Bu süreçte, dokümanlar ilgi düzeylerine göre filtrelenebilir ve sıralanır. Matematiksel olarak bu modül,

$$g_2 : T_Q \times C \rightarrow D$$

fonksiyonu ile ifade edilir. Burada $D = \{d_1, d_2, \dots, d_m\}$, C bilgi kaynağından seçilen ve T_Q ile en iyi eşleşen dokümanlar kümesini temsil eder. Bu modülün başarısı, T_Q ile d_i dokümanları arasındaki ilişkinin değerlendirilmesine ve dokümanların sıralama ve filtreleme etkinliğine bağlıdır; böylece nihai cevabın elde edilmesi için en uygun bağlam sağlanmış olur.

3. Cevap Seçim Modülü: İlgi düzeyine göre sıralanmış ve/veya filtrelenmiş dokümanlar kümesi D 'yi girdi olarak alan ve bu dokümanlardan aday cevapları çıkararak nihai cevabı A olarak döndüren modüldür. Bu modül, sorunun temsili T_Q ile dokümanların içeriklerini karşılaştırmak için çeşitli benzerlik ölçütleri ve doğrulama yöntemleri kullanır. Matematiksel olarak, bu süreç

$$g_3 : (T_Q, D) \rightarrow A$$

fonksiyonu ile ifade edilir. Burada A , D bağlamında bulunan en uygun cevabı temsil eder ve modülün temel amacı, dokümanların anlamsal içeriğini analiz ederek doğru cevabı seçmektir.

Soru işleme, doküman işleme ve cevap çıkarma gibi ayrık modülleri bünyesinde barındıran ve geleneksel bilgiye erişim tabanlı soru cevaplama sistemleri, son yıllarda yerini uçtan uca tasarlanmış sistemlere bırakmaya başlamıştır. Bu yenilikçi yaklaşımda, boru hattı tasarımları ile ara modüller baypas edilerek uçtan uca bir model kullanma yoluna gidilmektedir.

Haber arşivleri gibi büyük ölçekli veri barındıran sistemlerde de verimli bir şekilde soru cevaplamanın uygulanabilmesi noktasında, büyüklükleri ve karmaşıklıkları nedeniyle geleneksel bilgiye erişim tabanlı soru cevaplama sistemlerine benzer, ancak ara modüllerin en aza indirildiği 2 adımlı bir süreç yürütülür [6, 20, 21]. Bu süreçlerin ilki doküman işleme modülü ile aynı işleve sahip olan ve en genel ifade ile bilgiye erişim (IR) olarak isimlendirilen süreçtir [22]. Büyük ölçekli veri arşivlerinde bu sürecin boru hattının dışında tutularak, verinin türüne ve yapısına bağlı olarak özelleştirilmesi soru cevaplama probleminin çözülmesinde dar boğaz oluşturmaması açısından önem taşımaktadır [23]. Sürecin ikinci parçasında ise nöral bir okuduğunu anlama algoritması ile IR modülü tarafından elde edilen herbir dokümanın üzerinden geçilerek, ilgili soruyu yanıtlama olasılığı olan aralıklar bulunur [13].

2.1.1. Bilgiye Erişim Modülü

Özellikle 2018 yılı sonrasında doğal dil işleme alanında yeni bir çağın başlamasına sebep olan dil modelleri, doğrudan metinlerden soruların yanıtlanması için ara modüllere ihtiyaç duymaksızın uzmanlaşabilen sistemlerin gelişmesinde de önemli rol oynamaktadır. Öyle ki geliştirilen dil modelleri, değerlendirme deneylerinde kullanılan veri kümelerinde insan düzeyindeki performansın üstüne çıkmayı başarmıştır [24]. Büyük veri arşivleri üzerine kurulacak olan soru cevaplama sistemlerinde bilgiye erişim modülünün rolünü inceleyen son araştırmalar göstermiştir ki, IR Modülü'nün sağladığı gücülü veri soru yanıtlama sistemlerinin performansını büyük ölçüde etkileyebilecek ciddi bir darboğaz oluşturmaktadır. Bu sebeple birçok çalışma yalnızca IR modülünün iyileştirilmesi üzerine yoğunlaşmıştır [13, 23, 25, 26].

Önceleri DrQA [21] gibi bilinen en iyi büyük ölçekli soru cevaplama sistemlerinde, TF-IDF ve BM25 gibi terim tabanlı yaklaşımlar kullanılmıştır. Diğer taraftan Wang ve diğerleri IR modülünün ve MRC modülünün pekiştirmeli öğrenme ile birlikte eğitildiği R3 modeli ile daha önce benzerine rastlanmayan yenilikçi bir bakış açısı sunmuşlardır [27]. 2018 yılında sunulan bir diğer çalışmada ise [28], kullanıcıdan alınan sorunun temel terimlerini belirleyerek ve sorguyu yeniden formüle ederek IR modülünü geliştiren ve ET-RR olarak isimlendirilen bir modeli önermişlerdir. Belgeleri yinelemeli olarak erişen, yeniden sıralayan ve filtreleyen, erişim sürecinin ne zaman durdurulacağını da uyarlanabilir bir şekilde belirleyen dinamik yapıda bir IR sistemi ise 2020 yılında Nie ve diğerleri tarafından sunulmuştur [29]. IR modülünü BERT tabanlı okuduğunu anlama modülü ile entegre ederek BERTserini [23] ismi verilen yaklaşım ise, terim tabanlı seyrek erişim yaklaşımlarından farklılaşarak, son yıllarda hem soruların hem de belgelerin temsilde yoğun vektörlerin kullanılması ve yoğun geçiş erişim modellerine bir örnektir [22, 25, 26]. Bu modeller, geleneksel yöntemlere göre önemli iyileştirmeler sağlamakla birlikte, büyük miktarda eğitim verisine ihtiyaç duymaktadır.

Bilgi Erişiminde Zaman Bilgisi Kullanımı: Mevcut soru cevaplama sistemleri genel itibari ile Wikipedia gibi eşzamanlı belge koleksiyonları için tasarlandığından, uzun vadeli haber arşivlerindeki soruları cevaplarken, haber makalelerinde geçen olayların oldukça önemli olan zamansal bilgilerini kullanmazlar [30]. Ancak arama işlevlerinde zaman önemli bir rol oynar, çünkü çoğu veri kaynağı örtük ve/ya açık zamansal bil-

giler içerir ve kullanıcıdan gelen birçok sorgu da yine zamanla ilgilidir [31]. Konu ile ilgili yapılan bir araştırmada, web arama motorunun sorgu günlükleri incelendiğinde, sorguların neredeyse %14'ünün açık zamansal bilgiler içerdiği gösterilmiştir [32]. Daha önceki çalışmaların birkaçında soru ve belge içeriğinin zamansal bilgisinden yararlanabilen zamansal soru cevaplama sistemleri önerilmiş olsa da [33–36], bu önerilen sistemlerin tümü yine eşzamanlı belge koleksiyonları için tasarlanmış olan ve ayrıca geleneksel kural tabanlı, performansları oldukça düşük sistemlerdir. Diğer taraftan veri arşivleri içerisindeki belgelerde yer alan zamansal ifadelerden yararlanan ve o belgeleri zaman çizelgeleri boyunca kümeleyerek bilgiye erişimi kolaylaştıran, farklı bakış açısına sahip çalışmalar da literatürde mevcuttur [37]. Geleneksel kural tabanlı yapıdan farklı olarak, Kanhabua ve diğerleri kullanıcıdan alınan soruların örtük zamansal kapsamını belirlemek için üç farklı yöntem önermiştir ve bu bilgileri, belgelerin yeniden sıralanması ve dolayısıyla belgeye erişim etkinliğini artırmak için kullanmışlardır [38]. 2020 yılında yayınlanan çalışmada ise eşzamanlı olmayan dokümanların zaman bilgilerinden yararlanılmış ve IR modülü üzerinde oldukça büyük iyileştirmeler sağlamıştır [20]. Farklı açılardan zamansal bilgilerden yararlanarak doğru belgelerin alınmasını iyileştiren ve belgeleri yeniden sıralamak için kullanılabilecek ek bir modül öneren bu çalışma; sadece sorulardan çıkarılan zaman kapsamı bilgilerini kullanmakla kalmamakta, aynı zamanda bunu dokümanın zaman bilgileri ve doküman içeriğinde gömülü zamansal bilgilerle de birleştirmektedir [20]. Özellikle yalnızca eşzamanlı bilgi değil, farklı zaman dilimlerine ait birçok bilgiyi barındıran arşiv verilerinden bilgiye erişimde, dokümanların zamana dayalı sıralama ve filtrelemelerinin oldukça önemli olduğunu açıklayan araştırma çalışmaları da literatürde mevcuttur [39, 40]. Bu noktada, uzun vadeli zamansal haber makaleleri koleksiyonları üzerinde etkili bir şekilde çalışacak bir soru cevaplama sisteminin tasarlanmasında, tahmini soru zamanı bilgisinden faydalanarak, bu bilgiyi alınan her belgeden çıkarılan içerik zamansal bilgisi ile entegre eden bir modül içeren QANA modeli literatürde bir ilki temsil eder. Belgeleri yeniden sıralamak için kullanılan ve farklı açılardan zamansal bilgileri kullanarak doğru belgelerin alınmasını iyileştirir ve araştırmacılar bu probleme yönelik gelişmiş bir veri kümesi sunar [6].

Konu Tespiti ve Takibi Problemi: Konu Tespiti ve Takibi (Topic Detection and Tracking-TDT), farklı medya platformlarından ve diğer kaynaklardan haber niteliği

taşıyan metinlerin aranmasına ve düzenlenmesine yardımcı olmayı amaçlayan bir araştırma alanıdır. Literatürde ilk olarak 1998 yılında Allan ve diğerleri tarafından tanımlanan [41] Konu Tespiti ve Takibi (TDT) görevinin birincil amacı, bir dizi haber makalesini veya haber akışını “hikaye” adı verilen olay grupları halinde düzenlemektir. Belge kümeleme problemi, belirli bir belge koleksiyonundaki öğeler arasındaki benzerliklere dayalı olarak kümeleri tanımlamayı amaçlayan denetimsiz yöntemler gerektirir. Bu yöntemler, benzer belgeleri anlamsal bir vektör uzayında yoğun kümeler oluşturarak temsil etmeye çalışırken, benzer olmayan belgeler birbirinden daha uzakta yer alma eğilimindedir. Bu bakımdan, olay içeriği taşıyan haber metinlerinde kullanılan temsil yöntemi, uygun bir kümeleme yönteminin seçilmesinde önemli bir rol oynamaktadır [42].

Temelde incelenen problem bir metin kümeleme problemidir ve bu problemin çözümü için literatürde önerilen yaklaşımları birkaç gruba ayırarak incelemek mümkündür. Bu yaklaşımlar; vektör uzayı modelleri (vector space models), k-means varyasyonları, üretici (generative) algoritmalar, boyut indirgeme tabanlı (dimensionality reduction based) yöntemler ve ifade tabanlı (phrased-based) metotlardır [43, 44]. Metin tabanlı konu tespiti ve takibinde en yaygın kullanılan temsil modeli, bir belgedeki kelimeleri ve bunlara karşılık gelen terim ağırlıklarını [45–47] kullanan Vektör Uzayı Modelidir (VSM). Vektör uzayı modellerinde, kelimelerin sırası, birbirleriyle ilişkileri veya anlamsal bilgiler dikkate alınmadan yalnızca terim ağırlıkları önemli kabul edilir [45]. Sonuç olarak, VSM kullanan TDT yaklaşımları terim tabanlı olay tespit teknikleri olarak adlandırılır. Bu yöntemlerin uygulandığı çalışmaların [46, 48, 49] gözden geçirilmesine dayanarak, yaklaşım genel olarak aşağıdaki gibi tanımlanabilir: İlk olarak, bir dizi ön işleme adımıyla belgeler üzerinde kelime ve anahtar kelime çıkarımı gerçekleştirilir. Ardından dokümanlar, TDT görevinde sıklıkla kullanılan Terim Frekansı-Ters Doküman Frekansı (TF-IDF) [50, 51], İsimlendirilmiş Varlık Tanıma (NER) [52, 53] ve Kelime Torbası [54] gibi istatistik tabanlı vektör uzayı modellerine tabi tutularak bir kümeleme algoritması vasıtasıyla görev tamamlanır. Vektör uzayı modeller aslında otomatik indeksleme problemi için geliştirilmiş olan, ancak kümelerin homojen olarak dağıldığı ve küme sayısının bilindiği durumlarda kümeleme görevi için de başarılı sonuçlar veren yöntemlerdir. Geçmişte oldukça popüler olan bu yaklaşımlar, neredeyse tüm metin temsillerinde ve doğal dil işleme görevlerinde yüksek performansa sahip derin

öğrenme tabanlı yöntemlerin ortaya çıkmasıyla birlikte geri planda kalmıştır [45, 55]. Kelime, cümle ve belge gömme gibi farklı seviyelerdeki gömme teknikleri, anlamsal olarak benzer öğelerin vektör uzayında yakın bir şekilde konumlandırılmasını sağlar [56]. Çeşitli çalışmalar [55, 57, 58] son teknoloji kelime, cümle ve belge gömme tekniklerinin TDT görevlerinde VSM modellerinden daha iyi performans gösterdiğini ortaya koymuştur. Ayrıca, metin temsilindeki seyreklik ve boyutluluk sorunlarından da kaçınılabilirler [45], çok dilli temsil yöntemlerinin keşfi, dile özgü özelliklere bağlı temsil yöntemlerine göre önemli bir gelişme olarak kaydedilmiştir [59–61]. Bir diğer yenilikçi bakış açısı ise haber belgelerindeki olayları önceden belirlenmiş bir yapısal formda temsil etmek ve bu yapısal forma özgü kümeleme yaklaşımlarını uygulamaktır. Elemanların karmaşık ilişkilerini temsil etme yeteneği yüksek olan çizgeler, olay temsili için idealdir ve olay kümeleri doğrudan öznitelik veya yapı tabanlı çizge kümeleme yaklaşımları kullanılarak oluşturulabilir [62, 63]. Öznitelik tabanlı çizge kümelemede temel zorluk, düğüm bağlam bilgisini kullanmak ve yapısal öğelerin karşılıklı ilişkilerini yakalamaktır. Bu sorunu çözmek için son teknoloji derin öğrenme tabanlı çizge temsil yöntemleri geliştirilmiştir. Karmaşık çizge yapıları içindeki olayların zamansal, uzamsal ve aktör bileşenleri arasındaki zamansal, uzamsal ve doğrusal olmayan ilişkileri korudukları için TDT görevleri için büyük umut vaat etmektedirler [64, 65]. Çizge gömmeleri basittir, geleneksel kümeleme algoritmalarına doğrudan girdi olarak kullanılabilir ve yüksek seviyeli sonuçlar elde eder [66, 67].

Son yıllarda, alt çizgenin önemli özelliklerini korurken düğüm veya çizge gömmelerini öğrenmeyi amaçlayan çeşitli yaklaşımlar ortaya çıkmıştır. Bu yaklaşımlar arasında matris çarpanlarına ayırma tabanlı yöntemler [68, 69], skip-gram tabanlı yöntemler [70, 71] ve literatürde bulunan derin sinir ağı tabanlı yöntemler [72, 73] yer almaktadır. Bir bitişiklik matrisi ile temsil edilen bir çizge topolojisinde, N boyutundaki bir satır veya sütun vektörü (burada N düğüm sayısıdır) tipik olarak çizgedeki her bir düğümü temsil etmek için kullanılır. Bitişiklik matrislerine uygulanan matris çarpanlarına ayırma teknikleri, bir çizgeyi temsil edecek düşük boyutlu bir uzay bulmak için yaygın olarak kullanılmaktadır [74, 75]. Doğal dil işlemeye önemli katkıları olan Word2Vec yaklaşımı [56] da çizge gömme tekniklerine ilham vermiştir. Word2Vec'in temel konsepti, bir kelime için komşu kelimelerin vektörlerine dayanan ve birlikte oluşma örüntüleri tarafından belirlenen bir vektör temsili oluşturmaktır. Benzer şekilde, çizge yerleştirmede,

rastgele yürüyüş modellerinin ve düğüm komşuluklarının birlikte oluşma modelleri, çizgedeki her düğümün bir cümledeki bir kelime olarak ele alınmasıyla hesaplanır [76]. DeepWalk [71] ve Node2Vec [70] popüler skip-gram tabanlı yöntemlerdir. Bu yöntemler, çizge yerleştirme görevleri de dahil olmak üzere karmaşık sorunlara uçtan uca çözümler sağlamasıyla bilinen derin sinir ağı modellerinden yararlanır.

Çizge gömme yöntemleri aracılığıyla olay temsilleri elde etmeye odaklanan çalışmalar, öncelikle çizge topolojisini ve düğüm semantiğini korurken bireysel olay özelliklerini temsil edebilen ve olay ağı komşuluklarını yakalayabilen çizge modellerini veya kodlayıcıları araştırmıştır. Bir çalışma, olay özelliklerinden oluşturulan bir ağ üzerinde yaygın olarak kullanılan Node2Vec düğüm gömme yaklaşımından yararlanırken, başka bir yaklaşım her bir haber özelliğini bir vektörle eşleştirir ve olayı bu özelliklerin toplamı olarak temsil eder [77]. Zeng ve arkadaşları, derlem düzeyinde bir çizge kodlayıcı kullanarak bir olay ağını kodlayan ve böylece olay temsili için genel bir bağlam sağlayan yeni bir çerçeve önermiştir [78]. Buna ek olarak, yenilikçi bir çalışmada, araştırmacılar önemli olayları tespit etmek ve izlemek için tasarlanmış dinamik bir kelime çizgesi önermişlerdir [79].

Gazeteciler genellikle haber değeri taşıyan olayları metinsel belgeler şeklinde kamuoyuna sunarlar. Bu nedenle, TDT görevi birçok çalışmada bir belge kümeleme problemi olarak ele alınmıştır. Geleneksel kümeleme algoritmaları, haber belgelerini temsil eden farklı türdeki yoğun vektörlere yaygın olarak uygulanmış ve bunların avantajları ve dezavantajları daha önce tartışılmıştır. Bu algoritmalar arasında k-ortalamlar, hiyerarşik kümeleme ve DBSCAN literatürde en sık kullanılan yaklaşımlardır [61, 80, 81]. Geleneksel kümeleme algoritmalarının çoğu başlangıçta küme sayısının ('k') zorunlu bir parametre olduğunu varsaymıştır. Ancak, gerçek zamanlı haber sistemleri veya büyük haber yığınları halinde çalışan sistemler için küme sayısını belirlemek zordur.

Diğer taraftan literatürde yer alan hiyerarşik ve bölümsel algoritmaların kümeleme problemlerinin çözümünde her zaman dominant rol oynadığı bilinmektedir. Bu algoritmalar k-means ve onun varyasyonları sıklıkla kullanılmaktadır [82–86]. Ancak özellikle büyük korpuslar üzerinde etkinliklerinin azalması yönü bu algoritmaları dezavantajlı pozisyona getirmektedir. Üretici algoritmaların en büyük problemi ise; aykırı değerlere karşı olan hassasiyetlerinin, homojen olarak dağılmayan veriler üzerinde çalışmayı zorlaştırmasıdır. Küme sayısını girdi olarak alan bu üretici algoritmalar

en bilinenleri için Gaussian Model [87], Expectation Maximization [87, 88] ve von Mises-Fisher model [89, 90] listelenebilir. Küme sayısını girdi olarak almaya ihtiyaç duymayan ve verinin vektör modeli bipartite graph şeklinde ifade edilebilmesi durumunda oldukça başarılı çalışan kümeleme algoritmaları ise “spektral kümeleme” ismi ile anılırlar [91, 92]. Boyut azaltma kümeleme performansını artıran bir ön işleme adımı olduğu gibi, başlı başına kümeleme işlemi için de kullanılabilir. Görüntü işleme uygulamalarında sıklıkla kullanılan Principal Component Analysis [93] ve Non-negative Matrix Factorization [94, 95] gibi yöntemlerin metin kümeleme uygulamaları da dikkat çekmektedir. Ayrıca Soft Spectral Coclustering [96] ve Lingo’da [97] yine boyut azaltma yaklaşımı ile kümeleme yapabilen diğer yöntemlere örnek olarak gösterilebilir. Diğer kümeleme algoritmalarından farklı bir şekilde kelime dizilimlerine bağlı olarak kümeleme yapan metotlar ise Phrased-based olarak anılır ve genellikle kümeleme performansları diğer yöntemlerle kıyaslandığında daha düşüktür [98, 99]. Temel olarak kategorilere ayrılmış metin kümeleme yaklaşımlarına ait örnek çalışmaların neredeyse tamamı önceden kaydedilmiş haber metinleri koleksiyonları üzerinde çalışan uygulamalardır. Gerçek zamanlı çalışan sistemlerin sayısı oldukça azdır. Literatürde akan veri üzerinde çalışan kümeleme algoritmalarından ilklerden biri 2002 yılında yayınlanmıştır ve veri akışını parçalara bölerek her bir parçaya k-means kümeleme algoritmasını uygulamayı önermektedir [100]. Bu yönteme alternatif olarak önerilen bir başka yaklaşım ise bir sonraki yıl Brain B. ve diğerleri tarafından sunulmuştur ve bu yaklaşımdaki göze çarpan en büyük fark ‘kayan pencere’ yapısının kullanılıyor olmasıdır [101]. Linghui G., Jianping Z. ve Shiyong Z. tarafından 2011 yılında yapılan çalışmada, akan veri üzerinde adaptif metin öznitelik seçimi kullanarak metin kümelemesi gerçekleştirilmiştir [102]. 2018 yılında ise tek geçişli akan veri içerisindeki kümeleri eliptik fonksiyonla tespit edebilen BEstream metodu Niwan W., Saranya M. ve Chidchanok L. tarafından tanıtılmıştır [103]. Aynı yıl da sunulan başka bir çalışmada ise akan haber metinleri verileri ile çalışan ve Almanca, İspanyolca ve İngilizce haber metinleri üzerinde aynı anda çalışarak farklı dillerde kümeleme problemine çözüm üreten bir çalışma Miranda S., Zinotins A. ve diğerleri tarafından sunulmuştur [55].

2.1.2. Okuduğunu Anlama Modülü

Büyük ölçekli veri barındıran arşivler üzerinde çalışan soru cevaplama sistemlerinde, IR modülünden sağlanan filtrelenmiş ve/veya sıralanmış dokümanlardan nihai yanıtın elde edildiği süreç MRC modülü tarafından yürütülür. IR modülünün çıktısı olan bir veya birden fazla doküman içerisinde, kullanıcının sorusunun yanıtlanması diğer bir deyişle soru cevaplama işlevinin yürütüldüğü modül olarak da açıklanabilmektedir. Bu sebeple MRC modülünün gerçekleştirmeyi hedeflediği görevin kendisi de çoğu kaynakta “soru cevaplama” olarak adlandırılır [104, 105]. Formal bir tanımla, “MRC makinenin bir metni okuma, işleme, anlamını kavrama ve ilgili soruları yanıtlama yeteneğidir” [104, 106].

MRC sistemlerini geliştirmek için kullanılan yaklaşımları üç kategoride gruplandırılabilmesi mümkündür [18, 107]. İlk kategori, manuel olarak hazırlanmış kurallara bağlı çalışan ve farklı doküman türlerinin tamamı için geçerli olan kuralların tamamının elde edilmesinin zorluğundan muzdarip olan kural tabanlı yöntemleri içerir. QUARC ismi verilen, kısa hikaye metinlerine ait soruların yanıtlanması için tasarlanan kural tabanlı sistem örnek olarak verilebilir [108, 109]. Kullanıcının sorusunun yanıtlanmasında, en genel yaklaşımlardan biri [110, 111] olan soru cümlesinden yola çıkarak türlerinin adlandırılmış varlıklara (PERSON, LOCATION, vb.) göre sınıflandırılmasını temel alır. Örneğin; “Türkiye’nin başkenti neresidir?” sorusu için “LOCATION” türünde bir yanıt bekleniyor olması gibi. QUARC ile aynı temele dayanan ve Arapça dili için geliştirilmiş olan, ancak adlandırılmış varlıklar haricinde soruyu yeniden formüle etme, kökten ayırma ve kök çıkarma gibi birçok doğal dil işleme bileşeninden de yararlanan bir diğer örnek de QArabPro’dur [112].

MRC sistemleri için sıklıkla kullanılan ikinci yaklaşım ise makine öğrenmesine yöntemlerine dayanmaktadır. Kural tabanlı sistemlerde ki manuel olarak kural çıkarma ve bunu yapabilmek için uzman ihtiyacının olması durumunun bir benzeri makine öğrenmesi yöntemlerinde de karşımıza çıkmaktadır. Bu kez yine uzman bilgisi ile özniteliklerin belirlenmesi ihtiyacı vardır. Makine öğrenmesi yöntemleri uzman tarafından belirlenen bir dizi öznitelige dayanır ve ayrıca girdi ve çıktı özelliklerini eşleştirmek için bir model eğitilmektedir. 2000 yılında önerilen makine öğrenmesi tabanlı MRC sisteminde, yalnızca model önerilmemiş aynı zamanda bağlam cümlesinden çıkarılabilecek

yirmiye yakın öznitelik de tanıtılmıştır [113]. MiPACQ Klinik Soru Cevaplama Sistemi ise 2011 yılında geliştirilen ve Medpedia sağlık ve tıp ansiklopedisinden oluşturulan bir veri kümesi ile eğitilen ve değerlendirilen makine öğrenmesi tabanlı bir MRC sistemidir. Araştırmacıların daha önce tasarladıkları kural tabanlı sistemle kıyaslandığında önemli derecede gelişme gösterdiği görülmüştür [114]. IR sistemlerinde yaygın olarak kullanılan TF-IDF tekniği, MRC görevlerinde de kendisine yer bulmuştur. Aday cevapların sunulması durumunda (çoktan seçmeli sınav soruları gibi) özellikle IR sistemi ile desteklenmiş soru cevaplama sistemlerinde makine öğrenmesi algoritmaları ile önemli dereceye de yanıt elde etme kapasitesine sahiptir [115]. WIKIHOP gibi çok belgeli veri kümelerinde de elde ettiği yüksek başarımlarla yaygın olarak kullanılmaktadır [116]. SQUAD veri kümesinin tanıtıldığı akademik makalede önerilen yaklaşımda ise aday yanıtlardan uzunluklar, bigram frekansları, kelime frekansları, POS etiketleri, sözlüksel özellikler, bağımlılık ağacı yolu özellikleri vb. dahil olmak üzere çok sayıda özellik çıkarılmıştır ve tüm bu bilgilere dayanarak Lojistik Regresyon ile bir metin aralığının nihai cevap olup olmadığı tahmin edilmiştir [117].

Üçüncü yaklaşımda, MRC sistemi kullanıcıdan alınan ham girdilerden özellikleri otomatik olarak öğrenmek için derin öğrenme yöntemlerini kullanılmaktadır [107]. Bu yöntemler, yüksek doğrulukta modeller oluşturmak için makine öğrenmesi modellerine nazaran çok büyük miktarda eğitim verisine ihtiyaç duyar. Son yıllarda oluşturulan çok büyük miktarda veri içeren veri kümeleri, korpuslar ve hesaplama gücünün artması nedeniyle, derin öğrenme yöntemleri birçok görevde son teknoloji ürünü sonuçlar elde etmiştir. MRC görevinde, güncel araştırmaların çoğu da bu kategoriye girmektedir ve bu konu üzerinde çalışan araştırmacılar tarafından kullanılan iki ana derin öğrenme mimarisi tekrarlayan sinir ağı (RNN) ve evrişimli sinir ağıdır (CNN). RNN'ler genellikle sıralı verileri modellemek için kullanılır. MRC sistemlerinde tek yönlü ve çift yönlü versiyonlarda Long Short-Term Memory (LSTM) ve Gated Recurrent Unit (GRU) olmak üzere iki türü oldukça sık kullanılmaktadır [21, 118–120]. 2017 yılı sonrasında ise dikkat temelli transformer yapısı [121], RNN mimarisine güçlü bir alternatif olarak ortaya çıkmıştır. Derin öğrenme tabanlı MRC sistemleri gömme aşaması (embedding phase), akıl yürütme aşaması (reasoning phase) ve tahmin aşaması (prediction phase) olmak üzere 3 aşamalı bir yürütme sağlarlar.

İlk aşama olan gömme aşamasında, girdi karakterleri, sözcükler veya cümlelerin, anlamlı

bir uzayda gerçek değerli yoğun vektörlerle temsil edilmesi sağlanır. Yalnızca MRC görevi için değil, MRC'nin alt aşamalarından biri olan giriş temsilini oluşturmak için de derin sinir ağı modelleri yaygın olarak kullanılır. Örneğin, 2014 yılında yayınlanan çalışmanın [122] ardından birçok araştırmacı karakter gömmelerini elde etmek için CNN modellerini kullanmıştır [123, 124]. Benzer şekilde bazı çalışmalarda ise, LSTM / BiLSTM ve GRU / BiGRU gibi bir RNN modelinin son katmanından alınan karakter seviyesi bilgilerini kullanmıştır [125–127]. Bağlamsal kelime gömmelerinin (conceptual word embedding) öğrenilmesi için de genellikle RNN'den faydalanılır. 2017 yılında Chen ve diğerleri bu amaçla çok katmanlı bir BiLSTM modeli kullanmıştır [21]. 2020 yılında yapılan bir çalışmada ise ileri ve geri GRU'nın (backward and forward GRU) gizli durumları (hidden states) birleştirilerek, soru ve bağlam kelimelerinin temsilleri oluşturulmuştur [128]. ELMo, BERT ve GPT gibi önceden eğitilmiş dil modelleri büyük etiketlenmemiş korpuslar üzerinde eğitilmiş derin sinirsel dil modelleridir. Bu dil modelleri ile de diğer bahsedilen derin öğrenme modellerinde olduğu gibi vektör temsilleri elde edilebilmektedir. ELMo yöntemi, bağlamsallaştırılmış temsilleri 2 katmanlı bir Bi-LSTM ile elde ederken, BERT ve GPT sırasıyla çift yönlü ve tek yönlü transformatör tabanlı dil modelleridir. Dil modellerinden elde edilen bu temsiller, diğer temsillerin yanında [129, 130] veya tek başına kullanılabilirler [131].

Akıl yürütme aşamasının amacı ise girdi sorgusunu (soru) girdi doküman/dokümanları (bağlam) ile eşleştirmektir. Başka bir deyişle, soru ve bağlamın bölümleri arasındaki ilgiyi hesaplayarak, yanıtın çıkarılacağı bölümü belirlemektir. Makine çevirisi için tanıtılan ancak sonraları birçok dil işleme görevinde kullanılan dikkat mekanizması [132], bu aşama için de başarı ile kullanılmaktadır. RNN gibi yaygın sıralı yapılara alternatif olarak dikkat tabanlı transformatör mimarisinin [121] ortaya çıkmasıyla birlikte, BERT [24] ve XLNet [133] gibi yeni transformatör tabanlı dil modelleri tanıtılmıştır ve günümüzde literatüre geçen en başarılı sonuçlarda bu modellerin son katmanlarında ekleme/değiştirme veya bu dil modellerini hedef görevler için ince ayar yapılması ile elde edilmektedir [131, 134].

Son aşamada ise cevap tahmini yapılmaktadır. Çıktı (cevap) bağlamdan çıkarılabileceği gibi bağlama göre de oluşturulabilmektedir. Cevabın oluşturulması ihtiyacına yönelik bir çalışmada, kod çözücü modülü ile tek tek yanıt sözcükleri üretilmektedir [135]. Çoktan seçmeli soruların yer aldığı veri kümelerinde ise, sisteme birden fazla seçenek sunulmak-

tadır. Bu durumda, hem çıkarımsal hem de üretici modellerle soru ve paragrafa göre en iyi cevabın seçilmesi işlemi yapılabilir. Ancak SQUAD veri kümesinde olduğu gibi çıkarımsal bir cevaplama ihtiyacı mevcutsa, indekslerin tüm bağlam üzerindeki olasılık dağılımı tahmin edilerek aralığın başlangıç ve bitiş endeksleri tahmin edilir [120, 127]. Bazı çalışmalarda, önce eğitilmiş bir model tarafından sıralanmış aday parçalar (cevaplar) çıkarılır. Bu parçalar cümleler [136] veya varlıklar [137] olabilir.

Son teknoloji birçok yaklaşımda MRC'yi iyileştirmek için içsel veya dolaylı olarak bir anlamsal eşleştirme modülü kullanılmıştır [138, 139]. Bu modül, bir bağlam parçasının sorunun yanıtlanması ile ilgili olup olmadığına karar vermek için tasarlanmıştır ve bağlamdaki ilgisiz cümlelerin yok sayılmasına yardımcı olur [140–142]. Bu modülün diğer bir avantajı da, girdi bağlamının azaltılması yoluyla, doğruluk üzerinde neredeyse hiçbir etkisi olmadan, hesaplama maliyetlerini önemli ölçüde azaltılmasıdır [143]. Cevap adaylarının yeniden sıralanması ve puanlanması için, MRC modülünün çıktısını elde ettikten sonra bir doğrulama yöntemi olarak da anlamsal eşleştirme modülünü kullanan çalışmalar da literatürde mevcuttur [144, 145].

2.2. Bilgi Tabanlı Soru Cevaplama Problemi

1960'lı yıllarda doğal dil işleme araştırma alanındaki ilk ve öncü çalışmalardan biri olarak kabul edilen BASEBALL sistemine [146] ilham olan bilgi tabanlı soru cevaplama problemi (Knowledge Base Question Answering-KBQA), bir doğal dil sorusunu, bir veritabanı üzerinde yapılandırılmış bir sorguyla eşleştirerek yanıtlama görevi olarak tanımlanır [13]. Bu görevde ele alınan ontoloji tabanlı bilgi "bilgi çizgesi (Knowledge Graph-KG)" olarak adlandırılır ve [özne, ilişki, nesne] üçlüsü olarak temsil edilir. Bilgi çizgesi üzerindeki düğümler özne ve nesneleri, düğümler arasındaki kenarlar ise özne ve nesneler arasındaki ilişkiyi gösterir [147].

KBQA problemine ilişkin literatürde yer alan önceki çalışmalar, bir varlık ve yanıtın arasında yalnızca tek bir bağlantı bulunan basit soruları yanıtlamaya odaklanmaktadır ve genellikle önceden tanımlanmış manuel şablonları temel alarak yapılandırılmamış soruları yapılandırılmış mantıksal formlara dönüştürmek için bu şablonları kullanan yaklaşımları benimsemektedir [148–150]. Yoğun iş gücü gereksinimi ve dilbilgisi alan uzmanlığı bilgisine hakimiyet gibi oldukça büyük zorlukların yanı sıra, sabit

kurallara ve şablonlara bağımlı yapısı ile çoklu ilişki çıkarımı yapamaması nedeniyle, bu yaklaşımların uygulanabilirliğini kısıtlamaktadır . Gerçek uygulama senaryolarında esnek kullanıma olanak sağlayabilmek için, araştırmacılar bir ilişki zinciri üzerinde dolaşımı ve çok adımlı çıkarım gerektiren karmaşık soruları ele alarak çok adımlı (multi-hop) soru cevaplama problemi üzerinde yenilikçi çözümler bulmaya odaklanmışlardır [151]. Anlamsal ayrıştırma ve bilgiye erişim tabanlı yöntemler olmak üzere iki yaygın paradigma literatürde sıkça kullanılmaktadır [13].

2.2.1. Anlamsal Ayrıştırma Tabanlı Yöntemler

Anlamsal ayrıştırma tabanlı yaklaşımlarda (Semantic Parsing- SP) geleneksel olarak, doğal dilde verilen bir sorunun mantıksal forma dönüşümünü sağlamak için öncelikle soruda yer alan özne, nesne ve ilişkilerin ayrıştırılması, ardından sorunun yapılandırılmış ve çalıştırılabilir bir mantıksal forma dönüşümü ve son olarak ise, sorgu dilinde ifade edilen soru kullanılarak KG üzerinde arama prosedürünün tamamlanması ile yanıtın elde edilmesi adımlarını kapsayan bir işlem hattı kurgulanır [152–154]. Sorgunun adım adım inşasında, çoğunlukla soru sözdizimi analizi, varlık bağlantısı, ilişki çıkarımı veya pekiştirmeli öğrenme gibi yaklaşımlardan faydalanılır [155]. Ancak son teknoloji yaklaşımlarda anlamsal ayrıştırıcının tasarlanması bir sinirsel çeviri görevi olarak ele alınır ve böylece mantıksal, yapılandırılmış formdaki sorgular doğal dildeki sorulardan sıralı bir şekilde doğrudan üretilir [156, 157]. SP-tabanlı yaklaşımlar, her bir örnek için alt çizge çıkarımına gerek duymaması yönüyle IR-tabanlı yaklaşımlara göre avantaja sahiptir [158].

SP-tabanlı yaklaşımlarda farklı stratejiler uygulanarak üretilen mantıksal sorgular, bilgi tabanı üzerinde çalıştırılana kadar doğruluğu hakkında bir geri bildirim sağlamazlar ve aday mantıksal sorguların çalıştırılması için harici bir yürütücüye ihtiyaç duyması gibi bazı kısıtlamalara sahiptir. Diğer taraftan, model eğitimi için geniş etiketli veri kümelerine gereksinim duyarlar ve bu gereksinim yöntemlerin ölçeklenebilirliğinin yanı sıra akıl yürütme sürecinin yorumlanabilirliğini de sınırlar [159].

2.2.2. Bilgiye Erişim Tabanlı Yöntemler

IR-tabanlı yaklaşımlar çizge üzerinden soru ile ilgili varlıkların çıkarılıp, sıralanması ile yanıtı elde etmeye çalışırlar [151]. Soruya özgü çizge çıkarımı, sorunun semantik temsilinin elde edilmesi, çizge üzerinden çıkarım yapılması ve cevap adaylarının sıralanması adımlarını içeren yöntemlerdir [160]. Bir mantıksal sorgu üretmeksizin, doğrudan cevabı oluşturmaya çalışan tüm seq2seq yaklaşımlar IR-tabanlı olarak kabul edilebilirler [158]. SP-tabanlı yaklaşımlardan farklı olarak harici bir mantıksal sorgu yürütücüsüne bağımlı değildir, ancak genel olarak kamuya açık kıyaslamalarda doğruluk açısından SP-tabanlı yaklaşımların gerisinde kalmaktadır [161].

Farklı varlık türleri ve sayısal kısıtlamalarla tanımlanan karmaşık soruların işlenebilmesi, IR-tabanlı yaklaşımlar için temel zorluklardandır ve basit varlık tanımadan daha üstün bir anlayış gerektirir. Mevcut yöntemlerle bu türlü sorguları ele almak ve işleyebilmek zorlayıcıdır [159].

2.2.3. Büyük Dil Modelleriyle Bilgi Tabanlı Soru Cevaplama

Son teknoloji doğal dil araştırmalarında ele alınan birçok problem için umut vadeci çözümler üretme yeteneğine sahip LLM'ler, Wikipedia benzeri geniş kapsamlı bilgi kaynakları üzerinde eğitilmiş olmaları, gelişmiş akıl yürütme ve az/sıfır örnekli öğrenme yetenekleri ile soru cevaplama alanında gösterdikleri üstün performanslarla mevcut en iyi KBQA modellerinin yerini alıp alamayacağını anlamaya yönelik artan bir ilgiye yol açmıştır [162]. Yakın tarihe kadar mevcut KBQA araştırmalarında yalnızca belirteçleri tanımlamak için sınıflandırıcı olarak yada mantıksal sorgular veya sorular üretmek için taklit edici olarak kullanıldığı görülen LLM'ler [163–165], az veya sıfır örnekli bir bağlamda KBQA sistemlerinde akıl yürütme süreçlerinin doğruluğunun ve yorumlanabilirliğinin artırılmasında önemli bir potansiyele sahip olduğunu gösteren çalışmalarla literatürde yerini sağlamlaştırmaya başlamıştır [159]. Bu noktada bilgi alma sistemleriyle entegre edilerek KG'den ilgili bilgileri almak ve daha alakalı yanıtlar sağlamak, bir semantik ayrıştırıcı rolü ile doğal dil sorularını yapılandırılmış mantıksal formlara dönüştürmek veya çok adımlı akıl yürütmede hata tolerans mekanizmaları veya geri izleme stratejileri ile birlikte kullanılarak hata yayılımını azaltmak gibi amaçlarla kullanımın gelecek vadettiği görülmektedir [158, 159, 166].

LLM'ler, kapsamlı ve özelleştirilmiş bir eğitim gerektirmeden az sayıda örnek ile ince ayar yapılması yoluyla iyi performans elde edebilmeleri yönüyle KGQA sistemlerinin çeşitli alanlara hızlı bir şekilde uyarlanması gibi önemli bir avantaj sağlarlar [167]. Diğer taraftan, sorgu rehberliği veya düşünce zinciri (Chain-Of-Thought-CoT) gibi teknikler, LLM'lerin akıl yürütme sürecine ilişkin içgörüler sağlanmasına ve yanıtların daha yorumlanabilir hale getirilmesine olanak sağlar [166]. Bu noktada, LLM'lerin iyi tasarlanmış az sayıda örnek kümesiyle bile rekabetçi performans elde edebilmektedir ancak, LLM'lerin istem tasarımına yönelik hassasiyetlerinden dolayı, ilgili bilgi grafiği parçacıklarının dahil edilmesi benzeri etkili komut istemi mühendisliği tekniklerinin KBQA'de özellikle çok adımlı soruların yanıtlanabilme başarısı üzerinde önemli etkilerini gözardı etmemek önemlidir [167]. Performansa etkisi olan diğer bir unsur da dil modellerinin boyutu ve mimarisidir. Çalışmalar göstermektedir ki, çok adımlı KBQA için daha büyük boyutlu ve karmaşıklığı yüksek modeller genellikle daha iyi performans göstermektedir. Tüm bunların yanı sıra çalışmalar göstermektedir ki, LLM'ler özellikle sayısal hesaplama, özellik karşılaştırma veya kesin tarih belirleme gibi akıl yürütme ihtiyaçlarına karşılık vermekte zorluk yaşayabilmektedir [162]. Ayrıca LLM'lerin sınırlı bağlamsal penceresi, özellikle karmaşık çok adımlı akıl yürütmede daha büyük ve daha karmaşık KG'leri işlerken zorluklara neden olabilmektedir [158]. Uzun ve kompleks sorulardaki bir diğer zorluk da, komut isteminde tekrarlayan veya ilgisiz içeriğin çok sayıda belirteç kaplamasından dolayı LLM'ler tarafından oluşturulan yanıtın veya sorunun kalitesinin olumsuz etkilenmesidir [162].

3. ÖN ÇALIŞMALAR

Büyük dil modelleri kullanarak tasarlanan semantik ayrıştırma tabanlı bilgi grafiği soru cevaplama sisteminin Şekil 1.1’de sunulan genel mimarisinde yer alan 2 temel modül vardır. Bunlardan ilki Bilgi Erişim Modülü’dür. Bu modül öncelikle bir yığın halinde bulunan haber korpusu içerisindeki aynı içerikli haberlerin kümelenmesi görevini gerçekleştirecek şekilde tasarlanmıştır. Bu görev ‘Olay Tabanlı Haber Kümeleme’ olarak isimlendirilir ve Bölüm 3.1’de metodolojisi sunulan iki aşamalı benzerlik yaklaşımını benimseyen sistem ile gerçekleştirilir. Yapılan deneylerde Bcubed F-Skor metriğine göre %91,74 gibi yüksek bir başarımla aynı içeriğe sahip haber metinlerinin tespit edilmesini sağlanmaktadır.

Bölüm 3.2’de ise, aynı olay örgüsüne ait olayların yüksek başarımla kümelenmesini sağlayan ve Bölüm 4’te ayrıntıları detaylı olarak anlatılan ‘Olay Çizgesi’ modelinin tasarlanmasında temel oluşturan isimlendirilmiş varlıkların önemini tespit edilmesi amacıyla gerçekleştirilen bir dizi deney sunulmuştur.

3.1. Gerçek Zamanlı Haber Akışında Aynı İçerikli Haberlerin Tespiti

3.1.1. Giriş ve Problem Tanımı

Günümüzde bilgiye hızlı erişim gereksinimi, özellikle yoğun veri akışının yaşandığı haber sektöründe büyük bir önem taşımaktadır. Bu ihtiyaca yanıt olarak, çeşitli kaynaklardan veri toplayan haber toplayıcı sistemler geliştirilmiştir. Ancak, bu sistemlerin haberleri toplama, gruplama ve sunma süreçleri, veri akışının yoğunlaştığı ve özellikle savaş, doğal afet gibi hızlı gelişen olaylarda yetersiz kalabilmektedir. Aynı zamanda, aynı içerikteki haberlerin farklı kaynaklarda neredeyse aynı ifadelerle tekrar tekrar sunulması, kullanıcıların bilgi tekrarlarıyla karşılaşmasına neden olarak öne çıkan haberlerin görünürlüğünü düşürmektedir. Bu çalışma, benzer içerikteki haberleri yüksek doğruluk ve hızla gruplandırarak, bilgiye erişim deneyimini optimize etmeyi amaçlayan bir gerçek zamanlı sistem geliştirmektedir.

Bu çalışma, metin kümeleme problemi bağlamında ele alınmaktadır. Teknik olarak,

metin kümeleme, etiketlenmemiş bir veri kümesini içerik benzerliklerine göre gruplara ayırmayı amaçlayan denetimsiz öğrenme problemidir. Bu tür kümeleme, metinlerin çok boyutlu vektör temsilleri aracılığıyla yapısal özelliklerini analiz ederek benzer metinleri aynı gruplarda toplamayı hedefler. Metin kümeleme sürecinde her bir metin örneği, veri setinin içerik özelliklerini temsil eden bir vektöre dönüştürülür ve bu vektörler, genellikle mesafe ölçüleri kullanılarak karşılaştırılır. Bu çalışmada metin verisi, özellikle Türkçe dilinde, kelime gömme (embedding) yöntemleriyle yüksek boyutlu uzaylarda temsil edilmiştir. Bu çalışmanın amacı, klasik kümeleme algoritmalarını metin kümeleme problemine uyarlamak ve özellikle çok boyutlu Türkçe haber metinleri üzerinde gerçek zamanlı kümeleme yeteneğini optimize etmektir [11].

Bu çalışma kapsamında sunulan yenilikçi katkılar aşağıda özetlenmektedir:

- **Gerçek Zamanlı Haber Gruplama Modeli Tasarımı:** Farklı kaynaklardan gelen haber akışlarını $N = \{n_1, n_2, \dots, n_n\}$ şeklinde temsil edilen haber kümesi üzerinden alarak, yüksek doğruluk ve hızla $S = \{s_1, s_2, \dots, s_k\}$ şeklinde tanımlanan gruplara ayıran bir fonksiyon G 'dir. Bu model, geniş çapta veri akışını verimli bir şekilde yönetmek için gerçek zamanlı bir işlem altyapısı kullanır ve kolay entegrasyon sağlayacak şekilde optimize edilmiştir. Matematiksel olarak model şu şekilde ifade edilir:

$$G : H \rightarrow C$$

Burada N , farklı kaynaklardan gelen haberlerin anlık veri akışını, S ise bu haberlerin belirli gruplara atanmış temsilini ifade eder. Modelin gerçek zamanlı çalışmasını sağlayan mekanizma, G fonksiyonunun hem işlem doğruluğunu hem de hızı maksimize edecek şekilde tasarlanmıştır.

- **Türkçe Haber Derlemi ve Ön Eğitilmiş Vektör Modelleri:** 975.809 haber metni, 17.470.827 cümle ve 1.673.884.673 kelime içeren kapsamlı bir Türkçe haber derlemi oluşturulmuştur. Bu derlem üzerinden alana özgü word2vec ve doc2vec modelleri eğitilmiş olup, önceden eğitilmiş bu vektör temsilleri Türkçe dilindeki metin işleme çalışmalarında kullanılmak üzere açık erişime sunulmuştur.
- **Kümeleme Performansı İçin Çok Katmanlı Temsil Performans Analizi:** Haber metinlerinde kelime, cümle ve doküman seviyesindeki temsillerin hız ve doğruluk performansları açısından etkisi analiz edilmiştir. Bu analiz, farklı temsil

seviyelerinin hangi koşullarda daha etkili olduğunu detaylandırarak kümeleme süreçlerinde en verimli yöntemin belirlenmesine katkı sağlamaktadır.

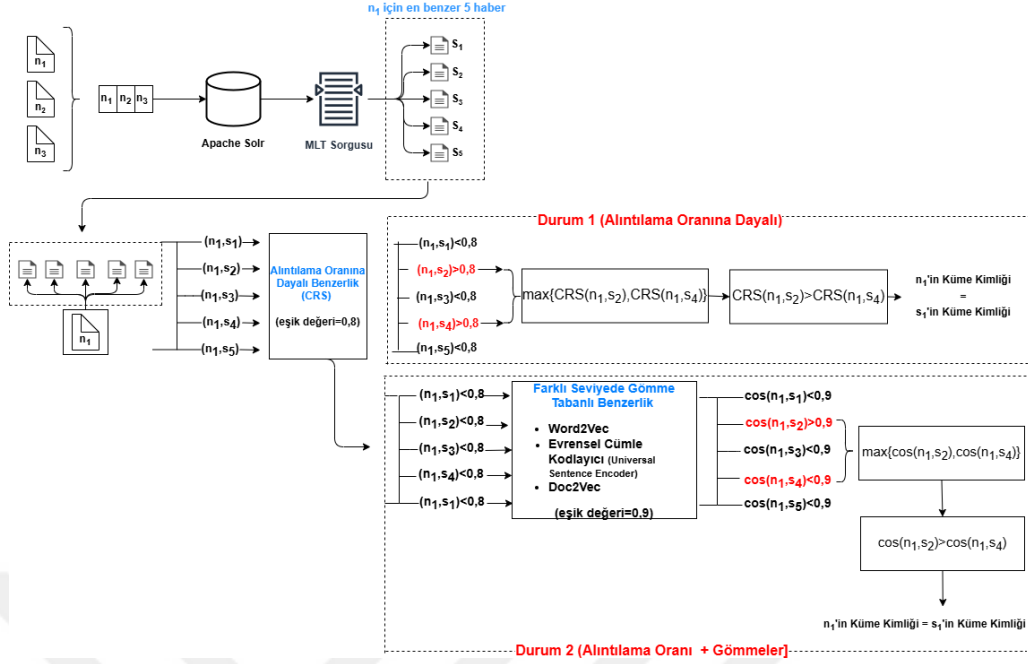
- **Türkçe Haber Kümeleme İçin Özgün Bir Veri Kümesi Geliştirilmesi:** Sistemin performans değerlendirmesi amacıyla, 5223 grup ve toplamda 40.000 haber içeren (*Grouped Turkish News Texts Test Set*) geliştirilmiştir. Bu veri kümesi, alanda yapılacak diğer araştırmalara ve performans değerlendirme çalışmalarına katkı sağlayacak niteliktedir.
- **Metin Temsillerinin Uzunluk Duyarlılığı Analizi:** Kümeleme görevi sırasında kullanılan temsillerin metin uzunluklarına karşı duyarlılığını analiz etmek amacıyla, “atıf oranı” yaklaşımını içeren ve içermeyen metin temsilleri karşılaştırılmıştır. Bu analiz ile, metin uzunluğunun temsiller üzerindeki etkisi, performans farkları bağlamında incelenmiştir.

3.1.2. Metodoloji

Ön işleme: Çalışmada kullanılan haber metinleri, çevrim içi haber kaynaklarından toplanan içerikleri ile basılı yayınlardan optik karakter tanıma (OCR) sistemi aracılığıyla dijitalleştirilen metinleri içermektedir. İlk aşamada, bu metinler jetonlarına ayrılmış (tokenization) ve kelime seviyesinde işlem yapılabilir hale getirilmiştir. Sonrasında, metinlerin dilsel ve yapısal tutarlılığını sağlamak amacıyla tüm karakterler küçük harfe dönüştürülmüş ve sayısal karakterler ile semboller metinlerden arındırılmıştır. Ayrıca, anlamsal gürültüyü azaltmak için durak kelimeler (stopwords) temizlenmiştir. Bu çalışmada kullanılan durak kelime listesi, Türkçe Haber Metinleri Derlemi (*Turkish News Texts Corpus*) üzerinde yapılan analizler sonucu belirlenmiş olup 438 farklı kelimedenden oluşmaktadır.

Önerilen Yaklaşım: Aynı içeriğe sahip haberleri gruplamak amacıyla dört farklı yöntem geliştirilmiştir: kelime vektörü tabanlı kümeleme, cümle vektörü tabanlı kümeleme, belge vektörü tabanlı kümeleme ve iki aşamalı kümeleme.

- **Kelime Vektörü Tabanlı Kümeleme:** Çalışmada, haber alanına özgü kelime vektörlerinin eğitilmesi amacıyla geniş çaplı bir Türkçe Haber Metinleri Derlemi (*Turkish News Texts Corpus*) oluşturulmuştur. Bu derlem, kapsamlı ön işleme adımlarından geçirilmiş ve ardından literatürde sıklıkla kullanılan Word2Vec



Şekil 3.1: İki Aşamalı Benzerlik Yaklaşımı ile Gerçek Zamanlı Haber Kümeleme Sistem Mimarisi

modeli [56] ile eğitilmiştir. Bu işlem sonucunda, derlemde bulunan 422.346 farklı kelimenin yoğun temsillerini sağlamak için kelime vektörleri oluşturulmuştur. Word2Vec modeli ile her bir kelime için 100 boyutlu vektörler üretilmiş ve bu vektörler her haber metni için genel bir temsil elde etmek amacıyla kullanılmıştır. Bu süreçte, her haber metni, o metindeki kelime vektörlerinin ortalaması alınarak tek bir vektörle temsil edilmiştir. Sisteme anlık olarak dahil edilen her yeni haberin vektörü, “MoreLikeThis” (MLT) sorgusu yardımıyla son üç gün içinde kaydedilmiş en benzer beş haberin vektörleri ile karşılaştırılmıştır. Bu karşılaştırmada kosinüs benzerliği hesaplanmış ve %90 benzerlik oranı eşik değer olarak belirlenmiştir. Yeni bir haber, en yüksek benzerlik skoruna sahip haberin bulunduğu kümeye eklenmiştir. Eğer karşılaştırılan haberlerin hiçbirinde %90 benzerlik eşiği aşılmıyorsa, bu yeni haber bağımsız bir küme olarak değerlendirilmiş ve sistemde yeni bir küme kimliği atanmıştır.

- **Cümle Vektörü Tabanlı Kümeleme:** Çalışmada cümle temsillerini oluşturmak amacıyla, Google araştırmacıları Yinfei Yang, Daniel Cer ve diğerleri tarafından 2018 yılında önerilen, Transformer ve CNN mimarilerine dayanan çok dilli bir

cümle kodlama modeli tercih edilmiştir. Bu model, “Multi-task Dual-Encoder Model” olarak adlandırılan bir derin öğrenme tekniği ile Türkçe dahil olmak üzere toplam 16 dilde eşzamanlı olarak eğitilmiştir [168]. Bu eğitim süreci, farklı dillerdeki cümlelerin aynı model kullanılarak vektör uzayında temsillerinin oluşturulmasını sağlamaktadır. Böylece modelin desteklediği dillerdeki cümleler arasındaki anlamsal benzerlik veya farklılıklar, herhangi bir ek işleme gerek kalmadan doğrudan incelenebilmektedir. Bu yöntemde, önceki kelime vektör tabanlı temsillerden farklı olarak, dokümanlar cümle vektörleri kullanılarak temsil edilmiştir. Cümle seviyesindeki bu gömlemeler, kosinüs benzerliği ölçütü kullanılarak %90 eşik değerine göre gruplandırılmıştır. Yeni bir haber, cümle vektörü temsili ile en yüksek benzerlik oranını sağlayan kümeye dahil edilmiştir. Eğer benzerlik eşiği karşılanmıyorsa, yeni bir küme oluşturulmuş ve yeni haber bu kümeye atanmıştır. Modelin önceden eğitilmiş cümle vektörleri, doğrudan TensorFlow Hub’dan elde edilerek kullanılmaktadır.

- **Belge Vektörü Tabanlı Kümeleme:** Çeşitli görevlerde kullanılan birçok makine öğrenmesi yöntemi, sabit uzunluktaki özellik vektörlerini girdi olarak gerektirmektedir; bu vektörler, uygulamalarda yaygın olarak Bag-of-Words (BoW) yöntemi ile elde edilmektedir. Ancak, BoW yöntemi, kelimelerin sırasını ve anlamsal ilişkilerini göz ardı ettiğinden dolayı sınırlamalara sahiptir ve geliştirilmeye açıktır. Bu bağlamda, doğal dil işleme alanında önemli ilerlemelere olanak sağlayan Word2Vec modeli, koşullu olasılık ilkesi ve anlamsal yakınlık ölçütü ile kelimeler arasındaki ilişkileri korumakta ve böylece metinlerin anlamsal bütünlüğünü daha etkili bir şekilde ifade edebilmektedir. Word2Vec yöntemini takiben, aynı araştırma ekibi 2014 yılında “Cümlelerin ve Belgelerin Dağıtık Temsilleri” (*“Distributed Representations of Sentences and Documents”*) adlı çalışmayı sunmuş ve belge seviyesindeki metinlerin doğrudan sabit boyutlu vektörlerle temsil edilmesine imkan sağlamıştır [169]. Doc2Vec olarak bilinen bu yöntem, Word2Vec yapısını büyük ölçüde koruyarak, yalnızca metinleri benzersiz şekilde temsil eden paragraf simgeleri eklemiştir. Bu geliştirme, kelimelerin tekil temsillerinden ziyade, bütünsel bir belge temsili sunarak farklı uzunluktaki dokümanların sabit boyutlu vektörlerle ifade edilmesini mümkün kılmıştır. Bu kümeleme yaklaşımında, Doc2Vec yöntemi ile her haber metni için 100 boyutlu

belge vektörleri oluşturulmuştur. Haber metninin ait olduğu küme kimliği, önceki kelime ve cümle seviyesinde kümeleme yaklaşımlarında olduğu gibi, kosinüs benzerlik metriğine ve %90 benzerlik eşiğine dayalı olarak belirlenmiştir. Yeni bir haber metni, kosinüs benzerlik eşiğine göre mevcut kümelere dahil edilmekte; eşiğin karşılanmadığı durumlarda ise, bağımsız bir küme olarak yeni bir kimlik ile sisteme eklenmektedir.

- **İki Aşamalı Kümeleme:** Farklı platformlardan sürekli olarak akan haber metinleri incelendiğinde, pek çok haberin orijinal metnin kısıltılması ya da genişletilmesi yoluyla oluşturulduğu gözlemlenmektedir. Haber metninin temsili, daha önceki yaklaşımlarda kelime, cümle ve belge seviyelerinde ele alınmıştır. Ancak, bu yöntemlerin metin boyutuna karşı hassasiyetleri nedeniyle, orijinal bir metin ile türetilmiş yeni bir haber metni arasındaki benzerlik doğru şekilde değerlendirilememektedir. Bu hassasiyeti azaltmak ve metinlerin farklı uzunluklardan kaynaklanan benzerlik problemini gidermek için Şekil 3.1’de gösterilen iki aşamalı bir kümeleme yaklaşımı uygulanmıştır. Bu bağlamda, gruplandırılacak bir haber metninin daha önce sisteme dahil edilmiş haberlerle karşılaştırılmasında, bir ”atıf oranı” eşik değeri belirlenmesi anlamlı bulunmuştur. “Atıf oranı” yaklaşımı doğrultusunda, yeni bir haber metni ile son üç gün içinde sisteme eklenen haberler arasından en çok benzeyen beş haber (s_1, s_2, s_3, s_4, s_5) seçilmiştir. $A = n, s_1, s_2, s_3, s_4, s_5$ kümesi, her bir haber metninin kelime sayısına göre incelenerek kısa ve uzun metinler belirlenmiştir. Bu aşamada, uzun metindeki benzersiz kelimelerin kısa metindeki benzersiz kelimelerle bire bir eşleşme oranı, “atıf oranı” olarak tanımlanmıştır.

Atıf oranı için eşik değeri %80 olarak belirlenmiş olup, bu eşiğin üzerinde en yüksek atıf oranına sahip metinler aynı kümede kabul edilmiştir. İkinci aşamada, eğer hiçbir metin çifti atıf oranı eşiğine ulaşamazsa, metinler farklı seviyelerdeki vektör temsilleri (kelime, cümle, belge) kullanılarak kosinüs benzerliği açısından karşılaştırılmış ve %90 benzerlik eşiğine göre kümeleme işlemi yapılmıştır. Bu iki aşamalı kümeleme yaklaşımı, farklı boyutlardaki haber metninin içeriksel benzerliklerinin daha doğru bir şekilde değerlendirilmesine olanak tanıyarak, sistemin özgün haberlerin türevlerinden ayırt edilmesini sağlamaktadır.

3.1.3. Deneyler

Veri Kümesi:

- **Türkçe Haber Metinleri Derlemi (Turkish News Texts Corpus):** Çalışma kapsamında, haber alanına özgü olarak kelime, cümle ve doküman seviyelerinde üç farklı metin temsili oluşturulmuştur. Kelime gömmeleri için Word2Vec, doküman gömmeleri için ise Doc2Vec modelleri kullanılmış olup, her iki model de 975.809 haber metnini içeren Türkçe Haber Metinleri Derlemi üzerinde eğitilmiştir. Bu derlem, 17.470.827 cümle ve 422.346 tekil kelime içermekle birlikte, toplamda 1.673.884.673 kelime barındırmaktadır. Hem kelime gömme hem de belge gömme modellerinin etkili bir şekilde eğitilmesi amacıyla, derlem, öncelikle kapsamlı bir ön işleme sürecinden geçirilmiştir.
- **Gruplandırılmış Türkçe Haber Metinleri Test Kümesi (Grouped Turkish News Texts Test Set):** Test veri kümesi, gerçek zamanlı akışta aynı içeriğe sahip haberlerin kümeleme performansını değerlendirmek amacıyla 40.000 haber metni ile oluşturulmuştur. Bu veri kümesindeki haberler, öncelikle çalışmada önerilen iki aşamalı kümeleme yöntemi kullanılarak gruplandırılmıştır. Kümeleme işleminin ardından, haber kümeleri alanında uzman 3 adet haber editörü tarafından paralel olarak incelenmiş ve hatalı kümelemeler manuel olarak düzeltilmiştir. Sonuç olarak, 40.000 haberden ve 5223 farklı gruptan oluşan test veri kümesi, gerçek zamanlı bir sistemin performansını simüle etmek için kullanılmıştır. Bu yapı, sistemin metin kümeleme doğruluğunun güvenilir bir şekilde ölçülmesine katkı sağlamaktadır.

Değerlendirme Metriği: Kümeleme görevi, nesneleri yapısal benzerliklerine göre gruplamayı amaçlarken, birbirine benzeyen nesneleri aynı gruba toplamak ve benzersiz nesneleri farklı gruplarda tutmak gibi zorlu hedeflere sahiptir. Bu kapsamda, kümeleme performansını değerlendirirken iki temel metrik sınıfı kullanılır: içsel metrikler ve dışsal metrikler. İçsel metrikler, kümeler içindeki elemanların birbirine yakınlığı ve diğer kümelerle olan uzaklık gibi özellikleri analiz ederken, dışsal metrikler, sonuçları insanlar tarafından oluşturulmuş altın standartlarla kıyaslamayı amaçlar. Bu çalışmada, önerilen sistemin performansını değerlendirmek için, literatürde küme performansını ölçmede en başarılı dış metriklerden biri olarak kabul edilen B-Cubed metriği kullanılmıştır [170,

171]. B-Cubed, her bir öge için kesinlik (precision) ve duyarlılık (recall) tahmin ederek değerlendirme sürecini detaylandırır. Bir ögenin kesinliği, aynı kümede bulunan diğer ögelerin ne kadarının aynı kategoriye ait olduğunu gösterirken, duyarlılık (recall) değeri, ögenin kendi kategorisindeki diğer ögelerin ne kadarının aynı kümede bulunduğunu ifade eder.

3.1.4. Sonuçlar ve Tartışma

Bu çalışmada ele alınan haber gruplama problemi, gerçek dünya uygulamaları için akan veri üzerinde çalışmak üzere tasarlanmıştır. Önerilen sistem, iki aşamalı bir yapıya sahiptir. İlk aşamada, haber kaynağından gelen anlık haberler Apache Solr veritabanına kaydedilmekte ve ardından “MoreLikeThis” (MLT) sorgusu aracılığıyla son üç gün içinde kaydedilmiş en benzer 5 haber veritabanından çıkarılmaktadır. İkinci aşamada ise, hızlı ve güvenilir bir kümeleme işlemi gerçekleştirmek için çeşitli yöntemler kullanılmıştır. Bu aşamada, kümeleme işlemi anlamsal benzerlik üzerine kurulmuş ve bu amaçla doküman temsilleri, kelime, cümle ve doküman seviyelerinde olmak üzere üç farklı yöntemle dayandırılmıştır.

Referans haber ve en benzer 5 haber, belirli ön işleme adımlarından geçirildikten sonra doküman vektörleri elde edilmiştir. Referans haberin gömmesi ile en benzer 5 haberin gömmesi arasında kosinüs benzerliği hesaplanarak metin çiftlerinin benzerlik oranı belirlenmiştir. Bu benzerlik için %90 eşik değeri uygulanmıştır; eğer metin çifti %90 veya üzerinde bir benzerlik gösteriyorsa, aynı kümede olma ihtimalleri yüksek kabul edilmiştir. Birden fazla metin bu eşiği aşarsa, referans metin en yüksek benzerlik oranına sahip metinle aynı kümeye dahil edilmiştir. Eğer hiçbir metin çifti %90 benzerlik oranına ulaşamazsa, referans metine yeni bir küme kimliği atanmıştır.

Farklı haber kaynaklarından gelen metinlerin çoğunlukla aynı içeriği birebir veya kısaltılmış/genişletilmiş şekilde sunduğu göz önüne alınarak, sistem tasarımında iki aşamalı bir kümeleme yaklaşımı uygulanmıştır. Bu kapsamda, metin çifti arasındaki benzerliğin belirlenmesinde “birebir atıf oranı” önemli bir ölçüt olarak kullanılmıştır. Atıf oranı, kısa ve uzun metinlerdeki benzersiz kelimelerin birebir eşleşme oranına dayalı olarak hesaplanmıştır. Eğer atıf oranına göre karşılaştırılan metin çiftlerinden hiçbirisi belirlenen eşiği geçemezse, metinlerin sayısal vektör temsilleri çıkarılarak, bu

vektörlerin kosinüs benzerliği hesaplanmıştır. Bu aşamada, daha önce olduğu gibi %90 benzerlik eşiği kullanılmıştır.

Tablo 3.1: Farklı Gömme Seviyeleri için Kümeleme Performansının B-Cubed Değerlendirme Metriği ile Karşılaştırılması

Gömme Seviyesi	Benzerlik	Kesinlik	Duyarlılık	F-Skor	Hız (sn)
Kelime	Kosinüs	0.7999	0.9176	0.8547	0.8
Cümle	Kosinüs	0.9818	0.7885	0.8746	0.72
Belge	Kosinüs	0.9996	0.7259	0.8410	0.59

Tablo 3.2: 2-adımlı Kümeleme Yaklaşımında Farklı Gömme Seviyelerinin Karşılaştırılması

Gömme Seviyesi	Benzerlik	Kesinlik	Duyarlılık	F-Skor	Hız (sn)
Kelime+Alıntılama	Kosinüs	0.8042	0.9550	0.8731	0.76
Cümle+Alıntılama	Kosinüs	0.9833	0.7929	0.8779	0.68
Belge+Alıntılama	Kosinüs	0.9952	0.8508	0.9174	0.62

Tablo 3.1’de kelime, cümle ve doküman seviyesindeki gömlemeler kullanılarak, B-Cubed değerlendirme metriği ile hesaplanan kesinlik, duyarlılık ve F-skor değerleri sunulmaktadır. Eşik değerin üzerinde atıf oranına sahip bir haber çiftinin aynı içerikte olduğu kabul edilmesi, sistemin kümeleme başarısını artırıcı bir etken olarak gözlemlenmiştir. Bu etkinin olumlu etkisi Tablo 3.2’de de görülmektedir. Gerçek zamanlı bir sistem için en önemli unsurlardan biri işlem hızıdır. Bu nedenle, önerilen yöntemlerin başarısının yanı sıra çalışma hızları da dikkate alınmıştır. Tablo 3.1 ve Tablo 3.2 incelendiğinde, iki aşamalı kümeleme yaklaşımının sistem performansı üzerinde ihmal edilebilir düzeyde bir değişime yol açtığı görülmektedir.

Bu yöntemlerin sonucunda elde edilen bulgular aşağıda özetlenmiştir:

1. Alıntılama skoru yaklaşımından bağımsız olarak, yalnızca kelime, cümle ve belge

temsilleri kullanılarak gerçekleştirilen kümeleme yaklaşımında, “Çok Görevli Çift Kodlayıcı Modeli” (Multi-task Dual-Encoder Model) F-skor açısından diğer yöntemlere göre daha yüksek bir performans sergilemiştir. Ancak hız açısından değerlendirildiğinde, Doc2Vec en hızlı çalışan yöntem olarak öne çıkmıştır.

2. İki adımlı kümeleme yaklaşımında Doc2Vec yöntemi ile elde edilen kümeleme performansında %7,6 oranında önemli bir F-skor artışı sağlanmıştır. Bu yöntemle %91,74 F-skoru elde edilerek en yüksek kümeleme performansı gözlenmiştir.
3. Kelime, cümle ve belge gömmeleri oluşturmak için kullanılan yöntemler arasında, alıntılama skoru yaklaşımıyla en büyük performans artışı Doc2Vec yönteminde elde edilmiştir. Bu durum, Doc2Vec’in metin uzunluklarına duyarlı bir yöntem olduğunu da göstermektedir.
4. İki aşamalı kümeleme yönteminin işlem süresine etkisi ihmal edilebilir düzeydedir ve pozitif bir etki sağlamaktadır. Bu nedenle, gerçek zamanlı sistemde atıf oranının kullanılması hızda bir kayba neden olmamış; aksine, kümeleme performansında belirgin bir iyileşme sağlamıştır.
5. F-skor değerleri karşılaştırıldığında, Word2Vec hem hız hem de performans açısından önerilen yöntemler arasında en düşük performansı göstermiştir. Bu sonuç, iki adımlı kümeleme performanslarında da değişmemiştir.
6. Tüm testler sonucunda, önerilen sistem için hız ve performans açısından en uygun yaklaşımın, Doc2Vec doküman temsilleri kullanılarak gerçekleştirilen iki adımlı kümeleme yöntemi olduğu tespit edilmiştir.

3.2. Adlandırılmış Varlıkların Olay Tabanlı Haber Kümelemeye Etkisi

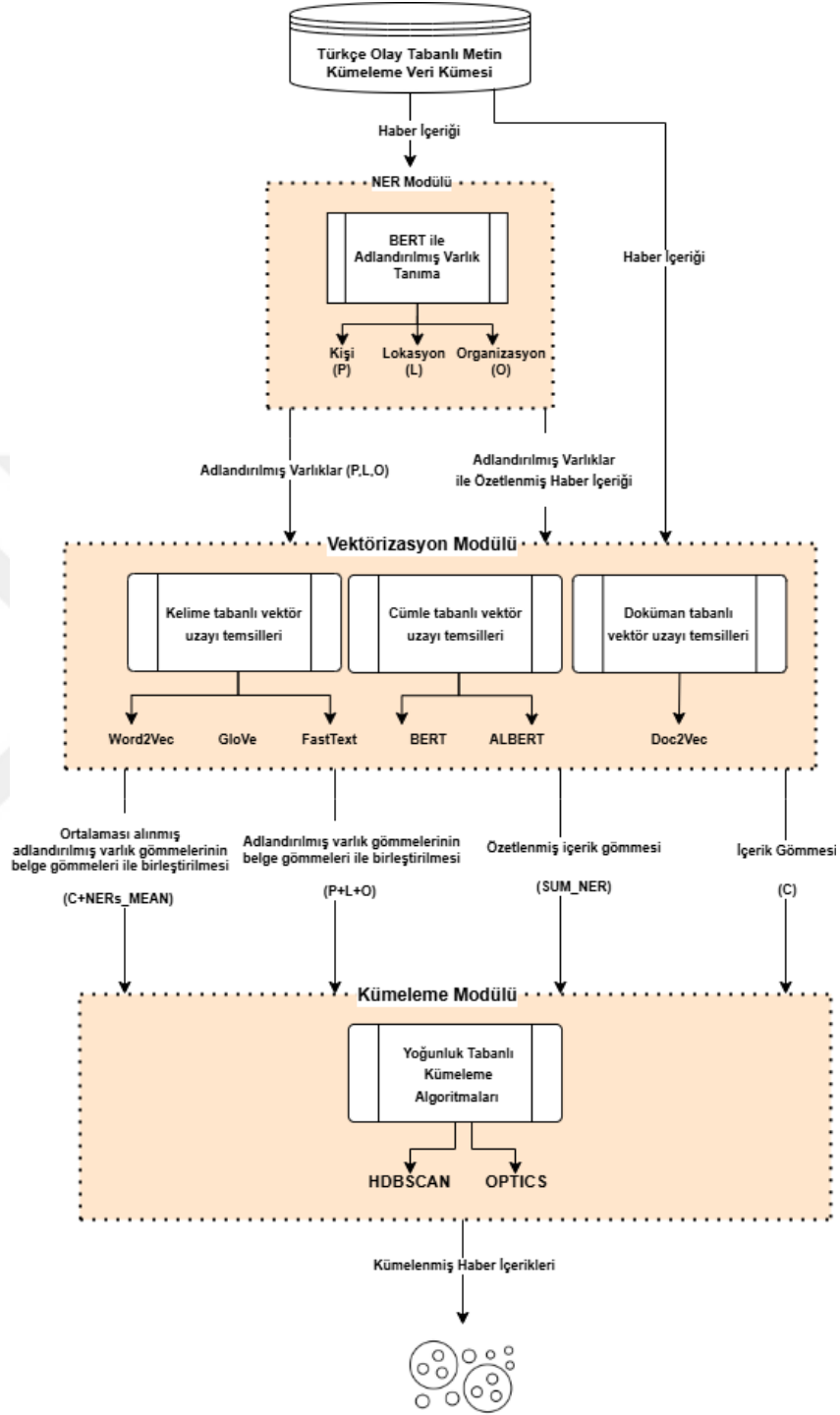
3.2.1. Giriş ve Problem Tanımı

Web ortamı, farklı coğrafi bölgelerdeki büyük ve küçük ölçekli haber kaynaklarından yayılan içeriklerle sürekli genişlemektedir. Bu genişleyen haber yığını, gündeme dair gelişmeleri kapsayan içeriklerin bireysel bir çabayla düzenli olarak izlenmesini zorlaştırmış, hatta imkânsız hale getirmiştir [172, 173]. Bu bağlamda, verileri anlamlandırmak, düzenlemek ve kullanıcılar için bilgiye erişimi kolaylaştırmak amacıyla çeşitli veri

işleme teknikleri geliştirilmiştir. Bu tekniklerden biri de, içerik benzerliklerine dayalı kümeleme yöntemleridir [172].

Metin kümeleme (Text Clustering-TC), metin belgelerinin belirli bir veri kümesini içerik benzerliklerine göre alt kümelere ayırarak organize eden bir süreçtir ve bilgi keşfi ile veri yoğunluklarını anlamlandırmada yetkin bir araçtır [174]. Özellikle Bilgi Erişimi (Information Retrieval - IR) sistemlerinde, arama ve filtreleme işlemlerinin daha hızlı ve güvenilir bir şekilde yapılabilmesi açısından veri hacminin azaltılması kritik bir rol oynamaktadır. Kümeleme çözümleri, IR sistemlerinin doğruluk (precision) ve duyarlılık (recall) performanslarını iyileştirmek amacıyla kullanılarak, kullanıcıya daha anlamlı ve karar destek süreçlerini kolaylaştıran içeriklerin sunulmasını sağlamaktadır [175, 176].

Yüzlerce haber kaynağından elde edilen yoğun veri akışı içinde, belirli olayları ve gelişmeleri tespit edebilmek için, haber içeriklerinin çok katmanlı bir hiyerarşi ile kümelenmesi gerekmektedir. Klasik haber kümeleme yaklaşımları, metinleri genellikle olaydan bağımsız genel kategorilere ayırır; örneğin, spor haberleri, yangın haberleri veya seçim haberleri gibi geniş gruplar oluşturur [177, 178]. Bunun yanı sıra, haber metinleri üzerinde uygulanan bir diğer yaygın yöntem ise konu modellemesidir. Konu modelleme teknikleri, belgelerdeki kelimelere belirli bir konuyla ilişkili olasılıklar atayarak, her bir konuya en uygun kelimelerin listelerini çıkarır ve böylece metinlerin konu bazında ayrıştırılmasını sağlar [178, 179]. Ancak, bu yaklaşımlar gündemin olay bazında ayrıntılı bir şekilde takibini gerektiren ihtiyaçlara yanıt verememektedir. Gündem oluşturan olayların hızlıca izlenmesi, ilgililik derecesine göre filtrelenmesi ve kamuoyu için önem arz eden gelişmelerin öne çıkarılması, olay tabanlı daha ayrıntılı çözümler gerektirmektedir. Bu ihtiyaç doğrultusunda, klasik kategori tabanlı yaklaşımlar yerine olayların içsel yapısını dikkate alan olay tabanlı kümeleme yöntemleri gereklidir. Örneğin, spor haberleri genel bir küme olarak değil, “24.10.2021 tarihinde oynanan Barselona - Real Madrid maçı” gibi spesifik olaylara odaklanarak gruplandırılmalıdır. Bu doğrultuda, Şekil 3.2’de yapısı verilen olay tabanlı bir haber kümeleme sistemi önerilmektedir [5].



Şekil 3.2: Önerilen Olay Tabanlı Haber Kümeleme Yaklaşımı

3.2.2. Metodoloji

Vektör uzayı temsillerinin oluşturulması : Yapılan deneylerde haberlerin temsili için kelime, cümle ve doküman gibi farklı seviyelerde gömmeler sinir ağı modelleri ile elde edilmiş ve metin kümeleme performansına etkileri karşılaştırılmıştır.

1. Kelime tabanlı vektör uzayı temsilleri: Çalışma kapsamında, Türkçe Haber Metinleri Derlemi kullanılarak eğitilen Word2Vec modeli ile önceden eğitilmiş kelime vektörleri kullanılmıştır. Word2Vec modeli kelimeleri en küçük birim olarak kabul ederken, kelimelerin vektörel gösterimi için n-gram'lardan yararlanan FastText [180], bir diğer model olarak tercih edilmiştir. FastText, özellikle Türkçe gibi sondan eklemeli dillerdeki yüksek temsil yeteneği ile öne çıkmaktadır. Öte yandan, sayma tabanlı matris çarpanlarına ayırma (count-based matrix factorization) ve lokal içerik penceresi (local content window) yöntemlerini birleştirerek anlamsal olarak daha güçlü bir kelime temsili oluşturabileceğini savunan GloVe modeli [181] çalışma kapsamında tercih edilen son kelime tabanlı vektör temsidir. GloVe ve FastText modelleri için Türkçe dilinde önceden eğitilmiş modeller kullanılmış ve 300 boyutlu kelime vektörlerinin ortalaması alınarak haber metinlerinin gömmeleri elde edilmiştir.
2. Cümle tabanlı vektör uzayı temsilleri: Cümle temsillerini oluşturmak için önceden eğitilmiş dil modelleri olan BERT [24] ve ALBERT [182] kullanılmıştır. BERT, etiketlenmemiş metinden derin çift yönlü temsilleri eğitmek için tasarlanmıştır ve tüm katmanlarında hem sol hem de sağ bağlamda birlikte koşullandırılmıştır. Cümle-BERT (Sentence-BERT-SBERT), önceden eğitilmiş BERT dil modelinin bir modifikasyonudur. Model, anlamsal olarak benzer cümleler için vektör uzayında yakın boyutlu cümle gömmeleri türetir. BERT modelini geliştiren Google araştırmacıları tarafından katman sayısı (L), gizli katman boyutu (H) ve dikkat mekanizması başlık boyutu (A) için farklı parametre değerlerine sahip iki model sunulmuştur. Bu modellerden ilki BERT Base olarak adlandırılmış ve $L=12$, $H=768$ ve $A=12$ parametre değerleri ile tasarlanmıştır. Diğer model ise $L=24$, $H=1024$ ve $A=16$ parametre değerlerine sahip BERT Large modelidir. Bu çalışmada, Türkçe dilinde 35 GB boyutunda ve 44.04.976.662 jeton içeren önceden eğitilmiş BERT base modeli olan BERTurk ve ALBERT modelleri ile cümle

vektörleri oluşturulmuştur.

3. Doküman tabanlı vektör uzayı temsilleri: Doc2Vec olarak adlandırılan ve belge vektörlerinin doğrudan elde edilmesini sağlayan yöntem çalışma kapsamında kullanılmıştır. Bu yaklaşımda da Word2Vec yaklaşımında olduğu gibi Doc2Vec yöntemi ile Türkçe Haber Metinleri Derlemi üzerinde doküman vektörleri ön eğitime tabi tutularak haber metinlerinin gömmeleri elde edilmiştir.

Metin kümeleme yöntemlerinin karşılaştırılması: Aynı içeriğe sahip haberlerin 'hikayeler' olarak kümelenmesi temelde doğal dil işleme araştırma alanı içerisindeki metin kümeleme problemine işaret etmektedir. Kümeleme problemi için geliştirilen modellerde kümelerin homojen dağılması, küme sayısının başlangıçta bilinmesi, aykırı değerlere duyarlılık ve kümelerin boyutu performans üzerinde önemli bir etkiye sahiptir. Bu nedenle verinin yapısına yönelik model seçimi de büyük önem taşımaktadır. Haber içeriklerinin uzun metinsel yapısının getirdiği bir diğer temel zorluk ise, çok fazla gürültü içermesidir. Bu nedenle geliştirilecek yaklaşım aynı zamanda gürültüyü en aza indirmeye ve olayın temel unsurlarını çıkarmaya odaklanmalıdır. Ayrıca olay tabanlı haber kümeleme problemlerinde her bir 'hikaye' bir küme olarak ele alındığı için küme sayısı doğası gereği oldukça fazladır. Tüm bu zorluklarla baş edebilmek için küme sayısını bir parametre olarak kabul etmeyen yoğunluk tabanlı HDBSCAN ve OPTICS kümeleme yöntemleri tercih edilmiştir.

DBSCAN (Density Based Spatial Clustering of Application with Noise) kümeleme algoritması [183] 1996 yılında tasarlanmıştır ve literatürdeki ilk yoğunluk tabanlı kümeleme algoritması olarak bilinmektedir. Bir kümedeki her nesnenin Eps yarıçapı içinde komşuluğunda en az $MinPts$ kadar nesne içermesi gerektiği temel varsayımına dayanır. Temel olarak gürültülü ve yüksek boyutlu veri yığınlarında rastgele şekillendirilmiş veri kümelerinin tespiti için tasarlanmış bir algoritmadır. D veri tabanındaki rastgele bir s noktasının ϵ komşuluğu Denklem 3.1'deki gibi tanımlanır. Eğer s noktasının komşuları en az sayıda nokta içeriyorsa, s noktası Denklem 3.2'deki gibi çekirdek noktası olarak tanımlanır. DBSCAN kümeleme algoritması, kullanıcı tarafından belirlenen Eps ve $MinPts$ parametrelerine karşı çok hassastır ve parametrelerin değiştirilmesi kümeleme sonuçlarında büyük değişikliklere yol açar. Ayrıca bir noktanın çekirdek olup olmadığının belirlenmesi de I/O yükünü oldukça artıran bir unsurdur. Hiyerarşik DBSCAN kümeleme algoritması (HDBSCAN) bu iki dezavantajı ortadan kaldırmak

için yapılan çalışmalar sonucunda ortaya çıkan en iyi alternatif yaklaşımlardan biridir [184].

HDBSCAN kümeleme algoritması, geleneksel yoğunluk tabanlı DBSCAN algoritmasını değişken epsilon değerleri üzerinde gerçekleştirir ve kümeleme sonuçlarını ε üzerinden en kararlı sonuç ile birleştirir. DBSCAN'ın aksine, HDBSCAN algoritması değişken yoğunluklu kümeleri desteklerken, mesafe ölçeği parametresini ayarlama ihtiyacını ortadan kaldırır. Hız açısından da oldukça iyi performans gösteren HDBSCAN algoritması, yoğunluk tabanlı kümeleme için oldukça önemli bir gelişme olarak değerlendirilmekte ve sıklıkla kullanılan DBSCAN algoritmasına iyi bir alternatif olarak çalışmalarda aktif olarak kullanılmaktadır.

$$N_{Eps} = \{n \in D / \text{dist}(s, n) \prec Eps\} \quad (3.1)$$

$$N_{Eps}(s) \succ MinPts \quad (3.2)$$

HDBSCAN kümeleme için kullanılan Python paketinde sunulan *min_cluster_size* parametresi, ideal olarak küme olarak değerlendirilmek istenen en küçük boyutlu grupta göre ayarlanmaktadır. Türkçe Olay Tabanlı Metin Kümeleme Veri Kümesi, aynı içeriğe sahip en az 6, en fazla 105 haber içermektedir. Bu nedenle *min_cluster_size* = 6 olarak belirlenmiştir.

OPTICS (Ordering points to identify the clustering structure), Ankerst, Breunig ve diğerleri tarafından 1999 yılında sunulan bir diğer yoğunluk tabanlı kümeleme algoritmasıdır [185]. DBSCAN kümeleme algoritması ile yakından ilişkili olan OPTICS algoritması, yüksek yoğunluklu çekirdek örneği bulur ve kümeleri bu örneklerle göre genişletir. DBSCAN'den farklı olarak, değişken bir komşuluk yarıçapı için küme hiyerarşisini korumaya çalışır. OPTICS, HDBSCAN'e kıyasla işlenecek veri noktasına en yakın bir sonraki veri noktasını belirlemek için öncelik kuyruğu (*Min_heap*) boyutunda daha fazla belleğe ihtiyaç duyar. Böylece DBSCAN'ın en büyük zayıflıklarından biri olan değişken yoğunluktaki verilerde anlamlı kümeler tespit etme sorununu giderir.

OPTICS kümeleme algoritması, DBSCAN'de olduğu gibi iki başlangıç parametre-

sine ihtiyaç duyar. Bu parametreler, dikkate alınacak maksimum mesafeyi (yarıçap) tanımlayan ε ve bir küme oluşturmak için gereken minimum nokta sayısını tanımlayan *MinPts*'dir. OPTICS algoritmasında kullanılan *MinPts* parametresi, çalışmada OPTICS kümelemesi için kullanılan Python paketindeki *min_samples* parametresi ile tanımlanmıştır. Bu parametre varsayılan olarak 5'tir, ancak çalışılan veri kümesinde küme olarak değerlendirilecek en küçük boyutlu grubun eleman sayısı 6 olduğu için 6 olarak belirlenmiştir. Ayrıca ε parametresi de kümeleme performansı açısından en tutarlı sonucu verdiği için 3 olarak belirlenmiştir. Çekirdek nokta tanımı HDBSCAN algoritması ile aynıdır ve OPTICS algoritmasında da her noktaya *MinPts*-inci en yakın noktaya olan uzaklığı tanımlayan bir çekirdek uzaklık atanır. Veri setindeki başka bir noktanın s , p noktasından ulaşılabilirlik mesafesi, s ile p arasındaki mesafe veya p 'nin çekirdek mesafesidir. Burada büyük olan değer ulaşılabilirlik mesafesi olarak belirlenir. OPTICS, HDBSCAN'a kıyasla işlenecek veri noktasına en yakın bir sonraki veri noktasını belirlemek için bellekte öncelik kuyruğu *Min_heap* büyüklüğünde bir alana ihtiyaç duymaktadır. Aynı zamanda OPTICS algoritmasının en yakın komşu sorgu maliyeti HDBSCAN'daki yarıçap sorgusuna kıyasla hesaplama maliyetinden daha yüksektir. Ancak ihtiyaç duyduğu parametre sayısı daha azdır.

3.2.3. Deneyler

Türkçe olay tabanlı metin kümeleme veri kümesi: Çalışmada önerilen olay tabanlı haber kümeleme yaklaşımının performansını değerlendirmek için Türkçe dilinde bir kıyaslama veri kümesi oluşturulmuştur. Bu veri kümesi literatürde ilk olmakla birlikte, diğer araştırmacıların da üzerinde çalışmasına olanak sağlayacak standartlardadır. İçerisinde kişi/kurum/kuruluş, zaman, yer ve olay bütünlüğü sağlanmış ve küme sayısı (olay sayısı) 930 olan, 19.848 haber metni ve toplam 317.646 cümle barındırır. Veri kümesinin oluşturulmasında veri sağlayıcı, Türkiye'nin en büyük medya takip şirketlerinden biri olan Interpress'tir. Sağlanan 40.305 haber metninden tam olarak tekrar eden metinler çıkarılmış ve belge uzunluğu en az 500, en fazla 10.000 karakter olan metinler kalacak şekilde filtre uygulanmıştır. Ardından 2021 yılında araştırmacılar tarafından akademik çalışmada [11] önerilen ve %90'ın üzerinde başarı ile çalışan atıf tabanlı gerçek zamanlı haber gruplama motoru ile haberler otomatik olarak gruplanmıştır.

Haber gruplama motorunun önerdiği kümeler arasından en az 6, en fazla 105 habere sahip kümeler tekrar filtrelenmiştir. Son aşamada ise insan doğrulayıcılar tarafından küme içeriğindeki haberlerin aynı olaya ait olup olmadığı kontrol edilerek doğrulama sağlanmıştır. Veri kümesinde yer alan orijinal haber metin kümelerine ait bazı örnekler Tablo 3.3’de sunulmuştur.

Adlandırılmış varlıkların metin kümeleme performansı üzerindeki etkileri: Haber metinleri özne (kişi/kurum/kuruluş), zaman, yer ve eylemin kendisini içerir. Tüm bu varlıklar ’olayların’ birbirinden ayrılması için ayırt edici unsurlardır. Bu nedenle haber metinlerinin ’olay’ temelinde kümelenmesinde isimlendirilmiş varlıklardan yararlanılarak kümeye dahil olan örneklerin birbirine yaklaşması, farklı kümeleri temsil eden örneklerin ise diğer kümelerden mümkün olduğunca ayrılması amaçlanmıştır. Büyük veri arşivlerini barındıran sistemlerde sadece metin kümeleme performansı değil, hız da büyük önem taşımaktadır. Bu noktada hedeflerimizden bir diğeri de işlenecek veri hacmini azaltarak, hız kazanmaktır. Çalışmalar kapsamında adlandırılmış varlıkları tanımlamak için Bidirectional Encoder Representations from Transformers (BERT) dil modeli kullanılmıştır. Türkçe için önceden eğitilmiş olan BERTurk modeli üzerinde kişi, organizasyon ve konum için adlandırılmış varlık etiketleri ile 282 farklı dil için Wikipedia sayfalarından bağlantılı varlıklar kullanılarak oluşturulan WikiAnn veri kümesi ile ince ayar yapılmıştır. Eğitilen BERT tabanlı NER modelinin performansı, uygulanan 2 farklı test veri kümesine göre %94’ün üzerindedir. Metin içerisindeki isimlendirilmiş varlıklar aşağıda listelenen 3 farklı şekilde kullanılmış ve performansları karşılaştırılmıştır.

- **Metni adlandırılmış varlıklarla özetleme (SUM_NER):** Metindeki kişi, konum veya kuruluş varlıkları, NER görevi için eğitilen BERT dil modeli tarafından belirlenir. En az bir adlandırılmış varlık etiketi içeren cümleler belirlenmiş, metin içeriğinde sadece bu cümleler tutularak ve diğer cümleler (herhangi bir adlandırılmış varlık içermeyen) temizlenerek özetlenmiştir. Böylece haberin tanımlayıcı bilgileri metin içinde tutularak, genel ifadeler içeren cümlelerin azaltılması ve üzerinde çalışılacak metnin boyutunun küçültülmesi sağlanmıştır. Toplam 19.848 haber için toplam cümle sayısı 317.646 iken, en az bir isimli varlık içeren toplam cümle sayısı 158.784’tür. Böylece cümle sayısında yaklaşık %50 oranında bir azalma olmuştur.

Tablo 3.3: Türkçe Olay Tabanlı Metin Kümeleme Veri Kümesinden Örnek Haber Kümeleri

Küme 1	
Haber 1	UEFA Avrupa Ligi'nde sıfır puanla grubunda son sırada yer alan Beşiktaş, bu akşam deplasmanda Braga ile oynayacağı maça önemli isimlerinden yoksun çıkacak. Siyah-beyazlı takımda sakatlıktan yeni çıkan Burak Yılmaz, Victor Ruiz, Gökhan Gönül ve N'Koudou Portekiz'e götürülmedi. Takımın parlayan iki yıldızı Atiba ve Diaby de dinlendirilme amacıyla kadroya alınmadı. Kulüpten yapılan açıklamada, "Teknik Direktörümüz Abdullah Avcı geçirdiği ağır grip nedeniyle devam eden tedavisini aksatmamak için takımımızın Braga kafilesinde yer alamamıştır. Hocamız, tedaviye vereceği cevaba göre yarın Portekiz'e gidecektir" denildi.
Haber 2	UEFA Avrupa Ligi'ndeki temsilcimiz Beşiktaş, bu akşam deplasmanda Braga ile karşı karşıya gelecek. Grupta 0 puanla son sırada yer alan siyah-beyazlılarda zorlu mücadelesi öncesi birçok eksik bulunuyor. Burak Yılmaz, Gökhan Gönül, Diaby, N'Koudou ve Ruiz'i kadroya almayan teknik direktör Abdullah Avcı, geçirdiği ağır gribal enfeksiyon nedeniyle kendisini de kafiyele dahil etmedi. Dün, iyileşmesi durumunda Braga'ya gideceğini belirten tecrübeli teknik adam, bugün kendisini iyi hissetmediğini söyleyerek İstanbul'da kalma kararı aldı.
Haber 3	Avrupa Ligi'nde Braga'nın konuğu olacak Beşiktaş'ta, teknik direktör Avcı, geçirdiği ağır grip nedeniyle Portekiz'e gidemedi. Siyah Beyazlılar ilk olarak Avcı'nın düzelmesi durumunda bu ülkeye gideceğini duyurdu. Ancak Siyah Beyazlılar'ın teknik patronunun durumunda bugün de düzelme olmadı. Avcı, Portekiz'e gitmedi. Beşiktaş futbol şube sorumlusu Erdal Torunoğulları da geçirdiği ameliyat nedeniyle kafiyelede yer alamadı. Safra kesesi rahatsızlığı nedeniyle bıçak altına yatan Torunoğulları, tedavisi nedeniyle İstanbul'da kaldı.
Küme 2	
Haber 1	Son dönemde sık sık suya mikrop karışması sonucu mühürlenmesi ile gündeme gelen Şifalı Su için nihayet çalışmalar başlıyor. Akçakoca Belediyesi'nin şubat 2017'de yıllık 20 bin TL bedel ile orman işletmeden 29 yıllığına kiraladığı alanda uygulanacak sosyal proje meclis gündemine taşındı. Kasım ayı son meclis toplantısında Şifalı Su mesire yerinin yatırım ve işletmesinin belediye şirketi ABİTAŞ'a verilmesi kararlaştırıldı.
Haber 2	Akçakoca Belediyesinin kiraladığı D-655 Karayolu üzerindeki Şifalısu Mesire yerinin yatırım ve işletmesini belediyeye ait şirket yapacak. Akçakoca Belediye Başkanı Okan Yanmaz'ın başkanlığında kasım ayı meclis toplantısı, belediye meclis salonunda yapıldı. İmar Planı değişiklik taleplerinin ağırlıklı olarak ele alındığı toplantıda Akçakoca Belediyesinin kiraladığı şifalı su mesire yerinin yatırım ve işletmesinin belediye şirketi ABİTAŞ'a verilmesi kararlaştırıldı.
Haber 3	Orman Bakanlığı tarafından Akçakoca Belediyesi'ne devri yapılan Şifalısu'da artık düğmeye basılıyor. Şubat ayında, Akçakoca Belediye Başkanlığı ile Akçakoca Orman İşletme Müdürlüğü arasında imzalanan protokolle 50 dönümlük orman alanını 29 yıllığına kiralayan Akçakoca Belediyesi Şifalısu ile ilgili önemli karar aldı. Belediye meclisinin bu akşamki oturumunda gündeme gelen Şifalısu orman içi dinlenme alanının geleceğini görüşerek karara bağlayan belediye meclisi, tesisin belediye işbirliği olan ABİTAŞ'a devrine karar verdi. Şifalısu'yu ihaleye çıkarmama kararı alan belediye, tesisi kendi işbirliği olan ABİTAŞ vasıtası ile çalıştıracak.

- **Adlandırılmış varlık gömmelerinin belge gömmesi ile birleştirilmesi (Content+P+L+O):** Cümle gömmeleri, farklı seviyede kelime, cümle ve doküman tabanlı uzay gösterim yöntemleri kullanılarak elde edilmiştir. Ancak olayların ayırt edilmesinde büyük rol oynayan kişi/kurum/kuruluş, zaman ve yer gibi bilgilerin bir anlamda temsile dahil edilmesi, haber metni temsiline yanı sıra isimlendirilmiş varlıklardan elde edilen gömmelere de dahil edilmesi için çalışmalar yürütülmektedir. Bu kapsamda haber metninde yer alan cümlelerde kişi, lokasyon ve organizasyon olarak tanımlanan varlıklar BERT modeli ile belirlenmiş ve kategorilerine göre ayrıştırılmıştır. Her bir varlığın kendi temsili (kelime/cümle/-belge tabanlı belge vektörünü elde etmek için kullanılan temsil yöntemi ile) elde edilmiştir ve her bir kategoriye dahil olan varlıkların gömmelerinin ortalamaları hesaplanarak kişi, lokasyon ve organizasyonun ayrı ayrı temsil vektörleri elde edilmiştir. Elde edilen temsiller belge vektörü ile uç uca eklenerek “İçerik (Content) + Kişi(P) + Konum(L) + Organizasyon (O)” şeklinde formüle edilebilecek bir haber temsili elde edilmiştir.
- **Ortalaması alınan adlandırılmış varlık gömmelerinin, belge gömmesi ile birleştirilmesi (Content+NERs MEAN):** Önceki yöntemden farklı olarak, bu kez kişi, lokasyon ve organizasyon vektörlerinin ortalaması alınarak tek bir isimlendirilmiş varlık gömmesi elde edilmiştir ve elde edilen bu gömme belge temsili ile uç uca eklenmek sureti ile birleştirilmiştir.

3.2.4. Sonuçlar ve Tartışma

Tablo 3.4’de B-cubed değerlendirme metriğine göre kesinlik, duyarlılık ve f-skoru değerleri verilmiştir. Sonuçlara göre elde edilen bulgular;

1. Elde edilen sonuçlara göre OPTICS kümeleme yöntemi, tüm deneylerde HDBSCAN kümeleme yöntemine göre daha düşük performans göstermiştir.
2. Adlandırılmış varlıklar ile metin özetleme yönteminde işlenecek cümle sayısında %50 kazanç sağlanırken, sadece ALBERT modelinde performans düşüşü gözlemlenmiştir. Hiçbir model için kayda değer bir performans kaybı söz konusu değildir.

Tablo 3.4: Adlandırılmış Varlıkların Metin Kümeleme Performansı Üzerindeki Etkileri

İçerik (Content)							
		HDBSCAN			OPTICS		
	Gömme	Kesinlik	Duyarlılık	F-Skoru	Kesinlik	Duyarlılık	F-Skoru
Kelime	Word2Vec	0.87724	0.766252	0.81799	0.81899	0.76078	0.78881
	FastText	0.8653	0.7789	0.8198	0.794898	0.77633	0.785507
	Glove	0.87429	0.77668	0.8226	0.82491	0.7596	0.79093
Cümle	Bert	0.89440	0.78746	0.83753	0.82093	0.76946	0.79436
	Albert	0.91212	0.83792	0.87345	0.78683	0.79256	0.78968
Doküman	Doc2Vec	0.86908	0.76813	0.815495	0.81518	0.765404	0.7895
SUM_NER							
	Gömme	Kesinlik	Duyarlılık	F-Skoru	Kesinlik	Duyarlılık	F-Skoru
Kelime	Word2Vec	0.86487	0.76373	0.81116	0.80754	0.75969	0.78288
	FastText	0.855617	0.776264	0.81401	0.788189	0.770814	0.779404
	Glove	0.86412	0.7732	0.81614	0.81508	0.765883	0.78971
Cümle	Bert	0.8843	0.79295	0.83614	0.79492	0.77566	0.78518
	Albert	0.88701	0.81096	0.84728	0.76682	0.78857	0.77755
Doküman	Doc2Vec	0.86097	0.763103	0.80909	0.80412	0.7701	0.78674
Content+P+L+O							
	Gömme	Kesinlik	Duyarlılık	F-Skoru	Kesinlik	Duyarlılık	F-Skoru
Kelime	Word2Vec	0.76196	0.84333	0.800588	0.72543	0.84098	0.77894
	FastText	0.7789486	0.83206	0.804629	0.728065	0.838327	0.779315
	Glove	0.77629	0.83934	0.80658	0.728117	0.843074	0.78139
Cümle	Bert	0.8067	0.841362	0.82366	0.71517	0.83622	0.77097
	Albert	0.82619	0.8463	0.83613	0.71251	0.82881	0.76627
Doküman	Doc2Vec	0.7723	0.83771	0.80367	0.72298	0.8388	0.7766
Content+NERs_MEAN							
	Gömme	Kesinlik	Duyarlılık	F-Skoru	Kesinlik	Duyarlılık	F-Skoru
Kelime	Word2Vec	0.82693	0.84773	0.8372	0.76433	0.84311	0.801799
	FastText	0.826292	0.8407367	0.833452	0.746501	0.842684	0.79168
	Glove	0.83825	0.84866	0.84342	0.762207	0.84196	0.800101
Cümle	Bert	0.86895	0.8372	0.85278	0.75816	0.80805	0.78231
	Albert	0.84562	0.83315	0.83934	0.718425	0.805039	0.75927
Doküman	Doc2Vec	0.83346	0.84773	0.840536	0.75289	0.84468	0.79615

3. Adlandırılmış varlık gömmelerinin doküman gömmesi ile birleştirilmesi yöntemi vektör boyutunun 3 kat artmasına neden olmuş, ayrıca modelin adlandırılmış varlıklara karşı hassasiyetinin arttığı gözlemlenmiştir. Azalan kesinlik değerleri ve artan duyarlılık değerleri, bu hassasiyetin aynı kişi/kurum veya aynı konum bilgisini içeren haberlerin olay farklı olsa da aynı kümeye dahil edilmesine yol açtığını göstermektedir.

4. Adlandırılmış varlıkların hassasiyetini azaltmak amacıyla doküman gömmesi ve

u uca ortalama adlandırılmıř varlık gmme ekleme yntemi ALBERT modeli hari diğr tm modeller ve her iki kmeleme algoritması iin performans artıřı ile sonulanmıřtır.

Yukarıda sıralanan bulgulara ek olarak alıřma kapsamında yapılan testlerde metinlerdeki isimlendirilmıř varlıkların TDT problemleri iin nemli olduėu grlmřtr. zellikle isimlendirilmıř varlıklar ile metin zetleme yaklařımında sadece isimlendirilmıř varlıkları ieren cmlelerin gmmesi ile elde edilen kmeleme performansı ile ieriğın tamamının gmmesi ile elde edilen performans arasında anlamlı bir fark olmaması bunu aıka ortaya koymaktadır. Adlandırılmıř varlık tespitinde 'olay' olgusunda kiři, kurum, konum gibi nemli unsurların belirlenmesi, bu olguyu temelden inceleyen tm problemler iin nem arz etmektedir.

4. OLAY ÇİZGESİ TABANLI HABER KÜMELEME: İSİMLENDİRİLMİŞ VARLIK MERKEZLİ ALT ÇİZGELERİN ROLÜ

4.1. Giriş

Çevrimiçi veya çevrimdışı platformlar aracılığıyla okuyuculara ulaştırılan haber metinleri, ana olay hakkında hızlı bir şekilde bilgi aktarmak amacıyla ilk birkaç cümlede temel gazetecilik sorularını (kim, ne, ne zaman, nerede ve neden) yanıtlamak üzere tasarlanmıştır. Bu yanıtlar ana olayın özelliklerini tasvir eder ve bir açıklama sağlar [186, 187]. Diğer bilimsel disiplinlerden farklı olarak, ele alınan olayın devam eden dönemleriyle ilgili alt olaylarla nedensel ilişkilerin tartışılması gerekli değildir. Bir haber metni uzun bir süreci tasvir etse bile, kapsadığı olay genellikle şimdiki zaman ya da yakın geçmiş olarak tanımlanır [188]. Sonuç olarak, devam eden bir olayla ilgili haberlerle uğraşırken, okuyucular genellikle güncel haberleri önceki makalelerle ilişkilendirmede zorluklarla karşılaşır ve bu da olayın gelişiminin analizini engeller.

Dijital çağda haber kaynaklarının sayısının katlanarak artması, haber içeriğindeki olayları takip edebilen ve bunları hikayelere dönüştürebilen akıllı haber toplama sistemlerine olan talebin artmasına neden olmuştur. Saygın küresel ve/veya ulusal haber kaynaklarından çeşitli başlıkları toplamayı, aynı olayla ilgili haber metinlerini gruplandırmayı ve verilerden çıkarılan özelliklere dayalı içgörüler sunmayı amaçlayan bu sistemler, doğal dil işleme araştırmalarındaki Konu Tespiti ve Takibi (TDT) görevlerine odaklanan çözümler gerektirmektedir [41, 49]. Bu göreve odaklanan araştırmalarda yürütülen süreç, geleneksel olarak olay tespiti ile başlar ve nihayetinde bir hikaye çizgisi oluşturmayı amaçlar [42, 80]. Literatürde birçok farklı yaklaşım önerilmiş olsa da, haber metinleri çok boyutlu uzayda vektörler olarak temsil edildikten sonra kümeleme çözümleri yaygın olarak kullanılmaktadır [42, 189, 190].

Mevcut haber toplama sistemleri arama tabanlı olay izleme yaklaşımlarını benimsemektedir. Bu yaklaşımlar, bir hikayenin farklı zamanlarda ortaya çıkan ayrık parçalarını etkili bir şekilde birleştirmekte, bunları tutarlı bir zaman çizelgesi halinde düzenlemekte, çatışmaları ayrıştırmakta ve nihayetinde insan beyninin bilişsel süreçleriyle tutarlı bir

anlatı yapısı oluşturmakta başarısız olmaktadır [191]. Ayrıca, neredeyse tüm doğal dil işleme problemlerinde olduğu gibi, çoğu yaklaşım İngilizceye odaklanmaktadır [60]. Geliştirilen çok dilli yaklaşımlar ise dile özgü özelliklere büyük ölçüde bağımlıdır, tek bir dile odaklanan yaklaşımlara kıyasla daha düşük performans sergiler ve Türkçe gibi düşük kaynaklı dillere genişletilmesi zordur [59, 60]. Bu sınırlamalara dayanarak, dilden bağımsız olarak bir haber metnindeki ana olayın özelliklerini temsil edebilen ve benzer olaylarda sıklıkla ortaya çıkan ortak kalıpları birleştirerek genel hikaye bağlamını yakalayabilen bir çizge modeli öneriyoruz. Önerilen bu çizge modelini kullanarak, haber içeriğindeki olayları bağlamından koparmadan ait oldukları hikaye zincirlerine göre kümeleme görevine odaklanıyoruz.

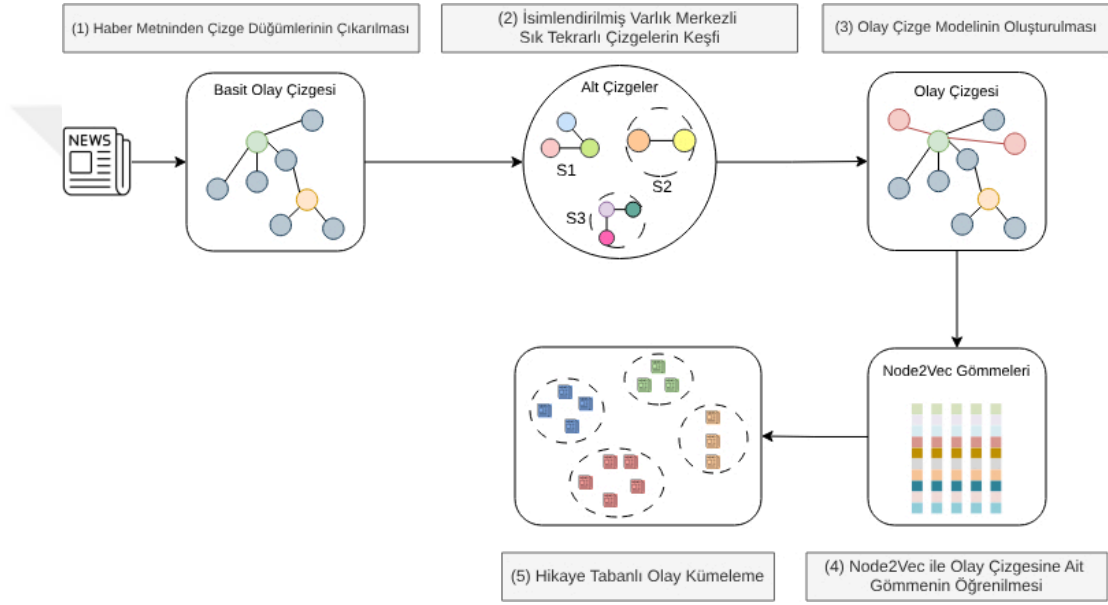
Bu çalışma, alana aşağıdaki şekilde özetlenebilecek katkılar sunmaktadır:

- Bir haber metnindeki olayın özelliklerini temsil etmenin ötesine geçen ve benzer olaylarda sıklıkla ortaya çıkan ortak kalıpları birleştirerek hikayenin genel bağlamını yakalayabilen bir olay çizge modeli önerilmiştir. Önerilen çizge modeli, haber metinlerinin dilden bağımsız olarak yapısal olarak temsil edilmesini sağlar.
- İsimlendirilmiş varlık merkezli sık tekrarlı alt çizgelerin keşfi için bir yaklaşım önerilmektedir. Bu yaklaşım, adlandırılmış varlıkları içeren yinelenen örüntüleri tanımlayarak geleneksel kümeleme algoritmalarının performansını artırmaktadır.
- Düşük kaynaklı bir dil olan Türkçe’de oluşturulan hikaye tabanlı bir haber kümeleme veri kümesi tanıtılmıştır. Bu veri kümesi, başta metin kümeleme olmak üzere doğal dil işlemede çok çeşitli görevler için kullanılabilir.
- Önerilen yaklaşım son teknoloji tek dilli belge kümeleme yaklaşımlarıyla karşılaştırılmıştır. İngilizce ve İspanyolca dillerinde News Clustering veri kümesi için daha yüksek kümeleme performansı ve Almanca dili için ise rekabetçi sonuçlar elde edilmiştir.

4.2. Yöntem

Hikaye tabanlı olay kümeleme problemi için sunulan çözümler, belirli bir olayı (haberi) temsil eden bir $n_1, n_2, \dots, n_n \in N$ çizge kümesi verildiğinde, aynı hikayeye ait çizgeleri içeren $s_1, s_2, \dots, s_m \in S$ alt kümeleri oluşturmayı amaçlamaktadır. Buradaki zorluk, olay

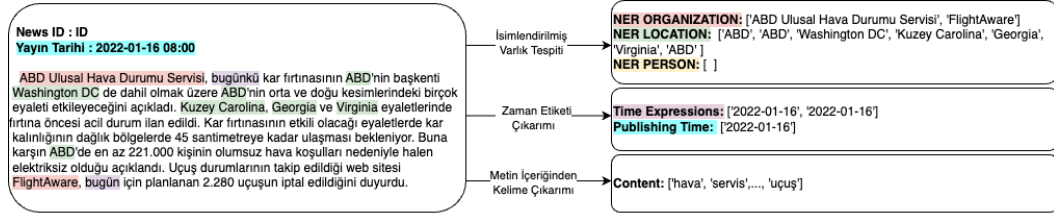
çizge modelini ve bunun gömmesini yüksek ayrıştırma ve temsil gücüyle hesaplamak ve böylece bu alt kümelerin doğru bir şekilde tespit edilmesini sağlamaktır. Şekil 4.1, bu sorunun üstesinden gelmek için bu çalışmada önerilen beş adımlı metodolojinin genel çerçevesini göstermektedir: (1) Haber metninden çizge düğümlerinin çıkarılması, (2) İsimlendirilmiş varlık merkezli sık tekrarlı çizgelerin keşfi, (3) Olay çizge modelinin oluşturulması, (4) Node2Vec ile olay çizgesine ait gömmenin öğrenimi ve (5) Hikaye tabanlı olay kümeleme.



Şekil 4.1: Olay Grafiği Tabanlı Haber Kümeleme Yaklaşımının Genel Yapısı [2]

Tanım 1 (Olay (Event)). Olay, habere konu olan birincil eylem ya da olguyu ifade eder. Bu çalışma kapsamında her bir haber metni bir olay olarak değerlendirilmiştir. Her olay, $n_1, n_2, \dots, n_n \in N$ şeklinde gösterilen bir çizge ile temsil edilir; bu çizge modelindeki hiperdüğüm, olayı tanımlayan benzersiz bir kimlik numarasını temsil eder.

Tanım 2 (Hikaye (Story)). Hikaye, bir haber kaynağında yayınlanan ve bir olayı $n_i \in N$ takip eden belirli bir zamanda $t \in T$ meydana gelen ilgili olayların bir koleksiyonudur. Bir hikaye, $s(s_i)$ kümesinin kardinalitesinin ikiden büyük veya eşit olduğu $s_1, s_2, \dots, s_m \in S$ şeklinde gösterilen olaylar kümesi ile temsil edilir. Hikaye kümesindeki her bir öğe benzersizdir ve aynı hikayedeki diğer öğelerden farklı bilgiler içermektedir.



Şekil 4.2: Haber Metninden Çizge Düzümlerinin Çıkarılması Örneği [2]

Haber metninden çizge düğümlerinin çıkarılması: TDT görevi, kamuoyu üzerinde etkisi olan gerçek dünya olaylarından oluşan bir külliyat kullanır ve birbiriyle oldukça ilişkili olayların hikaye kümeleri halinde düzenlenmesini gerektirir. Bu görevde, bir olay <yer, zaman, ilgili kişiler> üçlülere ile temsil edilir ve zaman içinde bir dizi haber makalesine yol açar [80]. Doğal dil işleme çözümleri tarafından metinsel olarak sunulan haber makalelerinde gizlenen bu üçlülerin doğru ve eksiksiz bir şekilde keşfedilmesi ve olay çizgesi modellerinde düğümler olarak entegrasyonu, olayı temsil edebilecek gömmelerin oluşturulması için çok önemlidir. Önerilen yöntemde örnek bir haber metninden çizge düğümleri elde etme prosedürü Şekil 4.2’te gösterilmiştir.

- **Adlandırılmış Varlık Tanıma (Named Entity Extraction-NER):** Metinsel içerikteki kişi, kurum/kuruluş ve konum gibi varlıkların tanımlanması, doğal dil işleme araştırma alanında adlandırılmış varlık tanıma görevi ile tanımlanmaktadır. Türkçe’ye özgü testler için Bidirectional Encoding Representations from Transformers (BERT) dil modeli kullanılmıştır. Özellikle, Türkçe için önceden eğitilmiş olan BERTurk modeli, kişiler, kuruluşlar ve konumlar için adlandırılmış varlık etiketleri içeren Türkçe NER veri kümesi kullanılarak ince ayarlanmıştır [192]. Eğitilen BERT tabanlı NER modeli %96’yı aşan bir F1 skoru elde etmiştir. Bildiğimiz kadarıyla bu model, Türkçe’de adlandırılmış varlık tespitine yönelik en yüksek performanslı yaklaşımı temsil etmektedir. İngilizce, Almanca ve İspanyolca dilleri için yapılan testlerde Spacy kütüphanesindeki CPU için optimize edilmiş işlem hatları kullanılmıştır. İngilizce [193] için OntoNotes, Almanca [194] için TIGER ve WikiNER derlemleri ve İspanyolca [195] için UD Spanish AnCora ve WikiNER ile önceden eğitilmiş Multi-CNN modelleri sunulmuştur ve NER görevindeki performansları %85 F1 puanının üzerindedir.

- **Zaman Etiketlerinin Çıkarılması:** Geçmişte belirli zaman noktalarında mey-

dana gelen olaylardan oluşan veri kümelerinde “zaman” kavramı, farklı olaylar arasındaki zamansal ilişkileri anlamak için çok önemlidir. Haber metnlerinin analizinde iki tür zaman etiketi dikkate alınmıştır: Açık/örtük zaman ifadeleri ve haberin yayınlanma zamanı.

Açık/örtük zaman ifadeleri, haber içeriğinde bir olayı diğerinden anlamsal olarak ayıran zaman işaretleri olarak tanımlanmaktadır. Bu çalışmada, Heidelberg Üniversitesi Veritabanı Sistemleri Araştırma Grubu tarafından geliştirilen çok dilli zamansal ifade etiketleyicisi HeidelTime, dört tür zamansal ifadeyi çıkarmak için kullanılmıştır: tarih, saat, süre ve tekrarlanan zaman. Bu kural tabanlı açık kaynaklı zaman etiketleyici, metinden zamansal ifadeleri çıkarır ve bunları TIMEX3 standardına göre temsil eder. İngilizce, Almanca ve İspanyolca dahil olmak üzere 13 farklı dil için manuel olarak oluşturulmuş kaynaklar içerir ve 200’den fazla dil için otomatik olarak kaynak oluşturur. HeidelTime, zamansal ifadeleri İngilizce, İspanyolca ve Türkçe için %90’ın üzerinde ve Almanca için %85’in üzerinde bir F1 puanı ile tanımlayabilir [196, 197].

Bir haber makalesinin ne zaman yayınlandığını gösteren zaman etiketini ifade eder. Yayınlanma zamanı, özellikle bir hikayeyi oluşturan olayların zamansal sırasını belirlemek için çok önemli bir bilgidir. Bu çalışmada kullanılan veri kümeleri, her bir haberin yayınlanma zamanının çıkarılmasına olanak tanıyan haber metni, kaynaklar ve tarih bilgilerini içermektedir. Bu iki tür zaman damgasını dikkate alarak, bu çalışma hem haber içeriğinin zamansal yönlerini hem de haber makalelerinin yayın tarihleri aracılığıyla olayların zamansal sıralamasını yakalar.

- **Metin İçeriğindeki Kelimelerin Çıkarılması:** Haber metninde bulunan kelimeler, olay hakkında bilgi aktaran temel unsurları oluşturmaktadır. Her bir haber ögesinin içeriğini birleştirmek için, durak sözcüklerin (stop words), bağlantıların, noktalama işaretlerinin ve sayıların temizlenmesi gibi adımları içeren bir ön işleme hattı uygulanmıştır. Elde edilen kelimeler TF-IDF puanına göre elenmiştir (veri kümesindeki en yüksek TF-IDF puanına sahip 500 kelimelik bir kısıtlama uygulanmıştır) ve kalan kelimelerin her biri olay çizgesinde bir düğüm görevi görür ve standardizasyon için küçük harflere dönüştürülür. Bu kelimeler arasındaki sıralı ilişkiyi kurmak için, metinde görünme sıralarına göre ‘NEXT’ olarak etiketlenen bir kenar eklenmiştir. Haber içeriğini temsil etmek

için kullanılan bu çizge modeli “SnakeGraph” olarak adlandırılır [198]. Snake-Graph, haber metninin yapısal ve sıralı yönlerini yakalayarak olay içindeki kelime düzeyindeki ilişkilerin analiz edilmesini ve görselleştirilmesini sağlar. Bu çizge modelinin dahil edilmesiyle, haber içeriğinin önemli dilsel bileşenleri genel olay temsiline entegre edilir.

Tanım 3 (Adlandırılmış Varlıklar). Bir olay $n_i \in N$ bağlamında, adlandırılmış varlıklar NER modeli kullanılarak metinden tanımlanır. Bu adlandırılmış varlıklar kişi varlıklarından $p_1, p_2, \dots, p_i \in P$, konum varlıklarından $l_1, l_2, \dots, l_k \in L$ ve organizasyon varlıklarından $o_1, o_2, \dots, o_x \in O$ oluşur.

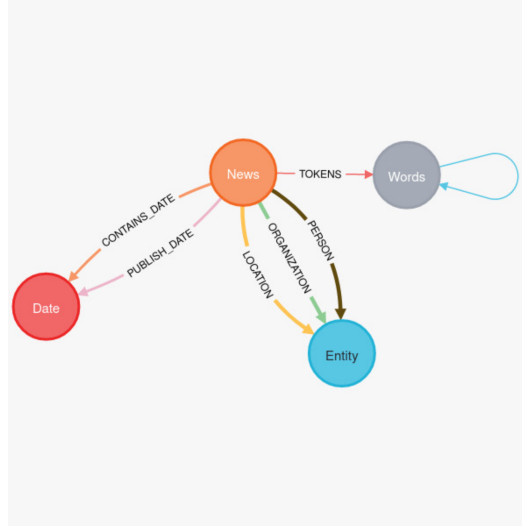
Tanım 4 (Açık/Örtük Zaman İfadeleri). Bir olay $n_i \in N$ bağlamında, zaman ifadeleri metinde tanımlanır. Bu zaman ifadeleri $t_1, t_2, \dots, t_q \in T$ olarak gösterilir ve kural tabanlı bir zamansal ifade etiketleyici kullanılarak belirlenir. “Tarih” türü olarak sınıflandırılır ve zamansal bilgilerin standart bir gösterimini sağlayan TimeML standartlarına uyarlar.

Tanım 5 (Yayınlanma Zamanı). Haber kaynağındaki $n_i \in N$ olayının yayınlanma zamanı $t_1, t_2, \dots, t_q \in T$ tarihleri ile temsil edilir. Bu tarihler, açık/örtük zaman ifadelerine benzer şekilde “tarih” türleri olarak sınıflandırılmıştır. Bunlar, olayların yayınlanma tarihlerini belirtmek için uygun etiketler atayan zamansal ifade etiketleyicisi ile aynı metodoloji kullanılarak tanımlanmıştır.

Tanım 6 (İçerik). Bir olayın içeriği $n_i \in N$ bir dizi benzersiz kelime $w_1, \dots, w_z \in W$ ile temsil edilir. Bu kelimeler, bir olayın ayrıntılarını ve özelliklerini aktaran metinsel bilgileri kapsar.

Adlandırılmış Varlık Merkezli Sık Tekrarlı Alt Çizgelerin Belirlenmesi: Haber kaynaklarından farklı zamanlarda elde edilen olay metinlerinin, önceki ve/sonraki dönemlerde meydana gelen ilgili olaylarla benzerlik göstermesi beklenir. Örneğin, 1999 yılında Türkiye’de meydana gelen büyük depremin ilk günündeki haberler, depremle ilgili birkaç gün devam eden olayları tetiklemiş ve bu olayla ilgili haber metinleri doğrudan ya da dolaylı olarak benzer unsurlar içermiştir. Bu nedenle, aynı hikayeye ait olayların kümelenmesinde sık tekrarlı alt çizgelerin tespiti ile aynı örüntüleri içeren olay çizgelerinin temsillerinin benzerliğinin artırılması amaçlanmıştır.

Aynı hikayeye ait haber metinleri, olayın temel unsurları olarak kabul edilen kişi/kurum,



Şekil 4.3: Basit Olay Çizge Modeli

konum ve zaman etiketleri açısından ortak örüntüler içermeye eğilimindeyken, haber içeriğinde aynı kelime dizilimini kullanarak bir anlatı gerçekleştirmek yaygın bir yaklaşım değildir. Bu nedenle, metin içeriğine odaklanmak yerine, adlandırılmış varlıklara dayalı olarak sık tekrar eden alt çizgelerin keşfine odaklanılmıştır.

Basit Olay Çizge Modeli: Çalışmada, 'Olay Çizge Modeli' için temel teşkil eden ağırlıklandırılmamış yönlendirilmiş bir çizge kullanılarak her bir olay için bir model oluşturulmuştur. Bu basit çizge modeli, benzersiz bir olay kimliğini temsil eden bir hiper düğüm (NEWS) içerir. Buna ek olarak, çizge, olayın metinsel içeriğinden türetilen ardışık benzersiz kelimeler (WORDS), NER modeli tarafından tanımlanan kişi, organizasyon ve konum terimleri (ENTITY) ve zaman etiketleri (DATE) gibi çeşitli düğüm türleri içerir. Hiper düğüm, yedi kenar türü aracılığıyla çeşitli düğümlere bağlanır: TOKENS, NEXT, PERSON, LOCATION, ORGANISATION, PUBLISH DATE ve CONTAINS DATE. Benzersiz olay kimliği hiper düğümü (NEWS), metinde görünme sırasına göre TOKENS ilişkisi kullanılarak grafikteki ilk WORDS düğümüne bağlanır. Kalan WORDS düğümleri, haber metnindeki terimlerin sırasını gösteren NEXT ilişkisi kullanılarak SnakeGraph yapısı ile birbirine bağlanır. Ayrıca NEWS hiper düğümü, varlık türüne bağlı olarak PERSON, LOCATION ve ORGANISATION gibi kenarlar kullanılarak adlandırılmış varlık düğümlerine (ENTITY) bağlanır. İki farklı kenar tipi olan PUBLISH DATE ve CONTAINS DATE, haber metninin yayın tarihi ile NEWS düğümü içeriğindeki açık/örtük zaman ifadelerini birbirine bağlar. Şekil 4.3 basit bir

olay grafiđi yapısını göstermektedir.

Algorithm 1: Simple Event Graph Construction

Input: News Text (*news_text*)

Output: Simple Event Graph representation of the news content (G_{SE})

Function *PreprocessNewsText* (*news_text*):

```
    cleaned_text ← RemoveStopWords (RemoveHyperlinks (RemovePunctuation (RemoveNumbers(news_text))))
    tokens ← TokenizeText(cleaned_text)
    normalized_tokens ← ConvertToLowercase(tokens)
    return normalized_tokens
```

Function *CreateEventGraph* (*normalized_tokens*, *named_entities*, *time_stamps*, *publication_date*):

```
    Initialize EventGraph as an empty graph
    AddNode(EventGraph, "NEWS", event_id)
    AddNode(EventGraph, "WORDS", normalized_tokens[0])
    AddEdge(EventGraph, "NEWS", normalized_tokens[0], label='TOKENS')
    for i from 1 to length(normalized_tokens) - 1 do
        AddNode(EventGraph, "WORDS", normalized_tokens[i])
        AddEdge(EventGraph, normalized_tokens[i-1], normalized_tokens[i], label='NEXT')
    end
    foreach entity, entity_type in named_entities do
        AddNode(EventGraph, entity_type, entity)
        AddEdge(EventGraph, "NEWS", entity, label=entity_type)
    end
    foreach time_stamp in time_stamps do
        AddNode(EventGraph, "DATE", time_stamp)
        AddEdge(EventGraph, "NEWS", time_stamp, label="CONTAINS DATE")
    end
    AddNode(EventGraph, "DATE", publication_date)
    AddEdge(EventGraph, "NEWS", publication_date, label="PUBLISH DATE")
```

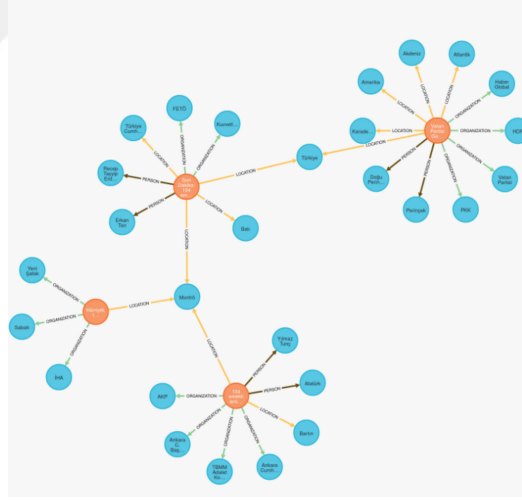
Function *Main* (*news_text*):

```
    normalized_tokens ← PreprocessNewsText (news_text)
    named_entities ← ExtractNamedEntities (news_text)
    time_stamps ← ExtractTimeStamps (news_text)
    publication_date ← ExtractPublicationDate (news_text)
    EventGraph ← CreateEventGraph (normalized_tokens, named_entities, time_stamps, publication_date)
    return EventGraph
```

return

Sık Tekrarlı Alt Çizge Madenciliđi (Frequent Subgraph Mining): Sık tekrarlı alt çizgeler, bir çizge koleksiyonu içinde kullanıcı tarafından belirlenen eşik değerden (*minsup*) daha sık görünür. Bu çalışmada, veri kümesindeki haber metinlerinin çizge modelleri kullanılarak bir çizge kümesi oluşturulmuştur. Bu çizge kümesinden çıkarılan sık tekrarlı alt çizgeler örüntü görevi görür ve bir küme içindeki haber metinleri tarafından paylaşılan karakteristik özellikler hakkında bilgi sağlar. Veri kümesindeki her bir

haber metni farklı uzunluklarda kelime dizilerinden oluştuğu için, ilgili çizgeler farklı sayıda düğüm ve kenara sahip olabilir. Bununla birlikte, her haber metni bir çizge ile temsil edilir ve bu da çok sayıda küçük çizge kümesinden sık tekrarlı alt çizgeleri verimli bir şekilde bulabilen algoritmaların kullanılmasını gerektirir. Yan ve arkadaşları tarafından önerilen gSpan algoritması bu görev için tercih edilen seçim olmuştur [199]. Depth First Search (DFS) kullanır ve verimli çalışma zamanı, gereksiz aday üretiminden kaçınma ve çizge izomorfizmi ile ilişkili hesaplama karmaşıklığının azaltılması gibi önemli zorlukları ele alır. “*minsup*” parametresi, her kümedeki minimum haber metni sayısına göre belirlenir. Bu parametre, bir alt çizgenin sık olarak kabul edilmesi için minimum oluşum sayısı eşiğini belirler ve böylece veri kümesinde anlamlı kalıpların keşfedilmesine olanak tanır. Keşfedilen sık tekrarlı alt çizgeler için 2 aşamalı bir filtreleme uygulanmıştır. İlk olarak, çizge boyutuna göre ve daha sonra destek değerine göre iki aşamalı bir sıralama yapılmıştır. Buna göre, en yüksek değerlere sahip ilk 200 sık tekrarlı alt çizge olarak kabul edilmiş ve diğerleri elenmiştir.

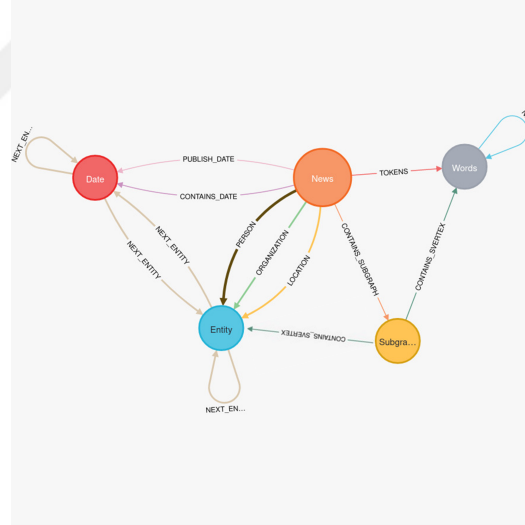


Şekil 4.4: Yıldız Çizge Yapısı Örneği

Sık Tekrarlı Alt Çizge Madenciliğinde Adlandırılmış Varlık Merkezli Yaklaşım:

Basit Olay Çizge modelinde, bir olayı tanımlayan ve diğerinden ayıran temel unsurlar adlandırılmış varlıklar (kişiler, konumlar ve organizasyonlar) ve zamansal bilgilerdir. Farklı metinler arasında benzer sırayla tekrarlanan adlandırılmış varlıkların ve zaman etiketlerinin varlığı, bu metinlerin aynı veya benzer haber içeriğini paylaştığını gösterir.

Çizge modelinde, tüm ENTITY ve DATE düğümleri, farklı kenar türleriyle benzersiz bir olay kimliği (NEWS) hiper düğümüne doğrudan bağlıdır. Ancak, Basit Olay Çizge modelinde ENTITY veya DATE düğümleri arasında doğrudan bir yol yoktur. Bu da hem ENTITY hem de DATE düğümlerini içeren alt çizgeler oluşturmanın mümkün olmadığı anlamına gelir. Bu sınırlama “Yıldız Çizge Problemi” olarak adlandırılır. Şekil 4.4 bir yıldız çizge yapısı örneğini göstermektedir. Bu modelde gSpan algoritması kullanılarak elde edilen alt çizgeler, çoğunlukla haber metni içeriğindeki sıralı ve tekrar eden kelime dizilerinden oluşmaktadır. Bu kelime dizileri aynı hikayeye ait olsalar da, haber metinleri boyunca sık sık ortaya çıkmaları pek olası değildir. Yıldız Çizge Problemini ele almak için, ‘NEXT ENTITY’ adlı bir kenar ekleyerek ENTITY ve DATE düğümleri arasında bir yol oluşturmak için bir yaklaşım önerilmiştir. Bu ekleme, bir olayın temel unsurları olarak kabul edilen hem adlandırılmış varlıkları hem de zaman etiketlerini içeren daha anlamlı ve ayrılabilir örüntülerin oluşturulmasına olanak tanır.



Şekil 4.5: Olay Çizgesi - Adlandırılmış Varlık Merkezli Sık Tekrarlı Alt Çizgelerle Genişletilmiş Olay Çizgesi

Olay Çizge Modeli: Bu çalışmada, hikaye tabanlı olay kümeleme yaklaşımı için “Olay Çizgesi” veya “Adlandırılmış Varlık Merkezli Sık Tekrarlı Alt Çizgelerle Genişletilmiş Olay Çizgesi” olarak adlandırılan bir çizge modeli kullanılmıştır. Bu çizge modeli, basit olay çizgesi modelinin bir uzantısıdır ve Adlandırılmış Varlık Merkezli Sık Tekrarlı Alt

Çizge düğümlerini içerir.

Olay çizgesi beş farklı düğüm tipinden oluşmaktadır: *NEWS*, *WORDS*, *ENTITY*, *DATE* ve *SUBGRAPH*. Bu düğümler benzersiz olay kimliğini, haber metnindeki kelimeleri, adlandırılmış varlıkları, zamansal bilgileri ve alt çizgeleri temsil eder. Çizge ayrıca dokuz farklı kenar türü içerir: *TOKENS*, *NEXT*, *PERSON*, *LOCATION*, *ORGANIZATION*, *PUBLISH DATE*, *CONTAINS DATE*, *CONTAINS SUBGRAPH* ve *CONTAINS VERTEX*. Bu kenarlar, kelimelerin sırası, adlandırılmış varlıklar ile olay kimliği arasındaki bağlantılar ve olay kimliği ile zamansal bilgiler arasındaki bağlantılar gibi düğümler arasındaki ilişkileri temsil eder. Olay çizgesindeki düğüm türleri aşağıdaki gibi tanımlanır:

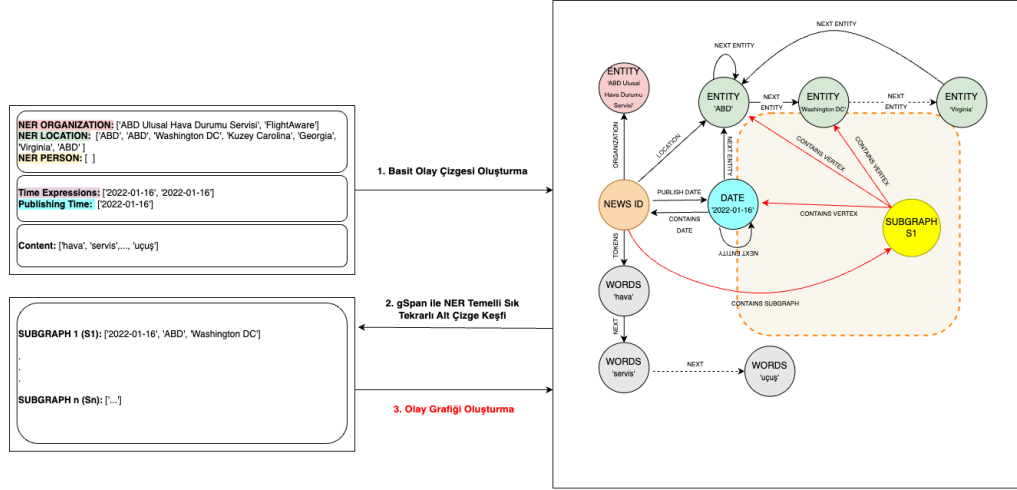
Tanım 7 (NEWS). Olay çizgesindeki *NEWS* düğümü, olay $n_i \in N$ olarak adlandırılan haber metnini diğer haber metinlerinden ayıran benzersiz kimlik numarasını temsil eder.

Tanım 8 (WORDS). Olay çizgesindeki *WORDS* düğümü, haber metni içeriğinde bulunan ve TF-IDF puanına göre filtrelenmiş benzersiz kelimeler $w_1, w_2, \dots, w_z \in W$ kümesini temsil eder. *WORDS* düğümü, olayı tanımlayan ve olay hakkında bilgi veren metinsel içeriği yakalar.

Tanım 9 (ENTITY). Olay çizgesindeki *ENTITY* düğümleri, NER modeli tarafından $n_i \in N$ olay metninden tanımlanan adlandırılmış varlıkları temsil eder. Bu adlandırılmış varlıklar kişileri $p_1, p_2, \dots, p_j \in P$, $l_1, l_2, \dots, l_k \in L$ konumları ve organizasyonları $o_1, o_2, \dots, o_x \in O$ içerebilir.

Tanım 10 (DATE). Olay çizgesindeki *DATE* düğümleri, haber metninin yayın tarihini (olay $n_i \in N$ olarak adlandırılır) ve metindeki tüm zaman ifadelerini temsil eder. Bu zaman ifadeleri, kural tabanlı bir zamansal ifade etiketleyici kullanılarak belirlenmiş ve TimeML standartlarına göre “tarih” olarak kategorize edilmiştir.

Tanım 11 (Adlandırılmış Varlık Merkezli Sık Tekrarlı Alt Çizge). Adlandırılmış varlıkları (*ENTITY*) ve/veya zaman ifadelerini (*DATE*) temsil eden düğümler tarafından yapılandırılan sık tekrarlı örüntüleri temsil eder. gSpan algoritması, önceden tanımlanmış bir minimum destek (minsup) eşiğini aşan bu örüntüleri 2 aşamalı bir filtreleme süreciyle tespit etmek için kullanılır. Tanımlanan her örüntüye *SUBGRAPH* düğümlerinde saklanan benzersiz bir kimlik numarası atanır. Minsup eşiği, bir alt çizgenin sık olarak kabul edilmesi için gereken minimum oluşum sayısını belirleyen kullanıcı tanımlı bir parametre görevi görür.



Şekil 4.6: Haber Metninden Olay Grafiği Oluşturma [2]

Olay çizge modelinde, gSpan algoritması tarafından minsup eşiği ve 2 adımlı filtreleme dayalı olarak tespit edilen her bir yönlendirilmemiş alt çizgeye benzersiz bir kimlik numarası atanır. Bu kimlik numarası, her bir olayın temel çizge yapısı içinde alt çizgeyi temsil eden bir *SUBGRAPH* düğümü oluşturmak için kullanılmıştır. Daha sonra *SUBGRAPH* düğümü ile *NEWS* hiper düğümü arasında, alt çizgenin genel olayın bir parçası olduğunu gösteren bir *CONTAINS SUBGRAPH* kenarı oluşturulur. Ayrıca, tespit edilen alt çizgenin yapısını oluşturan *ENTITY* ve/veya *WORDS* düğümleri ile ilgili *SUBGRAPH* düğümü arasında *CONTAINS VERTEX* kenarları oluşturulur. Bu, bireysel unsurlar (varlıklar veya kelimeler) ile ait oldukları alt çizge arasında bir bağlantı kurulmasını sağlar. Bu ek düğüm ve kenarların dahil edilmesiyle Olay çizge modeli, Şekil 4.5'te gösterildiği gibi her bir olaydaki tekrarlı alt çizgeleri yakalayacak şekilde genişletilmiştir. Buna ek olarak, örnek bir haber metninden bir olay grafiği oluşturma metodolojisi Şekil 4.6'da gösterilmektedir.

Node2Vec ile Olay Temsili Öğrenme: Yönlendirilmiş ve ağırlıklandırılmamış bir olay çizgesindeki $G = (V, E)$ her bir düğüm u , node2Vec modeli ile hesaplanan $ev(u)$ gömme vektörü ile düşük boyutlu bir uzayda temsil edilir. G olay çizgesindeki bir düğümle temsil edilen her bir varlığın (haber, kelime, varlık, tarih ve alt çizge) öğrenilmesi, maksimum olabilirlik optimizasyon problemi olarak formüle edilir. Gömme vektörünün (özellik temsili) $f : V \rightarrow R$ boyutu, gömme vektörleri için öğrenmeyi amaçladığımız düğümlerden eşleme fonksiyonu olarak tanımlanır. Olay grafiğindeki

her kaynak düğüm $u \in V$ için $N_S(u) \subset V$, komşuluk örnekleme stratejisi S tarafından oluşturulan u düğümünün ağ komşuluğudur. f tarafından verilen özellik temsiline göre Denklem 4.1 tarafından verilen amaç fonksiyonunu optimize etmek amaçlanmaktadır:

$$\max_f \sum_{u \in V} \log Pr(N_S(u) | f(u)) \quad (4.1)$$

Belirli bir u düğümünün, ağ rolüne ve/veya ait olduğu topluluğa göre düzenlemek üzere özelleştirilmiş G grafiğindeki temsili öğrenmek için, farklı komşuluklarını keşfetmek üzere önyargılı l uzunluğunda rastgele yürüyüşler oluşturulmuştur. π_{vx} , v ve x düğümleri arasındaki normalleştirilmemiş geçiş olasılığını temsil ederken, Z normalleştirme sabitini ve c_i , $c_0 = u$ düğümü ile başlayan i -inci düğüm olduğu göz önüne alındığında, Denklem 4.2'deki dağılım rastgele bir yürüyüş oluşturmak için kullanılır:

$$P(c_i = x | c_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z} & : \text{eğer } (v, x) \in E \\ 0 & : \text{aksi halde} \end{cases} \quad (4.2)$$

Dönüş parametresi p ve giriş-çıkış parametresi q temelinde ikinci dereceden bir yanlı yürüyüş tanımlanmıştır. Bir sonraki düğümün seçimi statik kenar ağırlıkları $\pi_{vx} = w_{vx}$ kullanılarak gerçekleştirilmiştir. Olay çizgesi G ağırlıksızdır, burada w_{vx} 1'e eşittir. Bir sonraki yürüme adımını belirlemek için normalleştirilmemiş geçiş olasılığı, rastgele bir yürüyüş G olay çizgesindeki (t, v) kenarını geçtiğinde ve şu anda v düğümünde olduğunda $\pi_{vx} = \alpha_{pq}(t, x) \cdot w_{vx}$ şeklindedir. Denklem 4.3'te görüldüğü gibi d_{tx} değerinin t ve x düğümleri arasındaki en kısa yol mesafesini temsil eden 0, 1 veya 2 olması gerektiğine dikkat etmek önemlidir:

$$\alpha_{pq}(t, x) = \begin{cases} \frac{1}{p} & : \text{eğer } d_{tx} = 0 \\ 1 & : \text{eğer } d_{tx} = 1 \\ \frac{1}{q} & : \text{eğer } d_{tx} = 2 \end{cases} \quad (4.3)$$

Bir olay çizgesi G için $ev_{event}(G)$ olay gömmesini elde etmek için, Denklem 4.4'te gösterildiği gibi tüm düğümlerin Node2Vec düğüm gömmelerinin aritmetik ortalaması hesaplanmıştır; burada grafikteki toplam düğüm sayısı $|V|$ ile gösterilmektedir:

$$ev_{event}(G) = \frac{1}{|V|} \sum_{u \in V} ev(u) \quad (4.4)$$

Bu işlem, olay grafiğindeki düğümlerin kolektif bilgi ve özelliklerinin yakalanmasını sağlayarak tüm olay için temsili bir gömme elde edilmesini sağlar. Ortak kelimelere, varlıklara, zamanlara veya alt çizge düğümlerine sahip olay çizgelerinin benzer vektör temsillerine sahip olması dikkat çekicidir. Vektör gösterimlerindeki bu benzerlik, ilgili olay çizgeleri arasındaki anlamsal ve yapısal ilişkileri gösterir.

Hikaye Tabanlı Olay Kümeleme: Çalışmada önerilen olay grafiği modeli ve olay $ev_{event}(G)$ gömmesi, olayların temel özelliklerini ve bu olayların ait olduğu hikayeleri temsil etme ve ilgili hikayelerin kümelerini tanımlamak için benzer özelliklere sahip vektörleri gruplama yeteneklerini analiz etmek amacıyla BIRCH ve HDBSCAN olmak üzere iki farklı geleneksel kümeleme algoritması üzerinde test edilmiştir.

Büyük veri kümelerini işlemek için özel olarak tasarlanmış olan BIRCH algoritması, kümeleme görevlerinin doğasında bulunan hesaplama karmaşıklığını ele alır. Bunu, bir mesafe matrisine dayanmadan önemli kümeleme özelliklerini kapsüllemek için oluşturulmuş ağaç benzeri bir yapı olan Küme Özellik Ağacı (CF Ağacı) kullanarak gerçekleştirir. Bu hiyerarşik ağaç, doğrudan kümeleme işlemlerine olanak tanıyarak büyük hacimli verilerin verimli ve ölçeklenebilir bir şekilde kümelenmesini kolaylaştırır. BIRCH, kümeleri hiyerarşik olarak düzenlemek için dallanma faktörü ' B' ' ve eşik ' T' ' gibi parametreleri kullanır. CF ağaç yapısı, Denklem 4.5'te gösterildiği gibi matematiksel gösterimleri içerir:

$$CF = (N, LS, SS) \quad (4.5)$$

Burada N bir kümedeki veri noktalarını, LS doğrusal toplamı ve SS veri noktası değerlerinin karesel toplamını gösterir. Artımlı bir öğrenme yaklaşımı benimseyen

BIRCH, girdi verilerini sürekli olarak tarar, CF ağacını günceller, yakın mikro kümeleri birleştirir ve kümeleri hiyerarşik olarak düzenler, bu da onu metin kümeleme görevleri için çok uygun hale getirir.

HDBSCAN, hiyerarşik kümeleme ilkelerini kullanan ve farklı şekil ve yoğunluktaki kümeleri tespit etme yeteneği ile karakterize edilen DBSCAN algoritmasının bir evrimidir. Yoğunluğu hesaplamak için minimum nokta sayısı (*MinPts*) ve kümenin minimum boyutu gibi parametreleri kullanır. HDBSCAN kümeleme algoritması ilk olarak veri noktaları arasındaki çift yönlü mesafeleri kodlayan bir karşılıklı erişilebilirlik çizgesi oluşturur. Bu çizge, hiyerarşik bir yapı oluşturmak için sağlam tek bağlantılı kümeleme ile daha da rafine edilen bir minimum yayılma ağacının (*MST*) temelini oluşturur. Hiyerarşik yapıdan türetilen bir kararlılık ölçütü, ulaşılabilirlik mesafelerinin sağlamlığını tahmin ederek küme kararlılığını değerlendirmek için kullanılmıştır. Bu metrik, bir noktanın bir kümeye ait olma olasılığını ve karşılıklı erişilebilirlik mesafelerinin dağılımını ölçerek bir kararlılık puanı elde etmektedir. HDBSCAN'ın yüksek boyutlu uzaylardaki karmaşık yapılara uyum sağlama özelliği, kümeleme görevleri için güçlü bir araçtır.

Haber metinleri, içerdikleri olayın etkisi ve sürekliliğine bağlı olarak değişen büyüklüklerde hikayeler halinde kümelenmektedir. Dolayısıyla bu çalışma, küme sayısı büyük olmasına ve veri kümesi boyutu büyük olmamasına rağmen çok farklı küme yoğunluklarına sahip veri kümeleriyle ilgilenmektedir. Bu nedenle, küme şekli ve yoğunluğundaki farklılıklara duyarlı olan kümeleme algoritmaları bu görev için uygun değildir. Tüm bu kısıtların yanı sıra zaman ve bellek karmaşıklığı da göz önünde bulundurulduğunda, küme sayısını bir giriş parametresi olarak gerektirmeyen BIRCH ve HDBSCAN algoritmaları tercih edilmiştir. Önerilen olay çizge modeli ve vektör temsili olayı temsil etme, ait olduğu hikayeyi algoritmadan bağımsız olarak diğer hikayelerden ayırt etme ve çok yönlü bir bakış açısı sunma kabiliyetini değerlendirmek için iki farklı algoritma üzerinde kıyaslama testleri gerçekleştirilmiştir. Test sonuçları incelendiğinde, BIRCH kümeleme algoritmasının hikaye kümelerini tespit etmek için B-cubed $F1$ skor değerlendirme metriği açısından HDBSCAN'dan daha iyi performans gösterdiği görülmüştür. Ayrıca, BIRCH algoritması $O(n)$ zaman karmaşıklığına sahipken, HDBSCAN $O(n^2)$ zaman karmaşıklığına sahiptir. Kümeleme performansı, zaman ve uzay karmaşıklığı göz önüne alındığında, BIRCH algoritması bu çalışmada tercih

edilen yöntem olarak önerilmektedir.

4.3. Sonuçlar ve Tartışma

Bu bölümde, çalışmada önerildiği gibi adlandırılmış varlık merkezli sık tekrarlı alt çizgelerle genişletilen olay çizgesinin temsil kabiliyetini değerlendirmek için kümeleme görevi üzerinde yapılan çeşitli deneysel çalışmalar ve sonuçları sunulmaktadır. Deneylerde kullanılan değerlendirme veri kümeleri, değerlendirme kriterleri ve stratejileri, deneysel ortamlar ve ortam ayrıntılı olarak açıklanmıştır. Tablo 4.7’de, çizge modelimizin ve kullandığımız kümeleme algoritmasının, aynı değerlendirme veri kümesi ve görevi üzerinde, dilden bağımsız olarak, son teknoloji yaklaşımlara kıyasla üstün performans gösterdiğini ortaya koyan açık bir karşılaştırma sunulmuştur. Ek olarak, basit bir olay grafiğindeki farklı düğüm ve kenar bağlantılarının etkilerini ve bunların çizge modelinin bir olayı temsil etme yeteneği üzerindeki etkilerini ayırt etmek için titiz bir araştırma yapılmıştır. Bu araştırma, bir olayın ve ilgili hikayenin önemli özelliklerini ve ilişkilerini yakalamada adlandırılmış varlık merkezli sık tekrarlı alt çizge düğümlerinin etkinliğini değerlendirmek için çok önemlidir.

Olay Çizgesi Tabanlı Haber Kümeleme yaklaşımının değerlendirilmesi aşağıdaki araştırma sorularının ele alınmasına odaklanmaktadır:

- RQ1. Olayı modelleyen bir Basit Olay Çizgesinin düğümlerinde hangi bilgiler saklanmalıdır ve bunların kümeleme performansı üzerindeki etkileri nelerdir?
- RQ2. Adlandırılmış varlık merkezli sık tekrarlı alt çizge düğümleri ile genişletilmiş bir olay çizge modeli kümeleme performansını ne ölçüde artırır?
- RQ3. Olay çizgesi tabanlı haber kümeleme yaklaşımı ile son teknoloji yöntemler arasındaki performans karşılaştırması nedir?

4.3.1. Deneysel Kurulum

Bu çalışmada yapılan deneylerde Hikaye Tabanlı Haber Veri Kümesi kullanılmış ve News Clustering veri kümesi çok dilli TDT görevi için kıyaslama ölçütü olarak kullanılmıştır. Tüm deneyler 64 GB belleğe, Intel(R) Xeon(R) CPU E5-2680 v3 2.50GHz (24 çekirdek) ve NVIDIA RTX A5000 (24 GB) ekran kartına sahip bir bilgisayarda

gerçekleştirilmiştir. Deney ortamında kurulu işletim sistemi Windows 10 Pro ve programlama dili Python 3.9.15'dir.

Tablo 4.1: News Clustering Veri Kümesi'ne ait İstatistikler

Dil	Eğitim		Test	
	#Doküman	#Hikaye	#Doküman	#Hikaye
İngilizce	12233	593	8726	222
Almanca	4043	377	2101	118
İspanyolca	4527	416	2177	149
Slovene	-	-	37	3
Fransızca	-	-	61	2
İtalyanca	-	-	88	2
Çince	10	1	440	9
Rusça	-	-	231	1
Hırvatça	-	-	13	2

4.3.2. Veri Kümeleri

Hikaye Tabanlı Haber Veri Kümesi: Farklı zamanlarda ve farklı haber kaynaklarında yayınlanan, aynı veya farklı hikayelerden olayları temsil eden haber metinlerinin kümelmesini ele almak için özel olarak oluşturulmuştur. Veri kümesi, Türkçe yazılmış ve çevrimiçi haber platformlarından alınan haber metinleri içeren 2031 olaydan oluşmaktadır. Bu olaylar 168 hikâye halinde düzenlenmiştir. Her hikâye en az iki, en fazla 44 olay içermektedir.

Tablo 4.2'de veri kümesindeki 2 farklı hikayeye ait olan olaylardan bazıları örnek olarak sunulmaktadır. Bu hikayelerde ele alınan konular araştırmacılar tarafından manuel olarak seçilmiştir ve 2021'den bu yana kamu gündemini etkileyen olaylara odaklanılmıştır. Haber verilerini toplamak için araştırmacılar Google News platformunu kullanmış ve her bir konuyla ilgili Türkçe haber makalelerini otomatik olarak alan bir komut dosyası geliştirmiştir. Veri kümesi, doğal dil işleme alanında uzmanlığa ve ilgili uygulamaların geliştirilmesinde pratik deneyime sahip iki bağımsız insan doğrulayıcı

tarafından yürütülen titiz bir doğrulama sürecinden geçmiştir.

”Hikaye Tabanlı Haber Veri Kümesi”nin oluşturulması, araştırmacıların haber metnlerinin belirli hikayeler bağlamında kümelenmesini araştırmasına ve analiz etmesine olanak sağlamaktadır. Veri kümesi, metin kümeleme için yeni yaklaşımlar geliştirmek ve değerlendirmek için bir temel oluşturmaktadır.

İlk aşamada, haberle ilgisi olmayan, eksik veya yanlış içerik içeren ya da mükerrer olan haber metinlerini filtrelemek için bir arındırma işlemi gerçekleştirilmiştir. Önceden işlenmiş haber kümeleri yayın tarihlerine göre sıralanmıştır. İlk haberin yayınlanmasından sonraki üç saate kadar uzanan zaman pencereleri oluşturulmuştur. Aynı haber kümesi ve zaman penceresi içindeki haber metnlerinin benzer içerikte olabileceği ancak farklı haber kaynaklarından gelebileceği öngörülmüştür.

Bir hikayenin olaylarındaki olası yinelemeleri veya tekrarları ele almak için, insan doğrulayıcılar kullanılarak ikinci bir doğrulama süreci gerçekleştirilmiştir. Amaç, veri kümesinin gereksiz veri içermediğinden emin olmaktır. Bu öngörüye uyan haber metinleri arasından sadece karakter sayısı bakımından en uzun metinler hikaye kümesinde tutulmuş, diğer tekrar eden haber metinleri ise elenmiştir. Bu yaklaşım, bir hikaye hakkında farklı bilgiler sağlayan olayların belirlenmesine olanak tanımaktadır.

Veri kümesinin özellikle Türkçe olarak oluşturulmuş olması, Türkçe gibi daha az kaynaklı dillerde çalışan araştırmacılar için önem taşımaktadır. Bu bağlamda, Türkçe doğal dil işleme araştırmacıları için tasarlanan “Hikaye Tabanlı Haber Veri Kümesi” literatürde bir ilk olma niteliği taşımaktadır.

Değerlendirme Veri Kümeleri: Bu makalede önerilen yöntemleri değerlendirmek için “Hikaye Tabanlı Haber Veri Kümesi” ve çok dilli metin kümeleme araştırmalarında yaygın olarak kullanılan bir haber veri kümesi [200] kullanılmıştır.

1. Hikaye Tabanlı Haber Veri Kümesi (SBN): Türkçe veri kümesi 168 farklı hikaye ve 2031 olay (haber makalesi) içermektedir.
2. News Clustering Veri Kümesi (NC): Çok dilli TDT görevinde ölçüt olarak kullanılan veri kümesi İngilizce, İspanyolca, Almanca, Çince, Rusça, Fransızca ve diğer birçok dilde belge içermektedir [200]. Tablo 4.1 veri kümesinin istatistiklerini sunmaktadır. Bu çalışmada yapılan deneylerde üç temel dilde (İngilizce, Almanca ve İspanyolca) sunulan test kümeleri kullanılmıştır. Veri kümesinde

bulunan diğer 6 dil için küme ve doküman sayısı oldukça düşüktür ve literatürde ilgili dillerde tek dilli metin kümeleme performanslarına dair sonuç sunulmamıştır.

4.3.3. Performans Değerlendirme

Değerlendirme Metrikleri: Kümeleme görevlerinde yaygın olarak iki tür değerlendirme ölçütü kullanılır: içsel ve dışsal. İçsel ölçütler, aynı küme içindeki öğelerin yakınlığını ve diğer kümelerden ayrılmasını ölçer. Buna karşılık, dışsal ölçütler, kümeleme sonuçlarının kalitesini ve etkinliğini değerlendirmek için insan yapımı altın standart olarak adlandırılan veri örneklerinin etiketlerinden veya benzer harici temel gerçeklerden yararlanır [201]. Bu çalışmada, insan doğrulayıcılar kullanılarak hazırlanan altın standart veri kümeleri kullanılmıştır. Bu nedenle, önerilen kümeleme algoritmasının performansını değerlendirmek için harici değerlendirme ölçütlerinin kullanılması uygun görülmüştür. B-cubed, değerlendirme ölçütlerinin sezgisel olması, sınırlamaları netleştirmesi, resmi olarak kanıtlanabilir olması ve matematiksel ilkelere göre gruplandırılmış ölçüt aileleri arasında ayırım yapması gerektiği yönündeki tüm resmi kısıtlamaları karşılayan tek harici kümeleme ölçütüdür [202]. Ayrıca, araştırmamızda ele alınan haber kümeleme problemi için literatürde belgelenmiş standart bir değerlendirme ölçütü olarak hizmet etmektedir [80].

Değerlendirme veri kümeleri üzerinde yapılan deneylerde, farklı yapıdaki olay çizgeleri ve adlandırılmış varlık merkezli sık tekrarlı alt çizge düğümlerinin kümeleme performansı üzerindeki etkisini değerlendirmek için B-Cubed Precision, B-Cubed Recall ve B-Cubed F-skorları kullanılmıştır.

Sırasıyla $L(e)$ ve $C(e)$ kategoriye ve e eleman kümesini temsil etsin. Dağılımdaki e ve e' arasındaki ilişkinin doğruluğu ve diğer B-Cubed metrikleri Denklem 4.6, Denklem 4.7 ve Denklem 4.8'de gösterildiği gibi tanımlanır [171, 203]:

$$\text{Doğruluk}(e, e') = \begin{cases} 1 & : \text{eğer ve sadece eğer } L(e') \leftrightarrow (C(e) = C(e')) \\ 0 & : \text{aksi halde} \end{cases} \quad (4.6)$$

Tablo 4.2: Hikaye Tabanlı Haber Veri Kümesi

Hikaye 1 (Küme 1)	
Olay 1	”Son dakika haberi: Uçak Trabzon Havalimanı’nda pistten çıktı” - Ankara-Trabzon seferini yapan Pegasus’a ait yolcu uçağı Trabzon Havalimanı’na inişte pistten çıktı. İlk belirlemelere göre uçaktaki yolcuların durumunun iyi olduğu bildirilmiş ve soğutma işlemleri yapılmıştır. Ankara-Trabzon seferini yapan uçaktaki tüm yolcular güvenli bir şekilde tahliye edildi. Trabzon Valisi, teknik ve idari soruşturma yapıldığını belirterek, ”Herhangi bir yaralanma yok.” dedi.
Olay 2	”Trabzon’daki o uçağın akıbeti 200 gün sonra belli oldu” - Ankara-Trabzon seferini yaparken 13 Ocak’ta Trabzon Havalimanı’na inişi sırasında pistten çıkan Pegasus Havayolları’na ait ‘Zeynep’ isimli uçağın söküme işlemine başlandı. Yaklaşık 20 saat süren söküm işlemi başladı. Uçağın söküm işleminin bir ay içerisinde tamamlanacağı ve tüm parçalarının İngiltere’ye gönderileceği bildirildi.
Olay 3	”Trabzon’da pistten çıkan uçak pide salonu olacak” - Yomra Belediye Başkanı Mustafa Bıyık yaptığı yazılı açıklamada, uçak enkazının belediyeye hiçbir maddi yük getirmeden tüm Karadeniz bölgesinin kullanımına sunulacağını duyurdu. Bıyık, uçağın öntümüzdeki günlerde gerçekleştirilecek bir proje kapsamında pide salonuna dönüştürüleceğini belirtti. Bıyık, uçağın Sancak Mahallesi’nde belediyeye ait çay bahçesine yerleştirileceğini ve kiracılar tarafından işletileceğini açıkladı. Toplam 4 milyon lirayı bulan kurulum ve inşaat masraflarının tamamı işletmeciler tarafından karşılanacak.
Hikaye 2 (Küme 2)	
Olay 1	”Rize-Artvin Havalimanı yıl sonunda hizmete açılıyor” - Vali Çeber, yılda 3 milyon yolcu kapasiteli yeni havalimanının sadece bir ulaşım merkezi olarak değil, Rize’nin, Artvin’in, Karadeniz Bölgesi’nin ve tüm ülkenin ticaret, turizm, tarım, sanayi ve kültürel açıdan kalkınmasına önemli katkılar sağlayacağını vurguladı. Rize’nin eşsiz iklim ve coğrafi özelliklerine dikkat çeken Çeber, “Zaman zaman doğal afetler gibi olağanüstü durumlarla karşılaşsak da, öngörülemeyen bir durum olmazsa havalimanını bu yılın sonunda açmayı öngörüyoruz. Cumhurbaşkanı RTE gelişmeleri yakından izliyor ve günlük ya da haftalık olarak düzenli güncellemeler alıyor.” dedi.
Olay 2	”Rize-Artvin Havalimanı Türkiye’nin yüzölçümünü değiştirdi” - Ulaştırma ve Altyapı Bakanlığı tarafından projelendirilen Rize-Artvin Havalimanı, Rize’nin Pazar ilçesine bağlı Yeşilköy’de 1000 hektarlık bir alan üzerine inşa edildi. Havalimanının temeli 3 Nisan 2017 tarihinde Cumhurbaşkanı RTE tarafından atıldı. Yıllık tahmini 3 milyon yolcu kapasiteli ve 1 milyar 78 milyon lira altyapı maliyetli bu önemli proje, Türkiye’nin çehresini değiştirdi. Deniz dolgusu için yaklaşık 100 milyon ton taş kullanıldı ve ülke yüzölçümü 2,8 kilometrekare genişledi. Bu dolgu işlemiyle oluşturulan arazi, uydu görüntülerinde açıkça görülebilmekte ve bu olağanüstü başarının ölçeğini gözler önüne sermektedir.
Olay 3	”Geceleri ıslıl ıslıl aydınlatılan Rize-Artvin Havalimanı, çay fincanı şeklinde görsel bir şölen sunuyor.” - Türkiye’nin ve Karadeniz Bölgesi’nin deniz dolgusu üzerine inşa edilecek ikinci havalimanı olacak Rize-Artvin Havalimanı projesinin inşaatı istikrarlı bir şekilde ilerliyor. Temelinin 3 Nisan 2017’de atılmasıyla başlayan ve 100 milyon ton taşın kullanıldığı 2,8 milyon metrekareselik geniş bir alanda deniz dolgusunun tamamlanmasının ardından, yöresel mimariye sahip terminal binaları, çay bardağı kulesi ve çay yaprağı şeklindeki giriş süslemesi tamamlandı. Havalimanının öne çıkan özelliklerinden biri de, geceleri çarpıcı bir görsel şölenle büyüleyen ve ışıklandırılmış dış cephesiyle Türk Bayrağı’nı sergileyen çay fincanı şeklindeki kulesi.

$$\text{Kesinlik Bcubed} = \text{Ortalama}_{e'} \left[\text{Ortalama}_{e'.C(e)=C(e')} \left[\text{Doğruluk}(e, e') \right] \right] \quad (4.7)$$

$$\text{Duyarlılık Bcubed} = \text{Ortalama}_{e'} \left[\text{Ortalama}_{e'.L(e)=L(e')} \left[\text{Doğruluk}(e, e') \right] \right] \quad (4.8)$$

Değerlendirme Metodolojisi ve Parametre Ayarları: Bu çalışmada, büyük veri kümelerini verimli bir şekilde işlemek üzere tasarlanmış, az sayıda parametre gerektiren ve küme sayısının önceden belirlenmesini gerektirmeyen, ancak zaman karmaşıklığı ve çalışma metodolojisi açısından birbirinden tamamen farklı iki geleneksel kümeleme algoritması olan BIRCH ve HDBSCAN kullanılmıştır. Deneylerde iki farklı kümeleme algoritmasının kullanılması, farklı düğüm ve kenar tipleri ile oluşturulan çizge modellerinin olay çizge modellemesinin kümeleme performansı üzerindeki etkilerini ve adlandırılmış varlık merkezli sık tekrarlı alt çizge düğümlerinin olay çizgesinin temsil kabiliyeti üzerindeki etkisini kümeleme algoritmalarından bağımsız olarak değerlendirme fırsatı sunmaktadır.

Tablo 4.3: Kümeleme Algoritmalarının Parametre Ayarları

HDBSCAN Kümeleme Algoritması				
Parametreler	Veri Kümeleri			
	SBN	NC	NC	NC
	(Türkçe)	(İngilizce)	(Almanca)	(İspanyolca)
Min Örnek	1	4	3	3
Min Küme Boyutu	2	5	5	4
BIRCH Kümeleme Algoritması				
Dallanma Faktörü	30	34	27	28
Eşik	5	6	5	5
Küme Sayısı	Yok	Yok	Yok	Yok

BIRCH, 2006 SIGMOD Test of Time Ödülü'nü kazanmış olup, bugüne kadarki en hızlı entegre hiyerarşik kümeleme algoritmalarından biri olarak kabul edilmektedir. Önemi, özellikle büyük ölçekli verilerin işlenmesindeki etkinliği nedeniyle sonraki kümeleme metodolojileri üzerindeki derin etkisinde yatmaktadır [204]. BIRCH'in mekanizmasının merkezinde, küme özelliklerinin (CF) düğüm görevi gördüğü küme özellik ağacı (CF Ağacı) yer almaktadır. Bu ağacın mimarisi üç temel parametre tarafından şekillendirilir: eşik (threshold), dallanma faktörü (branching factor) ve küme sayısı (n cluster) [205]. Bu çerçevede, gerçek kümeler CF ağacının yaprak düğümleri tarafından temsil edilir. (T) olarak sembolize edilen eşik parametresi, veri noktası sayısı veya bir yaprak düğüm girişinin kapsayabileceği maksimum çap üzerindeki sınırları belirler[201]. Bir veri noktası mevcut bir kümeye ancak kümenin yarıçapı bu eşik değerin içinde kalırsa eklenir; eşik değerin aşılması, başlangıçta boş olan yeni bir kümenin oluşturulmasını gerektirir. Dolayısıyla bu parametre, kümelerin boyutunu kontrol etmede çok önemli bir rol oynamaktadır. (Br) olarak gösterilen dallanma faktörü, her bir iç düğümün tutabileceği maksimum CF alt kümesi sayısını belirler ve bir düğümün destekleyebileceği çocuk sayısını etkili bir şekilde kısıtlar [204]. Yeni bir kümenin eklenmesi çocuk sayısının dallanma faktörünü aşmasına neden olursa, düğümün ayrılması gerekir. CF ağacı içindeki yapısal dengeyi korumak için, bu tür bir bölünmenin daha yüksek düğümlerde özyinelemeli olması gerekebilir. Ayrıca, önceden tanımlanmış bir küme sayısı parametresi belirlenirse, küme kalitesini optimize etmek için komşu kümelerin birleştirilmesine izin verir. Nihayetinde algoritma ya tamamlandığında önceden belirlenmiş sayıda küme üretecek şekilde ya da küme sayısı parametresini 'yok' olarak ayarlayarak nihai bir kümeleme adımı gerçekleştirilmeden ara kümelerin çıktısını alacak şekilde yapılandırılabilir [201, 204]. Kümelemenin son durumunu tanımlamadaki bu esneklik, BIRCH'in tasarımının önemli bir yönüdür ve belirli verilere veya operasyonel gereksinimlere göre uyarlanabilirliğe izin verir.

HDBSCAN, farklı yoğunluklardaki iç içe geçmiş kümeleri etkili bir şekilde belirleme kapasitesi nedeniyle yoğunluk tabanlı kümeleme alanında önde gelen bir algoritma olarak ortaya çıkmıştır [206]. Diğer yoğunluk tabanlı kümeleme yöntemlerinin aksine, HDBSCAN önemli bir dizi kritik girdi parametresi gerektirmez. Bunun yerine, iki sezgisel parametreye duyarlıdır: "min cluster size" ve "min samples". "min cluster size", bir küme [207] oluşturmak için gereken minimum veri noktası sayısını tanım-

ladığından HDBSCAN’ın temel parametresidir. Sonuç olarak, bu parametre arttıkça, ortaya çıkan kümeler daha büyük olma eğilimindedir ve potansiyel olarak daha küçük kümeleri birleştirerek toplam küme sayısını azaltır [208]. Minimum küme boyutu kriterini karşılamayan veri noktaları gürültü olarak sınıflandırılır. “min samples” parametresi HDBSCAN algoritması için isteğe bağlı bir parametredir ve bir çekirdek noktasının yoğun bir bölge oluşturmak için *eps* yapı çapı olarak adlandırılan belirli bir yarıçap içinde içermesi gereken minimum komşu veri noktası sayısını belirtir [207]. Değerinin artırılması, yüksek yoğunluklu bölgelerin belirlenmesi için eşiği yükseltirken, gürültü olarak kategorize edilen veri noktalarının sayısının artırılması kümeleri giderek daha yoğun bölgelere yoğunlaştırır. Tersine, bu parametre için düşük bir ayar, algoritmanın yerel yoğunluk değişimlerine karşı hassasiyetini artırır ve bu da büyük kümelerin birçok küçük kümeye bölünmesine neden olabilir. ”Min samples” için varsayılan değer ”min cluster size” parametresine eşittir, ancak bu, özellikle gerçek hayat verilerinde veri kümesine bağlı olarak değişebilir [207].

Optimum parametre seçimini sağlamak için, her iki algoritma için dört farklı veri kümesi kullanılarak bir Grid Search gerçekleştirilmiştir. Optimum parametre değerleri belirlenmiş ve Tablo 4.3’te listelenmiştir.

4.3.4. Sonuçlar

Farklı Basit Olay Çizge Modellerinin Kümeleme Performansı Üzerindeki Etkisi: Haber makaleleri <yer, zaman, ilgili kişiler> üçlülere ile temsil edilen olay içerikleridir. RQ1’i ele almak için, metin içeriğini, adlandırılmış varlıkları ve zaman etiketlerini içeren dört farklı temel çizge modeli tasarlanmıştır:

- **İçerik Çizgesi (G_C):** Olay metnini oluşturan benzersiz kelimeler $\{w_1, w_2, \dots, w_n\}$ ve benzersiz bir olay tanımlayıcısı *NEWS* kümesi verildiğinde, $G_C = (V_C, E_C)$ grafiği Denklem 4.9’da formüle edilir:

$$G_C = (\{w_1, w_2, \dots, w_n, NEWS\}, \{e_{NEXT}(w_i, w_{i+1}), e_{TOKENS}(w_1, NEWS)\}) \quad (4.9)$$

- **İsimlendirilmiş Varlık ve Zaman Etiketleri Çizgesi (G_{NeT}):** İsimlendirilmiş varlıklar $\{en_1, \dots, en_m\}$, zaman etiketleri $\{d_1, \dots, d_k\}$ ve benzersiz bir olay tanım-

layıcısı $NEWS$, çizge $G_{NerT}(V_{NerT}, E_{NerT})$ Denklem 4.10'da formüle edilmiştir:

$$G_{NerT} = (\{en_1, \dots, en_m, d_1, \dots, d_k, NEWS\}, \{e_{entityType}(NEWS, en_j), e_{dateType}(NEWS, d_l)\}) \quad (4.10)$$

Burada entityType 'PERSON', 'LOCATION' veya 'ORGANIZATION' olabilir ve dateType 'CONTAINS DATE' veya 'PUBLISH DATE' olabilir.

- **NER ile Özetlenen İçerik Çizgesi (G_{SNer}):** G_{SNer} , G_C ile aynı düğüm ve kenar yapısına sahip olan, ancak en az bir adlandırılmış varlık etiketi içeren cümlelere odaklanan bir çizge modelidir. Bu nedenle $G_{SNer} \subseteq G_C$ olarak ifade edilir.
- **İçerik, Adlandırılmış Varlıklar ve Zaman Etiketleri Çizgesi (Basit Olay Çizgesi- G_{SE}):** Kapsamlı ve yüksek boyutlu çizge modeli, olay metnindeki benzersiz kelimeleri, NER modeli tarafından tanımlanan adlandırılmış varlıkları, cümleler içindeki zaman etiketlerini ve haber yayın tarihini içerir. Bu yapı ile G_{SE} , $G_{SE} \cup G_{NerT}$ şeklinde formüle edilir.

Tüm veri kümeleri ve her iki kümeleme algoritması için bulgular incelendiğinde aşağıdaki gözlemler yapılabilir:

1. Yalnızca olay metnindeki benzersiz kelimelerle oluşturulan G_C çizge modeli, adlandırılmış varlıkları ve zamansal bilgilerini temsil eden ek düğümlerle genişletildiğinde, B-cubed F1 metriği açısından kümeleme performansında %5'e varan önemli bir iyileşme göstermektedir.
2. Yalnızca adlandırılmış varlıklar ve zaman etiketlerinden oluşan G_{NerT} çizge modeli dikkate alındığında, G_C çizge modeline kıyasla B-cubed F1 puanında %10'a varan gözle görülür bir düşüş vardır.
3. Olay içeriğinin en az bir isimli varlık içeren cümlelerle özetlenmesiyle oluşturulan G_{SNer} çizgesi, G_C çizge modeline kıyasla çizge düğüm ve kenar sayısında önemli bir azalmaya yol açmakta, ancak kümeleme performansında önemli bir düşüşe neden olmamaktadır.

Tablo 4.4'ten elde edilen ve yukarıda listelenen tüm sonuçlar göz önünde bulundurulduğunda, olaylar için tanımlayıcı olarak kabul edilen yer, zaman, kişi ve organizasyon unsurları çizge modelinin oluşturulması için güçlü bir temel oluştursa da, metin içeriğinden bağımsız olarak yalnızca bu unsurlarla oluşturulan bir olay çizgesi temsil kabiliyeti açısından yeterli olmadığı görülmektedir.

Tablo 4.4: Farklı Basit Çizge Modellerinin Kümeleme Performansı Üzerindeki Etkisi

HDBSCAN Kümeleme Algoritması												
	SBN			News Clustering			News Clustering			News Clustering		
	(Türkçe)			(İngilizce)			(Almanca)			(İspanyolca)		
	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1
G_C	0.727	0.768	0.745	0.814	0.786	0.800	0.768	0.788	0.778	0.785	0.794	0.789
G_{NerT}	0.664	0.661	0.661	0.722	0.687	0.704	0.691	0.683	0.687	0.706	0.688	0.697
G_{Sner}	0.705	0.799	0.745	0.819	0.826	0.823	0.760	0.824	0.790	0.776	0.830	0.802
G_{SE}	0.773	0.793	0.783	0.861	0.859	0.860	0.815	0.838	0.826	0.833	0.844	0.838
BIRCH Kümeleme Algoritması												
	SBN			News Clustering			News Clustering			News Clustering		
	(Türkçe)			(İngilizce)			(Almanca)			(İspanyolca)		
	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1
G_C	0.839	0.749	0.789	0.881	0.844	0.862	0.829	0.849	0.839	0.842	0.859	0.850
G_{NerT}	0.755	0.679	0.714	0.790	0.751	0.770	0.745	0.754	0.750	0.757	0.763	0.760
G_{Sner}	0.909	0.670	0.772	0.879	0.823	0.850	0.849	0.826	0.838	0.862	0.836	0.849
G_{SE}	0.841	0.808	0.822	0.921	0.917	0.919	0.859	0.914	0.886	0.872	0.925	0.898

Bu araştırma sorusunu yanıtlamak için yapılan deneyler, G_{SE} Basit Olay Çizge Modeli'nin oluşturulması sırasında yapılan analizleri de ortaya koymuştur. Ancak bu analizlerin sonuçları, isimlendirilmiş varlık merkezli sık tekrarlı alt çizge ile genişletilen ve nihai olarak çalışmada önerilen 'Olay Grafiği' modelinin oluşturulmasına temel teşkil edecek bir taban çizgisi oluşturulmasının önünü açmaktadır.

Adlandırılmış Varlık Merkezli Sık Tekrarlı Alt Çizge Düğümlerinin Kümeleme Performansı Üzerindeki Etkisinin Analizi: Toplumsal önemi olan olaylar, haber değerini yitirmeden önce öncelikle metinler aracılığıyla kamuoyuyla paylaşılır. Olayla ilgili gelişmeler ya da olayın tetiklediği diğer olaylar gibi aynı kompozisyonun (hikâyenin) parçası olan diğer metinler, devam eden süreçlerde haber akışının bir parçası haline gelir. Aynı olayın farklı parçalarını oluşturan metinler ortak örüntüler içerme eğilimindedir. Bu noktadan hareketle çalışmada, olay temsilde özel bir öneme sahip olan RQ1 araştırma sorusu için tasarlanan deneylerde tespit edilen, adlandırılmış varlıkların

sık tekrar eden alt örüntülerini tespit etmeye odaklanan bir yaklaşım sunulmaktadır.

G_{SE} olarak adlandırılan Basit Olay Çizge Modeli ile geleneksel kümeleme algoritmalarında hemen her dil için %78'in üzerinde bir performans elde edilmiştir. Haber metni ile birlikte isimlendirilmiş varlıkların ayrı düğümler olarak çizge modeline dahil edilmesinin, olayların ayırt edilebilirliğini artırdığı ve böylece kümeleme performansında önemli bir iyileşmeye yol açtığı gözlemlenmiştir.

RQ2'yi ele almak için, G_{SE} temel olarak ele alındığında, bu çizge modeline adlandırılmış varlık merkezli sık tekrarlı alt çizge düğümlerinin dahil edilmesinin kümeleme performansı üzerindeki etkisini araştırmak için üç farklı çizge modeli geliştirilmiştir. Dört farklı dildeki değerlendirme veri kümesi için deneysel sonuçlar B-cubed Precision, Recall ve F1 değerlendirme metrikleri açısından Tablo 4.5'te sunulmuştur.

Sık Tekrarlı Alt Çizge Düğümlü Basit Olay Grafiği (G_{SE+FS}): G_{SE} çizge kümesinden türetilen sık tekrarlı alt çizgelerden herhangi birini veya birkaçını içeren bir olay çizgesi, bu alt çizgelerin düğüm olarak eklenerek genişletilmiştir ve G_{SE+FS} çizge modeli oluşturulmuştur. G_{SE} çizge modelinde, adlandırılmış varlıklar ve zaman etiketleri doğrudan *NEWS* hiper düğümüne bağlıdır ve aralarında kenar yoktur; bu nedenle, sık tekrarlı alt çizgeler yalnızca haber içeriğindeki tekrar eden kelimelerden türetilir. $\{sg_1, sg_2, \dots, sg_p\}$ sık alt grafiklerin düğümleri olmak üzere, $G_{SE+FS} (V_{SE+FS}, E_{SE+FS})$ Denklem 4.11'de formüle edilmiştir:

$$G_{SE+FS} = \left(\{en_1, en_2, \dots, en_m, d_1, d_2, \dots, d_k, NEWS\}, \right. \\ \left. \{e_{entityType}(NEWS, en_j), e_{dateType}(NEWS, d_l)\} \right) \quad (4.11)$$

“NEXT ENTITY” kenarları ile Basit Olay Grafiği ($G_{SE+NEXTENTITY}$): G_{SE} çizge modelinde, tüm adlandırılmış varlıklar ve zaman etiketleri *NEWS* hiper düğümüne bir kenarla bağlıdır ve kendi aralarında bir kenar yoktur. Sonuç olarak, adlandırılmış varlıklar veya zaman etiketleri arasında bir yol olmadığından, sık tekrarlı alt çizgeler yalnızca metin içeriğinden türetilir. Bu sorunun çözümünde, $G_{SE+NEXTENTITY} (V_{SE+NEXTENTITY}, E_{SE+NEXTENTITY})$ çizge modeli, olayı karakterize eden varlıkları veya zaman etiketlerini temsil eden düğümler arasına *NEXTENTITY* kenarları eklenerek tasarlanmıştır. Çizge modeli Denklem 4.12'de formüle edilmiştir:

$$(G_{SE+NEXTENTITY}) = (\{V_{SE}\}, \{E_{SE} \cup e_{NEXTENTITY}(en_i, en_{i+1}|d_j)\}) \quad (4.12)$$

İsimlendirilmiş Varlık Merkezli Sık Tekrarlı Alt Çizge Düzümlerine Sahip Basit Olay Grafiği (Olay Grafiği- G_E): $G_{SE+NEXTENTITY}$ çizge modeli, adlandırılmış varlıklar veya zaman etiketleri arasında yollar oluşturarak bunların oluşturduğu örüntülerin türetilmesini sağlayan bir temel oluşturur. Bu çalışmada önerilen Olay Grafiği G_E , G_{SE+FS} çizge modelinin Denklem 4.13'te gösterildiği gibi $G_{SE+NEXTENTITY}$ 'den elde edilen sık tekrarlı alt çizgelerle geliştirilmiş güncellenmiş bir varyasyonu olarak tanımlanabilir:

$$(G_{SE+NEXTENTITY}) = (\{V_{SE+FS}\}, \{E_{SE+FS} \cup e_{NEXTENTITY}(en_i, en_{i+1}|d_j)\}) \quad (4.13)$$

Tablo 4.5: Adlandırılmış Varlık Merkezli Sık Tekrarlı Alt Çizge Düzümlerinin Kümeleme Performansı Üzerindeki Etkisi

HDBSCAN Kümeleme Algoritması												
	SBN			News Clustering			News Clustering			News Clustering		
	(Türkçe)			(İngilizce)			(Almanca)			(İspanyolca)		
	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1
G_{SE}	0.773	0.793	0.783	0.861	0.859	0.860	0.815	0.838	0.826	0.833	0.844	0.838
G_{SE+FS}	0.729	0.793	0.756	0.833	0.827	0.830	0.742	0.766	0.754	0.762	0.801	0.781
$G_{SE+NEXTENTITY}$	0.715	0.807	0.768	0.843	0.826	0.835	0.833	0.858	0.845	0.805	0.839	0.822
G_E	0.828	0.823	0.825	0.912	0.904	0.908	0.858	0.923	0.889	0.891	0.9505	0.920
BIRCH Kümeleme Algoritması												
	SBN			News Clustering			News Clustering			News Clustering		
	(Türkçe)			(İngilizce)			(Almanca)			(İspanyolca)		
	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1
G_{SE}	0.841	0.808	0.822	0.921	0.917	0.919	0.859	0.914	0.886	0.872	0.925	0.898
G_{SE+FS}	0.804	0.852	0.815	0.942	0.874	0.907	0.820	0.849	0.834	0.844	0.894	0.869
$G_{SE+NEXTENTITY}$	0.825	0.850	0.834	0.919	0.924	0.922	0.829	0.883	0.855	0.838	0.935	0.884
G_E	0.892	0.890	0.890	0.957	0.941	0.949	0.931	0.940	0.935	0.949	0.923	0.936

RQ2'yi ele almak için tasarlanan çizge modelleri kullanılarak tüm veri kümeleri üzerinde ve iki farklı algoritma ile kümeleme performansları incelendiğinde aşağıdaki sonuçlar

elde edilmiştir:

1. G_{SE+FS} grafik modeli, G_{SE} grafik modelinin genişletilmiş bir varyasyonudur ve haber metni külliyyatında sıklıkla tekrar eden örüntüleri temsil eden düğümlerle güçlendirilmiştir. Ancak, bu düğümlerin kümeleme performansı üzerinde olumlu bir etkisi olmamıştır.
2. $G_{SE+NEXTENTITY}$ çizge modeli, sık tekrarlı alt çizgelere ait düğümlerden bağımsız olarak, basit olay çizgesinde aralarında bağlantı bulunmayan adlandırılmış varlıklar ve zaman düğümleri arasına kenar eklemenin temsil kabiliyetine katkısını değerlendirmek üzere tasarlanmıştır. Bu çizge modeli, veri kümelerinin hiçbirinde olayın temsil kabiliyetini artırma açısından önemli bir fayda göstermemiştir.
3. Haber derleminde, farklı haber metinlerinde tamamen aynı kelime dizilimiyle tekrar eden örüntülerin oluşturduğu alt çizgeler düşük destek değerlerine ve grafik boyutlarına sahiptir. Yüksek destek değerlerine sahip (sık tekrar eden) alt çizgeler incelendiğinde, bu örüntülerin bir olayı diğerinden ayırt edebilecek ya da benzerlik gösterebilecek kelime dizileri değil, çevrimiçi haber kaynaklarının sayfa yapısında standart olan ve/veya aynı kaynaktan gelen tüm haberlerde sık tekrar eden kelime grupları olduğu anlaşılmaktadır.
4. G_E çizge modeli, dilden bağımsız olarak tüm veri kümelerinde ve deneylerde kullanılan kümeleme algoritmalarında diğer çizge modellerine kıyasla en yüksek performansa sahip olmuştur. Bu çalışmada kümeleme görevinde kullanılması önerilen BIRCH kümeleme algoritması ile yapılan deneylerde Hikaye Tabanlı Haber Kümeleme (Türkçe) veri kümesinde %89, Haber Kümeleme (İngilizce) veri kümesinde %94,9, Haber Kümeleme (Almanca) veri kümesinde %93,5 ve Haber Kümeleme (İspanyolca) veri kümesinde %93,6 B-cubed F1 skoru elde edilmiştir.
5. G_E çizge modeli, HDBSCAN kümeleme algoritması için B-cubed F1 puanında %8,2'ye ve BIRCH kümeleme algoritması için %6,8'e varan iyileşme ile G_{SE} çizge modeline kıyasla önemli bir performans artışı göstermiştir.
6. Tablo 4.6'da görüldüğü gibi, Adlandırılmış Varlık Merkezli Sık Tekrarlanan Alt Çizgelerin oranı Hikaye Tabanlı Haber Kümeleme (Türkçe) veri kümesi için %14,14, NC (İngilizce) veri kümesi için %9,28, NC (Almanca) veri kümesi

için %11,46 ve NC (İspanyolca) veri kümesi için %13,45'tir. Bu yoğunluk yüzdeleri, G_E ve G_{SE} grafik modellerini karşılaştırırken B-cubed F1 puanına göre geleneksel kümeleme algoritmalarının performansındaki iyileşme ile doğru orantılıdır. Özellikle, BIRCH Kümeleme Algoritması için, NC (İngilizce) veri kümesi için yaklaşık %3'lük bir performans artışı olurken, Hikaye Tabanlı Haber Kümeleme veri kümesi (Türkçe) için yaklaşık %6,8'lik bir artış olmuştur.

Tablo 4.6: Veri Kümelerindeki Adlandırılmış Varlık Merkezli Sık Alt Çizge Sayıları

	SBN (Türkçe)	NC (İngilizce)	NC (Almanca)	NC (İspanyolca)
G_{SE+FS}	1572	1817	1969	1293
G_E	1831	2003	2224	1494
#NER	259	186	255	201
Yüzde	14.14%	9.28%	11.46%	13.45%

Son Teknoloji Yöntemlere Karşı Kümeleme Performans Karşılaştırması: RQ3 araştırma sorusunu yanıtlamak için, bu çalışmada önerilen yöntemin kümeleme performansı çok dilli News Clustering (NC) kıyaslama veri kümesi üzerinde test edilmiş ve literatürde aynı veri kümesi üzerinde test edilen en gelişmiş (State Of The Art, SOTA) yöntemlerin performansı ile karşılaştırılmıştır. News Clustering veri kümesi dokuz dilden haber metinleri içermektedir. Ancak, İngilizce, Almanca, İspanyolca ve Çince için eğitim kümeleri bulunmaktadır (Çince için eğitim kümesi tek bir hikaye için sadece 10 haberden oluşmaktadır). Tablo 4.1 'de görüldüğü gibi, Almanca, İspanyolca ve İngilizce hariç, diğer tüm diller için test kümeleri bir ila dokuz hikaye içermektedir. Bu nedenle, literatürde NC veri kümesini kullanan SOTA metin kümeleme çalışmaları İngilizce [42, 55, 59, 60, 80], Almanca ve İspanyolca [55, 59, 60] test veri kümeleri üzerinde deneyler yapmış ve B-cubed performans metriklerine dayalı kümeleme performansı karşılaştırmaları sunmuştur.

Tablo 4.7, literatürde yer alan News Clustering veri kümesinin İngilizce, Almanca ve

İspanyolca dilleri için test kümeleri üzerindeki tek dilli metin kümeleme performanslarını sunmaktadır. Tablonun son satırında ise bu çalışmada haber gösteriminde kullanılması önerilen Olay Çizgesi G_E modeli kullanılarak BIRCH kümeleme algoritması üzerinde elde edilen haber kümeleme performansı B-cubed değerlendirme metriği cinsinden sunulmuştur.

Tablo 4.7: Önerilen Yöntem ile SOTA Yöntemlerin Kümeleme Performansı Karşılaştırması

	SBN (Türkçe)			News Clustering (İngilizce)			News Clustering (Almanca)			News Clustering (İspanyolca)		
	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1	Kesinlik	Duyarlılık	F1
Miranda vd. [55]	-	-	-	0.942	0.902	0.923	0.989	0.889	0.936	0.964	0.872	0.916
Staykovski vd. [42]	-	-	-	0.951	0.936	0.944	-	-	-	-	-	-
Linger vd. [59]	-	-	-	0.941	0.935	0.938	0.951	0.943	0.946	0.937	0.900	0.917
Saravanakumar vd. [80]	-	-	-	0.942	0.952	0.947	-	-	-	-	-	-
Santos vd. [60]	-	-	-	0.927	0.921	0.924	0.976	0.900	0.937	0.950	0.862	0.903
Önerilen Yöntem	0.892	0.890	0.890	0.957	0.941	0.949	0.931	0.940	0.935	0.949	0.923	0.936

Miranda vd.[55] haber metinlerini kelimeler, kelime öbekleri ve adlandırılmış varlıkları içeren çeşitli TF-IDF alt vektörlerinden oluşan bir vektör ile temsil etmektedir. TF-IDF ağırlıklı seyrek vektör gösterimlerini doc2vec tabanlı yoğun gösterimler ve bunların kombinasyonlarıyla karşılaştıran Staykovski ve diğerleri, iyi bilinen NewsLens [209] yaklaşımını yeniden uygulayarak haber kümeleme sorununu ele almış ve yalnızca İngilizce dili için değerlendirme sonuçları sunmuştur. Ancak NewsLens, özellikle haber kümeleme araştırma alanındaki öncü çalışmalardan biri olmasına rağmen, toplu işleme metodolojisi nedeniyle dil bağımlılığı ve ölçeklenebilirlik gibi bazı kritik kısıtlamalara sahiptir. Miranda ve diğerleri [55] ve Staykovski ve diğerleri [42] aynı dildeki belgeleri kümelemek için yoğun özelliklerin verimsizliğini savunmuşlardır. Literatürde, her belgenin başlık ve içeriğindeki varlık ve kelimelerin TF-IDF ağırlıklı bir kelime torbası alt vektörü ile temsil edildiği, her belge için tek dilli gösterimler yaygın olarak kullanılmıştır. Bu çalışmanın ardından Linger ve arkadaşları, NewsLens algoritmasının özellikle dil bağımlılığına odaklanan yenilikçi bir uzantısını önermiştir. Haber makalelerini yerel

konular halinde kümelemek için, düğümleri haber makalelerini temsil eden ve kenarları birbirlerine olan benzerliklerine göre ağırlıklandırılan yönlendirilmemiş bir çizge oluşturma yaklaşımını benimsemişlerdir. Bu benzerlik puanı, haber makalelerinin 9 TF-IDF tabanlı temsil alt vektörünün her biri arasındaki kosinüs benzerlikleri ile belirlenir. Bu yaklaşımların hiçbiri olayın ait olduğu hikayenin özelliklerinden faydalanarak bağlamsal ve anlatısal yönlerin derinlemesine analiz edilmesine olanak tanımamaktadır ki bu da çalışmamızın temel katkısıdır. Önceki çalışmalardaki iddiaların aksine, Saravanakumar ve arkadaşları [80] göreve özgü ince ayar, harici bilgiler ve seyrek ve zamansal temsillerin bir kombinasyonu ile güçlendirildiğinde yoğun gömmelerin kümeleme performansını artırdığını deneysel olarak göstermiştir. Burada parametrik olmayan bir k-means algoritmasına dayalı bir kümeleme yaklaşımı önermekte ve küme oluşturma kararlarını vermek için sık bir sinir ağı yaklaşımı kullanmaktadırlar. Sadece İngilizce dili için değerlendirme sonuçları sunan bu çalışmada, haber metni ile kümeler arasındaki benzerlik değerlerini girdi olarak alan sinir ağı için girdi uzayının küçük boyutluluğu aşırı uyum riskini artırmaktadır. Santos ve arkadaşlarının [60] son çalışması, her dil için özel bireysel modellerin kullanımına olan bağımlılığın üstesinden gelen çok dilli belge temsillerini kullanmaktadır. Dolayısıyla, Saravanakumar ve diğerleri [80] tarafından önerilen yaklaşımın aksine, bu çözüm İngilizce diliyle sınırlı kalmamaktadır. Bu makalenin, önerilen yaklaşımın düşük kaynaklı dillere uygulanabilirliğini analiz etmediği unutulmamalıdır. Çok dilli belge temsillerinin eğitimi bağlamında, düşük kaynaklı dillerin dezavantajlı konumunu göz önünde bulundurmak önemlidir.

İncelenen tüm çalışmalardan farklı olarak bu çalışmada, geleneksel metin temsil yöntemleri yerine, haber metinlerinin 'olay' yapısını içeren ve olayın ait olduğu hikayenin özelliklerini de temsil edebilen bir çizge modeli tasarlanmıştır. Diğer çalışmalarda olduğu gibi, sadece haber metni içeriğindeki unsurlardan değil, metinlerdeki açık/gizli varlıklardan, zaman unsurlarından ve sık tekrar eden örüntülerden de yararlanılarak zengin ilişkili bir olay temsili oluşturulmuştur. Bu temsil sadece ayrı olayları değil, aynı zamanda olayların ait olduğu hikaye bağlamını da yakalayabilmektedir. Ayrıca, bu yaklaşımın dilden bağımsız doğasının kümeleme görevi üzerindeki etkili performansı, News Clustering veri kümesinin İngilizce, Almanca ve İspanyolca test kümeleri üzerinde yapılan deneylerle gösterilmiştir. Öte yandan, düşük kaynaklı dillere uygulanabilirliği açısından da öne çıkmaktadır. Sonuçlar Tablo 4.7'de sunulmuştur. Önerilen

yaklaşım İngilizce ve İspanyolca’da en son teknoloji (SOTA) performansına ulaşırken Almanca’da rekabetçi bir performans sergilemektedir.

4.3.5. Kısıtlar

Adlandırılmış Varlık Merkezli Sık Tekrarlı Alt Çizgeler, olay çizgelerinde sıklıkla ortaya çıkan ve olay öğelerine özgü örüntülerdir. Bu alt çizgeler, olay çizgelerindeki adlandırılmış varlıklar ve zamansal düğümler arasına kenarlar eklenerek elde edilebilir. Alt çizgelerin sayısı, öğelerin bağlanma sırasına bağlı olarak değişebilir. Bu sorunu ele almak için bir yaklaşım, tüm elemanlar arasına kenarlar eklemektir; ancak bu, rastgele yürüyüş algoritmalarının yol uzunluğunda ve zaman karmaşıklığında bir artışa yol açar. Bu sınırlamanın üstesinden gelmek için, her haber metninde bulunması gereken yayın tarihinden başlayarak zaman etiketleri, kişiler, konumlar ve kuruluşlar arasına “NEXT ENTITY” bağlantısından bir kenar eklenmiştir. Bu unsurların bağlanma sırası, veri kümesindeki varlıkların sıklık sırasını dikkate almaktadır.

Hikaye Tabanlı Haber Veri Kümesi, gerçek haber kaynaklarından alınan sentetik olmayan haber metinlerinden oluşur ve her bir hikayedeki olay sayısı, hikayenin karmaşıklığına ve önemine bağlı olarak değişir. Örneğin, 2021 yılında Türkiye’de meydana gelen Büyük Manavgat yangını on günden fazla sürmüş, yangın sırasında ve sonrasında yaşanan gelişmeleri kapsayan çok sayıda haber makalesi ortaya çıkmış ve toplamda 60’tan fazla olay meydana gelmiştir. Bununla birlikte, Artemis Programı’nın bir parçası olan Artemis 1 uzay uçuşu görevi, iki yıl boyunca yaklaşık beş yeni olay üretmiştir. Dengesiz sayıda küme elemanına sahip gerçek veri kümelerinde, kümeleme algoritmasının ve parametre değerlerinin dikkatle seçilmesi çok önemlidir. Bu çalışmada, büyük ve dengesiz veri kümeleri üzerindeki performanslarıyla bilinen BIRCH ve HDBSCAN kümeleme algoritmaları kullanılmıştır. En uygun parametre değerleri Grid Search kullanılarak seçilmiştir. Ancak, veri kümesindeki küme boyutları, algoritmaların ve parametrelerin seçiminde sınırlamalar getirmektedir.

Önerilen metodoloji çeşitli dillerde umut vaat etse de, farklı dilsel bağlamlara adaptasyonunda bazı potansiyel sınırlamalar veya zorluklar da sunmaktadır. Olay Grafiği Tabanlı Haber Kümeleme yaklaşımı, haber metinlerini temsil etmek için bir metodoloji kullanmakta, çizge yapısına ve düğümlerin mekansal organizasyonuna dayalı olarak

dilden bağımsız bir şekilde yoğun bir vektör oluşturmaktadır. Bununla birlikte, temel bir adım, haber metinlerinden çıkarılan önemli unsurların düğüm olarak dahil edildiği temel bir olay çizgesinin oluşturulmasını gerektirir. Bu önemli unsurlar arasında adlandırılmış varlıklar da bulunmaktadır. Adlandırılmış Varlık Tanıma (NER) görevi için kullanılan dönüştürücü tabanlı yaklaşımlar, modellerin karmaşık yapısı nedeniyle genellikle büyük veri kümeleri kullanılarak her dil için bağımsız olarak eğitilir. Bununla birlikte, Türkçe gibi düşük kaynaklı diller için önceden eğitilmiş modeller, kapsamlı akademik incelemelerden geçmiş İngilizce gibi dillere kıyasla daha az kapsamlı verilerle mevcuttur. Sonuç olarak, önerilen metodolojinin önemli bir yönü olan NER tespiti, diller arasında farklı performans sonuçları verme potansiyeline sahiptir. Ayrıca, Türkçenin son ekleri ile örneklenen yüksek morfolojik karmaşıklık ile karakterize edilen diller, zaman etiketi tespitinden içerik belirtecine kadar çok sayıda zorluk sunmaktadır.

5. OLAY ÖRGÜSÜ BİLGİ GRAFİKLERİ VE BÜYÜK DİL MODELLERİ İLE ÇOK ADIMLI SORU CEVAPLAMA

5.1. Giriş

Google’ın 2012 yılında Bilgi Grafiği projesini başlatmasıyla terminolojiye giren ”bilgi grafiği (Knowledge Graph-KG)” terimi, özellikle semantik arama ile ilgili olarak çok sayıda akademik makaleye konu olmuştur [210, 211]. Bilgi grafiği, yapılandırılmamış verilerin yapılandırılmış çizge tabanlı bir temsilini sunmanın ötesinde, gerçek dünya varlıklarından oluşan geniş bir anlamsal ağ oluşturmada merkezi bir rol oynamaktadır. Çeşitli kaynaklardan gelen bilgilerin entegrasyonu, bu entegrasyondan doğan ilişkilerden bilginin çıkarılması ve bunun bir çizge olarak görselleştirilebilmesi, yüksek düzeyde açıklanabilirliğe katkıda bulunur [212]. Bilgi grafikleri arama motorlarında önemli bir rol oynamakta ve bir bilgi veritabanına dayandıkları için soru cevaplamayı da kolaylaştırmaktadır.

Bilgi Tabanı Soru Cevaplama (KBQA) sistemleri, son kullanıcıların çizgeyi ve yapısını tam olarak anlamalarını gerektirmeden bilgi grafiğinden veri almalarını sağlar. Bu, kullanıcılar tarafından sorulan yapılandırılmamış bir doğal dil sorusuna (Natural Language Question-NLQ), yapılandırılmış bir formda olan ve oldukça karmaşık bir şema altında temsil edilen bilgi grafiğinde bir cevap bulmak için gerekli adımları başarıyla yürütme problemi olarak tanımlanabilir [211, 212]. Önceki çalışmalar incelendiğinde, bu problem için önerilen çözümlerde iki temel konunun ele alındığı görülmektedir: ’bilgi erişimi’ [213] ve ’anlamsal ayrıştırma’ [214]. Bilgi erişimi, bilgi grafiğinden soruyla en alakalı olgusal üçlülere çıkarmayı ve cevapları belirlemek için bir alt grafik oluşturmayı içerir. Anlamsal ayrıştırma yaklaşımları genellikle yapılandırılmamış doğal dil sorularını seq2seq modelleri [215] kullanarak SPARQL, sorgu grafikleri veya S-ifadeleri gibi bilgi tabanlarında çalıştırılabilir mantıksal formlara dönüştürmeye odaklanır. Seq2seq modelleri KBQA araştırma alanında önemli ilerlemeler kaydetmiş olsa da, hala önemli zorluklarla karşı karşıyadır. Bu zorluklardan ilki, seq2seq semantik ayrıştırma yaklaşımları kullanılarak bilgi grafiğinden çıkarılan ilgili varlıkları ve ilişkileri temsil eden

aday üçlülerin tespit edilmesindeki düşük doğruluktur. Bir diğer önemli zorluk ise yüksek karmaşıklık seviyesine sahip veya bilgi grafiğindeki varlıklar arasındaki karmaşık ilişkilerden kaynaklanan soruların doğru mantıksal formlara dönüştürülmesindeki zorluktur. Öte yandan, varlık bağlama, ilişki sınıflandırma ve mantıksal form oluşturma gibi birden fazla adımdan oluşan bir boru hattı olarak tasarlanan geleneksel KBQA sistemleri, yeniden uygulama, sürdürülebilirlik ve hata yayılma olasılığının artması gibi zorluklarla karşı karşıyadır. Ayrıca, seq2seq modellerinde uzun metin dizilerinin işlenmesi, bağlamsal ayrıntıların yakalanma etkinliğini azaltan bir sınırlamadır. Bu sınırlama, soruların yanlış anlamsal temsillerinin oluşturulmasına yol açabilmektedir [215, 216].

Son yıllarda, bu zorlukların üstesinden gelmek ve geleneksel KBQA yaklaşımlarından daha yüksek verimlilikle çalışan sistemler tasarlamak için çeşitli NLP görevlerinde etkileyici yetenekler sergileyen LLM'lerden yararlanmaya yönelik artan bir ilgi vardır. LLM'lerin doğal dil anlama yeteneklerinin, KBQA'da en gelişmiş performansı elde etmek için bilgi tabanlarının bilgi zenginliğini nasıl tamamlayabileceğini vurgulayan çalışmalar büyük ilgi görmüştür [163, 216, 217]. Bu çalışmalara ek olarak, literatürdeki son makaleler LLM'lerin geleneksel KBQA modellerinin yerini alıp alamayacağını araştırmıştır. Bu araştırmalar, LLM'lerin geleneksel KBQA modellerine göre anlamsal ayrıştırmada üstünlük, gelişmiş verimlilik, görev basitleştirme ve tak-çalıştır esnekliği gibi çeşitli avantajlar sunduğunu belirtmiştir [162]. Bununla birlikte, özellikle açık kaynaklı olmayan ve ince ayar için erişilemeyen bazı LLM'lerin sınırlamalarının ve potansiyel dezavantajlarının ele alınması gerektiği de bildirilmektedir [162, 216].

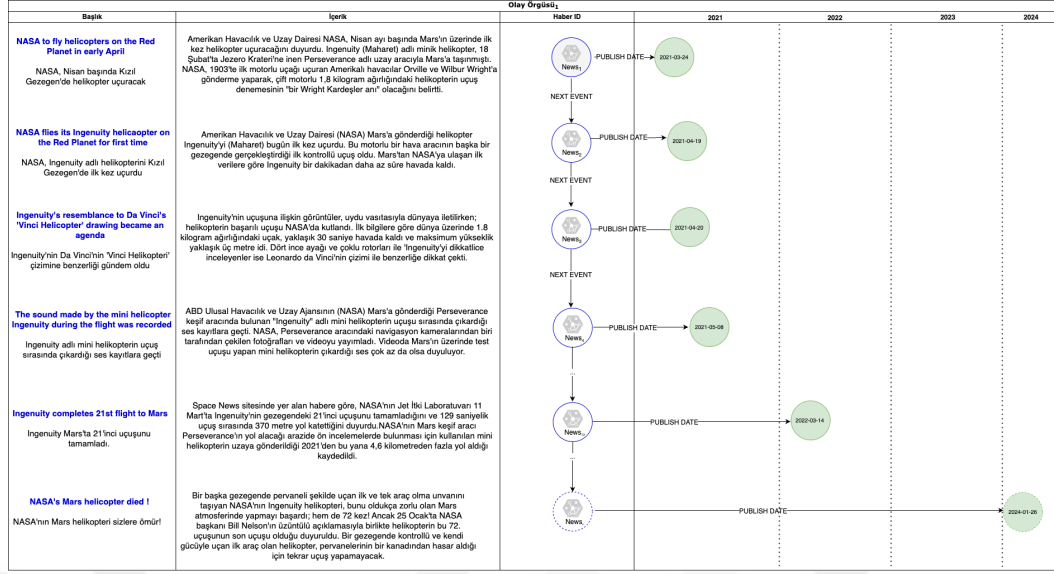
Bu çalışmada, Türkçe dilinde yazılmış ve çevrimiçi haber kaynaklarında farklı zamanlarda yayınlanmış 168 hikayeden 2.031 ayrık olay içeren “Hikaye Tabanlı Haber Veri Kümesi” [2] kullanılmıştır. Veri kümesinden bir bilgi grafiği oluşturmak için çok adımlı bir prosedür her bir olay örgüsüne (hikayeye) uygulanmıştır. Bu işlem hattı sayesinde 168 hikayeyi ve bunlarla ilişkili olayları kapsayan bir haber bilgi tabanı oluşturulmuştur. Bilgi tabanındaki üçlüler temel alınarak oluşturulan hikâye tabanlı KBQA veri kümesi bu çalışmada kullanılmıştır. Ayrıca, anlamsal ayrıştırma tabanlı bir soru cevaplama sistemi tasarlanmış ve yapılandırılmış mantıksal çizge veritabanı sorguları oluşturmak için büyük dil modelleri kullanılmıştır. Büyük dil modellerindeki istemlerin anlamsal zenginliği, görevdeki performansı önemli ölçüde etkilemektedir.

Bu nedenle, bu çalışma kapsamında yapılan deneyler, istem tasarımı sırasında ilgili örneklerin bir Bağlam İçi Öğrenme stratejisine entegre edilmesinin performansı daha da artırabileceğini göstermiştir. Büyük dil modelleri kullanarak tasarlanan bu semantik ayrıştırma tabanlı soru cevaplama sistemi, Bölüm 1'in 3. altbaşlığında listelenen hipotez 2 ve hipotez 3'ün kanıtlanmasına yönelik deneylerin gerçekleştirilmesinde kullanılmıştır. Burada tezin temel katkılarından olan olay grafiğinden hareketle oluşturulan 'olay örgüsü' yapısının çok adımlı soru cevaplama performansına olan etkisi ve yine olay çizge yapısına zaman bilgisi entegre etmenin zamansal soru cevaplama performansına olan etkileri bu çalışmada tartışılmıştır. Tüm bunların yanı sıra, semantik ayrıştırma tabanlı soru cevaplama sistemlerinin tasarımında büyük dil modellerinin kullanım potansiyeli de yapılan deneylerle ele alınmıştır. Tasarlanan sistemin etkinliği ayrıca MuSiQue-Ans Veri Kümesi'nin doğrulama kümesi [3] üzerinde yapılan testlerle detaylı olarak incelenmiştir.

5.2. Metodoloji

Bilgi grafiği (KG), gerçek dünya bilgisini toplamayı ve aktarmayı amaçlayan, varlıkları temsil eden düğümlerden ve bu varlıklar arasındaki ilişkileri temsil eden kenarlardan oluşan bir çizge türü olarak tanımlanır. Bilgi grafikleri geleneksel olarak yönlendirilmiş, kenar etiketli bir çizge veri modeli kullanılarak ifade edilir. Bu veri modelinde K , (s, p, o) üçlülerinden oluşan bir önermeler kümesidir; burada öğeler sırasıyla bir ifadenin öznesini, yüklemine ve nesnesini temsil eder. Bu üçlülerde, özne ve nesne varlıkları temsil eden tanımlayıcılardır, yüklem ise özne ile nesne arasındaki ilişkiyi ifade eden bir tanımlayıcıdır. Bu veri modeli W3C tarafından Kaynak Tanımlama Çerçevesi (RDF) olarak standartlaştırılmıştır. Bir bilgi grafiği olarak K verildiğinde, $E_k = \{e_1, \dots, e_n\}$, K içinde tanımlanan tüm varlıkların kümesidir ve $R_k = \{r_1, \dots, r_m\}$ K içinde tanımlanan tüm ilişkilerin kümesidir. Bilgi grafiğindeki K varlıklar ve ilişkiler onun kelime hazinesini oluşturur [211].

Olay Örgüsü Bilgi Grafiği Oluşturma: Bölüm 4'de isimlendirilmiş varlık merkezli alt çizgelerden faydalanarak, olay örgüsünü oluşturan hikaye parçacıklarının yüksek kümeleme başarımı ile tespit edilmesi sağlanmıştır. Tespit edilen küme elemanlarının, Şekil 5.1'da gösterildiği gibi haberlerin yayınlanma tarihi esas alınarak, haberlere ait kimlik numarasını tutan NEWS hiperdüğümleri arasına eklenen (NEXT EVENT)



Şekil 5.1: Olay Örgüsü Örnek Grafik Gösterimi

kenarları ile birbirine bağlanması sağlanmıştır. Ancak yalnızca haber içeriklerine ait sıralı kelimeler, isimlendirilendirilmiş varlıklar, zaman etiketleri gibi temel düğüm ve kenarlarla inşa edilmiş bu çizge yapısının daha bağlamsal zengin ilişkileri de bünyesinde barındıran ve çok adımlı soru cevaplamaya olanak sağlayan Olay Örgüsü Bilgi Grafiği yapısına dönüştürülmesi için uygulanan adımlar aşağıda listelenmiştir.

- **Bağlantı Çözümleme (Coreference Resolution):** Bağlantı çözümleme görevinde zamirler, atıfta bulundukları varlıklara dönüştürülür. Genellikle, metinlerdeki bağlantı çözümleme örnekleri kişi zamirleridir ve sonuç olarak, bu görev için kullanılan kişi zamirleri atıfta bulundukları varlıkların adlarıyla değiştirir [218]. Bilgi grafiği oluşturmada önemli bir ön işleme adımı olarak kabul edilen bu görev, varlık eşleştirmeye yardımcı olmakta, düğüm sayısını azaltmakta ve sorgulanması kolay anlamlı düğümlerin çizgede görünmesini sağlamaktadır. Çalışmada bağlantı çözümleme adımı, haber metinleri için Crosslingual Coreference Python kütüphanesi ¹ kullanılarak gerçekleştirilmiştir. Bu süreçte, bağımlılık ayrıştırma, adlandırılmış varlık tanıma, AttributeRuler kullanarak ince taneli POS etiketlerini kaba taneli POS etiketlerine eşleme ve lemmatizasyon gibi alt görevler uygulanmaktadır. Bu alt görevler, Spacy kütüphanesinde CPU için optimize edilmiş

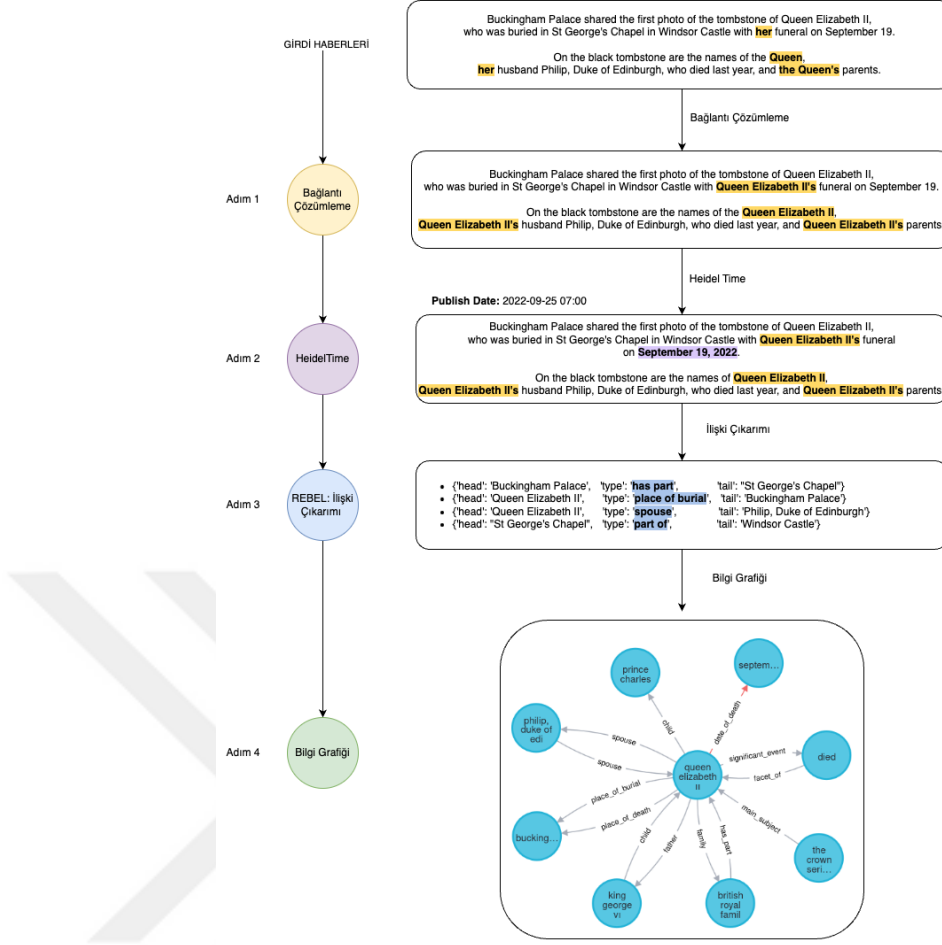
¹<https://pypi.org/project/crosslingual-coreference/>

“en_core_web_lg” İngilizce dil işlem hattı tarafından sağlanmaktadır. Tüm alt görevlerden sonra, AllenNLP’nin SpanBERT gömmelerini kullanan önceden eğitilmiş modeli “coref_spanbert” kullanılarak bağlantı çözümleme gerçekleştirilmektedir.

- **İlişki Çıkarımı (Relation Extraction):** İlişki tespiti, metin dizilerindeki varlıklar arasında anlamlı ilişkiler üretmeyi amaçlar [219]. Önceden tanımlanmış ilişkiler ve gösterge ifadelerinden oluşan arama listelerini kullanan yaklaşımlar ölçeklenebilirlik açısından sınırlamalara sahiptir. Doğal olarak oluşturulmuş metinlerdeki varlıklar arasındaki ilişkiler sonsuz çeşitlilik gösterebilir. Bilgi çıkarma gibi öğrenme görevleri için hassas bir şekilde ayarlanmış önceden eğitilmiş büyük modeller, farklı cümlelerdeki benzer ilişkileri tanımlarken başarılı bir şekilde ilişki etiketi tahminleri üretme yeteneğine sahiptir. BART’a dayanan ve 200’den fazla farklı ilişki türü için uçtan uca ilişki çıkarımı gerçekleştiren REBEL modeli, ince ayarlama yoluyla bir dizi ilişki çıkarımı ve sınıflandırma ölçütünde en gelişmiş performansı elde etmiştir [220]. REBEL modeli, bu çalışmada geliştirilen bilgi grafiği yapısı için ilişki tespit modeli olarak benimsenmiştir. REBEL modeli ile elde edilen ilişki sayısı girdi uzunluğuna bağlı olarak belirlenmiştir ve uygulanan eşik değerleri aşağıdaki gibidir.

- Kelime sayısı 500 veya daha az ise 10 ilişki;
- Kelime sayısı 1000 veya daha az ise 15 ilişki;
- Kelime sayısı 1500 veya daha az ise, 30 ilişki;
- Kelime sayısı 1000 veya daha az ise 45 ilişki;
- Kelime sayısı 2800 (maksimum kelime sayısı) veya daha az ise 50 ilişki.

Büyük Dil Modeli İstemi ile Cypher Query Oluşturma: Bilgi Çıkarımı İşlem Hattı, içerikteki varlıkları ve ilişkileri tanımlayarak yapılandırılmamış haber metninin bir bilgi grafiğine dönüştürülmesini sağlar. Bu çizge koleksiyonu Neo4j çizge veritabanında saklanır. İlişkisel ve zengin bilgi veritabanına dayalı soruları yanıtlamak için Neo4j Graph Cypher Sorguları gereklidir. Kullanıcılardan gelen yapılandırılmamış doğal dil sorularını Cypher Sorgularına dönüştürmek ve bu sorgular üzerinden cevaplar elde etmek için 6 büyük dil modeli kullanılmıştır.



Şekil 5.2: Olay Örgüsü Tabanlı Bilgi Grafiği İşlem Hattı

q kullanıcıdan gelen ve $q[w_1, \dots, w_T]$ şeklinde kelime torbası formatında temsil edilen bir doğal dil sorusu olmak üzere; K 'ya göre anlamsal temsiline, K 'nın kelime dağarcığını kullanan ve q 'nın amacını yakalayan üçlü kalıpları içeren bir q_0 sorgusu olduğunu varsayıyoruz. Örneğin, "Türkiye'nin başkenti neresidir?" sorusu, "başkent" ve "Türkiye"nin K bilgi grafiğindeki sözlüksel öğeler olduğu bir üçlü kalıp ($?x$, başkent, Türkiye) ile temsil edilebilir. Kullanıcıdan gelen bir doğal dil sorusu q verildiğinde, bilgi grafiği soru cevaplama, q 'ya cevap olan varlıkların bir kümesini $A_q \subseteq E_k$ bulma problemini içerir; yani her $e \in A_q$, q 'yu anlamsal olarak temsil eden bir sorgu q_0 için bir cevaptır. Bilgi grafiği soru cevaplama, yapılandırılmamış metin kullanan Bilgi Erişim KG'den farklı olarak, q sorusundaki $N \geq 1$ sıralı kelime dizileri dikkate alınır. Öte yandan, q sorusu yalnızca tek bir üçlü örüntü içeriyorsa, q basit bir soru olarak tanımlanır. Aksi takdirde, q karmaşık bir sorudur. Kullanıcıdan gelen doğal dil sorusu yapılandırılmamış

bir metin biçimindedir. Bu soruya dayalı bir bilgi grafiğinden bilgi çıkarmak için, Neo4j Cypher Sorgularına dönüştürülmesi gerekir. Bu dönüşümdeki ilk adım soru metninde bulunan varlıklar ve ilişkilerin tespitidir. Soru metinleri genellikle kısadır ve bilgi çıkarımı için haber metinlerine kıyasla genellikle basit ilişkilere sahip ve basit cümle kalıplarıyla ifade edilir. Bu cümleler için beklenen ilişki sayısı arttıkça ilişki tespit modelinin performansı düşme eğilimi göstermekte ve çoğu zaman ilgisiz ilişkiler ortaya çıkmaktadır. Bu doğrultuda, her bir doğal dil sorusundan üç adet üçlü elde edilmiştir. Bu üçlüler, sorudan Neo4j Cypher Sorgusuna dönüşüm için komut isteminde kullanılır. Bilgi grafiklerinin oluşturulması ve daha sonra bunlardan bilgi alınması veya çıkarılması, yakın zamana kadar geleneksel olarak denetimli öğrenme metodolojilerine dayanıyordu. Ancak bilgi tabanlarının büyük boyutu, geleneksel KBQA sistemlerinde kullanılan sinir ağı modelleri için büyük bir eğitim ve çıkarım süresine yol açıyor olması sebebiyle, çalışmalar karmaşıklığı azaltmak için bilgi tabanlarında arama aralığını sınırlamak benzeri çeşitli optimizasyon yaklaşımlarına odaklanma eğilimine girmiştir [156, 221]. Ancak tüm bu çalışmalara rağmen iyi performans gösteren bir KBQA modeli için eğitim süresi 2 haftayı aşmaktadır. Performansı büyük oranda etkileyen optimizasyon yaklaşımlarına rağmen, çok uzun eğitim sürelerine sahip olan bu sistemler için son çalışmalarda önceden eğitilmiş dil modellerinin sağladığı önemli kazanımlar, bu dil modellerinin sistemin güçlü bir bileşeni haline gelmesinin önünü açmıştır [222]. Son literatür, büyük dil modellerinin birçok görevde etkileyici bir başarı göstermesine rağmen, bilgi grafiği oluşturma görevinde son teknoloji (SOTA) modellerin gerisinde kaldığını göstermektedir. Öte yandan, büyük dil modellerinin özellikle akıl yürütme ve soru cevaplama görevlerinde nispeten daha üstün performans sergilediği de ortaya konmuştur. Bu değerlendirmeler ve literatürde çeşitli görevler üzerine yapılan son çalışmalar, çok atlamalı soruların üstesinden gelme, bağlamsal ilişkileri anlama ve ön eğitim bilgisinden yararlanma becerileri açısından dil modellerindeki önemli ilerleme düzeyini vurgulamaktadır.

LLM'ler büyük miktarda metin verisi üzerinde ön eğitime tabi tutulur ve metin özetleme, soru yanıtlama ve makine çevirisi de dahil olmak üzere çok çeşitli NLP görevlerinde olağanüstü performans gösteren modern NLP araştırmalarının temel bileşenleridir. Bu dil modelleri bir yanıt üretmek için kullanıcı tarafından sağlanan ve model tarafından üretilen çıktının kalitesini önemli ölçüde etkileyen bir ipucuna ihtiyaç duyar. Bununla

Tablo 5.1: Büyük Dil Modelleri Teknik Detay Tablosu

Model	Mimari	Bağlam Penceresi	Tarih	Kaynak	Maliyet 1M Jeton (USD)
Gpt-3.5-turbo-0125	Kod Çözücü	16.385	Ocak'24	Kapalı	Girdi: 0.50, Çıktı: 1.50
Gpt-4o	Kod Çözücü	128.000	Mayıs'24	Kapalı	Girdi: 5.00, Çıktı: 15.00
Claude 3.5 Sonnet	Kod Çözücü	200.000	Haziran'24	Kapalı	Girdi: 3.00, Çıktı: 15.00
Gemini 1.5 Flash	Kodlayıcı-Kod Çözücü	1.000.000	Mayıs'24	Kapalı	Girdi: 0.07, Çıktı: 0.30
Gemini 1.5 Pro	Kodlayıcı-Kod Çözücü	2.000.000	Şubat'24	Kapalı	Girdi: 3.50, Çıktı: 10.50
Qwen2.5-Coder-32B-Instruct	Kod Çözücü	131.072	Kasım'24	Açık	Girdi: ~0.50, Çıktı: ~0.50

Tablo 5.2: Büyük Dil Modelleri Ortalama Girdi ve Çıktı Karakter Sayısına göre API Çağrı Maliyetleri

Sağlayıcı	Model	Girdi Maliyeti USD (1M)	Çıktı Maliyeti USD (1M)	1 API Çağrı Maliyeti
OpenAI	Gpt-3.5-turbo-0125	0.50	1.50	0.0034
OpenAI	Gpt-4o	5.00	15.00	0.0342
Anthropic	Claude 3.5 Sonnet	3.00	15.00	0.0212
Google	Gemini 1.5 Flash	0.07	0.30	0.0005
Google	Gemini 1.5 Pro	3.50	10.50	0.0239
Alibaba	Qwen2.5-Coder-32B-Instruct	~0.50	~0.50	0.00082

birlikte, literatürdeki bazı yeni çalışmalar, komut isteminde çok az veya hiç gösterim olmadan bile alt görevlerde kayda değer sonuçlar elde edilebileceğini göstermiştir [223, 224]. Bu tür bulgular aslında dil modellerinin genellenebilirliği hakkında daha fazla kanıt sağlamaktadır. Bilgi grafiği oluşturma için, dil modelleri hem sıfır atış hem de birkaç atış ayarlarında mevcut son teknoloji modellerin performansını aşmamaktadır. Bununla birlikte, bilgi grafiği muhakeme görevlerinde, tüm dil modelleri tek atışlı ortamlarda en son teknolojiye sahip performansa ulaşır ve GPT-4 bunu sıfır atışlı ortamlarda başarır [225]. Bu çalışmada, cevaplanacak soruların Neo4j Cypher Sorgularına dönüştürülmesi adımı için sıfır atış ayarında aşağıda listelenen ve detay bilgileri Tablo 5.1'de sunulan 6 farklı büyük dil modeli kullanılmıştır ve her bir dil modeli ile gerçekleştirilen deney 3 kez tekrarlanarak başarımlarının ortalaması alınarak sonuçlar sunulmuştur. Bu amaçla tasarlanan basit komut istemi Şekil 5.3'de gösterilmektedir. Tüm dil modelleri için aynı

komut istemi kullanılmıştır.

OpenAI tarafından sağlanan, kapalı kaynaklı ve çeşitli NLP görevlerinde olduğu gibi semantik ayrıştırma görevinde de büyük dil modellerinin kıyaslanmasında ölçüt olarak kullanılması yaygın olan gpt-3.5-turbo-0125 ve gpt-4o kullanılmıştır [226]. 16.385 jetonluk (token) geniş bir bağlam çerçevesine sahip olan gpt-3.5-turbo-0125 modeli, Eylül 2021'e kadar olan verilerle eğitilmiştir ve parametre sayısı yaklaşık 175 milyardır. Bunun yanı sıra, gpt-4o 128.000 jetonluk bağlam çerçevesine sahipken, resmi olarak açıklanmamış olmakla birlikte parametre sayısının 300 milyar olduğu tahmin edilmektedir ve Ekim 2023'e kadar olan verilerle eğitilmiştir. Gpt-4o çok görevli dil anlayışı testlerinde gpt-3.5-turbo-0125 modeline göre kayda değer bir iyileşme gösterirken, yanlış bilgi üretme (halüsinasyon) oranı da daha düşüktür. Daha karmaşık, bağlamsal ve uzun görevlerde açıkça üstün bir performans sunan gpt-4o modeli, gpt-3.5-turbo-0125 modeline oranla (girdi ve çıktı için milyon jeton başına ödenen ücret bazında) yaklaşık 10 kat daha yüksek maliyetlidir [227].

Task:
Generate Cypher statement to query a graph database.

Instructions:
Use only the provided relationship types and properties in the schema.
Do not use any other relationship types or properties that are not provided.

Schema:

Node properties are the following: {node properties in graph schema}
Relationship properties are the following: {relations properties in graph schema}
The relationships are the following: {relations in graph schema}

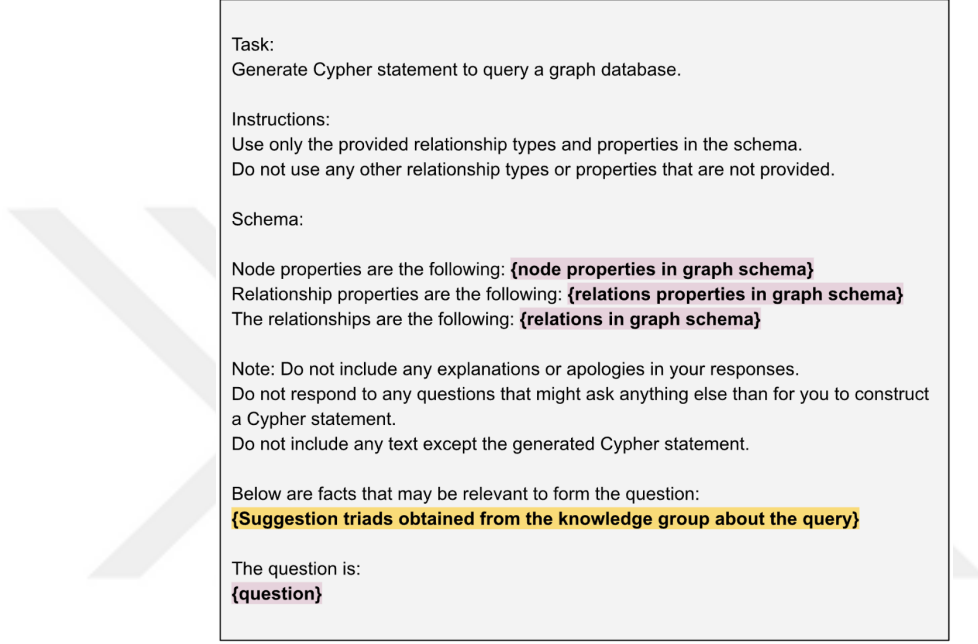
Note: Do not include any explanations or apologies in your responses.
Do not respond to any questions that might ask anything else than for you to construct a Cypher statement.
Do not include any text except the generated Cypher statement.

The question is:
{question}

Şekil 5.3: Basit Neo4j Cypher Sorgu Oluşturma İstemi

Claude 3.5 Sonnet, Anthropic tarafından geliştirilen ve Haziran 2024'de kullanıma sunulan bir dil modelidir, kapalı kaynak kodludur ve model ağırlıkları kamuya açık

sunulmamaktadır. 200.000 jetonluk bağlam penceresine sahiptir ve Nisan 2024’e kadar olan kamuya açık veriler ve veri kümeleri ile eğitilmiştir. Sıfır atış düşünce zinciri tekniği altında, lisans düzeyinde geniş ölçekli çok görevli dil anlama ve matematiksel problem çözme haricindeki diğer tüm değerlendirme görevlerinde gpt-4o’a göre daha üstün performans sergilediği belirtilmektedir [228].



Task:
Generate Cypher statement to query a graph database.

Instructions:
Use only the provided relationship types and properties in the schema.
Do not use any other relationship types or properties that are not provided.

Schema:

Node properties are the following: {node properties in graph schema}
Relationship properties are the following: {relations properties in graph schema}
The relationships are the following: {relations in graph schema}

Note: Do not include any explanations or apologies in your responses.
Do not respond to any questions that might ask anything else than for you to construct a Cypher statement.
Do not include any text except the generated Cypher statement.

Below are facts that may be relevant to form the question:
{Suggestion triads obtained from the knowledge group about the query}

The question is:
{question}

Şekil 5.4: Neo4J Cypher Sorgu Oluşturma için Bilgi Eklentili LM İstemi

Google’ın yapay zeka araştırma birimi DeepMind tarafından geliştirilen ve kapalı kaynak kodlu olarak kullanıma sunulan Gemini 1.5 Flash ve Gemini 1.5 Pro çalışma kapsamında kullanılmıştır. Gemini 1.5 Flash hız ve verimlilik odaklı bir dil modelidir, düşük gecikme süresiyle hızlı yanıtlar üretmek ve çeşitli görevlerde yüksek performans sunmak üzere optimize edilmiştir. Uzun içerikler üzerinde hatırlama ve akıl yürütme gibi yeteneklere odaklanan Gemini 1.5 Pro modeli ise, 1 milyon jetona kadar bugüne kadar ki en geniş bağlam penceresini sunar [229, 230].

Qwen2.5-Coder-32B-Instruct, Alibaba tarafından geliştirilen ve açık kaynak olarak sunulan ileri düzey bir kodlama dil modelidir. Bu model, kod üretimi, hata düzeltme ve kod anlama gibi görevlerde üstün performans sergilemektedir. 32,5 milyar parametreye sahiptir ve 64 katmanlı bir Transformer mimarisi kullanır. Model, 5,5 trilyon jeton

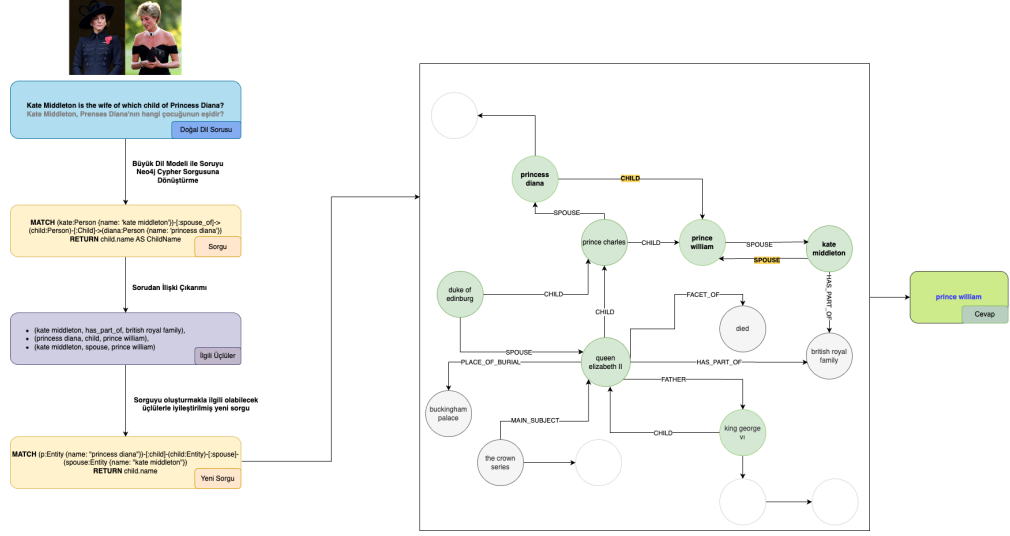
içeren geniş bir veri kümesi üzerinde eğitilmiştir. Bu veri kümesi, kaynak kodları, metin-kod eşleştirmeleri ve sentetik verileri içermektedir. Bu sayede model, kodlama yeteneklerinin yanı sıra genel ve matematiksel becerilerde de üstün performans göstermektedir. 128.000 jetona kadar uzun bağlamları işleyebilme kapasitesine sahiptir ve 40'tan fazla programlama dilinde ve çeşitli benchmark testlerinde açık kaynaklı modeller arasında en yüksek performansı göstermiştir. Model, Hugging Face platformu üzerinden erişilebilir durumdadır [231–233].

Çalışma kapsamında Şekil 5.4 ve Şekil 5.3'de gösterilen dil modeli istemleri ile Neo4J Cypher Query oluşturma süreci tasarlanmıştır. Burada ortalama girdi karakter sayısı 6469 ve çıktı karakter sayısı 121'dir. Her bir sorgu yaratma için 1 API çağrısına ihtiyacı vardır. Bu bilgiler ışığında ortalama maliyet bilgileri Tablo 5.2'da sunulmuştur.

Tablo 5.3: Bilgi ile Güçlendirilmiş Komut İstemi ile İyileştirilmiş Sorgu Örneği

Soru :	"What is the mouth of watercourse for the body of water where Bartram's Covered Bridge is located?"	
Cevap:	"Delaware River"	
Komut İstemi İyileştirme:	Var	Yok
Soru ile İlgili Olabilecek Üçlüler:	[("Bartram's Covered Bridge", 'instance', 'covered bridge'), ("Bartram's Covered Bridge", 'crosses', 'Crum Creek'), ("Bartram's Covered Bridge", 'instance_of', 'covered bridge')]	
Üretilen Cypher Sorgusu:	"MATCH (bridge:Entity {name: 'Bartram's Covered Bridge'})- [:crosses]->(water)-[:mouth_of_the_watercourse]-> (mouth) RETURN mouth.name"	"MATCH (bridge:Entity {name: 'Bartram's Covered Bridge'})- [:located_in_or_next_to_body_of_water]->(watercourse:Entity)- [:mouth_of_the_watercourse]->(mouth:Entity) RETURN mouth.name"
Üretilen Cevap:	"Delaware River"	No match entity

Büyük Dil Modeli İstemi Mühendisliği ile Sorgu İyileştirme: Kullanıcı tarafından sağlanan her girdi bir 'komut' olarak tanımlanır ve dil modeli tarafından üretilen çıktı bir 'yanıt' olarak tanımlanır. Dil modeli bir yanıt üretmek için kullanıcıdan bir komut ister ve tersine, kullanıcı tarafından bir komut verildiğinde ise dil modeli bir yanıt üretir. Karmaşık görevlerde performansı artırmak için birçok gelişmiş yönlendirme tekniği tasarlanmıştır. Örneğin, birkaç örnek içeren "birkaç vuruşluk standart istemler" bu tekniklerden biridir. Bu yaklaşım, bağlam içinde öğrenmeyi kolaylaştırdığı için faydalıdır. Bir diğer başarılı yaklaşım ise belirli bir örnek sunmadan "Adım adım düşün-



Şekil 5.5: Bilgi ile Güçlendirilmiş Dil Modeli İstem Hattı

lim” gibi bir ipucu ekleyerek muhakeme yeteneğini geliştirmeyi amaçlayan “Sıfır Atış Düşünce Zinciri (Zero-Shot CoT)” tekniğidir [234].

Şekil 5.3’te gösterilen istem, Neo4j Cypher sorgularının dil modeli tarafından doğru bir şekilde oluşturulmasını iyileştirmek için bilgi grafiğindeki sorularla ilgili üçlüler kullanılarak geliştirilmiştir. Sorudan elde edilen üçlülerdeki ‘head’ varlıkları ile eşleşen, veritabanındaki düğümlerde ‘head’ veya ‘tail’ varlıkları olan üçlüler komut istemine ek bilgi olarak verilmiştir. Bilgi grafiğindeki olguların dil modeline faydalı bilgi parçaları olarak sunulmasıyla cevaba ulaşmayı daha mümkün kılan sorgular üretilmiştir. Buna ilişkin örnek bir soru cevap sistemi çıktısı Tablo 5.3’de sunulmuştur.

5.3. Deneyler

5.3.1. Değerlendirme Veri Kümeleri

Hikaye Tabanlı KBQA Veri Kümesi: Bu çalışmada, “Hikaye Tabanlı Haber Veri Kümesi” [2] Google Translate API’sini kullanan ücretsiz ve sınırsız bir Python kütüphanesi olan Googletrans kullanılarak otomatik olarak İngilizce diline çevrilmiştir. Daha sonra, Şekil 5.2’de gösterilen boru hattı kullanılarak her bir hikaye için bir bilgi grafiği oluşturulmuştur ve bu bilgi grafiklerinden oluşan bir bilgi tabanı yaratılarak Neo4J çizge veritabanında saklanmıştır. Bu bilgi tabanı temel alınarak, 8518 farklı varlık ve 221

farklı ilişki türü içeren bir QA değerlendirme veri kümesi literatüre kazandırılmıştır. Bu ilişki ve varlık türlerinin oluşturduğu üçlülerden üretilen soru sayısı 767'dir. Bu sorular yaklaşık olarak eşit oranda tek adımlı ve birden fazla adımlı soruları içerecek şekilde tasarlanmış ve Tablo 5.4'deki dağılımlara göre 'kim, ne, hangi, nerede, ne zaman' sorularını kapsayacak ve en fazla 3 adımlı sorular bulunacak, Tablo 5.6'da Musique-Ans veri kümesi üzerindeki örnekleri sunulan ilk üç muhakeme yeteneğini gerektirecek şekilde tasarlanmıştır. Bu veri kümesi aynı haber metninin farklı cümlelerindeki bilgilerin sentezlenmesi ile cevaplanabilir çok atlamalı soruları da bünyesinde barındırmaktadır ve bazı örnek sorular Tablo 5.7 de sunulmuştur.

Tablo 5.4: Hikaye Tabanlı KBQA Veri Kümesi İstatistikleri

Soru Türleri		
	Toplam	%
Ne Zaman (When)	156	20,34
Kim (Who)	299	38,98
Hangi (Which)	169	22,03
Ne (What)	91	11,86
Nerede (Where)	52	6,78
Toplam	767	
Karmaşıklık		
	Toplam	%
Tek Adımlı	455	59,32
Çok Adımlı	312	40,68
Toplam	767	

MuSiQue-Ans Veri Kümesi: Çok adımlı soru cevaplama görevleri için özel olarak tasarlanmıştır ve bünyesinde tek adımlı sorular mevcut değildir. Mevcut çok adımlı QA veri kümelerinin, destekleyici gerçekleri birleştirme ihtiyacı olmadan çözülebilmesi ve bağlantısız akıl yürütmeyi teşvik etmesi gibi zorlukları nedeniyle, gerçek çok adımlı akıl yürütmeyi gerektiren bir veri kümesi oluşturmak amacıyla geliştirilmiştir. Literatürde sıklıkla kullanılan ve kıyaslama veri kümesi olarak kabul edilen SQuAD [117], Natural Questions [235], MLQA [236], TReX [237] ve Zero Shot RE [238] gibi veri kümelerinden tek adımlı soruları seçip birleştirilmesi ve her çok adımlı sorunun doğru cevaplar için destekleyici gerçeklerin bağlantısını gerektirmesini sağlayacak şekilde

düzenlenmesiyle oluşturulmuştur. Veri kümesi Tablo 5.6’da gösterildiği gibi farklı muhakeme yetenekleri gerektiren ve karmaşıklık açısından 2-4 adımlı soru arasında değişen yaklaşık 25 bin soru içerir ve önceki veri kümelerinden önemli ölçüde daha zor olacak şekilde tasarlanmıştır. Bu zorluk, insan ve makine performansı arasındaki üç kat artan fark ve bağlantısız akıl yürütme puanlarında önemli bir azalma ile kanıtlanmıştır [3]. Veri kümesinin yalnızca cevaplanabilir soruları içeren MuSiQue-Ans ve hem cevaplanabilir hem de cevaplanamayan soruları içeren MuSiQue-Full olmak üzere iki çeşidi vardır. Bu çalışmada, deneyler için Tablo 5.5’de istatistikleri sunulan MuSiQue-Ans versiyonunun 1252 soru içeren doğrulama kümesi ve yine bu doğrulama kümesinin yalnızca 2 adımlı sorulardan rastgele 100 soru seçimi ile oluşturulan ve daha önceki çalışmalarca kullanılmış olan ”validation-mini” versiyonu kullanılmıştır. MuSiQue-Ans veri kümesinin ”validation-mini” versiyonunda yer alan soru kimlik numaraları ve diğer tüm ayrıntılar referans çalışmada sunulmuştur [239].

Tablo 5.5: MuSiQue-Ans Veri Kümesi İstatistikleri

	2-Adımlı	3-Adımlı	4-Adımlı	Toplam (24814)
Eğitim	14,376	4,387	1,175	19,938
Doğrulama	1,252	760	405	2,417
Test	1,271	763	425	2,459

5.3.2. Değerlendirme Metrikleri

Doğal dil sorularından yapılandırılmış Neo4J çizge veritabanı sorgu dili Cypher Query sorgusuna dönüşüm başarımının ölçülmesinde, literatürde kullanılması önerilen ve aşağıda hesaplama yöntemi sunulan üç farklı değerlendirme metriği kullanılmıştır.

Yürütme Oranı (Execution Rate - ER): İlgili görev için kullanılan modelin çıktı sorgusunun çizge veritabanı üzerinde çalıştırılabilirliğini değerlendiren metriktir [240]. Bu metrik Denklem 5.1 ile tanımlanmıştır.

$$ER = \frac{\text{Çalıştırılabilir sorgu sayısı}}{\text{Toplam çıktı sorgusu sayısı}} \quad (5.1)$$

Altın Sorgu Doğruluğu (Gold Query Accuracy-GQA): İlgili görev için kullanılan modelin çıktı sorgusu ile altın sorgu arasındaki benzerliği BERTScore [241] kullanarak değerlendiren ve Denklem 5.2 ile tanımlanmış olan metriktir. Burada O_i modelin i . çıktı sorgusunu, G_i ise i . altın sorgusunu temsil etmektedir [240].

$$GQA = \frac{1}{N} \sum_{i=1}^N \text{BERTScore}(O_i, G_i) \quad (5.2)$$

Tam Eşleşme (Exact Match-EM): İlgili görev için kullanılan modelin çıktı sorgusunun çizge veritabanında yürütülmesi sonucunda elde edilen cevap ile, altın sorgunun yürütülmesi sonucunda elde edilen cevabın tam eşleşmesini ölçen metriktir [162].

BERTScore, yalnızca tam kelime eşleşmelerine odaklanmanın ötesinde, anlamsal benzerlik ve bağlam uyumunu da değerlendiren bir metrik olarak öne çıkar. Bu çalışmada önerilen üç metriğin kullanılması durumunda, yüksek puanlar tercih edilir. ER metriğinin yüksek değerleri, çıktı sorgularının veritabanı üzerinden doğru bir şekilde yürütüldüğüne işaret eder. Öte yandan, yüksek GQA ve EM skorları, çıktı sorgularının altın standart sorgularla veya cevaplarla uyumlu olduğunu ve doğru erişim sonuçlarının elde edildiğini göstermektedir [240]

5.4. Sonuçlar ve Tartışma

Olay Örgüsü Bilgi Grafiğinin Çok Adımlı Soru Cevaplama Performansına Etkisi:

Haber metinlerindeki olayları, ayrık olaylardan ziyade bir hikayenin parçası olarak gören ve sadece QA problemi için değil aynı zamanda olayların gelişimini analiz etmek için de bilgi erişimine izin veren, Bölüm 4’de ayrıntıları sunulan olay çizge modeli üzerine kurgulanmış olan Olay Örgüsü Bilgi Grafiği’nin çok adımlı soru cevaplama problemi üzerindeki etkinliği analiz edilmiştir. Olay çizgesi, hikaye tabanlı kümeleme yeteneğini artırarak, aynı hikayenin farklı zaman dilimlerinde yayınlanmış farklı içerikteki olaylarının yüksek doğruluk oranı ile belirlenmesine olanak tanır. Daha önceki çalışmada önerilen yöntem ile aynı hikayeye ait olduğu tespit edilen olaylar [2], yayınlanma zamanı dikkate alınarak ve Olay Çizge yapısındaki hiperdüğüm olarak yer alan haber kimlik numarası düğümleri (NEWS) arasına (NEXT EVENT) kenarlarının eklenmesi

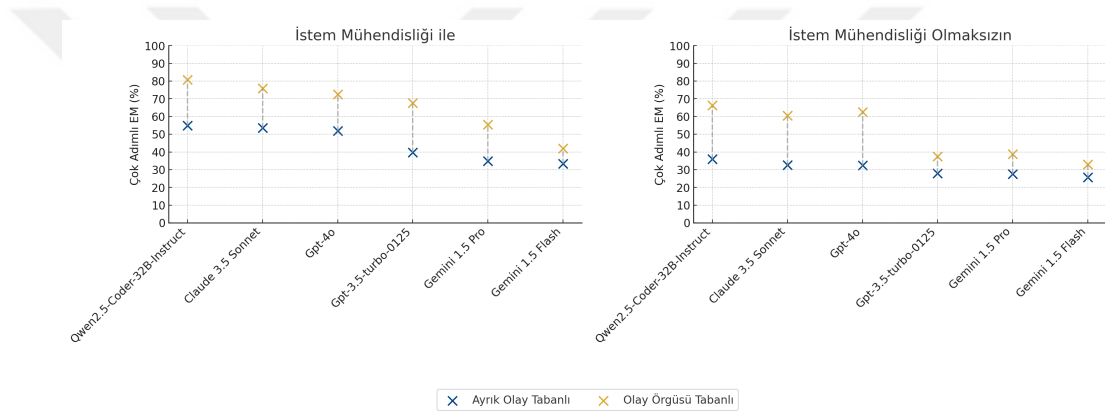
Tablo 5.6: MuSiQue-Ans Veri Kümesindeki Muhakeme Grafikleri [3]

Grafik	Soru	Ayrıştırma
	Namibya'nın ilk Cumhurbaşkanı'nın halefi kimdir? Hifikepunye Pohamba	S1. Namibya'nın ilk Cumhurbaşkanı kimdir? C: Sam Nujoma S2. Sam Nujoma 'nın halefi kimdir? C: Hifikepunye Pohamba
	Billy Giles'in öldüğü yerde kullanılan para birimi nedir? pound sterling	S1. Billy Giles nerede öldü? C: Belfast S2. Belfast , Birleşik Krallık'ın hangi bölgesinde? C: Kuzey İrlanda S3. Kuzey İrlanda'nın para birimi nedir? C: pound sterling
	McDonaldization'ın adını aldığı ilk kuruluş, Horn-dean'ın bulunduğu ülkede ne zaman açıldı? 1974	S1. McDonaldization neyin adını almıştır? C: McDonald's S2. Horn-dean hangi ülkededir? C: İngiltere S3. İlk McDonald's İngiltere'de ne zaman açıldı? C: 1974
	Napolyon, Louis XVI stilini mahkemeye getiren kadının annesinin öldüğü şehri ne zaman işgal etti? 1805	S1. Louis XVI stilini mahkemeye kim getirdi? C: Marie Antoinette S2. Marie Antoinette 'in annesi kimdir? C: Maria Theresa S3. Maria Theresa hangi şehirde öldü? C: Viyana S4. Napolyon Viyana 'yı ne zaman işgal etti? C: 1805
	Prazeres'in ülkesinin yönettiği Aruba'nın kıtasındaki sömürge kaç Alman yaşıyor? 5 milyon	S1. Aruba hangi kıtadadır? C: Güney Amerika S2. Prazeres'in ülkesi nedir? Portekiz S3. Güney Amerika'daki Portekiz'in yönettiği sömürge nedir? C: Brezilya S4. Brezilya'da kaç Alman yaşıyor? C: 5 milyon
	Malakoff'u ele geçiren insanlar, Philipsburg'un bulunduğu bölgeye ne zaman geldi? 1625	S1. Philipsburg hangi yerin başkentidir? C: Saint Martin S2. Saint Martin hangi coğrafi bölgede bulunur? C: Karayipler S3. Malakoff'u kim ele geçirdi? C: Fransızlar S4. Fransızlar Karayipler 'e ne zaman geldi? C: 1625

Tablo 5.7: Hikaye Tabanlı KGQA Veri Setinden Örnek Sorular

N-hop	Hikaye Kümesi (TR)	Soru (TR)	Neo4J Cypher Sorgusu	Soru Türü	GT Cevap
1	2022-italya-genel-seçimleri	Giorgia Meloni'nin politik ideolojisi nedir?	MATCH (p1:Entity name: "giorgia meloni")-[:political_ideology]->(p2:Entity) RETURN p2.name	Ne(What)	neo-fascist
1	kur-korumali-tl-vadeli-mevduat	Döviz korumalı mevduat (kkm) hesabı uygulaması ne zaman başlamıştır?	MATCH (account:Entity name: 'currency protected deposit') MATCH (account)-[:start_time]->(start) RETURN start.name AS start_date	Ne zaman(When)	23 march 2022
2	2022-kirgizistan-tacikistan-catismalari	Rusya'nın hangi sınır komşusu, Hindistan'ın da sınır komşusudur?	MATCH (r:Entity name: "russia")-[:shares_border_with]->(b:Entity)-[:shares_border_with]->(i:Entity name: "india") RETURN b.name	Hangi(Which)	china
2	mahmut-yazicioglu-cinayeti	Ece Erken'in kocası hangi spor kulübünün üyesidir?	MATCH (eee:Entity name: "ece erken")-[:spouse]->(Entity)-[:member_of_sports_team]->(sportsClub:Entity) RETURN sportsClub.name	Hangi(Which)	beşiktaş
3	94-oscar-odulleri	Academy tarafından verilen en iyi film ödülüne aday olan "the power of the dog"un yönetmeni kimdir?	MATCH (:Entity name: 'the power of the dog')-[:director]->(director:Entity)-[:nominated_for]->(Entity name: 'best picture')-[:conferred_by]->(Entity name: 'academy') RETURN director.name	Kim(Who)	jane campion

ile her bir hikaye için Şekil 5.1’da gösterildiği gibi bir ”Olay Örgüsü” çizge yapısının oluşturulması sağlanmıştır. Ayrıca Bölüm 5.2.1’de bahsedilen, bağlantılı çözümleme ve ilişki çıkarımı adımları ile haber metninden elde edilen kenarların, ilgili düğümler ile entegrasyonu oluşturulan ”Olay Örgüsü” bilgi grafiği sayesinde her bir hikaye bütüncül ve yapılandırılmış formda vektör uzayında temsil edilmektedir. Bu çizge, aynı hikayeye ait tüm olaylar ve olay aktörleri, zaman etiketleri, vb. varlıkların zengin ilişkilerini yapısında barındırmaktadır. Olay örgüsü bilgi grafiği üzerindeki yolları takip ederek çok adımlı sorulara cevap vermenin uygulanabilirliği deneylerle test edilmiştir.



Şekil 5.6: Ayırık Olay ve Olay Örgüsü Tabanlı Bilgi Grafiği Yapılarının Çok Adımlı QA Performansına Etkileri

Yapılan testlerde öncelikle her bir olay (haber metni) vektör uzayında olay örgüsünden bağımsız, ayırık olay tabanlı bir bilgi grafiği olarak temsil edilmiştir. Ardından aynı varlıkları içeren olaylara ait ortak varlıkların aynı ID’ye sahip olmaları sağlanmış ve grafikte daha önce görülmeyen bir varlık varsa, bu varlık ”STORY” düğümüne ’CONTAINS ENTITY’ bağlantısı ile bağlanarak her hikaye için bir olay örgüsü bilgi grafiği oluşturulmuştur. Her iki yaklaşım türü için de Hikaye Tabanlı KBQA Veri Kümesi’ndeki soruları yanıtlama performansı ölçülmüştür. Tablo 5.10’de Bölüm 5.2.3’de bahsedilen komut istemi iyileştirmesi uygulanarak ve uygulanmaksızın, ayırık olay tabanlı ve olay örgüsü tabanlı bilgi grafiği yapılarının tek adımlı ve çok adımlı soru cevaplama performansını Tam Eşleşme (EM) değerlendirme metriği açısından göstermektedir.

6 farklı büyük dil modeli ile Hikaye Tabanlı KBQA Veri Kümesi üzerinde uygulanan testlerde en yüksek başarımlar Qwen2,5-Coder-32B-Instruct modeli ile elde edilirken, en düşük başarımlar ise Gemini 1,5 Flash modeline aittir. Ayrıca deney sonuçlarından ve Şekil 5.6’da açıkça görüldüğü üzere, tüm modellerde Olay Örgüsü çizge yapısı çok adımlı soruların yanıtlanabilirliği üzerinde %27,89’a varan (EM metriği cinsinden) önemli bir iyileşme sağlamıştır. İki varlık arasındaki yalnızca tek bir bağlantı ile cevaplanabilen tek adımlı sorular için çizge yapısının anlamlı bir etkisi olmadığı görülmüştür.

Tablo 5.8: MuSiQue-Ans (validation-mini) Veri Kümesi üzerinde Büyük Dil Modelleri ile Çok adımlı QA Performansı

Model	İstem Mühendisliği ile		İstem Mühendisliği Uygulanmaksızın	
	EM	ER	EM	ER
Gpt-3,5-turbo-0125	0,51	0,68	0,38	0,63
Gpt-4o	0,53	0,70	0,42	0,63
Claude 3,5 Sonnet	0,54	0,71	0,42	0,65
Qwen2,5-Coder-32B-Instruct	0,56	0,72	0,45	0,66

Musique-Ans veri kümesi bünyesinde, literatürdeki diğer çok adımlı soru cevaplama veri kümelerinde olduğu gibi çoklu doküman (paragraf, belge,vb.) üzerinde çok adımlı muhakeme yeteneği gerektiren sorular barındırır. Dolayısı ile bu veri kümelerinde ayrı ayrı temsil edilen referans dokümanlar üzerinde çok adımlı bir soruyu cevaplayabilmek mümkün değildir. Bu sebeple, aynı deney diğer veri kümeleri üzerinde uygulanamamıştır. Ancak, Musique-Ans veri kümesinde, her bir soru için referans paragraflar listesi sunulmaktadır ve bu paragrafların soruyu yanıtlamak üzere destekleyici bilgi barındırıp barındırmadığı bilgisi bulunmaktadır. Buradan hareketle Musique-Ans validation-mini veri kümesi üzerindeki her bir soru için sunulan bu paragraf metnlerinin, birer ’olay’ gibi ele alınması ile her soru için üretilen ’Olay Örgüsü’ bilgi grafiği üzerinden 4 farklı büyük dil modeli ile çok adımlı soru cevaplama performansı ölçülmüştür. EM ve ER metrikleri ile elde edilen sonuçlar 5.8’de sunulmuştur. Sunulan sonuçlarda daha önceki deney sonuçlarına benzer şekilde, en yüksek sonuç Qwen2,5-Coder-32B-

Instruct dil modeli üzerinde, bilgi tabanı üzerindeki soru ile ilgili üçlülerin komut istemine eklenmesi istem mühendisliği yöntemi ile elde edilmiştir. Elde edilen en iyi sonuç (Qwen2,5-Coder-32B-Instruct + İM), son teknoloji soru cevaplama performansları ile Tablo 5.9’de karşılaştırılmıştır.

MuSiQue-Ans veri kümesi üzerinde çalışmalar yürüten daha önceki araştırmacıların literatüre kazandırdığı temel modeller ve liderlik tablosu modellerinin, tez çalışması kapsamında kullanılan mini-validation versiyonu üzerinde elde edilen başarımlar EM başarıım metriği cinsinden Tablo 5.9 gibidir. Bu sonuçlar daha önce aynı veri kümesi üzerinde çalışılmış olan [239] araştırmasından ve Aralık 2024 itibari ile veri kümesini sunan araştırmacılar tarafından sunulan liderlik tablosundan [242] faydalanılarak derlenmiştir.

Tablo 5.9: MuSiQue-Ans validation-mini Veri Kümesi Üzerinde Temel Soru Cevaplama Performansları

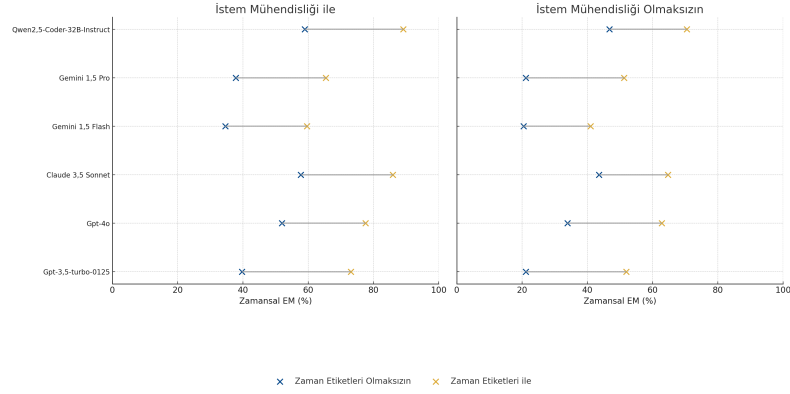
Model	EM
Question decomp.+ dense retrieval + CTE [239]	0,654
Step Execution by Select+Answer (EX(SA)) [3]	0,62
Qwen2,5-Coder-32B-Instruct + İM	0,56
Step Execution by End2End (EX(Ee)) [3]	0,53
Question decomp.+ dense retrieval [239]	0,447
Select+Answer (SA) [3]	0,43
End2End (EE) [3]	0,29

End2End (EE) Modeli, bir soru ve ilgili bağlamı giriş olarak alarak, bunları Longformer-Large transformer modeli [243] aracılığıyla işleyerek doğrudan cevap ve destekleyici paragrafları çıkarır. Bu model, özellikle uzun metinleri etkili bir şekilde işleme konusunda başarılıdır ve uçtan uca bir çözüm sunar. Select+Answer (SA) Modeli ise, ilk aşamada bağlamdaki en alakalı paragrafları sıralamak ve seçmek için RoBERTa-large modelini [244] kullanan ve daha sonra seçilen paragraflardan Longformer-Large modeli yardımı ile cevap tahmini yapar. Step Execution (EX) Modeli, çok atlamalı bir soruyu döngüsel

Tablo 5.10: Ayrık Olay ve Olay Örgüsü Tabanlı Bilgi Grafiği Yapılarının Çok Adımlı QA Performansı Karşılaştırması

Model	İstem Mühendisliği ile				İstem Mühendisliği Uygulanmaksızın			
	Ayrık Olay Tabanlı EM (%)		Olay Örgüsü Tabanlı EM (%)		Ayrık Olay Tabanlı EM (%)		Olay Örgüsü Tabanlı EM (%)	
	Tek Adımlı	Çok Adımlı	Tek Adımlı	Çok Adımlı	Tek Adımlı	Çok Adımlı	Tek Adımlı	Çok Adımlı
Gpt-3,5-turbo-0125	68,57	39,74	68,57	67,63	45,49	27,88	45,49	37,50
Gpt-4o	76,92	51,92	77,14	72,44	63,96	32,37	64,40	62,50
Claude 3,5 Sonnet	81,10	53,53	81,10	75,96	67,91	32,69	67,91	60,58
Gemini 1,5 Flash	65,71	33,33	65,71	41,99	39,12	25,64	39,78	33,01
Gemini 1,5 Pro	68,13	34,94	67,69	55,45	41,76	27,56	42,42	38,78
Qwen2,5-Coder-32B-Instruct	82,42	54,81	82,86	80,77	71,21	35,90	71,21	66,35

olmayan bir grafik (DAG) kullanarak modeller ve BART-large modeli ile tek atlamalı sorulara bölerek adım adım bir akıl yürütme süreci gerçekleştirir. EX modeli, her adımda tek atlamalı akıl yürütme için EE modelini kullanan EX(EE) ve her adımda SA modelini kullanarak daha odaklı bir analiz sağlayan EX(SA) olmak üzere iki farklı varyasyona sahiptir [3]. Çok adımlı soru cevaplama probleminde, küçük ve orta ölçekli büyük dil modellerinin kullanımını ve bilgi grafiklerinin LLM’lerin çok adımlı soru yanıtlama yeteneklerini geliştirdiği konusunu araştıran [239] çalışmasında, soru ayrıştırma kullanan modellerin farklı geri getirme yöntemleri ve komut yönlendirme teknikleri ile yüksek başarımlar elde edebileceğini göstermektedir. Bu amaçla yaptığı deneylerde, soru ayrıştırma ve yoğun geri getirmeyi bir arada kullanan yaklaşımda MuSiQue-Ans validation-mini Veri Kümesi’nde %44,7 EM başarıma ulaşırken; soru ayrıştırma, yoğun geri getirme ve CTE istem mühendisliği yaklaşımı ile %65,4 EM başarımına ulaşmıştır. Son literatürde yer alan bu çalışmalar, ilgili paragrafların tespiti veya soru ayrıştırma gibi ek modüllerle Tablo 5.9’de sunulan başarımları elde etmiştir. Bu çalışma kapsamında, sorular yalnızca veri kümesinde yer alan soru ile ilgili olabilecek paragraflardan oluşturulan bilgi grafiği üzerinden, doğrudan büyük dil modelleri komut istemi ile yapılandırılmış Neo4J Cypher Query ile yanıtlanmıştır. Buna rağmen, önerilen olay örgüsü çizge modeli ve istem mühendisliği yaklaşımı ile yapılan diğer çalışmalarla rekabetçi bir performans elde edilmiştir.



Şekil 5.7: Hikaye Tabanlı KBQA Veri Kümesi üzerinde Zaman Etiketli Düşümlerinin Zamansal Soru Cevaplama Performansına Etkisi

Açık/Örtük Zaman Etiketli Düşümlerinin Zamansal Soru Cevaplama Performansına Etkisi: Zamansal sorular, istatistiksel olarak kullanıcılar tarafından en fazla yönlendirilen soru türüdür. Örtük ve açık tüm zamansal bilgilerin ilişkisel bir biçimde bilgi grafiğine dahil edilmesinin soru yanıtlama sisteminin performansını önemli ölçüde etkilediğini savunuyoruz. Çalıştığımız alanın doğası gereği, olayların zaman içindeki dönüşüm ve değişimini gözlemlemek ancak zamansal unsurların bilgi grafiğinde sistematik ve somut bir şekilde temsil edilmesiyle sağlanabilir.

Tablo 5.11: Hikaye Tabanlı KBQA Veri Kümesi Zaman Etiketli Düşümlerinin Zamansal Soru Cevaplama Performansına Etkisi

Model	İstem Mühendisliği ile		İstem Mühendisliği Uygulanmaksızın	
	Zaman Etiketleri Olmaksızın EM (%)	Zaman Etiketleri ile EM (%)	Zaman Etiketleri Olmaksızın EM (%)	Zaman Etiketleri ile EM (%)
Gpt-3,5-turbo-0125	39,74	73,08	21,15	51,92
Gpt-4o	51,92	77,56	33,97	62,82
Claude 3,5 Sonnet	57,69	85,90	43,59	64,74
Gemini 1,5 Flash	34,62	59,62	20,51	41,03
Gemini 1,5 Pro	37,82	65,38	21,15	51,28
Qwen2,5-Coder-32B-Instruct	58,97	89,10	46,79	70,51

Bu hipotezi analiz etmek için bilgi grafiğini iki şekilde oluşturuyoruz: HEIDEL Time mimarisi aracılığıyla işlenen ve zaman etiketleri ile düzenlenmiş metinlerle ve işlenmemiş metinlerle. Bilgi grafiği oluşturma boru hattının ilişki çıkarma adımında,

Hikaye Tabanlı KBQA Veri Kümesi için zamansal etiketler tespit edilmeksizin “ölüm tarihi” ve “doğum tarihi” gibi zamansal ilişkilerin sayısı 956’ken, zamansal etiketlerin belirlenmesi ve haber metinlerinin güncellenmesi ile bu sayı 2172’ye kadar yükselmiştir. Bu önemli artış, zamansal referanslar içeren soruların cevaplanabilirliğine de yansımıştır. Tespit edilen örtük ve açık zamansal referansların belirlenen formatta metne eklenmesiyle zamansal ilişki çıkarımının performansı önemli ölçüde arttığı Tablo 5.11’de açıkça görülmektedir. Hikaye Tabanlı KBQA Veri Kümesi’nde yer alan ve soru dağılımının %20,34’sini oluşturan zamansal soruların komut istemi mühendisliği olmaksızın en fazla %46,79’u cevaplanabilirken, zaman etiketleri entegrasyonu ile %70,51’e kadar yükselmiştir. Tüm dil modelleri için artışın dikkat çekici olduğu Şekil 5.7’de görülmektedir.

Tablo 5.12: MuSiQue-Ans (validation-mini) Veri Kümesi Zaman Etiketli Düşümlerinin Zamansal Soru Cevaplama Performansına Etkisi

Model	İstem Mühendisliği ile		İstem Mühendisliği Uygulanmaksızın	
	Zaman Etiketleri Olmaksızın EM (%)	Zaman Etiketleri ile EM (%)	Zaman Etiketleri Olmaksızın EM (%)	Zaman Etiketleri ile EM (%)
Gpt-3,5-turbo-0125	25,00	37,50	12,50	25,00
Gpt-4o	25,00	37,50	25,00	25,00
Claude 3,5 Sonnet	25,00	50,00	25,00	37,50
Qwen2,5-Coder-32B-Instruct	37,50	50,00	25,00	37,50

Benzer şekilde MuSiQue-Ans validation-mini veri kümesindeki paragrafların bilgi grafiği oluşturma hattından geçirilmesi sonucu oluşturulan bilgi tabanında; Heidel Time zaman etiketleyicisi uygulanmaksızın doğrudan ilişki çıkarımı ile elde edilen zamansal ilişkilerin sayısı 124’ken, bu sayının etiketleyici kullanılarak 213’e kadar arttığı görülmektedir. Ancak Hikaye Tabanlı KBQA Veri Kümesi’nde %20,54 oranında yer alan zamansal sorular, bu veri kümesinde yalnızca %8 oranında bulunmaktadır. Zamansal soruların düşük yoğunluğuna karşın gösterilen iyileşmeler en yüksek başarıma sahip 4 dil modeli için Tablo 5.12’de sunulmuştur. Zaman etiketlerinin zamansal soru performansına olumlu etkisi, tabloda verilen deney sonuçlarından görülmektedir. Ancak veri kümesinde yer alan soru sayısı bir kısıtlılık oluşturmaktadır.

Tablo 5.13: MuSiQue-Ans Doğrulama Veri Kümesi Zaman Etiketli Düşümlerinin Zamansal Soru Cevaplama Performansına Etkisi

Model	İstem Mühendisliği ile		İstem Mühendisliği Uygulanmaksızın	
	Zaman Etiketleri Olmaksızın EM (%)	Zaman Etiketleri ile EM (%)	Zaman Etiketleri Olmaksızın EM (%)	Zaman Etiketleri ile EM (%)
Gpt-3,5-turbo-0125	23,90	33,46	20,08	27,34
Gpt-4o	29,25	35,95	23,52	29,25
Claude 3,5 Sonnet	31,17	39,01	25,24	31,36
Qwen2,5-Coder-32B-Instruct	32,12	40,92	26,39	33,65

Bu kısıtlılığı aşmak için MuSiQue-Ans veri kümesindeki doğrulama kümesindeki zamansal sorular tespit edilerek, aynı deney tekrarlanmıştır. MuSiQue-Ans doğrulama kümesinde yer alan 2417 sorunun 523 tanesi zamansal sorulardan oluşmaktadır. Bu sorulara ilişkin detaylı istatistikler Tablo 5.14’de sunulmuştur. Heidelberg zaman etiketleyicisi metin içerisinde bulunan tarih bilgilerini referans kabul ederek, örtük bilgileri tam tarih formatına dönüştürmeyi hedefler. Metin içerisinde referans tarihlendirmeyi sağlayacak zamansal bilginin yer almaması durumu örtük zaman etiketlerinin elde edilememesi ile sonuçlanmaktadır. Bu kısıtlılığı aşmak için Hikaye Tabanlı KBQA Veri Kümesi’nde yer alan haberlerin yayınlanma tarihi referans tarih bilgisi olarak eklenmiştir. Ancak MuSiQue-Ans veri kümesinde benzer bir bilgi yer almaması zamansal etiketlerin elde edilmesini sınırlandıran bir unsurdur. Bu sınırlılıklara ve veri kümesinin doğası gereği daha karmaşık bir çıkarım ihtiyacı rağmen, zamansal etiketlerin soru cevaplama başarımı üzerine olumlu etkisi oldukça belirgindir. İstem mühendisliği uygulanması durumunda yaklaşık %10’lara varan zamansal soru cevaplama performans artışı görülmektedir.

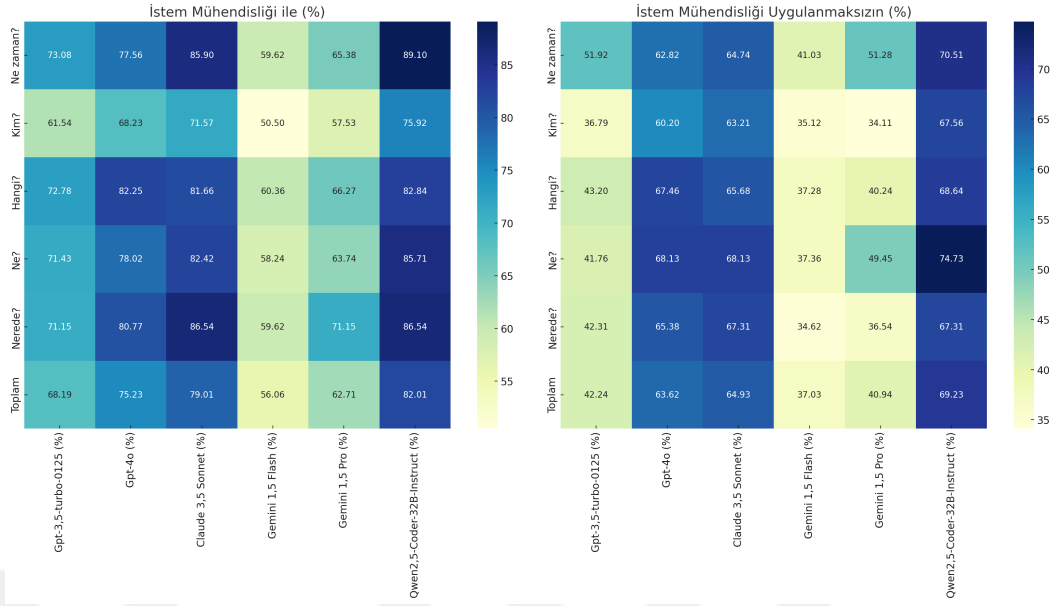
Tablo 5.14: MuSiQue-Ans Doğrulama Kümesi Zamansal Soru İstatistikleri

	2-adımlı	3-adımlı	4-adımlı	Toplam
Doğrulama Kümesi Soru Sayısı	1252	760	405	2417
Doğrulama Kümesi Zamansal Soru Sayısı	183	225	115	523
Yüzde (%)	14,62	29,61	28,40	21,64

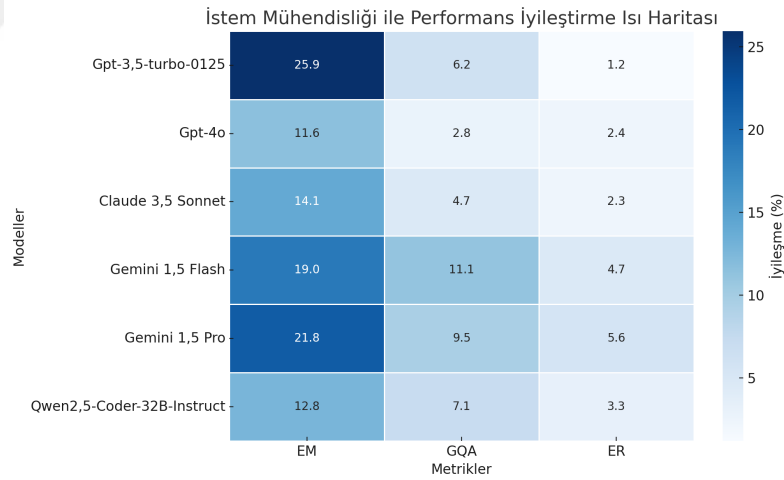
MuSiQue-Ans veri kümesi, doğrudan metin içerisinde çıkarım yapılabilecek bilgilerle yanıtlanabilir sorular içerdiği gibi, mantıksal akıl yürütme seviyesi oldukça yüksek soruları da kapsamaktadır. Zamansal sorular içerisinde, ”önce-sonra” çıkarımlarının yapılması, kişi (doğum tarihi,ölüm tarihi) formatında verilen metinden bu zaman bilgilerinin ölüm yada doğum eylemiyle ilişkili olduğunun anlaşılması, vb. kompleks ilişkilerin elde edilmesini gerektiren örnekler mevcuttur. Bu örnekler görevi zorlaştırmakta ve yalnızca zaman etiketleyicisi ile elde edilen örtük bilgilerle sağlanan iyileşmeyi sınırlandırmaktadır. Bu sebeple, Hikaye Tabanlı KBQA Veri Kümesi üzerinde görülen iyileşme oranı, MuSiQue-Ans doğrulama veri kümesine oranla oldukça yüksektir.

İstem Mühendisliğinin Soru Cevaplama Performansı Üzerindeki Etkisi: Geliştirilen KBQA sisteminde, bilgi grafiği şemasını tanımlayan ve ardından verilen doğal dil sorusu için bir Neo4j Cypher Sorgusu üreten bir komut istemi oluşturulmuştur. Üretilen sorgunun anlamsal ve sözdizimsel olarak doğru olmasını sağlamak, doğru cevaplar elde etmek için kritiktir. Bu nedenle, sorguyu iyileştirmek için yol gösterici teknikler kullanılmıştır. Sorudan elde edilen ”baş (head)” varlıklara dayalı bilgi grafiğindeki üçlüler arasında filtreleme yapılarak, soruyla ilgili en sık tekrar eden (en fazla üç) üçlüler istemdeki modele ek bilgi olarak sağlanmıştır. Bu yaklaşım, geliştirilen sorgu ile soru yanıtlama performansında farklı metriklerle göre önemli bir iyileşmeye yol açmıştır. Hikaye Tabanlı KBQA Veri Kümesi üzerinde yapılan deneylere ilişkin sonuçlar Şekil 5.9’de 1s1 haritası ile sunulmuştur. Ayrıca soru türüne ve karmaşıklığına bağlı olarak soru cevaplama görevindeki soru sayısı ve doğruluk değerleri Şekil 5.8’te “tam eşleşme oranı” metriği ile elde edilen skorlarla karşılaştırmalı olarak gösterilmiştir.

Önerilen istem mühendisliği yaklaşımı, dil modellerinin soru cevaplama performansında EM metriği ile %25.9’a varan iyileşme sağlarken, bu iyileşme GQA metriği açısından %11.1, ER metriği açısından %5.6 olmuştur. ”Ne?” ve ”Nerede?” gibi bilgi tabanlı sorular, tüm modellerde daha yüksek performansa sahiptir. ”Kim?” sorusu gibi bağlamsal anlayış gerektiren sorularda ise performans genelde düşüktür. Bu soru türü, istem mühendisliği uygulandığında bile nispeten zorlayıcı bir alan olarak kalmaktadır. Hem istem mühendisliği ile hem de istem mühendisliği olmadan açık ara liderlik eden Qwen2,5-Coder-32B-Instruct ve Claude 3,5 Sonnet modelleri, istem mühendisliği ile performanslarını da artırarak diğer modellerden belirgin oranda sıyrılmaktadır. Bu artış, özellikle ”Kim?” ve ”Nerede?” sorularında daha belirgindir.



Şekil 5.8: İstem Mühendisliğinin Soru Cevaplama Performansı Üzerindeki Etkisi



Şekil 5.9: İstem Mühendisliği ile Soru Cevaplama Performans İyileştirme Isı Haritası

MuSiQue-Ans validation-mini Veri Kümesi üzerinde, çalışmada önerdiğimiz istem mühendisliği yaklaşımı ile elde ettiğimiz sonuçlar EM metriği cinsinden Tablo 5.8’de gösterilmiştir. Bu sonuçlar sorulardan elde edilen üçlüler ile elde edilen ve büyük dil

modeli istemine ek bilgi olarak sunulan ilgili üçlülerin soru cevaplama performansını artırdığı açıkça ortaya koymaktadır.

Hikaye Tabanlı KBQA Veri Kümesi üzerinde yapılan deneylerin sonuçlarında, dil modellerinin istem mühendisliği kullanarak veya kullanmayarak %68,21 ile %91,13'e varan oldukça yüksek yürütme doğruluğuna (EM) sahip olduğu görülmektedir. Diğer taraftan yine dil modellerinin çıktı sorgusu ile altın sorgu arasındaki benzerliği BERTScore ile ölçen altın sorgu doğruluğu (GQA) metriğine göre %89,22'e varan başarıma sahip olduğu görülmektedir. Yapılandırılmış Neo4J sorgusunun, veritabanı üzerinde çalıştırılabilir formata uygun ve altın sorgu benzerliği oldukça yüksek sorgular oluşturabiliyor olması umut vericidir.

Tablo 5.15: Hikaye Tabanlı KBQA Veri Kümesi Üzerinde Komut İstemi Mühendisliğinin Büyük Dil Modellerinin QA Performansına Etkisi

Model	İstem Mühendisliği ile (%)			İstem Mühendisliği Uygulanmaksızın (%)		
	EM	GQA	ER	EM	GQA	ER
Gpt-3,5-turbo-0125	68,19	77,31	81,23	42,24	71,08	80,05
Gpt-4o	75,23	82,18	86,31	63,62	79,43	83,96
Claude 3,5 Sonnet	79,01	85,71	87,74	64,93	80,98	85,40
Gemini 1,5 Flash	56,06	69,23	73,01	37,03	58,18	68,32
Gemini 1,5 Pro	62,71	71,61	75,62	40,94	62,09	70,01
Qwen2,5-Coder-32B-Instruct	82,01	89,22	91,13	69,23	82,08	87,87

5.5. Kısıtlar

Çalışmada tasarlanan semantik ayrıştırma tabanlı bilgi grafiği soru cevaplama sistemi önemli yenilikler sunmasına rağmen, çeşitli sınırlılıklar içermektedir. Öncelikle, çalışmada önerilen hipotezleri test etmek ve sistemin etkinliğini geniş bir yelpazede değerlendirmek için uygun bir değerlendirme veri kümesinin bulunmaması, en büyük kısıtlardan biridir. Özel bir benchmark eksikliği, sistemin doğruluğunu farklı veri türlerinde ve senaryolarda test etme yeteneğini sınırlamaktadır. Bunun yanı sıra, doğrudan Türkçe gibi düşük kaynaklı diller için eğitilmiş varlık bağlama ve ilişki çıkarımı modellerinin eksikliği, bilgi grafiği oluşturma süreçlerini daha zor hale getirmektedir. Mevcut model-

lerin çoğunun yüksek kaynaklı diller için optimize edilmiş olması, Türkçe gibi dillerde performans kayıplarına yol açmakta ve bilgi grafiği oluşturma sürecini daha manuel ve zaman alıcı bir hale getirmektedir.

Bir diğer sınırlılık, önerilen sistemin etkinliği büyük ölçüde büyük dil modellerinin mevcut performansına bağımlı olmasıdır. Ancak, büyük dil modelleri genellikle yüksek hesaplama gücü gerektirir ve bu durum, hem donanım kaynakları açısından hem de pratik uygulamalar için erişilebilirlik açısından kısıtlamalar oluşturabilir. Ayrıca, bazı açık kaynaklı olmayan modellerin eğitimi ve özelleştirilmesi için erişim kısıtlamaları, sistemin esnekliğini azaltmaktadır. Bu sınırlamalar, çalışmanın kapsamını ve uygulama alanını daraltmakla birlikte, gelecekteki araştırmalarda ele alınabilecek iyileştirme ve genişletme alanlarına işaret etmektedir. Özellikle, daha kapsamlı değerlendirme veri kümeleri oluşturulması, düşük kaynaklı diller için özel olarak optimize edilmiş modellerin geliştirilmesi ve sistemin farklı soru türlerine uyarlanabilirliğinin artırılması, bu alandaki çalışmaları daha ileriye taşıyacaktır.

6. SONUÇLAR VE GELECEK ÇALIŞMALAR

Doktora çalışması kapsamında, büyük ölçekli ve bağlamsal olarak zengin veri kaynaklarından olay tabanlı bilgi çıkarımı ve bu bilgilerin çok atlamalı soru cevaplama problemlerinde etkin bir şekilde kullanılmasını sağlamak amacıyla özgün bir olay çizgesi modeli geliştirilmiştir. Çalışma, haber metinlerindeki olayların bağlamsal ve zamansal ilişkilerini dikkate alarak, yalnızca bağımsız bilgi birimleri yerine olay örgüsü zincirleri oluşturulmasını sağlamıştır. Zaman bilgisi entegrasyonu, isimlendirilmiş varlık odaklı alt çizge madenciliği ve kümelenmiş metin grupları üzerinden bilgi çıkarımı gibi yaklaşımlar, özellikle çok adımlı soruların yanıtlanmasında sistemin doğruluğunu ve verimliliğini artırmıştır. Ayrıca, Türkçe doğal dil işleme alanına katkı sağlamak amacıyla ilk kez bilgi tabanlı soru cevaplama ve olay odaklı metin kümeleme için uygun veri kümeleri literatüre kazandırılmıştır. Tüm bunların yanı sıra semantik ayrıştırma tabanlı bilgi grafiği soru cevaplama araştırmalarında, büyük dil modellerinin etkinliği de ele alınmıştır. Büyük dil modelleri için kritik öneme sahip komut isteminin bilgi grafiğinden elde edilen ve sorularla ilintili bilgilerle iyileştirmesi önerisi ile, soru cevaplama performansını önemli ölçüde iyileştiren bir komut istemi mühendisliği yaklaşımı sunulmuştur.

Türkçe doğal dil işleme alanında değerlendirilebilir standartlara uygun veri kümelerinin geliştirilmesi, bu dildeki araştırmalar için önemli bir temel oluşturmuştur. Ancak, gelecekte bu veri kümelerinin genişletilmesi ve İngilizce’de mevcut olan değerlendirme veri kümeleriyle benzer örneklem büyüklüğüne ulaşması hedeflenmelidir. Bunun yanı sıra, olay örgüsü yapısının daha sağlam ve zengin ilişkilerle desteklenmesi için ilişki çıkarımı modellerinin geliştirilmesi önem taşımaktadır. Bu kapsamda, ileri düzey grafik sinir ağlarının kullanımı değerlendirilebilir. Büyük dil modellerinin performans ve maliyet açısından daha kapsamlı ve çeşitli veri kümeleriyle test edilmesi, gelecek çalışmalara ışık tutabilir. Ayrıca, veri kümesi çeşitliliğinin artırılmasıyla farklı bağlamlarda modelin genellenebilirliğinin incelenmesi ve sistemin dinamik bilgi güncellemelerine uyumlu hale getirilmesi, bu alanda önemli ilerlemeler sağlayacaktır.

KAYNAKLAR

- [1] Abdel-Nabi H., Awajan A., Ali M. Z. (2023), “Deep learning-based question answering: a survey”, *Knowledge and Information Systems*, 65(4), 1399–1485.
- [2] Kömeçoğlu B. B., Yılmaz B. (2024), “Event graph-based news clustering: The role of named entity-centered subgraphs”, *IEEE Access*.
- [3] Trivedi H., Balasubramanian N., Khot T., Sabharwal A. (2022), “Musique: Multi-hop questions via single-hop question composition”, *Transactions of the Association for Computational Linguistics*, 10, 539–554.
- [4] Nielsen R., Selva M. (2019), “More important, but less robust? five things everybody needs to know about the future of journalism”, *Reuters Institute for the Study of Journalism Report*.
- [5] Komecoglu B. B., Yılmaz B. (2022), Topic detection and tracking towards determining public agenda items: The impact of named entities on event-based news clustering, in “Machine Learning for Societal Improvement, Modernization, and Progress”, IGI Global, 158–179.
- [6] Wang J., Jatowt A., Färber M., Yoshikawa M. (2021), “Improving question answering for event-focused questions in temporal collections of news articles”, *Information Retrieval Journal*, 24, 29–54.
- [7] Bouziane A., Bouchiha D., Doumi N., Malki M. (2015), “Question answering systems: survey and trends”, *Procedia Computer Science*, 73, 366–375.
- [8] Allam A. M. N., Haggag M. H. (2012), “The question answering systems: A survey”, *International Journal of Research and Reviews in Information Sciences (IJRRIS)*.
- [9] Soares M. A. C., Parreiras F. S. (2020), “A literature review on question answering techniques, paradigms and systems”, *Journal of King Saud University-Computer and Information Sciences*, 32(6), 635–646.

- [10] Mavi V., Jangra A., Jatowt A., others (2024), “Multi-hop question answering”, *Foundations and Trends® in Information Retrieval*, 17(5), 457–586.
- [11] Kömeçoğlu Y., Kömeçoğlu B. B., Yılmaz B. (2021), Real-time news grouping: Detecting the same-content news on turkish news stream, in “Progress in Intelligent Decision Science: Proceeding of IDS 2020”, Springer, 14–25.
- [12] Kömeçoğlu B. B., Yılmaz B. (2024), Knowledge-augmented large language model prompting for cypher query construction, in “2024 9th International Conference on Computer Science and Engineering (UBMK)”, IEEE, 187–192.
- [13] Jurafsky D., Martin J. H. (2024), “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition with Language Models”, 3rd Edition, Prentice Hall PTR. Online manuscript released August 20, 2024.
- [14] Woods W. A. (1973), Progress in natural language understanding: an application to lunar geology, in “Proceedings of the National Computer Conference and Exposition”, 441–450.
- [15] Jr B. F. G., Wolf A. K., Wolf C., Laughery K. (1961), Baseball: an automatic question-answerer, in “Proceedings of the IRE-AIEE-ACM’61 (Western)”, 219–224.
- [16] Mishra A., Jain S. K. (2016), “A survey on question answering systems with classification”, *Journal of King Saud University-Computer and Information Sciences*, 28(3), 345–361.
- [17] Caballero M. (2021), “A brief survey of question answering systems”, *International Journal of Artificial Intelligence & Applications (IJAIA)*.
- [18] Baradaran R., Ghiasi R., Amirkhani H. (2022), “A survey on machine reading comprehension systems”, *Natural Language Engineering*, 28(6), 683–732.
- [19] Qiu L., Xiao Y., Qu Y., Zhou H., Li L., Zhang W., Yu Y. (2019), Dynamically fused graph network for multi-hop reasoning, in “Proceedings of the 57th annual

meeting of the association for computational linguistics”, 6140–6150.

- [20] Wang J., Jatowt A., Färber M., Yoshikawa M. (2020), Answering event-related questions over long-term news article archives, in “Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42”, Springer, 774–789.
- [21] Chen D., Fisch A., Weston J., Bordes A. (2017), “Reading wikipedia to answer open-domain questions”, arXiv preprint arXiv:1704.00051.
- [22] Qu Y., Ding Y., Liu J., Liu K., Ren R., Zhao W. X., Dong D., Wu H., Wang H. (2020), “Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering”, arXiv preprint arXiv:2010.08191.
- [23] Yang W., Xie Y., Lin A., Li X., Tan L., Xiong K., Li M., Lin J. (2019), “End-to-end open-domain question answering with bertserini”, arXiv preprint arXiv:1902.01718.
- [24] Kenton J. D. M.-W. C., Toutanova L. K. (2019), Bert: Pre-training of deep bidirectional transformers for language understanding, in “Proceedings of naacL-HLT”, Vol. 1, Minneapolis, Minnesota, 2.
- [25] Guu K., Lee K., Tung Z., Pasupat P., Chang M. (2020), Retrieval augmented language model pre-training, in “International conference on machine learning”, PMLR, 3929–3938.
- [26] Karpukhin V., Oğuz B., Min S., Lewis P., Wu L., Edunov S., Chen D., Yih W.-t. (2020), “Dense passage retrieval for open-domain question answering”, arXiv preprint arXiv:2004.04906.
- [27] Wang S., Yu M., Guo X., Wang Z., Klinger T., Zhang W., Chang S., Tesauro G., Zhou B., Jiang J. (2018), R 3: Reinforced ranker-reader for open-domain question answering, in “Proceedings of the AAAI Conference on Artificial Intelligence”, Vol. 32.
- [28] Ni J., Zhu C., Chen W., McAuley J. (2018), “Learning to attend on essential terms:

An enhanced retriever-reader model for open-domain question answering”, arXiv preprint arXiv:1808.09492.

- [29] Zhang Y., Nie P., Ramamurthy A., Song L. (2021), Answering any-hop open-domain questions with iterative document reranking, in “Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval”, 481–490.
- [30] Wang J. (2022), “A study on effective approaches for exploiting temporal information in news archives”.
- [31] Lin S., Jin P., Zhao X., Yue L. (2014), “Exploiting temporal information in web search”, *Expert Systems with Applications*, 41(2), 331–341.
- [32] Zhang R., Konda Y., Dong A., Kolari P., Chang Y., Zheng Z. (2010), Learning recurrent event queries for web search, in “Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing”, 1129–1139.
- [33] Pasca M. (2008), Towards temporal web search, in “Proceedings of the 2008 ACM symposium on Applied computing”, 1117–1121.
- [34] Harabagiu S., Bejan C. A. (2005), Question answering based on temporal inference, in “Proceedings of the AAAI-2005 workshop on inference for textual question answering”, 27–34.
- [35] Saquete E., Vicedo J. L., Martínez-Barco P., Munoz R., Llorens H. (2009), “Enhancing qa systems with complex temporal question processing capabilities”, *Journal of Artificial Intelligence Research*, 35, 775–811.
- [36] Saquete Boró E., Martinez-Barco P., Munoz R., Vicedo J.-L., others (2004), Splitting complex temporal questions for question answering systems, *Association for Computational Linguistics (ACL)*.
- [37] Alonso O., Gertz M., Baeza-Yates R. (2009), Clustering and exploring search results using timeline constructions, in “Proceedings of the 18th ACM conference on Information and knowledge management”, 97–106.

- [38] Kanhabua N., Nørvåg K. (2010), Determining time of queries for re-ranking search results, in “International conference on theory and practice of digital libraries”, Springer, 261–272.
- [39] Campos R., Dias G., Jorge A. M., Jatowt A. (2014), “Survey of temporal information retrieval and related applications”, *ACM Computing Surveys (CSUR)*, 47(2), 1–41.
- [40] Alonso O., Strötgen J., Baeza-Yates R., Gertz M. (2011), “Temporal information retrieval: Challenges and opportunities.”, *Twaw*, 11, 1–8.
- [41] Allan J., Carbonell J. G., Doddington G., Yamron J., Yang Y. (1998), “Topic detection and tracking pilot study final report”.
- [42] Staykovski T., Barrón-Cedeno A., Da San Martino G., Nakov P., others (2019), Dense vs. sparse representations for news stream clustering, in “CEUR WORKSHOP PROCEEDINGS”, Vol. 2342, CEUR-WS, 47–52.
- [43] Andrews N. O., Fox E. A. (2007), “Recent developments in document clustering”.
- [44] Blokh I., Alexandrov V. (2017), “News clustering based on similarity analysis”, *Procedia computer science*, 122, 715–719.
- [45] Xiao K., Qian Z., Qin B. (2022), “A survey of data representation for multi-modality event detection and evolution”, *Applied Sciences*, 12(4), 2204.
- [46] Li C., Liu M., Cai J., Yu Y., Wang H. (2020), “Topic detection and tracking based on windowed dbscan and parallel knn”, *IEEE Access*, 9, 3858–3870.
- [47] Xu G., Meng Y., Chen Z., Qiu X., Wang C., Yao H. (2019), “Research on topic detection and tracking for online news texts”, *IEEE access*, 7, 58407–58418.
- [48] Vanetik N., Litvak M., Levi E. (2020), “Real-world events discovering with twist”, *Natural Language Processing for Electronic Design Automation*, 71–107.
- [49] Xiao K., Qian Z., Qin B. (2021), “A graphical decomposition and similarity measurement approach for topic detection from online news”, *Information Sciences*,

570, 262–277.

- [50] Lin Y., Miao Z., Ni M., Jiang H., Wang C., Gao J., Lu J., Shi G. (2020), Online topic detection and tracking system and its application on stock market in china, in “Information Retrieval: 26th China Conference, CCIR 2020, Xi’an, China, August 14–16, 2020, Proceedings 26”, Springer, 137–148.
- [51] Bernard G., Suire C., Faucher C., Doucet A., Rosso P. (2022), Tracking news stories in short messages in the era of infodemic, in “International Conference of the Cross-Language Evaluation Forum for European Languages”, Springer, 18–32.
- [52] Mamo N., Azzopardi J., Layfield C. (2021), “An automatic participant detection framework for event tracking on twitter”, *Algorithms*, 14(3), 92.
- [53] Liu W., Jiang L., Wu Y., Tang T., Li W. (2020), “Topic detection and tracking based on event ontology”, *IEEE Access*, 8, 98044–98056.
- [54] Zhang Y., Jin R., Zhou Z.-H. (2010), “Understanding bag-of-words model: a statistical framework”, *International journal of machine learning and cybernetics*, 1, 43–52.
- [55] Miranda S., Znotiņš A., Cohen S. B., Barzdins G. (2018), “Multilingual clustering of streaming news”, *arXiv preprint arXiv:1809.00540*.
- [56] Mikolov T. (2013), “Efficient estimation of word representations in vector space”, *arXiv preprint arXiv:1301.3781*.
- [57] Du Q., Li N., Liu W., Sun D., Yang S., Yue F. (2022), “A topic recognition method of news text based on word embedding enhancement”, *Computational Intelligence and Neuroscience*, 2022(1), 4582480.
- [58] Hu L., Yu S., Wu B., Shao C., Li X. (2020), “A neural model for joint event detection and prediction”, *Neurocomputing*, 407, 376–384.
- [59] Linger M., Hajaiej M. (2020), “Batch clustering for multilingual news streaming”, *arXiv preprint arXiv:2004.08123*.

- [60] Santos J., Mendes A., Miranda S. (2022), “Simplifying multilingual news clustering through projection from a shared space”, arXiv preprint arXiv:2204.13418.
- [61] Gusev I., Smurov I. (2021), “Russian news clustering and headline selection shared task”, arXiv preprint arXiv:2105.00981.
- [62] Hu D., Feng D., Xie Y. (2022), “Egc: A novel event-oriented graph clustering framework for social media text”, *Information Processing & Management*, 59(6), 103059.
- [63] Frisoni G., Moro G., Carlassare G., Carbonaro A. (2021), “Unsupervised event graph representation and similarity learning on biomedical literature”, *Sensors*, 22(1), 3.
- [64] Santos B. N., Rossi R. G., Rezende S. O., Marcacini R. M. (2020), “A two-stage regularization framework for heterogeneous event networks”, *Pattern Recognition Letters*, 138, 490–496.
- [65] Mattos J. P. R., Marcacini R. M. (2021), Semi-supervised graph attention networks for event representation learning, in “2021 IEEE International Conference on Data Mining (ICDM)”, IEEE, 1234–1239.
- [66] Mu L., Jin P., Zhao J., Chen E. (2021), “Detecting evolutionary stages of events on social media: A graph-kernel-based approach”, *Future Generation Computer Systems*, 123, 219–232.
- [67] Wu Z., Pan S., Chen F., Long G., Zhang C., Philip S. Y. (2020), “A comprehensive survey on graph neural networks”, *IEEE transactions on neural networks and learning systems*, 32(1), 4–24.
- [68] Agibetov A. (2023), “Neural graph embeddings as explicit low-rank matrix factorization for link prediction”, *Pattern Recognition*, 133, 108977.
- [69] Qiu J., Dong Y., Ma H., Li J., Wang C., Wang K., Tang J. (2019), Netsmf: Large-scale network embedding as sparse matrix factorization, in “The World Wide Web Conference”, 1509–1520.

- [70] Grover A., Leskovec J. (2016), node2vec: Scalable feature learning for networks, in “Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining”, 855–864.
- [71] Perozzi B., Al-Rfou R., Skiena S. (2014), Deepwalk: Online learning of social representations, in “Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining”, 701–710.
- [72] Chen Y., Wu L., Zaki M. (2020), “Iterative deep graph learning for graph neural networks: Better and robust node embeddings”, *Advances in neural information processing systems*, 33, 19314–19326.
- [73] Chikwendu I. A., Zhang X., Agyemang I. O., Adjei-Mensah I., Chima U. C., Ejiyi C. J. (2023), “A comprehensive survey on deep graph representation learning methods”, *Journal of Artificial Intelligence Research*, 78, 287–356.
- [74] Hamilton W. L. (2020), “Graph representation learning”, Morgan & Claypool Publishers.
- [75] Liu X., Murata T., Kim K.-S., Kotarasu C., Zhuang C. (2019), A general view for network embedding as matrix factorization, in “Proceedings of the Twelfth ACM international conference on web search and data mining”, 375–383.
- [76] Mohammed S. N., Gündüç S. (2022), “Degree-based random walk approach for graph embedding”, *Turkish Journal of Electrical Engineering and Computer Sciences*, 30(5), 1868–1881.
- [77] Ma Y., Zong L., Yang Y., Su J. (2019), News2vec: News network embedding with subnode information, in “Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)”, 4843–4852.
- [78] Zeng Q., Li M., Lai T., Ji H., Bansal M., Tong H. (2021), Gene: Global event network embedding, in “Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)”, 42–53.

- [79] Deng S., Rangwala H., Ning Y. (2019), Learning dynamic context graphs for predicting social events, in “Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining”, 1007–1016.
- [80] Saravanakumar K. K., Ballesteros M., Chandrasekaran M. K., McKeown K. (2021), “Event-driven news stream clustering using entity-aware contextual embeddings”, arXiv preprint arXiv:2101.11059.
- [81] Weng M.-H., Wu S., Dyer M. (2022), “Identification and visualization of key topics in scientific publications with transformer-based language models and document clustering methods”, *Applied Sciences*, 12(21), 11220.
- [82] Dangre N., Bodke A., Date A., Rungta S., Pathak S. (2016), “System for marathi news clustering”, *Procedia Computer Science*, 92, 18–22.
- [83] Bisandu D. B., Prasad R., Liman M. M. (2018), “Clustering news articles using efficient similarity measure and n-grams”, *International Journal of Knowledge Engineering and Data Mining*, 5(4), 333–348.
- [84] Lwin M. T., Aye M. M. (2017), “A modified hierarchical agglomerative approach for efficient document clustering system”, *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, 29(1), 228–238.
- [85] Bouras C., Tsogkas V. (2016), “Assisting cluster coherency via n-grams and clustering as a tool to deal with the new user problem”, *International Journal of Machine Learning and Cybernetics*, 7, 171–184.
- [86] Llewellyn C., Grover C., Oberlander J. (2014), Summarizing newspaper comments, in “Proceedings of the International AAAI Conference on Web and Social Media”, Vol. 8, 599–602.
- [87] Liu X., Gong Y., Xu W., Zhu S. (2002), Document clustering with cluster refinement and model selection capabilities, in “Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval”, 191–198.

- [88] Neal R. M., Hinton G. E. (1998), A view of the em algorithm that justifies incremental, sparse, and other variants, in “Learning in graphical models”, Springer, 355–368.
- [89] Banerjee A., Ghosh J. (2002), Frequency sensitive competitive learning for clustering on high-dimensional hyperspheres, in “Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN’02 (Cat. No. 02CH37290)”, Vol. 2, IEEE, 1590–1595.
- [90] Banerjee A., Dhillon I. S., Ghosh J., Sra S., Ridgeway G. (2005), “Clustering on the unit hypersphere using von mises-fisher distributions.”, *Journal of Machine Learning Research*.
- [91] Kannan R., Vempala S., Vetta A. (2004), “On clusterings: Good, bad and spectral”, *Journal of the ACM (JACM)*, 51(3), 497–515.
- [92] Kummamuru K., Dhawale A., Krishnapuram R. (2003), Fuzzy co-clustering of documents and keywords, in “The 12th IEEE International Conference on Fuzzy Systems, 2003. FUZZ’03.”, Vol. 2, IEEE, 772–777.
- [93] Zhang D., Dong Y. (2004), Semantic, hierarchical, online clustering of web search results, in “Asia-Pacific Web Conference”, Springer, 69–78.
- [94] Ding C., He X., Simon H. D. (2005), On the equivalence of nonnegative matrix factorization and spectral clustering, in “Proceedings of the 2005 SIAM international conference on data mining”, SIAM, 606–610.
- [95] Ding C., Li T., Peng W. (2006), Nmf and plsi: Equivalence and a hybrid algorithm, in “Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval”, 641–642.
- [96] Greene D., Cunningham P. (2005), Producing accurate interpretable clusters from high-dimensional data, in “European conference on principles of data mining and knowledge discovery”, Springer, 486–494.
- [97] Osiński S., Stefanowski J., Weiss D. (2004), Lingo: Search results clustering

algorithm based on singular value decomposition, in “Intelligent Information Processing and Web Mining: Proceedings of the International IIS: IIPWM ‘04 Conference held in Zakopane, Poland, May 17–20, 2004”, Springer, 359–368.

- [98] Zu Eissen S. M., Stein B., Potthast M. (2005), The suffix tree document model revisited, in “Proceedings of the 5th international conference on knowledge management”, 596–603.
- [99] Hammouda K. M., Kamel M. S. (2004), “Efficient phrase-based document indexing for web document clustering”, IEEE Transactions on knowledge and data engineering, 16(10), 1279–1296.
- [100] O’callaghan L., Mishra N., Meyerson A., Guha S., Motwani R. (2002), Streaming-data algorithms for high-quality clustering, in “Proceedings 18th international conference on data engineering”, IEEE, 685–694.
- [101] Babcock B., Datar M., Motwani R., O’Callaghan L. (2003), Maintaining variance and k-medians over data stream windows, in “Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems”, 234–243.
- [102] Gong L., Zeng J., Zhang S. (2011), “Text stream clustering algorithm based on adaptive feature selection”, Expert Systems with Applications, 38(3), 1393–1399.
- [103] Wattanakitrunroj N., Maneeroj S., Lursinsap C. (2018), “Bestream: Batch capturing with elliptic function for one-pass data stream clustering”, Data & Knowledge Engineering, 117, 53–70.
- [104] Chen D. (2018), “Neural reading comprehension and beyond”, Stanford University.
- [105] Liu S., Zhang X., Zhang S., Wang H., Zhang W. (2019), “Neural machine reading comprehension: Methods and trends”, Applied Sciences, 9(18), 3698.
- [106] Li W., Li Y., Pandelea V., Ge M., Zhu L., Cambria E. (2022), “Ecpec: Emotion-cause pair extraction in conversations”, IEEE Transactions on Affective Comput-

ing, 14(3), 1754–1765.

- [107] Kazi S., Khoja S., Daud A. (2023), “A survey of deep learning techniques for machine reading comprehension”, *Artificial Intelligence Review*, 56(Suppl 2), 2509–2569.
- [108] Jurafsky D. (2020), “Speech and language processing”, Prentice Hall.
- [109] Riloff E., Thelen M. (2000), A rule-based question answering system for reading comprehension tests, in “ANLP-NAACL 2000 workshop: reading comprehension tests as evaluation for computer-based language understanding systems”.
- [110] Lin K., Bogin B., Neumann M., Berant J., Gardner M. (2019), “Grammar-based neural text-to-sql generation”, arXiv preprint arXiv:1905.13326.
- [111] Yu T., Li Z., Zhang Z., Zhang R., Radev D. (2018), “Typesql: Knowledge-based type-aware neural text-to-sql generation”, arXiv preprint arXiv:1804.09769.
- [112] Akour M., Abufardeh S., Magel K., Al-Radaideh Q. (2011), “Qarabpro: A rule based question answering system for reading comprehension tests in arabic”, *American Journal of Applied Sciences*, 8(6), 652–661.
- [113] Ng H. T., Teo L. H., Kwan J. L. P. (2000), A machine learning approach to answering questions for reading comprehension tests, in “2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora”, 124–132.
- [114] Cairns B. L., Nielsen R. D., Masanz J. J., Martin J. H., Palmer M. S., Ward W. H., Savova G. K. (2011), The mipacq clinical question answering system, in “AMIA annual symposium proceedings”, Vol. 2011, American Medical Informatics Association, 171.
- [115] Clark P., Etzioni O., Khot T., Sabharwal A., Tafjord O., Turney P., Khashabi D. (2016), Combining retrieval, statistics, and inference to answer elementary science questions, in “Proceedings of the AAAI conference on artificial intelligence”, Vol. 30.

- [116] Welbl J., Stenetorp P., Riedel S. (2018), “Constructing datasets for multi-hop reading comprehension across documents”, *Transactions of the Association for Computational Linguistics*, 6, 287–302.
- [117] Rajpurkar P. (2016), “Squad: 100,000+ questions for machine comprehension of text”, *arXiv preprint arXiv:1606.05250*.
- [118] Ghaeini R., Fern X. Z., Shahbazi H., Tadeipalli P. (2018), “Dependent gated reading for cloze-style question answering”, *arXiv preprint arXiv:1805.10528*.
- [119] Liu C., Zhao Y., Si Q., Zhang H., Li B., Yu D. (2018), Multi-perspective fusion network for commonsense reading comprehension, in “Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 17th China National Conference, CCL 2018, and 6th International Symposium, NLP-NABD 2018, Changsha, China, October 19–21, 2018, Proceedings 17”, Springer, 262–274.
- [120] Clark C., Gardner M. (2017), “Simple and effective multi-paragraph reading comprehension”, *arXiv preprint arXiv:1710.10723*.
- [121] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. (2017), “Attention is all you need.(nips), 2017”, *arXiv preprint arXiv:1706.03762*, 10, S0140525X16001837.
- [122] Kim Y. (2014), Convolutional neural networks for sentence classification, in “Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)”, Association for Computational Linguistics, Doha, Qatar, 1746–1751.
- [123] Gong Y., Bowman S. R. (2017), “Ruminating reader: Reasoning with gated multi-hop attention”, *arXiv preprint arXiv:1704.07415*.
- [124] Zhang J., Zhu X., Chen Q., Ling Z., Dai L., Wei S., Jiang H. (2017), Exploring question representation and adaptation with neural networks, in “2017 3rd IEEE International Conference on Computer and Communications (ICCC)”, IEEE, 1975–1984.

- [125] Wang S., Yu M., Chang S., Jiang J. (2018), “A co-matching model for multi-choice reading comprehension”, arXiv preprint arXiv:1806.04068.
- [126] Du X., Cardie C. (2018), “Harvesting paragraph-level question-answer pairs from wikipedia”, arXiv preprint arXiv:1805.05942.
- [127] Yang Z., Dhingra B., Yuan Y., Hu J., Cohen W. W., Salakhutdinov R. (2016), “Words or characters? fine-grained gating for reading comprehension”, arXiv preprint arXiv:1611.01724.
- [128] Yang Y., Kang S., Seo J. (2020), “Improved machine reading comprehension using data validation for weakly labeled data”, IEEE Access, 8, 5667–5677.
- [129] Ren Q., Cheng X., Su S. (2020), Multi-task learning with generative adversarial training for multi-passage machine reading comprehension, in “Proceedings of the AAAI Conference on Artificial Intelligence”, Vol. 34, 8705–8712.
- [130] Lee H.-g., Kim H. (2020), “Gf-net: Improving machine reading comprehension with feature gates”, Pattern Recognition Letters, 129, 8–15.
- [131] Zhang C., Luo C., Lu J., Liu A., Bai B., Bai K., Xu Z. (2020), Read, attend, and exclude: multi-choice reading comprehension by mimicking human reasoning process, in “Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval”, 1945–1948.
- [132] Bahdanau D. (2014), “Neural machine translation by jointly learning to align and translate”, arXiv preprint arXiv:1409.0473.
- [133] Yang Z. (2019), “Xlnet: Generalized autoregressive pretraining for language understanding”, arXiv preprint arXiv:1906.08237.
- [134] Tu M., Huang K., Wang G., Huang J., He X., Zhou B. (2020), Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents, in “Proceedings of the AAAI conference on artificial intelligence”, Vol. 34, 9073–9080.
- [135] Hewlett D., Jones L., Lacoste A., Gür I. (2017), Accurate supervised and semi-

- supervised machine reading for long documents, in “Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing”, 2011–2020.
- [136] Min S., Seo M., Hajishirzi H. (2017), “Question answering through transfer learning from large fine-grained supervision data”, arXiv preprint arXiv:1702.02171.
- [137] Sachan M., Xing E. (2018), Self-training for jointly learning to ask and answer questions, in “Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)”, 629–640.
- [138] Liu Q., Geng X., Huang H., Qin T., Lu J., Jiang D. (2021), “Mgrc: An end-to-end multigranularity reading comprehension model for question answering”, IEEE Transactions on Neural Networks and Learning Systems, 34(5), 2594–2605.
- [139] Liu Q., Mao R., Geng X., Cambria E. (2023), “Semantic matching in machine reading comprehension: An empirical study”, Information Processing & Management, 60(2), 103145.
- [140] Iyer S., Min S., Mehdad Y., Yih W.-t. (2021), Reconsider: improved re-ranking using span-focused cross-attention for open domain question answering, in “Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies”, 1280–1287.
- [141] Nie Y., Wang S., Bansal M. (2019), “Revealing the importance of semantic retrieval for machine reading at scale”, arXiv preprint arXiv:1909.08041.
- [142] Mrini K., Farcas E., Nakashole N. (2021), Recursive tree-structured self-attention for answer sentence selection, in “Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)”, 4651–4661.
- [143] Bonadiman D., Moschitti A. (2020), “A study on efficiency, accuracy and document structure for answer sentence selection”, arXiv preprint arXiv:2003.02349.
- [144] Zhang Z., Yang J., Zhao H. (2021), Retrospective reader for machine reading

comprehension, in “Proceedings of the AAAI conference on artificial intelligence”, Vol. 35, 14506–14514.

- [145] Kratzwald B., Eigenmann A., Feuerriegel S. (2019), “Rankqa: Neural question answering with answer re-ranking”, arXiv preprint arXiv:1906.03008.
- [146] Green Jr B. F., Wolf A. K., Chomsky C., Laughery K. (1961), Baseball: an automatic question-answerer, in “Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference”, 219–224.
- [147] Pandya H. A., Bhatt B. S. (2021), “Question answering survey: Directions, challenges, datasets, evaluation matrices”, arXiv preprint arXiv:2112.03572.
- [148] Lan Y., Wang S., Jiang J. (2019), “Knowledge base question answering with a matching-aggregation model and question-specific contextual relations”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(10), 1629–1638.
- [149] Abujabal A., Yahya M., Riedewald M., Weikum G. (2017), Automated template generation for question answering over knowledge graphs, in “Proceedings of the 26th international conference on world wide web”, 1191–1200.
- [150] Bast H., Haussmann E. (2015), More accurate question answering on freebase, in “Proceedings of the 24th ACM international on conference on information and knowledge management”, 1431–1440.
- [151] Jin W., Zhao B., Yu H., Tao X., Yin R., Liu G. (2023), “Improving embedded knowledge graph multi-hop question answering by introducing relational chain reasoning”, Data Mining and Knowledge Discovery, 37(1), 255–288.
- [152] Chen Y., Subburathinam A., Chen C.-H., Zaki M. J. (2021), Personalized food recommendation as constrained question answering over a large-scale food knowledge graph, in “Proceedings of the 14th ACM international conference on web Search and data mining”, 544–552.
- [153] He G., Lan Y., Jiang J., Zhao W. X., Wen J.-R. (2021), Improving multi-hop

knowledge base question answering by learning intermediate supervision signals, in “Proceedings of the 14th ACM international conference on web search and data mining”, 553–561.

- [154] Lan Y., Jiang J. (2020), Query graph generation for answering multi-hop complex questions from knowledge bases, Association for Computational Linguistics.
- [155] Li Z., Wang H., Zhang W. (2022), “Translational relation embeddings for multi-hop knowledge base question answering”, Journal of Web Semantics, 74, 100723.
- [156] Gu Y., Kase S., Vanni M., Sadler B., Liang P., Yan X., Su Y. (2021), Beyond iid: three levels of generalization for question answering on knowledge bases, in “Proceedings of the Web Conference 2021”, 3477–3488.
- [157] Cao S., Shi J., Pan L., Nie L., Xiang Y., Hou L., Li J., He B., Zhang H. (2020), “Kqa pro: A dataset with explicit compositional programs for complex question answering over knowledge base”, arXiv preprint arXiv:2007.03875.
- [158] Chakraborty A. (2024), “Multi-hop question answering over knowledge graphs using large language models”, arXiv preprint arXiv:2404.19234.
- [159] Xiong G., Bao J., Zhao W. (2024), “Interactive-kbqa: Multi-turn interactions for knowledge base question answering with large language models”, arXiv preprint arXiv:2402.15131.
- [160] Lan Y., He G., Jiang J., Jiang J., Zhao W. X., Wen J.-R. (2022), “Complex knowledge base question answering: A survey”, IEEE Transactions on Knowledge and Data Engineering, 35(11), 11196–11215.
- [161] Gu Y., Pahuja V., Cheng G., Su Y. (2022), “Knowledge base question answering: A semantic parsing perspective”, arXiv preprint arXiv:2209.04994.
- [162] Tan Y., Min D., Li Y., Li W., Hu N., Chen Y., Qi G. (2023), Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family, in “International Semantic Web Conference”, Springer, 348–367.

- [163] Sun J., Xu C., Tang L., Wang S., Lin C., Gong Y., Shum H.-Y., Guo J. (2023), “Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph”, arXiv preprint arXiv:2307.07697.
- [164] Gu Y., Deng X., Su Y. (2023), Don’t generate, discriminate: A proposal for grounding language models to real-world environments, in “The 61st Annual Meeting Of The Association For Computational Linguistics”.
- [165] Li T., Ma X., Zhuang A., Gu Y., Su Y., Chen W. (2023), Few-shot in-context learning on knowledge base question answering, in “Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)”, 6966–6980.
- [166] Wang C., Xu Y., Peng Z., Zhang C., Chen B., Wang X., Feng L., An B. (2023), “keqing: knowledge-based question answering is a nature chain-of-thought mentor of llm”, arXiv preprint arXiv:2401.00426.
- [167] Shah M., Tian J. (2024), Improving llm-based kgqa for multi-hop question answering with implicit reasoning in few-shot examples, in “Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)”, 125–135.
- [168] Chidambaram M., Yang Y., Cer D., Yuan S., Sung Y.-H., Strope B., Kurzweil R. (2018), “Learning cross-lingual sentence representations via a multi-task dual-encoder model”, arXiv preprint arXiv:1810.12836.
- [169] Le Q., Mikolov T. (2014), Distributed representations of sentences and documents, in “International conference on machine learning”, PMLR, 1188–1196.
- [170] Bagga A., Baldwin B. (1998), Entity-based cross-document coreferencing using the vector space model, in “COLING 1998 Volume 1: The 17th international conference on computational linguistics”.
- [171] Amigó E., Gonzalo J., Artiles J., Verdejo F. (2009), “A comparison of extrinsic clustering evaluation metrics based on formal constraints”, *Information retrieval*, 12, 461–486.

- [172] Piriayatamwong P., Kalloori S., Zünd F. (2024), Improving german news clustering with contrastive learning, in “Proceedings of the 33rd ACM International Conference on Information and Knowledge Management”, 3979–3983.
- [173] Chen C. C., Wang H.-C. (2023), “Adapting the influences of publishers to perform news event detection”, *Journal of Information Science*, 49(5), 1277–1292.
- [174] Al-Betar M. A., Abasi A. K., Al-Naymat G., Arshad K., Makhadmeh S. N. (2023), “Bare-bones based salp swarm algorithm for text document clustering”, *IEEE Access*.
- [175] Abasi A. K., Khader A. T., Al-Betar M. A., Naim S., Makhadmeh S. N., Alyasseri Z. A. A. (2020), “Link-based multi-verse optimizer for text documents clustering”, *Applied Soft Computing*, 87, 106002.
- [176] Bouras C., Tsogkas V. (2012), “A clustering technique for news articles using wordnet”, *Knowledge-Based Systems*, 36, 115–128.
- [177] Gwadera R., Crestani F. (2009), Mining and ranking streams of news stories using cross-stream sequential patterns, in “Proceedings of the 18th ACM conference on Information and knowledge management”, 1709–1712.
- [178] Lu Y., Mei Q., Zhai C. (2011), “Investigating task performance of probabilistic topic models: an empirical study of plsa and lda”, *Information Retrieval*, 14, 178–203.
- [179] Shomanov A., Mansurova M. (2021), Parallel news clustering and topic modeling approaches, in “*Journal of Physics: Conference Series*”, Vol. 1727, IOP Publishing, 012018.
- [180] Bojanowski P., Grave E., Joulin A., Mikolov T. (2017), “Enriching word vectors with subword information”, *Transactions of the association for computational linguistics*, 5, 135–146.
- [181] Pennington J., Socher R., Manning C. D. (2014), Glove: Global vectors for word representation, in “Proceedings of the 2014 conference on empirical methods in

natural language processing (EMNLP)”, 1532–1543.

- [182] Lan Z., Chen M., Goodman S., Gimpel K., Sharma P., Soricut R. (2019), “Albert: A lite bert for self-supervised learning of language representations”, arXiv preprint arXiv:1909.11942.
- [183] Ester M., Kriegel H.-P., Sander J., Xu X., others (1996), A density-based algorithm for discovering clusters in large spatial databases with noise, in “kdd”, Vol. 96, 226–231.
- [184] McInnes L., Healy J., Astels S. (2017), “hdbscan: Hierarchical density based clustering”, Journal of open source software, 2(11), 205.
- [185] Ankerst M., Breunig M. M., Kriegel H.-P., Sander J. (1999), “Optics: Ordering points to identify the clustering structure”, ACM Sigmod record, 28(2), 49–60.
- [186] Hamborg F., Breiteringer C., Gipp B. (2019), “Giveme5w1h: A universal system for extracting main events from news articles”, arXiv preprint arXiv:1909.02766.
- [187] Keith B., Horning M., Mitra T. (2020), “Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization”, Computational Journalism C+ J.
- [188] Park R. E. (1940), “News as a form of knowledge: A chapter in the sociology of knowledge”, American journal of Sociology, 45(5), 669–686.
- [189] Hu X., Ma W., Chen C., Wen S., Zhang J., Xiang Y., Fei G. (2022), “Event detection in online social network: Methodologies, state-of-art, and evolution”, Computer Science Review, 46, 100500.
- [190] Cekineli R. F., Karagoz P. (2022), “Event prediction from news text using subgraph embedding and graph sequence mining”, World Wide Web, 25(6), 2403–2428.
- [191] Vossen P., Caselli T., Kontzopoulou Y. (2015), Storylines for structuring massive streams of news, in “Proceedings of the First Workshop on Computing News Storylines”, 40–49.

- [192] “dbmdz Turkish bert model” (2020), <https://huggingface.co/dbmdz/bert-base-turkish-cased> , Accessed: Oct. 19,2022.
- [193] “spaCy English Trained Pipelines” (n.d.), https://spacy.io/models/en#en_core_web_lg , Accessed: Nov. 5,2022.
- [194] “spaCy German Trained Pipelines” (n.d.), https://spacy.io/models/de#de_core_news_lg , Accessed: Nov. 5,2022.
- [195] “spaCy Spanish Trained Pipelines” (n.d.), https://spacy.io/models/es#es_core_news_lg , Accessed: Nov. 5,2022.
- [196] Strötgen J., Gertz M. (2015), A baseline temporal tagger for all languages, in “Proceedings of the 2015 conference on empirical methods in natural language processing”, 541–547.
- [197] Camcı H., Eryiğit G. (2021), “Türkçe zamansal ifadelerin yakalanması ve tanımlanması”, Bilişim Teknolojileri Dergisi, 14(3), 337–343.
- [198] Genc H., Yilmaz B. (2019), Text-based event detection: deciphering date information using graph embeddings, in “Big Data Analytics and Knowledge Discovery: 21st International Conference, DaWaK 2019, Linz, Austria, August 26–29, 2019, Proceedings 21”, Springer, 266–278.
- [199] Yan X., Han J. (2002), gspan: Graph-based substructure pattern mining, in “2002 IEEE International Conference on Data Mining, 2002. Proceedings.”, IEEE, 721–724.
- [200] Rupnik J., Muhic A., Leban G., Skraba P., Fortuna B., Grobelnik M. (2016), “News across languages-cross-lingual document similarity and event tracking”, Journal of Artificial Intelligence Research, 55, 283–316.
- [201] Lorbeer B., Kosareva A., Deva B., Softić D., Ruppel P., Küpper A. (2018), “Variations on the clustering algorithm birch”, Big data research, 11, 44–53.
- [202] Bagga A., Baldwin B. (1998), Algorithms for scoring coreference chains, in “The first international conference on language resources and evaluation workshop

on linguistics coreference”, Vol. 1, Citeseer, 563–566.

- [203] Heusden R., Kamps J., Marx M. (2022), Bcubed revisited: elements like me, in “Proceedings of the 2022 acm sigir International Conference on Theory of Information Retrieval”, 127–132.
- [204] Lang A., Schubert E. (2022), “Betula: Fast clustering of large data with improved birch cf-trees”, Information Systems, 108, 101918.
- [205] Zhang T., Ramakrishnan R., Livny M. (1997), “Birch: A new data clustering algorithm and its applications”, Data mining and knowledge discovery, 1, 141–182.
- [206] Stewart G., Al-Khassaweneh M. (2022), “An implementation of the hdbscan* clustering algorithm”, Applied Sciences, 12(5), 2405.
- [207] Campello R. J., Moulavi D., Sander J. (2013), Density-based clustering based on hierarchical density estimates, in “Pacific-Asia conference on knowledge discovery and data mining”, Springer, 160–172.
- [208] Wang D., Huang Y., Cai Z. (2023), “A two-phase clustering approach for traffic accident black spots identification: integrated gis-based processing and hdbscan model”, International journal of injury control and safety promotion, 30(2), 270–281.
- [209] Laban P., Hearst M. A. (2017), newslens: building and visualizing long-ranging news stories, in “Proceedings of the Events and Stories in the News Workshop”, 1–9.
- [210] Zou X. (2020), A survey on application of knowledge graph, in “Journal of Physics: Conference Series”, Vol. 1487, IOP Publishing, 012016.
- [211] Hogan A., Blomqvist E., Cochez M., d’Amato C., Melo G. D., Gutierrez C., Kirrane S., Gayo J. E. L., Navigli R., Neumaier S., others (2021), “Knowledge graphs”, ACM Computing Surveys (Csur), 54(4), 1–37.
- [212] Yani M., Krisnadhi A. A. (2021), “Challenges, techniques, and trends of simple knowledge graph question answering: a survey”, Information, 12(7), 271.

- [213] Yao Y., Zeng Y., Zhong N., Huang X. (2007), Knowledge retrieval (kr), in “IEEE/WIC/ACM International Conference on Web Intelligence (WI’07)”, IEEE, 729–735.
- [214] Berant J., Chou A., Frostig R., Liang P. (2013), Semantic parsing on freebase from question-answer pairs, in “Proceedings of the 2013 conference on empirical methods in natural language processing”, 1533–1544.
- [215] Jiang P., Cai X. (2024), “A survey of semantic parsing techniques”, *Symmetry*, 16(9), 1201.
- [216] Luo H., Tang Z., Peng S., Guo Y., Zhang W., Ma C., Dong G., Song M., Lin W., Zhu Y., others (2023), “Chatkbqa: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models”, arXiv preprint arXiv:2310.08975.
- [217] Jiang J., Zhou K., Dong Z., Ye K., Zhao W. X., Wen J.-R. (2023), “Structgpt: A general framework for large language model to reason over structured data”, arXiv preprint arXiv:2305.09645.
- [218] Liu R., Mao R., Luu A. T., Cambria E. (2023), “A brief survey on recent advances in coreference resolution”, *Artificial Intelligence Review*, 56(12), 14439–14481.
- [219] Detroja K., Bhensdadia C., Bhatt B. S. (2023), “A survey on relation extraction”, *Intelligent Systems with Applications*, 19, 200244.
- [220] Cabot P.-L. H., Navigli R. (2021), Rebel: Relation extraction by end-to-end language generation, in “Findings of the Association for Computational Linguistics: EMNLP 2021”, 2370–2381.
- [221] Ye X., Yavuz S., Hashimoto K., Zhou Y., Xiong C. (2021), “Rng-kbqa: Generation augmented iterative ranking for knowledge base question answering”, arXiv preprint arXiv:2109.08678.
- [222] Hu N., Wu Y., Qi G., Min D., Chen J., Pan J. Z., Ali Z. (2023), “An empirical study of pre-trained language models in simple knowledge graph question answering”,

World Wide Web, 26(5), 2855–2886.

- [223] Wei X., Cui X., Cheng N., Wang X., Zhang X., Huang S., Xie P., Xu J., Chen Y., Zhang M., others (2023), “Zero-shot information extraction via chatting with chatgpt”, arXiv preprint arXiv:2302.10205.
- [224] Nori H., King N., McKinney S. M., Carignan D., Horvitz E. (2023), “Capabilities of gpt-4 on medical challenge problems”, arXiv preprint arXiv:2303.13375.
- [225] Ma Y., Cao Y., Hong Y., Sun A. (2023), “Large language model is not a good few-shot information extractor, but a good reranker for hard samples!”, arXiv preprint arXiv:2303.08559.
- [226] Schneider P., Klettner M., Jokinen K., Simperl E., Matthes F. (2024), “Evaluating large language models in semantic parsing for conversational question answering over knowledge graphs”, arXiv preprint arXiv:2401.01711.
- [227] “OpenAI models” (2024), <https://platform.openai.com/docs/models> , Accessed: Nov. 24,2024.
- [228] “Claude 3.5 Sonnet” (2024), <https://www.anthropic.com/news/claude-3-5-sonnet> , Accessed: Nov. 24,2024.
- [229] “Gemini Models” (2024), <https://deepmind.google/technologies/gemini/> , Accessed: Nov. 24,2024.
- [230] Anil R., Borgeaud S., Wu Y., Alayrac J.-B., Yu J., Soricut R., Schalkwyk J., Dai A. M., Hauth A., Millican K., others (2023), “Gemini: A family of highly capable multimodal models”, arXiv preprint arXiv:2312.11805.
- [231] “Qwen’s Collections” (2024), <https://qwen2.org/qwen2-5/> , Accessed: Nov. 24,2024.
- [232] “Qwen2.5-Coder-32B-Instruct Model Card” (2024), <https://huggingface.co/Qwen/Qwen2.5-Coder-32B-Instruct> , Accessed: Nov. 24,2024.
- [233] Hui B., Yang J., Cui Z., Yang J., Liu D., Zhang L., Liu T., Zhang J., Yu

- B., Dang K., others (2024), “Qwen2. 5-coder technical report”, arXiv preprint arXiv:2409.12186.
- [234] Wei J., Wang X., Schuurmans D., Bosma M., Xia F., Chi E., Le Q. V., Zhou D., others (2022), “Chain-of-thought prompting elicits reasoning in large language models”, *Advances in neural information processing systems*, 35, 24824–24837.
- [235] Kwiatkowski T., Palomaki J., Redfield O., Collins M., Parikh A., Alberti C., Epstein D., Polosukhin I., Devlin J., Lee K., others (2019), “Natural questions: a benchmark for question answering research”, *Transactions of the Association for Computational Linguistics*, 7, 453–466.
- [236] Lewis P., Oğuz B., Rinott R., Riedel S., Schwenk H. (2019), “Mlqa: Evaluating cross-lingual extractive question answering”, arXiv preprint arXiv:1910.07475.
- [237] Elsahar H., Vougiouklis P., Remaci A., Gravier C., Hare J., Laforest F., Simperl E. (2018), T-rex: A large scale alignment of natural language with knowledge base triples, in “*Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*”.
- [238] Levy O., Seo M., Choi E., Zettlemoyer L. (2017), “Zero-shot relation extraction via reading comprehension”, arXiv preprint arXiv:1706.04115.
- [239] Sağlam B. D. (2024), Knowledge graph augmented multi-hop question answering using large language models, Master’s thesis, Middle East Technical University.
- [240] He B., Liu Y., Hildebrandt M., Ding Z., Han Y., Tresp V. (2024), An automated evaluation framework for graph database query generation leveraging large language models, in “*Proceedings of the Third International Workshop on Linked Data-driven Resilience Research (D2R2’24)*”.
- [241] Zhang T., Kishore V., Wu F., Weinberger K. Q., Artzi Y. (2019), “Bertscore: Evaluating text generation with bert”, arXiv preprint arXiv:1904.09675.
- [242] Artificial Intelligence T. A. I. (2022), “Musique question answering competition leaderboard”, https://leaderboard.allenai.org/musique_ans/submissions/public ,

Accessed: November 30, 2024.

- [243] Beltagy I., Peters M. E., Cohan A. (2020), “Longformer: The long-document transformer”, arXiv preprint arXiv:2004.05150.
- [244] Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. (2019), “Roberta: A robustly optimized bert pretraining approach. arxiv [preprint](2019)”, arXiv preprint arXiv:1907.11692.



ÖZGEÇMİŞ

Başak BULUZ KÖMEÇOĞLU, lisans derecesini 2014 yılında İstanbul Aydın Üniversitesi Matematik-Bilgisayar ve Bilgisayar Mühendisliği bölümlerini çift anadal programını tamamlayarak aldı. Yüksek lisansını Gebze Teknik Üniversitesi, Bilgisayar Mühendisliği bölümünde “Akademik Başarının Modellenmesinde Çizge Madenciliği Yaklaşımı” tez çalışması ile 2018 yılında tamamlamıştır. Doktora eğitimine 2018 yılından beri Gebze Teknik Üniversitesi Bilgisayar Mühendisliği bölümünde devam etmektedir.

2014-2020 yılları arasında İstanbul Aydın Üniversitesi Anadolu Bil Meslek Yüksekokulu, İstanbul, Türkiye’de öğretim görevlisi ve aynı zamanda 5 yıl boyunca Bilgisayar Programcılığı (İngilizce) Program Başkanı olarak görev yapmıştır. 2020 yılından beri Gebze Teknik Üniversitesi Bilişim Teknolojileri Enstitüsü’nde araştırma görevlisidir. Araştırma alanları arasında makine öğrenmesi, derin öğrenme, metin madenciliği, veri madenciliği, doğal dil işleme ve çizge madenciliği ve uygulamaları yer almaktadır. Buluz Kömeçoğlu ve ekibi, veri merkezli nesne algılama tekniği ile İstanbul’da düzenlenen ulusal Teknofest 2021 Aerial Object Detection Challenge’da birincilik ödülüne layık görülmüştür.

TEZ ÇALIŞMASI KAPSAMINDA YAPILAN YAYINLAR

Kömeçoğlu, B. B., Yılmaz, B., (2024) "Event Graph-Based News Clustering: The Role of Named Entity-Centered Subgraphs," in IEEE Access, vol. 12, pp. 105613-105632, doi: 10.1109/ACCESS.2024.3435343.

Kömeçoğlu, B. B., Yılmaz, B., (2022), "Topic Detection and Tracking Towards Determining Public Agenda Items: The Impact of Named Entities on Event-Based News Clustering," In Machine Learning for Societal Improvement, Modernization, and Progress, pp. 158-17, doi: 10.4018/978-1-6684-4045-2.ch008.

Kömeçoğlu, Y., Kömeçoğlu, B. B., Yılmaz, B. (2021), "Real-Time News Grouping: Detecting the Same-Content News on Turkish News Stream," In Progress in Intelligent Decision Science: Proceeding of IDS 2020, pp. 14-25, doi: 10.1007/978-3-030-66501-2_2.

Kömeçoğlu, B. B., Yılmaz, B. (2024), "Knowledge-Augmented Large Language Model Prompting for Cypher Query Construction," In 9th International Conference on Computer Science and Engineering (UBMK) 2024, pp. 187-192, doi: 10.1109/UBMK63289.-2024.10773450

EKLER

Ek-A : Veri Kümeleri

Tez çalışması sırasında literatüre sunulan Hikaye Tabanlı KBQA Veri Kümesi, Hikaye Tabanlı Haber Veri Kümesi ve Gruplandırılmış Türkçe Haber Metinleri Test Kümesi github sayfasında paylaşılacaktır.

Çalışmada değerlendirmeler için kullanılan Musique-Ans veri kümesi ve liderlik tablosu araştırmacıların oluşturduğu resmi web sitesinde yer almaktadır. Ayrıca Musique-Ans validation-mini kümesi için örneklem kimliklerinin listesine ilgili yüksek lisans tezinin APPENDIX B.1 bölümünden ulaşılabilir [239].

Bölüm 5.4.2’de sunulan ”Açık/Örtük Zaman Etiketli Düşümlerinin Zamansal Soru Cevaplama Performansına Etkisi” deneyleri için kullanılan Musique-Ans doğrulama veri kümesinde yer alan zamansal soruların kimlik numaraları aşağıda paylaşılmıştır.

2hop__103889_86452	2hop__151455_120171	2hop__23459_35123
2hop__107601_110222	2hop__153532_72380	2hop__2416_86208
2hop__107905_110222	2hop__157737_46621	2hop__24661_81961
2hop__109242_113442	2hop__158275_35739	2hop__253612_110882
2hop__123104_77615	2hop__159146_120065	2hop__26642_66233
2hop__129499_85379	2hop__159741_161674	2hop__26642_81961
2hop__130404_59180	2hop__159741_40169	2hop__272714_113442
2hop__130867_54221	2hop__159903_154896	2hop__278022_71701
2hop__131061_54221	2hop__161151_42429	2hop__286621_67370
2hop__132472_120537	2hop__16267_25719	2hop__286621_84856
2hop__132909_33564	2hop__16581_16525	2hop__293131_110222
2hop__132929_120537	2hop__16632_110014	2hop__30032_46621
2hop__133241_33564	2hop__167577_31122	2hop__31091_31122
2hop__13548_13529	2hop__16844_20510	2hop__31094_31122
2hop__135822_46621	2hop__16844_42173	2hop__31124_31122
2hop__135993_160249	2hop__170823_120171	2hop__315072_120259
2hop__136067_160249	2hop__17508_17335	2hop__32254_84601
2hop__136179_13529	2hop__175168_110222	2hop__326140_9237
2hop__13986_161674	2hop__176_51164	2hop__326404_120065
2hop__13986_40169	2hop__176712_8311	2hop__327688_56873
2hop__14092_8311	2hop__198650_84856	2hop__329676_119915
2hop__142699_67465	2hop__199216_33564	2hop__337223_56873
2hop__142881_127399	2hop__203955_25719	2hop__34594_160249
2hop__144317_120065	2hop__20433_157939	2hop__34595_160249
2hop__145282_21711	2hop__215898_67465	2hop__34734_92852

2hop__347686_127399	2hop__59612_42429	2hop__85463_33311
2hop__35136_35117	2hop__603365_120259	2hop__85463_66122
2hop__35173_17335	2hop__603598_25719	2hop__857493_17335
2hop__35209_17335	2hop__606439_127399	2hop__861047_110949
2hop__370564_71701	2hop__609313_20273	2hop__86458_20273
2hop__37097_86208	2hop__61714_89309	2hop__87953_17335
2hop__3737_13529	2hop__619265_72380	2hop__9174_9237
2hop__3739_13529	2hop__62020_158262	2hop__9223_9237
2hop__388467_110882	2hop__622637_120171	2hop__9238_9237
2hop__391857_162018	2hop__62311_62279	2hop__94629_1110
2hop__39231_17335	2hop__629030_56873	3hop1__106845_5188_86687
2hop__39420_39345	2hop__635544_110949	3hop1__120881_2053_52946
2hop__394720_120259	2hop__645448_77615	3hop1__127821_157828_162309
2hop__40697_46621	2hop__655505_110949	3hop1__131820_59747_43101
2hop__41158_35123	2hop__67057_52871	3hop1__131820_59747_60748
2hop__418360_67370	2hop__677160_110882	3hop1__132795_40769_64047
2hop__433694_20273	2hop__6827_55848	3hop1__132798_40769_64047
2hop__435184_67370	2hop__6827_75783	3hop1__132967_40769_64047
2hop__435184_84856	2hop__685930_110949	3hop1__132968_40769_64047
2hop__442175_56873	2hop__691639_154896	3hop1__133264_40769_64047
2hop__445963_6095	2hop__711639_110949	3hop1__135392_87694_124169
2hop__450600_158262	2hop__73615_20510	3hop1__135392_87694_64412
2hop__451574_120065	2hop__73615_42173	3hop1__135659_87694_124169
2hop__46862_68042	2hop__748615_123956	3hop1__135659_87694_64412
2hop__47071_162399	2hop__75207_46837	3hop1__135793_87694_64412
2hop__47712_25719	2hop__75207_7363	3hop1__135794_87694_64412
2hop__486392_35739	2hop__753260_71701	3hop1__136016_87694_124169
2hop__500379_126537	2hop__756620_127399	3hop1__136016_87694_64412
2hop__503085_110222	2hop__75714_21969	3hop1__136083_87694_124169
2hop__50432_56873	2hop__765180_123956	3hop1__136083_87694_64412
2hop__511454_120259	2hop__780238_110949	3hop1__136129_87694_124169
2hop__516918_126537	2hop__781079_9237	3hop1__136129_87694_64412
2hop__518354_67465	2hop__78401_158262	3hop1__139862_2053_52946
2hop__51881_20590	2hop__794999_113442	3hop1__140345_2053_52946
2hop__520033_123956	2hop__797443_120537	3hop1__140554_2053_52946
2hop__530579_13529	2hop__799936_20273	3hop1__140786_2053_52946
2hop__53433_59035	2hop__806869_56873	3hop1__140855_2053_52946
2hop__536283_67465	2hop__80834_21969	3hop1__140994_2053_52946
2hop__539184_119915	2hop__814856_110882	3hop1__141138_2053_52946
2hop__540035_113442	2hop__818302_25719	3hop1__145194_160545_34751
2hop__544760_35739	2hop__818692_113442	3hop1__145427_106426_77199
2hop__556426_9237	2hop__826864_17335	3hop1__161433_33952_34099
2hop__557284_160249	2hop__829474_123956	3hop1__163654_503371_21711
2hop__567588_67465	2hop__83405_158262	3hop1__17192_78396_157843
2hop__575282_114648	2hop__8347_8311	3hop1__17192_78396_54865
2hop__57543_158262	2hop__837900_71701	3hop1__186307_503371_21711
2hop__576118_25719	2hop__846921_160249	3hop1__220023_503371_21711
2hop__580231_162399	2hop__84847_21969	3hop1__222497_309482_89834
2hop__590631_110882	2hop__84937_21969	3hop1__222979_40769_64047

3hop1__268450_503371_21711	3hop1__567955_503371_21711	3hop1__90327_78276_68042
3hop1__273571_834494_34053	3hop1__573244_339990_15538	3hop1__92251_34179_23375
3hop1__273571_834494_34099	3hop1__579562_629431_124169	3hop1__9285_5188_23307
3hop1__295815_40769_64047	3hop1__579562_629431_64412	3hop1__9285_5188_63702
3hop1__302383_503371_21711	3hop1__593597_40769_64047	3hop1__9285_5188_86687
3hop1__305282_282081_10972	3hop1__595371_40769_64047	3hop1__92991_73181_68042
3hop1__305282_282081_73772	3hop1__603558_87694_124169	3hop1__92991_78276_68042
3hop1__315504_629431_124169	3hop1__603558_87694_64412	3hop2__104979_23308_28343
3hop1__315504_629431_64412	3hop1__604991_339990_15538	3hop2__106842_30645_84681
3hop1__347858_503371_21711	3hop1__609479_503371_21711	3hop2__11282_57233_56883
3hop1__353661_629431_124169	3hop1__617062_127905_80487	3hop2__132957_132798_40768
3hop1__353661_629431_64412	3hop1__617062_127905_87812	3hop2__132957_132967_40768
3hop1__354480_834494_34053	3hop1__617062_127905_89859	3hop2__132957_132968_40768
3hop1__354480_834494_34099	3hop1__617062_127905_92225	3hop2__132957_133264_40768
3hop1__358853_339990_15538	3hop1__626214_503371_21711	3hop2__132957_133292_40768
3hop1__36497_73181_68042	3hop1__637555_339990_15538	3hop2__132957_222979_40768
3hop1__36497_78276_68042	3hop1__639955_834494_34053	3hop2__132957_295815_40768
3hop1__365934_87694_124169	3hop1__639955_834494_34099	3hop2__132957_379231_40768
3hop1__365934_87694_64412	3hop1__640171_228453_10972	3hop2__132957_433432_40768
3hop1__379231_40769_64047	3hop1__64694_34179_23375	3hop2__132957_483843_40768
3hop1__38738_73181_68042	3hop1__64957_87694_124169	3hop2__132957_51933_40768
3hop1__38738_78276_68042	3hop1__64957_87694_64412	3hop2__132957_52222_40768
3hop1__390673_228453_10972	3hop1__652852_2053_52946	3hop2__132957_593597_40768
3hop1__395211_503371_21711	3hop1__672543_503371_21711	3hop2__132957_595371_40768
3hop1__398232_326948_10972	3hop1__680374_503371_21711	3hop2__132957_76377_40768
3hop1__398232_326948_59893	3hop1__698501_503371_21711	3hop2__132957_800948_40768
3hop1__404541_629431_124169	3hop1__707569_228453_10972	3hop2__222979_132477_40768
3hop1__404541_629431_64412	3hop1__73796_87694_124169	3hop2__222979_132536_40768
3hop1__423001_503371_21711	3hop1__73796_87694_64412	3hop2__222979_132628_40768
3hop1__446612_160545_34751	3hop1__751782_2053_52946	3hop2__222979_132957_40768
3hop1__462960_160545_34751	3hop1__76377_40769_64047	3hop2__222979_133099_40768
3hop1__465684_160545_34751	3hop1__764280_834494_34053	3hop2__222979_133154_40768
3hop1__46571_87694_124169	3hop1__764280_834494_34053	3hop2__222979_380281_40768
3hop1__46571_87694_64412	3hop1__764280_834494_34053	3hop2__222979_840908_40768
3hop1__483634_720378_15538	3hop1__765565_326948_10972	3hop2__27069_92991_76291
3hop1__483843_40769_64047	3hop1__773623_87694_124169	3hop2__28727_92991_76291
3hop1__491648_339990_15538	3hop1__783711_503371_21711	3hop2__304722_203717_63959
3hop1__49236_89752_54221	3hop1__786067_228453_10972	3hop2__304722_230978_63959
3hop1__497845_629431_124169	3hop1__800948_40769_64047	3hop2__304722_238553_63959
3hop1__497845_629431_64412	3hop1__813285_774554_124169	3hop2__304722_297619_63959
3hop1__516535_834494_34053	3hop1__813285_774554_64412	3hop2__304722_330033_63959
3hop1__516535_834494_34099	3hop1__818422_160545_34751	3hop2__304722_397371_63959
3hop1__516535_834494_34099	3hop1__823336_228453_10972	3hop2__304722_398767_63959
3hop1__516535_834494_34099	3hop1__831499_228453_10972	3hop2__304722_423353_63959
3hop1__51933_40769_64047	3hop1__840353_503371_21711	3hop2__304722_466316_63959
3hop1__52222_40769_64047	3hop1__847420_720378_15538	3hop2__304722_526346_63959
3hop1__545624_2053_52946	3hop1__849312_503371_21711	3hop2__304722_582839_63959
3hop1__5544_34179_23375	3hop1__857241_503371_21711	3hop2__304722_590317_63959
3hop1__558449_503371_21711	3hop1__90327_73181_68042	3hop2__304722_591230_63959

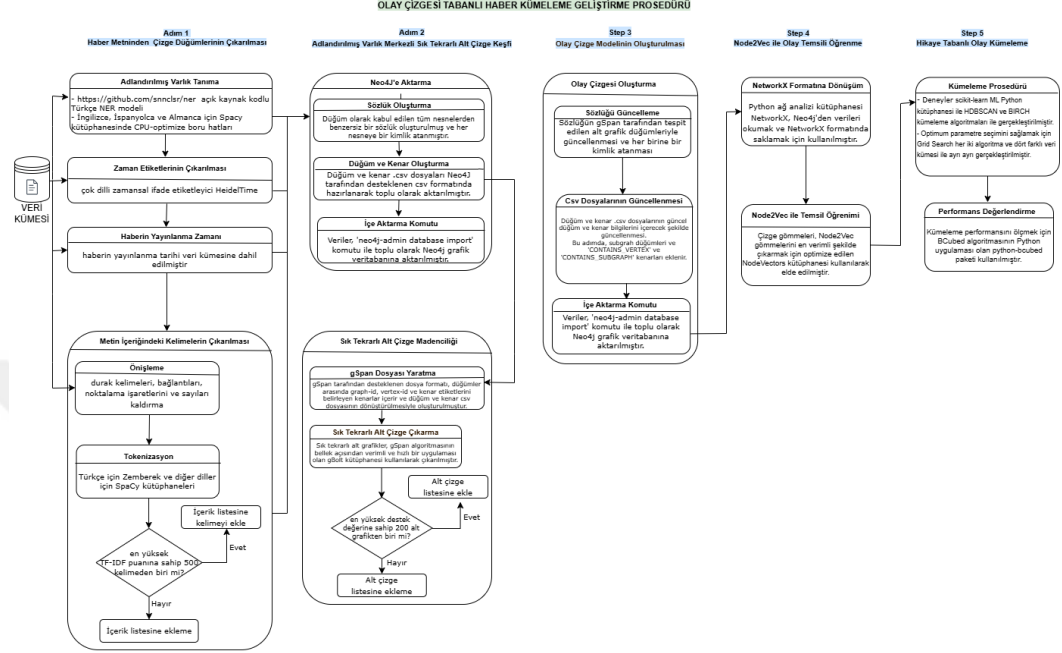
3hop2__304722_602987_63959	162309	4hop2__103790_14670_8987_8529
3hop2__304722_667199_63959	4hop1__145494_698949_157828_-	4hop2__103790_39078_8987_8529
3hop2__304722_689786_63959	162309	4hop2__105527_14670_8987_8529
3hop2__304722_692815_63959	4hop1__146971_698949_157828_-	4hop2__160585_14670_8987_8529
3hop2__304722_740784_63959	162309	4hop2__161324_14670_8987_8529
3hop2__304722_747004_63959	4hop1__151650_5274_458768_33632	4hop2__161602_426860_88460_21034
3hop2__304722_781203_63959	4hop1__151650_5274_458768_33633	4hop2__161602_426860_88460_63559
3hop2__304722_796034_63959	4hop1__151650_5274_458768_33677	4hop2__161602_426860_88460_86452
3hop2__30956_65244_84681	4hop1__151903_5274_458768_33633	4hop2__161602_474028_88460_21034
3hop2__36497_89854_76291	4hop1__151903_5274_458768_33677	4hop2__161602_474028_88460_63559
3hop2__38242_92991_76291	4hop1__152028_141441_458768_33633	4hop2__161602_474028_88460_86452
3hop2__38326_92991_76291	4hop1__152146_5274_458768_33632	4hop2__5206_14670_8987_8529
3hop2__43283_12039_93690	4hop1__152146_5274_458768_33633	4hop2__567956_14670_8987_8529
3hop2__43892_23308_28343	4hop1__152146_5274_458768_33677	4hop2__71753_158279_70784_61381
3hop2__57233_105140_56883	4hop1__152562_5274_458768_33632	4hop2__71753_158279_70784_79935
3hop2__57233_106716_56883	4hop1__152562_5274_458768_33633	4hop2__71753_158985_70784_61381
3hop2__57233_11282_56883	4hop1__152562_5274_458768_33677	4hop2__71753_158985_70784_79935
3hop2__6989_37759_74563	4hop1__17192_17130_70784_61381	4hop2__71753_261673_70784_61381
3hop2__72083_92991_76291	4hop1__17192_17130_70784_61381	4hop2__71753_261673_70784_79935
3hop2__79512_16214_84681	4hop1__17192_17130_70784_79935	4hop2__71753_623626_70784_61381
3hop2__87184_38738_76291	4hop1__342858_131850_33952_34053	4hop2__71753_623626_70784_79935
3hop2__87184_90327_76291	4hop1__342858_131850_33952_34099	4hop2__71753_648517_70784_61381
3hop2__87184_92991_76291	4hop1__343844_17130_70784_61381	4hop2__71753_648517_70784_79935
3hop2__88342_93066_47738	4hop1__343844_17130_70784_61381	4hop2__71753_685934_70784_61381
3hop2__88342_93066_90766	4hop1__343844_17130_70784_79935	4hop2__71753_685934_70784_79935
3hop2__89854_36497_76291	4hop1__3496_5274_458768_33633	4hop2__71753_729371_70784_61381
3hop2__89854_38738_76291	4hop1__3496_5274_458768_33677	4hop2__71753_729371_70784_79935
3hop2__89854_90327_76291	4hop1__39871_314549_131976_44892	4hop2__71753_73205_70784_79935
3hop2__89854_92991_76291	4hop1__443274_17130_70784_61381	4hop2__9988_158279_70784_61381
3hop2__90327_87184_76291	4hop1__443274_17130_70784_79935	4hop2__9988_158279_70784_79935
3hop2__90327_89854_76291	4hop1__567737_141375_458768_33632	4hop2__9988_158985_70784_61381
3hop2__92991_27069_76291	4hop1__567737_141375_458768_33633	4hop2__9988_158985_70784_79935
3hop2__92991_28727_76291	4hop1__567737_141375_458768_33677	4hop2__9988_261673_70784_61381
3hop2__92991_38242_76291	4hop1__58323_375563_161848_53331	4hop2__9988_261673_70784_79935
3hop2__92991_38326_76291	4hop1__58323_375563_161848_84681	4hop2__9988_685934_70784_61381
3hop2__92991_72083_76291	4hop1__638988_17130_70784_61381	4hop2__9988_729371_70784_61381
3hop2__92991_87184_76291	4hop1__638988_17130_70784_79935	4hop2__9988_729371_70784_79935
3hop2__92991_89854_76291	4hop1__726391_153080_33952_34053	4hop3__160230_470287_128875_72134
3hop2__93066_88342_90766	4hop1__749065_698949_157828_-	4hop3__160230_470287_345497_72134
4hop1__105002_17130_70784_61381	162309	4hop3__160230_470287_862998_72134
4hop1__105401_17130_70784_61381	4hop1__88342_49853_128008_80487	4hop3__193820_466199_128875_72134
4hop1__105401_17130_70784_79935	4hop1__88342_49853_128008_87812	4hop3__193820_466199_695123_72134
4hop1__105688_17130_70784_61381	4hop1__88342_49853_128008_89859	4hop3__322354_466199_695123_72134
4hop1__105688_17130_70784_79935	4hop1__88342_49853_128008_92225	4hop3__324979_836463_161616_77103
4hop1__115977_582428_157828_-	4hop1__88342_75218_128008_80487	4hop3__373866_5189_276146_86687
162309	4hop1__88342_75218_128008_87812	4hop3__373866_5189_31809_86687
4hop1__118068_582428_157828_-	4hop1__88342_75218_128008_89859	4hop3__373866_5189_38229_86687
162309	4hop1__88342_75218_128008_92225	4hop3__524186_219173_128875_72134
4hop1__145422_698949_157828_-	4hop1__9007_698949_157828_162309	4hop3__524186_219173_548463_72134

4hop3__524186_219173_638950_72134	4hop3__771862_466199_695123_72134	4hop3__857_846_613770_7713
4hop3__524186_219173_862998_72134	4hop3__773800_152874_161616_77103	4hop3__862_846_326964_7713
4hop3__531933_836463_161616_77103	4hop3__794915_466199_695123_72134	4hop3__862_846_613770_7713
4hop3__601548_836463_161616_77103	4hop3__800262_466199_695123_72134	
4hop3__678814_466199_695123_72134	4hop3__857_846_326964_7713	



Ek-B: Olay Çizgesi Tabanlı Haber Kümeleme Geliştirme Prosedürü

Bölüm 4’de deney sonuçları sunulan ”Olay Çizgesi Tabanlı Haber Kümeleme” yaklaşımına dair geliştirme prosedürleri Şekil 6.1’de sunulmuştur.



Şekil 6.1: Olay Çizgesi Tabanlı Haber Kümeleme Geliştirme Prosedürü.