T.C.
İSTANBUL ÜNİVERSİTESİ-CERRAHPAŞA
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

DOKTORA TEZİ

DERİN ÖĞRENME YÖNTEMLERİ İLE İLİŞKİSEL DOKÜMAN SINIFLANDIRILMASI

Halil İbrahim OKUR

DANIŞMAN
Prof. Dr. Ahmet SERTBAŞ
(danışman imzalı olacaktır)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği, Doktora Programı

Haziran, 2024

TEZ KABUL VE ONAYI

Halil İbrahim OKUR tarafından, **Prof. Dr. Ahmet SERTBAŞ** danışmanlığında hazırlanan **"DERİN ÖĞRENME YÖNTEMLERİ İLE İLİŞKİSEL DOKÜMAN SINIFLANDIRILMASI"** başlıklı bu çalışma, jürimiz tarafından **12/06/2024** tarihinde yapılan sınav sonucunda **oy birliği** ile başarılı bulunarak **Doktora Tezi** olarak kabul edilmiştir.

Tez Jürisi

		İmza	Sonuç
DANIŞMAN	Prof. Dr. Ahmet SERTBAŞ İstanbul Üniversitesi - Cerrahpaşa Bilgisayar Mühendisliği Anabilim Dalı		☒ Kabul ☐ Ret
ÜYE	Doç. Dr. Atakan KURT İstanbul Üniversitesi - Cerrahpaşa Bilgisayar Mühendisliği Anabilim Dalı		☒ Kabul ☐ Ret
ÜYE	Doç. Dr. Can EYÜPOĞLU Milli Savunma Üniversitesi Bilgisayar MühendisliğiAnabilim Dalı		☒ Kabul ☐ Ret
ÜYE	Doç. Dr. Muhammed Ali AYDIN İstanbul Üniversitesi - Cerrahpaşa Bilgisayar Mühendisliği Anabilim Dalı		☒ Kabul ☐ Ret
ÜYE	Dr. Öğr. Üyesi Kadir TOHMA İskenderun Teknik Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı		☒ Kabul ☐ Ret

Okumaya, araştırmaya ve bilgiye tutkulu, saygın her bireye ve ayrıca 6 Şubat 2023 tarihinde gerçekleşen depremde vefat eden tüm vatandaşlarımıza ithaf ediyorum…

BÜTÇE DESTEKLERİ

DERİN ÖĞRENME YÖNTEMLERİ İLE İLİŞKİSEL DOKÜMAN SINIFLANDIRILMASI

Bu tez çalışması için herhangi bir kurumdan bütçe desteği alınmamıştır.

TEŞEKKÜR

Desteğini hiçbir zaman eksik etmeyen sevgili eşim İlknur OKUR'a ve her zaman enerjimi ve neşemi arttıran oğlum Ahmet Erdem OKUR'a, annem Medine OKUR, babam Özkan Deniz OKUR ve kardeşim Mehmet Erdem OKUR'a ve her zaman bana yol gösteren saygıdeğer hocam Prof. Dr. Ahmet SERTBAŞ'a, destekleri ve arkadaşlıkları ile her zaman yanımda olan Kadir TOHMA, Hüseyin ATASOY, Ebu Yusuf GÜVEN ve Mehmet Yavuz YAĞCI hocalarıma, akademik ve kişisel gelişimimde önemli payı olan İstanbul Üniversitesi – Cerrahpaşa Bilgisayar Mühendisliğindeki tüm hocalarıma ve İskenderun Teknik Üniversitesi Bilgisayar Mühendisliği Bölümündeki tüm çalışma arkadaşlarıma saygı ve sevgilerimle teşekkür ederim.

Haziran 2024 Halil İbrahim OKUR

İÇİNDEKİLER

Sayfa No

ŞEKİL LİSTESİ

Sayfa No

TABLO LİSTESİ

Sayfa No

SİMGE VE KISALTMA LİSTESİ

Kısaltmalar	Açıklama
BERT	: Bidirectional Encoder Representations from Transformers
CNN	: Evrişimli Sinir Ağları
DNN	: Derin Sinir Ağları
GNN	: Graf Sinir Ağları
KB	: Knowledge Bases (Bilgi Tabanları)
LR	: Lojistik Regresyon
NER	: Named Entity Recognition (Adlandırılmış Varlık Tanıma)
NLP	: Natural Language Processing (Doğal Dil İşleme)
RE	: İlişki Çıkarma (Relation Extraction)
RNN	: Tekrarlayan Sinir Ağları
SPARQL	: SPARQL Protocol and RDF Query Language (RDF Sorgu Dili)
S-R-O	: Subject – Relation – Object (Özne – İlişki – Nesne)
SVM	: Destek Vektör Makineleri

ÖZET

DOKTORA TEZİ

DERİN ÖĞRENME YÖNTEMLERİ İLE İLİŞKİSEL DOKÜMAN SINIFLANDIRILMASI

Halil İbrahim OKUR

İstanbul Üniversitesi-Cerrahpaşa

Lisansüstü Eğitim Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği, Doktora Programı

Danışman: Prof. Dr. Ahmet SERTBAŞ

Metin sınıflandırması, geniş ve karmaşık metin havuzlarından anlamlı bilgiler çıkarmak ve bu metinleri otomatik olarak kategorilere ayırmak için kullanılan doğal dil işleme (NLP) uygulamalarından biridir. Ancak, zengin morfolojik yapıya sahip dillerde ve çeşitli veri kaynaklarında bu yöntemler yetersiz kalabilir. Bu sorunu ele almak için, bu çalışma Named Entity Recognition (NER) modelleri ve ön eğitimli BERT tabanlı modeller ile Wikidata'dan elde edilen varlık-ilişki bilgilerinin metinlere entegrasyonunu araştırmaktadır. Türkçe ve İngilizce metin veri setleri üzerinde yapılan değerlendirmeler, çeşitli makine ve derin öğrenme sınıflandırma algoritmalarının performansını incelemiştir. Bulgular, varlık-ilişki bilgilerinin metinlere eklenmesi ile algoritmaların başarımını önemli ölçüde artırdığını göstermektedir. Özellikle BBC-News ve TTC-4900 veri setlerinde, ilişkisel bilgi entegrasyonu sınıflandırma başarısını belirgin şekilde iyileştirmiştir. Türkçe metinlerin zengin yapılarını etkin bir şekilde işleyebilmek için TRT-Haber web sayfasından elde edilen veri seti üzerinde graf tabanlı derin öğrenme teknikleri uygulanmış, bu tekniklerle dikkate değer bir doğruluk oranı elde edilmiştir. Araştırmanın sonuçları, ilişkisel bilginin eklenmesinin metin sınıflandırmasında etkinliğini ve potansiyelini göstermekte olup, Türkçe metinlerin daha iyi anlaşılmasına katkıda bulunmaktadır. Bu çalışma, ilişkisel metin sınıflandırma sistemlerinin geliştirilmesine önemli katkılar sağlamış ve gelecek araştırmalar için önemli bir referans noktası oluşturmuştur.

Haziran 2024 , 111 sayfa.

Anahtar kelimeler: İlişkisel metin sınıflandırma, derin öğrenme, graf sinir ağları, ilişki çıkarımı, bilgi tabanları

ABSTRACT

Ph.D. THESIS

RELATIONAL DOCUMENT CLASSIFICATION WITH DEEP LEARNING METHODS

Halil Ibrahim OKUR

İstanbul University-Cerrahpaşa

Institute of Graduate Studies

Department of Computer Engineering

Computer Engineering, PhD Program

Supervisor : Prof. Dr. Ahmet SERTBAŞ

Text classification, used to extract meaningful information from vast and complex pools of text and to categorically organize it automatically, is one of the applications of natural language processing (NLP). However, these methods can be insufficient for languages with rich morphological structures and diverse data sources. To address this issue, this study investigates the integration of entity-relation information from Wikidata into texts using Named Entity Recognition (NER) models and pre-trained BERT-based models. Evaluations conducted on Turkish and English text datasets have examined the performance of various machine learning and deep learning classification algorithms. Findings indicate that the addition of entity-relation information significantly enhances the performance of algorithms. Particularly, the integration of relational information has notably improved classification success in the BBC-News and TTC-4900 datasets. For effectively processing the complex structures of Turkish texts, graph-based deep learning techniques have been applied to a dataset obtained from the TRT-News website, achieving a noteworthy accuracy rate. The results demonstrate the efficacy and potential of adding relational information in text classification, contributing to a better understanding of Turkish texts. This research has made significant contributions to the development of relational text classification systems and has established an important reference point for future studies.

June 2024, 111 pages.

Keywords: Relational text classification, deep learning, graph neural networks, relationship extraction, knowledge bases

1. GİRİŞ

Doğal Dil İşleme (NLP), insan dilinin karmaşıklığını ve zenginliğini bilgisayar sistemleri aracılığıyla yorumlama çabasını temsil eder. Bu alanda, son yıllarda (Abdullah, 2023) özellikle belirgin olan bir sorun, geniş doküman havuzlarından istenilen anlamlı bilgilerin çıkarılması veya sınıflandırılması olmuştur. Özellikle sosyal medya, web, iş dünyası ve eğitim gibi alanlarda metin ve belge birikimi, günümüzün en büyük bilgi işleme zorluklarından biri haline gelmiştir. Bu geniş metin yığınlarından anlamlı bilgilerin çıkarılması ve belgelerin etkin bir şekilde sınıflandırılması, insan gücüyle pratik olarak imkansız hale gelmiştir.

Doğal Dil İşleme (NLP), bilgisayarların insan dilini anlayıp işleyebilmesini sağlayan bir bilim dalı olarak, bu meydan okumayı ele almakta ve belgeleri, dilin anlamsal ve dilbilgisel özelliklerine dayalı olarak otomatik olarak sınıflandırabilen sistemler geliştirmekte büyük rol oynamaktadır. Bu sistemler, metinlerin içeriğine göre sınıflandırma işlemi gerçekleştirirken, kelimelerin anlamları, kelime öbekleri ve dilbilgisi kuralları gibi dil özelliklerini dikkate alarak daha yüksek doğruluk oranları ve daha hızlı sınıflandırma süresi sunarak, kullanıcı dostu bir deneyim sunmaktadır. Bu bağlamda metin sınıflandırma, metinlere, cümlelere, paragraflara ve dokümanlara sistematik bir şekilde etiketler atamayı amaçlayan, Doğal Dil İşleme' nin temel bir yaklaşımı olarak öne çıkmaktadır (Hotho, 2005).

Metinler, bireylerin bilgi ve düşüncelerini diğerlerine aktarabilmelerini sağlayarak anlamlı ve düzenli bir iletişimde bilgi alışverişini kolaylaştırır. Buradaki bir cümle veya ifade, kelimelerin söz dizimsel ve anlamsal özelliklerinin kombinasyonu ile oluşur. Cümleler arası ilişkiler paragrafları, paragraflar arası ilişkiler ise metinleri oluşturur. Bu süreçte, dokümanların bileşimsel özellikleri ve heterojen bilgiler, Doğal Dil İşleme için önemli bir zorluk teşkil eder. Her dokümanın benzersiz yapısal karmaşıklığı ve içerdiği bilgi türlerinin çeşitliliği, NLP tekniklerinin bu metinlerden anlamlı bilgi çıkarma yeteneğini önemli ölçüde etkiler. Dolayısıyla, metinlerin söz dizimsel ve anlamsal özelliklerini, ve bu özelliklerin nasıl bir araya gelerek tutarlı bir anlam oluşturduğunu anlamak, etkili bir metin sınıflandırma sistemi geliştirmek için kritik öneme sahiptir (Zhu, 2023).

Metin sınıflandırma ise, dokümanların benzer içerik ve bilgilere göre tanımlanıp sıralanması işlemidir, bu süreçte dokümanlar etiket bilgisiyle karakterize edilir. Bu etiket bilgisi, dokümanın özetini, başlığını, içeriğini veya dokümanlar arasındaki ilişkileri temsil eden bilgileri içerebilir. Bu bilgiler sayesinde metinlerin analizi, sınıflandırma sürecinde önemli bir rol oynar (Han ve diğ., 2022). Doğal dil işleme alanında temel bir sorun olan metin sınıflandırma, dilin yapısal, bölgesel, kültürel ve zamansal çeşitliliklerine uygun tekniklerin uygulanmasını gerektirir. Metinler, insanlar arasında bilgi ve düşüncelerin aktarımını kolaylaştırırken, doğal dilin bilgisayar sistemlerine entegrasyonu, etkili bir insan-bilgisayar etkileşimi sağlamak için önemlidir. Bu bağlamda, metinlerin yapay zeka sistemleri tarafından işlenebilir hale getirilmesi ve ön işlem aşamalarından geçirilmesi, sistemlerin performansını belirleyen temel faktörler arasındadır (Mooney ve Roddick, 2013). Dolayısıyla kullanılan yöntemler, dilin morfolojisine uygun olmalı ve metinler buna göre hazırlanmalıdır.

Uzun yıllardan beri doğal dilin söz dizimsel ve anlamsal yapısının bilgisayar sistemlerine kazandırılmasına yönelik çalışmalar yapılmaktadır. Bilgi çıkarma, özetleme, soru cevaplama, metin sınıflandırma ve diğer çeşitli doğal dil işleme problemleri bu çalışmalara örnek olarak verilebilir. Bu sorunları çözmenin ilk adımı, metinleri yapay zeka sistemlerinin kullanabileceği bir formata dönüştürmek, metinlerin ön işleme tabi tutulmasıdır. İyi bir metin ön işleme aşaması, yapay zeka sistemlerinin performansını etkileyebilir. Ancak karmaşık etiketlere sahip web sayfalarını metin ön işlemeye ihtiyaç duymadan doğrudan okuyabilen GPT-4 (Achiam ve diğ., 2023) ve Llama2 (Touvron ve diğ., 2023) gibi modellerin geliştirilmesi ön işlemenin önemini azaltmıştır. Özellikle bu yeni nesil yapay zeka sistemleri, doğal dilin yapısal ve anlamsal karmaşıklığını daha önce gerekli olan yoğun ön işleme adımlarına gerek kalmadan işleyebilme yeteneğine sahiptir. Öte yandan, özellikle belirli uygulamalar ve karmaşık dil işleme gereksinimleri için metin ön işlemenin sağladığı derinlemesine temizlik ve yapılandırma, değerli bir adım olmayı sürdürüyor.

Bu tez çalışmasında metinler içerisinde yer alan varlıklar tespit edilerek ve bu varlıklara ait ilişkilerin bir bilgi bankasından eklenmesi ve zenginleştirilmesiyle metin sınıflandırma performansının arttırılması amaçlanmıştır. Motivasyonumuz, hızla gelişen metin tabanlı yapay zeka araçlarının kullanıldığı alanlarda otomatik ve uzaktan denetim yoluyla bilgiye daha kolay ve kapsamlı erişimi desteklemektir. Önceden eğitilmiş BERT (Bidirectional Encoder Representations from Transformers) tabanlı isimlendirilmiş varlık tespiti (Named Entity Recognition – NER) modelleri, metin içindeki varlıkları tespit etmek için kullanılır. Varlık

tespitinden sonra bu varlıklarla ilgili ilişkiler, SPARQL sorgu dili kullanılarak Wikidata veri tabanında sorgulanır. Elde edilen varlık-ilişki bilgisi, metinler ile entegre edilerek metin verileri zenginleştirilir. Türkçe ve İngilizce metin veri setleri üzerinde sınıflandırma performansını ortaya koymak amacıyla farklı makine öğrenme algoritmaları, derin öğrenme algoritmaları ve graf tabanlı yöntemler kullanılmıştır. Sonuçlar, metinleri varlık-ilişki bilgileriyle zenginleştirmenin, metin sınıflandırma görevlerinde tatmin edici performans sağladığını göstermektedir. Bu tez çalışmasının katkıları aşağıda sıralanmıştır:

- İlişkisel Türkçe metin sınıflandırma yaklaşımı: Türkçe metin sınıflandırma görevleri için varlık-ilişki bilgisine dayalı yeni bir metin sınıflandırma yaklaşımı tanıtılmıştır. Sınıflandırma performansını ortaya koymak amacıyla varlık ilişkileriyle zenginleştirilmiş metinlere çeşitli makine öğrenmesi, derin öğrenme algoritmaları, graf tabanlı yöntemler uygulanmıştır.

- Veri tabanı oluşturma ve uzaktan denetim: SPARQL (SPARQL Protocol and RDF Query Language) sorgulama dili ve Wikidata veri tabanı kullanılarak varlık-ilişki bilgi tabanları oluşturulmuştur. İngilizce ve Türkçe için varlık-ilişki bilgileri toplanarak NLP çalışmalarında uzaktan denetim yoluyla ilişkisel metin sınıflandırma görevlerine kaynak sağlanması amaçlanmıştır. Ayrıca, bir haber sitesinden veri kazıma (web scraping) yöntemi ile içerdikleri etiketlere dayalı bir ilişkisel metin veri seti elde edilmiştir.

- Metin Sınıflandırma Performansı: Tanıtılan ilişkisel model, literatürde daha önce aynı görev için kullanılan metin veri kümeleri üzerinde metin sınıflandırma performansında iyileşme göstermektedir.

Katkılar incelendiğinde ilişkisel metin sınıflandırma yaklaşımıyla Türkçe çalışmaları için yeni bir yaklaşım sunulmaktadır. Tez çalışmasının 2. Bölümünde çalışmanın amacı, hedefleri ve kapsamı hakkında bilgiler verilmiş ve incelenen ilgili literatür çalışmaları detaylı olarak anlatılmıştır. 3. bölümde bu çalışmada önerilen ilişkisel sınıflandırma yöntemi ve graf sinir ağı modeli ayrıntılı olarak anlatılmaktadır. Bölüm 4'te bu çalışmanın deneysel bulguları ve performans ölçümleri sunulmaktadır. Bölüm 5'te ise çalışmanın bulguları ve mevcut literatür bilgisi değerlendirilerek yorumlanmıştır. Son olarak 6. bölümde çalışmanın sonuçları ve çıkarımlarının özeti ve sentezi sunulmaktadır. Bu bölümde ayrıca araştırmanın sınırlılıkları tartışılmakta ve gelecek araştırmalara yönelik öneriler sunulmaktadır.

2. KAVRAMSAL ÇERÇEVE

Metin sınıflandırma, doğal dil işleme (NLP) alanının en önemli araştırma konularından biridir ve çeşitli uygulama alanlarında yaygın olarak kullanılmaktadır. Bu disiplin, metinlerin otomatik olarak belirli kategorilere veya etiketlere atanmasını sağlayarak, bilgi yönetimi, duygu analizi, spam filtreleme, belge sınıflandırma gibi alanlarda temel bir rol oynar. Özellikle günümüzde, dijitalleşmenin getirdiği bilgi çağında, metin verisinin hacmi hızla artmaktadır. Bu artış, verilerden anlamlı bilgiler çıkarma ihtiyacını da beraberinde getirir. Metin sınıflandırma çalışmaları, bu geniş veri havuzlarını analiz ederek, verileri yönetilebilir ve işlenebilir bilgilere dönüştürme amacı güder (Kaur ve diğ., 2018).

Literatürde metin sınıflandırma üzerine çok sayıda yöntem ve teknik önerilmiş olmasına karşın, metinlerdeki varlık ilişkilerinin kullanımı ve bu ilişkilerin sınıflandırma performansı üzerindeki etkisine dair sınırlı sayıda çalışma mevcuttur. Bu durum, özellikle Türkçe metin sınıflandırmasında, varlık ilişkilerinin yeterince araştırılmamış bir alan olduğunu gösterir. Bu çalışma, Türkçe metin sınıflandırmasında varlık ilişkilerinin kullanılmasının önemini vurgulayarak, mevcut literatürü bu açıdan ele almakta ve kapsamlı bir inceleme sunmaktadır. Ayrıca, çalışma metin sınıflandırma performansını artırmak için makine öğrenimi ve derin öğrenme yöntemlerinin nasıl kullanılabileceğini araştırmaktadır. Bu kapsamda, Derin Sinir Ağları (DNN) (Magalhães ve diğ., 2023), Tekrarlayan Sinir Ağları (RNN) (Liu ve diğ., 2016), Evrişimli Sinir Ağları (CNN) (Sarasu ve diğ., 2023), Graf Sinir Ağları (GNN) (Huang ve diğ., 2019) ve hibrit modeller gibi çeşitli derin öğrenme modelleri (Akpatsa ve diğ., 2021), metin sınıflandırıcılarında kullanılarak etkili sonuçlar elde edilmektedir. Aynı zamanda, Destek Vektör Makineleri (SVM) (Gunawan ve diğ., 2023), Karar Ağaçları (DT) gibi makine öğrenimi algoritmaları (Agarwal ve diğ., 2023) da metin sınıflandırma modelleri arasında değerlendirilmektedir. Metin sınıflandırmada kullanılan kelime gömme teknikleri (Incitti ve diğ., 2023) ve dil modeli tabanlı yaklaşımlar (Sun ve diğ., 2023), varlık ilişkilerinin kullanımı için uygun bir zemin hazırlamaktadır. Bu teknikler, metinlerdeki kelimelerin anlamsal özelliklerini ve bağlamsal ilişkilerini dikkate alarak, metinlerin daha anlamlı bir şekilde temsil edilmesini sağlar (Asudani ve diğ., 2023).

Bu literatür taraması, Türkçe metin sınıflandırmasının mevcut durumunu özetlemekte ve varlık ilişkilerinin bu alandaki gelişmelere nasıl katkıda bulunabileceğini vurgulamaktadır. Literatürde metin sınıflandırma veya ilişkisel metin çalışmalarına ilişkin bilgiler detaylı bir

şekilde incelenmiş, en iyi sonuçları gösteren sınıflandırma yöntemleri belirlenmiştir. Bu kapsamlı inceleme, "Metin Sınıflandırma Çalışmaları", "İlişkisel Metin Sınıflandırması" ve "Graf Tabanlı Modeller ve Uygulamaları" başlıklar altında yapılan çalışmaları detaylandırarak, metin sınıflandırma ve ilişkisel analiz çalışmalarının akademik literatürdeki yerini sağlam bir şekilde tanımlamaktadır. Bu bölüm, alandaki bilgi birikimine katkıda bulunmayı amaçlayarak, gelecekteki çalışmalara yön verecek temel bir kaynak oluşturur.

2.1. Metin Sınıflandırma Çalışmaları

Metin sınıflandırma, büyük metin veri kümelerinden anlamlı bilgileri çıkarmak amacıyla kullanılan kritik bir yöntemdir. Çeşitli dillerde ve alanlarda uygulanabilirliği ile dikkat çeken bu metodoloji, haber kategorisi bulma, duygu analizi, spam filtreleme, e-posta kategorizasyonu, müşteri yorumları analizi, makine çevirisi, soru cevaplama, metin özetleme, tıbbi metin analizi ve yasal belge analizi gibi geniş bir uygulama yelpazesi sunar. Dil ve metin tabanlı bilgi işlemenin merkezinde yer alan metin sınıflandırma, sürekli gelişim göstermekte ve yeni tekniklerle zenginleştirilmektedir (Kowsari, 2019).

Metin sınıflandırmanın gücüne rağmen, veri eksikliği, veri dengesizliği ve anlamsal belirsizlik gibi önemli zorluklarla karşılaşılmaktadır. Veri eksikliği, özellikle nadir görülen sınıflar için yeterli etiketli veri bulunmaması durumunda ortaya çıkar ve modelin genelleme yeteneğini kısıtlar. Veri dengesizliği, bazı sınıfların diğerlerine göre daha fazla temsil edilmesiyle, modelin az temsil edilen sınıfları doğru tanımasını zorlaştırır. Anlamsal belirsizlik ise kelime ve cümlelerin farklı bağlamlarda farklı anlamlar taşıması durumunda meydana gelir ve doğru sınıflandırmayı zorlaştırır.Bu zorlukların üstesinden gelmek için, veri toplama ve zenginleştirme, dengeli örnekleme teknikleri, kelime gömme modelleri ve bağlamsal dil modelleri gibi çeşitli yöntemler kullanılmaktadır. Kelime gömme modelleri, kelimelerin anlamlarını sayısal vektörlerle temsil ederek anlamsal ilişkileri yakalarken, bağlamsal dil modelleri kelimelerin ve cümlelerin bağlamını dikkate alarak daha doğru ve anlamlı temsiller öğrenir. Bu yaklaşımlar, metin sınıflandırma süreçlerinin etkinliğini artırarak daha güvenilir sonuçlar elde edilmesini sağlar (Incitti ve diğ., 2023).

Metin sınıflandırma, belge seviyesi, paragraf seviyesi, cümle seviyesi ve cümle altı seviyesi olmak üzere dört ana kapsamda uygulanabilir (Kowsari, 2019). Her bir seviye, sınıflandırmanın ihtiyaç duyduğu detay ve derinliği sağlar. Özellik çıkarma ise, metinlerin

yapılandırılmamış veri kümelerinden yapılandırılmış özellik uzayına dönüştürülmesi sürecidir (Li, 2022). Bu süreçte TF-IDF, Word2Vec, GloVe gibi kelime gömme teknikleri ve BERT gibi dil modeli tabanlı yaklaşımlar, metinlerdeki kelimelerin sıralama ve anlamsal özelliklerini dikkate alarak, metin sınıflandırma görevlerinde önemli bir rol oynar. Özellikle BERT modeli, metinlerin daha kapsamlı bir temsilini yakalayarak, kelime bağlamı ve ilişkilerinin daha iyi anlaşılmasını sağlar (Okur ve diğ., 2021).

Metin veya belge veri kümelerinin içerdiği geniş benzersiz kelime havuzu, veri ön işleme adımlarını zaman ve bellek karmaşıklığı açısından zorlayabilir. Bu durum, basit algoritmalar kullanarak boyut azaltma yöntemi ile çözülmeye çalışılırken, bazı durumlarda bu yöntemler beklenen performansı gösteremeyebilir. Bu nedenle, birçok araştırmacı, uygulamalarında zaman ve bellek karmaşıklığını azaltmak için boyut azaltma tekniklerini tercih etmektedir (Basha, 2019).

Metin sınıflandırma çalışmaları, geleneksel sınıflandırma algoritmalarından derin öğrenme yaklaşımlarına kadar geniş bir yelpazeyi kapsar. Geleneksel yöntemler arasında lojistik regresyon ve Naïve Bayes Classifier gibi hesaplaması ucuz ve az bellek gerektiren modeller yer alırken, SVM ve K-en yakın komşu (KNN) gibi teknikler, belirli görevlerde sınıflandırıcı olarak kullanılmaktadır (Li, 2022). Ağaç tabanlı sınıflandırıcılar ve koşullu rastgele alanlar (CRFs) gibi grafiksel sınıflandırma yöntemleri, özellikle belge özetleme ve otomatik anahtar kelime çıkarma gibi özel görevlerde ön plana çıkar (Augustyniak ve diğ., 2021).

Son yıllarda, derin öğrenme yaklaşımları, doğal dil işleme ve görsel sınıflandırma gibi alanlarda geleneksel makine öğrenimi algoritmalarına göre üstün sonuçlar elde etmiştir (Khan ve diğ., 2023). Bu başarı, derin öğrenme algoritmalarının karmaşık ve doğrusal olmayan ilişkileri modelleme kapasitesine dayanır. Metin sınıflandırma görevlerinin yanı sıra, doğal dil anlama (NLU) görevleri de, derin öğrenme tabanlı sınıflandırıcılar kullanılarak etkin bir şekilde ele alınabilir (Bharadiya, 2023). Duygu analizi, haber kategorizasyonu ve konu sınıflandırması, metin sınıflandırmanın başlıca uygulama alanlarındandır (Dawei, 2021).

Derin öğrenme modelleri, metin sınıflandırma için birçok modelin önerildiği bir alanı temsil eder. Bu modeller, RNN, CNN, kapsül ağları gibi çeşitli mimarilere dayanır ve metin içerisindeki önemli ifadeleri, kelime bağımlılıklarını ve metin yapılarını yakalama yeteneğiyle dikkat çeker. Dikkat mekanizmaları, metin içindeki ilişkili kelimeleri tanımlamada etkili olmuş ve model geliştirmede önemli bir araç haline gelmiştir. Graf Sinir Ağları(GNN) gibi yenilikçi

yaklaşımlar da doğal dilin içsel graf yapılarını anlama konusunda yeni perspektifler sunar (Li, 2022).

Transformers modelleri, derin öğrenme dünyasında önemli bir yenilik olarak öne çıkmaktadır. RNN' lerle karşılaştırıldığında, bu modeller yüksek düzeyde paralelleştirme kabiliyeti sayesinde, dil modellerinin GPU' lar aracılığıyla daha etkin bir şekilde eğitilmesine olanak tanır. Bu, özellikle geniş veri setleri üzerinde derin öğrenme modellerini eğitme sürecinde büyük bir avantaj sağlar. Okur ve arkadaşlarının yaptığı çalışmada (2021), Türkçe metin sınıflandırmada kullanılan ön eğitimli sinirsel modellerin verimliliği ve performansı detaylı bir şekilde incelenmiştir. Bu araştırmada, Word2Vec ve GloVe gibi klasik kelime gömme yöntemlerinin yanı sıra BERT, Electra ve FastText gibi daha yeni ve gelişmiş modellerin Türkçe metin sınıflandırma görevlerinde nasıl kullanılabileceği ve bu tekniklerin sunduğu avantajlar üzerine odaklanılmıştır.

Metin sınıflandırma çalışmalarının incelendiği bu bölüm, doğal dil işleme (NLP) alanındaki temel yöntemlerden en gelişmiş derin öğrenme modellerine kadar geniş bir yelpazeyi kapsamaktadır. Geleneksel makine öğrenimi algoritmalarından, karmaşık ve doğrusal olmayan ilişkileri başarıyla modelleyebilen derin öğrenme tekniklerine kadar olan süreç, metin sınıflandırma alanında kaydedilen ilerlemelerin çeşitliliğini ve derinliğini gözler önüne sermektedir.

Bu geniş çerçevede, metin sınıflandırmanın çeşitli uygulama alanlarına ve karşılaşılan zorluklara yönelik stratejiler geliştirilmiştir. Geliştirilen bu yöntemler, metin sınıflandırmanın sadece mevcut durumunu değil, aynı zamanda gelecekteki potansiyelini de belirleyen temel faktörlerdir. Bu temel üzerine inşa edilecek olan sonraki bölümler, metin sınıflandırma alanındaki spesifik araştırma konularının ve uygulama alanlarının daha detaylı bir şekilde ele alınmasına imkan tanıyacaktır. Bu şekilde, metin sınıflandırmanın çeşitli boyutları arasındaki bağlantıları ve bu disiplinin akademik literatürde nasıl bir yer edindiğini daha net bir şekilde anlayabileceğiz.

2.2. İlişkisel Metin Sınıflandırması

Metin sınıflandırma, doğal dil işleme (NLP) alanının en önemli araştırma konularından biridir ve çeşitli uygulama alanlarında yaygın olarak kullanılmaktadır. Bu disiplin, metinlerin otomatik olarak belirli kategorilere veya etiketlere atanmasını sağlayarak metin tabanlı

alanlarda temel bir rol oynar. Özellikle günümüzde, dijitalleşmenin getirdiği bilgi çağında, metin verisinin hacmi hızla artmaktadır. Bu artış, verilerden anlamlı bilgiler çıkarma ihtiyacını da beraberinde getirir. Metin sınıflandırma çalışmaları, bu geniş veri havuzlarını analiz ederek, verileri yönetilebilir ve işlenebilir bilgilere dönüştürme amacı güder (Sinoara ve diğ., 2019).

Literatürde metin sınıflandırma üzerine çok sayıda yöntem ve teknik önerilmiş (Kowsari ve diğ., 2019) olmasına karşın, metinlerdeki varlık ilişkilerinin kullanımı ve bu ilişkilerin sınıflandırma performansı üzerindeki etkisine dair sınırlı sayıda çalışma mevcuttur. Bu durum, özellikle Türkçe metin sınıflandırmasında, varlık ilişkilerinin yeterince araştırılmamış bir alan olduğunu gösterir. Bu çalışma, Türkçe metin sınıflandırmasında varlık ilişkilerinin kullanılmasının önemini vurgulayarak, mevcut literatürü bu açıdan ele almakta ve kapsamlı bir inceleme sunmaktadır. Ayrıca, çalışma metin sınıflandırma performansını artırmak için makine öğrenimi ve derin öğrenme yöntemlerinin nasıl kullanılabileceğini de araştırmaktadır.

Bu kapsamda literatürde metin sınıflandırma veya ilişkisel metin çalışmalarına ilişkin bilgiler ve performansları incelenmiştir. İncelenen literatür çalışmaları arasında en iyi sonucu gösteren sınıflandırma yöntemi tabloda sunulmaktadır. Literatürde incelenen çalışmalara ilişkin bilgiler Tablo 1'de gösterilmektedir.

Tablo 1. Literatürdeki metin sınıflandırma çalışmaları

Referans	Veriseti	Sınıflandırma Modeli	Doğruluk (Accuracy)	F1-Skor
(Kılınç, 2017)	TTC-3600	Random Forest	%91.03	-
(Flisar, 2018)	Web News	SVM	-	%90.10
(Kuyumcu, 2019)	TTC-3600	Multinomial NB	%90	-
(Parlak, 2023)	TTC-3600, 20Newsgroup	SVM	-	%78.1, %98.0
(Acı, 2019)	TTC-3600	CNN	%93.3	-
(Pittaras, 2021)	20Newsgroup	DNN	%82.6	-
(Lezama, 2022)	20Newsgroup	CNN	%79	-
(Nagumothu, 2021)	20Newsgroup	BERT ve Glove	%83.9	%83.1

Kılınc ve arkadaşları (2019) tarafından yürütülen bir çalışmada, Türkçe haber makaleleri ile metin kategorizasyonu araştırmaları için özel olarak TTC-3600 adında yeni bir veri seti geliştirilmiştir. Araştırmacılar, TTC-3600 üzerinde beş farklı metin sınıflandırıcısı ve iki özellik

seçim yöntemi değerlendirmiştir. Sonuçlar, özellik sıralama tabanlı özellik seçimiyle birleştirilen Rastgele Orman sınıflandırıcısının %91.03 doğruluk oranı elde ettiğini göstermiştir.

Flisar ve arkadaşları (2018) tarafından yürütülen çalışma, kısa metin sınıflandırma performansını, kısa metin doküman temsil modelini DBpedia bilgi tabanı ile zenginleştirerek artırmayı amaçlamıştır. Metin zenginleştirme yöntemleri, DBpedia Spotlight çerçevesi kullanılarak uygulanmış ve etkinlikleri değerlendirilmiştir. Bu çalışmada yürütülen deneyler, bu yaklaşımın çeşitli veri kümelerinde farklı sınıflandırıcılar kullanılarak temel yöntemleri geride bıraktığını göstermiştir.

Kuyumcu ve arkadaşları (2019) tarafından yürütülen bir çalışmada, Türkçe TTC-3600 veri seti, FastText kelime vektör temsil modeli ve klasik makine öğrenimi algoritmaları kullanılarak metin sınıflandırması için kullanılmıştır. Çok Sınıflı NB sınıflandırması kullanılarak yaklaşık %90 doğruluk oranı elde edilmiştir.

Parlak (2023) tarafından yürütülen çalışmada, ön işleme tekniklerinin sınıflandırma performansına katkıları değerlendirilmiştir. Çalışmada, sık kullanılan gereksiz kelimelerin (stop-word) temizlenmesi, kök indirgeme, harflerin küçültülmesi ve metin öbeklerine (token) ayırma gibi temel ön işleme adımlarının, metin sınıflandırma performansına olan katkılarını ve dil özelliklerine göre değişen performans farklılıklarını ortaya koymuştur. İki farklı ön işleme tekniğinin farklı haber veri kümeleri üzerindeki etkisi incelenmiştir. Türkçe veri kümesi (TTC-3600) için en yüksek F1 skoru 0.781 iken, İngilizce veri kümesi (20Newsgroup) için en yüksek F1 skoru 0.980 olmuştur.

Aci ve arkadaşları (2019) tarafından yürütülen metin sınıflandırma çalışması, TTC-3600 veri seti üzerinde Evrişimli Sinir Ağları ve Word2Vec metodu kullanılarak ve metinler Zemberek yazılımı kullanılarak ön işleme tabi tutularak gerçekleştirilmiştir. Önerilen yöntem %93.3 performans doğruluğu elde etmiştir.

Pittaras ve arkadaşları (2021), derin sinir ağları (DNN'ler) kullanarak metin sınıflandırma görevleri için girdiye anlamsal bilgi zenginleştirmesi yapmanın yöntemlerini araştırmıştır. Metinlerden anlamsal çıkarımlar yapılarak, ağırlıklı kavram terimlerinden oluşan bir anlamsal frekans vektörü oluşturulmuştur. Kavramlar, çeşitli anlamsal çözümleme teknikleri kullanılarak seçilmiş ve kelime vektörleriyle anlamsal ilişkiler göz önünde bulundurularak birleştirilmiştir.

Deneysel sonuçlar, bu anlamsal zenginleştirmenin sınıflandırma performansını önemli ölçüde artırdığını ve birleştirme yönteminin en iyi sonuçları verdiğini göstermiştir. Çalışma ayrıca, anlamsal vektörler üzerinde terim frekansı-ters belge frekansı normalizasyonunun etkis ini ve boyut azaltma potansiyelini de incelemiştir.

Lezama ve arkadaşları (2022), Wikipedia'dan ilişkiler çıkarma üzerine dayalı yenilikçi bir yaklaşım sunmaktadır. Sinonimlik, hiponimlik ve hiperonimlik gibi semantik ilişkileri çıkarmaya odaklanan bu çalışma, önerilen ilişkisel gömme modellerini CNN kullanarak metin sınıflandırma performansını değerlendirmiştir ve 20-Newsgroup veri kümesi üzerinde maksimum %79 doğruluk elde edilmiştir.

Nagumothu ve arkadaşları (2021) tarafından gerçekleştirilen bir çalışmada, doküman ilgisini sınıflandırmada bağlantılı veri üçlülerinin (konu-yüklem-nesne yapıları) kullanılması incelenmiştir. Araştırmacılar, özel olarak geliştirilen bir haber akışı web sitesi için özelleştirilmiş bir veri seti üzerinde, bağlantılı veri üçlülerinin eklenmesinin GloVe temsillerinin F-skorunu %6 oranında artırdığını bulmuşlardır. Çalışma, bağlantılı veri üçlülerinin, derin öğrenme modelleri ve metin temsilleri gibi mevcut yaklaşımlara ek olarak doküman sınıflandırma performansını önemli ölçüde geliştirebileceğini göstermiştir. Sonuçlar, bağlantılı veri üçlülerinin, metin sınıflandırması için zenginleştirilmiş belge temsilleri oluşturmakta etkili olduğunu ortaya koymuştur.

YAGO (Rebele ve diğ., 2016), Freebase (Bollacker ve diğ., 2007), DBpedia (Mendes ve diğ., 2011) ve Wikidata (Vrandečić ve diğ., 2014) gibi bilgi tabanları, metinlerden varlıklar ve ilişkiler çıkarmak için kullanılabilmektedir. Bu bilgi tabanları üzerinde çeşitli NLP görevleri için birçok çalışma yapılmıştır. Bu çalışmalar, metinlerdeki varlık ve ilişki çıkarma (Zhang ve diğ., 2021) veya bir bilgi tabanından uzak denetim kullanarak varlık ve ilişkileri etiketleme (Smirnova ve diğ., 2018) (Shang ve diğ., 2022) üzerine odaklanmaktadır. Uzak denetimle metinlerde varlık ve ilişki bilgisinin tespiti, metin sınıflandırma (TC) görevleri için de faydalı olabilir (Yang, 2019) (Tang ve diğ., 2021).

Tablo 1'de görüldüğü üzere, Türkçe TTC-3600 veri kümesi üzerinde hem geleneksel makine öğrenimi algoritmaları hem de derin öğrenme modelleri kullanılarak çeşitli metin sınıflandırma görevleri gerçekleştirilmiştir. Ayrıca, farklı dillerdeki performans karşılaştırması için, İngilizce haber makalelerinden oluşan 20Newsgroup ve BBC-News veri kümeleri de

literatürde kullanılmıştır. Bu çalışmaların yanı sıra, ilişkisel metin analizi için akademik yayınlar (Flisar, 2018), (Pittaras, 2021), (Lezama, 2022), (Nagumothu, 2021) da mevcuttur.

Literatür taraması neticesinde, bu tez kapsamında varlık tanıma ve ilgili bilgilerin bilgi tabanlarından edinilerek metinlerin iyileştirilmesi suretiyle sınıflandırma başarısının artırılması süreci ele alınmıştır. Araştırmanın temel hedefi, metin bazlı zeka araçlarının daha etkin kullanımını sağlayacak yöntemler geliştirmektir. BERT gibi gelişmiş modellemeler ile varlık tespiti gerçekleştirilmiş, SPARQL ve Wikidata üzerinden elde edilen ilişkisel bilgilerle metinler güçlendirilmiştir. Hem Türkçe hem de İngilizce veri setleri üzerinde yürütülen analizler, bu entegrasyonun sınıflandırma performansını olumlu yönde etkilediğini ortaya koymuştur. Bu çalışma, Türkçe metin işleme alanında varlık ve ilişki bilgilerini kullanarak yeni bir sınıflandırma metodolojisi sunmakta ve performans artışını somut örneklerle desteklemektedir. Ayrıca, bilgi tabanı inşası ve web kazıma gibi tekniklerle zenginleştirilmiş veri setlerinin oluşturulması, sınıflandırma işlemleri için yeni kaynaklar sunmaktadır. Bu bulgular, alanında yenilikçi adımlar olarak değerlendirilmekte ve ileriki çalışmalara ışık tutmaktadır.

2.3. Graf Tabanlı Modeller ve Uygulamaları

Metin sınıflandırması, Doğal Dil İşlemenin (DDİ) en önemli bileşenlerinden biri olarak, dil bazlı verileri anlama ve işleme süreçlerinde merkezi bir öneme sahiptir. Bu süreç, metinlerin etkin bir biçimde temsil edilmesini temel alır ve genellikle büyük veri setlerinin ve ileri düzey derin öğrenme tekniklerinin entegrasyonunu içerir. Özellikle BERT gibi ileri düzey modellerin bu alanda kaydettiği başarılar, dil işleme teknolojilerindeki gelişmelerin bir yansımasıdır. Büyük ölçekli veri setleri üzerinde eğitilmiş önceden hazırlanmış modeller, yeni model geliştirme süreçlerinde karşılaşılan zorlukları aşmada kritik bir rol oynar(Okur ve diğ., 2021).

Metin sınıflandırmasının uygulama yelpazesi geniş olup, sosyal medya analizlerinden sağlık sektörüne kadar pek çok alanda değerlendirilmektedir. Geleneksel yöntemlerin ötesine geçen modern yaklaşımlar, metin içerisindeki tokenlerin birbirleriyle olan ilişkilerini ve dizilimlerini derinlemesine analiz eder(Gasparetto ve diğ., 2022). Metinler arası ilişkileri modellemede graf yapılarının sağladığı avantajlar, bu tekniklerin NLP içerisindeki popülerliğini artırmakta ve metin sınıflandırması ile ilişkisel analiz çalışmalarında vazgeçilmez bir hale getirmektedir (Zhang ve diğ., 2020).

Son yıllarda, Doğal Dil İşleme (DDİ) için özellikle derin öğrenme (DL) alanında önemli başarılar elde edilmiştir (Nguyen ve diğ., 2019). Bu gelişmeler, bilgisayarların metinlerden özellikleri otomatik olarak öğrenip çıkarmada, metinsel özellik çıkarma sürecinde insan çabasını en aza indirgeyebileceklerini göstermektedir.

Metinler arası ilişkilerin anlaşılması ve analizi, NLP' nin en zorlayıcı yönlerinden biridir ve bu, özellikle ilişkisel metin sınıflandırma görevlerinde daha da belirgindir (Cervetti, 2020). Metinler arası ilişkiler, dilin doğasından kaynaklanan karmaşık ve çok katmanlı yapıları içerir; bu ilişkiler, metin içindeki ve metinler arasındaki bağlam, anlam ve niyeti ortaya çıkarır (Kowsari, 2019). Graf Sinir Ağları (GNN), bu tür karmaşık yapıları modellemek için giderek daha fazla tercih edilen bir yöntem haline gelmiştir. GNN' ler, metinlerdeki ve metinler arasındaki ilişkileri, graf tabanlı bir yapıda etkin bir şekilde temsil edebilir. Bu yaklaşım, metinler arasındaki doğrudan ve dolaylı bağlantıları, ayrıca kelime, cümle ve paragraf düzeyindeki ilişkileri görselleştirmekte ve analiz etmekte son derece başarılıdır (Wu ve diğ., 2023).

Bu bölümde, çalışmanın merkezinde yer alan metin sınıflandırması, ilişkisel metin sınıflandırması ve Graf Sinir Ağları (GNN) gibi anahtar konular üzerine literatürde yapılmış olan çalışmalar detaylı bir şekilde ele alınmış ve incelenmiştir. Bu inceleme sürecinde, söz konusu alanlardaki öncü ve etkili araştırmalar kapsamlı bir şekilde değerlendirilerek, çalışmanın konuları hakkında geniş bir perspektif sunulmuştur.

Malekzadeh ve arkadaşları (2021) çalışmasında, Graf Sinir Ağları (GNN) kullanımının literatürdeki yerini ve bu yaklaşımın metin sınıflandırmasındaki performansını ayrıntılı bir şekilde incelemiştir. Bu çalışma, GNN tabanlı yöntemlerin geleneksel derin öğrenme yaklaşımlarına kıyasla metin sınıflandırması ve çeşitli NLP görevlerinde nasıl etkili olduğunu gözler önüne sermiştir. Tablo 2' de ise, literatürde graf tabanlı metodolojileri kullanan çeşitli araştırmalar derlenmiştir.

Bu araştırmalar, metin sınıflandırma görevlerinde graf tabanlı yaklaşımların çeşitli veri setleri üzerindeki etkinliklerini ve performanslarını sergilemektedir. Tablo 2’ de bu konuda bulunan mevcut çalışmaların geniş bir yelpazesini sunmakta ve graf tabanlı metodolojilerin metin sınıflandırma alanındaki uygulamalarının kapsamlı bir özetini sağlamaktadır. Bu çizelgedeki çalışmaların detayları ve metodolojik yaklaşımları, bu bölümün devamında ayrıntılı bir şekilde ele alınmaktadır.

Tablo 2. Literatürdeki graf tabanlı metin sınıflandırma çalışmaları

Referans	**Model**	**Karşılaştırılan Modeller**	**Veri Seti**	**Sonuçlar** (Veriseti Sırası ile Doğruluk / F1-Skor %)
Huang ve diğ. (2019)	GCN+Soft max	CNN, LSTM, fastText, Graph-CNN, Text-GCN	R8, R52, Ohsumed	Doğruluk: 97.8, 94.6, 69.4
Yao ve diğ. (2019)	GCN	TF-IDF + LR LSTM, Bi-LSTM, PV-DBOW, PV-DM, PTE, FastText, SWEM, LEAM,Graph-CNN-C, Graph-CNN-S, Graph-CNN-F	20NG R8, R52, Ohsumed, MR	Doğruluk: 86.3, 97.1, 93.6, 68.4, 76,7
C. Liu ve diğ. (2022)	Document-relational GCN	Text GCN, Synonym augmented text GCN	20NG R8, R52, Ohsumed, MR	Doğruluk: 86.8, 96.9, 93.2, 65.9, 76.6
Chen ve diğ. (2022)	Relational-GCN	Decision trees, KNN, Naive Bayes, Logistic regression, LSTM, GRU, SGRU	OSHA dataset	F1-Score: (0.77)
X. Liu ve diğ. (2023)	Feature-based GCN	RNN, CNN, GCN, RGCN,R-BERT	SemEval-2010 Task 8, KBP37	Doğruluk: 89.75, 70.28

Huang ve arkadaşlarının (2019) gerçekleştirdiği çalışma, Graf Sinir Ağları (GNN) kullanılarak metin sınıflandırmasındaki yenilikçi uygulamaları ve bu tekniklerin mevcut zorlukların üstesinden gelmedeki etkinliğini incelemiştir. Geleneksel yöntemlerin aksine, her giriş metni için ayrı graf yapısı oluşturarak bellek kullanımını önemli ölçüde azaltmayı başarmışlardır. Araştırma, GNN modelinin, CNN, LSTM ve diğer yaygın modellere kıyasla üstün performans sergilediğini, R8, R52, Ohsumed veri setlerinde sırasıyla %97.8, %94.6, %69.4 doğruluk oranlarına ulaştığını ortaya koymuştur. Bu bulgular, GNN' nin metin sınıflandırma alanında önemli bir potansiyele sahip olduğunu göstermektedir.

Yao ve arkadaşlarının (2019) çalışması, metin sınıflandırması için Graf Konvolüsyonel Ağları (GCN) kullanımını incelemiştir. Araştırmada, kelime eşzamanlılığı ve doküman-kelime ilişkilerine dayalı bir metin grafiği oluşturularak Text GCN modeli eğitilmiştir. Bu model, dokümanlar için bilinen sınıf etiketleriyle kelime ve doküman gömülülerini ortaklaşa öğrenirken, harici kelime gömme veya ek bilgi gerektirmeden mevcut yöntemlere göre daha iyi performans gösterdiği bulunmuştur. Çalışma, 20NG, R8, R52, Ohsumed ve MR gibi çeşitli veri setlerinde yapılan deneylerle, Text GCN'nin diğer sınıflandırma modellerine kıyasla üstün

olduğunu ortaya koymuştur. Bu, GCN tabanlı yaklaşımın, global kelime ko-oluşumları ve doküman-kelime ilişkileri üzerinden etkili bir metin sınıflandırma modeli sağladığını gösterir.

Kumar ve arkadaşlarının (2022) yürüttüğü çalışma, sosyal medya ağlarındaki metinlerin sınıflandırılmasında Graf Sinir Ağı (GNN) teknolojisinin kullanımını derinlemesine inceler. Bu araştırma, sosyal medya platformlarında dolaşan büyük miktarda metin ve içeriğin doğruluğunu sınıflandırmada GNN'nin potansiyelini ortaya koyar. Araştırmacılar, metinleri iki boyutlu vektörler olarak işleyebilen GNN'nin, metin sınıflandırmasında geleneksel yöntemlere kıyasla üstün performans sergilediğini bulmuşlardır. Kendi kendini organize eden haritalar (SOM) kullanılarak en yakın komşular hesaplanmış ve bu süreç, metinler arasındaki gerçek mesafeleri belirleyerek sınıflandırmayı iyileştirmiştir.

Liu C. ve diğ. (2022), metin sınıflandırması için Graf Konvolüsyonel Ağları (GCN) kullanımını incelerken, dokümanlar arası ilişkileri içeren karmaşık ve zengin ilişki tabanlı matris grafikleri oluşturarak yüksek doğruluk elde etmiştir. Bu çalışmada, doküman-doküman ve kelime-kelime ilişkilerini kapsayan zengin matris grafikleri oluşturulmuş, bu sayede mevcut yöntemlere kıyasla daha yüksek doğruluk elde edilmiştir. Önerilen document-relational GCN modeli, TF-IDF doküman-doküman ilişkilerini içerecek şekilde tasarlanmış ve beş popüler veri seti üzerinde test edilerek, karşılaştırmalı modellere göre daha üstün sonuçlar göstermiştir. Özellikle 20NG, R8, R52, Ohsumed, MR veri setlerinde sırasıyla %86.8, %96.9, %93.2, %65.9, %76.6 doğruluk oranlarına ulaşılmıştır, bu da metin sınıflandırmada dokümanlar arası ilişkilerin gücünü ortaya koymaktadır.

Chen ve arkadaşlarının (2022) çalışması, kaza araştırma raporlarının metinlerinden kazaların nedenlerini ve süreçlerini analiz etmek amacıyla otomatik metin analizine odaklanır. Araştırma, ilişkisel graf konvolüsyonel ağ (R-GCN) ve önceden eğitilmiş BERT modeline dayalı yenilikçi bir metin madenciliği tabanlı kaza nedeni sınıflandırma yöntemi önermiştir. Bu yöntem, OSHA veri seti üzerinde gerçekleştirilen testlerde, karar ağaçları, KNN, Naive Bayes, Lojistik regresyon, LSTM, GRU ve SGRU gibi mevcut modellere kıyasla göze çarpan bir performans sergilemiş, özellikle F1-Skoru bakımından %77 gibi etkileyici bir sonuç elde edilmiştir.

Liu X. ve diğ. (2023) çalışması, doğal dil işleme görevlerinde ilişki sınıflandırmasının önemini vurgulamış ve ilişkisel metin sınıflandırma modeli önermiştir. Bu model, etiketlenmiş

Çince metin veri seti üzerinde yapılan deneylerle ve SemEval-2010 ve KBP37 gibi genel İngilizce veri setleri üzerinde gerçekleştirilen deneylerle test edilmiştir.

Bu bölümde metin sınıflandırması ve ilişkisel analiz alanlarındaki graf yapısının kullanımına ait mevcut araştırmaların geniş kapsamlı bir özetini sunulmuştur, bu da bu alanlardaki bilgi birikimini ve çeşitli araştırma yöntemlerini ayrıntılı bir şekilde gözler önüne sermiştir. Ancak, bu inceleme sırasında, özellikle Türkçe metinlerin sınıflandırılmasında ilişkisel graf sinir ağları kullanımı açısından belirgin bir boşluk fark edilmiştir. Tez çalışmasının bu kısmı, belirtilen boşluğu doldurmayı ve Türkçe metinler için yenilikçi ve etkili metodolojik yaklaşımlar geliştirmeyi hedeflemektedir.

Bu bağlamda, haber metinlerinin sınıflandırılması ve ilişkisel yapılarının derinlemesine analizi üzerine odaklanılmıştır. Türkçe'nin zengin ve karmaşık yapısını detaylı bir şekilde yansıtabilmek için özel olarak oluşturulan bir Türkçe haber metin veri seti kullanılmıştır. Temel olarak, haber metinleri arasındaki etiket bazlı ilişkiler, detaylı bir graf yapısı içinde modellenmiştir. Bu modellemede, düğümler haber metinlerini temsil ederken, kenarlar ise bu metinler arasındaki ortak etiketler temelinde kurulan ilişkileri ifade etmektedir. Çalışmada, metinlerin semantik gösterimleri için gelişmiş bir doğal dil işleme modeli olan Türkçe BERT kullanılmıştır. Bu sayede, metinlerin yüksek boyutlu ve anlamlı gösterimleri elde edilerek, doğal dil anlama yeteneği geliştirilmiştir. Ayrıca, metinler arası ilişkilerin sayısal bir formata dönüştürülmesi için, metin çiftlerinin kenar komşuluk matrisleri hesaplanmıştır. Bu yöntem, metinler arasındaki bağlantı sıklığını ve gücünü vektörler halinde kodlayarak, ilişkisel analizin daha detaylı yapılmasını sağlamıştır. Elde edilen zengin ve çok boyutlu graf vektör yapısı, bir graf sinir ağı (GNN) tabanlı sınıflandırma modeline beslenmiştir. Bu model, graf teorisini ve derin öğrenme tekniklerini bütünleştirerek, metinlerin hem sınıflandırılmasını hem de ilişkisel bağlamda anlamlandırılmasını hedeflemiştir.

Bu doğrultuda, geliştirilen yöntemler, gerçekleştirilen deneyler ve elde edilen sonuçlar derinlemesine ve ayrıntılı bir şekilde ele alınarak tartışılmıştır. Hedeflenen kapsamlı ve çok yönlü yaklaşım, haber metinlerinin sınıflandırılması ve ilişkisel yapıların çözümlenmesi alanında önemli bir potansiyel ortaya koymuştur.

3. YÖNTEM

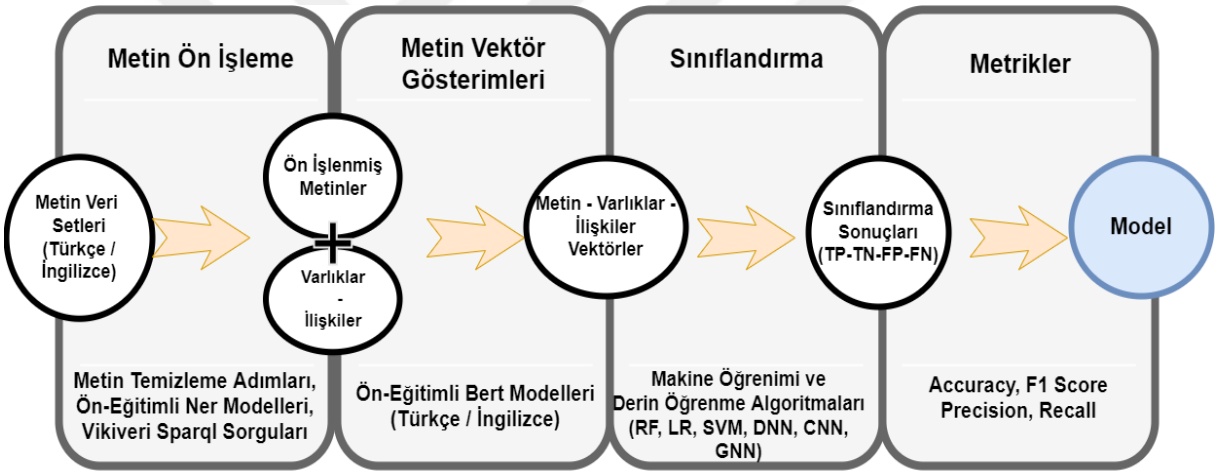
Metin sınıflandırmasında son yıllarda ilişkisel ve graf bazlı metodolojiler giderek önem kazanmaktadır. Bu metodolojiler, metinler arası ilişkileri ve dilin anlamsal yapılarını daha derinlemesine analiz etmek amacıyla ilişkisel veri yapıları ve graf teorisinden yararlanır. Özellikle Türkçe gibi zengin morfolojik ve söz dizimsel özelliklere sahip dillerde, literatürde bu yaklaşımlar yeterince derinlemesine ele alınmamıştır. Bu eksiklikten yola çıkarak, bu tez çalışması, Türkçe metinler için etiket bazlı ilişkisel graf yapısını kullanarak yeni bir sınıflandırma modeli geliştirmeyi amaçlamaktadır. Bu model, metinler arası etiket ve anlamsal bağlantıları modelleyerek, metin sınıflandırmasında doğruluk ve etkinlik sağlamayı hedefler. Bu yaklaşımın, metin sınıflandırma literatüründeki mevcut boşlukları doldurması ve Türkçe doğal dil işleme alanına yenilikçi katkılarda bulunması beklenmektedir.

Bu tez çalışmasının yöntem bölümünde, Türkçe haber metinlerinin sınıflandırılması ve ilişkisel yapıların analizi için kullanılan çeşitli metodolojik yaklaşımlar ayrıntılı bir şekilde incelenmektedir. Araştırmanın temel amacı, zengin morfolojik özelliklere sahip Türkçe metinler üzerinde etiket bazlı ilişkisel graf yapısını kullanarak daha doğru ve etkin bir sınıflandırma modeli geliştirmektir. Ayrıca dil farklılıklarının gözlemlenen etkisi, metodolojinin etkinliğini ve doğruluğunu arttırmak amacıyla Türkçe metin veri setlerinin yanı sıra İngilizce metin veri setleri üzerinde de benzer tekniklerin uygulanmasının, yararlı olabileceğini göstermektedir. Bu yaklaşım, farklı dil yapılarındaki sınıflandırma modellerinin genel uygulanabilirliğini ve adaptasyon kabiliyetini artıracaktır.

Türkçenin morfolojik zenginliği ve İngilizcenin söz dizimsel basitliği, her dilin benzersiz özelliklerine uygun varlık-ilişki entegrasyon stratejilerinin geliştirilmesini zorunlu kılar. Örneğin, Türkçe' deki "eğitim" (education) kelimesinin çeşitli çekimleri, 'eğitimi' (the education), 'eğitimde' (in education), ve 'eğitimden' (from education) gibi, varlıkların ve ilişkilerin daha karmaşık yapılar içinde nasıl ortaya çıktığını gösterir. Bu karmaşıklık, varlık-ilişki entegrasyon sürecinde dikkate alınmalı ve daha sofistike doğal dil işleme teknikleri gerektirir. Buna karşın, İngilizce' deki "education" kelimesi tipik olarak sabit bir forma sahiptir, bu da modellerin daha tutarlı ve doğru sonuçlar vermesini sağlar. Bu, İngilizce' de, özellikle geniş kategoriler içeren veri setlerinde sınıflandırma performansını artıran, daha basit ve doğrudan varlık- ilişki entegrasyonunu kolaylaştırır.

Bu çerçevede, yöntem bölümünde öncelikle kullanılan veri setlerini, metin ön işlem süreçlerini, ilişkisel yapının kurulmasını, modelleme için hazırlanma süreçlerini ve metin sınıflandırma modellerini detaylandırılmıştır. Metinlerin semantik temsili için kullanılan BERT gibi doğal dil işleme teknolojileri ve bu teknolojilerin entegrasyon süreci açıklanmaktadır. Modelin yapılandırılması, eğitilmesi ve test edilmesi süreçleri, kullanılan graf sinir ağı teknikleriyle birlikte ele alınmış, bu süreçlerin teknik detayları üzerinde durulmuştur. Ayrıca, modelin performansını objektif bir şekilde değerlendirebilmek için kullanılan metrikler ve karşılaştırmalı analiz yöntemleri detaylı bir şekilde izah edilmiştir. Bu metodolojik yaklaşımın, Türkçe metin sınıflandırma ve ilişkisel yapı analizi konusunda ne gibi yenilikler ve iyileştirmeler sunduğu, yöntem bölümünün kapsamlı bir değerlendirmesi ile takip edilmektedir.

3.1. Önerilen İlişkisel Metin Sınıflandırma Modeli



Şekil 1. Önerilen İlişkisel Metin Sınıflandırma Modeli

Bu bölümde, Türkçe metin veri setlerinde varlık tanımlama, uzaktan denetim ile varlık ilişkilerinin eklenmesi ve işaretlenmiş metinlerin sınıflandırılması üzerine odaklanılmıştır. Araştırmada kullanılan yöntemler, Wikidata gibi Wikipedia tabanlı bir veritabanından çıkarılan varlık-ilişki bilgileri içeren bir bilgi tabanı ile desteklenmiştir. Python araçları ve SPARQL sorgu dilli kullanılarak elde edilen bu bilgiler, sınıflandırma işlemleri için kullanılan metinlere entegre edilmiştir. Ayrıca, metinleri vektör temsillerine dönüştürmek için Türkçe ve İngilizce önceden eğitilmiş BERT modelleri (Schweter, 2020) tercih edilmiştir. Bu bölümde, makine öğrenimi ve derin öğrenme algoritmalarının kullanımı ile metin sınıflandırma performanslarının nasıl optimize edildiği ele alınacaktır. İlişkisel metin sınıflandırma modelimiz, bu tekniklerle zenginleştirilmiş metinler üzerinde uygulanmış ve elde edilen sonuçlar, algoritmalara özgü iyileştirmelerle birlikte tartışılacaktır.

Şekil 1 'de sunulan sınıflandırma modeli çerçevesinde, metinleri ön işleme ve vektör hale getirme süreçleri detaylı bir şekilde ele alınmıştır. Modelin bir sonraki adımı, derin öğrenme teknikleri ve makine öğrenimi algoritmalarını kullanarak vektörleştirilmiş metinleri sınıflandırmaktır. Bu adımda, metinlerin sınıflandırılması için destek vektör makineleri (SVM), lojistik regresyon (LR), derin sinir ağları (DNN), evrişimli sinir ağları (CNN) ve graf tabanlı sinir ağları (GNN) gibi algoritmalar kullanılmıştır. Bu algoritmalar, sınıflandırma aşamasında farklı yapısal ve ilişkisel özellikleri değerlendirebilmekte ve bunları sınıflandırma performansına dönüştürebilmektedir.

Bu çalışma, ayrıca, metin sınıflandırma yöntemleri arasında ilişkisel bilginin kullanım eksikliğini gidermeyi hedeflemektedir. İlişkisel bilginin sınıflandırma başarısına etkisi, önerilen modelin hem Türkçe hem de İngilizce veri setlerine uygulanması ve bu iki dilde sınıflandırma performanslarının karşılaştırılması ile incelenmiştir. Elde edilen sonuçlar, modelin doğruluk, F1 skoru, kesinlik ve yeniden çağırma gibi performans metrikleri üzerinden değerlendirilmiştir.

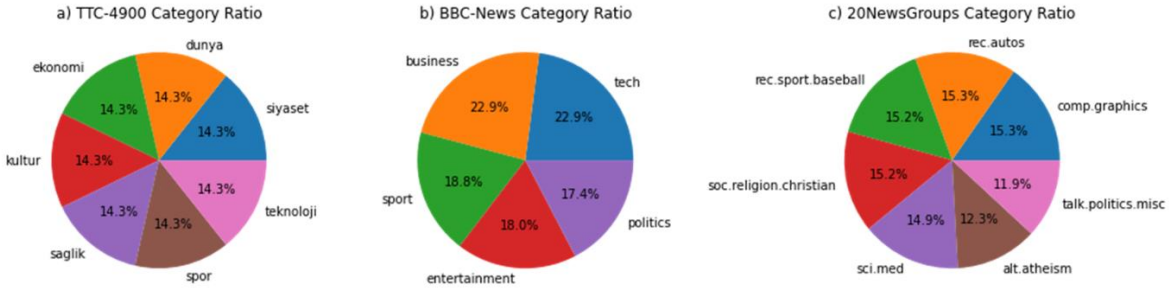
Son olarak, modelimiz, gerçek dünya verilerine uygulanarak elde edilen sonuçların yanı sıra, bu sonuçların literatürdeki mevcut çalışmalarla nasıl karşılaştırıldığı tartışılmıştır. Modelin uygulandığı Türkçe ve İngilizce metin veri setlerindeki performansı, sınıflandırma algoritmalarının etkinliği açısından önemli farklılıklar göstermiştir. Önerilen modelin, metin sınıflandırmada ilişkisel bilgi kullanımının önemini vurgulayan ve bu alandaki araştırmalar için yeni yollar açan sonuçlar doğurduğu görülmüştür. Ayrıca, bu çalışma, farklı dillerdeki metinlerin sınıflandırılması konusunda derinlemesine bir karşılaştırmalı analiz sağlayarak, dil özelliklerinin ve yapısal farklılıkların model performansı üzerindeki etkisini ortaya koymaktadır.

3.2. Kullanılan Veri Setleri

Bu tez çalışmasında kullanılan veri setleri, metin sınıflandırma modellerimizin eğitim ve değerlendirme süreçlerinde temel veri kaynağı olarak hizmet etmektedir. Türkçe ve İngilizce metinler üzerinde modelimizin genel performansını ölçmek ve karşılaştırmalı analiz yapabilmek için üç ana veri seti seçilmiştir: TTC-4900 (Kılınç ve diğ., 2017), BBC-News (BBC Datasets, 2006) ve 20Newsgroups (Lang, 2008).

Tablo 3. Kullanılan veri setleri hakkında bilgi.

Veriseti	Doküman Sayısı	Kategori Sayısı
TTC-4900	4900	7
BBC-News	2224	5
20Newsgroups	6517	7



Şekil 2. a) TTC-4900, b) TRT-Haber, c) 20NewsGroups veri setleri kategori dağılımları.

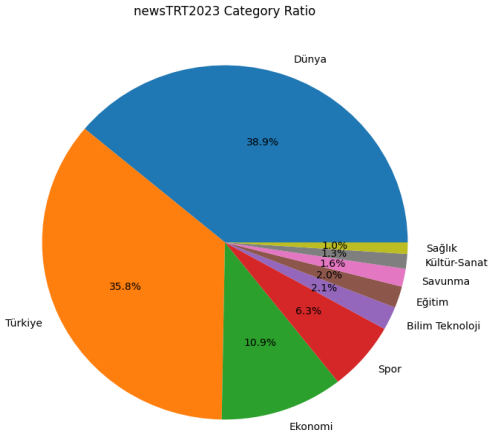
TTC-4900 Veri Seti: Bu veri seti, Türkçe metin sınıflandırma çalışmalarında sıklıkla kullanılan bir kaynaktır. Toplamda 4900 belge içeren bu set, 7 farklı kategoriye ayrılmıştır. Bu kategoriler arasında politika, spor ve teknoloji gibi alanlar bulunmaktadır. Veri setinin temizlenmesi ve ön işleme adımları, Türkçe'nin zengin morfolojik yapısına uygun detaylı teknikler kullanılarak yapılmıştır.

BBC-News Veri Seti: İngilizce metin sınıflandırma için kullanılan bu veri seti, BBC News tarafından sağlanan 2224 haber makalesini içermekte ve 5 ana kategori altında düzenlenmiştir. Bu set, daha resmi ve düzenli bir dil kullanımı sergileyen metinleri içerir, bu da modelimizin farklı dil yapılarında nasıl performans gösterdiğini değerlendirmek için idealdir.

20Newsgroups Veri Seti: Bu set, geniş kapsamlı bir İngilizce haber grubu veri setidir ve orijinalinde 18846 belge bulunmaktadır. Çalışmamız için bu set, 6517 belgeye indirgenmiş ve 7 kategoriye ayrılmıştır. Bu indirgeme ve kategorilendirme, çalışmamızın özel ihtiyaçları doğrultusunda yapılmıştır ve modelimizin geniş ve çeşitli veri kümeleri üzerindeki etkinliğini test etmek amacı taşır.

Ayrıca, bu çalışmada özellikle Türkiye ve dünya gündeminden toplanan haber içeriklerini kapsamlı bir şekilde analiz etmek amacıyla oluşturulan "**newsTRT2023**" (TRT Haber, 2023) adında detaylı bir Türkçe haber metni veri seti kullanılmıştır. Bu set, TRT Haber' in resmi dijital platformlarından web kazıma (web scraping) yöntemi ile elde edilen ve belirli bir tarih aralığını

kapsayan 4027 haber metninden oluşmaktadır. Haberler, 'Dünya', 'Türkiye', 'Ekonomi' gibi 9 farklı kategoriye ayrılmış ve her haber, yayın tarihi, başlık, metin, etiketler gibi çeşitli bilgilerle zenginleştirilmiştir. Bu veri seti, metin madenciliği ve doğal dil işleme tekniklerini kullanarak haber metinlerindeki yaygın kelimeleri, ana konuları ve ilişkisel bağları analiz etme imkanı sunmaktadır.



Şekil 3. newsTRT2023 Veri Seti Kategori Oranları

Açık kaynak olarak literatüre sunulacak olan 'newsTRT2023' veri seti, akademik araştırmalardan uygulamalı projelere kadar geniş bir kullanım alanına hitap eden zengin bir kaynak olarak ve zengin kategoride (Şekil 2) öne çıkmaktadır.

Tablo 4. newsTRT2023 veri setine ait örnekler

Haber Başlığı	İçerik (Metin)	Etiketler	Kategori
AKSUNGUR 30 bin feet yükseklikte uçuş gerçekleştirdi	Türk Havacılık Uzay Sanayii AŞ (TUSAŞ) tarafından yerli ve milli imkanlarla geliştirilen Aksungur silahlı insansız hava aracı (SİHA), TUSAŞ Motor Sanayii (TEI) tarafından üretilen … (haber içerik devami)	'TUSAŞ', 'AKSUNGUR', 'SİHA', 'Yerli ve Milli Teknolojiler'	Savunma
Geçen yıl en çok turist ağırlayan 5. ülke Türkiye oldu	Söz konusu raporda, turizm endüstrisi alanında değişen dinamiklerin ve trendlerin öncelikleri etkilediği belirtilirken … (haber içerik devami)	'gemi', 'Şampiyonlar Ligi', 'Muğla', 'dolar', 'Antalya', 'İran','turizm', 'Turizm', 'EY web sitesinden', 'Dubai'	Ekonomi

Medya içerik analizinden haber trendlerini izlemeye, kamuoyu araştırmalarından yapay zeka destekli haber sınıflandırma sistemlerine kadar çeşitli uygulamalar için idealdir. Bu veri seti, metin madenciliği ve doğal dil işleme tekniklerini kullanarak haber metinlerindeki yaygın

kelimeleri, ana konuları analiz etme imkanı sunar. Ayrıca, bu veri seti, makine öğrenmesi modellerinin eğitilmesi ve test edilmesi için de değerli bir kaynak olarak kullanılabilir. Veri setinden seçilmiş örnekler Tablo 4' te sunulmuştur, bu örnekler veri setinin çeşitliliğini ve zenginliğini göstermektedir. Bu örnekler, ilişkisel metin sınıflandırma görevlerinde kullanılmak üzere, veri setinin metin (text), varlık (entity) ve etiket (label) bilgilerini içermektedir. Varlıklar, metinleri graf olarak temsil etmek ve metinler arasındaki ilişkileri görselleştirmek için kullanılmıştır. Bu yaklaşım, veri setindeki metinler arasındaki ilişkisel bağlantıları anlamak ve sınıflandırmak için etkili bir metod sunmaktadır. Tablo 4, veri setinin nasıl işlendiğini ve ilişkisel metin sınıflandırmada kullanılacak temel öğeleri örneklendirerek göstermektedir.

Bu bölümde belirtilen veri setlerinin kullanımı, çalışmanın metodolojisine derinlemesine bir bakış açısı sunmakta ve Türkçe metin sınıflandırma modellerinin yanı sıra İngilizce metinler üzerinde de karşılaştırmalı bir değerlendirme yapılmasına olanak sağlamaktadır. Her bir veri setinin detaylı analizi, modelin dil özelliklerini ve çeşitli sınıflandırma görevlerindeki performansını anlamak için kritik öneme sahiptir.

3.3. Metin Ön-İşlem

Metin sınıflandırma çalışmalarında ön işleme süreci, model performansını doğrudan etkileyen bir adımdır. Metinlerin temizlenmesi, standardizasyonu ve işlenmesi, yapay zeka modellerinin daha verimli çalışmasını sağlar. Bu süreç, metinlerdeki gürültünün azaltılması, yazım hatalarının düzeltilmesi, argo ve kısaltmaların standart dile uygun hale getirilmesi gibi çeşitli işlemleri kapsar. Ayrıca, metinlerin tokenlara ayrılması, köklere indirgenmesi (stemming ve lemmatization) ve etkisiz kelimelerin (stop words) çıkarılması, metinleri daha temiz ve analiz için uygun bir hale getirir.

Metinler, sosyal medya, web içerikleri, iş dünyası kaynakları ve kullanıcı yorumları gibi çeşitli platformlardan elde edilebilir. Bu platformlar, metinlerin dil bilgisine uygunluk açısından büyük farklılıklar gösterebilir. Örneğin, sosyal medya metinleri genellikle konuşma diline daha yakın olduğundan daha fazla dil bilgisi hatası içerebilir, iş dünyası veya resmi metinler ise, dil bilgisi kurallarına daha sıkı uyulduğu için genellikle daha az hata içerir.

Belirtilen bu çeşitlilik, metinlerin kaynaklarına bağlı olarak temizleme süreçlerinin farklılaşmasını zorunlu kılar. Sosyal medya metinleri için, uygun olmayan kısaltmalar, argo

ifadeler veya mantıksız kelimelerin düzeltilmesi ya da çıkarılması gerekmektedir; oysa resmi metinlerde bu tür düzeltmelerin etkisi daha sınırlı olabilir. Bu ön işleme süreçleri ve teknolojik entegrasyonlar, Türkçe metin sınıflandırma ve ilişkisel yapı analizi alanında yapılan çalışmalara önemli yenilikler ve iyileştirmeler katmaktadır. Bu çalışmalar, metinlerin daha doğru ve etkin bir şekilde sınıflandırılmasını sağlayarak Türkçe Doğal Dil İşleme alanına değerli katkılarda bulunmaktadır.

Bu çalışmada, özellikle Türkçe metin sınıflandırma ve ilişkisel analiz üzerine odaklanılmıştır. Metinler, çeşitli ön işleme süreçlerinden geçirildikten sonra, geleneksel makine öğrenimi algoritmaları ve derin öğrenme modelleri kullanılarak işlenmektedir. Türkçe'nin zengin morfolojik yapısını ve metinler arası ilişkileri etkili bir şekilde modellemek amacıyla Grafik Sinir Ağları (GNN) ve Derin Sinir Ağları (DNN) gibi teknolojiler entegre edilmiştir. Bu teknolojiler, metinler arası karmaşık ilişkileri graf tabanlı yapılar kullanarak başarılı bir şekilde modellemek için tercih edilmektedir.

3.3.1. Metin Temizleme Adımları

Metin ve belge veri setleri, sıklıkla gereksiz kelimeler barındırabilir; bu durum, özellikle istatistiksel ve olasılık odaklı öğrenme algoritmalarını kullanırken sistem performansını olumsuz etkileyebilir. Bu bölüm, metin temizleme ve ön işleme yöntemlerine dair kapsamlı bir genel bakış sağlayacaktır.

Metinlerin Küçük Parçalara Ayrılması (Tokenization): Metnin, kelime, sembol, ifade veya diğer anlamlı unsurlar gibi daha küçük birimlere, yani tokenlara ayrılması sürecidir. Bu işlem, hem metin sınıflandırması hem de metin madenciliği uygulamaları için belgelerin ayrıştırılmasını zorunlu kılar. Örnek olarak;

"Eşim bana her zaman destek olur." cümlesi, parçalara ayrıldığında {"Eşim", "bana", "her", "zaman", "destek", "olur} şeklinde bir dizi oluşturur.

Metinlerin parçalara (token) ayrılması sürecinde, kelimenin kendisi, kökü, morfolojik yapısı ya da alt parçaları (karakter seviyesinde parçalama) gibi çeşitli yöntemler kullanılabilir. Al Nahas ve diğerleri (2020) tarafından yapılan bir çalışmada, Türkçe metinler üzerinde farklı tokenizasyon teknikleri kullanılarak derin öğrenme tabanlı sınıflandırma işlemi gerçekleştirilmiş ve bu çeşitli tekniklerin performans üzerindeki etkileri değerlendirilmiştir.

Etkisiz Kelimelerin Temizlenmesi (Stop Words Removal): Model performansını olumsuz etkileyebilen ve dilde sıkça kullanılan kelimelerin metin veri setlerinden çıkarılması sürecidir. Örneğin, Türkçede yaygın olarak kullanılan 've', 'ile', 'gibi' gibi kelimeler etkisiz kelime kategorisine girer.

Büyük-Küçük Harf Duyarlılığı (Capitalization): Metinlerdeki tutarsızlıkları gidermek amacıyla kelime veya küçük parça metinlerin (token) büyük-küçük harf kullanımını standart bir forma dönüştürmek sürecidir. Bu işlem, metin analizinde tutarlılığı artırır.

Argo ve Kısaltmalar (Slang and Abbreviations): Argo ifadeler ve kısaltmalar, metin ön işleme sürecinde ele alınması gereken önemli anormalliklerdir. Örneğin, "Türk Dil Kurumu" teriminin kısaltması olan "TDK" metin içinde yer alabilir. Ayrıca, argo ifadeler, gayri resmi konuşma dilinde sıkça kullanılan sözcüklerdir. Bu tür sözcüklerin yaygın işleme yöntemi, onları resmi dile (formal yapıya) çevirmektir.

Gürültü Temizleme (Noise Removal): Metinlerdeki özel karakterler, noktalama işaretleri ve benzeri gereksiz unsurların çıkarılmasını ifade eder. Bu unsurlar insanlar için anlam taşısa da, sınıflandırma algoritmaları açısından zararlı olabilirler.

Yazım Hatalarını Düzeltme (Spelling Correction): Yazım yanlışlarının düzeltilmesi, sosyal medya gibi metinlerinde çok karşılaşılan bir durumdur. Doğal dil işleme (NLP) alanında, bu sorunu çözmek için çeşitli algoritmalar ve teknikler geliştirilmiştir (Bölücü, 2019).

Eklerden Arındırma (Stemming): Doğal dil işlemede (NLP), aynı anlamı taşıyan ancak farklı biçimlerde olan kelimelerin, temel formuna indirgenmesi sürecidir. Örneğin, Türkçe'de 'adaletli' kelimesinin eklerden arındırılmış hali 'adalet'tir. İngilizce metinler için Snowball stemmer (Porter, 2001) ve NLTK kütüphanesi (Perkins, 2010) kullanılırken, Türkçe metinler için Snowball'ın Türkçe versiyonu (Çilden, 2006) ve Zemberek Kütüphanesi (Akın, 2007) tercih edilebilir.

Kök Bulma (Lemmatization): Bu işlem, bir kelimenin temel formunu (lemma) elde etmek için son eklerin değiştirilmesi veya kaldırılması üzerine kurulmuş bir doğal dil işleme (NLP) sürecidir. Kelimenin kökünü belirleme aşamasını içerir. NLTK ve Zemberek kütüphaneleri, kök bulma işlemleri için sıklıkla tercih edilen araçlardandır.

Kılınç ve arkadaşları tarafından 2017 yılında yürütülen çalışmada, TTC-3600 Türkçe metin sınıflandırma veri seti tanıtılmış ve metinlerin sınıflandırılmadan önce eklerden temizlenmesi veya köklerinin çıkarılmasının performans üzerindeki etkisi değerlendirilmiştir. Çalışmada, metinlerdeki kelimeler farklı yöntemlerle (ham halleriyle, frekansı koruyan örnekleme (FPS) ile kelime sayısının azaltılmasıyla veya Zemberek kütüphanesi kullanılarak köklerin bulunmasıyla) işlenmiştir. Bu çeşitli ön işleme tekniklerinin metin sınıflandırma başarısına olan etkileri araştırılmıştır. Bu tür çalışmalar, kök çıkarma ve ek temizleme işlemlerinin metin sınıflandırmadaki önemini ve etkisini ortaya koymaktadır.

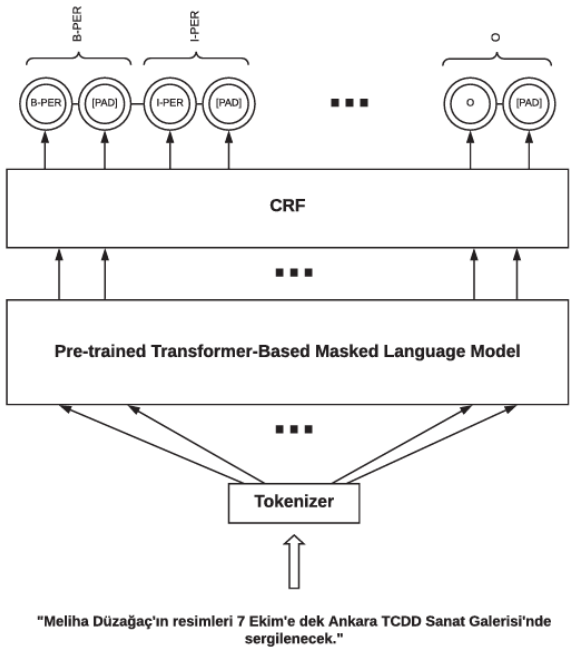
Morfolojik yapılar: Morfoloji, dil biliminin bir dalıdır ve kelimelerin farklı biçimlerini, kelime türlerini inceleyerek sözcük yapısını ele alır. Bu disiplin, özellikle morfolojik olarak zengin dillerde, dilin yapısını derinlemesine anlamak için büyük bir önem taşır.

Güngör ve arkadaşları (2019) tarafından gerçekleştirilen araştırmada, adlandırılmış varlık tanıma sürecinde morfolojik analizin değerlendirilmesi hedeflenmiştir. Bu çalışmada, çeşitli dillerden veriler tek bir doğal dil işleme modeli altında incelenmiş, kelime ve karakter seviyesindeki gömmelerin (embeddings) yanı sıra morfolojik gömmeler kullanılarak dil temsilleri oluşturulmuştur. Türkçe, Çekçe, Macarca, Fince ve İspanyolca gibi dillerde yapılan deneyler, bu dillerin morfolojik yapılarının karşılaştırmalı analizlerini sunmaktadır. Bu tip araştırmalar, farklı dillerin morfolojik özelliklerinin doğal dil işleme modellerine entegrasyonunun önemini ve bu entegrasyonun dil modellemesi üzerindeki etkisini ortaya koymaktadır.

3.3.2. İsimlendirilmiş Varlık Tanıma (NER - Named Entity Recognition)

İsimlendirilmiş Varlık Tanıma (NER - Named Entity Recognition), metin içindeki isimlendirilmiş varlıkları önceden belirlenmiş varlık kategorilerine göre bulma ve sınıflandırma sürecidir. NER, metin parçalarında etiketlenmiş varlıkları tanır ve bunları kişi, yer ve organizasyon gibi önceden belirlenmiş kategorilere ayırır. NER, çeşitli doğal dil uygulamalarına dayanarak soru cevaplama, metin özetleme, makine çevirisi ve metin sınıflandırma gibi görevler için kullanılabilir. NER sistemleri makul tanıma doğruluğu sağlamada başarılı olsa da, genellikle kuralları veya özellikleri dikkatli bir şekilde tasarlamak için çok fazla insan çabası gerektirir. İngilizce için etiketlenmiş corpus veri kümeleri olarak OntoNotes (Hovy ve diğ., 2006)ve Winer (Ghaddar, 2017) bulunmaktadır. Türkçe çok katmanlı corpus yayınında (Yıldız ve diğ., 2018), İngilizce-Penn-Treebank'ten 9600 cümle ve kelimelerin

morfolojileri, isimlendirilmiş varlıklar, anlamlar ve anlamsal rol etiketleri içerir şekilde tanıtılmıştır. Ayrıca, CRF tabanlı bir Türkçe NER (Şeker ve diğ., 2017) yaklaşımı da vardır.



Şekil 4. Transformer tabanlı NER modeli (Aras ve diğ., 2021)

Son yıllarda, transformer sinir ağı tabanlı BERT (Devlin ve diğ., 2017) modeli ve önceden eğitilmiş modellerin NER uygulamalarında kullanıldığı transfer öğrenme yaklaşımı kullanılmaktadır. Bu yaklaşımla, BERT modeli kullanılarak bir veri kümesinde varlık kelimeleri içeren bir temsil yapılabilir. Büyük bir korpus üzerinde önceden eğitilmiş transformer modelleri kullanarak, geniş bir kelime yelpazesini kapsayan temsiller oluşturur. Önceden eğitilmiş bir NER modeli kullanarak metinlerdeki varlıkları elde etmek, performans ve zaman maliyetlerini azaltır. Şekil 4' te görüldüğü üzere Türkçe için de BERTurk modeli (Aras ve diğ., 2021) gibi önceden eğitilmiş modeller mevcuttur. Bu gelişmeler, NER'in doğal dil işlemedeki rolünü ve etkinliğini önemli ölçüde artırmaktadır (Okur ve Sertbaş, 2021).

Sonuç olarak, İsimlendirilmiş Varlık Tanıma (NER) teknolojisinin gelişimi, doğal dil işleme alanında dikkate değer bir gelişmedir. Yapay zeka ve makine öğrenimi modellerinin evrimi, özellikle BERT gibi transformer tabanlı yaklaşımlar sayesinde, NER uygulamaları daha da hassas ve etkin hale gelmiştir. Bu ilerlemeler, metin bazlı veri kaynaklarından maksimum bilgi çıkarma yeteneğini artırarak, dil modellerinin kapsam ve doğruluğunu genişletmektedir. Türkçe dahil olmak üzere birçok dil için geliştirilen önceden eğitilmiş modeller ve veri setleri,

NER'in uygulama alanlarını genişletiyor ve daha fazla dilde yüksek kaliteli doğal dil işleme çözümlerinin önünü açıyor. Özellikle, metin sınıflandırma gibi alanlarda NER teknolojisinin kullanımı, verileri daha iyi anlama ve yönetme yeteneğimizi önemli ölçüde artırmaktadır, bu da NER'i geleceğin teknoloji ekosistemlerinde vazgeçilmez bir araç haline getiriyor.

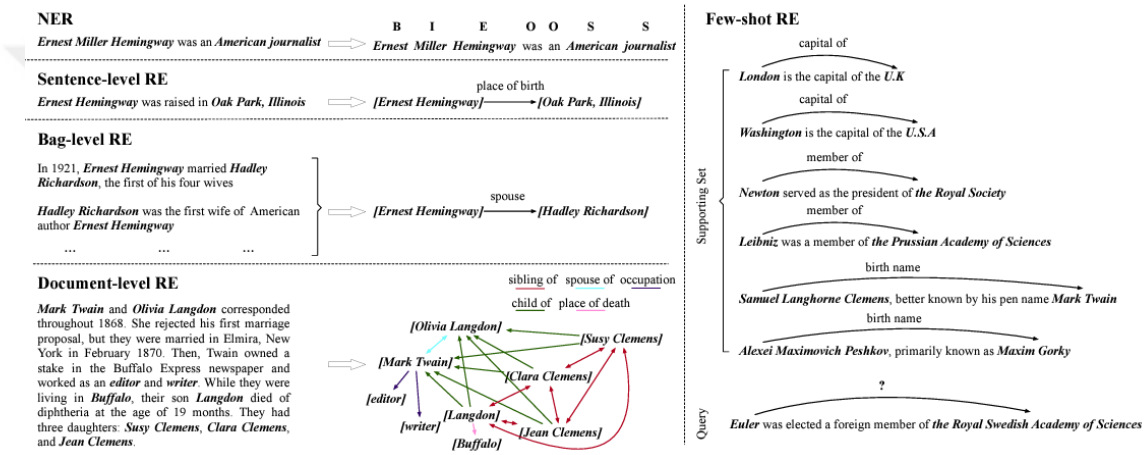
3.3.3. İlişki Çıkarımı (Relation Extraction)

İlişki (relation), iki veya daha fazla varlık arasındaki bağlantı olarak tanımlanabilir. Örneğin, iki birey arasında bir evlilik ilişkisi olabilir veya bir kişi ile bir organizasyon arasında üyelik veya sahiplik ilişkileri bulunabilir. **İlişki Çıkarma (Relation Extraction - RE)**, metin içindeki varlıklar arasındaki ilişkileri bulmak için kullanılan bir yöntemdir. Bu süreç, metinden kişi, yer, organizasyon, nesne ve fiil yapıları gibi varlıkların çıkarılmasını ve ardından bu varlıklar arasında anlamlı bağlantılar kurulmasını içerir. Özne ve nesne arasındaki konu-ilişki-nesne ilişki yapısı genellikle S-R-O (subject – relation - object) formatında ifade edilir. Örneğin, "Ankara, Türkiye'nin başkentidir" cümlesinde Ankara ve Türkiye arasındaki başkent ilişkisi (Ankara – Başkent – Türkiye) olarak modellenebilir. Amaç, metin içerisindeki varlıklar arasında bağlantılar kurarak, bu bilgilerden yararlanabilecek şekilde bilgi tabanları oluşturmaktır.

Varlıklar arasındaki ilişkiler, vektör formuna dönüştürülerek klasik makine öğrenimi veya derin öğrenme algoritmaları kullanılarak sınıflandırılabilir (Ren, 2018). Son zamanlarda yapılan çalışmalarda bu ilişki yapıları genellikle bir graf olarak temsil edilmekte ve graf sinir ağları ile sınıflandırılmaktadır. Graf üzerindeki her bir düğüm, kelime varlıkları veya belgelerin kendisi olarak düşünülebilir. Graf düğümleri arasındaki bağlantılar, varlık veya belgeler arasındaki ilişkileri temsil eder ve graf sinir ağları bu ilişkilerin sınıflandırılmasında kullanılmaktadır (Han, 2018).

İlişki çıkarımı süreci, doğal dil işleme (NLP) alanında iki temel alt kategoriye ayrılır ve bu kategoriler, metinlerin derinlemesine analizinde önemli bir rol oynar (Sahu, 2019). Bu iki ana alt kategori, **cümle seviyesinde ilişki çıkarımı** ve **belge seviyesinde ilişki çıkarımı**dır. Cümle seviyesinde ilişki çıkarımı, metinlerin cümle bazında ayrıştırılarak cümle içindeki varlıklar arasındaki ilişkilerin tespit edilmesine odaklanırken, belge seviyesinde ilişki çıkarımı, bireysel belgeler arasındaki ilişkilere dikkat çeker ve bu ilişkiler üzerinden belgeleri sınıflandırmaya veya etiketlemeye yarar. Bu iki yöntem, NLP' nin çeşitli uygulamalarında kritik öneme sahiptir.

Cümle seviyesinde ilişki çıkarımı, metinlerin cümle bazında incelenmesi ve bu cümleler içindeki varlıklar arasındaki ilişkilerin belirlenmesi üzerine kuruludur. Bu işlem sırasında, cümle içinde doğal olarak mevcut olan ilişkiler tespit edilebilir ya da cümle içinde herhangi bir ilişki olmaması durumunda, önceden tanımlanmış ilişkilere sahip varlıkların bir bilgi tabanından (**knowledge-based** - KB) otomatik ilişki etiketleri kullanılabilir. Bu bilgi tabanı, potansiyel ilişkileri içeren varlıkların bir koleksiyonu olarak düşünülebilir. Bu tür bir yöntem **uzaktan denetim (distant supervision)** olarak adlandırılır ve çeşitli araştırmalar tarafından desteklenmektedir (Ji, 2017) (Han, 2019). Şekil 5'te, uzaktan denetim (distant supervision) yöntemi kullanılarak tüm ilişki çıkarımı senaryolarına ait örnekler görülmektedir.



Şekil 5. Uzaktan denetim (distant supervision) yöntemi ile tüm ilişki çıkarımı senaryolarının örnekleri (Han, 2019).

Uzaktan denetim yöntemi kullanıldığında, metin içerisindeki varlıklara otomatik olarak atanan ilişkiler bazı problemlere neden olabilir. Bunlardan biri, varlıkların yanlış etiketlenmesi sorunudur. Örneğin, bir metin ekonomiyle ilgili ise ve içerisinde ekonomiye dair önemli bir isim geçiyorsa, bu isme yanlışlıkla siyaset gibi farklı bir alandan etiketler atanabilir. Bu durum yanlış etiketlemeye sebep olur ve bu yanlış etiketlerin temizlenip düzeltilmesi gereklidir (Ru, 2018). Cümle seviyesinde yapılan ilişki çıkarımı, metin içerisindeki varlıklar arası ilişkileri daha ayrıntılı ve hassas bir şekilde incelemeye olanak tanır, böylece doğal dil işleme uygulamalarında kritik bir işlevi yerine getirir. Uzaktan denetim yöntemi, büyük veri setleriyle çalışırken zaman ve kaynak tasarrufu sağlasa da, dikkatli bir şekilde uygulanmalıdır.

Belge seviyesi ilişki çıkarımı, bireysel belgeler arasındaki ilişkileri inceleme üzerine odaklanır. Bu tür ilişkiler, akademik yayınlarda gözlemlenen atıflar gibi belgeler arasında oluşturulan belirli bağlantılar veya referanslar şeklinde ortaya çıkabilir. Ayrıca, belgelerin

başlıkları veya özetlerinden elde edilen etiketler ve nesneler de ilişki çıkarımında kullanılabilir. Son dönemde bu alandaki çalışmalar, artan belge ve doküman miktarının otomatik olarak etiketlenmesi veya sınıflandırılması gibi konular üzerinde yoğunlaşmaktadır (Yao, 2019) (Han, 2019). Bu yaklaşım, belge içeriklerini daha iyi anlamak ve organize etmek için değerli bir yöntem sunar.

Belge seviyesi ilişki çıkarımının temel amacı, büyük miktarda belgeyi otomatik olarak kategorize etmek ve içlerinden anlamlı bilgileri ayıklamaktır. Bu süreç, metinleri daha geniş bir perspektiften analiz etmeye imkan tanıyarak, belgelerin içeriği hakkında daha derinlemesine bilgi sağlar. Belge seviyesindeki ilişki çıkarımı, bilgi yönetimi, bilgi elde etme ve metin madenciliği gibi alanlarda kritik bir öneme sahiptir ve büyük veri kümelerinin etkin bir şekilde işlenmesine yardımcı olur. Bu metodoloji, metin bazlı verilerin daha kapsamlı bir şekilde incelenmesini mümkün kılarak, doğal dil işleme ve metin analizi alanlarında yenilikler sunar.

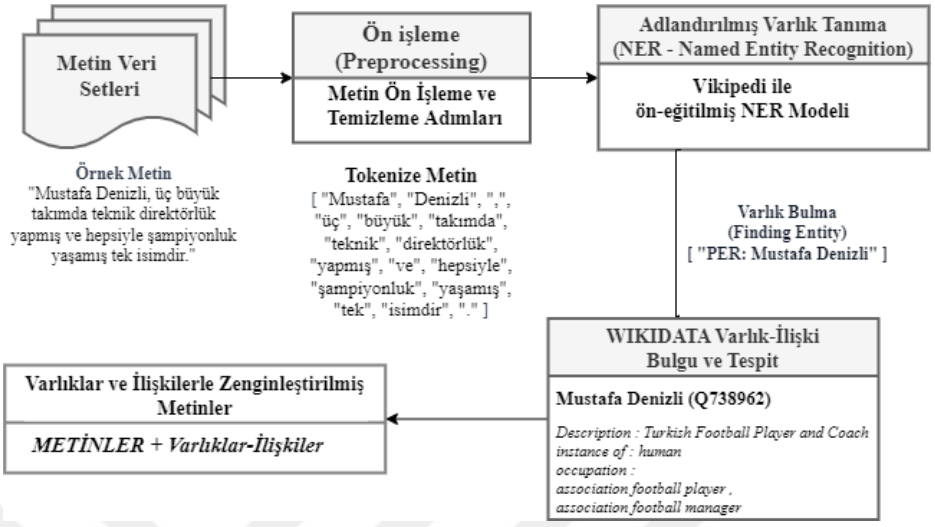
Özetle, ilişki çıkarımı, doğal dil işleme teknolojilerinin merkezinde yer almakta ve metinlerdeki varlıklar arasındaki ilişkileri belirlemek için hayati bir araç olarak hizmet etmektedir. Cümle ve belge seviyesindeki ilişki çıkarımı teknikleri, büyük ve karmaşık veri setlerinden anlamlı bilgiler çıkarılmasını sağlar, bu da bilgi tabanlarının genişletilmesi, metin analizi ve daha geniş çapta bilgi yönetimi için kritik öneme sahiptir. Bu tekniklerin gelişimi, yapay zekâ ve makine öğrenimi uygulamalarının daha da ileri götürülmesine olanak tanıyarak, bilgisayarla dil işlemenin sınırlarını genişletmektedir.

3.4. Metinlerde Varlık ve İlişkilerin Bulunması

Bu bölümde, metinlerdeki varlıkların ve ilişkilerin nasıl tespit edildiği (Şekil 6) ve bu sürecin ilişkisel metin sınıflandırması üzerindeki etkileri detaylı bir şekilde incelenmektedir. Öncelikle, metin ön işleme adımları ile veri setlerinden gürültünün arındırılması, yazım hatalarının düzeltilmesi, dil bilgisel tutarlılığın sağlanması gibi süreçler gerçekleştirilmiştir ve bu süreçler, veri kalitesini artırarak sınıflandırma modellerinin daha etkili çalışmasını sağlamıştır.

Ardından, Named Entity Recognition (NER) teknikleri kullanılarak metinlerdeki varlıkların nasıl tanımlandığı ve sınıflandırıldığı ele alınmıştır. NER, metin içindeki isimlendirilmiş varlıkları (kişi, yer, kurum gibi) tanımlayıp kategorilere ayırarak, bu varlıklar

arasındaki ilişkilerin belirlenmesi sürecini kolaylaştırmıştır. Bu süreç, metinler arası bağlantıları ve yapıları daha iyi anlamamıza yardımcı olmuştur.



Şekil 6. Metinlerdeki varlıkları (entities) bulma ve Vikiveri varlık ilişkisi (entity-relation) tespiti.

Son olarak, Relation Extraction (RE) yöntemlerinden faydalanılarak, tanımlanan varlıklar arasındaki ilişkilerin nasıl çıkarıldığı açıklanmıştır. RE, varlık çiftleri arasında var olan veya olası ilişkileri keşfetmek için kullanılmıştır ve bu, özellikle karmaşık veri setlerinde bilgi erişimi ve bilgi yönetimi açısından kritik öneme sahip olmuştur.

Bu bölüm, metin sınıflandırma ve ilişkisel analiz çalışmalarında kullanılan yöntemlerin temel prensipleri ve uygulamaları üzerine odaklanarak, metinlerden maksimum bilgiyi çıkarma stratejilerini derinlemesine tartışmaktadır. Bu süreçler, özellikle büyük ve çeşitli veri setlerinin analiz edilmesinde, elde edilen bulguların doğruluğunu ve derinliğini artırarak, daha bilinçli ve verimli kararlar alınmasına imkan tanımıştır.

3.4.1. NER Modeli İle Varlıkların Tespiti

Bu çalışmada, Named Entity Recognition (NER) görevi için Türkçe metinlerde "savasy/bert-base-turkish-ner-cased" ve İngilizce metinlerde "Jean-Baptiste/roberta-large-ner-english" adlı ön eğitimli dönüştürücü modeller kullanılmıştır. Her iki model de açık kaynaklıdır ve geniş bir kullanım alanına sahiptir. "savasy/bert-base-turkish-ner-cased" modeli, Wikipedia sayfalarından çıkarılan LOC (yer), PER (kişi) ve ORG (organizasyon) etiketlerine sahip varlıkları içeren 176 dilde mevcut olan çok dilli WikiANN veri seti kullanılarak eğitilmiştir. Bu veri seti, hem Türkçe hem de İngilizce dillerinde varlık elde etmek için bir bilgi tabanı olarak

hizmet etmiştir. Dolayısıyla kullanılan her iki model de, geniş çapta kullanılan WikiANN veri seti ile eğitilmiş mevcut açık kaynaklı modellerdir. Bu yaklaşım, tanınmış ve güvenilir kaynaklardan faydalanarak NER görevinde yüksek doğruluk ve tutarlılık elde etmemizi sağlamıştır.

Bu modellerin kullanımı, metinlerimizdeki varlıkların doğru bir şekilde tanımlanmasını ve sınıflandırılmasını mümkün kılmış, böylece metinler arasındaki ilişkilerin daha etkin bir şekilde analiz edilmesine olanak tanımıştır. Ayrıca, bu modellerin sağladığı yüksek doğruluk oranları, sonraki analiz aşamalarının güvenilirliğini artırmış ve daha geniş kapsamlı ilişkisel analizler için sağlam bir temel oluşturmuştur.

3.4.2. Varlıklara Ait İlişkilerin Elde Edilmesi

Bu çalışmada, metinlerdeki varlıkların tanımlanmasının ardından, bu varlıkların potansiyel ilişkilerle zenginleştirilmesi süreci gerçekleştirilmiştir. Varlık bilgi tabanı modeli oluşturmak için Wikipedia' ya ait çok dilli bir korpus olan Wikidata üzerinde SPARQL sorgu dili kullanılmıştır. Bu sorgulama yöntemiyle, Wikidata veritabanına yapılan sorgular sonucunda, varlıklar için bir Wikidata ID' si döndürülmekte, bu ID ise bireyler için meslek gibi çeşitli nitelik ilişkilerini içeren sayfalara erişim sağlamaktadır. Bu süreç, SPARQL sorguları aracılığıyla ilgili varlıkların çıkarılmasını ve Wikidata' nın bir bilgi tabanı (KB) olarak kullanılmasını etkinleştirir, böylece tespit edilen varlıklar Wikidata KB' den ek ilgili varlıklarla zenginleştirilir.

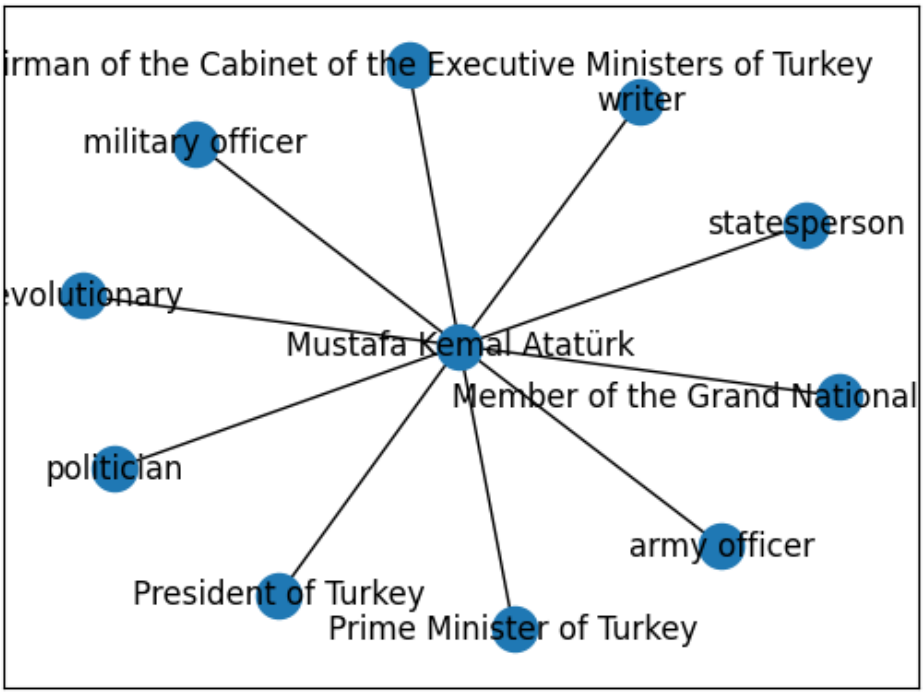
Uzaktan denetim (distant supervision), Wikidata gibi veritabanları kullanılarak otomatik etiketleme için bir yöntem olarak kullanılmaktadır ve büyük veri kümelerini zenginleştirmeyi sağlar. Diğer yandan, 'varlık bağlama' işlemi, metinlerde tanımlanan varlıkların dış kaynaklardaki benzersiz tanımlayıcıları ile eşleştirilmesi sürecini ifade eder. Çalışmamızda, varlık bağlama, metinlerdeki varlıkları Wikidata' daki karşılık gelen girişlerle ilişkilendirmek için kullanılmış, uzaktan denetim ise bu varlıklar ve ilgili ilişkilerin otomatik olarak etiketlenmesi için kullanılmıştır.

```
# Varlık listesinden ilişki bilgilerini almak için fonksiyonu tanımla
def get_EntitiesRelationsList(entityList):
    # SONUÇLARI SAKLAMAK İÇİN BOŞ BİR DATAFRAME VE METİN DİZESİ OLUŞTUR
    df = pd.DataFrame(columns=['source', 'target', 'relation', 'description'])
    df_entity_relations_text = ""

    # HER VARLIK İÇİN İŞLEM YAP
    for entity in entityList:
        searchEntity = "\"" + entity + "\""  # VARLIK ARAMA DİZESİ OLUŞTUR

        try:
            # WIKIDATA SPARQL ENDPOINT URL'Sİ
            endpoint_url = "https://query.wikidata.org/sparql"

            # VARLIK ARAMA SORGUSU OLUŞTUR
            query_entity_search = """
                # VARLIK ARAMAK İÇİN WIKIDATA'DA SPARQL KULLANILIR.
                # BELİRLİ BİR VARLIK ADINI (ENTITY) VE DİL KODUNU (LANG) ALARAK,
                # İLGİLİ VARLIĞIN WIKIDATA'DAKİ TANIMLARINI VE AÇIKLAMALARINI ARAR.
                # ARAMA SONUÇLARI, VARLIĞIN BİR KİMLİK NUMARASI (ITEM), ETİKETİ (ITEMLABEL)
                # VE OPSİYONEL OLARAK BİR AÇIKLAMA (DESCRIPTION) İÇERİR.
            """

            # SONUÇLARI AL VE VARLIK BİLGİLERİNİ ELDE ET
            results_entity_search = get_results(endpoint_url, query_entity_search)
            # VARLIK BİLGİLERİNİ İŞLE

            # İLİŞKİ SORGUSU OLUŞTUR
            query_relations = """
                # İLGİLİ VARLIĞIN WIKIDATA'DAKİ İLİŞKİLERİNİ ARAMAK İÇİN SPARQL KULLANILIR.
                # BELİRLİ BİR VARLIK KİMLİĞİ (SEARCHID) VE DİL KODUNU (SEARCHLANG) ALARAK,
                # İLGİLİ VARLIĞIN WIKIDATA'DAKİ İLİŞKİLERİNİ VE AÇIKLAMALARINI ARAR.
                # İLİŞKİ SONUÇLARI, VARLIK ADI (WDLABEL), İLİŞKİ ADI (PS_LABEL) VE
                # İLİŞKİ AÇIKLAMALARINI (DESCRIPTION) İÇERİR.
            """

            # SONUÇLARI AL VE İLİŞKİ BİLGİLERİNİ ELDE ET
            results_relations = get_results(endpoint_url, query_relations)
            # İLİŞKİ BİLGİLERİNİ İŞLE

        except:
            continue  # HATA DURUMUNDA DEVAM ET

    return df, df_entity_relations_text
```

Şekil 7. Wikidata'dan varlık-ilişki bilgilerini almayı gösteren sözde kod

Şekil 7' te gösterilen bu fonksiyon (***def get_EntitiesRelationsList(entityList):***), belirli bir varlık listesinde bulunan her bir varlık için ilişki bilgilerini çekmek amacıyla kullanılır. İşlevi şu adımları içerir: İlk olarak, ilişki bilgilerini saklamak için bir veri yapısı oluşturulur. Her bir varlık için, ilgili varlığın tanımları ve ilişkileri Wikidata'dan çekilebilmek üzere özel SPARQL sorguları oluşturulur ve gönderilir. Alınan sonuçlar işlenir, varlık adları arasındaki ilişkiler belirlenir ve ilişki açıklamaları dikkate alınarak uygun bir formata dönüştürülür. Hata durumunda işlem devam ettirilir ve bir sonraki varlık için işlem tekrarlanır. Tüm varlıklar için işlemler tamamlandıktan sonra, elde edilen ilişki bilgileri ve açıklamalar birleştirilir ve dönüş değeri olarak verilir. Bu dönüş değeri, varlık listesindeki her bir varlığın Wikidata'daki ilişkilerini ve ilişki açıklamalarını içerir.

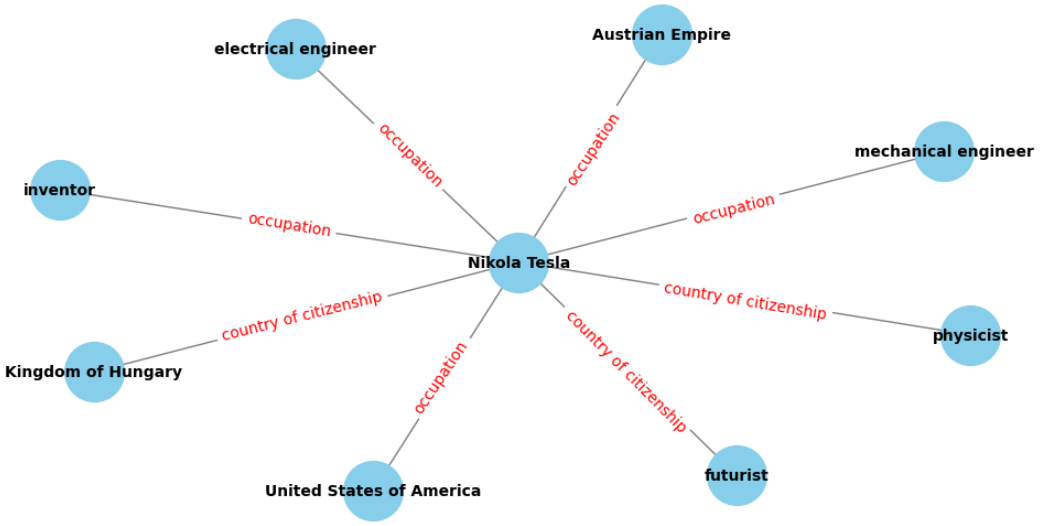
Metodolojimiz, Wikidata'dan alınan yapılandırılmış bilgileri yapılandırılmamış verilerle bütünleştirir, varlıkların tanımları, ilgili metinler ve yapılandırılmış üçlülerin ötesine geçen diğer açıklayıcı bilgileri kapsar. Yapılandırılmış ve yapılandırılmamış verilerin bu karışımı, Kepler (Wang ve diğ., 2021) modeline benzer bir bütünleştirme yaklaşımı izler ve varlıklar ile ilişkiler hakkında daha derin bir anlayış sağlar. Kepler modeli, bilgi gömme ve dil modelleme tekniklerini birleştirerek, yapılandırılmış bilgi tabanı bilgileri ile yapılandırılmamış metin verilerinin bütünleştirilmesini kolaylaştırır ve böylece varlık-ilişki verilerinin zenginliğini ve bağlamını artıran kapsamlı bir çerçeve sunar.



Şekil 8. Varlık\Entity (Mustafa Kemal Ataturk) and İlişkiler\Relations

Entity: Mustafa Kemal Ataturk – WikidataID: Q5152: (Mustafa Kemal Ataturk – occupation - military officer - member of an armed force or uniformed service who holds a position of authority), (Mustafa Kemal Ataturk - occupation - statesperson - civil servant or politician in high government offices), ... (Şekil 8)

Örnek olarak, "Mustafa Kemal Atatürk" veya "Atatürk" ismi, önceden eğitilmiş bir NER modeli kullanılarak tespit edilen bir metni ele alalım. Tespit edilen varlık ismi "Atatürk" için Wikidata üzerinde bir SPARQL sorgusu gerçekleştirilir. Eşleşen bir sonuç bulunursa, bu varlık ismine karşılık gelen Wikidata ID değeri "Q5152" olarak sorgu sonucunda elde edilir. Ardından, bu varlık (entity) ID değeri, "Mustafa Kemal Atatürk" Wikidata sayfasından kişiye ait istenilen nitelikleri, yani ilişkiler (relations) ve tanımları elde etmek için yeni bir sorguda kullanılabilir. Elde edilen varlık ilişkilerinin bir örneği Şekil 8'de gösterilmiştir.



Şekil 9. Varlık\Entity (Nicolas Tesla) and İlişkiler\Relations

Entity: Nikola Tesla – WikidataID: Q9036: (Nikola Tesla – physicist - a person who specializes in physics), (Nikola Tesla - futurist - a person who studies or is expert in futurism), (Nikola Tesla - mechanical engineer - a person who designs, builds, or maintains engines, machines, or structures), (Nikola Tesla - electrical engineer - a person who designs, develops, or maintains electrical systems and equipment), (Nikola Tesla - inventor - a person who creates or discovers a new method, form, device, or other useful means), (Nikola Tesla - United States of America - a country where Nikola Tesla was a citizen), (Nikola Tesla - Kingdom of Hungary - a country where Nikola Tesla was a citizen), (Nikola Tesla - Austrian Empire - a country where Nikola Tesla was a citizen). Bu ilişkilerin bir örneği Şekil 9'da gösterilmiştir.

Örneğin, "Nikola Tesla" ismi, önceden eğitilmiş bir NER modeli kullanılarak tespit edilen bir metni ele alalım. Tespit edilen varlık ismi "Nikola Tesla" için Wikidata üzerinde bir SPARQL sorgusu gerçekleştirilir. Eşleşen bir sonuç bulunursa, bu varlık ismine karşılık gelen Wikidata ID değeri "Q9036" olarak sorgu sonucunda elde edilir. Ardından, bu varlık (entity) ID değeri, "Nikola Tesla" Wikidata sayfasından kişiye ait istenilen nitelikleri, yani ilişkiler (relations) ve tanımları elde etmek için yeni bir sorguda kullanılabilir. Elde edilen varlık ilişkilerinin bir örneği yukarıda gösterilmiştir.

Bu örnekler incelendiğinde, Wikidata' dan elde edilen ilişkiler kişilerin mesleği, tuttuğu pozisyon ve kişiyle ilgili tanımları içerir. Metindeki kişi hakkında bu tanımlayıcı ilişki bilgileri, metinlerin kategorisinin belirlenmesini kolaylaştırıcı bir etkiye sahiptir. Diğer tür ilişkilerle

ilgili bilgiler kullanılmamıştır. İlişkilerin artırılması, sınıflandırma görevini daha karmaşık ve zorlu hale getirir. Tablo 5' te görüldüğü gibi, veri kümeleri içinde tespit edilen varlık (entity) sayıları ve bu varlıklar aracılığıyla Wikidata' dan elde edilebilecek ilişki (relation) sayıları gösterilmektedir.

Tablo 5. Önerilen model ile veri kümelerine eklenen varlık (entity) ve ilişki (relation) sayısı.

Veriseti	**Doküman Sayısı**	**NER tarafından tespit edilen Varlıkların (Entities) Sayısı (Per, Org, Loc, Misc)**	**Sadece PER (kişi) etikteli Varlıklar (Entities) ile Wikidata İlişkilerinin (Relationships) Sayısı**
TTC-4900	4900	32070	4952
BBC-News	2224	15304	16527
20Newsgroups	6517	38179	14992

Çalışmamızda, metinlerde Named Entity Recognition (NER) ile 'kişi' olarak tanımlanan varlıklar, SPARQL kullanılarak Wikidata veritabanında sorgulanmıştır. Bu hedef odaklı yaklaşım, bireylere özgü ilişkileri içeren sınırlı bir ilişki setine odaklanmaktadır, bu da potansiyel gürültüyü azaltır. Bu süreç başlangıçta veri toplama kapsamını sınırlayıcı gibi görünse de, ilgisiz bilgilerin dahil edilmesini önemli ölçüde azaltarak analiz sürecini hızlandırır ve sonuçların yorumlanmasını kolaylaştırır. Özellikle 'kişi' etiketi kullanımı, sosyal dinamikler, etki ilişkileri ve bireylerin toplumdaki rolleri hakkında derinlemesine bir inceleme yapılmasını sağlar, araştırmamızın odak noktasını netleştirir ve ilgili bilgilerin daha etkili bir şekilde çıkarılmasını mümkün kılar.

Sonuç olarak, sınıflandırma için metin veri kümeleri, metinler içindeki varlıkların ve ilişkilerin belirlenmesiyle zenginleştirilmiştir. Böylece, metinler içindeki varlıklardan sağlanan yeni bilgilerin eklenmesinin sınıflandırma performansı üzerindeki olumlu etkisi gösterilmeye çalışılmıştır. Zenginleştirilmiş metinler, sınıflandırma için hazırlanmak üzere vektörlere dönüştürülmüştür.

3.5. Metin Temsilleri

Metinlerdeki söz dizimsel ve anlamsal ilişkileri çözümlemek için kullanılan metin özellikleri çıkarma teknikleri, sürekli olarak geliştirilmekte ve iyileştirilmektedir. Bu teknolojiler, metinler arasındaki karmaşık bağlantıları daha iyi anlamak ve çözümlemek amacıyla tasarlanmıştır. Gelişim süreci devam etmekle birlikte, mevcut metodlar bazı kısıtlamalar içermesine rağmen, bu alanda kaydedilen ilerlemeler dikkate değerdir. Tekniklerin evrimi, daha doğru ve etkili analizler yapılabilmesi için önemli potansiyeller sunmaktadır (Li, 2017).

Bu bağlamda, metin analizi süreçlerini daha da ileri taşıyan kelime temsil yöntemleri büyük bir öneme sahiptir. Bu yöntemler, metin içerisindeki kelimelerin ve ifadelerin matematiksel ve sayısal değerlere dönüştürülmesini sağlayarak, bilgisayarların bu verileri işleyebilmesine olanak tanır. Başlangıçta, basit ve etkili bir yöntem olan Kelime Torbası (Bag of Words - BoW) modeli kullanılmıştır. Ancak zamanla, metinlerin daha derinlemesine analizi için Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF) gibi yöntemler geliştirilmiştir. Daha sonra, kelime ilişkilerini ve anlamsal benzerlikleri daha detaylı yakalayabilen word embeddings, word2vec, GloVe ve fastText gibi modeller ortaya çıkmıştır. En sonunda, derin öğrenme temelli transformer modelleri ve BERT gibi mimariler, metin işleme ve dil modelleme alanında devrim yaratmıştır. Bu teknikler, kelimelerin ve ifadelerin bağlamlarını çok daha iyi anlamamızı sağlayarak, metin analizindeki başarıyı artırmıştır.

3.5.1. Ağırlıklı Kelime Çıkarma ve Terim Frekansı-Ters Belge Frekansı (TF-IDF)

Ağırlıklı kelime çıkarımı, özellik çıkarma süreçlerinin temel bir yöntemi olarak kabul edilir ve her kelimenin bir belgede kaç kez geçtiğini temel alan Terim Frekansı (TF) metodunu kullanır. Bu yöntem ile her belge, içindeki kelime sıklıklarını temsil eden bir vektöre dönüştürülür. Ancak, bu yaklaşım dili sık kullanılan bazı kelimelerin belgelerde aşırı temsil edilmesi sorununa neden olabilir (Ramos, 2003).

Terim Frekansı (TF), belirli bir terimin bir dokümanda ne kadar sık geçtiğini ölçer. TF, terimin belirli bir dokümanda gözlemlenen frekansının, dokümandaki toplam terim sayısına oranıyla hesaplanır (Denklem 1).

$$TF = \frac{dok\ddot{u}mandaki\ terime\ ait\ sayı}{dok\ddot{u}mandaki\ toplam\ terim\ sayısı} \quad (1)$$

Ters Doküman Frekansı (IDF), bir terimin tüm doküman koleksiyonunda ne kadar nadir olduğunu gösterir. Daha nadir terimler, daha yüksek bir IDF değeri alır, bu da onların bir dokümanda daha belirleyici olduğu anlamına gelir. IDF, belirli bir terimi içeren dokümanların sayısının, tüm dokümanlar içindeki oranının logaritmasının alınmasıyla hesaplanır (Denklem 2).

$$\text{IDF} = \log\left(\frac{korpusdaki\ toplam\ doküman\ sayısı}{terimi\ içeren\ doküman\ sayısı}\right) \quad (2)$$

Bir terimin TF-IDF değeri, ilgili TF ve IDF değerlerinin çarpımıyla elde edilir. Bu değer, terimin belirli bir dokümanda ne kadar önemli olduğunu belirler (Denklem 3).

$$TF - IDF = TF \times IDF \quad (3)$$

Bu ölçütler, doküman içerisindeki kelimelerin önemini değerlendirmede ve metin bazlı sorgularda, öneri sistemlerinde, doküman sınıflandırmada ve benzeri pek çok alanda kullanılır. Nadir kullanılan fakat doküman için özel öneme sahip kelimeler bu yöntemle tespit edilebilir.

Yaygın kullanılan kelimelerin etkisini azaltmak amacıyla Ters Belge Frekansı (IDF) tekniği geliştirilmiştir. IDF, belgedeki kelimelerin ne sıklıkta kullanıldığına göre ağırlıklar atayarak, terim sıklığına bir denge unsuru getirir. Bu şekilde, TF ve IDF bir araya gelerek Terim Frekansı-Ters Belge Frekansı (TF-IDF) olarak adlandırılan kombinasyonu oluşturur. TF-IDF, belgelerdeki yaygın kelimelerin aşırı temsil edilmesi sorununu hafifletmeye yardımcı olur, fakat bu yöntem kelimeler arasındaki bağlantıları göz ardı ederek her kelimeyi bağımsız bir indeks olarak sunar ve bu da bazı analiz eksikliklerine yol açabilir.

3.5.2. Kelime Torbası Yöntemi (Bag Of Words)

Kelime torbası modeli (BoW), bir metnin basit ve öz bir biçimde ifade edilmesini sağlayan, metin içerisindeki kelime frekanslarına dayanan bir yöntemdir. Bu model, doğal dil işleme (NLP), spam tespiti, belge sınıflandırması ve bilgi çıkarımı gibi birçok alanda etkin olarak kullanılmaktadır. BoW tekniğinde, metin bir kelime koleksiyonu olarak ele alınır; burada kelimeler listelenir ve bir matriste düzenlenir. Bu süreçte, kelimeler arasındaki anlamsal bağlar genellikle ihmal edilir. Kelimelerin frekansları hesaplanır ve bu veriler, belgelerin temel içeriğini belirlemek için kullanılır. Bu yaklaşım, dilbilgisi ve kelime sıralamasını göz ardı ederken, kelime sıklığını öne çıkarır. Bu model, metinlerin özetlenmesi ve sınıflandırılması gibi

işlemlerde temel bir araç olarak kabul görmekte ve geniş bir kullanım alanına sahiptir (Zhang, 2010).

3.5.3. N-Gram

N-gram yöntemi, metin analizi için kullanılan bir tekniktir ve bir dizi metinde, belirli bir sıra ile yan yana gelen n adet kelimenin birleşiminden oluşan gruplardır. Bu gruplar, doğrudan bir metin temsili sağlamasa da, bir metni ifade etmek için kullanılacak özellikler topluluğu olarak işlev görür. Öte yandan, Kelime Torbası (BoW) modeli, kelimelerin sırasını dikkate almayan daha basit bir metin temsili sunar ve bu modelin oluşturulması nispeten kolaydır. Metinler, genellikle yönetilebilir boyutta bir vektör aracılığıyla ifade edilir. N-gram, farklı uzunluklardaki kelime gruplarını (örneğin, 1-gram, 2-gram, 3-gram) kullanarak daha kapsamlı metin özellikleri çıkarır ve böylece 1-grama kıyasla daha zengin bilgiler sunar. (Cavnar, 1994).

Örneğin, "Öğrenci kütüphanede sessizce ders çalışıyor." cümlesi için;

- 2-gram: {"Öğrenci kütüphanede", "kütüphanede sessizce", "sessizce ders", "ders çalışıyor"}
- 3-gram: {"Öğrenci kütüphanede sessizce", "kütüphanede sessizce ders", "sessizce ders çalışıyor"}.

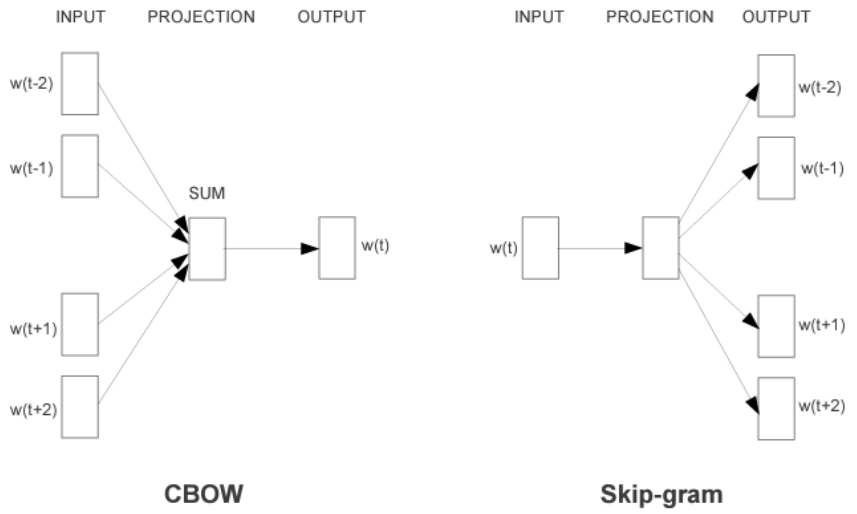
Bu örnekler, n-gram tekniğinin, metinlerin daha bağlamsal ve detaylı bir şekilde temsil edilmesini nasıl sağladığını açıkça gösterir. Kelimelerin sıralamasını ve bağlamını göz önünde bulunduran bu metod, metin analizinde daha kapsamlı bir anlayış elde etmemize yardımcı olur. N-gram, özellikle dil modelleme ve metin madenciliği uygulamalarında, kelimelerin ve ifadelerin içerdiği anlamların daha derinlemesine çözümlenmesine olanak tanır.

3.5.4. Kelime Gömme Yöntemleri (Word Embeddings)

Word embeddings, kelimeleri yoğun sayısal vektörler şeklinde ifade eden ileri bir doğal dil işleme (NLP) metodudur. Bu teknik, kelimeleri çok boyutlu bir uzayda temsil ederek, anlamlarını ve kullanıldıkları dil bağlamlarını görsel olarak noktalarla ifade eder. Her bir kelime, onun semantik özelliklerini yansıtan bir dizi sayısal değerle tanımlanır. Kelimeler arasındaki semantik ve sözdizimsel ilişkileri sayısal olarak modelleyerek, dilin daha derinlemesine analiz edilmesini sağlar. Örneğin, "kral" ve "kraliçe" gibi semantik olarak benzer kelimeler, vektör uzayında birbirlerine yakın yerleştirilir. Bu, kelimelerin anlamlarını ve kullanıldıkları bağlamları yansıttığı için dil analizinde büyük bir doğruluk sağlar (Qaiser, 2018).

Kelime gömme, kelime dağarcığındaki her kelimenin gerçek sayılardan oluşan N boyutlu bir vektöre eşlendiği bir özellik öğrenme tekniğidir. Bu teknik, derin öğrenme modellerinde başarıyla kullanılmakta olup, Word2Vec, GloVe ve FastText gibi yöntemler en yaygın kullanılan kelime gömme örnekleridir. Bu yöntemler, kelimelerin bağlamsal anlamlarını ve birbirleriyle olan ilişkilerini daha iyi yakalayarak, metin analizinde ve sınıflandırmada daha yüksek performansa ulaşmayı sağlar.

Word2Vec, etiketsiz (unsupervised) ve tahmine dayalı bir modeldir ve kelimeleri vektör uzayında temsil etmeyi amaçlar. Google araştırmacısı Tomas Mikolov ve ekibi tarafından 2013 yılında geliştirilmiştir (Mikolov, 2013). Bu model, **sürekli kelime torbası (CBOW)** ve **Skip-gram** modelleri üzerine inşa edilmiştir. CBOW modeli, birden fazla kelimenin input olarak alındığı ve bu kelimelerin merkezinde yer alabilecek bir kelimenin output olarak tahmin edilmeye çalışıldığı bir yapıdır. Skip-Gram modeli ise, tek bir kelimenin input olarak alınıp bu kelimenin çevresinde yer alabilecek kelimelerin output olarak tahmin edilmeye çalışıldığı bir yapı sunar (Şekil 10).



Şekil 10. CBOW ve Skip-gram mimarisi (Mikolov, 2013).

Word2Vec modeli, anlamca ilişkili kelimeleri vektörel uzayda birbirlerine yakın olarak yerleştirir. Mesela, "kahve" ve "fincan", "bilgisayar" ve "klavye", "orman" ve "ağaç" gibi kelimeler Word2Vec modelinde semantik yakınlıkları nedeniyle birbirine komşu konumda bulunur. Bu durum, kelimelerin benzer bağlamlarda kullanıldığını ve aralarında güçlü anlamsal bağlar olduğunu yansıtır. Word2Vec, kelimeler arasındaki bu ilişkileri başarıyla modelleyerek,

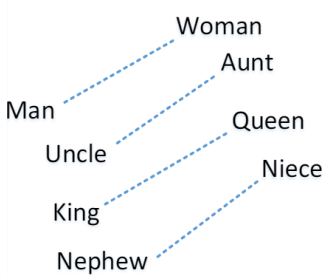
doğal dil işleme alanında çeşitli uygulamalarda kullanılır. Model, metinlerdeki kelimeleri anlamlandırma, dil yapılarını tanıma ve derinlemesine dil analizi yapma gibi işlemlerde etkili bir araçtır.

GloVe (global vectors for word representation), metin sınıflandırmasında kullanılan etkili bir kelime gömme tekniğidir. Pennington ve ekibi tarafından (2014) geliştirilen bu unsupervised öğrenme algoritması, kelimelerin vektör temsillerini elde etmek için tasarlanmıştır. Bu model, bir korpustan toplanan kelimeler arası küresel birlikte bulunma istatistiklerine dayanarak eğitilir. Anlamsal olarak ilişkili kelimelerin vektör temsilleri bu modelde birbirine yakın konumlandırılır.

Burada bahsedilen formül, kelime vektörleri ve bu vektörlerin bir dil modelinde nasıl kullanılabileceği hakkında bilgi veriyor. Formül, özellikle bir kelimenin bağlamında başka bir kelimenin olasılığını hesaplamak için kullanılıyor. Denklem 4:

$$f(w_i - w_j, \widetilde{w}_k) = \frac{P_{ik}}{P_{jk}} \tag{4}$$

Bu fonksiyon, iki kelime vektörü arasındaki farkın, belirli bir bağlam kelimesi ile ilişkisini değerlendirir. Fonksiyonda parametre olarak kelime i'nin kelime vektörü ve kelime j'nin kelime vektörü verilmiştir. Sonucunda, kelime k'nin kelime i'nin bağlamında olasılığını gösterir. Yani, kelime k'nin kelime i ile aynı cümle ya da belirli bir kelime aralığında birlikte bulunma ihtimalidir. Benzer şekilde, kelime k'nin kelime j'nin bağlamında olasılığını da gösterir. Formülün genel amacı, iki kelime vektörü arasındaki ilişkinin, bir bağlam kelimesine göre nasıl değiştiğini ölçmektir.



Şekil 11. GloVe: Global Vectors for Word Representation (Kowsari, 2019)

Bu modelde (Şekil 11), semantik olarak ilişkili kelimeler vektörel uzayda birbirlerine yakın olarak konumlandırılır.

Örnek olarak; "güneş" ve "ışık", "okyanus" ve "deniz", "mutluluk" ve "sevinç" gibi kelimeler, GloVe modelinde anlamsal bağları nedeniyle birbirlerine yakın yerlerde yer alırlar. Bu yakınlık, kelimelerin anlam açısından birbirine benzer olduğunu ve metinlerde benzer bağlamlar içinde kullanıldıklarını gösterir. GloVe, kelimelerin anlamlarını ve kullanıldıkları bağlamları daha iyi kavramamızı sağlayan, doğal dil işleme alanında sıkça tercih edilen etkili bir kelime gömme yöntemidir.

FastText, Facebook AI Araştırma laboratuvarı tarafından geliştirilmiş, Word2Vec modeline benzer yapısıyla dikkat çeken bir kelime gömme yöntemidir (Joulin ve diğ., 2016). Bu teknik, kelimelerin morfolojik özelliklerini de içerecek şekilde tasarlanmıştır. Her kelime, w, bir dizi n-gram karakterli çanta olarak temsil edilir ve kelimeleri tek tek değil, bunları birkaç harflik n-gramlara bölerek yapay sinir ağına girdi olarak sunar. FastText modeli, kelimelerin anlamını ve yapısını derinlemesine analiz etmek için kullanılır ve özellikle morfolojik olarak zengin dillerde etkili olduğu için, bu dillerin incelenmesinde ve metin işlemede kritik bir rol oynamaktadır.

Örneğin, "mutluluk" kelimesi kullanıldığında, tri-gram yapısı ile analiz edilir ve yapay sinir ağına "mut", "utl", "tlu", "lul", "ulu", "luk" şeklinde girdiler sunulur. Burada "n", n-gramın tekrar derecesini ifade eder ve bir kelimenin kaç harflik parçalara bölüneceğini belirler. "mutluluk" kelimesinin kelime vektörü, bu n-gram vektörlerinin birleşimi olarak hesaplanır.

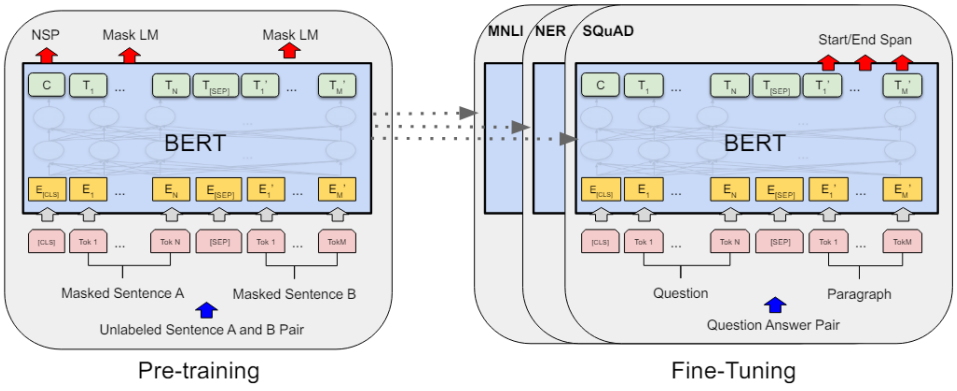
Facebook, FastText modelini geniş bir dil yelpazesi kapsayacak şekilde tasarlamıştır ve model, Wikipedia veri seti üzerinde 300 boyutlu vektörler kullanılarak eğitilmiştir. Bu eğitim sonucunda, 294 dil için önceden eğitilmiş kelime vektörleri yayınlanmıştır. Bu geniş kapsamlı model, kelimelerin anlamını ve yapısını derinlemesine analiz etmek için kullanılır ve doğal dil işlemede değerli bir araç olarak kabul edilir. Özellikle morfolojik olarak zengin dillerde etkili olan FastText, bu dillerin incelenmesinde ve metin işlemede kritik bir rol oynar.

Bu geniş kapsamlı model ve teknikler, metin sınıflandırma, duygu analizi ve makine çevirisi gibi alanlarda performansı önemli derecede iyileştirebilir, metin tabanlı verilerin işlenmesinde ve anlamlandırılmasında derin öğrenme yaklaşımlarının temel bileşenleri arasında yer alır.

3.5.5. BERT: Bidirectional Encoder Representations from Transformers

BERT, Google AI tarafından geliştirilmiş ve doğal dil işleme alanında önemli bir yenilik getiren bir modeldir. Bu modelin tam adı "Bidirectional Encoder Representations from Transformers" olup, metinlerden derinlemesine çift yönlü temsiller öğrenmek üzere tasarlanmıştır. BERT modeli, etiketlenmemiş metinler üzerinden eğitilerek her bir katmanda hem sol hem de sağ bağlamı değerlendirebilen bir yapı sunar. Bu yapı, çeşitli doğal dil işleme görevlerinde yüksek performans elde edilmesini sağlar ve modelin sadece ek bir çıktı katmanı ile özelleştirilerek kullanılabilmesine olanak tanır. BERT' in uygulama alanları arasında metin sınıflandırma, adlandırılmış varlık tanıma ve soru-cevap sistemleri yer alır.

BERT modeli, başlangıçta İngilizce dili üzerinde eğitilmiş olmasına rağmen, kısa sürede çoklu dil desteği ile genişletilmiştir. Bu genişleme, modelin 100'den fazla dilde etkili dil temsilleri öğrenmesini mümkün kılmıştır. BERT' in çift yönlü Transformer mimarisi, metinleri tamamıyla ve bağlamsal olarak değerlendirme yeteneği sayesinde dil işleme görevlerinde önemli başarılar elde edilmiştir. Öncül eğitim sürecinde kullanılan maskeleme işlemi ve sonraki cümle tahmini görevleri, modelin dilin yapısal ve bağlamsal özelliklerini daha iyi kavramasını sağlar. BERT modelinin başarısı, birçok türevinin geliştirilmesine ilham kaynağı olmuştur.



Şekil 12. BERT için genel ön eğitim ve ince ayar prosedürleri (Devlin, 2018).

BERT (Bidirectional Encoder Representations from Transformers) modelinin eğitim sürecini iki ana aşamada Şekil 12' de görülmektedir: ön eğitim ve ince ayar. Ön eğitim aşamasında model, Masked Language Model (Mask LM) ve Next Sentence Prediction (NSP) görevleri kullanılarak büyük miktarda etiketlenmemiş metin üzerinde eğitilir. Bu süreç, modelin dil yapısını genel anlamda öğrenmesini sağlar. İnce ayar aşamasında ise, model özel bir görev için özelleştirilir, örneğin, belirli bir metin parçasından cevap üretme veya

adlandırılmış varlık tanıma gibi. Model bu aşamada, belirli bir göreve özel veri seti ile eğitilerek, daha spesifik görevlerde yüksek performans göstermesi için ayarlanır. İşte BERT' in bazı önemli türevleri ve bunların özellikleri:

DistilBERT: BERT'in daha hafif ve daha hızlı bir versiyonu olan DistilBERT, öğrenci-öğretmen mimarisi kapsamında eğitilmiştir. BERT modelinin temel özelliklerini korurken, model boyutunu ve işlem gereksinimlerini azaltmak için tasarlanmıştır. DistilBERT, BERT'in performansının yaklaşık %95'ini, parametre sayısını ise yarı yarıya indirerek sunar. Bu model, kaynak kısıtlı ortamlarda veya gerçek zamanlı uygulamalarda kullanılmak üzere uygundur.

RoBERTa: BERT'in bir başka türevi olan RoBERTa, BERT'in eğitim sürecinde yapılan bazı değişikliklerle geliştirilmiştir. Özellikle, dinamik maskeleme, daha büyük mini gruplar ve daha fazla veri üzerinde daha uzun süre eğitim yapılması gibi optimizasyonlar içerir. RoBERTa, birçok doğal dil işleme görevinde BERT'ten daha üstün performans göstermiştir.

ALBERT (A Lite BERT): Bellek kullanımını optimize ederek daha büyük modellerin eğitilmesine olanak tanıyan bir başka BERT türevidir. ALBERT, parametre paylaşımı tekniği sayesinde model boyutunu önemli ölçüde azaltır ve özellikle uzun belgelerle çalışırken etkilidir.

Türkçe için özelleştirilmiş olan BERTurk modeli, Türkiye'deki dil yapılarına uygun olarak geliştirilmiştir. Bu model, Kemal Oflazer tarafından sağlanan ve Türkçe OSCAR Korpusu, Türkçe Vikipedi verileri ve OPUS korpuslarından oluşan geniş bir veri seti üzerinde eğitilmiştir. Toplamda 35 GB büyüklüğünde ve 44 milyar token içeren bu veri seti, BERTurk'ün Türkçe metinlerde yüksek doğrulukla çalışmasını sağlamaktadır. BERTurk, metin sınıflandırma, duygu analizi gibi çeşitli doğal dil işleme görevlerinde kullanılmakta ve Türkçe dilinin zengin yapısal ve semantik özelliklerini başarıyla modellemektedir.

Turkish ELECTRA Model: BERT ve BERTurk modelleri gibi, ELECTRA da Türkçe metinler üzerinde önceden eğitim yapmak için kullanılmıştır. ELECTRA, verimli dil temsilleri öğrenmek için jeneratör ve diskriminatör adında iki ayrı model kullanır. Bu yaklaşım, özellikle ince taneli dil özelliklerinin yakalanmasında etkili olmuştur.

Bu modeller, BERT'in başarılı mimarisini temel alarak, çeşitli iyileştirmeler ve spesifik uygulama gereksinimlerine uygun adaptasyonlar ile doğal dil işleme alanında kullanılmaktadır. Her bir türev, belirli görevler ve operasyonel ihtiyaçlar göz önünde bulundurularak

tasarlanmıştır, böylece BERT teknolojisinin geniş bir yelpazede etkin bir şekilde kullanılmasına imkan tanınmıştır. Sonuç olarak, BERT ve onun Türkçe adaptasyonu olan BERTurk, doğal dil işleme teknolojilerindeki sınırları genişletmiş ve dil işleme uygulamalarının doğruluğunu artırmıştır. Bu modeller, çoklu dillerde etkili bir şekilde çalışabilme kapasitesi ile dil işleme alanında standartları yükseltmiştir.

3.6. Metin Vektör Temsillerinin Elde Edilemesi

Metin, varlık (entity) ve ilişki (relationship) vektör temsilleri, Türkçe BERT (BERTurk) modeli (Aras, 2021) ve İngilizce BERT temel modeli (Devlin, 2018) kullanılarak elde edilmiştir. BERT modeli, metinlerin, varlıkların ve ilişkilerin vektör temsillerini elde etmede temel bir araç olarak öne çıkmaktadır. BERTurk ve İngilizce BERT modelleri, sırasıyla Türkçe ve İngilizce metinler için derin öğrenme tabanlı bir yaklaşım sunmaktadır. Geniş ve çeşitli veri setleri üzerinde eğitilen bu modeller, dilin karmaşık yapısını ve anlamsal özelliklerini yakalayan güçlü temsiller üretmektedir. Özellikle BERTurk modeli, Kemal Oflazer tarafından sağlanan özel bir korpus kullanılarak eğitilmiştir; bu korpus Türkçe OSCAR Corpus, Türkçe Wikipedia Dump ve OPUS korpuslarını içermekte olup toplamda 35 GB ve 44 milyar token içermektedir. BERT modelleri, metinlerin sağlam vektör temsillerini sağlamakla kalmaz, dilbilgisi etiketleme, adlandırılmış varlık tanıma ve soru cevaplama gibi doğal dil işleme alt görevlerinde de yararlıdır.

Her metin için, n-boyutlu bir vektör temsili elde etmek amacıyla bir BERT modeli kullanılır. Bu temsil, metnin dilbilimsel ve anlamsal özelliklerini kapsamlı bir şekilde yansıtır. Aynı zamanda, Wikidata'dan türetilen her varlık-ilişki için m-boyutlu BERT vektör temsilleri üretilir. Bu süreç, metinlerin yanı sıra ilişkili varlıklar ve ilişkilerin kapsamlı bir analizini sağlar, böylece hem dilbilimsel hem de ontolojik özellikleri içeren zengin bir vektör temsili oluşturur.

Denklem (5)' deki her metin için, n boyutlu bir vektör temsili elde etmek amacıyla bir BERT modeli kullanırız:

$$\mathbf{V}_{\mathrm{BERT}} = [\, v_1, v_2, \ldots, v_{\mathrm{n}}] \tag{5}$$

Ek olarak, her metinde Wikidata' dan elde edilen her varlık (entity) ve onunla ilişkili ilişkiler (relations) için, denklem (6)'de m-boyutlu BERT vektör temsillerini türetiyoruz:

$$\mathbf{V}_{\mathrm{ER}}^{\mathrm{i}} = \left[\, er_1^{\mathrm{i}} + er_2^{\mathrm{i}} + \cdots + er_{\mathrm{m}}^{\mathrm{i}} \right] \quad for\ i = 1,2,\ldots,j \tag{6}$$

Bu m-boyutlu vektörlerin ortalaması (average entity-relation vector) denklem (7) ile hesaplanır:

$$\mathbf{V}_{\mathrm{ER-avg}} = \frac{1}{j} \sum_{i=1}^{j} \mathbf{V}_{\mathrm{ER}}^{\mathrm{i}} \tag{7}$$

Denklem (8)'teki tek bir metin için, hem metnin n-boyutlu BERT temsilini hem de m-boyutlu ortalama varlık-ilişki vektörünü (entity-relation vector) içeren birleştirilmiş vektör (combined vector) dolayısıyla n+m boyutludur:

$$\mathbf{V}_{\mathrm{combined}} = \left[\mathbf{V}_{\mathrm{BERT}}, \mathbf{V}_{\mathrm{ER-avg}} \right] \tag{8}$$

Bu işlem veri kümesindeki k metnin her biri için tekrarlanarak denklem (9) elde edilir:

$$\mathbf{V}_{\mathrm{combined}}^{\mathrm{k}} = \left[\mathbf{V}_{\mathrm{BERT}}^{\mathrm{k}}, \mathbf{V}_{\mathrm{ER-avg}}^{\mathrm{k}} \right] \tag{9}$$

Burada, $\mathbf{V}_{\mathbf{BERT}}^{\mathbf{k}}$ k numaralı metnin **n-boyutlu** BERT vektör temsilini ifade ederken, $\mathbf{V}_{\mathbf{ER-avg}}^{\mathbf{k}}$ ise **k numaralı** metin için ortalama **m-boyutlu** varlık-ilişki (entity-relationship) vektörünü belirtir. Böylece her birleştirilmiş $\mathbf{V}_{\mathbf{combined}}^{\mathbf{k}}$, veri setindeki **k numaralı** metin için **n+m boyutlu** birleştirilmiş vektör temsilini temsil eder.

Sonuç olarak, her metnin BERT temsili ortalama varlık-ilişki vektörü ile birleştirilerek, her metin için n+m boyutlu birleştirilmiş vektör temsili elde edilir. Bu birleşik temsil, metnin dilbilimsel özelliklerini ve varlık-ilişki bilgilerini kapsar ve çeşitli doğal dil işleme görevlerinde uygulanabilir. Bu yaklaşım, metinler arası ilişkiler, anlamsal benzerlikler ve farklılıklar hakkında daha derin bir anlayış sağlar. Özellikle büyük veri setleriyle çalışırken, bu tür birleşik vektör temsilleri, metinlerin karmaşık yapılarını ve ilişkilerini modellemek için güçlü bir araç sunar.

3.6.1. Boyut Azaltma (Dimension Reduction)

Terim bazlı vektör modellemelerinde, metinler genellikle geniş özellik setleri ile temsil edilir. Bu durum, hem zaman karmaşıklığı hem de bellek kullanımı açısından maliyetli olabilmektedir. Bu maliyeti azaltmak için araştırmacılar, özellik uzayını küçültecek boyut indirgeme tekniklerine başvurmaktadırlar.

En yaygın boyut azaltma yöntemleri arasında Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Random Projection ve Autoencoder yer almaktadır. Örneğin, PCA, veri setindeki varyansı maksimum düzeyde koruyarak daha az sayıda boyuta indirgemeyi amaçlar, böylece orijinal veri setinin en önemli özelliklerini daha küçük bir uzayda temsil etmeyi sağlar. LDA ise sınıflandırma problemlerinde sınıflar arası ayırt ediciliği maksimize edecek şekilde boyut indirger. Random Projection, veri boyutunu rastgele düşürerek hızlı ve etkili bir çözüm sunarken, Autoencoder derin öğrenme temelli bir yaklaşımla verinin daha düşük boyutlu ancak zengin temsillerini öğrenir. Bu teknikler, veri işleme süreçlerini optimize ederek daha hızlı ve maliyet-etkin çözümler sunar.

3.6.2. Metin Graf Temsili

Metin sınıflandırmasında graf temsili, metinlerdeki yapısal ve anlamsal öğeler arasındaki ilişkileri görselleştirmek ve daha derinlemesine analizler yapabilmek için kullanılan güçlü bir araçtır. Metinlerde yer alan kelimeler, cümleler, belgeler ve diğer öğeler düğümler olarak, bu öğeler arasındaki ilişkiler ise kenarlar olarak modellenir. Bu temsil, metinleri daha karmaşık ve bağlamsal ilişkiler üzerinden değerlendirmeyi mümkün kılarak, sınıflandırma algoritmalarının daha hassas ve etkili çalışmasını sağlar.

Örneğin, bir haber metninde geçen "ekonomi" ve "bütçe" kelimeleri arasındaki ilişki, her iki kelimeyi düğüm olarak gösteren ve aralarındaki semantik ilişkiyi bir kenarla ifade eden bir graf ile modellenebilir. Bu tür bir görselleştirme, kelimeler arasındaki anlamsal ilişkileri ve metin içi yapıları net bir şekilde ortaya koyar.

Özellikle co-occurrence grafikleri, metin içinde sıkça birlikte geçen kelimeleri analiz eder. Diyelim ki bir sağlık makalesinde "diyet" ve "kalori" kelimeleri sıkça yan yana kullanılmış. Bu iki kelime, belirli bir pencere boyutu dikkate alınarak düğüm olarak eklenir ve aralarında, birlikte geçme sıklıklarına bağlı olarak bir kenar oluşturulur. Bu bağlantı, diyet ve kalori kelimelerinin metin içinde nasıl ilişkilendirildiğini görsel olarak sergiler.

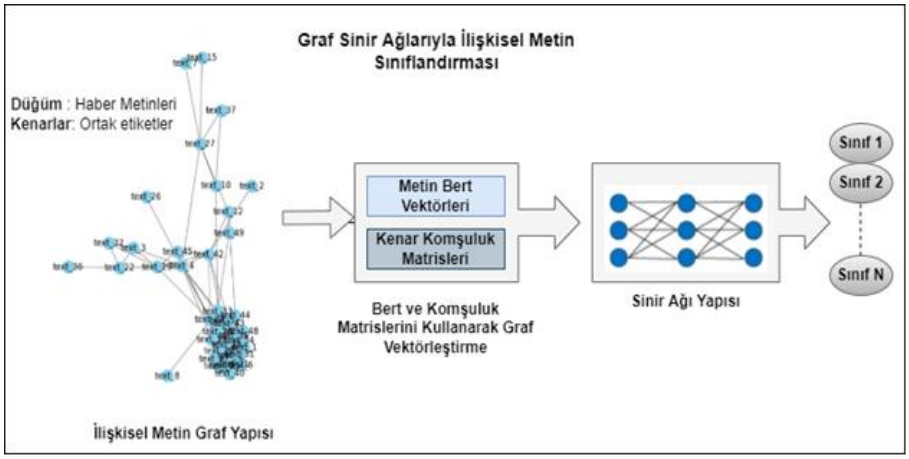
Graf temsili, metinlerdeki kavramsal ilişkileri derinlemesine analiz etmek için de kullanılır. Örneğin, bir metinde geçen "küresel ısınma" ve "karbon emisyonları" gibi çevresel terimler arasındaki ilişkiler, bu kavramların her birini birer düğüm olarak gösteren ve aralarındaki etkileşimleri kenarlarla ifade eden bir graf ile incelenebilir. Bu, özellikle çevresel politikaların tartışıldığı metinlerde, terimler arasındaki ilişkilerin ve öneminin anlaşılmasına yardımcı olur.

Ek olarak, metinlerde sıkça karşılaşılan varlıklar (entiteler), haber makalelerindeki etiketler, URL'ler, akademik yayınlardaki referanslar, anahtar kelimeler ve metin özetleri gibi öğelerin graf temsilindeki kullanımı, metinlerin daha kapsamlı bir şekilde işlenmesini ve anlamlandırılmasını sağlar. Örneğin, bir akademik makaledeki referanslar birbirleriyle ve makalenin ana konusuyla olan ilişkileri gösteren düğümler ve kenarlar ile gösterilebilir. Bu, makalenin hangi çalışmalarla daha yakın ilişkili olduğunu ve bilimsel bağlamını anlamada kritik bir rol oynar.

Bu süreçler ve tekniklerin entegrasyonu, metin sınıflandırma sistemlerinin genel performansını artırarak, daha doğru sonuçlar elde etmeye yardımcı olur ve algoritmaların daha etkin ve doğru çalışmasına olanak tanır.

3.7. Graf Sinir Ağı ile İlişkisel Metin Sınıflandırması

Bu bölümde anlatılan araştırmada ise, Türkçe metinlerin sınıflandırılması ve ilişkisel analizine yönelik olarak metin graf yapısı ve Derin Sinir Ağları (DNN) teknolojilerini birleştiren yenilikçi bir metodoloji geliştirmiştir. Bu metodoloji, haber metinlerinin derinlemesine analizi için graf yapısı ve DNN teknolojilerinin avantajlarını entegre etmektedir. Graf Tabanlı Sinir Ağı (GNN) kullanılarak, metinlerdeki anahtar kelimeler, varlıklar ve bu unsurlar arasındaki ilişkiler grafiksel bir modelle temsil edilmektedir. Bu strateji, haber metinlerinin yapısal ve bağlamsal detaylarını kapsamlı bir şekilde analiz etmekte ve metinlerin çok boyutlu özelliklerini kelime seçiminden cümle yapısına ve genel tona kadar işleyerek detaylı bir sınıflandırma sağlamaktadır.



Şekil 13. Graf Sinir Ağı ile İlişkisel Metin Sınıflandırması

Araştırma kapsamında kullanılan "newsTRT2023" adlı veri seti, 4027 adet Türkçe haber içermekte ve her haber öğesi yayın tarihi, başlık, web adresi, kategori, kısa özet, tam metin, ilişkili etiketler ve metinde geçen varlıklar gibi özellikler barındırmaktadır. Bu derinlemesine veri seti, modelin geliştirilmesi ve test edilmesi için mükemmel bir altyapı sunar ve GNN, haber metinlerini etkin bir şekilde sınıflandırma imkânı sağlar. Çalışmada kullanılan graf modelinin görsel temsili Şekil 13' te sunulmuştur.

Bu belirtilen graf yapısında, düğümler haber metinlerini temsil ederken, kenarlar ise bu metinler arasındaki etiket bazında ilişkileri (ortak etiketlerin varlığı durumunda) ifade eder. Her bir haber makalesine ilişkili etiketler ve varlıklar eklenmiş olup, bu unsurlar üzerinden metinler arası ilişkisel bağlantıları ve yapıları anlamak için detaylı grafiksel analizler yapılmaktadır. Bu analizler, metinlerin ilişkisel ve yapısal özelliklerini detaylı bir şekilde incelemeyi mümkün kılar.

Ayrıca, haber metinleri üzerinde ilişkisel etiketler kullanılarak oluşturulan graf yapısı matematiksel aşağıdaki adımlar ile modellenmiştir ve bu modelin matematiksel ifadesi Denklem 10' da gösterilmiştir.

$$f(G) = GNN(X, A) \tag{10}$$

Bu formül, Matematiksel işlem, Graf notasyonu, düğüm özellikleri, kenar ağırlıkları, kenar komşuluk matrisi ve graf vektörleştirme adımlarını içermektedir. İlgili adımlar maddeler halinde gösterilmiştir:

1. **Graf Notasyonu:**

   - **G = (V, E)** olarak gösterilen denklemde graf **G**, düğümler kümesi **V** ve kenarlar kümesi **E** olmak üzere.
   - **V = {v1, v2, ..., vn},** burada her **vi** bir haber metnini temsil eder.
   - **E ⊆ { (vi, vj) | vi, vj ∈ V }**, burada **(vi, vj)** düğümleri arasındaki bir kenarı temsil eder.

2. **Düğüm Özellikleri:**

   - Her düğüm **vi** için, BERT modeli kullanılarak elde edilen semantik vektör **xi** 'dir.

3. **Kenar Ağırlıkları:**

   - **Wij**, düğüm çiftleri **vi** ve **vj** arasındaki ilişkinin gücünü ve sıklığını temsil eder.

4. **Kenar Komşuluk Matrisi:**

   - **A**, grafın topolojik yapısını temsil eden kenar komşuluk matrisidir.

     Eğer **(vi, vj) ∈ E** ise **Aij = wij**, aksi halde **Aij = 0.**

5. **Graf Vektörleştirme:**

   - Grafın vektörleştirilmesi için, her düğüm için semantik vektörler ve kenar komşuluk matrisi kullanılarak özellik matrisi **X** oluşturulur.

6. **Graf Sinir Ağı (GNN):**

   - GNN modeli, grafı ve onun özelliklerini kullanarak sınıflandırma ya da diğer tahminler yapar ve **f(G)** olarak ifade edilir.

7. **Model Eğitimi ve Testi:**

   - Model eğitimi, belirlenen kayıp fonksiyonu L ve hiperparametrelerle optimize edilir.

   - Test işlemleri, modelin performansını ölçmek için çeşitli metriklerle değerlendirilir.

Bu adımlara göre; G, haber metinlerini ve bunlar arasındaki ilişkileri temsil eden bir graf yapısıdır. X, haber metinlerinin özelliklerini temsil eden matristir. Bu özellikler, BERT modeli kullanılarak elde edilen semantik vektörlerdir. Her bir vektör, bir haber metninin içeriğini yüksek boyutlu bir uzayda temsil eder. A, grafın kenar komşuluk matrisidir ve grafın topolojik yapısını temsil eder. Bu matris, haber metinleri arasındaki ilişkilerin varlığını ve gücünü belirten sayısal değerler içerir. Matristeki her bir değer (Aij), i ve j düğümleri (haber metinleri) arasındaki bağlantının varlığını ve şiddetini gösterir. Bu matematiksel model, haber metinlerinin karmaşık yapılarını çözümlemek ve anlamak için kullanılmıştır.

Elde edilen zengin ve çok boyutlu graf vektör yapısı, graf sinir ağı tabanlı sınıflandırma modeline beslenmiştir. Bu model, graf teorik temsilleri ve derin öğrenme tekniklerini entegre

ederek, metinlerin sınıflandırılmasını ve ilişkisel bağlamda anlamlandırılmasını amaçlar. Modelin eğitimi, çeşitli hiperparametre ayarları yapılarak optimize edilmiştir, test işlemleri ise modelin genelleştirme kabiliyetini doğrulamak amacıyla yapılmıştır. Bu süreçte, model performansını değerlendirmek için çeşitli metrikler (doğruluk, F1 skoru, Precision ve Recall) kullanılmıştır. Buna göre makine öğrenimi algoritmalarının da karşılaştırmalı analizleri gerçekleştirilmiştir.

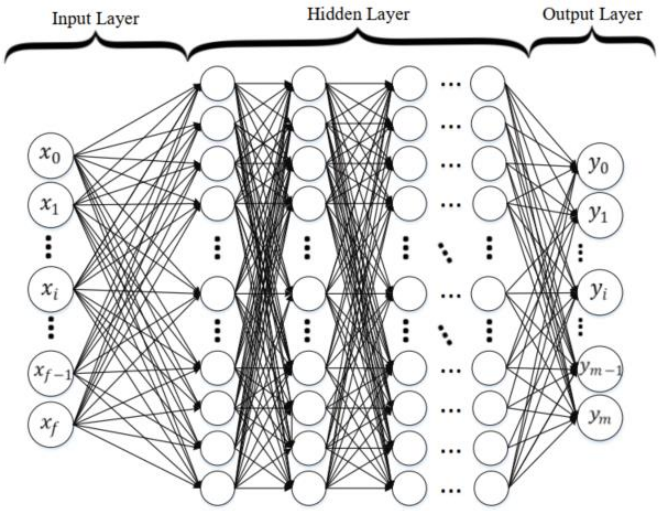
3.8. Sınıflandırma Modelleri

İlişkisel metin sınıflandırması, metinlerdeki varlıklar (entiteler) ve bu varlıklar arasındaki ilişkileri tanımlayarak derinlemesine analiz yapma kapasitesine sahip sınıflandırma tekniklerini kullanır. Bu süreçte, metinler öncelikle ön işlemeden geçirilerek temizlenir, normalleştirilir ve sonrasında vektörel bir forma dönüştürülür. Bu dönüşüm, metinlerin makine öğrenimi modelleri tarafından işlenebilir hale getirilmesini sağlar. İlişkisel metin sınıflandırmasında kullanılan temel sınıflandırma modelleri ve bunların özellikleri şu şekildedir:

Destek Vektör Makineleri (SVM): SVM, karar sınırlarını belirleyerek sınıflandırma yapan bir makine öğrenmesi modelidir. Yüksek boyutlu öznitelik uzaylarında, veri setlerini etkin bir şekilde temsil edebilme kapasitesine sahip olan SVM, özellikle metin sınıflandırma görevlerinde, metinlerin temsil edilmesi gereken karmaşık yapılarda oldukça etkilidir. SVM modelleri, sınıflar arasındaki en geniş marjı bulmaya çalışır ve bu sayede ayrımı netleştirir.

Lojistik Regresyon (LR): Lojistik Regresyon, bağımsız değişkenlerin belirli bir sınıfa ait olma olasılığını hesaplayarak, bu olasılıkları sigmoid fonksiyonu kullanarak 0 ile 1 arasında bir değere dönüştüren bir sınıflandırma modelidir. Çoğunlukla iki sınıflı sınıflandırma problemlerinde kullanılır ve sonuçlar olasılık olarak ifade edilir. Metin sınıflandırma gibi alanlarda, belirli kategorilere ayrım yapma görevlerinde yaygın olarak tercih edilir.

Derin Sinir Ağları (DNN): DNN'ler, birden fazla gizli katmandan oluşur ve bu katmanlar aracılığıyla verilerdeki karmaşık özellikleri ve ilişkileri öğrenir. Metin sınıflandırma görevlerinde, büyük veri kümeleri üzerinde karmaşık sınıflandırma ve regresyon problemlerini çözmede kullanılır (Kowsari, 2019). Bu modeller, özellikle derin öğrenme tekniklerinin gelişmesiyle birlikte metinlerdeki ince nüansları ve bağlamları yakalama konusunda üstün performans göstermektedir (Kowsari, 2019).



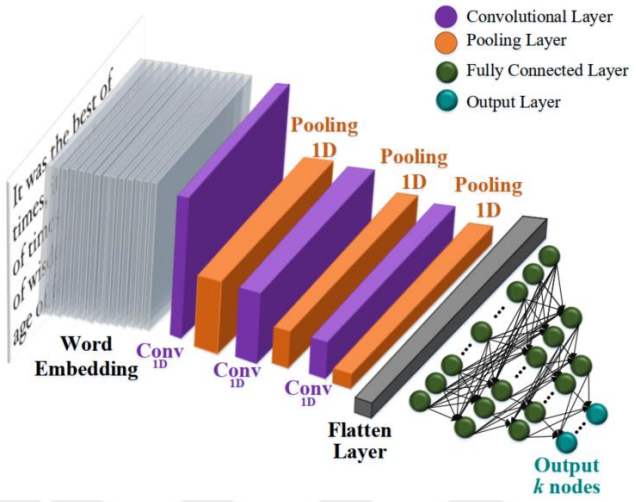
Şekil 14. Derin Sinir Ağı (DNN) modeli mimarisi (Kowsari, 2019)

Bu bağlamda, Şekil 14'te gösterilen DNN modeli mimarisi, DNN'lerin nasıl yapılandırıldığını ve verilerdeki karmaşık ilişkileri nasıl öğrendiğini görsel olarak özetlemektedir. Ayrıca DNN'in çok katmanlı yapısını ve katmanlar arasındaki yoğun bağlantıları açıkça göstermektedir. DNN'ler, bu çok katmanlı yapıları sayesinde giriş katmanından itibaren verileri işleyerek, giderek daha yüksek düzeyde özellikler çıkarabilirler. Derin öğrenme tekniklerinin gelişmesiyle birlikte, DNN'ler metin sınıflandırma gibi görevlerde daha etkili ve verimli hale gelmiş, farklı dillerde ve bağlamlarda yüksek doğrulukla çalışabilir duruma gelmiştir. Bu çalışmada kullanılan DNN modeli için belirlenen parametreler hidden_layer_sizes=(100,), activation='relu', solver='adam', ve random_state=42 şeklindedir.

Evrişimli Sinir Ağları (CNN): CNN, özellikle görüntü işleme için tasarlanmış olsa da, metin sınıflandırma görevlerinde de başarıyla kullanılmaktadır. Metin verilerinde, evrişim katmanları yerel özellikleri çıkararak ve havuzlama (pooling) katmanlarıyla bu özellikleri yoğunlaştırarak metinlerin yapısal ve bağlamsal özelliklerini başarıyla tespit eder. CNN modelleri, özellikle metinlerdeki desenleri ve yapıları saptama konusunda oldukça etkilidir (Kowsari, 2019).

Şekil 15'te görüldüğü gibi, Evrişimli Sinir Ağları (CNN) metin sınıflandırma görevlerinde de etkili bir mimari sunar. Bu mimaride, kelime gömme (word embedding) katmanı, metin verilerini sayısal vektörlere dönüştürerek CNN modeline giriş sağlar. Ardından, evrişim katmanları (convolutional layers) metin içerisindeki yerel özellikleri çıkarır ve havuzlama (pooling) katmanları bu özellikleri yoğunlaştırarak önemli bilgileri öne çıkarır. Bu süreç,

metinlerin yapısal ve bağlamsal özelliklerinin daha iyi anlaşılmasını sağlar. Son olarak, düzleştirme (flatten) katmanı ve tam bağlantılı (fully connected) katmanlar, çıkarılan özellikleri kullanarak sınıflandırma görevini gerçekleştirir.

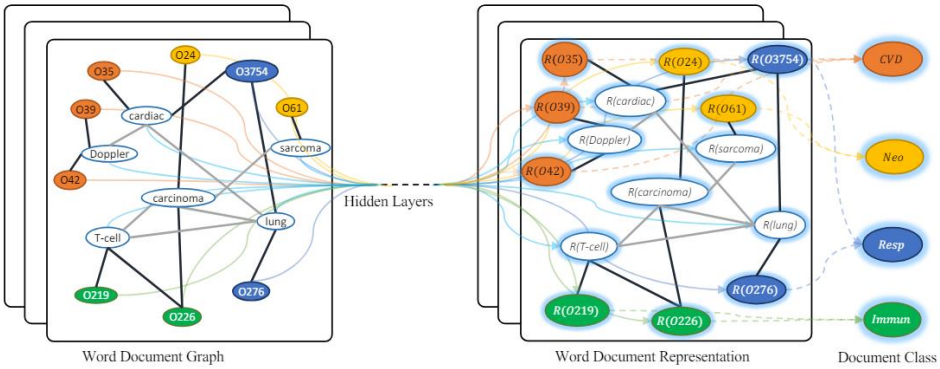


Şekil 15. Metin Sınıflandırmada Evrişimli Sinir Ağları (CNN) Mimarisi (Kowsari, 2019)

CNN modelleri, metinlerdeki desenleri ve yapıları saptama konusundaki yüksek performanslarıyla, metin sınıflandırmada başarılı sonuçlar elde ederler. Bu çalışmada kullanılan CNN modeli için belirlenen parametreler ise epochs=10, batch_size=64, optimizer='adam', ve üç evrişim katmanı (128, 64, 32 filtre sayılarıyla) şeklindedir.

Graf Sinir Ağları (GNN): GNN, metinler gibi yapısal verilerdeki düğümler (varlıklar) ve kenarlar (ilişkiler) arasındaki bağlantıları modellemek için kullanılır. Bu model, metinlerin graf olarak temsil edilmesiyle, bu graf üzerinde düğümler ve kenarlar arasındaki etkileşimleri derinlemesine analiz eder. GNN, özellikle karmaşık yapıların ve ilişkisel bağlantıların olduğu metinlerde, bu bağlantıları anlamlandırma ve sınıflandırma konusunda önemli avantajlar sağlar.

Her bir modelin seçimi, kullanılacak olan veri setinin özellikleri, sınıflandırılacak metinlerin türü ve projenin gereksinimlerine göre yapılmalıdır. Şekil 16'da gösterilen GNN yapısı, metinlerin graf olarak temsil edilmesiyle düğümler ve kenarlar arasındaki ilişkileri derinlemesine analiz etme yeteneğini vurgulamaktadır. Bu yaklaşım, özellikle karmaşık yapıların ve ilişkisel bağlantıların olduğu metinlerde, anlamlı ve doğru sınıflandırmalar yapmada önemli avantajlar sunar. GNN'ler, metinlerdeki varlıklar ve bu varlıklar arasındaki ilişkiler üzerine odaklanarak, metinlerin daha derinlemesine ve bağlamsal analizini sağlar.



Şekil 16. Metin Graf Sinir Ağı (GNN) Yapısı (Yao, 2019)

Bu modellerin uygulanabilirliği ve başarısı, veri setinin büyüklüğü, karmaşıklığı ve sınıflandırma görevinin özellikleri gibi faktörlere bağlı olarak değişkenlik gösterebilir. Örneğin, geniş ve çeşitli veri setlerinde DNN ve CNN modelleri güçlü performans sergilerken, GNN'ler daha spesifik ve ilişki odaklı veri kümelerinde öne çıkabilir. Ayrıca, veri setinin karmaşıklığı ve modelin eğitim süresi de model seçimini etkileyen kritik faktörlerdir. Derin sinir ağları ve evrişimli sinir ağları, büyük veri kümeleri üzerinde daha uzun eğitim sürelerine ihtiyaç duyarken, graf sinir ağları, veri yapısının karmaşıklığına bağlı olarak daha esnek ve adaptif olabilir. Bu nedenle, model seçimi yapılırken projenin gereksinimleri, veri setinin yapısı ve sınıflandırma hedefleri dikkatlice değerlendirilmelidir.

3.9. Metrikler

Metin sınıflandırma performansını değerlendirmek için çeşitli metrikler kullanılır. Bu metrikler, modelin doğruluğunu, hatalarını ve genel etkinliğini ölçmek için kritik öneme sahiptir. Bu bölümde, doğruluk (precision), geri çağırma (recall), F1-Skor ve doğruluk (accuracy) gibi temel performans ölçütleri tanıtılacak ve bu metriklerin hesaplanmasında kullanılan formüller açıklanacaktır. Ayrıca, metin sınıflandırma süreçlerinde kullanılan karşıtlık matrisi (confusion matrix) ve kayıp-doğruluk eğrisi (loss-accuracy curve) gibi araçlar da detaylandırılacaktır. Belirtilen bu metrikler, TP (True Positive - Doğru Pozitif), TN (True Negative - Doğru Negatif), FP (False Positive - Yanlış Pozitif) ve FN (False Negative - Yanlış Negatif) terimleri üzerinden hesaplanır. TP, modelin doğru olarak Pozitif tahmin ettiği örnekleri, TN doğru olarak Negatif tahmin ettiği örnekleri, FP yanlış olarak Pozitif tahmin ettiği örnekleri ve FN yanlış olarak Negatif tahmin ettiği örnekleri ifade eder. Bu temel terimler, sınıflandırma modelinin performansını doğru bir şekilde değerlendirmek için kritik öneme

sahiptir ve bu sayede modelin hangi sınıfları doğru veya yanlış tahmin ettiğini anlamak mümkün olur. Performans ölçütleri ve formüller:

Kesinlik (Precision): Kesinlik, Pozitif olarak tahmin edilen değerlerin aslında ne kadarının Pozitif olduğunu gösterir (Denklem 11).

$$\text{Precision} = \frac{\text{TP}}{TP+FN} \tag{11}$$

Duyarlılık (Recall): Duyarlılık (Recall), aslında Pozitif olması gereken örneklerin ne kadarının model tarafından doğru bir şekilde Pozitif olarak tahmin edildiğini gösteren bir metriktir (Denklem 12).

$$\text{Recall} = \frac{\text{TP}}{TP+FP} \tag{12}$$

F1 Skor: F1 Skoru, Kesinlik ve Duyarlılık değerlerinin harmonik ortalamasını gösterir ve bu metrik, özellikle dengesiz veri kümelerinde model performansını değerlendirmek için kullanılır (Denklem 13).

$$\text{F1 Score} = \frac{2 \times \text{Recall} \times \text{Precision}}{Recall+Precision} \tag{13}$$

Doğruluk (Accuracy): Doğruluk değeri, modelde doğru tahmin edilen alanların toplam veri setine oranıyla hesaplanır (Denklem 14).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{TP+TN+FP+FN} \tag{14}$$

Karşıtlık Matrisi (Confusion Matrix): Karşıtlık Matrisi, sınıflandırma modelinin performansını değerlendirmek için kullanılan bir tablodur. Her sınıf için gerçek ve tahmin edilen sınıf sayısını gösterir. Bu matris, modelin hangi sınıfları doğru veya yanlış tahmin ettiğini anlamak için önemli bir araçtır.

Bir karmaşıklık matrisi C, $C_{i,j}$ olarak tanımlanır ve i grubunda bilinen ve j grubunda tahmin edilen gözlem sayısına eşittir. İkili sınıflandırmada, true negative (doğru negatif) sayısı $C_{0,0}$ false negative (yanlış negatif) sayısı $C_{1,0}$ true positive (doğru pozitif) sayısı $C_{1,1}$ ve false positive (yanlış pozitif) sayısı $C_{0,1}$ olarak tanımlanır (Tablo 6).

Tablo 6. İkili Sınıflandırma İçin Karmaşıklık Matrisi Yapısı

		Tahmin	
		Pozitif	**Negatif**
Gerçek	**Pozitif**	TP ($C_{1,1}$)	FN ($C_{1,0}$)
	Negatif	FP ($C_{0,1}$)	TN ($C_{0,0}$)

Karmaşıklık matrisi, modelin performansını değerlendirmek ve hangi tür hataların daha sık yapıldığını belirlemek için kritik öneme sahiptir. Bu sayede, modelin iyileştirilmesi gereken alanlar belirlenebilir.

Kayıp-Doğruluk Eğrisi (Loss-Accuracy Curve): Kayıp-Doğruluk Eğrisi, modelin eğitim sürecindeki kaybını (hata oranını) ve doğruluğunu (accuracy) gösterir. Kayıp eğrisi, modelin ne kadar hatalı tahminler yaptığını, doğruluk eğrisi ise modelin ne kadar doğru tahminler yaptığını gösterir. Kayıp fonksiyonu genellikle Denklem 15' teki gibi tanımlanır:

$$L(yg, yt) = \frac{1}{n} \sum_{i=1}^{n} (\mathrm{yg}_i - \mathrm{yt}_i)^2 \qquad (15)$$

Bu formül, gerçek değerler **yg** ile tahmin edilen değerler **yt** arasındaki ortalama karesel hatayı ifade eder. Doğruluk ise genellikle Denklem 15' teki şekilde hesaplanır.

Grafik olarak bu eğriler şu şekilde görselleştirilebilir:

- **Kayıp Eğrisi:** Yatay eksen (Eğitim epoch sayısı), Dikey eksen (Kayıp değeri)
- **Doğruluk Eğrisi:** Yatay eksen (Eğitim epoch sayısı), Dikey eksen (Doğruluk oranı)

Bu grafikler, modelin eğitim sürecindeki performansını net bir şekilde görselleştirir ve değerlendirmemizi sağlar. Ek olarak, modelin aşırı uyum (overfitting) veya yetersiz uyum (underfitting) yapıp yapmadığını belirlemek için kullanılır. Eğitim ve doğrulama setleri için ayrı ayrı çizilen bu eğriler, modelin genel performansını değerlendirmede kritik rol oynar.

4. BULGULAR

Bu çalışma, Türkçe ve İngilizce metin sınıflandırma veri setleri üzerinde yoğunlaşarak, metin ön işleme süreçlerinin detaylı bir şekilde nasıl uygulandığını ve bu süreçlerin sonucunda metinlerdeki varlıkların ve ilişkilerin nasıl tespit edildiğini kapsamlı bir şekilde ele almaktadır. Elde edilen metinler, önceden eğitilmiş Named Entity Recognition (NER) modelleri yardımıyla tanımlanan varlıklar ile zenginleştirilmiş ve Wikidata'dan elde edilen bilgilerle daha da derinlemesine işlenmiştir. Bu metinler, daha sonra BERT tabanlı ön eğitimli bir model kullanılarak vektörlendirilmiştir. Sınıflandırma süreçleri için çeşitli yöntemler tercih edilmiştir; bu yöntemler arasında Support Vector Machines (SVM), Logistic Regression, Deep Neural Networks (DNN), Convolutional Neural Networks (CNN) ve Graph-based Neural Networks (GNN) bulunmaktadır. Her bir modelin performans doğruluk oranları ve yapılan iyileştirmeler, geniş çapta analiz edilerek değerlendirilmiştir.

Bu araştırma, metinler arası ilişkileri daha etkin bir şekilde anlamak ve bu bilgileri sınıflandırma algoritmalarıyla başarılı bir şekilde işleyebilmek amacıyla yapılandırılmıştır. Farklı metin veri setlerine uygulanan sınıflandırma tekniklerinin metinler arası ilişkisel ve yapısal özellikleri nasıl çıkardığı ve bu bilgileri nasıl işlediği detaylı olarak incelenmiştir. Araştırma sonuçları, kullanılan metodolojilerin etkinliğini göstermekte ve metin sınıflandırma alanında ilişkisel analizin potansiyelini vurgulamaktadır.

Bulgular bölümü, elde edilen verilerin analizi ve bu verilerden çıkarılan sonuçların grafikler, tablolar ve ilgili istatistiksel analizlerle desteklenmiş şekilde sunulmuştur. Her bir sınıflandırma deneyinin metodolojisi, uygulanan işlemler, elde edilen sonuçlar ve bu sonuçların çalışmanın genel amaçlarına nasıl katkıda bulunduğu detaylı bir şekilde irdelenmiştir. Çalışmanın literatürdeki diğer yöntemlerle karşılaştırıldığında, metinler arası ilişkisel sınıflandırmanın önemini ve potansiyelini daha da belirginleştirerek bu alanda önemli bir ilerleme kaydettiği gösterilmektedir.

Araştırmanın bulguları, kullanılan sınıflandırma modellerinin etkinliklerini ve sınırlılıklarını değerlendirirken, aynı zamanda tezin temel amaçlarına yönelik derinlemesine bilgiler sunmaktadır. Bu sonuçlar, analiz süreci boyunca formüle edilen çeşitli sorulara ve varsayımlara ışık tutarak, çalışmanın bilimsel katkısını daha geniş bir perspektiften ortaya koymaktadır. Önerilen modelin geliştirilme süreci, yapılandırılması ve uygulanması ile elde

edilen sonuçlar, tezin temel amaçları ile nasıl uyum sağladığı derinlemesine incelenmiştir. Ayrıca, modelin potansiyel uygulamaları, pratik sınırlamaları ve gelecekteki çalışmalar için öneriler de detaylı bir şekilde ele alınmıştır. Bu sonuçlar, sınıflandırma modellerinin pratik uygulamalarına yönelik yol gösterici nitelikte olup, bu alanda devam eden ve gelecekteki araştırmalar için temel bir referans noktası sunmaktadır.

4.1. İlişkisel Metin Sınıflandırma Modellerinin Performans Analizi

Bu bölüm, araştırmanın çeşitli sınıflandırma modelleri üzerindeki odaklanmasını detaylandırmaktadır. Özellikle, Türkçe ve İngilizce metin veri setlerinde uygulanan ön işleme ve vektörleştirme süreçleri sonrasında, SVM, Lojistik Regresyon, DNN, CNN ve GNN modelleri kullanılarak gerçekleştirilen sınıflandırma işlemlerinin sonuçları sunulmaktadır. Her bir modelin etkinliği, metinler arası ilişkisel ve yapısal özellikleri çıkarabilme kapasiteleri ile değerlendirilmiş ve bu modellerin performansı, belirli metriklerle ölçülerek karşılaştırmalı bir analizle incelenmiştir.

Sınıflandırma süreçlerinin sonuçları, her bir modelin doğruluk oranları, F1 skorları ve diğer ilgili performans metrikleri üzerinden değerlendirilmiştir. Ayrıca, bu modellerin metin veri setlerindeki varlıkları ve ilişkileri nasıl işlediği, özellikle de ilişkisel bilgileri sınıflandırma başarısına etkisini gösteren bulgular detaylı bir şekilde tartışılmıştır. Analizler, metin sınıflandırma tekniklerinin, kompleks metin yapılarını ve dilin anlamsal zenginliklerini nasıl modele ettiğini ve bu yaklaşımların veri bilimi ve doğal dil işleme disiplinlerinde nasıl stratejik önem taşıdığını ortaya koymaktadır.

Bu analitik inceleme, algoritmalardaki performans farklılıklarını ortaya koyarak, hangi sınıflandırma modelinin hangi tür veri setleri ve senaryolar için daha uygun olduğuna dair kritik bilgiler sağlamaktadır. Bulgular, sınıflandırma modellerinin pratik uygulamaları ve gelecekteki araştırmalar için önemli ipuçları sunmakta, bu modellerin sınırlılıkları ve geliştirilme potansiyelleri üzerine önerilerde bulunmaktadır.

Bu bölümde ayrıca, sınıflandırma tekniklerinin geliştirilmesi ve iyileştirilmesi için önerilen stratejiler, pratik sınırlamalar ve modelin uygulama alanlarına yönelik tartışmalar yer almaktadır. Araştırma, bu tekniklerin mevcut ve potansiyel kullanım alanlarını genişletmeye yönelik önerilerle son bulmaktadır, bu da metin sınıflandırma alanında devam eden ve gelecekteki araştırmalar için temel bir referans noktası sunmaktadır.

4.1.1. TTC-4900 Veri Seti ile Performans Analizi

Bu bölüm, Türkçe metin sınıflandırma için kullanılan TTC-4900 veri seti üzerinde gerçekleştirilen deneysel çalışmaların sonuçlarını detaylandırmaktadır. Çalışmada, Logistic Regression, Linear SVM, CNN ve DNN olmak üzere çeşitli sınıflandırma algoritmaları kullanılmıştır. Bu algoritmalar, ilişkisel bilgilerin entegrasyonu öncesi ve sonrasında değerlendirilmiş, elde edilen sonuçlar karşılaştırmalı olarak analiz edilmiştir (Tablo 7).

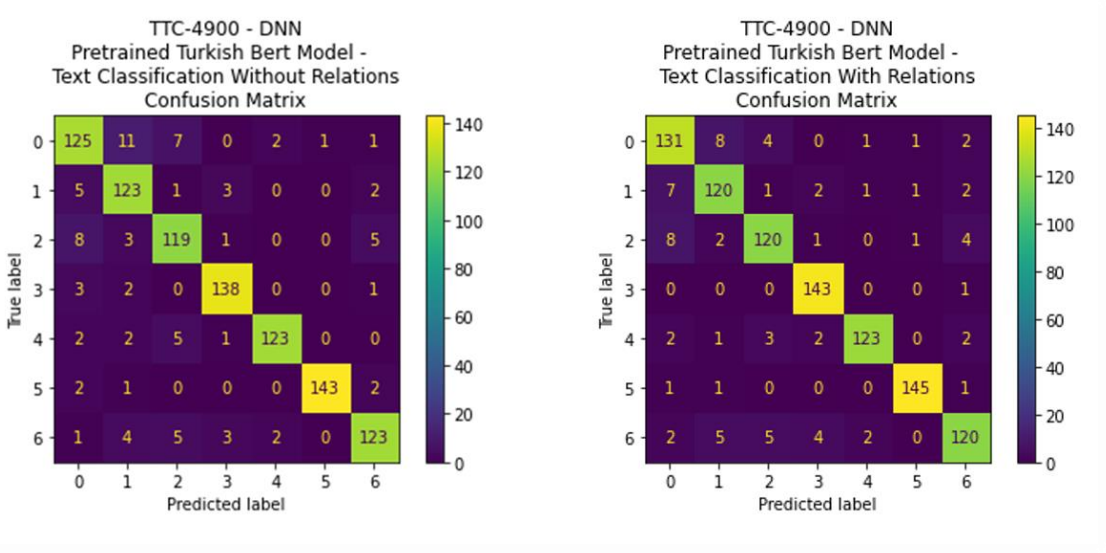
Tablo 7. TTC-4900 Veri Seti İçin İlişkisel Metin Sınıflandırma Sonuçları

	İlişkisiz Durum (Without Relations)				**İlişkili Durum (With Relations)**				**İyileşme (Accuracy Improvement) (%)**
Sınıflandırma	**Prec.**	**Rec.**	**F1-Score**	**Acc.**	**Prec.**	**Rec.**	**F1-Score**	**Acc.**	
LogisticReg.	88.2	87.8	87.9	**87.8**	90.8	90.8	90.8	**90.8**	**3.42**
LinearSVM	91.4	91.3	91.3	**91.3**	92.0	92.0	92.0	**92.0**	**0.77**
CNN	86.8	86.0	86.0	**86.0**	89.2	89.1	89.1	**89.1**	**3.61**
DNN	91.4	91.3	91.3	**91.3**	92.1	92.1	92.1	**92.1**	**0.88**

Her model için ilişkisel ve ilişkisiz durumlarda elde edilen Precision, Recall, F1-Score ve Accuracy metrikleri, modelin metinleri nasıl sınıflandırdığını ve ilişkisel bilginin sınıflandırma doğruluğuna etkisini göstermektedir. İlişkisel bilgi entegrasyonunun sınıflandırma performansına etkisi, her bir model için yüzdelik iyileşme oranlarıyla ölçülmüştür.

İlişkisel bilgilerin dahil edilmesiyle özellikle CNN ve Logistic Regression modellerinde gözlemlenen performans artışları dikkat çekicidir. Logistic Regression modeli ilişkisiz durumda %87.8 olan doğruluk oranını, ilişkili bilgilerle %90.8'e çıkarmış, böylece %3.42'lik bir iyileşme sağlamıştır. CNN modelinde ise %3.61'lik bir iyileşme ile doğruluk %86.0'dan %89.1'e yükselmiştir.

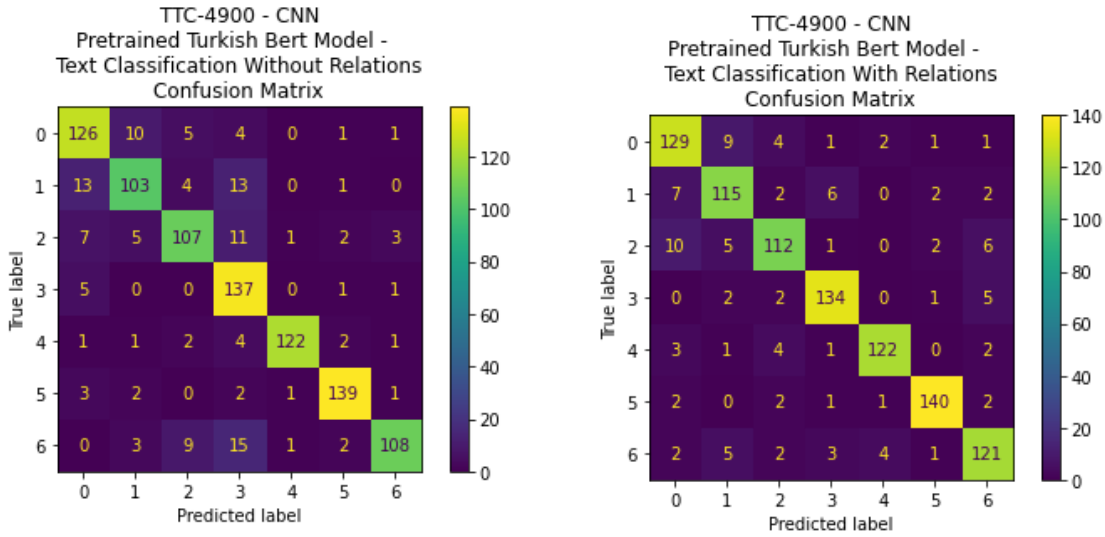
Karmaşıklık matrisleri, her bir modelin sınıflandırma doğruluğunu görsel olarak sunar ve hangi sınıfların daha başarılı sınıflandırıldığını, hangilerinde zorluklar yaşandığını ortaya koyar. Bu matrisler, modelin tahminlerinin gerçek değerlerle ne kadar uyumlu olduğunu ve sınıflandırma hatalarının dağılımını detaylı bir şekilde gösterir.



Şekil 17. TTC-4900 Metin Veri Seti ve DNN Sınıflandırıcı Karmaşıklık Matrisleri

Şekil 17' te, TTC-4900 metin veri seti için bir Derin Sinir Ağı (DNN) kullanılarak yapılan metin sınıflandırma sonuçları karşımıza çıkıyor. İki farklı karmaşıklık matrisi, önceden eğitilmiş bir Türkçe BERT modeli kullanılarak yapılan sınıflandırmanın etkinliğini göstermektedir. Soldaki matris, ilişkisel bilgilerin dahil edilmeden elde edilen sonuçları sergilerken, sağdaki matris ilişkisel bilgilerin eklenmesiyle elde edilen sınıflandırma sonuçlarını gösteriyor. Her iki matriste de dikey eksen modelin tahminlerini, yatay eksen ise gerçek etiketleri ifade ediyor. Diagonal üzerindeki değerler her sınıf için doğru sınıflandırılan örnekleri (true positives) temsil ederken, diagonal dışındaki değerler yanlış sınıflandırmaları (false positives ve false negatives) göstermektedir.

Matrislerin karşılaştırılmasında, ilişkisel bilginin kullanımının bazı sınıflar için doğru sınıflandırma sayısını arttırdığı gözlemleniyor. Özellikle 0. ve 3. sınıfta doğru tahmin sayısını dikkate değer bir artış gösteriyor. Bu, ilişkisel bilginin eklenmesinin sınıflandırma başarısını belirli sınıflarda iyileştirebileceğini işaret ediyor. Buna karşın, diğer sınıflarda doğru sınıflandırma sayılarında değişkenlik gözleniyor ki bu, ilişkisel bilginin tüm sınıflarda aynı etkiyi göstermeyebileceğini gösteriyor. Bu iki matrisin incelenmesi, modelin bazı sınıflar için performansını geliştirirken, diğerlerinde nötr veya negatif bir etki yaratabileceğini göstermektedir. Bu gözlemler, doğru bir şekilde sınıflandırılmış ve sınıflandırılmamış örneklerin sayısını görselleştiren matrislerin, modelin performansını anlamak ve iyileştirmeler yapmak için nasıl değerli araçlar olabileceğine işaret etmektedir.

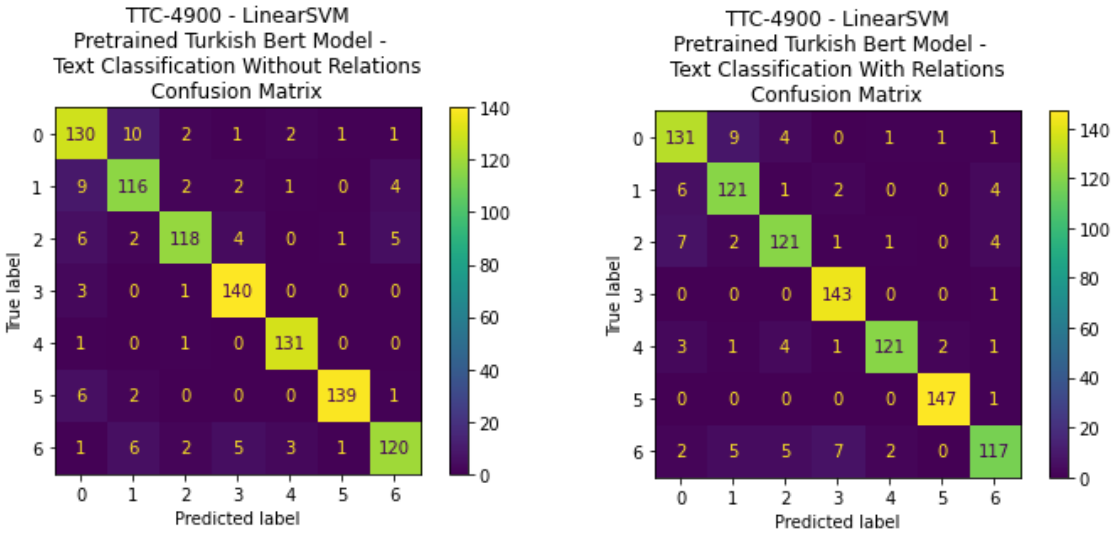


Şekil 18. TTC-4900 Metin Veri Seti ve CNN Sınıflandırıcı Karmaşıklık Matrisleri

Şekil 18, TTC-4900 metin veri setinde gerçekleştirilen sınıflandırma işleminin iki farklı karmaşıklık matrisiyle temsil edilmiş sonuçlarını gözler önüne sermektedir. Sol taraftaki karmaşıklık matrisi, ilişkisel bilgi olmadan uygulanan CNN modelinin performansını, sağ taraftaki matris ise ilişkisel bilgi eklenmiş durumda modelin nasıl performans gösterdiğini gösteriyor. Her iki durumda da, matrislerin yatay eksenlerindeki sayılar, gerçek sınıf etiketlerini, dikey eksenlerindeki sayılar ise modelin tahmin ettiği sınıf etiketlerini ifade ediyor.

Diagonal üzerindeki sayılar doğru sınıflandırma sayılarını (true positives), diagonalin dışındaki sayılar ise modelin yaptığı yanlış tahminleri (false positives ve false negatives) gösteriyor. İlişkisel bilginin eklenmesi sonucunda 0, 1, 2, 5 ve 6 gibi belirli sınıflarda doğruluk oranlarında artış gözleniyor ki bu, söz konusu bilginin modelin sınıflandırma kabiliyetini güçlendirdiğini düşündürmektedir.

İlişkisel bilginin genel performans üzerindeki etkisi pozitif gibi görünse de, sınıflandırma başarısının her bir sınıf düzeyinde detaylı bir şekilde incelenmesinin önemi bu sonuçlarla bir kez daha vurgulanmış oluyor.



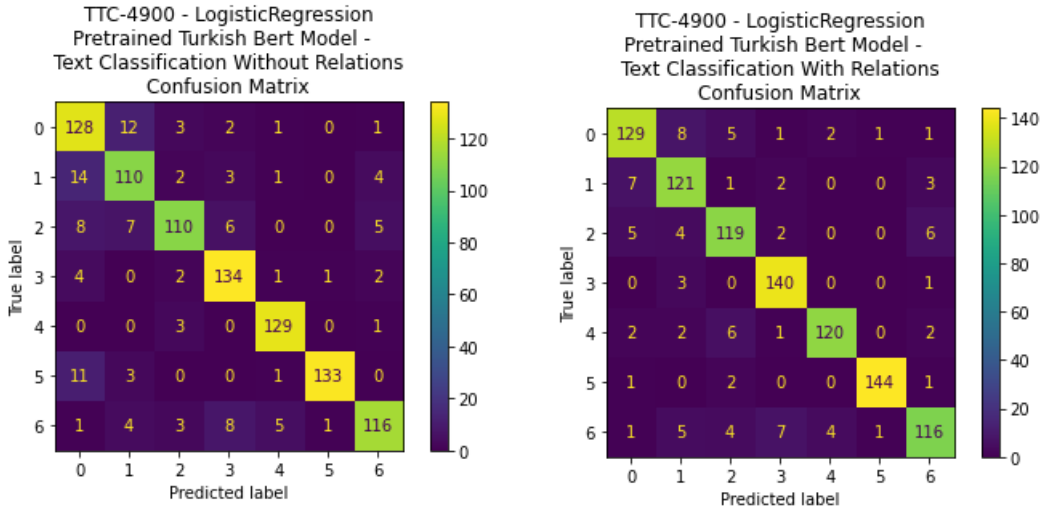
Şekil 19. TTC-4900 Metin Veri Seti ve LinearSVM Sınıflandırıcı Karmaşıklık Matrisleri

Şekil 19, önceden eğitilmiş Türkçe BERT modeliyle desteklenen Lineer SVM sınıflandırıcısının, TTC-4900 veri setindeki performansını iki farklı koşul altında sergileyen karmaşıklık matrislerini içermektedir. Soldaki matris, sınıflandırma işleminin ilişkisel bilgi olmaksızın nasıl sonuçlandığını; sağdaki matris ise ilişkisel bilgi eklendiğinde elde edilen sonuçları göstermektedir.

Matrisler, modelin sınıf tahminlerinin doğruluğunu renk yoğunluğu ve sayısal değerlerle sunarken, dikey eksenler modelin tahmin ettiği, yatay eksenler ise gerçek sınıf etiketlerini temsil ediyor. Diagonaldeki değerler modelin her sınıf için doğru tahminlerini (true positives), diagonal dışındaki değerler yanlış tahminleri (false positives ve false negatives) ifade ediyor.

İlişkisel bilginin kullanımının, doğru tahminlerde bir artışa sebep olduğu gözlemlenirken; bazı sınıfların doğru tahmin sayılarında bir düşüş yaşandığı göze çarpıyor.

Genel olarak, matrislerden alınan verilerin analizi, model geliştirme sürecinde ilişkisel veri entegrasyonunun önemini ve sınıflandırıcı seçiminin sonuçlar üzerindeki etkisini net bir şekilde göstermektedir.



Şekil 20. TTC-4900 Metin Veri Seti ve LogisticRegression Sınıflandırıcı Karmaşıklık Matrisleri

Şekil 20'de, TTC-4900 metin veri seti kullanılarak, Lojistik Regresyon modeli ile yapılan sınıflandırmanın sonuçları iki karmaşıklık matrisi aracılığıyla sunulmuştur. İlişkisel bilgi kullanılarak ve kullanılmadan elde edilen sonuçlar karşılaştırılıyor; ilişkisel bilginin eklenmesiyle, sınıflandırma doğruluğunda sınıflara özel iyileşmelerin olduğu görülüyor. Özellikle, 1, 3 ve 5 numaralı sınıfların doğru sınıflandırma sayılarında artış dikkat çekiyor, bu da modelin ilişkisel bağlamı dikkate aldığında daha isabetli tahminler yapabildiğine işaret ediyor.

Bu sonuçlar, ilişkisel bilginin sınıflandırma modellerinin performansını iyileştirme konusunda belirgin bir etkiye sahip olabileceğini göstermektedir. Özellikle, ilişkisel bilgilerin entegrasyonu, modelin belirli sınıflar üzerindeki doğruluk oranlarını artırarak sınıflandırma sürecinin genel etkinliğini güçlendirmiştir. Bu iyileşmeler, ilişkisel bilginin sınıflandırma algoritmalarına dahil edilmesinin önemini ortaya koymakta ve metin sınıflandırma görevlerinde daha hassas ve doğru sonuçlar elde etmek için gerekli olabileceğini işaret etmektedir.

Elde edilen bulgular, bu tür bilgilerin sınıflandırma sistemlerine entegrasyonunun pratikte nasıl bir katkı sağlayabileceğini daha iyi anlamak için sağlam bir temel teşkil ediyor. Bu çalışma, metin sınıflandırma algoritmalarının geliştirilmesinde ilişkisel bilgilerin kullanımı yönünde atılabilecek adımlar için değerli bir referans noktası sunmaktadır.

4.1.2. BBC-News Veri Seti ile Performans Analizi

Bu bölüm, İngilizce metin sınıflandırma için kullanılan BBC-News veri seti üzerinde gerçekleştirilen sınıflandırma deneylerini ele almakta ve ilişkisel bilginin sınıflandırma performansına etkilerini değerlendirmektedir. Çalışmada, Logistic Regression, Linear SVM, CNN, ve DNN gibi çeşitli sınıflandırma modelleri kullanılmıştır. Her modelin performansı, ilişkisel bilgiler entegre edilmeden önce ve sonra karşılaştırmalı olarak incelenmiştir (Tablo 8).

Tablo 8. BBC-News Veri Seti İçin İlişkisel Metin Sınıflandırma Sonuçları

	İlişkisiz Durum (Without Relations)				**İlişkili Durum (With Relations)**				**İyileşme (Accuracy Improvement) (%)**
Sınıflandırma	**Prec.**	**Rec.**	**F1-Score**	**Acc.**	**Prec.**	**Rec.**	**F1-Score**	**Acc.**	
LogisticReg.	83.2	77.6	75.6	**77.6**	97.7	97.6	97.6	**97.6**	**25.78**
LinearSVM	94.4	94.2	94.1	**94.2**	98.7	98.7	98.7	**98.7**	**4.78**
CNN	90.1	89.7	89.6	**89.7**	96.7	96.7	96.7	**96.7**	**7.81**
DNN	97.6	97.6	97.6	**97.6**	97.8	97.8	97.8	**97.8**	**0.21**

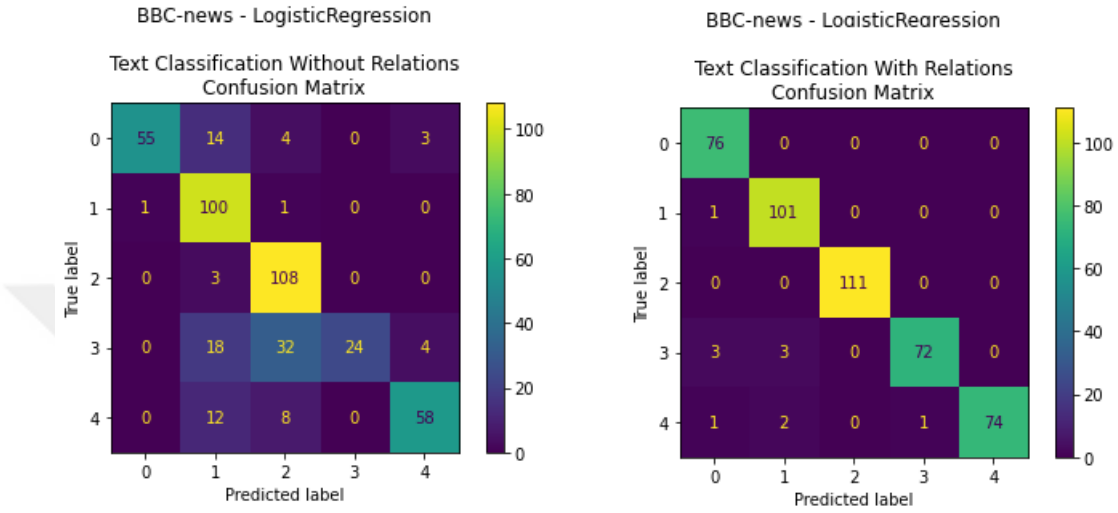
Linear SVM ve CNN modelleri, ilişkisel bilgilerin sınıflandırma süreçlerine entegrasyonu ile önemli performans artışları göstermiştir. Özellikle Linear SVM modeli %94.2 doğruluk oranından, %98.7'ye yükselerek %4.78'lik bir iyileşme sağlamıştır. CNN modeli ise %89.7'den %96.7'ye ulaşarak, %7.81'lik etkileyici bir performans artışı sergilemiştir. Bu sonuçlar, ilişkisel bilginin modellerin doğruluk oranlarını önemli ölçüde artırabileceğini ve sınıflandırma işlemlerinde bu tür bilgilerin entegrasyonunun, model performansını optimize etmede kritik bir rol oynayabileceğini göstermektedir.

Logistic Regression modeli ise %25.78'lik olağanüstü bir iyileşme ile dikkat çekmektedir. İlişkisiz durumda %77.6 olan doğruluk oranı, ilişkiler dahil edildiğinde %97.6'ya yükselmiştir. Bu büyük sıçrama, ilişkisel bilgilerin modelin sınıflandırma kabiliyetini ne ölçüde geliştirebileceğinin açık bir göstergesidir.

Öte yandan, DNN modeli yüksek başlangıç doğruluğuna rağmen (%97.6), ilişkisel bilgilerle sadece minimal bir iyileşme (%0.21) göstermiştir. Bu, bazı yüksek performanslı

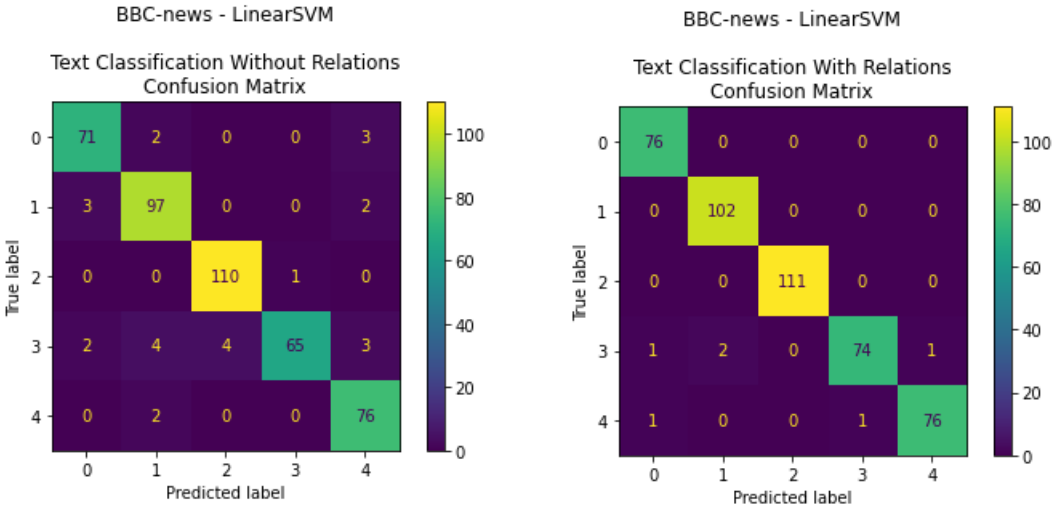
modellerin zaten var olan ilişkileri kendi iç dinamikleriyle tespit edebildiği veya farklı tür bilgilerden yararlanarak benzer sonuçlara ulaşabildiği anlamına gelebilir.

Bu analizler, ilişkisel metin sınıflandırmasının potansiyelini daha da belirginleştirerek, bu alanda daha ileri araştırmalar için sağlam bir temel oluşturmuştur.



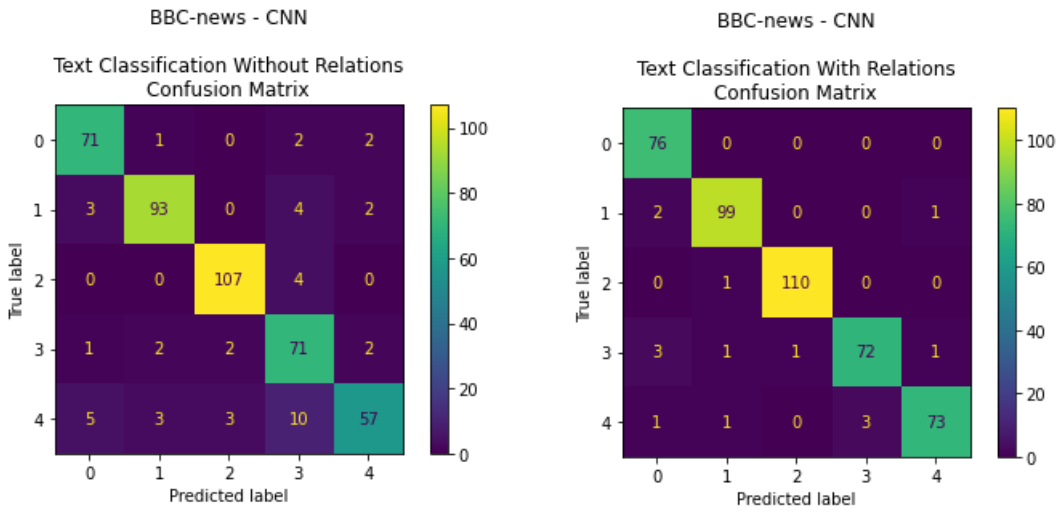
Şekil 21. BBC-news Metin Veri Seti ve LogisticRegression Sınıflandırıcı Karmaşıklık Matrisleri

Şekil 21' de sergilenen BBC-news veri seti üzerindeki Lojistik Regresyon sınıflandırma deneyleri, ilişkisel bilginin entegrasyonunun modelin performansını anlamak açısından kritik öneme sahip olduğunu vurgulamaktadır. İki karmaşıklık matrisi, ilişkisel öğeler eklenmeden önceki ve sonraki durumları görselleştirmekte ve önemli farklılıkları öne çıkarmaktadır. İlişkisel bilgilerin modellemeye dahil edilmesiyle, sınıf 0, 3 ve 4'te yanlış sınıflandırılan örneklerin azaldığı ve doğru sınıflandırılanların sayısının belirgin bir şekilde arttığı gözlenmektedir. Bu artış, modelin, sınıflar arası farkları daha açık bir şekilde ayırt ettiğini ve her bir sınıfa ait örnekleri daha doğru bir şekilde tanımlayabildiğini göstermektedir. Özellikle, modelin 3 ve 4 numaralı sınıflarda gösterdiği performans artışı, ilişkisel bilginin sadece doğruluk oranını artırmakla kalmayıp, modelin daha kompleks örnekleri bile başarıyla işleyebilmesini sağladığını işaret eder. Bu bulgular, sınıflandırma sürecine ilişkisel bilgi entegrasyonunun, algoritmaların daha derinlemesine öğrenmesine ve gelişmiş tahmin kabiliyetleri geliştirmesine olanak tanıdığını göstermektedir. Bu tür bir entegrasyonun, geniş bir yelpazedeki uygulamalar için metin sınıflandırma modellerinin etkinliğini artırabileceği ve model geliştirmenin gelecekteki yönlerine ışık tutabileceği sonucuna varılmaktadır.



Şekil 22. BBC-news Metin Veri Seti ve LinearSVM Sınıflandırıcı Karmaşıklık Matrisleri

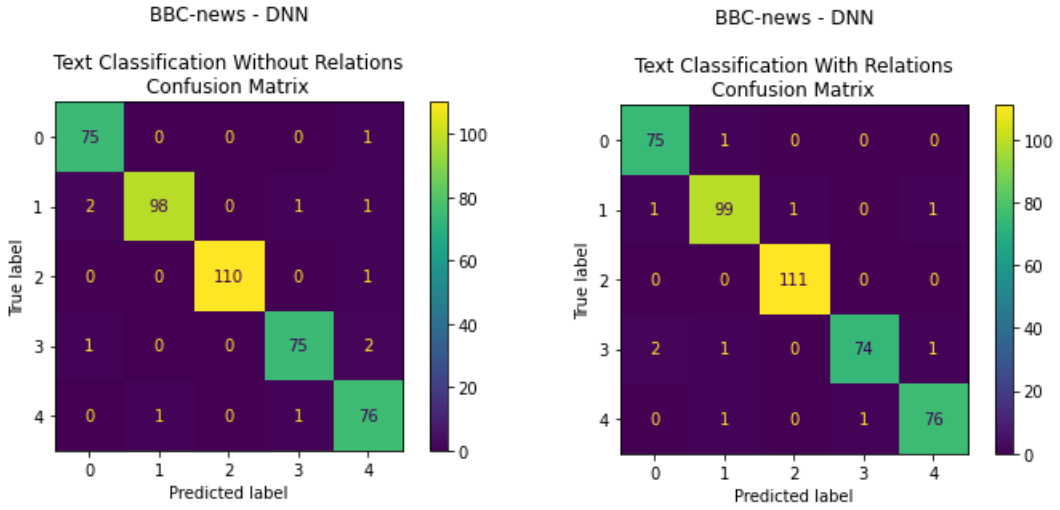
Şekil 22'te, BBC-news veri setinde kullanılan Lineer SVM modelinin karmaşıklık matrisleri aracılığıyla performansı gösterilmektedir. İlişkisel bilgi ile zenginleştirilen model (sağ matris), bazı sınıflarda sınıflandırma hatasının belirgin şekilde azaldığını gösteriyor; sınıf 0,1 ve 3' ün keskin bir şekilde iyileştiği, özellikle sınıf 3' teki doğru tahminlerin artışının, modelin bu sınıftaki örnekleri daha net bir şekilde ayırt edebildiğini gösteriyor. Bu, ilişkisel bilginin sınıflandırma modelinin hassasiyetini ve genel performansını artırma kapasitesine dair somut bir kanıt sunar.



Şekil 23. BBC-news Metin Veri Seti ve CNN Sınıflandırıcı Karmaşıklık Matrisleri

Şekil 23, BBC-news veri setindeki CNN modelinin performansını gösteren karmaşıklık matrislerini içerir. İlişkisel bilginin entegrasyonunun modelin sınıflandırma başarısını tüm

sınıflar boyunca nasıl iyileştirdiğini açıkça görüyoruz. Özellikle, sınıf 0,1,2 ve 4'teki yanlış pozitiflerin azaldığı ve doğru pozitiflerin sayısının arttığı belirgindir.



Şekil 24. BBC-news Metin Veri Seti ve DNN Sınıflandırıcı Karmaşıklık Matrisleri

Şekil 24, BBC-news veri seti üzerinde gerçekleştirilen Derin Sinir Ağı (DNN) sınıflandırma deneyinin sonuçlarını karşılaştırmalı karmaşıklık matrisleriyle sergiliyor. İlişkisel bilginin entegrasyonu sonrasında, sınıf 1 ve 3'teki doğru sınıflandırma sayılarında hafif bir iyileşme olduğunu görebiliyoruz. Bu iyileşme, modelin ilişkisel bilgiyi kullanarak belirli sınıflarda sınıflandırma doğruluğunu artırma potansiyeline işaret etmekle birlikte, genel olarak DNN modelinin zaten yüksek düzeyde doğru sınıflandırma yapabilme kapasitesine sahip olduğunu da gösteriyor. İlişkisel bilginin bu model üzerindeki etkisinin sınırlı olması, belki de modelin öğrenme kapasitesinin ve kendi içsel özelliklerinin, zaten bu bilgileri kısmen modellemiş olabileceğine dair bir işaret olabilir. Bu, ilişkisel bilginin model üzerindeki etkisinin her zaman açıkça gözlemlenemeyebileceğini, ancak spesifik durumlar ve sınıflar için incelendiğinde belirli iyileştirmelerin mümkün olduğunu göstermektedir.

4.1.3. 20Newsgroups Veri Seti ile Performans Analizi

Bu bölüm, çeşitli metin sınıflandırma modellerinin 20Newsgroups veri seti üzerindeki performansını değerlendirirken, ilişkisel bilgilerin entegrasyonunun sınıflandırma sonuçlarına olan etkisini inceler. Çalışmada kullanılan modeller arasında Logistic Regression, Linear SVM, CNN, ve DNN bulunmaktadır. Modellerin hem ilişkisel bilgiler dikkate alınmadan önceki hem de sonraki performansları, kapsamlı bir şekilde karşılaştırılmıştır (Tablo 9).

Tablo 9. 20Newsgroups Veri Seti İçin İlişkisel Metin Sınıflandırma Sonuçları

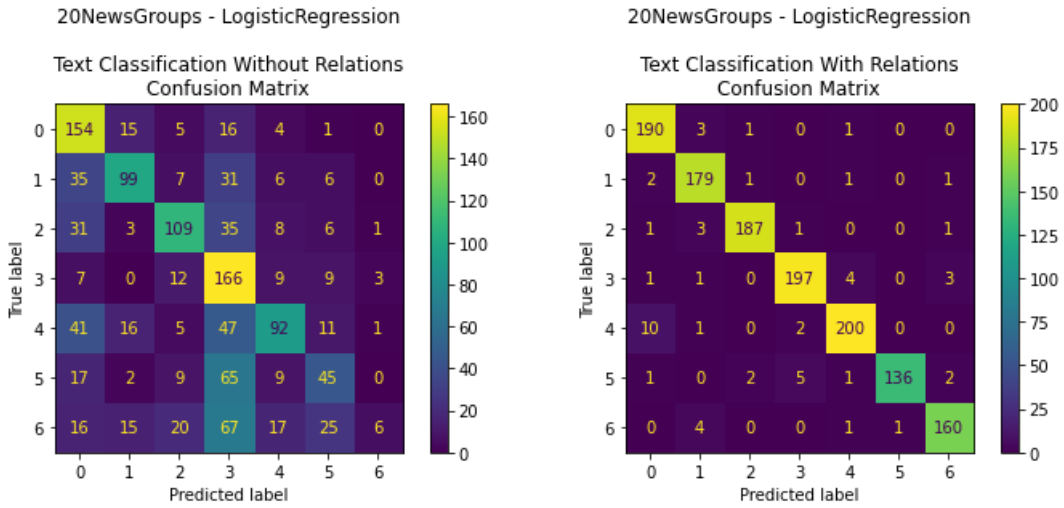
Sınıflandırma	İlişkisiz Durum (Without Relations)				İlişkili Durum (With Relations)				İyileşme (Accuracy Improvement) (%)
	Prec.	Rec.	F1-Score	Acc.	Prec.	Rec.	F1-Score	Acc.	
LogisticReg.	80.0	76.5	73.3	**76.5**	95.9	95.8	95.8	**95.8**	**25.23**
LinearSVM	80.0	78.8	78.3	**78.8**	96.9	96.9	96.9	**96.9**	**22.97**
CNN	75.1	73.3	73.3	**73.3**	91.6	91.5	91.5	**91.5**	**24.83**
DNN	90.6	90.5	90.5	**90.5**	96.6	96.6	96.6	**96.6**	**6.75**

Linear SVM modeli, ilişkisel bilgilerin dikkate alınmasıyla %78.8 doğruluk oranından %96.9'a yükselmiş ve %22.97'lik dikkate değer bir iyileşme kaydetmiştir. CNN modeli de benzer bir gelişme göstererek, %73.3 doğruluk oranından %91.5'e ulaşmıştır, bu da %24.83'lük büyük bir performans artışıdır.

Logistic Regression modelinin performansı, ilişkisel bilgiler dahil edildiğinde %76.5'ten %95.8'e yükselmiş, böylece %25.23'lük önemli bir iyileşme göstermiştir. Bu, ilişkisel bilgilerin sınıflandırma performansını önemli ölçüde artırabileceğini gösteren bir başka örnektir.

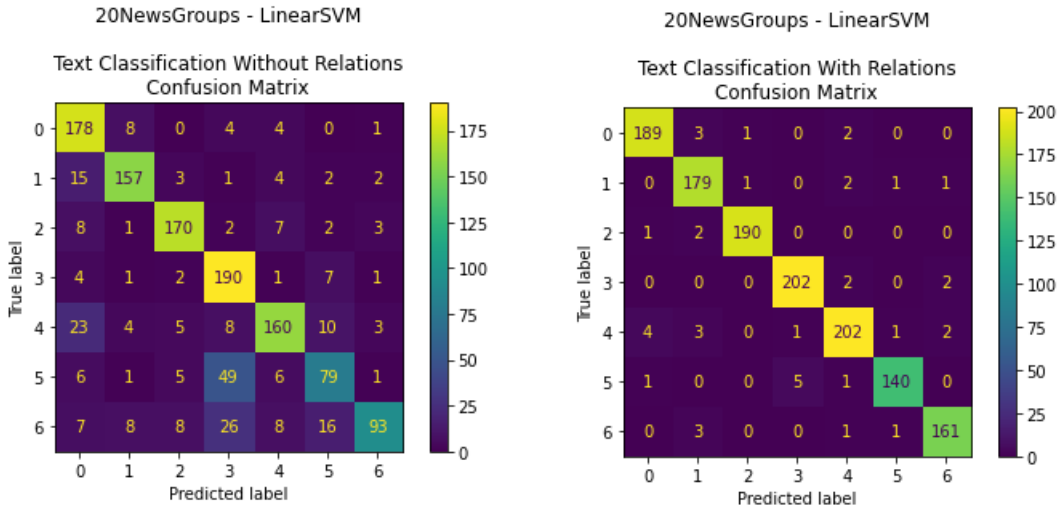
Diğer yandan, DNN modeli yüksek bir başlangıç doğruluğu sergilemesine rağmen (%90.5), ilişkisel bilgilerin entegrasyonundan sonra %96.6'lık bir doğrulukla sadece %6.75'lik bir iyileşme göstermiştir. Bu, bazı modellerin zaten yüksek doğruluk oranlarına sahip olduğunda, ek bilgilerden daha az yararlanabileceğini göstermektedir.

Bu sonuçlar, ilişkisel bilginin sınıflandırma modellerinin doğruluk ve etkinliğini nasıl önemli ölçüde artırabileceğini göstermektedir. Özellikle ilişkisel bilgilerin entegrasyonu, özellikle düşük performans gösteren modellerde büyük iyileştirmeler sağlayabilir, bu da metin sınıflandırma çalışmalarında yeni yaklaşımların araştırılmasını teşvik eder.



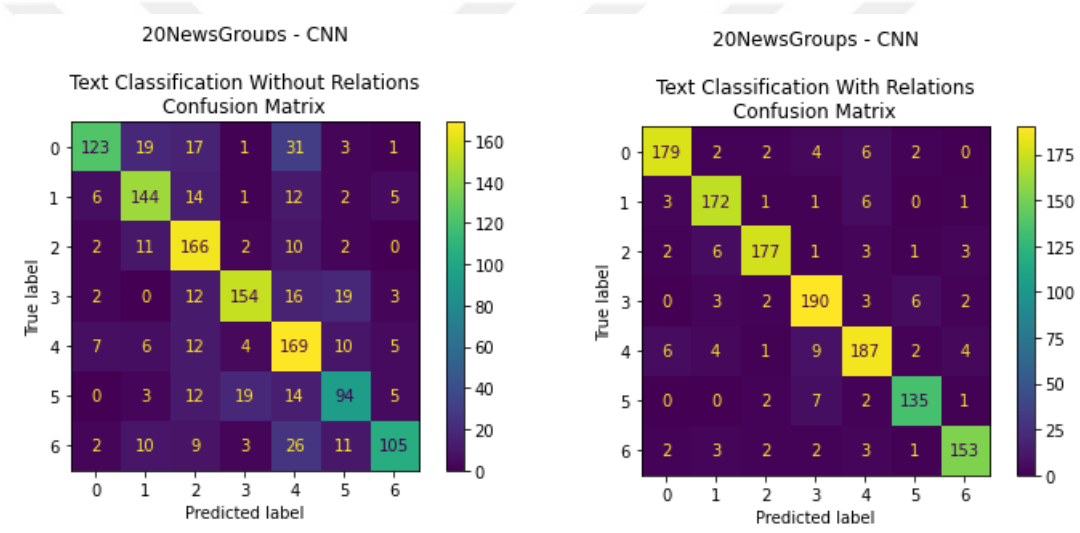
Şekil 25. 20Newsgroups Metin Veri Seti ve LogisticRegression Sınıflandırıcı Karmaşıklık Matrisleri

Şekil 25, 20Newsgroups veri setinde kullanılan Lojistik Regresyon modelinin performansını, ilişkisel bilgi ile ve olmadan iki ayrı karmaşıklık matrisi üzerinden değerlendirir. İlişkisel bilgi eklenmesinin ardından (sağdaki matris), bütün sınıflarda doğru sınıflandırmaların dikkate değer arttığı görülüyor. Genel olarak, bu iyileşmeler Lojistik Regresyon modelinin, özellikle bazı sınıflarda, ilişkisel bağlamı anlamada güçlü bir potansiyele sahip olduğunu işaret ediyor. Bu sonuçlar, metin sınıflandırma alanında ilişkisel bilgilerin daha stratejik bir şekilde kullanılmasının önünü açıyor ve bu bilgilerin dahil edilmesinin özellikle bazı sınıflandırma zorluklarının üstesinden gelmede kritik öneme sahip olabileceğini gösteriyor.



Şekil 26. 20Newsgroups Metin Veri Seti ve LinearSVM Sınıflandırıcı Karmaşıklık Matrisleri

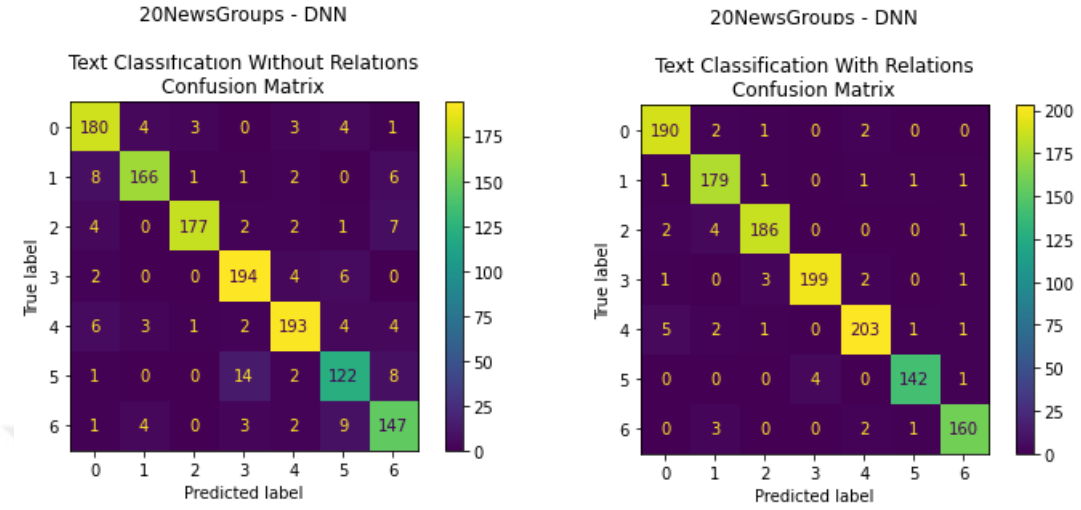
Şekil 26, 20Newsgroups veri setinde Linear SVM modeli ile gerçekleştirilen sınıflandırma sonuçlarını iki ayrı karmaşıklık matrisinde gösteriyor. İlişkisel bilgi entegre edildiğinde (sağdaki matris), tüm sınıflar için doğru sınıflandırma sayılarının arttığı görülmektedir, bu durum modelin bu kategorilerdeki örnekleri daha iyi anladığını ve daha az karışıklık yaşadığını göstermektedir. Diğer taraftan, sınıf 4, 5 ve 6' daki iyileşme göze çarpmaktadır. İlişkisel bilgi eklenmesiylemodelin performansındaki artış, ilişkisel bilginin modelin anlayışını ve sınıflandırma kabiliyetini güçlendirdiğinin bir göstergesi olarak yorumlanabilir. Bu sonuçlar, sınıflandırma modellerinin geliştirilmesinde ilişkisel bilginin entegrasyonunun potansiyel önemini gözler önüne seriyor ve sınıflandırma algoritmalarının doğruluğunu ve karar verme mekanizmalarını daha da geliştirebilecek yöntemlerin araştırılmasına olanak sağlıyor.



Şekil 27. 20Newsgroups Metin Veri Seti ve CNN Sınıflandırıcı Karmaşıklık Matrisleri

Şekil 27, CNN modelinin 20Newsgroups veri seti üzerinde gerçekleştirilen sınıflandırma işlemini iki farklı durumda ele almaktadır: İlişkisel bilgilerin dahil edilmemesi ve edilmesi sonrası. İlişkisel bilginin eklenmesiyle sağlanan performans artışı (sağdaki matris), her bir sınıfta doğru sınıflandırılan örnek sayısının arttığını gösteriyor, bu da modelin genel olarak her kategoride tahminlerini iyileştirdiğini işaret ediyor. Özellikle sınıf 5 ve 6' da görülen artışlar, modelin sınıflandırma yeteneğinin geniş bir bağlamsal anlayışla nasıl desteklenebileceğinin açık örnekleridir. Bu matris, ilişkisel bilgilerin entegrasyonunun, modelin daha genel ve kapsayıcı bir anlayışa ulaşmasına nasıl katkıda bulunduğunu ve yanlış pozitiflerin azaltılmasına yardımcı olduğunu göstermekte, böylece modelin doğruluk oranında %24.83 gibi önemli bir iyileşme sağlanmaktadır. Bu iyileşme, CNN modelinin, verilen ilişkisel bağlamların daha

derinlemesine anlaşılması ve sınıflandırma süreçlerinde etkin olarak kullanılması gerektiğinin altını çizmektedir.



Şekil 28. 20Newsgroups Metin Veri Seti ve DNN Sınıflandırıcı Karmaşıklık Matrisleri

Şekil 28'de, 20Newsgroups veri seti üzerinde Derin Sinir Ağı (DNN) kullanılarak yapılan metin sınıflandırma görevinin sonuçları iki farklı karmaşıklık matrisi ile gösteriliyor. İlişkisel bilginin entegrasyonuyla (sağdaki matris), hemen her sınıfta doğru sınıflandırma sayısında bir artış görülmekte ve bu, modelin genel anlamda sınıflandırma hassasiyetini artırdığını işaret ediyor. Özellikle sınıf 2,5 ve 6' daki iyileşmeler, modelin bu sınıfların karakteristik özelliklerini daha iyi kavradığını ve sınıflandırmada daha az hata yaptığını gösteriyor. Bu durum, DNN modelinin, verilen ilişkisel bilgileri işleyerek tahminlerini nasıl optimize edebileceğinin somut bir örneği olarak değerlendirilebilir. Bununla birlikte, diğer sınıflarda (örneğin sınıf 3 ve 4) daha mütevazı iyileşmeler olması, modelin zaten yüksek doğruluk oranına sahip sınıflarda, ilişkisel bilgiden daha az fayda sağladığını gösteriyor olabilir. Bu sonuçlar, ilişkisel bilginin sınıflandırma modellerine entegrasyonunun, belirli durumlarda ve sınıflarda model performansını önemli ölçüde artırabileceğine işaret etmektedir.

4.1.4. Mevcut Araştırmalar ile Bulgularımızın Karşılaştırmalı Değerlendirmesi

Bu bölümde, Türkçe ve İngilizce metin sınıflandırma veri setleri üzerinde yapılan mevcut araştırmalarla elde edilen bulgularımız karşılaştırılmaktadır. Araştırmamız, literatürde bulunan ön işleme adımlarını ve varlık ile ilişki tespiti yöntemlerini takip ederek, zenginleştirilmiş metin bilgilerinin sınıflandırma performansı üzerindeki etkisini derinlemesine incelemiştir. Bu süreçte, BERT tabanlı ön eğitimli bir model ve çeşitli sınıflandırma algoritmaları (SVM,

Lojistik Regresyon, DNN, CNN) kullanılmıştır. İlişkisel bilgilerin entegrasyonunun sınıflandırma doğruluğunu önemli ölçüde artırdığı gözlemlenmiştir.

TTC-4900, BBC-News ve 20Newsgroups veri setleri üzerinde yapılan deneyler, ilişkisel bilgilerin katkısını açıkça göstermiştir. Bu sonuçlar, özellikle kaynak sınırlaması olan dillerde ilişkisel bilginin önemini vurgulamaktadır. Araştırmamız, farklı dillerde metin sınıflandırma görevlerinin model performansını artırabileceğini kanıtlamaktadır. İngilizce metin setleri üzerinde yapılan literatür çalışmalarıyla karşılaştırıldığında, ilişkisel bilgi ile doğruluk ve hatırlama oranlarını artırma başarısı, bu yaklaşımın evrenselliğini ve uygulanabilirliğini ortaya koymaktadır.

Tez çalışmasında kullanılan derin öğrenme modelleri (RNN, CNN, BERT tabanlı modeller gibi) ve graf tabanlı modeller, literatürde sunulan benzer modellerle karşılaştırıldığında, ilişkisel bilgi entegrasyonu sayesinde belirgin bir performans artışı göstermiştir. Bu modeller, literatürde belirtilenlerle kıyaslandığında daha yüksek doğruluk oranları ve F1-Skorları sunarak, özellikle Türkçe metin veri setleri üzerinde yapılan deneylerde dilin anlamsal ve sözdizimsel özelliklerini daha etkin bir şekilde işleyebildiğini kanıtlamıştır.

Tezimizde benimsenen ilişkisel yöntemler, literatürdeki geleneksel metin sınıflandırma yaklaşımlarından ayrılarak, özellikle ön işleme teknikleri ve özellik çıkarma metodolojilerinde yenilikçi stratejiler uygulamıştır. Bu inovatif yaklaşımlar, bulgularımızın literatürdeki sonuçlardan daha üstün olmasına önemli katkıda bulunmuştur. Bu karşılaştırmalar, tezinizin literatürde var olan çalışmalarla olan bağlantısını ve sağladığı katkıları net bir şekilde gözler önüne sermek için mükemmel bir yol sunar.

Tablo 10, tezimizde elde edilen bulgular (tablo 7 ve 9) ile literatürdeki çalışmaların başarı oranlarını karşılaştırmaktadır. Bu karşılaştırma, literatürdeki veri setlerine uygulanan modellerin doğruluk oranlarını ve tezimizde aynı veri setlerine, ilişkisel bilgi eklenerek ve eklenmeden uygulanan modellerin doğruluk oranlarını yan yana sunmaktadır. Bu sayede, tezimizde uygulanan yöntemlerin ve ilişkisel bilginin etkilerini daha net bir şekilde görebilmekteyiz.

Tablo 10. Bulgular ile literatürdeki başarı oranları karşılaştırması

Veri Seti	Literatürdeki Çalışma	Literatürdeki Model ve Doğruluk (%)	Tez Çalışmasındaki İlişkisel Bilgi Eklenmeden Bulunan Doğruluk (%)	Tez Çalışmasındaki İlişkisel Bilgi Eklenmeden Bulunan Doğruluk (%)
TTC-3600	(Parlak, 2023)	SVM ile %78.1	SVM ile %91.3	SVM ile **%92.0**
TTC-3600	(Acı, 2019)	CNN ile **%93.3**	CNN ile %86.0	CNN ile %89.1
TTC-3600	(Kılınç, 2017)	Random Forest ile %91.03	DNN ile %91.3	DNN ile **%92.1**
20Newsgroup	(Pittaras, 2021)	DNN ile %82.6	DNN ile %90.5	DNN ile **%96.6**
20Newsgroup	(Lezama, 2022)	CNN ile %79	CNN ile %73.3	CNN ile **%91.5**

Örneğin, TTC-3600 veri seti üzerinde yapılan çalışmalar karşılaştırıldığında, literatürde SVM modeli ile elde edilen %78.1'lik doğruluk oranına karşın, tezimizde SVM modeli ile %91.3 doğruluk elde edilmiş ve ilişkisel bilgi eklenmesiyle bu oran %92.0'a yükseltilmiştir. Bu, tezimizde uygulanan metodolojinin ve veri işleme tekniklerinin literatürdeki çalışmalara göre daha etkili olduğunu göstermektedir.

Benzer şekilde, 20Newsgroup veri seti için literatürdeki CNN modeli %79 doğruluk sağlarken, tez çalışmamızda bu model %73.3 doğruluk elde etmiş, ancak ilişkisel bilgilerin eklenmesiyle %91.5 gibi önemli bir iyileşme sağlanmıştır. Bu, ilişkisel bilginin modellerin performansını nasıl önemli ölçüde artırabileceğini açıkça ortaya koymaktadır.

Bu karşılaştırmalar, tezimizin literatürde var olan çalışmalarla olan bağlantısını ve sağladığı katkıları net bir şekilde gözler önüne sermektedir. İlişkisel bilginin eklenmesinin, modellerin anlamsal ve sözdizimsel özellikleri daha iyi işlemesine ve dolayısıyla sınıflandırma doğruluğunun artmasına olanak tanıdığı görülmektedir. Bu sonuçlar, tezimizin alana sağladığı yenilikçi katkıların altını çizerken, gelecek çalışmalar için de bir referans noktası sunmaktadır.

Genel olarak, bu sonuçlar, çeşitli sınıflandırıcıların ilişkisel bilgiyi nasıl işlediği ve bu bilgiden nasıl yararlandığı konusundaki farklılıkları göstermektedir. İlişkisel bilginin dahil edilmesi genellikle performansı artırır, ancak iyileşme derecesi kullanılan modelin özelliklerine ve ilişkisel bilgiyi nasıl entegre ettiğine bağlıdır. Sonuç olarak, bu çalışma, ilişkisel bilginin, Türkçe ve İngilizce de dahil olmak üzere çeşitli dillerde metin sınıflandırma performansını artırmada kritik bir rol oynadığını göstermektedir.

4.2. Graf Tabanlı Sinir Ağları ile Türkçe Haber Metinlerinin Sınıflandırılması ve İlişkisel Analizi

Bu bölümde, Türkçe haber metinlerinin sınıflandırılması ve ilişkisel analizi için graf tabanlı sinir ağları (GNN) ve derin öğrenme tekniklerinin entegrasyonuna odaklanılmıştır. Buna göre, Türkçe "newsTRT2023" veri seti üzerindeki çeşitli makine öğrenimi modellerinin performansları, Graf Sinir Ağı (GNN) mimarisiyle kıyaslanarak değerlendirilmiştir. Araştırma kapsamında, geleneksel sinir ağı modelleri (DNN) ile birlikte, daha yenilikçi olan graf sinir ağı (GNN) modelleri karşılaştırmalı olarak değerlendirilmiştir. Bu yaklaşımların her ikisi de, Türkçe haber metinlerinin yapısal ve bağlamsal özelliklerini daha iyi anlamak ve sınıflandırmak için kullanılmıştır.

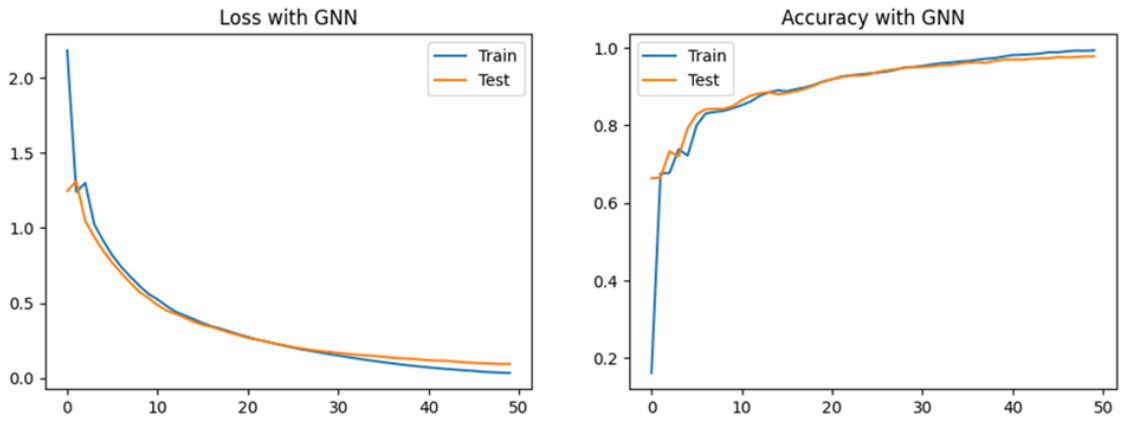
Tablo 11. Türkçe Haber Metinlerinin Graf ve Diğer Sınıflandırma Modelleri İle Değerlendirilmesi Sonuçları

Model	Accuracy	Recall	Precision	F1-Score
Logistic Regression	89.71	89.71	89.54	89.43
Light Gradient Boosting Machine	86.98	86.98	85.74	85.74
Extreme Gradient Boosting	86.66	86.66	86.24	85.69
Extra Trees Classifier	83.68	83.68	82.22	80.85
Random Forest Classifier	83.85	83.85	82.30	81.08
Gradient Boosting Classifier	83.82	83.82	82.09	82.29
Linear Discriminant Analysis	85.20	85.20	85.27	85.12
K Neighbors Classifier	86.62	86.62	86.45	85.97
SVM - Linear Kernel	89.11	89.11	88.87	88.59
Ridge Classifier	86.05	86.05	85.72	85.39
Deep Neural Network (DNN)	92.14	72.18	85.44	77.00
Graph Neural Network (GNN)	**97.93**	**92.22**	**95.08**	**93.59**

Çalışmamız, metinler arasındaki ilişkisel ve bağlamsal bilgilerin, graf tabanlı modeller tarafından nasıl işlendiğini ve bu modellerin sınıflandırma görevlerinde nasıl bir performans sergilediğini inceler. Bu modellerin metin sınıflandırmadaki etkinliğini, diğer makine öğrenimi yöntemleriyle karşılaştırmak amacıyla yapılan analizler, Tablo 11' de özetlenmiştir. Çalışmada ayrıca, graf tabanlı modellerin, metinler arası ilişkisel bilgileri anlamlandırma kapasitesine özel bir vurgu yapılmıştır.

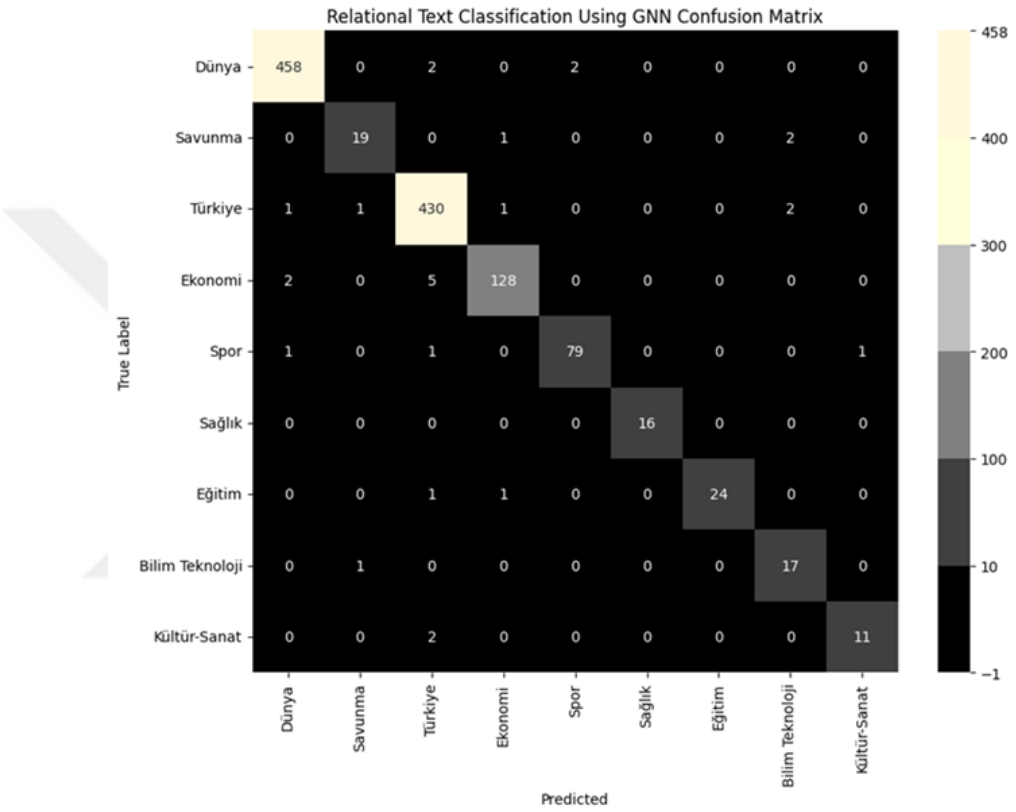
Tablo 11, GNN modelinin diğer modellere kıyasla dikkate değer bir şekilde daha yüksek doğruluk, geri çağırma, kesinlik ve F1 skoru ile üstün performans sergilediğini göstermektedir. Bu, GNN'nin haber metinlerinin yapısal ve bağlamsal özelliklerini derinlemesine analiz ederek karmaşık ilişkileri başarıyla çözümlediğini kanıtlar.

GNN, özellikle %97.93'lük etkileyici bir doğruluk oranı elde ederek, haber metinlerinin sınıflandırılmasında kullanılan diğer yöntemlere göre önemli avantajlar sağlamıştır. Detaylı performans analizi, modelin haber metinlerinin yapısal ve ilişkisel özelliklerini ne kadar iyi anladığını ve işlediğini gösteren kritik bir göstergedir. Bu analizlerin sonuçları, grafikler ve ilişkisel veri görselleştirmeler ile desteklenmiştir. Grafikler, farklı modellerin sınıflandırma başarısını ve ilişkisel bilgilerin sınıflandırma üzerindeki etkisini göstermek için kullanılmıştır. Bu görseller, model performanslarının yanı sıra, ilişkisel bilginin önemini ve graf tabanlı yaklaşımların potansiyel avantajlarını ortaya koymaktadır. Özellikle, Şekil 29' da gösterilen kayıp ve doğruluk grafikleri, modelin eğitim sürecinde nasıl geliştiğini ve yüksek doğruluk oranına nasıl ulaştığını ayrıntılı bir şekilde ortaya koymaktadır.



Şekil 29. GNN model eğitim-test sırasındaki kayıp ve doğruluk eğrisi

Eğitim ve test setleri için kayıp grafikleri, modelin öğrenme sürecinde kaybın nasıl azaldığını ve 10. epoch'tan sonra neredeyse sabit bir değere ulaştığını göstermektedir. Bu, modelin veri setine iyi uyum sağladığını ve genelleme kabiliyetinin yüksek olduğunu işaret eder. Bu, özellikle 50 epoch'a kadar devam eden istikrarlı performansla daha da pekişmektedir. Bu durum, modelin karmaşık haber metni veri setini anlama ve sınıflandırma konusunda başarılı olduğunu ve güvenilir sonuçlar üretebildiğini gösterir.



Şekil 30. GNN model Karşıtlık Matrisi

Aynı zamanda, doğruluk grafikleri, modelin eğitim ve test setlerinde benzer performans sergilediğini ve aşırı uyuma (overfitting) maruz kalmadığını gösterir. Bu durum, modelin sadece eğitim verilerini ezberlemek yerine genel bir öğrenme gerçekleştirdiğini ve bu bilgileri yeni, görülmemiş verilere uygulayabildiğini kanıtlar.

Şekil 30’ da sunulan karşıtlık matrisi, modelin sınıflandırma kabiliyetini ve tahminlerinin hassasiyetini detaylı bir şekilde gösterir. Karşıtlık matrisinde, diyagonal üzerindeki büyük değerler modelin çoğu durumda doğru tahminlerde bulunduğunu gösterirken, diyagonal dışındaki değerler modelin bazı kategorilerde yaptığı hataları göstermektedir. Ancak, genel olarak modelin sınıflandırma performansı yüksek ve hata oranları düşüktür.

Diyagonal üzerindeki büyük değerler (örneğin, 'Dünya' kategorisi için 458, 'Türkiye' için 430, 'Ekonomi' için 128 gibi), modelin çoğu durumda doğru tahminlerde bulunduğunu gösterir, yani bu kategorilerdeki haber metinlerini doğru bir şekilde sınıflandırabilmiştir. Diyagonal dışındaki değerler (örneğin, 'Dünya' kategorisindeki 2 'Ekonomi' yanlış sınıflandırması gibi), modelin bazı kategorilerde hatalı sınıflandırmalar yaptığını gösterir. Bununla birlikte, genel olarak modelin sınıflandırma performansı oldukça yüksek ve hata oranları düşüktür, bu da GNN'nin metin sınıflandırma konusunda etkili olduğunu göstermektedir. Özellikle yüksek hacimli (çok örnekli) kategorilerde ('Dünya' ve 'Türkiye') modelin iyi bir doğrulukla çalıştığı gözlemlenmiştir.

Bu bulgular, özellikle zorlu ve yapısal olarak karmaşık Türkçe metinler üzerinde yapılan deneylerde, graf tabanlı sinir ağlarının ve derin öğrenme modellerinin başarıyla entegre edilmesinin etkili olduğunu göstermektedir. Tablo 8 ve ilgili şekiller, bu modellerin metin sınıflandırma ve ilişkisel analiz görevlerinde nasıl üstün performans sergileyebileceğine dair değerli içgörüler sunar.

Grafik Sinir Ağı (GNN) mimarisinin, haber metinlerinin sınıflandırılması gibi karmaşık görevlerde, metinler arası ilişkisel ve yapısal özellikleri derinlemesine analiz edilerek bağlamsal anlayışı nasıl artırabildiğini gösterilmiştir. GNN' nin önerilen yapısı, literatürde sunulan diğer çalışmalarla karşılaştırıldığında, Türkçe metin sınıflandırma problemlerine yönelik yenilikçi bir yaklaşım sergileyerek dikkat çekmektedir.

Ayrıca, çok dilli ve İngilizce odaklı veri setleri üzerinde yapılan çalışmalarla karşılaştırıldığında, "newsTRT2023" veri seti üzerinde yapılan deneyler benzer veya daha iyi sonuçlar elde etmiştir. Bu durum, GNN mimarilerinin, Türkçe gibi morfolojik olarak zengin dillerin karmaşıklığını başarıyla ele alabilme kapasitesini ortaya koymaktadır. Derin öğrenme ve graf tabanlı analiz tekniklerinin bütünleşik kullanımı, dilin bağlamsal ve yapısal özelliklerini etkili bir şekilde modelleyerek yüksek doğruluk oranlarına ulaşılmasını sağlamıştır.

5. TARTIŞMA

Bu bölümde, tez çalışmasının ana bulguları değerlendirilmekte ve araştırma sonuçları, ilişkisel metin sınıflandırma yaklaşımının Türkçe metinler için ne ölçüde etkili olduğunu ve bu yaklaşımın potansiyel uygulama alanlarını ortaya koymaktadır. Ayrıca, çalışmada kullanılan yöntemlerin avantajları ve karşılaşılan zorluklar ile bu yöntemlerin sınıflandırma performansı üzerindeki etkileri de tartışılmaktadır.

Araştırmada elde edilen sonuçlar, derin öğrenme ve graf tabanlı modellerin ilişkisel doküman sınıflandırması görevinde ne derece başarılı olduğunu göstermektedir. Bu sonuçlar, literatürdeki benzer çalışmalarla karşılaştırıldığında, özellikle Türkçe metin veri setleri üzerinde yapılan deneylerde belirgin bir performans artışı sağladığını ortaya koymaktadır. Örneğin, çalışmada kullanılan BERT, CNN, DNN ve GNN modelleri, literatürde yer alan benzer modellerle karşılaştırıldığında, anlamsal ve sözdizimsel özelliklerin daha etkili bir şekilde işlenmesine olanak tanımış ve bu da sınıflandırma doğruluğunu artırmıştır.

Önceki çalışmalarda makine öğrenmesi ve derin öğrenme sınıflandırma algoritmaları kullanılmışken, bu tezde ele alınan derin öğrenme tabanlı ilişkisel sınıflandırma yaklaşımları, daha karmaşık ve heterojen doküman setlerinin üzerinde etkili olmuştur. Özellikle, metinlerdeki varlık ve ilişki bilgilerinin derinlemesine analizi, sınıflandırma performansında önemli bir iyileşme sağlamıştır. Bu iyileşme, varlık tanıma ve ilişki çıkarma tekniklerinin kullanımının, metinler arası bağlantıları ve anlam katmanlarını daha iyi modelleyebilmesine dayanmaktadır.

Ayrıca, bu çalışma, Türkçe gibi az kaynaklı diller için özel olarak geliştirilmiş algoritmaların geliştirilmesine olan ihtiyacı da vurgulamaktadır. Türkçe metinlerin özgün dil yapısının ve kullanım özelliklerinin, genel amaçlı modellere göre daha spesifik yaklaşımlar gerektirdiği gözlemlenmiştir. Bu durum, dil modellemesinde yerel dil özelliklerini dikkate alan ve bunları modelin ayrılmaz bir parçası haline getiren yeni yöntemlerin araştırılmasını gerektirir.

5.1. İlişkisel Metin Sınıflandırma Yaklaşımı

Araştırmamızda, BERT tabanlı modeller ve çeşitli sınıflandırma algoritmaları (SVM, Lojistik Regresyon, DNN, CNN) kullanılarak ilişkisel bilginin eklenmesinin metin sınıflandırma performansına etkisi bulgular bölümünde gösterilmiştir. İlişkisel bilgilerin entegrasyonu, sınıflandırma doğruluğunu önemli ölçüde artırmıştır. Örneğin, TTC-3600 veri setinde literatürde SVM modeli ile %78.1 doğruluk elde edilmiştir (Parlak, 2023). Tezimizde ise SVM modeli ile %91.3 doğruluk ve ilişkisel bilgi eklenmesiyle %92.0 doğruluk elde edilmiştir. Random Forest modeli ile yapılan başka bir çalışmada (Kılınç, 2017) %91.03 doğruluk sağlanmışken, tezimizde DNN modeli ile %91.3 ve ilişkisel bilgi eklenmesiyle %92.1 doğruluk elde edilmiştir. 20Newsgroup veri setinde ise, literatürde DNN modeli ile %82.6 doğruluk sağlanmıştır (Pittaras, 2021). Tez çalışmamızda ise DNN modeli ile %90.5 doğruluk ve ilişkisel bilgi eklenmesiyle %96.6 doğruluk elde edilmiştir. CNN modeli ile yapılan başka bir çalışmada (Lezama, 2022) literatürde %79 doğruluk sağlanırken, tez çalışmamızda %73.3 doğruluk ve ilişkisel bilgilerle %91.5 doğruluk sağlanmıştır.

Bu sonuçlar, ilişkisel bilginin metin sınıflandırma performansını artırmada kritik bir rol oynadığını ve özellikle Türkçe metin veri setlerinde önemli iyileşmeler sağladığını göstermektedir. Özetle, tezimizde kullanılan metodolojiler ve ilişkisel bilgi entegrasyonu, mevcut literatüre kıyasla daha yüksek doğruluk oranları sunarak, metin sınıflandırma görevlerinde önemli bir gelişme sağlamıştır.

5.2. Graf Tabanlı Sınıflandırma Yaklaşımı

Tez çalışmasının "newsTRT2023" veri seti üzerinde uygulanan Graf Sinir Ağları (GNN) tabanlı etiket bazlı ilişkisel metin sınıflandırma modelinin detaylı incelemesinde, Türkçe metin sınıflandırma ve ilişkisel analiz alanlarında kayda değer bir ilerleme gözlemlenmiştir. Graf Sinir Ağları, metinler arası ilişkisel ve yapısal özellikleri analiz ederek, dilin bağlamsal anlayışını önemli ölçüde artırabilmektedir.

GNN mimarisi, Türkçe haber metinlerinin sınıflandırılmasında ve ilişkisel analizinde oldukça yüksek bir başarı oranı göstermiştir. Modelin elde ettiği %97.93'lük doğruluk oranı, graf tabanlı analitik gücünün ve derin öğrenme kabiliyetlerinin başarılı birleşimini yansıtmaktadır. Bu yaklaşım, metinlerin karmaşık yapılarını daha iyi anlamamızı sağlamakta ve daha derinlemesine analizler yapılmasına olanak tanımaktadır.

Mevcut literatürdeki geniş kapsamlı incelemeler, metin sınıflandırması ve ilişkisel analiz alanlarında graf yapılarının kullanımını detaylı bir şekilde ele almıştır. Ancak bu incelemeler sırasında, Türkçe metinlerin sınıflandırılması konusunda ilişkisel graf sinir ağları kullanımı açısından belirgin bir eksiklik olduğu tespit edilmiştir. Bu bağlamda, belirtilen eksikliği doldurmayı ve Türkçe metinler için yenilikçi ve etkili metodolojik yaklaşımlar geliştirmeyi hedefleyen çalışmalar yapılmıştır.

Geliştirilen GNN mimarisi ile literatürde bulunan diğer modellere kıyasla Türkçe metin sınıflandırma problemlerine yenilikçi bir yaklaşım sergilemektedir. Özellikle, çok dilli ve İngilizce odaklı veri setleri üzerinde yapılan çalışmalarla karşılaştırıldığında benzer veya daha iyi sonuçlar elde edilmiştir. Bu durum, GNN mimarilerinin, Türkçe gibi morfolojik olarak zengin ve karmaşık dillerin özelliklerini başarıyla ele alabilme kapasitesini göstermektedir. GNN'nin bu özellikleri, Türkçe metin sınıflandırma ve ilişkisel analiz çalışmalarında önemli bir metodolojik ilerleme sağlamakta ve dilin zengin yapısal özelliklerinin daha etkili bir şekilde işlenmesine olanak tanımaktadır.

Modelimiz, Türkçe metin madenciliği ve doğal dil işleme tekniklerinin ilerlemesine önemli katkılarda bulunarak yeni yöntemlerin kapısını aralamaktadır. Özellikle medya analizi, kamuoyu araştırmaları ve yapay zeka destekli haber sınıflandırma sistemleri gibi çeşitli alanlarda pratik uygulamalar sunmaktadır. Bu çalışma, mevcut modeli daha da genişleterek Türkçe doğal dil işleme alanında daha kapsamlı perspektifler sunma potansiyeline sahiptir.

Türkçe metin sınıflandırma ve ilişkisel analiz alanlarına getirdiği yenilikçi ve önemli katkılarla, bu model dil işleme ve metin analizi alanlarında gelecekte yapılacak araştırmalara sağlam bir temel oluşturmaktadır. Modelin tam kapasitesini değerlendirebilmek ve algoritmanın performansını daha da artırabilmek için çeşitli ve geniş veri setleri üzerinde ek testlerin yapılması gereklidir. Bu sayede model, farklı kullanım senaryolarına daha uygun hale gelecek ve metodolojik yaklaşımlar daha da ileri taşınabilecektir.

5.3. Kısıtlamalar ve Zorluklar

Bu bölümde, tez çalışmamızda karşılaştığımız kısıtlamalar ve zorluklar ele alınmaktadır. Araştırmanın doğası gereği, veri erişimi ve kalitesi, hesaplama kaynakları, ve modelin karmaşıklığı gibi alanlarda bazı engellerle karşılaşılmıştır. Bu bölümde, bu kısıtlamaların

bulgular üzerindeki olası etkileri ve bu sorunlara yönelik olarak uyguladığımız çözümler detaylandırılmaktadır.

5.3.1. Veri Erişimi ve Kalitesi

Kısıtlama: Araştırmada kullanılan metinler; literatürde kullanılan metin veri setleri, çeşitli haber web sayfalarından elde edilen metin verileri ve ilişkisel bilgi kaynağı olarak kullanılan Wikidata bilgi bankasından toplanmıştır. Bu kaynaklardan elde edilen verilerin heterojen yapısı ve kalite kontrolündeki zorluklar, veri işleme ve model eğitimi süreçlerini zorlaştırmaktadır. Özellikle, WikiData'dan alınan ilişkisel metinler, farklı yapıda ve çeşitlilikte olabilmekte, bu da standart işleme metodlarını uygulamayı güçleştirmektedir.

Etki: Kalitesiz veya hatalı veriler, modelin öğrenme sürecini olumsuz etkileyebilir ve sınıflandırma performansını düşürebilir. Veri setlerinin çeşitliliği ve kapsamı, özellikle ilişkisel metinlerde, modelin genelleştirme yeteneğini sınırlayabilir. İlişkisel metinlerin doğru şekilde işlenmemesi, yanlış öğrenmeye ve modelin gerçek dünya verilerine uygulanabilirliğinin azalmasına neden olabilir.

Çözüm: Veri işleme ve ön işleme süreçleri, literatürdeki uygulamaları takip ederek gerçekleştirildi. Veri entegrasyonu ve bütünlüğünü sağlamak için çeşitli kaynaklardan toplanan verilerin doğru ve tutarlı bir şekilde işlenmesi sağlandı.

5.3.2. Hesaplama Kaynakları

Kısıtlama: Model eğitimleri için gerekli olan yüksek performanslı hesaplama kaynaklarına erişim, özellikle kaynak kısıtlı durumlarda zor olabilir.

Etki: Yetersiz hesaplama kaynakları model eğitim sürelerini uzatabilir ve optimizasyon süreçlerini kısıtlayabilir, bu da araştırma sürecini olumsuz etkileyebilir.

Çözüm: Hesaplama kaynaklarına erişim ihtiyacını karşılamak için Google Colab gibi ücretsiz hesaplama platformlarından faydalanılmıştır. Bu platformlar, model eğitimi ve test süreçlerinde ihtiyaç duyulan hesaplama gücünü sağlayarak süreçleri desteklemiştir.

5.3.3. Kullanılan Metodolojilerin ve Araçların Etkinliği

Derin öğrenme ve graf tabanlı modellerin kurulumu, yapılandırılması, eğitimi ve uygulanması karmaşık süreçlerdir. Bu karmaşıklık, model geliştirme ve eğitim süreçlerinin verimliliğini doğrudan etkileyebilir. Bu kapsamda, modellerin etkin bir şekilde

uygulanabilmesi için kapsamlı bir destek ve kaynak seti gerekmektedir. Eğitim süreçlerini desteklemek ve hızlandırmak amacıyla kullanılan araçlar, araştırmacılara modelleri daha iyi anlama ve daha etkin şekilde uygulama olanağı sunar. Yeni başlayanlar ve kaynakları sınırlı olan araştırmacılar için, çeşitli açık kaynaklı yazılım kütüphaneleri ve çevrimiçi platformlar, model kurulumu ve optimizasyon süreçlerinde büyük yardımcı olabilir. Bu araçlar arasında TensorFlow, PyTorch ve Scikit-learn gibi kütüphaneler, modellerin daha erişilebilir ve uygulanabilir hale gelmesine olanak tanırken, çevrimiçi eğitim kaynakları ve topluluk destekli forumlar, araştırmacılara karşılaştıkları sorunlarda yardımcı olabilir. Bu tür destekler, model geliştirme ve eğitim süreçlerinin daha verimli ve etkili olmasını sağlayarak, araştırmacıların zorlukları aşmalarına ve modellerini başarıyla uygulamalarına imkan verir.

Derin Öğrenme Modelleri (DNN ve CNN):

- **Avantajlar:** Bu modeller, karmaşık veri yapılarını işleyebilme ve yüksek seviyede özellik çıkarma kapasitesi sunma özellikleriyle dikkat çeker. Çalışmamızda, BERT, DNN ve CNN gibi modellerin kullanımı, metinlerdeki anlamsal ve sözdizimsel özelliklerin derinlemesine analiz edilmesini sağlamış ve sınıflandırma doğruluğunu önemli ölçüde artırmıştır.

- **Kısıtlamalar:** Bu tür modellerin eğitimi için gereken yüksek hesaplama gücü ve büyük veri setleri, özellikle kaynak kısıtlı ortamlarda büyük bir engel teşkil edebilir. Ayrıca, modelin genelleştirme yeteneğini sınırlayan overfitting sorununa karşı duyarlıdırlar.

Graf Tabanlı Modeller (GNN):

- **Avantajlar:** Graf tabanlı modeller, metinler arası ve metin içi ilişkileri etkili bir şekilde modelleyebilir. Bu çalışmada kullanılan GNN, ilişkisel doküman sınıflandırılmasında, bağlamsal ve yapısal ilişkileri dikkate alarak yüksek performans göstermiştir.

- **Kısıtlamalar**: Graf yapısının oluşturulması ve işlenmesi büyük miktarda bellek ve işlem gücü gerektirir. Ayrıca, bu modellerin uygulanması ve optimizasyonu, özellikle büyük veri setlerinde, teknik olarak zorlayıcı olabilir.

Makine Öğrenimi Teknikleri:

- **Avantajlar:** Destek Vektör Makineleri (SVM) ve Lojistik Regresyon gibi makine öğrenimi modelleri, metin sınıflandırma görevlerinde etkin çözümler sunabilir. Çalışmada, SVM ve lojistik regresyon modelleri özellikle lineer sınıflandırma problemlerinde güçlü performans göstermiş, çeşitlendirilmiş metin setleri üzerinde hızlı ve etkili denemeler yapılmasına olanak tanımıştır. Bu modeller, yüksek boyutlu veri setlerinde dahi stabil ve güvenilir sonuçlar üretebilir.

- **Kısıtlamalar**: Daha Düşük Karmaşıklıkla Gelen Sınırlılıklar: SVM ve lojistik regresyon modelleri, bazı durumlarda daha karmaşık ilişkisel özellikleri modellemekte yetersiz kalabilir. Özellikle, metinler arasındaki ince bağlantıları ve anlam katmanlarını çıkarmak konusunda sınırlı kalabilirler. Bu modeller, özellikle eğitim verilerine çok sıkı bir şekilde uyum sağladıklarında, yeni ve farklı veri setlerine genelleştirme yapma konusunda zorlanabilirler. Modelin overfitting'e yatkınlığı, veri setindeki değişikliklere karşı duyarlılığı artırabilir, bu da modelin farklı senaryolarda istikrarlı sonuçlar üretmesini zorlaştırabilir. Bu bağlamda, SVM ve lojistik regresyonun avantajları, belirli senaryolarda hızlı ve basit çözümler sunarken, karmaşık ilişkisel analizler ve geniş çaplı genelleştirmeler için daha ileri teknoloji modellerin tercih edilmesi gerekebilir.

Bu kapsamda, derin öğrenme ve graf tabanlı modellerin karmaşık kurulum ve yapılandırma süreçlerini yönetmek için etkin stratejiler uygulanmıştır. Özellikle, model optimizasyonu ve hata azaltma süreçleri, çeşitli otomatik araçlar ve algoritmalar kullanılarak iyileştirilmiştir. Bu araçlar arasında, parametre ayarlama ve model doğrulama için gelişmiş yazılım çözümleri yer almaktadır, bu sayede modellerin daha etkin ve verimli bir şekilde eğitilmesi sağlanmıştır. Ayrıca, modelin genelleştirme yeteneğini artırmak ve overfitting sorununu minimize etmek için, kapsamlı veri çeşitlendirme ve artırma teknikleri kullanılmıştır. Bu yaklaşımlar, modellerin farklı veri kümelerine ve gerçek dünya senaryolarına daha iyi adapte olmasını sağlayarak, sınıflandırma doğruluğunu ve modelin genel performansını önemli ölçüde iyileştirmiştir.

6. SONUÇ VE ÖNERİLER

Bu tez çalışması, ilişkisel metin sınıflandırması üzerine yapılan kapsamlı araştırmalara dayanarak, metinlerdeki varlık-ilişki bilgilerinin entegrasyonunun makine öğrenimi ve derin öğrenme modelleri kullanılarak sınıflandırma performansını nasıl artırdığını detaylı bir şekilde incelemiştir. Çalışma, kişi varlıklarına odaklanmış ve bu varlıklar için seçilen sınırlı sayıda özellik değeri ile daha özelleştirilmiş bir analiz yapılmasını sağlamıştır. Veri setindeki ilişkilerin tanımlanması, bir varlık-ilişki bilgi tabanının oluşturulmasına imkan tanıyarak, sınıflandırma sürecinde kullanılmıştır.

6.1. İlişkisel Metin Sınıflandırmasının Önemi

İlişkisel metin sınıflandırması, derin öğrenme ve makine öğrenimi modellerinin metinleri daha etkin bir şekilde anlamasını ve sınıflandırmasını sağlayan kritik bir rol oynamaktadır. Türkçe metinler üzerinde yapılan çalışmalar, varlık-ilişki entegrasyonunun %3.42'lik önemli bir doğruluk iyileştirmesi sağladığını ortaya koymuştur. Bu, dilin yapısal özelliklerinin sınıflandırma modellerinin geliştirilmesi ve uygulanmasında stratejik bir şekilde dikkate alınmasının gerekliliğini göstermektedir.

Dil işleme alanında bu tür entegrasyonlar, özellikle çok dilli ve kültürel metin veri kümeleri üzerinde uygulandığında, modellerin daha kapsamlı ve duyarlı olmasını sağlayabilir. İngilizce metinlerde bu iyileştirme %25.78'e kadar çıkarak, dil yapısının basitliği ve geniş çapta erişilebilir doğal dil işleme araçlarının varlık-ilişki entegrasyonunu kolaylaştırdığını göstermektedir. Bu büyük iyileştirme oranı, dilin morfolojik ve yapısal özelliklerinin sınıflandırma performansı üzerinde doğrudan bir etkisi olduğunu ve bu özelliklerin modeller tarafından nasıl işlendiğinin sınıflandırma başarısını önemli ölçüde etkileyebileceğini vurgular.

Bu çalışmada elde edilen bulgular, dilin yapısal ve morfolojik özelliklerinin modelleme süreçlerine ne kadar entegre edilmesi gerektiğini, ve bu entegrasyonun modelin genel başarımını nasıl iyileştirebileceğini ortaya koyar. Türkçe gibi aglütinatif bir dilde, varlık-ilişki bilgilerinin sınıflandırma sistemlerine entegrasyonu, metinleri daha derinlemesine analiz etme ve daha doğru sınıflandırma sonuçları elde etme fırsatı sunar. Bu nedenle, gelecekteki

çalışmaların dil özelliklerini modelleme ve sınıflandırma süreçlerine dahil etme yöntemlerini geliştirmesi büyük önem taşır.

Sonuç olarak, ilişkisel metin sınıflandırması, sadece metinleri daha iyi anlama ve sınıflandırma yeteneği sunmakla kalmaz, aynı zamanda dilin yapısal ve morfolojik zenginliklerinin bu süreçlerde nasıl kullanılacağına dair önemli içgörüler sağlar. Bu, özellikle farklı dillerde ve metin türlerinde çalışmalar yapılacaksa, dil özelliklerinin doğru bir şekilde anlaşılmasını ve modelleme süreçlerine entegre edilmesini zorunlu kılar. Bu entegrasyon, sınıflandırma modellerinin daha etkin ve doğru çalışmasını sağlar, dolayısıyla dil işleme teknolojilerinin daha geniş bir uygulama yelpazesi sunmasına olanak tanır.

6.2. Graf Sinir Ağları ve Türkçe Metin Sınıflandırması

Gelişen teknoloji ve artan veri hacmi, dil işleme ve metin analizi alanlarında yenilikçi ve etkili çözümler geliştirmenin gerekliliğini ortaya koymaktadır. Bu bağlamda, Türkçe haber metinlerinin sınıflandırılması ve ilişkisel analizine yönelik bu çalışma, dil işleme alanındaki mevcut metodolojileri geliştirmek için Graf Sinir Ağları (GNN) kullanımını önermektedir. Çalışmada kullanılan "newsTRT2023" veri seti, Türkçe'nin zengin morfolojik yapısını ve metinler arasındaki ilişkileri kapsayarak, dil işleme tekniklerinin doğruluğunu ve kapsamını test etmek için ideal bir platform sunmaktadır. GNN modeli, haber metinlerinin yapısal ve bağlamsal özelliklerini derinlemesine analiz ederek, %97.93 gibi dikkate değer bir doğruluk oranı elde etmiştir. Bu sonuç, GNN'nin karmaşık ve ilişkisel metin verilerini işleme konusunda üstün bir başarıya sahip olduğunu göstermektedir. Ayrıca, bu araştırma, Türkçe metin sınıflandırma ve ilişkisel analizde kullanılan geleneksel ve derin öğrenme tabanlı modellerle karşılaştırmalı bir değerlendirme yaparak, GNN'nin bu modellere kıyasla üstün performansını kanıtlamaktadır. Bu bulgular, Türkçe metin sınıflandırma çalışmalarında GNN'nin önemli bir potansiyel taşıdığını vurgulamakta ve bu alandaki metodolojik yaklaşımları daha ileriye taşıma potansiyeline sahiptir. Araştırma, Türkçe metin analizinde derin öğrenme ve graf tabanlı yöntemlerin bütünleştirilmesinin, dilin zengin ve karmaşık yapısını başarılı bir şekilde anlamada ve işlemede ne kadar etkili olduğunu göstermektedir. Bu çalışma, dil işleme ve metin analizi alanlarında gelecekte yapılacak araştırmalara katkıda bulunarak, bu alanlardaki bilgi birikimini ve uygulama yelpazesini genişletmeyi hedeflemektedir.

6.3. Dil Yapısının Modellere Etkisi

Türkçe'nin morfolojik zenginliği ve aglütinatif yapısı, varlık-ilişki bilgilerinin derinlemesine analizini sağlayarak sınıflandırma performansını artırabilir. İngilizce'de ise, dilin sözdizimsel yapısının basitliği ve geniş çapta erişilebilir doğal dil işleme araçları, varlık-ilişki entegrasyonunu kolaylaştırmıştır. Bu nedenle, farklı dillerin yapısal ve morfolojik özellikleri, metin sınıflandırma modellerinin geliştirilmesinde özel olarak dikkate alınmalıdır.

6.4. Gelecekteki Araştırma Yönleri

Dil Özelliklerinin Etkili Entegrasyonu: Gelecekteki çalışmaların dillerin morfolojik ve sözdizimsel özelliklerini daha etkin bir şekilde entegre etmesi gerekmektedir. Bu entegrasyon, özellikle dil özelliklerine özgü varlık-ilişki entegrasyon stratejilerinin özelleştirilmesi için kritik önem taşır. Bu çalışmamız, Türkçe metin sınıflandırmasına odaklanmıştır, zira Türkçenin benzersiz yapısal karmaşıklıkları genel amaçlı üretken modellerin doğrudan uygulanmasının ötesinde özel yaklaşımlar gerektirir. Bu bağlamda, Llama2 ve GPT-4 gibi modellerin belirli alanlara nasıl uyarlanabileceği ve bu uyarlamanın etkililiğinin nasıl ölçülebileceği önemli bir araştırma konusudur. Varlık-ilişki bilgisinin entegrasyonunun, spesifik alanlarda modellerin performansını nasıl geliştirebileceğini vurgulamaktadır. Gelecekteki çalışmalar, dillerin morfolojik ve sözdizimsel özelliklerinin etkili entegrasyonuna odaklanmalı, dil özelliklerine özgü varlık-ilişki entegrasyon stratejilerinin özelleştirilmesi üzerine yoğunlaşmalı ve yenilikçi doğal dil işleme tekniklerinin geliştirilip uygulanmasına ağırlık vermelidir.

Yenilikçi NLP Teknikleri: İnovatif doğal dil işleme tekniklerinin geliştirilmesi ve uygulanması, metin sınıflandırma performansını artırabilir.

Etkileşimli Öğrenme Sistemleri: Kullanıcı geri bildirimlerine dayalı etkileşimli öğrenme sistemleri, belirsiz veya tartışmalı ilişkilerin sınıflandırılmasını iyileştirebilir.

Bu çalışma, dil işleme ve metin analizi alanlarında yapılan araştırmalara güçlü bir katkı sağlamış ve çeşitli dillerin yapısal ve morfolojik özelliklerinin modelleme süreçlerine nasıl entegre edilmesi gerektiğinin altını çizmiştir. Bu temel üzerine inşa edilecek gelecek çalışmalar, daha etkin ve güvenilir metin sınıflandırma sistemlerinin geliştirilmesine katkıda bulunacak ve böylece dil işleme teknolojilerinin daha geniş bir uygulama yelpazesi sunmasını sağlayacaktır.

KAYNAKLAR

Abdullah, M. H. A., Aziz, N., Abdulkadir, S. J., Alhussian, H. S. A., & Talpur, N., 2023, Systematic Literature Review of Information Extraction From Textual Data: Recent Methods, Applications, Trends, and Challenges. *IEEE Access*.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B., 2023, Gpt-4 technical report, *arXiv preprint arXiv:2303.08774*.

Acı, Ç., & Çırak, A. (2019). Türkçe Haber Metinlerinin Konvolüsyonel Sinir Ağları ve Word2Vec Kullanılarak Sınıflandırılması. Bilişim Teknolojileri Dergisi, 12(3), 219-228. https://doi.org/10.17671/gazibtd.457917

Agarwal, J., Christa, S., Pai, A., Kumar, M. A., & Prasad, G. (2023, January). Machine learning application for news text classification. *In 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 463-466). IEEE.

Ahonen, H., Heinonen, O., Klemettinen, M., & Verkamo, A. I., 1997, Applying data mining techniques in text analysis, *Report C-1997-23, Dept. of Computer Science, University of Helsinki*.

Akın, A. A., & Akın, M. D. (2007). Zemberek, an open source NLP framework for Turkic languages. *Structure*, 10(2007), 1-5.

Akpatsa, S. K., Li, X., & Lei, H. (2021). A survey and future perspectives of hybrid deep learning models for text classification. *In Artificial Intelligence and Security: 7th International Conference, ICAIS 2021*, Dublin, Ireland, July 19–23, 2021, Proceedings, Part I 7 (pp. 358-369). Springer International Publishing.

Al Nahas, A., Kulunk, A., Gozutok, B., Kalkan, S. C., & Erdinc, H. Y. (2020, August). How to Segment Turkish Words for Neural Text Classification?. *In 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)* (pp. 1-5). IEEE.

Aras, G., Makaroğlu, D., Demir, S., & Cakir, A. (2021). An evaluation of recent neural sequence tagging models in Turkish named entity recognition. *Expert Systems with Applications*, 182, 115049.

Asudani, D. S., Nagwani, N. K., & Singh, P. (2023). Impact of word embedding models on text analytics in deep learning environment: a review. *Artificial intelligence review*, 56(9), 10345-10425.

Augustyniak, Ł., Kajdanowicz, T., & Kazienko, P. (2021). Comprehensive analysis of aspect term extraction methods using various text embeddings. *Computer Speech & Language*, 69, 101217.

Basha, S. R., & Rani, J. K. (2019). A comparative approach of dimensionality reduction techniques in text classification. *Engineering, Technology & Applied Science Research*, 9(6), 4974-4979.

BBC Datasets, Collected by BBC News Website & Homepage, URL: http://mlg.ucd.ie/datasets/bbc.html (Accessed: Feb. 19, 2024)

Bharadiya, J. (2023). A comprehensive survey of deep learning techniques natural language processing. *European Journal of Technology*, 7(1), 58-66.

Bollacker, K., Cook, R., & Tufts, P. (2007, July). Freebase: A shared database of structured general human knowledge. In AAAI (Vol. 7, pp. 1962-1963).

Bölücü, N., & Can, B. (2019, April). Context based automatic spelling correction for turkish. *In 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)* (pp. 1-4). IEEE.

Cavnar, W. B., & Trenkle, J. M. (1994, April). N-gram-based text categorization. *In Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval* (Vol. 161175, p. 14).

Cervetti, G. N., & Wright, T. S. (2020). The role of knowledge in understanding and learning from text. *Handbook of reading research*, 5.

Chen, Z., Huang, K., Wu, L., Zhong, Z., & Jiao, Z. (2022). Relational graph convolutional network for text-mining-based accident causal classification. *Applied Sciences*, 12(5), 2482.

Çilden, E. K. (2006). Stemming Turkish words using snowball.

Dawei, W., Alfred, R., Obit, J. H., & On, C. K. (2021). A literature review on text classification and sentiment analysis approaches. *Computational Science and Technology: 7th ICCST 2020, Pattaya, Thailand*, 29–30 August, 2020, 305-323.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805.

Flisar, J., & Podgorelec, V. (2018, June). Document enrichment using DBPedia ontology for short text classification. In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics (pp. 1-9).

Ghaddar, A., & Langlais, P. (2017, November). Winer: A wikipedia annotated corpus for named entity recognition. *In Proceedings of the Eighth International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) (pp. 413-422).

Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. Information, 13(2), 83

Gunawan, L., Anggreainy, M. S., Wihan, L., Lesmana, G. Y., & Yusuf, S. (2023). Support vector machine based emotional analysis of restaurant reviews. Procedia Computer Science, 216, 479-484.

Han, J., Pei, J., & Tong, H., 2022, *Data mining: concepts and techniques*, Morgan kaufmann.

Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., & Sun, M. (2018). Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv preprint arXiv*:1810.10147.

Han, Xu, et al. "OpenNRE: An open and extensible toolkit for neural relation extraction." *arXiv preprint arXiv*:1909.13078 (2019).

Hotho, A., Nürnberger, A., & Paaß, G., 2005, A brief survey of text mining, *Journal for Language Technology and Computational Linguistics*, 20(1), 19-62.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006, June). OntoNotes: the 90% solution. *In Proceedings of the human language technology conference of the NAACL*, Companion Volume: Short Papers (pp. 57-60).

Huang, L., Ma, D., Li, S., Zhang, X., & Wang, H. (2019). Text level graph neural network for text classification. *arXiv preprint arXiv:1910.02356*.

Incitti, F., Urli, F., & Snidaro, L. (2023). Beyond word embeddings: A survey. *Information Fusion*, 89, 418-436.

Ji, Guoliang, et al. "Distant supervision for relation extraction with sentence-level attention and entity descriptions." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. No. 1. 2017.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv*:1612.03651.

Kaur, B., & Bathla, G. (2018). Document classification using various classification algorithms: a survey. *Int J Fut Revol Comput Sci Commun Eng*, 4(2), 150-155.

Khan, W., Daud, A., Khan, K., Muhammad, S., & Haq, R. (2023). Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends. *Natural Language Processing Journal*, 100026.

Kılınç, D., Özçift, A., Bozyigit, F., Yıldırım, P., Yücalar, F., & Borandag, E. (2017). TTC-3600: A new benchmark dataset for Turkish text categorization. Journal of Information Science, 43(2), 174-185.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150.

Kumar, V. S., Alemran, A., Karras, D. A., Gupta, S. K., Dixit, C. K., & Haralayya, B. (2022, December). Natural Language Processing using Graph Neural Network for Text Classification. *In 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES)* (pp. 1-5). IEEE.

Kuyumcu, B., Aksakalli, C., & Delil, S. (2019, June). An automated new approach in fast text classification (fastText) A case study for Turkish text classification without pre-processing. In Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval (pp. 1-4).

Lang K., "The 20 Newsgroups Data Set," [Online]. Available: http://people.csail.mit.edu/jrennie/20Newsgroups/ , 2008. (Accessed: Feb. 19, 2024)

Lezama-Sánchez, A. L., Tovar Vidal, M., & Reyes-Ortiz, J. A. (2022). An approach based on semantic relationship embeddings for text classification. Mathematics, 10(21), 4161.

Li, B., Liu, T., Zhao, Z., Tang, B., Drozd, A., Rogers, A., & Du, X. (2017, September). Investigating different syntactic context types and context representations for learning word embeddings. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2421-2431).

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., ... & He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2), 1-41.

Liu, C., Wang, X., & Xu, H. (2022). Text Classification Using Document-Relational Graph Convolutional Networks. *IEEE Access*, 10, 123205-123211.

Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

Liu, X., Tian, J., Niu, N., Li, J., & Han, J. (2023). "Standard Text" Relational Classification Model Based on Concatenated Word Vector Attention and Feature Concatenation. *Applied Sciences*, 13(12), 7119.

Magalhães, D., Lima, R. H., & Pozo, A. (2023). Creating deep neural networks for text classification tasks using grammar genetic programming. *Applied Soft Computing*, 135, 110009.

Malekzadeh, M., Hajibabaee, P., Heidari, M., Zad, S., Uzuner, O., & Jones, J. H. (2021, December). Review of graph neural network in text classification. *In 2021 IEEE 12th annual ubiquitous computing, electronics & mobile communication conference (UEMCON)* (pp. 0084-0091). IEEE.

Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011, September). DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th international conference on semantic systems (pp. 1-8).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv*:1301.3781.

Mooney, C. H., & Roddick, J. F., 2013, Sequential pattern mining--approaches and algorithms. *ACM Computing Surveys (CSUR)*, 45(2), 1-39.

Nagumothu, D., Eklund, P. W., Ofoghi, B., & Bouadjenek, M. R. (2021). Linked data triples enhance document relevance classification. Applied Sciences, 11(14), 6636.

Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López García, Á., Heredia, I., ... & Hluchý, L. (2019). Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. Artificial Intelligence Review, 52, 77-124.

Okur, H. I., & Sertbaş, A. (2021, September). Pretrained neural models for turkish text classification. *In 2021 6th International Conference on Computer Science and Engineering (UBMK)* (pp. 174-179). IEEE.

Parlak, B. (2023). The effects of preprocessing on Turkish and English News Data. Sakarya University Journal of Computer and Information Sciences, 6(1), 59-66.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Perkins, J. (2010). Python text processing with NLTK 2.0 cookbook. PACKT publishing.

Porter, M. F. (2001). Snowball: A language for stemming algorithms.

Pittaras, N., Giannakopoulos, G., Papadakis, G., & Karkaletsis, V. (2021). Text classification with semantically enriched word embeddings. Natural Language Engineering, 27(4), 391-425.

Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, No. 1, pp. 29-48).

Rebele, T., Suchanek, F., Hoffart, J., Biega, J., Kuzey, E., & Weikum, G. (2016). YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames. In The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II 15 (pp. 177-185). Springer International Publishing.

Ren, F., Zhou, D., Liu, Z., Li, Y., Zhao, R., Liu, Y., & Liang, X. (2018, August). Neural relation classification with text descriptions. *In Proceedings of the 27th international conference on computational linguistics* (pp. 1167-1177).

Ru, Chengsen, et al. "Using semantic similarity to reduce wrong labels in distant supervision for relation extraction." *Information Processing & Management* 54.4 (2018): 593-608.

Sahu, Sunil Kumar, et al. "Inter-sentence relation extraction with document-level graph convolutional neural network." *arXiv preprint arXiv*:1906.04684 (2019).

Sarasu, R., Thyagharajan, K. K., & Shanker, N. R. (2023). SF-CNN: Deep Text Classification and Retrieval for Text Documents. *Intelligent Automation & Soft Computing*, 35(2).

Schweter, S. (2020). BERTurk-BERT models for Turkish. *Zenodo*, 2020, 3770924.

Shang, Y. M., Huang, H., Sun, X., Wei, W., & Mao, X. L. (2022). A pattern-aware self-attention network for distant supervised relation extraction. Information Sciences, 584, 269-279.

Sinoara, R. A., Camacho-Collados, J., Rossi, R. G., Navigli, R., & Rezende, S. O. (2019). Knowledge-enhanced document embeddings for text classification. Knowledge-Based Systems, 163, 955-971.

Smirnova, A., & Cudré-Mauroux, P. (2018). Relation extraction using distant supervision: A survey. ACM Computing Surveys (CSUR), 51(5), 1-35.

Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., & Wang, G. (2023). Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

Şeker, G. A., & Eryiğit, G. (2017). Extending a CRF-based named entity recognition model for Turkish well formed text and user generated content 1. *Semantic Web*, 8(5), 625-642.

Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1), 25-29.

Tang, M., Yang, B., & Xu, H. (2021, April). A Multi-grained Attention Network for Multi-labeled Distant Supervision Relation Extraction. In 2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS) (pp. 110-115). IEEE.

TRT Haber, TRT Haber Web Sayfası, URL: https://www.trthaber.com/tum-mansetler-sayfa-1.html (Accessed: Feb. 19, 2024)

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T., 2023, Llama 2: Open foundation and fine-tuned chat models, *arXiv preprint* arXiv:2307.09288.

Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., & Tang, J. (2021). KEPLER: A unified model for knowledge embedding and pre-trained language representation. Transactions of the Association for Computational Linguistics, 9, 176-194.

Wu, L., Chen, Y., Shen, K., Guo, X., Gao, H., Li, S., ... & Long, B. (2023). Graph neural networks for natural language processing: A survey. *Foundations and Trends® in Machine Learning*, 16(2), 119-328.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10), 78-85.

Yang, K., He, L., Dai, X., Huang, S., & Chen, J. (2019, June). Exploiting noisy data in distant supervision relation classification. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 3216-3225).

Yao, L., Mao, C., & Luo, Y. (2019, July). Graph convolutional networks for text classification. *In Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 7370-7377).

Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., ... & Sun, M. (2019). DocRED: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv*:1906.06127.

Yıldız, O. T., Ak, K., Ercan, G., Topsakal, O., & Asmazoğlu, C. (2018, April). A multilayer annotated corpus for turkish. *In 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)* (pp. 1-6). IEEE.

Zhang, N., Chen, X., Xie, X., Deng, S., Tan, C., Chen, M., ... & Chen, H. (2021). Document-level relation extraction as semantic segmentation. arXiv preprint arXiv:2106.03618.

Zhang, Y., Yu, X., Cui, Z., Wu, S., Wen, Z., & Wang, L. (2020). Every document owns its structure: Inductive text classification via graph neural networks. arXiv preprint arXiv:2004.13826.

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.

Zhu, H., Peng, H., Lyu, Z., Hou, L., Li, J., & Xiao, J. (2023). Pre-training language model incorporating domain-specific heterogeneous knowledge into a unified representation. Expert Systems with Applications, 215, 119369.

İNTİHAL RAPORU İLK SAYFASI

Halil İbrahim OKUR

ORİJİNALLİK RAPORU

%5	%4	%2	%3
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1	Submitted to The Scientific & Technological Research Council of Turkey (TUBITAK) Öğrenci Ödevi	%2
2	acikbilim.yok.gov.tr İnternet Kaynağı	<%1
3	dergipark.org.tr İnternet Kaynağı	<%1
4	avesis.istanbulc.edu.tr İnternet Kaynağı	<%1
5	www.odatv.com İnternet Kaynağı	<%1
6	openaccess.iste.edu.tr İnternet Kaynağı	<%1
7	Gencer, Mustafa. "Yapısal Olmayan Metinler İçin Adlandırılmış Varlık Tanıma Algoritmaları ve Uygulamaları", Dokuz Eylul Universitesi (Turkey), 2024 Yayın	<%1
8	mdbf.ksbu.edu.tr İnternet Kaynağı	<%1