**KÜMELEME TABANLI METİN SINIFLANDIRMA**

**Ensar ŞEHİTOĞLU**

**Yüksek Lisans Tezi  
Bilgisayar Mühendisliği Anabilim Dalı  
---  
2025**(Her hakkı saklıdır.)

T.C.

ATATÜRK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

**KÜMELEME TABANLI METİN SINIFLANDIRMA ÖRNEĞİ**

(AN EXAMPLE of CLUSTERING-BASED TEXT CLASSIFICATION)

YÜKSEK LİSANS TEZİ

Ensar ŞEHİTOĞLU

Danışman: ---

Erzurum  
Temmuz, 2025

# KABUL VE ONAY TUTANAĞI

… danışmanlığında, Ensar Şehitoğlu tarafından hazırlanan bu çalışma, …/…/20... tarihinde aşağıdaki jüri tarafından Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Bilim Dalı’nda Yüksek Lisans tezi olarak **oybirliği / oy çokluğu (…/…)** ile kabul edilmiştir.

|  |  |  |
| --- | --- | --- |
| Jüri Başkanı: | Unvan Ad SOYAD  *Üniversite Adı* | ………................. |
| Danışman: | Unvan Ad SOYAD  *Üniversite Adı* | ………................. |
| Jüri Üyesi: | Unvan Ad SOYAD  *Üniversite Adı* | ………................. |
|  |  | Enstitü Yönetim Kurulunun …./…/…. tarih ve …… sayılı kararı. |

Bu tezin Atatürk Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliği’nin ilgili maddelerinde belirtilen şartları yerine getirdiğini onaylarım.

**Prof. Dr. Saltuk Buğrahan CEYHUN**

**Enstitü Müdürü**

**Not:** Bu tezde kullanılan özgün ve başka kaynaklardan yapılan bildiriş, çizelge, şekil ve fotoğrafların kaynak olarak kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere tabidir.

# ETİK BİLDİRİM VE İNTİHAL BEYAN FORMU

Yüksek Lisans Tezi olarak … danışmanlığında sunulan “**Kümeleme Tabanlı Metin Sınıflandırma Örneği**” başlıklı çalışmanın tarafımızdan bilimsel etik ilkelere uyularak yazıldığını, yararlanılan eserlerin kaynakçada gösterildiğini, Fen Bilimleri Enstitüsü tarafından belirlenmiş olan Turnitin Programı benzerlik oranlarının aşılmadığını ve aşağıdaki oranlarda olduğunu beyan ederiz.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tez Bölümleri** | **Tezin Benzerlik Oranı (%)** | | **Maksimum Oran (%)** | |
| Giriş | | *Tezdeki benzerlik oranın yazınız* | | **30** |
| Kuramsal Temeller | | *Tezdeki benzerlik oranın yazınız* | | **30** |
| Materyal ve Metot | | *Tezdeki benzerlik oranın yazınız* | | **35** |
| Araştırma Bulguları ve Tartışma | | *Tezdeki benzerlik oranın yazınız* | | **20** |
| Sonuçlar | | *Tezdeki benzerlik oranın yazınız* | | **20** |
| Tezin Geneli | | *Tezdeki benzerlik oranın yazınız* | | **25** |

***Not:*** *Yedi kelimeye kadar benzerlikler ile Başlık, Kaynakça, İçindekiler, Teşekkür, Dizin ve Ekler kısımları tarama dışı bırakılabilir. Yukarıdaki azami benzerlik oranları yanında tek bir kaynaktan olan benzerlik oranlarının* ***%5’den*** *büyük olmaması gerekir.*

Beyan edilen bilgilerin doğru olduğunu, aksi halde doğacak hukuki sorumlulukları kabul ve beyan ederiz.

|  |  |
| --- | --- |
| **Tez Yazarı (Öğrenci)** | **Tez Danışmanı** |
| Ensar ŞEHİTOĞLU | --- |
| Tarih girmek için burayı tıklayınız | Tarih girmek için burayı tıklayınız |
| İmza: | İmza: |

**\*** Tez ile ilgili YÖKTEZ’de yayınlamasına ilişkin bir engelleme var ise aşağıdaki alanı doldurunuz.

Tezle ilgili patent başvurusu yapılması / patent alma sürecinin devam etmesi sebebiyle Enstitü Yönetim Kurulunun …./…/…. tarih ve …………. sayılı kararı ile teze erişim 2 (iki) yıl süreyle engellenmiştir.

Enstitü Yönetim Kurulunun …./…/…. tarih ve …………. sayılı kararı ile teze erişim 6 (altı) ay süreyle engellenmiştir.

# TEŞEKKÜR

…

# ÖZET

**YÜKSEK LİSANS TEZİ**

**Kümeleme Tabanlı Metin Sınıflandırma Örneği**

**Ensar ŞEHİTOĞLU**

**Danışman: …**

**Amaç:** Metin sınıflandırma birçok farklı probleme çözüm üretmek için kullanılmaktadır. Bu çalışmada, metin verilerinden yola çıkarak bireylerin cinsiyet tahmininin yapılması amaçlanmıştır. Geleneksel sınıflandırma yaklaşımlarının aksine, etiketlenmemiş veri kullanımını ön planda tutan kümeleme tabanlı yöntemler tercih edilmiştir. Bu yaklaşım ile etiketlenmiş veri eksikliğine rağmen anlamlı sonuçlar elde edilmesini sağlamak hedeflenmiştir. Bu bağlamda, hem kümeleme tabanlı tekniklerin hem de gözetimsiz öğrenme yöntemlerinin etkinliği değerlendirilmiştir.

**Yöntem:** Bu çalışmada, metin sınıflandırma ve kümeleme tabanlı bir yöntem kullanılarak cinsiyet tahmini yapılmıştır. Çalışmada Twitter mesajları ile elde edilen Türkçe ve İngilizce veri kümeleri üzerinde önişleme yapılarak ve önişleme yapılmadan 2 farklı senaryo uygulanmıştır. Metin temsili için geleneksel teknikler (TF-IDF, Bag-of-Words vb.) ile kelime gömme (Word2Vec, GloVe, FastText) ve dil modelleri (BERT, T5, RoBERTa vb.) kullanılmıştır. Kümeleme algoritmaları (K-Means, Agglomerative Clustering, Birch vb.) ile veriler gruplandırılmış, ardından çapraz doğrulama ile sınıflandırıcılar (Logistic Regression, SVM, Random Forest vb.) değerlendirilmiştir. Kümeleme performansı için Silhouette, Davies-Bouldin ve Calinski-Harabasz skorları hesaplanmıştır. Tüm süreç, model tutarlılığı için sabit bir rastgelelik değeri ile tekrarlanabilir şekilde tasarlanmıştır.

**Bulgular:** Çalışma kapsamında kullanılan veri kümesi ve yöntemler, farklı kümeleme algoritmalarının performanslarını detaylı bir şekilde analiz etmek için kullanılmıştır. Silhouette skoru ve diğer değerlendirme metrikleri, kümelerin ayrışma düzeyini başarılı bir şekilde ortaya koymuştur. Özellikle, HDBSCAN algoritması, veri setindeki karmaşık yapıları anlamada diğer yöntemlere kıyasla daha iyi performans göstermiştir. Az sayıda etiketli veri ile gerçekleştirilen bu deneyler sonucunda, cinsiyet tahmini sınıflandırmasında yüksek doğruluk oranlarına ulaşıldığı gözlemlenmiştir.

**Sonuç:** Kümeleme tabanlı yöntemlerin, etiketlenmiş veri eksikliğinde metin sınıflandırmada etkili bir alternatif olduğu ortaya konmuştur. Çalışma, cinsiyet çıkarımı örneği üzerinden bu yöntemlerin başarı potansiyelini göstermiştir. Çalışma, cinsiyet çıkarımı örneği üzerinden bu yöntemlerin başarı potansiyelini göstermiştir.

**Anahtar Kelimeler:** Kümeleme, Metin Sınıflandırma, Cinsiyet Çıkarımı, Doğal Dil İşleme, Gözetimsiz Öğrenme.

**Mart 2018, 360 sayfa**

# ABSTRACT

**MASTER THESIS**

**TEACHING THE CONCEPTS IN PARTICULATE NATURE OF MATTER TO 6TH GRADE VISUALLY IMPAIRED STUDENTS**

**Ensar ŞEHİTOĞLU**

**Supervisor: Assoc. …**

**Purpose:** This study aimed todevelop an instructional design model to teach basic concepts in ‘Particulate Nature of Matter’ unit to 6th grade visually impairment students, evaluate the practicality of the design and its effectiveness in concept learning.… …

**Method:** This study was carried out through a Design Based Research (DBR) approach in three stages. In the first stage, needs of visually impaired students regarding science learning and their understanding level of the concepts in ‘Particulate Nature of Matter’ unit were investigated… …

**Findings:** Based on the data collected, it is found that the level of students’ learning of the concepts in “Particulate Nature of Matter” unit is improved, as well as, the instructional design model developed contributed to students science learning… …

**Results:** Instructional design enhanced with tactile materials, which is developed according to the visually impaired students’ needs, improved students’ science learning … …

**Keywords:** science teaching, ADDIE Instructional Design Model, particulate nature of matter, visual impairment

**March 2018, 360 pages**

# İÇİNDEKİLER

[KABUL VE ONAY TUTANAĞI i](#_Toc185541815)

[ETİK BİLDİRİM VE İNTİHAL BEYAN FORMU ii](#_Toc185541816)

[TEŞEKKÜR iii](#_Toc185541817)

[ÖZET iv](#_Toc185541818)

[ABSTRACT v](#_Toc185541819)

[İÇİNDEKİLER vi](#_Toc185541820)

[TABLOLAR DİZİNİ vii](#_Toc185541821)

[ŞEKİLLER DİZİNİ viii](#_Toc185541822)

[KISALTMALAR VE SİMGELER DİZİNİ ix](#_Toc185541823)

[GİRİŞ 1](#_Toc185541824)

[Deneme Başlığı 1](#_Toc185541825)

[KURAMSAL TEMELLER 2](#_Toc185541826)

[MATERYAL VE METOT 3](#_Toc185541827)

[ARAŞTIRMA BULGULARI VE TARTIŞMA 4](#_Toc185541828)

[SONUÇ VE ÖNERİLER 5](#_Toc185541829)

[KAYNAKÇA 6](#_Toc185541830)

[ÖZGEÇMİŞ 7](#_Toc185541831)

# TABLOLAR DİZİNİ

[**Tablo 1.** Örnek Tablo 1](#_Toc185622601)

[**Tablo 2.** Örnek Tablo2 1](#_Toc185622602)

[**Tablo 4.** Örnek Tablo 6](#_Toc185622603)

# ŞEKİLLER DİZİNİ

[**Şekil 1.** Deneme şekil1 1](#_Toc185622546)

[**Şekil 2.** Deneme şekil2 6](#_Toc185622547)

# KISALTMALAR VE SİMGELER DİZİNİ

**AC :** Yığılımsal Kümeleme (Agglomerative Clustering)

**ALBERT :** Hafif BERT (A Lite BERT)

**API :** Uygulama Programlama Arayüzü (Application Programming Interface)

**BERT :** Transformatörlerden Çift Yönlü Kodlayıcı Temsilleri (Bidirectional Encoder Representations from Transformers)

**BoW :** Kelime Çantası (Bag of Words)

**CF :** Kümeleme Özelliği (Clustering Feature)

**CV :** K Katlı Çapraz Doğrulama (K Fold Cross Validation)

**DBSCAN :** Yoğunluk Bazlı Uzamsal Kümeleme Uygulamaları ile Gürültü (Density-Based Spatial Clustering of Applications with Noise)

**DT :** Karar Ağacı (Decision Tree)

**ELECTRA :** Token Değiştirmelerini Doğru Şekilde Sınıflandıran Bir Kodlayıcıyı Verimli Bir Şekilde Öğrenme (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)

**ERNIE :** Bilgi Entegrasyonu ile Geliştirilmiş Temsil (Enhanced Representation through Knowledge IntEgration)

**FastText :** Hızlı Metin Temsili (Fast Text Representation)

**FlauBERT** **:** Fransızca Dilinde BERT (French Language BERT)

**FN** **:** Yanlış Negatif (False Negative)

**FNR :** Yanlış Negatif Oranı (False Negative Rate)

**FP** **:** Yanlış Pozitif (False Positive)

**FPR :** Yanlış Pozitif Oranı (False Positive Rate)

**GloVe :** Küresel Vektörler ile Kelime Temsili (Global Vectors for Word Representation)

**HC** **:** Hiyerarşik Kümeleme (Hierarchical Clustering)

**K-Means** **:** K-Ortalamalar (K-Means)

**KNN :** K-En Yakın Komşular (K-Nearest Neighbors)

**LR :** Lojistik Regresyon (Logistic Regression)

**MarianMT :** MarianMT (Marian Neural Machine Translation)

**MLP :** Çok Katmanlı Algılayıcı (Multilayer Perceptron)

**MNB :** Çok Terimli Naive Bayes (Multinomial Naive Bayes)

**n-gram :** n-Uzunluklu Dilimleme (n-Length Slicing)

**NLP :** Doğal Dil İşleme (Natural Language Processing)

**NLTK :** Doğal Dil Araç Takımı (Natural Language Toolkit)

**RF :** Rastgele Orman (Random Forest)

**RoBERTa :** Güçlü Bir Şekilde Optimize Edilmiş BERT Ön Eğitim Yaklaşımı (A Robustly Optimized BERT Pretraining Approach)

**SC** **:** Spektral Kümeleme (Spectral Clustering)

**SVM :** Destek Vektör Makineleri (Support Vector Machines)

**T5 :** Metinden Metne Transfer Transformatörü (Text-to-Text Transfer Transformer)

**TF** **:** Terim Sıklığı (Term Frequency)

**TF-IDF** **:** Terim Sıklığı - Ters Belge Sıklığı (Term Frequency - Inverse Document Frequency)

**TN** **:** Gerçek Negatif (True Negative)

**TP** **:** Gerçek Pozitif (True Positive)

**TPR :** Gerçek Pozitif Oranı (True Positive Rate)

**Word2Vec :** Kelimeden Vektöre (Word to Vector)

**XLM :** Çapraz Dilli Dil Modeli (Cross-lingual Language Model)

**YSA :** Yapay Sinir Ağları (Artificial Neural Networks)

# GİRİŞ

Dil kullanımı, toplumsal cinsiyetin yansımalarını anlamak için önemli bir araçtır. Yazılı iletişimde bireylerin dil kullanımındaki kalıplar, cinsiyetin belirlenmesine yönelik değerli ipuçları sunmaktadır. Özellikle Enron e-posta veri seti üzerinde yapılan çalışmalar, yazılı dildeki ifadelerin cinsiyetlere göre farklılık gösterebileceğini ortaya koymuştur.

Yazılı iletişimde cinsiyet çıkarımı, bireylerin dil kullanımını analiz ederek cinsiyetlerinin otomatik olarak belirlenmesini amaçlayan bir yaklaşımdır. Doğal dil işleme teknikleri sayesinde, yazılı mesajlardaki dilsel kalıplar analiz edilmekte ve bu kalıpların cinsiyet tespiti üzerindeki etkileri değerlendirilmektedir. Bu tür çalışmalar, yazılı iletişimde cinsiyetin dil üzerindeki etkilerini anlamaya yönelik teorik ve uygulamalı katkılar sunmakta, aynı zamanda dilin toplumsal boyutlarına dair yeni perspektifler geliştirilmesine olanak sağlamaktadır (Prabhakaran et al. 2014).

Toplumsal cinsiyet ve dil arasındaki ilişki, sosyal medya platformlarının günlük hayatımızda önemli bir yer edinmesiyle birlikte, dilbilim, sosyoloji ve veri bilimi gibi çeşitli disiplinlerde dikkat çeken bir araştırma alanı haline gelmiştir. Sosyal medya, bireylerin dilsel tercihlerini toplumsal cinsiyet normları çerçevesinde şekillendirdiği dinamik bir ortam sunmaktadır. Özellikle Twitter gibi platformlar, kullanıcıların bireysel kimliklerini ve toplumsal aidiyetlerini dil yoluyla ifade ettikleri, aynı zamanda sosyal etkileşimlerin dilsel boyutunun incelenebileceği önemli bir bağlam oluşturur.

Ancak, toplumsal cinsiyetin dil üzerindeki etkisi, basit bir kadın/erkek ikiliğiyle sınırlandırılamaz. Bu durum, toplumsal cinsiyetin çeşitli dilsel stiller, duruşlar ve kişilikler aracılığıyla karmaşık bir şekilde ifade edildiğini göstermektedir. Bu çalışmada, Twitter kullanıcılarının dilsel benzerliklerine göre gruplandırıldığı bir kümeleme analizi yönteminden faydalanılmıştır. Araştırma, bu grupların çoğunun belirgin toplumsal cinsiyet eğilimleri sergilediğini ortaya koymaktadır. Bununla birlikte, dilsel kaynakların kullanımında, genel dil istatistikleriyle çelişen örüntüler de tespit edilmiştir. Örneğin, kadınların genel olarak daha fazla duygu terimi ve emotikon kullandıkları bilinse de, bazı kadın ağırlıklı kümelerde bu eğilim tersine dönebilmektedir. Benzer şekilde, erkeklerin genelde daha fazla isim ve spor terimi kullandıkları gözlemlenirken, bazı erkek ağırlıklı gruplarda bu kullanım oranlarının düşük olduğu belirlenmiştir.

Ayrıca, sosyal çevredeki toplumsal cinsiyet homojenliğinin, dilsel özelliklerin ana akım toplumsal cinsiyet normlarına uyumuyla güçlü bir ilişki içinde olduğu saptanmıştır. Sosyal çevresinde aynı cinsiyetten bireylerin yoğunlukta olduğu kullanıcıların, toplumsal cinsiyet normlarına daha uygun dilsel özellikler sergileme eğiliminde olduğu görülmüştür. Bu durum, dilin yalnızca bireysel bir ifade aracı olmadığını, aynı zamanda sosyal etkileşim ve kimlik inşasının bir aracı olduğunu da göstermektedir. Bu tür analizler, toplumsal cinsiyet eşitliği konularında farkındalık yaratmaya yardımcı olabilir. Sonuç olarak, sosyal medya platformlarındaki dil, toplumsal cinsiyetin karmaşık ve çok yönlü doğasını anlamak için önemli bir araştırma alanıdır (Bamman et al. 2014).

Twitter kullanıcılarının yaşını ve cinsiyetini tahmin etmek, sosyal medya verilerinin değerini ve uygulamasını artırmada önemli bir adımdır (O'Connor et al. 2024). Twitter, kullanıcılarının kısa mesajlar veya tweet'ler yayınladığı bir sosyal ağ ve mikro blog platformudur. 2010'un sonlarında, Twitter'ın dünya genelinde 175 milyon kayıtlı kullanıcıya sahip olduğu ve günde 65 milyon tweet üretildiği tahmin ediliyordu. Twitter, geniş hacmi, çeşitli ve çok dilli kullanıcı kitlesi ve hizmet şartlarının cömertliği nedeniyle sosyal medya araştırmaları için cazip bir platformdur. Bu durum, birçok araştırmacıyı Twitter verilerinden oluşan derlemeler oluşturmaya yönlendirmiştir (Burger et al. 2011).

Twitter'daki gayri resmi dil, geleneksel metin madenciliğine çok sayıda zorluk çıkarsa da, gayri resmi metnin ayırt edici özelliklerinin çoğu yazarlık analizi için yararlı bilgiler sağlayabilir. Geçmişte metin yoluyla cinsiyet tahmini, öncelikle cümle yapısı ve noktalama işaretleri veya kelime sayıları, sözcük türleri ve diğer sözlük tabanlı yöntemler kullanıyordu (Corney, 2003).

Cinsiyet çıkarımı, bir yazarın metinlerinden cinsiyetinin tahmin edilmesi sürecidir. Bu amaçla, metinlerdeki dil kullanımı, sözcük tercihleri ve yazım stili gibi özellikler analiz edilir. Özellikle Türkçe metinlerde, yazarın cinsiyetini belirlemeye yönelik çalışmalar yapılmıştır. Örneğin, n-gram ve kelime torbası yaklaşımları kullanılarak, metinlerin yazarını tanıma, türünü belirleme ve cinsiyet tespiti gibi problemler üzerine modeller geliştirilmiştir (Yıldırım & Yıldız, 2018).

Cinsiyet tahmini, bireylerin cinsiyetini belirlemek için kullanılan çeşitli yöntemleri ifade eder. Bu yöntemler, el yazısı analizi (Maken et al. 2019), isim analizi (Wais et al. 2016), sosyal medya gönderileri (Suman et al. 2022) ve akıllı telefon uygulama verileri (Wang et al. 2015) gibi farklı veri kaynaklarına dayanabilir. Kullanıcı yorumlarından fiiller, zamirler ve sıfatlar gibi dil özelliklerini analiz etmek, cinsiyeti orta düzeyde doğrulukla tahmin edebilir. Bu yöntem, belirli dil kalıpları ile cinsiyet arasındaki ilişkiyi kullanır (Hosseini et al. 2016).

Kümeleme tabanlı metin sınıflandırma, özellikle etiketlenmemiş büyük veri setlerinde, verilerin doğal yapısını ortaya çıkarmada etkilidir. Bu yöntem, metinlerin otomatik olarak sınıflandırılmasını sağlar ve etiketleme maliyetini düşürür (Akşit, 2020). Ayrıca, farklı yazarların metinleri arasındaki benzerlikleri belirleyerek, yazarların cinsiyetlerinin tahmin edilmesinde kullanılabilir. Bu bağlamda, kümeleme algoritmaları, metin madenciliğinde önemli bir araçtır (Vicente et al. 2015).

Cinsiyet çıkarımı yöntemin uygulama alanları arasında pazarlama stratejilerinin geliştirilmesi, hedef kitle analizleri ve ürün tasarımı bulunmaktadır. Örneğin, bir moda markası, müşteri yorumlarını analiz ederek, kadın ve erkek müşterilerin ürünler hakkındaki görüşlerini ayrı ayrı değerlendirebilir ve bu sayede ürün geliştirme süreçlerini optimize edebilir. Ayrıca, reklam ajansları, cinsiyet temelli algı farklılıklarını dikkate alarak, görsel tasarımları hedef kitleye uygun şekilde hazırlayabilirler. Cinsiyet, tüketici davranışlarının analizi, reklam stratejileri, e-ticaret uygulamaları, ürün tasarımı ve pazarlama, kültürel ve toplumsal yapıların anlaşılması gibi alanlarda kritik bir rol oynar. Feminen ve maskülen kimliklere göre farklı reklam yaklaşımları geliştirilirken, sosyo-ekonomik statüyle etkileşimler cinsiyet kimliklerini şekillendirir. Bu kavram, kültürel ve toplumsal yapıların anlaşılmasında da önemli bir araçtır (Yağcı & İlarslan, 2010; Alabay, 2011). Ayrıca cinsiyet tahmini, kimlik hırsızlığı ve diğer suç faaliyetleri için kullanılan sahte profilleri ve klon profil saldırılarını tespit etmede kullanılır (Vyawahare & Govilkar, 2022). Sosyal medyada cinsiyet tahmini, güvenlik ve pazarlama alanındaki uygulamaları nedeniyle dikkat çekmiştir (Hosseini & Tammimy, 2016).

Bu tez çalışmasının amacı, kümeleme tabanlı metin sınıflandırma tekniklerinin cinsiyet çıkarımı görevindeki etkinliğini incelemektir. Çalışmanın amacı, metinlerin içerik özelliklerine dayanarak cinsiyet tahmininin doğruluğunu artırmak ve bu süreçte kullanılan yöntemlerin etkinliğini değerlendirmektir. Ayrıca, cinsiyet çıkarımında kullanılan veri setleri ve performans ölçütleri de detaylı bir şekilde ele alınacaktır. Bu bağlamda, çalışmanın literatüre katkısı, cinsiyet çıkarımında kümeleme tabanlı yaklaşımların potansiyelini ortaya koymak olacaktır.

# 

# KURAMSAL TEMELLER

Onikoyi ve arkadaşları (Onikoyi et al., 2023) bu araştırmada, Twitter'da cinsiyet tahmini doğruluğunu artırmak amacıyla çeşitli yöntemler kullanarak mevcut veri setini genişletmiştir. Çalışma, Twitter Kullanıcı Cinsiyet Sınıflandırması veri setini 296.108 ek tweet ile genişletmiş ve veri ön işleme işlemleriyle modelin eğitilmesine uygun hale getirilmiştir. Ayrıca, GloVe, BERT, GPT2 ve Word2Vec gibi gelişmiş kelime gömme modelleri kullanılarak kelimeler arasındaki anlamsal ilişkiler yakalanmıştır. Özellikle GloVe modeli, 2 milyar tweet üzerinde eğitilen versiyonuyla en yüksek doğruluk oranını elde etmiştir. Profil açıklamalarının dahil edilmesi de doğruluğu artırmış ve cinsiyet tahmininde önemli bir katkı sağlamıştır. Çalışmada, Naive Bayes, Rastgele Orman, Karar Ağacı ve XGBoost gibi farklı makine öğrenimi algoritmalarının performansları değerlendirilmiş ve GloVe modeliyle birleştirilen Rastgele Orman algoritması %70 doğrulukla en iyi performansı göstermiştir.

Vicente ve arkadaşları (Vicente et al., 2015) bu çalışmada, Twitter kullanıcılarının cinsiyetini yalnızca profil bilgilerine dayanarak otomatik olarak tahmin etmeyi amaçlamıştır. Araştırmacılar, kullanıcı adları ve ekran adlarındaki isim tabanlı özellikleri kullanarak %97'ye varan doğruluk oranları elde etmişlerdir. Bu özellikleri değerlendirmek için Lojistik Regresyon, Destek Vektör Makineleri ve Naive Bayes gibi denetimli öğrenme algoritmaları kullanılmıştır. Çalışmada, Fuzzy c-Means kümeleme algoritması da kullanılarak etiketlenmemiş verilerle %96'lık bir doğruluk oranına ulaşılmıştır. Fuzzy c-Means, etiketlenmiş veri gerektirmediği ve performansı veri miktarı arttıkça iyileştiği için özellikle Twitter gibi büyük veri kümeleri için avantajlıdır. Bu sonuçlar, Fuzzy c-Means'in Twitter'da cinsiyet tespiti için etkili ve ölçeklenebilir bir yöntem olduğunu göstermiştir.

Bamman ve arkadaşları (Bamman et al., 2014) bu çalışmada, Twitter'da 14.000'den fazla kullanıcıdan oluşan bir veri seti kullanarak, dilin cinsiyet kimliğini nasıl ifade ettiğini ve sosyal bağlar üzerindeki etkisini araştırmıştır. Cinsiyetin dil üzerindeki etkisini anlamak için bir lojistik regresyon modeliyle 10.000 sözcükten oluşan bir kelime dağarcığı kullanılmış ve model %88 doğruluk oranıyla cinsiyeti metne dayalı olarak tahmin etmiştir. Ayrıca, sosyal ağ homofilisinin (aynı cinsiyet ilişkilerinin yoğunluğu) dil üzerindeki etkisinin güçlü olduğu görülmüştür. Ancak bazı bireylerin cinsiyet normlarına uymayan dil kullanımları, farklı kimlikler ve toplumsal bağlamların yansımaları olarak değerlendirilmiştir. Araştırma, cinsiyetin yalnızca ikili bir değişken olarak ele alınamayacağını, dil ve sosyal ağların cinsiyet kimliğinin çok boyutlu bir şekilde inşa edildiğini gösterdiğini savunmaktadır.

Alowibdi ve arkadaşları (Alowibdi et al., 2013) bu araştırmada, Twitter'da cinsiyet sınıflandırması için profil özelliklerinin kullanımını incelemiştir. Kullanıcı profillerinden elde edilen veriler, cinsiyet tahmini için etkili bir kaynak olarak değerlendirilmiştir. Araştırmada fonem tabanlı özellik çıkarımı, renk kuantizasyonu ve sıralama gibi teknikler kullanılarak isimler, kullanıcı adları ve profil renkleri analiz edilmiştir. Fonem tabanlı yaklaşım, isim analizine dayalı olarak %82.5 doğruluk oranı sağlarken, renk tabanlı özelliklerle yapılan analizlerde NB-Tree sınıflandırıcısı %74 doğruluk oranına ulaşmıştır.

Miller ve arkadaşları (Miller et al., 2012) bu çalışmada, sosyal ağların hızlı büyümesiyle elde edilen büyük kullanıcı verilerinin, yazarlık analizi ve cinsiyet gibi bilgilerin çıkarılmasında önemli fırsatlar sunduğunu vurgulanmıştır. Twitter, kullanıcı verilerini API aracılığıyla sunarak bu veriler için değerli bir kaynak oluşturmuştur. Çalışmada, tweet metninden seçilen 1 ila 5 gramlık n-gram özellikleriyle Perceptron ve Naive Bayes algoritmaları kullanılarak Twitter kullanıcılarının cinsiyeti belirlenmeye çalışılmıştır. N-gram özelliklerinin, tweetlerin hızlı akışını ve hacmini ele almak için akış uygulamalarında kullanıldığı, ayrıca geleneksel sözlük yöntemleriyle değerlendirilemeyen resmi olmayan tweetlerin n-gramlarla temsil edildiği belirtilmiştir. Seçilen en iyi n-gram özellikleri ile Naive Bayes ve Perceptron algoritmaları, %99'un üzerinde doğruluk, dengeli doğruluk ve F1-score elde etmiştir.

Rego ve arkadaşları (Rego et al., 2021) bu çalışmada, cinsiyeti isme göre tahmin etmek için çeşitli makine öğrenimi algoritmaları incelenmiş ve uygulanmıştır. Yabancı isimler üzerinde yapılan bu tür tahminler, özellikle doğal dil işleme (NLP) alanında önemli bir yer tutmaktadır. Extra Trees, KNN, Naive Bayes, SVM, Random Forest, Gradient Boosting, Light GBM, Logistic Regression, SVC Classifier ve derin sinir ağı modelleri olan MLP, RNN, GRU, CNN ve BiLSTM gibi yöntemler değerlendirilmiştir. Brezilya isimlerinden oluşan bir veri kümesi kullanılarak modeller eğitilmiş ve değerlendirilmiştir. Performans ölçümleri olarak accuracy, recall, precision, F1-score ve confusion matrix analiz edilmiştir. Sonuçlar, isimlere dayalı özellik çıkarımı kullanılarak cinsiyet tahmininin başarılı bir şekilde yapılabildiğini göstermektedir. Bazı modeller, vakaların %95'inden fazlasında doğru tahminler yapabilmiştir ve tekrarlayan modeller, bu ikili sınıflandırma probleminde ileri beslemeli modellere göre üstün performans sergilemiştir.

To ve arkadaşları (To et al., 2020) bu çalışmada, bireylerin biyolojik cinsiyetinin adlara dayalı sınıflandırılması üzerine birçok çalışma yapıldığını, ancak Vietnamca adlarla ilgili çalışmaların sınırlı olduğunu görmüştür. Vietnamca adlara dayalı cinsiyet tahmini için 26.000'den fazla tam ad içeren ve cinsiyet etiketleriyle anotasyon yapılmış yeni bir veri seti önerilmiştir. Bu veri seti araştırma amaçlı olarak erişime sunulmuştur. Ayrıca, altı makine öğrenimi algoritması (SVM, Multinomial Naive Bayes, Bernoulli Naive Bayes, Karar Ağacı, Rastgele Orman ve Lojistik Regresyon) ile bir derin öğrenme modeli (LSTM) üzerinde fastText kelime gömme kullanılarak deneyler yapılmıştır. Sonuçlara göre, en yüksek F1 puanı %96 ile LSTM modeliyle elde edilmiştir ve bu model temel alınarak bir web API geliştirilmiştir.

Basile ve arkadaşları (Basile et al., 2019) bu çalışmada, Evalita 2018 GxG çapraz tür/alan cinsiyet tahmini yarışmasında İtalyanca metinler üzerine bir araştırma gerçekleştirmiştir. N-gram özelliklerini kullanan doğrusal bir modelin farklı türler arası dayanıklılığı değerlendirilmiştir. Eğitim ve test türlerinin farklı olduğu durumlarda model performansının belirgin şekilde düştüğü gözlemlenmiştir. Tür bağımsız özellikler yakalamak için soyut özellikler de test edilmiştir. Çalışmada, resmi in-genre test setinde %55 F1 puanı ile birinci sırada yer alınmış, cross-genre test setinde ise %51 F1 puanı elde edilmiştir. Bu sonuçlar, türler arası cinsiyet tahmininde mevcut modellerin sınırlılıklarını göstermiştir.

Vanmassenhove ve arkadaşları (Vanmassenhove et al., 2019) bu çalışmada, Hollandaca için türler arası cinsiyet tahmini üzerine CLIN29 2019 ortaklaşa görevinde sunulan sistem açıklanmıştır. Çalışmada, sinirsel modeller (CNN, RNN, LSTM vb.), geleneksel yöntemler (SVM, RF, LogReg vb.), çeşitli özellik kümeleri ve veri ön işleme teknikleri ile deneyler yapılmıştır. Sonuçlar, belirteçleştirilmiş ve küçük harfe dönüştürülmemiş verilerin sinirsel modellerde en iyi performansı sağladığını, kelime kümeleri, karakter üçlüleri ve kelime listelerinin ise geleneksel modeller için en etkili kombinasyonu oluşturduğunu göstermiştir. Bu yaklaşımlar, önceki görevlerde kullanılan n-gramlar, karakter n-gramları ve sözcük türü etiketleri gibi özellikleri geride bırakmıştır. En iyi modellerin birleştirilmesiyle oluşturulan ağırlıklı topluluk modeli, cinsiyet tahmini ve türler arası zorluk görevlerinde birincilik elde etmiş, bu süreçte sırasıyla %64,93 ve %56,26 doğruluk oranlarına ulaşılmıştır.

Tripathi ve Faruqui (Tripathi & Faruqui, 2011) bu çalışmada, Hint isimlerinin cinsiyet tahmini için Destek Vektör Makinesi (SVM) tabanlı bir sınıflandırma yaklaşımı sunulmuştur. İlk olarak, morfolojik analizlere dayalı olarak bu tür bir sınıflandırma için faydalı olabilecek çeşitli özellikler belirlenmiş ve değerlendirilmiştir. Ardından, bu özelliklerle birlikte n-gram eklerinin kullanımına dayalı yeni bir yaklaşım önerilmiştir. Bu yöntem, temel yaklaşıma kıyasla önemli bir avantaj sağlamaktadır. Tahmin sistemlerinde tüm kelimeler yerine n-gram eklerinin kullanıldığı bu çalışmanın, bu alandaki ilk uygulama olduğuna inanılmaktadır. Sistem, eğitim verisi boyutundaki artışla daha da iyileşmesi beklenen %94,9'luk bir en iyi F1 puanı elde etmiştir.

Prabhakaran ve arkadaşları (Prabhakaran et al., 2014) bu çalışmada, otomatik olarak cinsiyet ataması yapmak için Sosyal Güvenlik İdaresi'nin bebek isimleri veri tabanını kullanarak Enron veri tabanındaki e-postaların %87'sinin göndereninin cinsiyetini güvenilir bir şekilde tanımlayan bir yöntem geliştirilmiştir. Bu, büyük ölçekli metin veri tabanlarında cinsiyet çıkarımının uygulanabilirliğini göstermektedir.

(Liu, 2023) bu çalışmada, sosyal ağ metinlerinin duygu analizi için iki yönlü kısa ve uzun vadeli bellek modeli (AT-BiLSTM) önermiştir. Bu yöntem, BERT modeliyle metinlerin vektörleştirilmesini sağlamış ve dikkat mekanizması ile entegre edilen BiLSTM ile yön düzeyinde duygu analizi gerçekleştirmiştir. Aynı veri kümesi üzerinde yapılan değerlendirmelerde, önerilen yöntemin %93,72 doğruluk, %93,91 geri çağırma ve %92,38 F1 puanı ile alternatif yöntemlere kıyasla daha yüksek performans sergilediği tespit edilmiştir.

Yu ve arkadaşları (Yu et al., 2021) bu araştırmada, metin sınıflandırmasının doğal dil işleme alanında yaygın bir uygulama olduğu ve derin öğrenmenin bu alandaki en sık kullanılan yöntemlerden biri olduğu tespit edilmiştir. Mevcut zorlukları aşmak için BERT-BiGRU modeli önerilmiştir. Bu modelde, kelime vektörleri için geleneksel word2vec yerine BERT kullanılmış ve bağlam bilgisine dayalı kelime temsiline olanak tanınmıştır. Ardından, iki yönlü özellik çıkarımı için BiGRU modeli BERT'e entegre edilmiştir. Deneysel sonuçlara göre, önerilen BERT-BiGRU modeli ile doğruluk, geri çağırma ve F1 puanı 0.9'un üzerinde değerler elde edilmiştir. Modelin, özellikle Çince metin sınıflandırma görevlerinde yüksek performans gösterdiği görülmüştür.

Bensahla ve arkadaşları (Bensahla et al., 2024) klinik PDF belgelerindeki gövde metninin denetimsiz olarak çıkarılması, doğal dil işleme (NLP) görevlerinin geliştirilmesi için önemli bir adım olarak kabul edilmiştir. Ancak, bu süreç hala bir zorluk olarak kalmıştır. Yapılan bir çalışmada, büyük veri setleri kullanarak gövde metnini çıkarmak için tasarlanmış gözetimsiz bir algoritma sunulmuştur. DBSCAN kümelemesi kullanılarak, yöntem metin bloklarını içerik ve koordinatlarına göre çıkarmış ve düzenlemiştir. Değerlendirme sonuçları, çeşitli tıbbi uzmanlık kaynakları arasında 0,82 ila 0,98 arasında değişen precision puanları, 0,62 ila 0,94 arasında recall puanları ve 0,71 ila 0,96 arasında F1-score puanları elde edilmiştir.

Urrutia ve arkadaşları (Urrutia et al., 2023) robot destekli müdahalelerde vibro-akustik sinyallerden (VA sinyalleri) özellik çıkarma yöntemleri incelenmiştir. Özellik çıkarma, Cepstrum Dönüşümü (CT), Mel-Frekans Cepstral Katsayıları (MFCC'ler) ve Hızlı Chirplet Dönüşümü (FCT) kullanılarak yapılmış, ardından boyut azaltma için PCA, t-SNE ve UMAP gibi teknikler uygulanmıştır. Sınıflandırma ise en yakın komşular algoritması ile gerçekleştirilmiştir. Sonuçlar, CT ve MFCC'nin boyut azaltma ile birleşiminde yüksek verimlilik göstermiş ve sınıflandırma metriklerinde %99 accuracy, recall ve F1-score değerleri elde edilmiştir. Kümeleme metriği ise 0,61 olarak bulunmuştur.

Taamneh ve arkadaşları (Taamneh et al., 2017) bu çalışmada, hiyerarşik kümeleme ve yapay sinir ağları (YSA) kullanılarak yol trafik kazalarının ciddiyeti tahmin edilmiştir. Verilerdeki çeşitliliği azaltmak için hiyerarşik kümeleme uygulanmış ve kazalar altı farklı kümeye ayrılmıştır. Ardından, her bir küme için iki YSA modeli oluşturulmuş ve ilk model tüm kazalarla eğitilmiş, ikinci modelde ise kazaların %66'sı eğitim, %34'ü ise test seti olarak kullanılmıştır. Sonuçlar, kümeleme uygulandıktan sonra elde edilen doğruluğun, kümeleme yapılmadan elde edilenden %11-%16 daha yüksek olduğunu, yüzdelik bölünme ile doğruluğun ise %2-%5 arttığını göstermiştir. Yapay sinir ağı modeli ile verilerin kümelenmesi sayesinde %99,8 doğruluk oranına ulaşılmıştır.

Davagdorj ve arkadaşları (Davagdorj et al., 2022) bu çalışmada, biyomedikal belgelerin tematik olarak tutarlı gruplarını keşfetmek için BioBERT dil temsilleri ve Gauss Karışım Modeli tabanlı bir kümeleme çerçevesi önerilmiştir. PubMed veritabanından alınan biyomedikal belgeler üzerinde, klasik metin ön işleme teknikleri kullanılmış, ardından BioBERT ile vektör temsiller elde edilmiştir. Kümeleme işlemi için Gauss Karışım Modeli kullanılmıştır. Deneysel sonuçlar, önerilen modelin sırasıyla Fowlkes-Mallows puanı 0,7817, silüet katsayısı 0,3765, ayarlanmış rand endeksi 0,4478 ve Davies-Bouldin puanı 1,6849 elde ederek, kıyaslama modellerinden daha iyi performans gösterdiğini ortaya koymuştur.

Woo ve arkadaşları (Woo et al., 2019) bu çalışmada, büyük ölçekli tıbbi metin verisi için verimli bir veri temizleme süreci önerilmiştir. Tekrarlanan yazım hataları ve tutarsızlıklar gibi sorunları gidermek amacıyla metin kümeleme yöntemleri ve değer dönüştürme tekniği kullanılmıştır. Süreç, metin kümeleme ve değer birleştirme aşamalarından oluşmaktadır. Kümeleme aşamasında, anahtar çarpışma ve en yakın komşu yöntemleri karşılıklı olarak kullanılmıştır. Yanlış değerlerin doğru değerlerine dönüştürülmesi için değer dönüştürme aşaması uygulanmıştır. Önerilen süreç, 1995-2015 yılları arasında Samsung Tıp Merkezi'nde parazit analizi için üretilen 574.266 dışkı muayenesi raporuna uygulanmıştır. Yazım hataları ve çoğaltma desenlerinde %98,61 ve %97,78 oranlarında yüksek düzeltme oranları elde edilmiştir. Sonuç olarak, veri doğruluğu toplam kelimeler üzerinden %100 civarında elde edilmiştir.

Wang ve arkadaşları (Wang et al., 2023) bu çalışmada, tıbbi metin sınıflandırması için, ERNIE-Health adlı ayırt edici bir önceden eğitimli dil modeli kullanılarak prompt-tuning tekniği ile sınıflandırma yapılması incelenmiştir. Yöntem, metni yeni bir dizeye sararak, kategori etiketini [UNK] token'ı ile değiştirir ve model, aday kategorilerin olasılık dağılımını hesaplamak için eğitilir. Yöntem, KUAKE-Question Intention Classification ve CHiP-Clinical Trial Criterion veri setlerinde test edilmiştir. Deneysel sonuçlar, sırasıyla 0.866 ve 0.861 accuracy değerleri elde edilmiştir. Ayrıca, modelin kayıp (loss) değerlerinin eğitim süresi boyunca fine-tuning yöntemine kıyasla daha hızlı düştüğü gözlemlenmiştir.

Yetisgen-Yildiz ve Pratt (Yetisgen-Yildiz & Pratt, 2005) bu çalışmada, tıbbi metin sınıflandırmasındaki genel performansı etkileyen metin temsili tekniklerini incelemiştir. Bu hedefe ulaşmak için, temel kelime temsili (bag-of-words) ve daha karmaşık tıbbi ifade temsili (bag-of-phrases) destekleyen bir metin sınıflandırma sistemi geliştirilmiştir. Ayrıca, kelime ve ifade temsillerinin birleşimi (hibrit) kullanılarak daha kapsamlı bir analiz yapılmıştır. Sistem, tıbbi bilgi tabanını ve doğal dil işleme tekniklerini entegre ederek metinden tıbbi ifadeler çıkarmaktadır. Farklı temsillerin etkilerini değerlendirmek için OHSUMED veri setindeki MEDLINE belgeleri ile yapılan deneylerde sınıflandırma performansındaki değişim ölçülmüştür. Sınıflandırma performansı, bilgi erişim metrikleriyle ölçülmüş olup, precision (p), recall (r) ve F1-score (F1) değerleri kullanılmıştır. Deneylerde hibrit yaklaşım (p=0.87, r=0.46, F1=0.60) ile bag-of-words yaklaşımından (p=0.85, r=0.44, F1=0.58) ve bag-of-phrases yaklaşımından (p=0.87, r=0.42, F1=0.57) daha iyi sonuçlar elde edilmiştir.

Tan ve arkadaşları (Tan et al., 2023) bu çalışmada, deri kanseri görüntülerinin sınıflandırılmasında benzerlik kümelemesi ve dikkat aktarımı yöntemlerinin etkisi araştırılmıştır. Çalışma kapsamında, derin metrik öğrenmeye dayalı denetimsiz kümeleme yöntemi kullanılarak benzer görüntüler bir araya getirilmiş ve bu süreçte elde edilen model ağırlıkları, sınıflandırma modelini eğitmek için öğretici model olarak kullanılmıştır. Dikkat aktarımı yöntemi, modelin hem benzerlik özelliklerini hem de kategori bilgilerini aynı anda öğrenmesini sağlayarak tanı doğruluğunu artırmıştır. Elde edilen sonuçlar, önerilen yöntemin doğruluk oranını %72.53'e yükselterek başlangıç modeline göre %5'lik bir iyileşme sağladığını göstermiştir. Özellikle, üç farklı kötü huylu lezyonun F1-Skorlarında önemli artışlar gözlenmiştir: Bazal hücreli karsinom (BCC) 0.65'ten 0.73'e, Skuamöz hücreli karsinom (SCC) 0.28'den 0.37'ye ve Melanom (MEL) 0.54'ten 0.58'e yükselmiştir. Bu bulgular, benzerlik kümelemesi ve dikkat aktarımı yöntemlerinin, deri lezyonlarının tanısında toplam doğruluğu ve F1-Skorunu artırarak daha başarılı bir sınıflandırma sağladığını ortaya koymuştur.

Ahne ve arkadaşları (Ahne et al., 2022) bu çalışmada, diyabetle ilgili biyomedikal literatürün keşfini ve klinik karar verme sürecini iyileştirmek amacıyla, kullanıcı dostu bir etkileşimli sınıflandırma ve konu keşfi metodolojisi geliştirilmiştir. Bu yaklaşım, sağlık profesyonellerinin sınırlı bilgisayar bilimi bilgisiyle bile metin verilerini keşfetmelerini kolaylaştırmak için tasarlanmıştır. Yöntem, hiyerarşik kümeleme, hedefe yönelik sınıflandırma, kullanıcı etkileşimlerini en aza indiren aktif öğrenme ve görsel bir kullanıcı arayüzü gibi dört ana bileşenden oluşmaktadır. Performans değerlendirmesi, PubMed veritabanından alınan 50.911 diyabetle ilgili özet üzerinde gerçekleştirilmiştir. Hiyerarşik kümeleme için elde edilen F1 skoru 0,73 olarak raporlanmış ve scikit-learn kütüphanesi ile karşılaştırıldığında rekabetçi bir performans sergilemiştir. Aktif öğrenme stratejisinde, ağırlıklı F1 skoru 0,62 olarak elde edilmiş ve bu, alternatif stratejilere benzer bir performans göstermiştir. Yöntem ayrıca, artan veri miktarına rağmen düşük bellek kullanımı sunmuş, bu da büyük veri kümelerinde avantaj sağlamıştır.

Workman ve arkadaşları (Workman et al., 2017) bu çalışmada, semantik olarak ilişkili ancak farklı temsillere sahip kelimelerin ayrımını yapmanın, özellikle biyomedikal literatürde zorluk oluşturduğu belirtilmiştir. Bu bağlamda, demans ve deliryum gibi semantik olarak ilişkili kavramların ayrımını yapabilmek için, semantik öğeler arasındaki etkileşim sıklığı verilerinden yararlanılmıştır. Çalışmada, semantik predikasyonlardan elde edilen veri kümeleri iki farklı Expectation Maximization (EM) kümeleme sürecinde kullanılmıştır: biri etiketlenmemiş (unlabeled), diğeri ise kavram etiketleriyle birlikte. Etiketlenmemiş veri kümelerinde, kümelerin %80'inin beklenen kümelenme sayısını ve benzer oranları sağladığı, etiketli veri kümelerinde ise benzeyen veya eşleşen oranlar gözlemlendiği belirtilmiştir. En iyi performans gösteren sınıflandırıcı, %89 doğruluk oranına ulaşmış ve bireysel kavram sınıflandırmaları için F1 skorları 0.69 ile 1 arasında değişiklik göstermiştir. Sonuçlar, klinik metinlerin doğal dil işleme süreçlerinde nasıl kullanılabileceğine dair bir tartışma ile sunulmuştur.

Dipnall ve arkadaşları (Dipnall et al., 2022) bu çalışmada, radyoloji metinlerini kullanarak proksimal humerus kırıklarını sınıflandırmak için çeşitli makine öğrenmesi (ML) ve derin öğrenme algoritmalarının performansları değerlendirilmiş. 2010-2019 yılları arasında büyük bir travma merkezine kabul edilen hastaların verileri kullanılmış ve bu veriler X-ray ve BT raporları gibi metin kaynaklarıyla birleştirilmiş. Sınıflandırma amacıyla Neer sınıf etiketleri kullanılmış. Modeller, doğruluk, geri çağırma, kesinlik, F1, ve "One-versus-rest" gibi metriklerle değerlendirilmiş. Bir BERT modeli, %61 doğruluk ve "One-versus-rest" metriklerinde 0.8'in üzerinde performans göstermiş. En yüksek kesinlik, geri çağırma ve F1 puanları sırasıyla %50, %39 ve %39 olarak rapor edilmiştir.

# 

# MATERYAL VE METOT

Tablo 1. Örnek Tablo

Şekil 1. Deneme şekil1

# ARAŞTIRMA BULGULARI VE TARTIŞMA

# SONUÇ VE ÖNERİLER

Tablo 4. Örnek Tablo

Şekil 2. Deneme şekil2

# KAYNAKÇA

Prabhakaran, V., Reid, E. E., & Rambow, O. (2014, October). Gender and power: How gender and gender environment affect manifestations of power. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1965-1976).

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. Journal of Sociolinguistics, 18(2), 135-160.

O'Connor, K., Golder, S., Weissenbacher, D., Klein, A. Z., Magge, A., & Gonzalez-Hernandez, G. (2024). Methods and Annotated Data Sets Used to Predict the Gender and Age of Twitter Users: Scoping Review. Journal of Medical Internet Research, 26, e47923.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011, July). Discriminating gender on Twitter. In Proceedings of the 2011 conference on empirical methods in natural language processing (pp. 1301-1309).

Corney, M. W. (2003). Analysing e-mail text authorship for forensic purposes (Doctoral dissertation, Queensland University of Technology).

Yıldırım, S., & Yıldız, T. (2018). Türkçe için karşılaştırmalı metin sınıflandırma analizi. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 24(5), 879-886.

Maken, P., Gupta, A., & Gupta, M. K. (2019). A study on various techniques involved in gender prediction system: a comprehensive review. Cybernetics and Information Technologies, 19(2), 51-73.

Wais, K. (2016). Gender prediction methods based on first names with genderizeR. R J., 8(1), 17.

Suman, C., Chaudhari, R., Saha, S., Kumar, S., & Bhattacharyya, P. (2022). Investigations in emotion aware multimodal gender prediction systems from social media data. IEEE Transactions on Computational Social Systems, 10(2), 470-479.

Wang, Y., Tang, Y., Ma, J., & Qin, Z. (2015). Gender prediction based on data streams of smartphone applications. In Big Data Computing and Communications: First International Conference, BigCom 2015, Taiyuan, China, August 1-3, 2015, Proceedings 1 (pp. 115-125). Springer International Publishing.

Hosseini, M., & Tammimy, Z. (2016). Recognizing users gender in social media using linguistic features. Computers in Human Behavior, 56, 192-197.

Akşit, M. (2020). Büyük veride hiyerarşik kümeleme yöntemlerinin kofenetik korelasyon ile karşılaştırılması (Master's thesis, Fen Bilimleri Enstitüsü).

Vicente, M., Batista, F., & Carvalho, J. P. (2015, August). Twitter gender classification using user unstructured information. In 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-7). IEEE.

Yağcı, M. İ., & İlarslan, N. (2010). Reklamlarin Ve Cinsiyet Kimliği Rolünün Tüketicilerin Satin Alma Davranişlari Üzerindeki Etkisi. Doğuş Üniversitesi Dergisi, 11(1), 138-155.

Alabay, M. N. (2011). Sosyal medyada tüketiciler ve pazar bölümleme uygulamalar.

Vyawahare, M., & Govilkar, S. (2022). Fake profile recognition using profanity and gender identification on online social networks. Social Network Analysis and Mining, 12(1), 170.

Onikoyi, B., Nnamoko, N., & Korkontzelos, I. (2023). Gender prediction with descriptive textual data using a machine learning approach. Natural Language Processing Journal, 4, 100018.

Vicente, M., Batista, F., & Carvalho, J. P. (2015, August). Twitter gender classification using user unstructured information. In 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-7). IEEE.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. Journal of Sociolinguistics, 18(2), 135-160.

Alowibdi, J. S., Buy, U. A., & Yu, P. (2013, December). Empirical evaluation of profile characteristics for gender classification on twitter. In 2013 12th international conference on machine learning and applications (Vol. 1, pp. 365-369). IEEE.

Miller, Z., Dickinson, B., & Hu, W. (2012). Gender prediction on twitter using stream algorithms with n-gram character features.

Rego, R.C., Silva, V.M., & Fernandes, V.M. (2021). Predicting Gender by First Name Using Character-level Machine Learning.

To, H.Q., Nguyen, K.V., Nguyen, N.L., & Nguyen, A.G. (2020). Gender Prediction Based on Vietnamese Names with Machine Learning Techniques. Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval.

Basile, A., Dwyer, G., & Rubagotti, C. (2019). Simple n-gram based models perform well for gender prediction . Sometimes.

Vanmassenhove, E., Moryossef, A., Poncelas, A., Way, A., & Sht, D. (2019). ABI Neural Ensemble Model for Gender Prediction. In GxG@ CLIN (pp. 53-61).

Tripathi, A., & Faruqui, M. (2011). Gender prediction of Indian names. IEEE Technology Students' Symposium, 137-141.

Prabhakaran, V., Reid, E. E., & Rambow, O. (2014, October). Gender and power: How gender and gender environment affect manifestations of power. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1965-1976).

Liu, J. (2023). Sentiment Classification of Social Network Text Based on AT-BiLSTM Model in a Big Data Environment. Int. J. Inf. Technol. Syst. Approach, 16, 1-15.

Yu, Q., Wang, Z., & Jiang, K. (2021). Research on Text Classification Based on BERT-BiGRU Model. Journal of Physics: Conference Series, 1746.

Bensahla, A., Zaghir, J., Gaudet-Blavignac, C., & Lovis, C. (2024). Unsupervised Extraction of Body-Text from Clinical PDF Documents. In Digital Health and Informatics Innovations for Sustainable Health Care Systems (pp. 214-215). IOS Press.

Urrutia, R., Espejo, D., Evens, N., Guerra, M., Sühn, T., Boese, A., ... & Poblete, V. (2023). Clustering Methods for Vibro-Acoustic Sensing Features as a Potential Approach to Tissue Characterisation in Robot-Assisted Interventions. Sensors, 23(23), 9297.

Taamneh, M., Taamneh, S., & Alkheder, S. (2017). Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks. International journal of injury control and safety promotion, 24(3), 388-395.

Davagdorj, K., Wang, L., Li, M., Pham, V. H., Ryu, K. H., & Theera-Umpon, N. (2022). Discovering thematically coherent biomedical documents using contextualized bidirectional encoder representations from transformers-based clustering. International Journal of Environmental Research and Public Health, 19(10), 5893.

Woo, H., Kim, K., Cha, K., Lee, J. Y., Mun, H., Cho, S. J., ... & Kang, M. (2019). Application of efficient data cleaning using text clustering for semistructured medical reports to large-scale stool examination reports: methodology study. Journal of medical Internet research, 21(1), e10013.

Wang, Y., Wang, Y., Peng, Z., Zhang, F., Zhou, L., & Yang, F. (2023). Medical text classification based on the discriminative pre-training model and prompt-tuning. Digital Health, 9, 20552076231193213.

Yetisgen-Yildiz, M., & Pratt, W. (2005). The effect of feature representation on MEDLINE document classification. In AMIA annual symposium proceedings (Vol. 2005, p. 849). American Medical Informatics Association.

Tan, Z., Lin, J., Chen, K., Zhuang, Y., & Han, L. (2023). Skin cancer image recognition based on similarity clustering and attention transfer. Journal of X-Ray Science and Technology, 31(2), 337-355.

Ahne, A., Fagherazzi, G., Tannier, X., Czernichow, T., & Orchard, F. (2022). Improving diabetes-related biomedical literature exploration in the clinical decision-making process via interactive classification and topic discovery: Methodology development study. Journal of medical Internet research, 24(1), e27434.

Workman, T. E., Weir, C., & Rindflesch, T. C. (2017, February). Differentiating sense through semantic interaction data. In AMIA Annual Symposium Proceedings (Vol. 2016, p. 1238).

Dipnall, J. F., Lu, J., Gabbe, B. J., Cosic, F., Edwards, E., Page, R., & Du, L. (2022). Comparison of state-of-the-art machine and deep learning algorithms to classify proximal humeral fractures using radiology text. European journal of radiology, 153, 110366.

# ÖZGEÇMİŞ