BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ TEZLİ YÜKSEK LİSANS PROGRAMI

METİN SINIFLANDIRILMASINDA ÖZETLEME VE YENİDEN İFADE ETME TEKNİKLERİNİN DEĞERLENDİRİLMESİ

HAZIRLAYAN

CEZMİ ERDÖR

YÜKSEK LİSANS TEZİ

ANKARA - 2025

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ TEZLİ YÜKSEK LİSANS PROGRAMI

METİN SINIFLANDIRILMASINDA ÖZETLEME VE YENİDEN İFADE ETME TEKNİKLERİNİN DEĞERLENDİRİLMESİ

HAZIRLAYAN

CEZMİ ERDÖR

YÜKSEK LİSANS TEZİ

TEZ DANIŞMANI

DR. ÖĞR. ÜYESİ DİDEM ÖLÇER

ANKARA - 2025

BAŞKENT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı çerçevesinde Cezmi Erdör tarafından hazırlanan bu çalışma, aşağıdaki jüri tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 15/08/2025

Tez Adı: Metin Sınıflandırılmasında Özetleme ve Yeniden İfade Etme Tekniklerinin Değerlendirilmesi

Tez Jüri Üyeleri (Unvanı, Adı - Soyadı, Kurumu)		**İmza**
Dr. Öğr. Üyesi Didem ÖLÇER	Başkent Üniversitesi	……...
Dr. Öğr. Üyesi Çağatay Berke ERDAŞ	Başkent Üniversitesi	……...
Dr. Öğr. Üyesi Ahmet Anıl MÜNGEN	Ostim Teknik Üniversitesi	……...

ONAY

Prof. Dr. Dilek ÇÖKELİLER SERDAROĞLU

Fen Bilimleri Enstitüsü Müdürü

Tarih: … / … / ……..

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
YÜKSEK LİSANS TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

Tarih: 12 / 08 /2025

Öğrencinin Adı, Soyadı: Cezmi Erdör

Öğrencinin Numarası: 22120279

Anabilim Dalı: Bilgisayar Mühendisliği

Programı: Tezli Yüksek Lisans

Danışmanın Unvanı/Adı, Soyadı: Dr. Öğr. Üyesi Didem Ölçer

Tez Başlığı: Metin Sınıflandırılmasında Özetleme ve Yeniden İfade Etme Tekniklerinin Değerlendirilmesi

Yukarıda başlığı belirtilen Yüksek Lisans/Doktora tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 54 sayfalık kısmına ilişkin, 12 / 08 /2025 tarihinde şahsım/tez danışmanım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 5 'tir. Uygulanan filtrelemeler:

1. Kaynakça hariç
2. Alıntılar hariç
3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları hariç

"Başkent Üniversitesi Enstitüleri Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Usul ve Esaslarını" inceledim ve bu uygulama esaslarında belirtilen azami benzerlik oranlarına tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Öğrenci İmzası: …….

ONAY

Öğrenci Danışmanı :

Dr. Öğr. Üyesi Didem ÖLÇER

Tarih: 15/08/2025

…….

TEŞEKKÜR

Bu tez çalışmasının her aşamasında bana rehberlik eden, bilgi ve deneyimleriyle katkı sağlayan kıymetli tez danışmanım Dr. Öğr. Üyesi Didem Ölçer'e en içten teşekkürlerimi sunarım. Ayrıca bu süreçte maddi ve manevi desteklerini hiçbir zaman esirgemeyen, bana her zaman güç veren annem Gülsüm Nazan ERDÖR'e ve babam Mehmet Hüsnü ERDÖR'e minnet ve teşekkür ederim.

ÖZET

Cezmi ERDÖR

METİN SINIFLANDIRILMASINDA ÖZETLEME VE YENİDEN İFADE ETME TEKNİKLERİNİN DEĞERLENDİRİLMESİ

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

2025

Bu tez çalışmasında, haber metinleri üzerinde gerçekleştirilen özetleme, yeniden ifade etme (parafrazlama) ve semantik benzerlik ölçme yöntemlerinin performansları kapsamlı biçimde incelenmiştir. Çalışma kapsamında öncelikle metinler, ön işleme adımlarından geçirilmiş ve Bag of Words (BoW) ile temsil edilmiştir. Ardından, farklı makine öğrenmesi modelleri (Lojistik Regresyon, Destek Vektör Makineleri, Rastgele Orman ve Tek Katmanlı Yapay Sinir Ağı üzerinde katmanlı k-kat çapraz doğrulama yöntemiyle sınıflandırma deneyleri yapılmış; en yüksek başarımı gösteren model sonraki aşamalarda kullanılmıştır. Özetleme aşamasında yalnızca çıkarımsal(extractive) yöntemler tercih edilmiş; TextRank, LexRank, TF-IDF tabanlı özetleme ve kural tabanlı (rule-based) yöntemler uygulanmıştır. Yeniden ifade etme aşamasında ise iki farklı yöntem—WordNet tabanlı ve Back-Translation—kullanılmıştır. Üretilen özetler ve yeniden ifade edilen metinler hem kosinüs benzerliği hem de BERTScore ölçütleriyle değerlendirilmiştir. Analizler, ham metinler ile özetlenmiş metinlerin; ham metinler ile özetlenip yeniden ifade edilmiş metinlerin ve ham metinler ile doğrudan yeniden ifade edilmiş metinlerin karşılaştırılması esas alınarak yürütülmüştür. Bulgular, bazı yöntemlerde özetlenip yeniden ifade edilmiş metinlerin semantik benzerlik skorlarının yalnızca özetlenmiş metinlerden yüksek olabildiğini, ancak genel olarak doğrudan yeniden ifade edilen metinlerin en yüksek benzerlik değerlerine ulaştığını ortaya koymuştur. Bu çalışma, özetleme ve yeniden ifade etme yöntemlerinin birlikte kullanıldığında semantik benzerlik üzerindeki etkilerini ortaya koymakta ve metin işleme uygulamalarında yöntem seçiminde yol gösterici bulgular sunmaktadır.

ANAHTAR KELİMELER: Metin özetleme, Yeniden ifade etme, Semantik benzerlik, Makine öğrenmesi, Doğal dil işleme.

ABSTRACT

Cezmi ERDÖR

EVALUATION OF SUMMARISING AND PARAPHRASING TECHNIQUES IN TEXT CLASSIFICATION

Başkent University Institute of Science

Department of Computer Engineering

2025

In this thesis, the performances of summarization, paraphrasing, and semantic similarity measurement methods applied to news articles are comprehensively examined. Initially, the texts were preprocessed and represented using the Bag of Words (BoW) model. Subsequently, classification experiments were conducted on different machine learning models (Logistic Regression, Support Vector Machines, Random Forest, and Perceptron) using the stratified k-fold cross-validation method; the model with the highest performance was employed in the subsequent stages. In the summarization stage, only extractive methods were preferred, and TextRank, LexRank, TF-IDF-based summarization, and rule-based methods were applied. In the paraphrasing stage, two different methods—WordNet-based and Back-Translation—were used. The generated summaries and paraphrased texts were evaluated using both the Cosine Similarity and BERTScore metrics. The analyses were carried out by comparing raw texts with summarized texts; raw texts with summarised and paraphrased texts; and raw texts with directly paraphrased texts. The findings revealed that, in some cases, the semantic similarity scores of summarised-and-paraphrased texts could be higher than those of solely summarized texts; however, in general, directly paraphrased texts achieved the highest similarity scores. This study demonstrates the effects of using summarization and paraphrasing methods together on semantic similarity and provides guiding insights for method selection in text processing applications.

KEYWORDS : Text summarization, Paraphrasing, Semantic similarity, Machine learning, Natural language processing.

İÇİNDEKİLER

TABLOLAR LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR

BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
BoW	Bag of Words
Kosinüs Benzerliği	İki vektör arasındaki açıyı temel alan benzerlik ölçütü
F1 Skoru	Kesinlik ve Duyarlılık değerlerinin harmonik ortalaması
MLP	Multi-Layer Perceptron (Çok Katmanlı Algılayıcı)
NLP	Natural Language Processing (Doğal Dil İşleme)
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SVM	Support Vector Machine (Destek Vektör Makineleri)
TF-IDF	Term Frequency - Inverse Document Frequency

1. GİRİŞ

Günümüzde dijitalleşmenin hız kazanmasıyla birlikte metin tabanlı veri miktarı olağanüstü bir şekilde artmıştır. Haber siteleri, sosyal medya platformları, çevrim içi içerik üreticileri ve bloglar gibi kaynaklardan akan büyük hacimli metin verilerinin işlenmesi, özetlenmesi ve yorumlanması, doğal dil işleme (Natural Language Processing- NLP) alanında kritik öneme sahip hale gelmiştir. Bu bağlamda, metin özetleme ve yeniden ifade etme (paraphrasing) gibi metin yoğunluklu görevler, bilgiye hızlı ve anlamlı erişimin sağlanmasında temel araçlar arasında yer almaktadır.

Metin özetleme, uzun metinlerin daha kısa ve öz ifadelerle yeniden yazılması sürecini ifade ederken; yeniden ifade etme, bir ifadenin anlamını koruyarak farklı biçimlerde yeniden ifade edilmesini amaçlamaktadır. Bu iki işlem, doğal dilin semantik boyutuna odaklandığı için bilgi kaybı, anlam bozulması veya dilsel bütünlük kayıpları gibi riskleri de beraberinde getirir. Özellikle özetlenmiş metinlerin yeniden ifade edilmesi durumunda hem içerik hem de biçimsel anlamda bütünlüğün korunması daha karmaşık hale gelmektedir.

Bu tez çalışması, ham metinlerden yola çıkarak elde edilen özetlerin farklı yeniden ifade etme teknikleriyle yeniden ifade edilmesi durumunda, anlamın ne ölçüde korunduğunu ortaya koymayı hedeflemektedir. Bununla birlikte, ham metinlerin doğrudan yeniden ifade edilmesi ile özetlenmiş halinin yeniden ifade edilmesi durumları karşılaştırılarak, hangi yöntemin daha yüksek semantik benzerlik sağladığı değerlendirilmektedir. Çalışmada, metin özetleme için TextRank, LexRank, TF-IDF ve kural tabanlı yaklaşımlar; yeniden ifade etme için WordNet tabanlı, çeviri temelli ve BERT tabanlı yaklaşımlar kullanılmıştır. Anlamsal benzerlik değerlendirmeleri ise kosinüs benzerliği ve BERTScore metrikleriyle yapılmıştır.

1.1. Problem Tanımı ve Hedefler

Metin özetleme ve yeniden ifade etme teknikleri doğal dil işleme uygulamalarında sıklıkla kullanılmaktadır. Ancak bu iki sürecin birleştirilerek birlikte değerlendirilmesi nadiren ele alınmıştır. Özellikle bir metnin özetlendikten sonra yeniden ifade edilmesi sürecinde, orijinal anlamın ne ölçüde korunabildiği sorusu, bu çalışmanın temel problemini oluşturmaktadır. Ayrıca,

ham metnin doğrudan yeniden ifade edilmesi ile özetlenmiş halinin yeniden ifade edilmesi arasında anlam benzerliği bakımından ne tür farklar bulunduğu da araştırma kapsamına alınmıştır.

Bu tez çalışmasıyla hedeflenen temel amaçlar şunlardır:

- Ham metinlerin farklı özetleme algoritmalarıyla özetlenmesi,
- Bu özetlerin çeşitli yeniden ifade etme teknikleriyle yeniden ifade edilmesi,
- Özetlerin anlam kaybına uğrayıp uğramadığının metriklerle değerlendirilmesi,
- Ham metnin doğrudan yeniden ifade edilmesi ile özetleme + yeniden ifade etme sonuçlarının karşılaştırılması,
- Doğal dil işleme uygulamaları için daha anlam koruyucu yöntemlerin belirlenmesi.

1.2. Tez Yapısının Özeti

Bu tez çalışması toplam 10 ana bölümden oluşmaktadır. İkinci bölümde, metin özetleme, yeniden ifade etme ve semantik benzerlik gibi kavramlarla ilgili literatürde yer alan çalışmalar sunulmuştur. Üçüncü bölümde kullanılan veri seti tanıtılmış ve veri ön işleme süreci açıklanmıştır. Dördüncü bölümde metinlerden öznitelik çıkarımı için kullanılan yöntemler ele alınmış, beşinci bölümde ise sınıflandırma algoritmaları açıklanmıştır. Altıncı bölümde farklı özetleme algoritmalarının detayları, yedinci bölümde yeniden ifade etme tekniklerinin çalışma prensipleri yer almaktadır. Sekizinci bölümde semantik benzerlik ölçüm teknikleri açıklanırken, dokuzuncu bölümde tüm yöntemlerin karşılaştırılması yapılmıştır. Onuncu bölümde bulgular tartışılmış ve son bölümde genel sonuçlar ile gelecek çalışmalar için öneriler sunulmuştur.

2. LİTERATÜR

Doğal dil işleme (NLP) alanında artan veri hacmi, bilgiye hızlı erişimi mümkün kılacak yöntemlerin geliştirilmesini zorunlu hale getirmiştir. Bu bağlamda metin özetleme, uzun metinlerden anlamlı ve bilgilendirici kısa içerikler üretmeyi hedefleyen önemli bir görev olarak öne çıkmaktadır. Son yıllarda, özellikle derin öğrenme modellerinin gelişimiyle birlikte hem çıkarımsal (extractive) hem de soyutlayıcı (abstractive) özetleme yöntemlerinde önemli ilerlemeler kaydedilmiştir [1].

Geleneksel çıkarımsal yöntemler, metindeki önemli cümleleri seçerek özet oluştururken; modern yöntemler artık kelime düzeyinde değil, anlam düzeyinde analiz yaparak içerik üretmektedir. Transformer mimarileri bu dönüşümde kilit rol oynamıştır. Örneğin, BERT tabanlı MatchSum modeli, kaynak metin ile aday cümleler arasında bağlamsal benzerlikleri karşılaştırarak özetleme yapmaktadır [2]. Benzer şekilde, TLM (Topic-aware Latent Model) ve BARTSUM gibi modeller, metnin semantik yapısını daha etkili analiz ederek çıkarımsal özet kalitesini artırmayı başarmıştır [3].

Soyutlayıcı özetleme alanında ise BART ve PEGASUS gibi transformer tabanlı modeller, insan tarafından üretilmiş özetlere yakın kalite sunabilmektedir. BART, encoder-decoder mimarisine sahip olup, maskelenmiş metinleri yeniden inşa ederek anlam koruyucu cümleler üretmektedir [4]. PEGASUS ise özel olarak özetleme için tasarlanmış bir pretraining stratejisi kullanmakta, “boşluk doldurma” (gap-sentence generation) üzerinden eğitim alarak kritik bilgi taşıyan ifadeleri tahmin etme becerisi kazanmaktadır [5].

Öte yandan, çıkarımsal ve soyutlayıcı tekniklerin hibrit yaklaşımlarla birleştirildiği modeller de geliştirilmektedir. T5 (Text-to-Text Transfer Transformer) modeli, tüm NLP görevlerini metin dönüştürme problemi olarak ele alarak, özeti doğrudan metinsel bir çıktı olarak üretmektedir [6]. T5, yeniden ifade etme, özetleme, soru-cevap gibi çok sayıda görevi aynı mimariyle gerçekleştirebilmesiyle dikkat çekmektedir.

Güncel araştırmalar, özetleme kalitesini yalnızca ROUGE skorlarıyla değil, aynı zamanda BERTScore, QAEval, BLEURT gibi bağlamsal ölçütlerle de değerlendirmeye yönelmiştir [7]. Bu sayede, sadece sözcük benzerliği değil, anlam tutarlılığı da dikkate alınmaktadır.

Sonuç olarak, metin özetleme sistemleri yüzeysel yapılardan bağlamsal ve anlam odaklı modellere doğru evrilmiştir. Bu evrim, özellikle haber özetleme, tıbbi belge özetleme, yasal metin özeti gibi alanlarda yüksek kaliteli çıktılar üretmeyi mümkün kılmaktadır. Bu tez kapsamında, çıkarımsal özetleme teknikleri tercih edilerek, özetlenmiş metinlerin daha sonra yeniden ifade edilmesi ve semantik açıdan değerlendirilmesi amaçlanmaktadır.

Metin işleme uygulamalarında iki ifadenin benzerliğini ölçmek, doğal dilin anlam boyutuna dair önemli çıkarımlar yapılmasını sağlar. Bu amaçla geliştirilen semantik benzerlik ölçütleri, özellikle metin özetleme, yeniden ifade etme ve makine çevirisi gibi görevlerin değerlendirilmesinde yaygın biçimde kullanılmaktadır. Literatürde bu tür benzerlik ölçümleri vektör tabanlı klasik yöntemlerden, bağlamsal derin öğrenme tabanlı modellere kadar geniş bir yelpazeye yayılmaktadır.

Kosinüs benzerliği, geleneksel olarak en sık başvurulan yöntemlerden biridir. Bu yaklaşımda metinler, genellikle TF-IDF veya Bag of Words (BoW) temelli vektörler aracılığıyla sayısallaştırılır. Ardından iki vektör arasındaki açı ölçülerek benzerlik skoru elde edilir. Skor 1'e yaklaştıkça cümleler benzer, 0'a yaklaştıkça farklı kabul edilir. Bu yöntemin basit ve yorumlanabilir olması önemli bir avantaj olmakla birlikte, bağlam farkındalığının sınırlı olması nedeniyle semantik yakınlık yerine kelime örtüşmesine dayalı değerlendirme yapar [8].

Son yıllarda bağlamsal bilgi işleme yetenekleri gelişen modellerin ortaya çıkmasıyla birlikte embedding tabanlı benzerlik ölçütleri öne çıkmıştır. Bu yöntemlerde cümleler, önceden eğitilmiş dil modelleri yardımıyla yüksek boyutlu vektör uzaylarına gömülür. Bu bağlamda geliştirilen BERTScore, her kelimeyi bağlamsal olarak kodlayarak referans ve aday cümleler arasındaki benzerliği çok daha hassas şekilde ölçebilmektedir [5]. BERTScore, yalnızca kelime örtüşmesine değil, aynı zamanda anlam eşdeğerliğine odaklanarak daha gerçekçi bir benzerlik analizi sunar.

Cümle düzeyindeki karşılaştırmalarda ise Sentence-BERT (SBERT) önemli bir çözüm sunar. Reimers ve Gurevych tarafından geliştirilen bu model, BERT mimarisi üzerine çift yönlü

(Siamese) yapılar ekleyerek, cümleler arası benzerliği ölçen bir embedding üretim süreci oluşturmuştur [9]. SBERT, klasik BERT'in aksine, cümle karşılaştırma görevlerinde hesaplama maliyetini azaltırken doğruluğu da artırmaktadır. Bu sayede, yeniden ifade etme, bilgi erişimi ve anlamsal eşleştirme gibi görevlerde yüksek başarı göstermektedir.

Özellikle soyutlayıcı özetleme ve yaratıcı dil üretimi gibi bağlamda çoklu doğru cevabın olabileceği görevlerde, BLEURT gibi insan yargılarını modellemeyi hedefleyen yeni nesil ölçütler kullanılmaktadır [10]. BLEURT, BERT üzerine ince ayar (fine-tuning) yapılarak geliştirilmiş bir değerlendirme sistemidir ve insan değerlendirmeleriyle güçlü korelasyon göstermektedir. Benzer şekilde, QAEval yöntemi de metinlerin içerdiği bilgiyi ölçmek için soru-yanıt eşleştirme temelli değerlendirme yaklaşımı sunmaktadır [11].

Güncel literatür, klasik benzerlik metriklerinin özellikle kısa ve bağlam bağımlı cümlelerde yetersiz kaldığını; BERTScore ve SBERT gibi bağlamsal yöntemlerin daha tutarlı sonuçlar sunduğunu ortaya koymaktadır. Bu tez çalışmasında, hem kosinüs benzerliği gibi geleneksel vektör tabanlı hem de BERTScore gibi bağlamsal yöntemler kullanılarak, yeniden ifade etme ve özetleme kalitesine dair çok yönlü analizler gerçekleştirilmiştir.

3. MATERYAL METOT

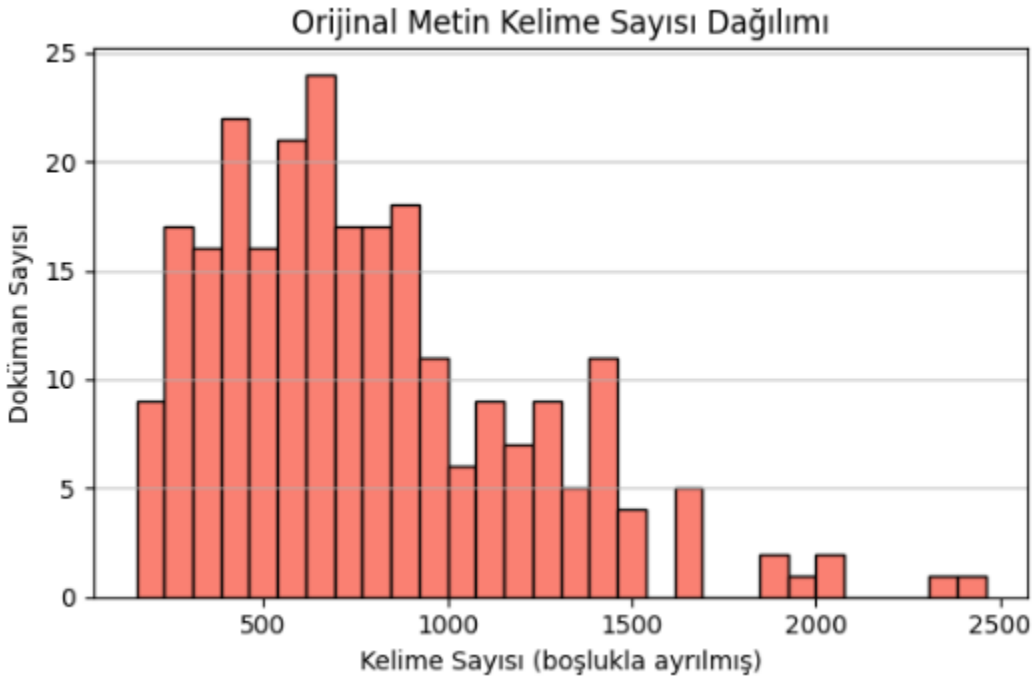
3.1. Veri Setinin Tanıtımı

Bu çalışmada kullanılan veri seti, çok kategorili ve çeşitli alanlara yayılmış haber metinlerinden oluşmaktadır. Veri kümesi; sağlık, iş, spor, teknoloji, eğlence, seyahat ve eğitim kategorilerinde sınıflandırılmış toplam 252 adet İngilizce haber metnini içermektedir. Metinler, New York Times [54], Reuters [55], ABC [56], BBC [57], CNN [58] ve The Guardian [59] gibi uluslararası güvenilir haber kaynaklarından manuel olarak toplanmıştır. Aşağıda veri kümesinin kategori bazlı dağılımı verilmiştir:

Tablo 3.1. Veri Kümesinin Kategori Bazlı Dağılımı

Kategori	**Etiket**	**Doküman Sayısı**
Sağlık	1	30
İş	2	44
Spor	3	47
Teknoloji	4	36
Eğlence	5	37
Seyahat	6	28
Eğitim	7	30
Toplam	—	**252**

Veri kümesinin ham metinlerine ilişkin uzunluk analizi yapılmış ve her dokümandaki kelime sayısı dağılımı aşağıdaki şekilde görselleştirilmiştir.



Şekil 3.1. Orijinal metinlerdeki kelime sayısı dağılım histogramı

Dağılım Özellikleri:

- X ekseni, belgelerin uzunluğunu (kelime sayısı); Y ekseni ise bu uzunluktaki belge sayısını göstermektedir.
- Belirgin yoğunluk aralığı 400 ila 900 kelime arasında yoğunlaşmakta olup bu aralıkta her binde 15–25 belge yer almaktadır.
- Ortalama kelime sayısının yaklaşık 800–850 civarında olduğu, medyanın ise 700–750 aralığında seyrettiği gözlemlenmiştir.
- Dağılım sağa çarpık (right-skewed) bir yapı göstermektedir; yani uzun metinler daha seyrek görülmektedir.

- Örneğin, 1500 kelimenin üzerindeki metinler oldukça az sayıda olup; 2200–2500 kelimeye kadar ulaşan sadece birkaç örnek mevcuttur.

- Bu durum, veri setindeki metinlerin büyük çoğunluğunun orta uzunlukta olduğunu, yalnızca az sayıda metnin uç değerlere sahip olduğunu ortaya koymaktadır.

Bu betimsel analiz, özetleme ve yeniden ifade etme algoritmalarının uygulanacağı metinlerin uzunluk dağılımı hakkında ön bilgi sunmakta ve özellikle özetleme çıktılarının değerlendirilmesinde önemli bir temel teşkil etmektedir.

3.2. Veri Ön İşleme

Bu çalışmada kullanılan metinlerin sağlıklı şekilde işlenebilmesi için çeşitli ön işleme adımları uygulanmıştır. Bu işlemler sayesinde metinler hem daha temiz hem de daha anlamlı hale getirilmiş ve modelleme sürecine hazır hâle getirilmiştir. Uygulanan ön işleme adımları aşağıda alt başlıklar hâlinde sunulmuştur.

3.2.1. Küçük Harfe Dönüştürme (Lowercasing)

Tüm metinler küçük harfe çevrilerek, büyük-küçük harf farkından kaynaklı anlamsız ayrımlardan kaçınılmıştır. Bu işlem sayesinde “Cat” ve “cat” gibi kelimeler aynı terim olarak değerlendirilmiştir.

- Örnek (öncesi): "Companion animals, and especially Cats, are..."
- Örnek (sonrası): "companion animals, and especially cats, are..."

3.2.2. Noktalama ve Özel Karakter Temizliği

Noktalama işaretleri ve özel karakterler metinden kaldırılmıştır. Bu sayede kelime temsilleri sadeleştirilmiş ve model için anlamlı olmayan işaretlerden arındırılmıştır.

- Örnek (öncesi): "H5N1 -- generally barn cats that lived on dairy farms..."
- Örnek (sonrası): "h5n1 generally barn cats that lived on dairy farms"

3.2.3. Tokenizasyon

Temizlenmiş metinler, boşluk karakterine göre kelimelere (tokenlara) ayrılmıştır. Bu işlem, metni model tarafından işlenebilir hâle getirmiştir.

- **Örnek:**
    - Girdi: "bird flu currently circulating has not adapted"
    - Tokenlar: ["bird", "flu", "currently", "circulating", "has", "not", "adapted"]

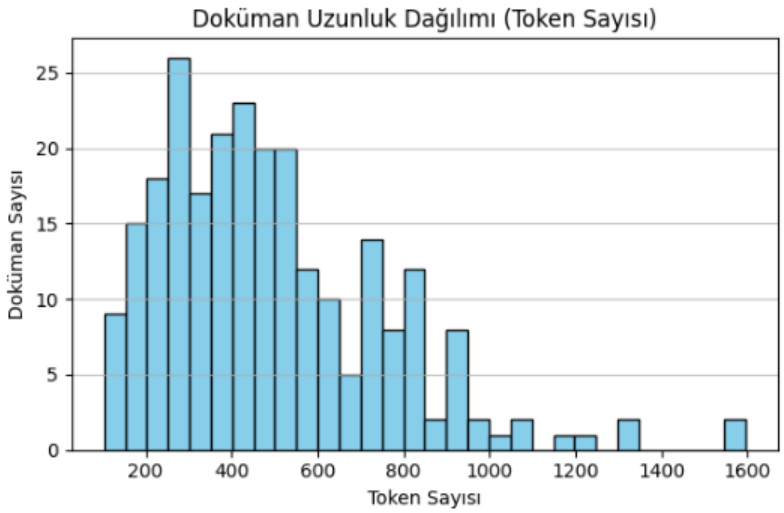
3.2.4. Stopword Temizliği

İngilizce'de çok sık kullanılan ancak anlam açısından katkısı olmayan sözcükler (ör. "the", "is", "in", "of" vb.) çıkarılmıştır. Bu adım, modelin sadece bilgi içeren kelimelere odaklanmasını sağlamıştır.

- **Örnek (öncesi):** "This is because we snuggle with and sleep in bed with our cats."
- **Örnek (sonrası):** "snuggle sleep bed cats"

3.2.5. Token Dağılım Analizi

Veri ön işleme adımları uygulandıktan sonra her belgedeki kelime (token) sayısı hesaplanmış ve dağılımı görselleştirilmiştir.

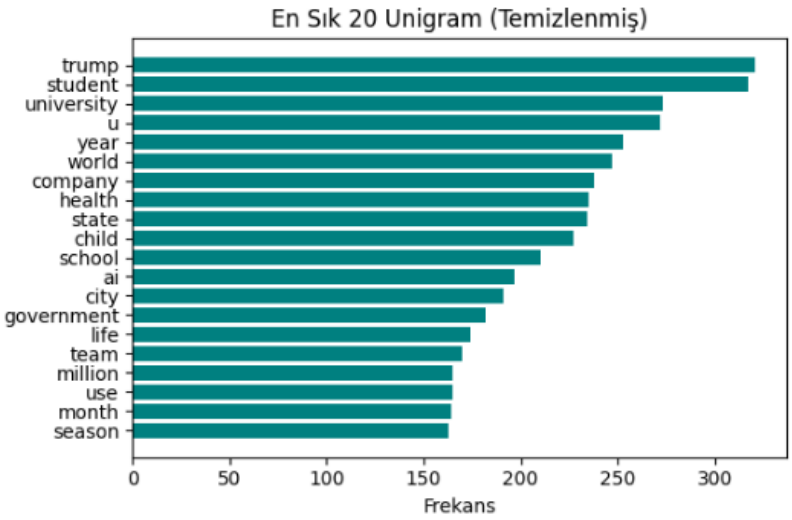


Şekil 3.2. Ön işleme sonrası dokümanlardaki kelime (token) sayısı dağılımı

- **250–500 kelime arası** belgeler en yoğun grubu oluşturur.
- **800 üzeri kelime içeren belgeler** nadir olup, dağılım sağa çarpık (skewed) yapıdadır.

3.2.6. En Sık Geçen Unigram Analizi

Her dokümanda geçen token'ların frekansı hesaplanarak, en sık kullanılan 20 kelime aşağıdaki grafikte sunulmuştur.

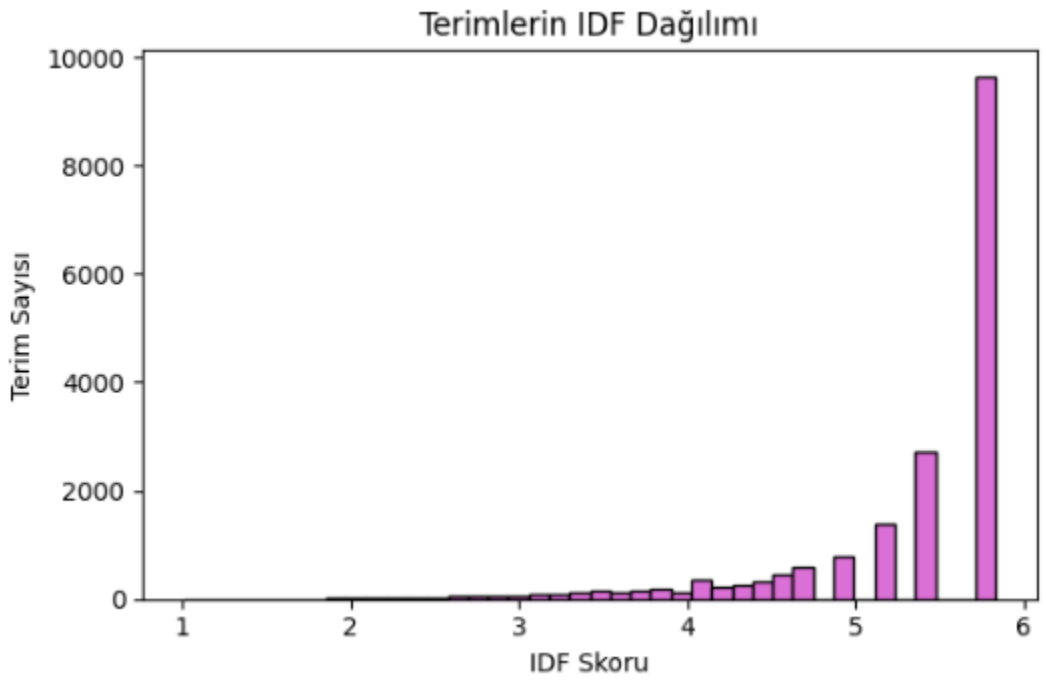


Şekil 3.3. En sık geçen 20 kelimenin (unigram) sıklık grafiği

Grafik, veri setinde öne çıkan kavramlar hakkında fikir verir. Özellikle “trump”, “student”, “health”, “ai” gibi kelimeler; eğitim, siyaset, sağlık ve teknoloji kategorilerine dair ipuçları taşımaktadır.

3.2.7. IDF Skor Dağılımı

Metinlerde geçen terimlerin ayırt edicilik gücünü görmek amacıyla IDF (Inverse Document Frequency) skorları hesaplanmış ve dağılım grafiği oluşturulmuştur.



Şekil 3.4. Terimlerin IDF skor dağılımı

IDF formülü (3.1) aşağıdaki gibi tanımlanmıştır:

$$IDF(t) = log\left(\frac{N+1}{df(t)+1}\right) + 1 \quad (3.1)$$

Bu formülde kullanılan değişkenler şu şekildedir:

- N: Veri setindeki toplam belge sayısıdır.
- df(t): Terim *t*'nin yer aldığı belge sayısıdır.

Formülde, df(t) değeri büyüdükçe, yani ilgili terim daha çok belgede geçtikçe, kesir küçülür ve sonuçta logaritma değeri daha düşük bir IDF skoru elde edilir. Bu durum, o terimin belirleyici gücünün azaldığını ve daha genel bir ifade olduğunu gösterir.

Şekil 3.4'te gösterilen grafik, hangi terimlerin yaygın (genel) hangilerinin seyrek (özel) olduğunu ortaya koymaktadır. Dağılım aşağıdaki şekilde yorumlanabilir:

- Düşük IDF (yaklaşık 1-2): Grafik neredeyse sıfıra yakın başlamaktadır, bu da çok az sayıda terimin tüm belgelerde veya büyük bir kısmında geçtiğini göstermektedir. Örneğin, "trump" ve "student" gibi kelimeler bu gruba dahildir.
- Orta IDF (yaklaşık 3-5): Bu aralıkta yüzlerce terim yer almaktadır. Eğitim, sağlık, siyaset gibi alanlara özgü ancak yaygın olmayan terimler bu bölgede toplanmaktadır.
- Yüksek IDF (yaklaşık 5-6): Dağılımın en yoğun olduğu bölgedir. Bu kısımda, yalnızca birkaç belgede geçen, dolayısıyla yüksek ayırt edicilik taşıyan on binlerce terim bulunmaktadır. Bunlar çoğunlukla nadir geçen, kontekst-spesifik kelimelerdir.

4. ÖZNİTELİK ÇIKARIMI

Makine öğrenmesi modellerinde metinsel verilerin sayısal olarak ifade edilebilmesi için öznitelik çıkarımı büyük önem taşır. Bu tez kapsamında, iki temel vektörleştirme yöntemi kullanılmıştır: Bag of Words (BoW) ve Term Frequency – Inverse Document Frequency (TF-IDF).

4.1. Bag of Words (BoW)

BoW yöntemi, her belgedeki kelime sıklıklarını dikkate alarak sabit uzunlukta vektörler oluşturur. Bu temsilde kelimenin metin içindeki konumu, gramer yapısı ya da anlam bağlamı dikkate alınmaz [12]. Yöntem basitliği nedeniyle çok sayıda NLP uygulamasında başarıyla kullanılmaktadır [13].

Bu çalışmada, önişleme adımları tamamlanmış metinler, Python'da CountVectorizer kullanılarak unigram düzeyinde BoW vektörlerine dönüştürülmüştür. Her vektör, kelime dağarcığında (vocabulary) yer alan terimlerin belge içinde kaç kez geçtiğini sayar. Bu temsiller, makine öğrenmesi modellerine girdi olarak sunulmuştur.

BoW yöntemi, nadir ama anlamlı kelimeleri, yaygın ancak bilgi taşımayan kelimelerle aynı şekilde ele aldığı için bazen ayırt edici gücü düşürebilir [14].

4.2. TF-IDF Ağırlıklı Temsil

TF-IDF (Term Frequency–Inverse Document Frequency), metin verilerinden anlamlı öznitelikler çıkarma amacıyla yaygın olarak kullanılan bir terim ağırlıklandırma yöntemidir. Bu yöntem, bir kelimenin bir belge içindeki önemini ölçmek için hem belge içi frekansını (TF) hem de tüm belgeler arasındaki yaygınlığını (IDF) dikkate alır.

TF-IDF temsili, daha önce Bölüm 3.2.8'de tanımlanan formül üzerinden hesaplanmıştır. Bu temsilde her kelimeye, belgedeki sıklığı ve tüm belge koleksiyonundaki yaygınlığı dikkate alınarak ağırlık atanmıştır.

Yöntem, metin sınıflandırma, bilgi geri getirimi ve doğal dil işleme görevlerinde sıkça kullanılan etkili bir araçtır. TF-IDF ile her kelimenin belgeler üzerindeki göreli önemi hesaplanır. Böylece çok yaygın ancak bilgi taşımayan kelimelerin etkisi azaltılırken, nadir fakat anlamlı kelimelere daha fazla ağırlık verilir [15].

Bu tez çalışmasında TfidfVectorizer sınıfı kullanılarak TF-IDF matrisleri oluşturulmuş ve bu matrisler sınıflandırma modellerine girdi olarak sunulmuştur. Ayrıca TF-IDF'in belge frekansının tersi olan IDF bileşeni, stop word'lerin baskın etkisini ortadan kaldırmakta etkili olmuştur.

TF-IDF yönteminin sınıflandırma dışında özetleme [16] ve anlamsal benzerlik hesaplamalarında [9] da yaygın olarak kullanıldığı bilinmektedir.

5. MAKİNE ÖĞRENMESİ MODELLERİ

Doğal dil işleme (NLP) problemlerinde ham metinlerin istatistiksel modellere uygun sayısal temsillere dönüştürülmesinin ardından, bu temsiller üzerinden çeşitli makine öğrenmesi (ML) algoritmaları ile sınıflandırma modelleri oluşturulabilir. Bu tez çalışmasında da öznitelik çıkarımı ve ön işleme adımlarıyla hazır hale getirilen haber metinleri, çeşitli makine öğrenmesi modellerine girdi olarak sunulmuş ve sınıflandırma başarımları değerlendirilmiştir.

Makine öğrenmesi; verilerden örüntüleri öğrenerek, gelecekteki gözlemler hakkında çıkarımda bulunmayı amaçlayan bir yapay zeka alt alanıdır. Bilgisayarların açık bir şekilde programlanmadan öğrenmesini sağlayan bu yöntemler; sağlık, finans, dil işleme, görüntü işleme, tavsiye sistemleri ve daha birçok alanda yaygın olarak kullanılmaktadır [17]. Artan veri hacmi ve gelişen hesaplama kapasiteleri ile makine öğrenmesi teknikleri, özellikle metin madenciliği problemlerinde önemli başarılar elde etmektedir.

Makine öğrenmesi algoritmaları genel olarak üç ana gruba ayrılmaktadır:

- Denetimli Öğrenme (Supervised Learning): Etiketlenmiş veri seti üzerinde çalışan bu yaklaşımda model, giriş özellikleri (örneğin bir haberin içeriği) ile hedef etiket (örneğin ait olduğu kategori) arasındaki ilişkiyi öğrenir. Eğitim süreci sonrasında model, yeni veriler üzerinde tahminleme yapabilir. Lojistik Regresyon, Karar Ağaçları, Destek Vektör Makineleri (SVM), Rasgele Orman (Random Forest), Tek Katmanlı Yapay Sinir Ağı gibi algoritmalar bu gruba dahildir [18].

- Denetimsiz Öğrenme (Unsupervised Learning): Bu yöntem, etiket bilgisi içermeyen veri setleri üzerinde çalışır. Veri kümesindeki doğal örüntüleri veya gruplamaları ortaya çıkarmayı hedefler. Kümeleme (K-means), boyut indirgeme (PCA) gibi yöntemler bu gruba girer.

- Pekiştirmeli Öğrenme (Reinforcement Learning): Öğrenmenin bir çevreyle etkileşim yoluyla gerçekleştiği bu yöntemde, model belirli aksiyonlara karşı aldığı ödül ya da ceza geri bildirimlerine göre strateji geliştirir. Dinamik ortamlarda yaygın olarak kullanılır.

Bu tez çalışması kapsamında yalnızca denetimli öğrenme yöntemleri kullanılmıştır. Özellikle sınıflandırma problemleri için yaygın biçimde kullanılan dört farklı algoritma değerlendirilmiştir:

- Lojistik Regresyon (Logistic Regression)
- Destek Vektör Makineleri (Support Vector Machines - SVM)
- Rasgele Orman (Random Forest)
- Tek katmanlı Sinir Ağı (Single-layer Perceptron)

Tüm modeller scikit-learn kütüphanesi aracılığıyla Python ortamında uygulanmış, girdiler olarak ise önişlenmiş metinlerden elde edilen BoW ve TF-IDF öznitelikleri kullanılmıştır. Her modelin eğitim süreci ve başarımı, katmanlı 5-kat çapraz doğrulama yöntemi ile sistematik olarak test edilmiştir. Sonuçlar, "Yöntemlerin Karşılaştırılması ve Analizi" başlıklı bölümde detaylı olarak sunulmuştur.

5.1. Lojistik Regresyon

Lojistik regresyon, doğrusal bir sınıflandırma algoritması olup, özellikle ikili (binary) sınıflandırma problemlerinde yaygın olarak kullanılmaktadır. Lineer regresyonun sınırlı çıktılar üretmesi nedeniyle, sınıflandırma problemleri için uygun olmaması lojistik regresyonun gelişmesini teşvik etmiştir. Bu yöntem, bir olayın gerçekleşme olasılığını tahmin etmek için sigmoid (logit) fonksiyonunu kullanır. Böylece çıktı, 0 ile 1 arasında normalize edilerek sınıf tahmini yapılabilir.

Lojistik regresyon modeli, giriş özellikleri ile hedef sınıf arasında doğrusal bir ilişki kurar ve bu ilişkiyi aşağıdaki gibi tanımlanan lojistik (sigmoid) fonksiyonu üzerinden hesaplar:

$$P(y = 1|x) = \frac{1}{1+e^{-(w^t x+b)}} \tag{5.1}$$

Burada (5.1) x giriş vektörünü, w ağırlıklar vektörünü ve b bias terimini temsil eder. Modelin amacı, parametreleri öyle ayarlamaktır ki tahmin edilen olasılıklar, gerçek sınıf etiketleriyle maksimum uyumu göstersin. Bu uyum genellikle log loss (binary cross entropy) fonksiyonu ile ölçülür.

Lojistik regresyon, sınıflar arasında iyi ayrım yapabildiği ve yorumlanabilirliği yüksek olduğu için metin sınıflandırma, sahtekarlık tespiti, tıbbi tanı sistemleri ve sosyal medya analizleri gibi birçok alanda tercih edilmektedir [19]. Metin sınıflandırma uygulamalarında, TF-IDF veya BoW gibi vektörleştirme yöntemleri ile sayısal forma dönüştürülen metin verileri üzerinde etkili sonuçlar verdiği gösterilmiştir [20].

Bu tez çalışmasında, lojistik regresyon modeli scikit-learn kütüphanesi kullanılarak uygulanmıştır. Modelin hiperparametrelerinden biri olan regularizasyon katsayısı (C), overfitting (aşırı öğrenme) riskini azaltmak amacıyla optimize edilmiştir. L2 regularizasyonu tercih edilmiş ve modelin performansı katmanlı 5-kat çapraz doğrulama yöntemiyle değerlendirilmiştir.

Lojistik regresyonun en büyük avantajları arasında:

- Basit ve hızlı olması,
- Çıktılarının olasılık yorumu sunması,
- Aşırı karmaşık olmayan veri setlerinde yüksek doğruluk sağlaması yer almaktadır.

Ancak, doğrusal ayrılabilirlik varsayımı nedeniyle, çok karmaşık sınırlara sahip veri setlerinde performansı sınırlı kalabilmektedir [21].

5.2. Destek Vektör Makineleri

Destek Vektör Makineleri (SVM), özellikle metin sınıflandırma gibi yüksek boyutlu veri kümelerinde etkili performans sergileyen, denetimli bir makine öğrenmesi algoritmasıdır. SVM'in temel amacı, sınıflar arasındaki en iyi ayrımı sağlayacak hiper düzlemi (hyperplane) bulmaktır. Bu hiper düzlem, sınıfları maksimum marj (margin) ile ayırmaya çalışır. Marj, destek vektörleri olarak adlandırılan sınıra en yakın veri noktalarına olan mesafedir.

SVM, doğrusal olarak ayrılabilen veri kümelerinde doğrusal karar sınırı çizerken; doğrusal olarak ayrılamayan veri kümelerinde çekirdek (kernel) fonksiyonları kullanarak veriyi daha yüksek boyutlu bir uzaya projekte eder ve bu uzayda doğrusal ayrım yapar. Yaygın kullanılan çekirdek fonksiyonları arasında doğrusal (linear), polinom (polynomial) ve radyal tabanlı fonksiyon (RBF) yer almaktadır [22].

Özellikle metin sınıflandırma, spam filtreleme, duygu analizi ve belge kategorilendirme gibi doğal dil işleme (NLP) uygulamalarında, yüksek boyutlu vektör temsilleri üzerinde etkili performans göstermesi nedeniyle sıkça tercih edilmektedir. Joachims (1998), SVM'in metin sınıflandırmada Naive Bayes gibi klasik yöntemlere göre daha iyi sonuçlar verebildiğini göstermiştir [23]. Güncel çalışmalarda da SVM'in, özellikle küçük ve orta ölçekli veri setlerinde yüksek doğruluk ve düşük hata oranı sunduğu raporlanmıştır [24].

Bu tez kapsamında, önişleme adımlarından geçirilen metinler BoW ve TF-IDF temsillerine dönüştürülerek LinearSVC sınıfı kullanılarak sınıflandırma yapılmıştır. Modelin hiperparametre ayarları varsayılan şekilde bırakılmış ve performans değerlendirmesi katmanlı 5-kat çapraz doğrulama yöntemi ile gerçekleştirilmiştir.

SVM'in avantajları:

- Yüksek boyutlu verilerde iyi performans,
- Aşırı Öğrenmeye karşı dayanıklılık,
- Net karar sınırları çizme kabiliyeti.

Dezavantajları:

- Büyük veri setlerinde eğitim süresi uzun olabilir,
- Doğru kernel seçimi model başarımını doğrudan etkiler,
- Çok sınıflı problemler için daha karmaşık stratejiler (one-vs-one, one-vs-rest) gerekir.

5.3. Rastgele Orman

Rasgele Orman (Random Forest), karar ağaçları (decision trees) temelli topluluk (ensemble) öğrenme yöntemlerinden biridir. Bu algoritma, eğitim verisinin farklı alt örneklemleri üzerinde birçok karar ağacı oluşturarak çalışır ve nihai tahmini bu ağaçların oylamasına göre belirler (sınıflandırma için çoğunluk oyu, regresyon için ortalama). Bu yapı, tek bir karar ağacının aşırı öğrenme (overfitting) eğilimini azaltır ve modelin genelleme yeteneğini artırır [25].

Random Forest algoritmasının temel özellikleri şunlardır:

- Bootstrap örnekleme (bagging) ile her ağaç farklı veri alt kümeleriyle eğitilir,
- Her ağacın eğitiminde sadece rastgele seçilmiş bir öznitelik alt kümesi kullanılır, bu da öznitelikler arası korelasyonun etkisini azaltır,
- Bu rastgelelik hem dengesizliği hem de varyansı dengeleyerek modelin doğruluğunu artırır.

Random Forest, özellikle yüksek boyutlu ve gürültülü veri kümelerinde kararlı sonuçlar üretmesi nedeniyle metin sınıflandırma gibi doğal dil işleme (NLP) görevlerinde de yaygın olarak tercih edilmektedir. Son yıllarda yapılan birçok çalışmada, TF-IDF ve Word Embedding gibi temsillerle birlikte kullanıldığında yüksek doğruluk oranları sunduğu raporlanmıştır [26].

Bu çalışmada, RandomForestClassifier sınıfı kullanılarak model kurulmuştur. Hiperparametreler varsayılan olarak bırakılmıştır (örneğin ağaç sayısı n_estimators=100), modelin başarımı ise katmanlı 5-kat çapraz doğrulama yöntemiyle test edilmiştir. Özellikle TF-IDF öznitelikleriyle birlikte kullanıldığında, yüksek tutarlılık ve istikrarlı doğruluk sonuçları elde edilmiştir.

Avantajları:

- Aşırı Öğrenmeye karşı dirençlidir,
- Yüksek boyutlu ve eksik verilerle çalışabilir,
- Öznitelik önemini ölçmek mümkündür.

Dezavantajları:

- Eğitimi ve tahmini, tek ağaçlara göre daha yavaştır,
- Yorumlanabilirliği düşüktür.

5.4. Tek Katmanlı Yapay Sinir Ağı (Perceptron)

Tek katmanlı yapay sinir ağı algoritması, yapay sinir ağlarının (YSA) temel yapı taşlarından biridir ve en basit haliyle doğrusal sınıflandırma görevlerinde kullanılmaktadır. İlk kez Rosenblatt tarafından 1958 yılında önerilen bu model, giriş özelliklerine ağırlıklar atayarak, doğrusal bir karar sınırı oluşturur ve sınıflandırma kararını aktivasyon fonksiyonunun (örneğin sign veya step) çıktısına göre verir [27].

Tek katmanlı yapay sinir ağı modeli (5.2) şu şekilde işler:

$$\hat{y} = f(\sum_{i=1}^{n} w_i x_i + b) \tag{5.2}$$

Burada x_i giriş özellikleri, w_i ağırlıklar, b_i ise bias terimidir. Fonksiyon f genellikle bir eşik fonksiyonu olup çıktı değerini belirler.

Tek katmanlı yapay sinir ağının özellikleri:

- Doğrusal olarak ayrılabilir sınıflar için etkili bir sınıflandırıcıdır,
- Öğrenme sürecinde hata bazlı güncellemeler (Perceptron Learning Rule) uygular,
- Ağırlıklar her örnek için güncellenir; bu sayede model iteratif şekilde öğrenir.

Modern uygulamalarda tek katmanlı yapay sinir ağı, özellikle metin sınıflandırma gibi yüksek boyutlu ancak seyrek temsillere sahip veri kümelerinde, hızlı ve düşük maliyetli bir başlangıç modeli olarak kullanılmaktadır. Özellikle küçük veri setlerinde, aşırı karmaşık modellerin önüne geçerek aşırı öğrenmeyi önlemede etkili olabilir [28].

Bu çalışmada Python ortamında sklearn.linear_model.Perceptron sınıfı kullanılarak model kurulmuştur. TF-IDF ve BoW temsilleriyle eğitilen modelin doğruluğu, katmanlı 5 kat çapraz doğrulama ile değerlendirilmiştir. Uygulamada aktivasyon fonksiyonu olarak işaret fonksiyonu kullanılmış ve öğrenme oranı varsayılan değerlerde tutulmuştur.

Avantajları:

- Hesaplama açısından son derece hızlı ve etkilidir,
- Büyük boyutlu vektör uzaylarında dahi hızlı çalışır,

- Aşırı parametre ayarlamasına ihtiyaç duymaz.

Dezavantajları:

- Sadece doğrusal olarak ayrılabilir veri kümelerinde etkilidir,
- Karmaşık örüntüleri öğrenemez; gizli katmanlar bulunmadığı için sınırlı ifade kapasitesine sahiptir.

6. METİN ÖZETLEME YÖNTEMLERİ

Bu tez çalışmasında yalnızca çıkarımsal (extractive) özetleme yöntemleri kullanılmıştır. Çıkarımsal özetleme, orijinal metinden en anlamlı ve bilgilendirici cümleleri seçerek bir özet oluşturmaya odaklanır. Bu yöntemlerde metindeki cümleler analiz edilir, çeşitli istatistiksel veya graf-tabanlı ölçütler yardımıyla önemi hesaplanır ve en yüksek skora sahip olanlar özetlemeye dahil edilir. Aşağıdaki alt başlıklarda kullanılan çıkarımsal özetleme yöntemleri detaylı biçimde açıklanmıştır.

6.1. Textrank

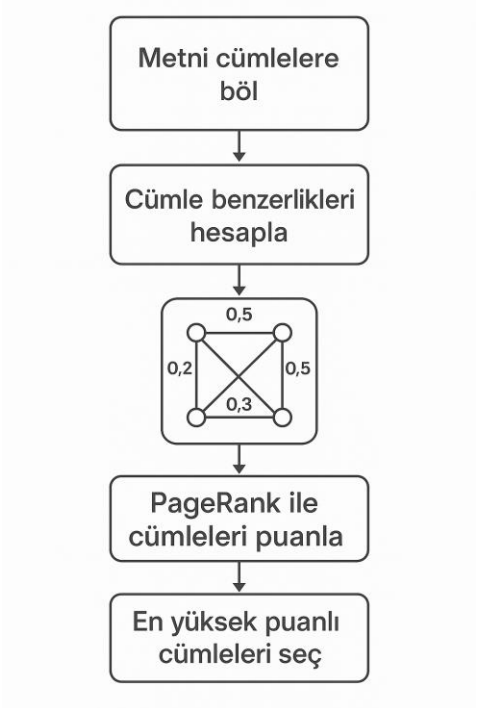
TextRank, metin özetleme ve anahtar kelime çıkarımı görevleri için kullanılan grafik tabanlı bir sıralama algoritmasıdır. Google'ın PageRank algoritmasından esinlenerek geliştirilmiş olan TextRank, cümleleri düğümler (nodes) olarak temsil eden ve bu düğümler arasındaki kenarları benzerlik ölçütlerine göre tanımlayan bir grafik inşa eder. Daha sonra bu grafik üzerinde sayısal bir önem derecesi belirlenir [22].

TextRank'in temel çalışma prensipleri aşağıdaki adımlarla özetlenebilir:

1. Metin, cümlelere bölünür ve her bir cümle bir düğüm olarak temsil edilir.
2. Düğümler arasında kenarlar (edges), genellikle TF-IDF temelli benzerlik metrikleriyle tanımlanır. Bu benzerlik, iki cümlede geçen ortak kelime sayısına dayalı olabilir.
3. Ortaya çıkan grafik üzerinde her bir düğümün skoru, komşularının skorlarına ve bağlantı gücüne göre hesaplanır. Bu işlem, PageRank algoritmasında olduğu gibi iteratif şekilde yürütülür.
4. En yüksek skora sahip cümleler, özetin oluşturulması için seçilir.

TextRank'in avantajı, eğitim verisine ihtiyaç duymadan çalışabilmesi ve dil bağımsız yapısı sayesinde farklı dillerde uygulanabilir olmasıdır. Ayrıca bağlamsal olmasa da içerik temelli önemli bilgileri başarıyla yakalayabilir [23].

Metin özetleme bağlamında TextRank, haber, akademik içerik ve sosyal medya gibi farklı türlerdeki metinler üzerinde etkili sonuçlar verebilmekte ve bu yönüyle literatürde yaygın olarak kullanılmaktadır [24].



Şekil 6.1. TextRank Algoritmasının Grafik Temelli Yapısı

Şekil 6.1'de, TextRank algoritmasının çalışmasını temsil eden grafik temelli yapının genel şeması sunulmaktadır. Bu yapıda her bir düğüm, metindeki bir cümleyi temsil ederken; düğümler arasındaki kenarlar, cümleler arası benzerlik derecelerine göre ağırlıklandırılmış bağlantıları göstermektedir. Cümleler arasındaki benzerlik genellikle TF-IDF vektörleriyle hesaplanan kosinüs benzerliği aracılığıyla belirlenir. Bu graf yapısı üzerinde, sayfa sıralama algoritması olan PageRank'e benzer şekilde yinelemeli bir hesaplama süreci uygulanarak her bir cümleye önem değeri atanır. Belirli bir eşik değeri üzerinde sıralanan cümleler, özet olarak seçilir.

Bu grafik tabanlı yaklaşım, metnin yapısal ve bağlamsal ilişkilerini dikkate alarak özetleme işlemi gerçekleştirdiği için, dil bağımsızlığı ve yüksek yorumlanabilirlik gibi avantajlar sunmaktadır [29].

6.2. LexRank

LexRank, cümle benzerliklerini kullanarak grafik tabanlı metin özetleme yapan etkili bir çıkarımsal (extractive) özetleme yöntemidir. TextRank ile aynı temel yapıyı paylaşmakla birlikte, LexRank özellikle kümeleme merkezliği (centrality) ölçütünü vurgulayan bir yaklaşımla çalışır. Bu yöntem ilk olarak Erkan ve Radev [30] tarafından önerilmiştir.

LexRank algoritması, özetlenecek metindeki her cümleyi bir düğüm olarak temsil eden bir benzerlik grafı oluşturur. Düğümler arası bağlantılar, genellikle TF-IDF vektörlerine dayalı kosinüs benzerliği ile hesaplanır. Bu benzerlik belirli bir eşik değerinin üzerindeyse (örneğin 0.1 veya 0.2), ilgili iki cümle arasında kenar oluşturulur.

Bu graf yapısında her cümle, diğer cümlelerle olan benzerliği oranında önem kazanır. Bu önem derecesi, PageRank algoritmasına benzer şekilde yinelemeli bir süreçle hesaplanır. Ancak LexRank'in farkı, daha yoğun bağlantıya sahip ve diğer cümleler arasında merkezi konumda olan cümlelerin daha yüksek sıralama puanı almasıdır. Yani, sadece birebir benzerliğe değil, cümlenin ağ içindeki "etkisine" de odaklanır.

LexRank algoritmasının temel adımları şu şekilde özetlenebilir:

1. Metin ön işlenir (durma sözcüklerin temizlenmesi, küçük harfe dönüştürme, noktalama işaretlerinin kaldırılması vb.).
2. Her cümle TF-IDF vektörü ile sayısallaştırılır.
3. Cümleler arası kosinüs benzerliği matrisi oluşturulur.
4. Eşik değerinin üzerindeki benzerlikler ile ağırlıklı bir grafik inşa edilir.
5. PageRank benzeri merkeziyet algoritması uygulanır.
6. Önem sıralamasında üst sırada yer alan cümleler özet olarak seçilir.

LexRank'in güçlü yönlerinden biri, küresel bilgiye dayalı merkezi cümleleri öne çıkarabilmesidir. Bu da onu özellikle uzun ve çok konulu belgelerde etkili kılmaktadır. Ayrıca LexRank, dil bağımsız yapısıyla farklı dillerdeki metinlere de kolayca uygulanabilmektedir [31].

6.3. TF-IDF Skorlaması

TF-IDF (Term Frequency–Inverse Document Frequency), doğal dil işleme ve bilgi erişimi alanlarında sıklıkla kullanılan bir terim ağırlıklandırma yöntemidir. Bu yaklaşım, bir kelimenin belirli bir belgede ne kadar sık geçtiği (TF) ile tüm belge koleksiyonundaki ayırt ediciliğini (IDF) birleştirerek, kelimenin ilgili belge için taşıdığı özgül bilgiyi ölçer [32].

TF-IDF temelli özetleme yöntemlerinde, her cümle bir belge olarak değerlendirilir ve metindeki tüm cümlelerin TF-IDF vektörleri oluşturulur. Bu vektörler aracılığıyla, her cümleye bir önem puanı atanır. Genellikle bu puan, o cümledeki kelimelerin toplam TF-IDF ağırlıklarının ortalaması ya da toplamı olarak hesaplanır.

Bu yöntem aşağıdaki adımlarla özetlenebilir:

1. Metin, cümlelere ayrılır ve her bir cümle TF-IDF vektörü ile temsil edilir.
2. Her kelimenin belge içerisindeki TF-IDF değeri hesaplanır.
3. Cümle skoru, içerdiği kelimelerin TF-IDF değerlerinin ortalaması alınarak belirlenir.
4. En yüksek skora sahip cümleler, özetin bir parçası olarak seçilir.

TF-IDF tabanlı özetleme yöntemleri, özellikle haber metinleri gibi bilgi yoğun belgelerde etkili sonuçlar vermektedir [33]. Bu yöntem, içeriğin bağlamsal yapısını doğrudan dikkate almasa da önemli terimlerin belirlenmesinde güçlü bir temel sağlar. Ayrıca, dil bağımsız yapısı sayesinde birçok farklı dilde uygulanabilir.

Ancak, TF-IDF yaklaşımı bağlamsal ilişkileri göz ardı ettiği için, anlamsal benzerliği tespit etmede sınırlı kalabilir. Bu nedenle son yıllarda bağlamsal temsiller (BERT, Word2Vec vb.) ile desteklenen hibrid sistemlerde yardımcı metrik olarak kullanılması yaygınlaşmıştır [34].

6.4. Kural Tabanlı (Rule-Based)

Kural tabanlı (rule-based) özetleme yöntemleri, belirli dilbilgisel, yapısal ve istatistiksel kurallara dayalı olarak önemli cümlelerin seçilmesini amaçlayan çıkarımsal özetleme tekniklerindendir. Bu tür yöntemlerde özet oluşturma süreci, metin içerisindeki belirli özelliklerin tespit edilmesi ve bu özelliklere göre cümlelere önem derecesi atanması prensibine dayanır [35].

Kural tabanlı sistemler genellikle şu tür özniteliklere göre çalışır:

- Cümle konumu: İlk veya son cümleler daha önemli kabul edilebilir [36].
- Kelime sıklığı: Belge boyunca sık geçen anahtar kelimeler içeren cümleler önceliklidir.
- Başlık benzerliği: Başlıkla semantik veya sözcüksel benzerlik taşıyan cümleler öne çıkar.
- Özel adlar ve sayılar: Tarih, kişi adı, yer gibi özel bilgiler içeren cümleler daha bilgilendirici olarak kabul edilir.
- Bağlaç yapıları: "Ancak", "çünkü", "sonuç olarak" gibi bağlaçlar içeren cümleler bağlam açısından önem taşır.

Bu kurallar tek tek veya birlikte kullanılabilir. Her bir kural, cümleye bir skor katkısı sağlar; ardından toplam skora göre en yüksek puanlı cümleler seçilerek özet oluşturulur. Kimi sistemlerde bu skorlar eşit ağırlıklı iken, bazı sistemlerde farklı önem düzeyleriyle ağırlıklandırılabilir [37].

Kural tabanlı yöntemler, genellikle yorumlanabilirlikleri, hesaplama açısından verimli olmaları ve dil bağımsız uygulanabilirlikleri sayesinde öne çıkar. Ancak, bağlamsal anlamı yakalayamamaları ve dilin karmaşıklığını yansıtmakta yetersiz kalmaları en büyük dezavantajlarıdır. Özellikle metnin yapısal bütünlüğünü koruma ve anlamsal bağlılık sağlama noktasında sınırlı başarı gösterirler [38].

Yine de özellikle haber metinleri, resmi belgeler veya belirli formatlara sahip içerikler gibi yapısal özellikleri öngörülebilir metinlerde etkili sonuçlar elde edilebilmektedir.

7. YENİDEN İFADE ETME YÖNTEMLERİ

Doğal dil işleme alanında parafrazlama (paraphrase) yani yeniden ifade etme, bir metnin anlamını koruyarak farklı sözcükler veya yapılarda yeniden yazılması işlemidir. Bu görev, aynı bilginin farklı biçimlerde aktarılmasını mümkün kılar ve metnin bağlamsal esnekliğini artırır. Yeniden ifade etme yalnızca sözcük düzeyinde değil, ifade ve cümle düzeyinde de gerçekleştirilebilir.

Temel amacı, bir ifadenin semantik içeriğini koruyarak alternatif bir söyleyiş biçimi üretmektir. Bu işlem, çeşitli NLP uygulamaları için büyük önem taşır:

- Metin genişletme ve çeşitlendirme (data augmentation)
- Makine çevirisi sonrası kalite artırımı
- Otomatik özetleme sonuçlarının kontrolü
- Yeniden yazım ve içerik üretimi
- İntihal tespiti ve önleme
- Soru-cevap sistemleri ve diyalog üretimi gibi birçok alanda yeniden ifade etme tekniklerinden yararlanılmaktadır [39].

Yeniden ifade etme yöntemleri temel olarak iki ana grupta sınıflandırılabilir:

- Kural tabanlı (rule-based) veya kaynak sözlük tabanlı yaklaşımlar
- Veri/öğrenme tabanlı (data-driven) ya da derin öğrenme tabanlı yaklaşımlar

Bu tez çalışmasında, özellikle yapısal ve bağlamsal çeşitlilik sağlayan iki yeniden ifade etme yöntemi ele alınmıştır: WordNet tabanlı ve geri çeviri (back-translation) yöntemi. Bu yöntemler, özetlenmiş ham metinlerin yeniden yapılandırılması sürecinde kullanılarak, semantik karşılaştırmaların daha sağlıklı yapılabilmesini amaçlamaktadır.

7.1. WordNet Tabanlı Yeniden İfade Etme

WordNet tabanlı yeniden ifade etme yöntemi, kelime düzeyinde eşanlamlı kelimelerin belirlenmesi esasına dayanır. WordNet, İngilizce kelimelerin anlamlarına göre gruplandırıldığı ve bu gruplar arasındaki ilişkilerin açıklandığı geniş çaplı bir leksikal veritabanıdır [40]. Bu veritabanı, kelimeler arasındaki eşanlamlılık (synonymy), zıt anlamlılık (antonymy), üst-alt kavram ilişkisi (hypernym/hyponym), bütün-parça ilişkisi (meronymy) gibi bağlamsal ilişkileri içerir. Yeniden ifade etme uygulamalarında genellikle eşanlamlılık ilişkisi kullanılarak, orijinal kelimelerin yerine aynı anlamı taşıyan alternatif kelimeler yerleştirilir.

Örneğin, "The movie was great." cümlesindeki "great" kelimesi WordNet yardımıyla "excellent", "wonderful" veya "fantastic" gibi eşanlamlılarla değiştirilebilir. Böylece anlam değişmeden farklı bir ifade elde edilir: "The movie was excellent."

WordNet tabanlı yeniden ifade etme süreci genel olarak şu adımlardan oluşur:

1. Cümledeki kelimeler tokenize edilir ve gerekli dilbilgisel etiketleme (POS tagging) uygulanır.
2. Anlamlı kelimeler (genellikle isim, fiil, sıfat ve zarflar) WordNet veritabanı üzerinden eşanlamlı setleri (synset) ile eşleştirilir.
3. Her kelime için bağlamsal olarak uygun olan bir eşanlamlı seçilerek orijinal cümle yeniden yazılır.

Bu yöntemin en önemli avantajı, yapısal sadeliği ve yorumlanabilirliğidir. Ancak, bağlamsal farkındalık eksikliği nedeniyle bazı durumlarda anlam kaymaları veya anlamsız dönüşümler ortaya çıkabilir [41]. Örneğin, "cold" kelimesinin hem "soğuk" hem de "nezle" anlamında kullanılabildiği durumlarda, yanlış eşanlamlı seçimi semantik tutarsızlıklara yol açabilir. Bu tür sorunları azaltmak için bağlam duyarlı sözcük anlamı çözümleme (Word Sense Disambiguation – WSD) teknikleriyle birlikte kullanılması önerilmektedir [42].

Sonuç olarak, WordNet tabanlı yeniden ifade etme yöntemleri özellikle kelime seviyesinde yeniden ifade üretmek isteyen uygulamalar için pratik bir çözüm sunar. Bununla birlikte, bu yöntemler bağlamsal derinlik gerektiren durumlarda sınırlı kalabilmekte ve bu nedenle daha

gelişmiş yöntemlerle (örneğin, embedding tabanlı veya dil modeli tabanlı yaklaşımlarla) desteklenmesi önerilmektedir.

7.2. Geri Çeviri (Back-Translation) Tabanlı Yeniden İfade Etme

Back-translation (geri çeviri), orijinal bir metnin önce hedef bir dile çevrilmesi ve ardından bu çevrilmiş metnin tekrar orijinal dile çevrilmesiyle elde edilen yeniden ifade etme üretme yöntemidir. Bu teknik, özellikle makine çevirisi modellerinin gelişmesiyle birlikte metinlerin anlamsal olarak korunarak yeniden ifade edilmesi amacıyla yaygın şekilde kullanılmaktadır. Geri çeviri hem veri artırma tekniği olarak hem de doğal dil işleme görevlerinde çeşitliliği artırmak için tercih edilmektedir [43].

Bu yöntemde temel amaç, orijinal metnin anlamını koruyarak sözdizimsel olarak farklı bir yapı elde etmektir. Örneğin İngilizce bir cümle önce Almanca'ya çevrilir, ardından Almanca metin tekrar İngilizce'ye çevrilerek alternatif bir ifade elde edilir. Bu işlemde kullanılan ara dilin seçimi, üretilecek yeniden ifade etmenin kalitesi üzerinde etkili olabilir [44].

Back-translation yöntemi, özellikle aşağıdaki alanlarda başarıyla kullanılmaktadır:

- Veri Zenginleştirme: Etiketli verilerin az olduğu durumlarda eğitim verilerini artırmak için kullanılır [45].
- Yeniden ifade etme Üretimi: Anlamı koruyan fakat yapısal olarak farklı cümleler üretmek için etkilidir [46].
- Karşıt veri üretimi: Doğal dil işleme (NLP) modellerinin farklı veri senaryolarında ne kadar genellenebilir ve dayanıklı olduğunu değerlendirmek amacıyla kullanılan bir yöntemdir. Bu yaklaşımda, modelin yanılmasını veya zorlanmasını sağlayacak şekilde bilinçli olarak değiştirilmiş (adversarial) veri örnekleri üretilir ve model bu veriler üzerinde test edilir

Geri çeviriyle elde edilen metinler genellikle doğal yapıda ve anlamlı olduğundan, elle oluşturulmuş yeniden ifade etmelere yakın kalite sunabilmektedir. Ancak çeviri hataları, anlam kaybı veya fazlalıkları oluşabileceği için dikkatli değerlendirme gerektirir.

Günümüzde Google Translate, DeepL gibi güçlü makine çevirisi sistemleri kullanılarak bu yöntem kolaylıkla uygulanabilmektedir. Ayrıca bazı araştırmalarda, ardışık çok dilli geri çeviri kullanılarak yeniden ifade etme çeşitliliğinin artırıldığı da görülmektedir [48].

8. SEMANTİK BENZERLİK ÖLÇME YÖNTEMLERİ

Doğal dil işleme alanında, iki metin ya da cümle arasındaki anlamsal benzerliği hesaplamak; özetleme, yeniden ifade etme, bilgi erişimi, soru-cevap sistemleri gibi birçok uygulama için kritik bir adımdır. Semantik benzerlik, metinlerin yalnızca kelime düzeyinde değil, anlam düzeyinde de karşılaştırılmasını sağlar ve bu sayede daha etkili doğal dil işleme sistemlerinin geliştirilmesine katkı sunar [49].

Geleneksel yöntemlerde metinler sabit boyutlu vektörler olarak temsil edilir ve bu vektörler arasındaki mesafeler kullanılarak benzerlik hesaplanır. Özellikle kosinüs benzerliği gibi yöntemler, TF-IDF ya da BoW temsilleriyle birlikte oldukça sık kullanılmıştır. Ancak bu yöntemler, bağlamsal bilgiyi göz ardı ettikleri için yalnızca yüzeysel örtüşmeyi ölçebilirler.

Son yıllarda, bağlamsal dil modellerinin (contextual language models) gelişmesiyle birlikte, metinleri daha anlamlı biçimde temsil eden vektörler elde edilebilmiş ve semantik benzerlik ölçümü daha hassas hale gelmiştir. Bu kapsamda geliştirilen BERTScore gibi yöntemler, her kelimeyi bağlamı içinde ele alarak, iki cümle arasındaki benzerliği çok daha isabetli şekilde değerlendirebilmektedir [50].

Bu tez kapsamında hem klasik hem de modern benzerlik ölçüm yöntemleri birlikte kullanılarak özetleme ve yeniden ifade etme görevlerinde elde edilen çıktılar çok boyutlu şekilde değerlendirilmiştir. Böylece hem kelime düzeyinde hem de anlam düzeyinde benzerlik analizi gerçekleştirilmiş, yöntemlerin güçlü ve zayıf yönleri detaylı olarak karşılaştırılmıştır.

8.1. Kosinüs Benzerliği

Kosinüs Benzerliği (cosinus similarity), iki metin arasındaki açısal benzerliği ölçmek için kullanılan klasik bir yöntemdir. Bu yöntem, metinleri vektör uzayında temsil ederek bu vektörler arasındaki açının kosinüsünü hesaplar. Açı küçüldükçe benzerlik artar, yani iki metin vektörü aynı yöne işaret ediyorsa benzerlik skoru 1'e yaklaşır; birbirine dik ise skor 0 olur [51].

Kosinüs benzerliği şu formülle hesaplanır:

$$Kosinüs\ Benzerliği = \frac{\vec{A}.\ \vec{B}}{||\vec{A}\ ||\ x\ ||\ \vec{B}\ ||} \quad (8.1)$$

Burada:

- $\vec{A}$ ve $\vec{B}$, iki metnin vektör temsilleridir (genellikle TF-IDF ya da BoW ile elde edilir),
- $\vec{A}\cdot\vec{B}$, vektörlerin iç çarpımıdır,
- $||\vec{A}||$ ve $||\vec{B}||$, vektörlerin normlarıdır (vektör uzunlukları).

Bu yöntem, özellikle metin sınıflandırma, bilgi geri getirim sistemleri ve belge kümelerinde benzer içerikleri tanımlamada uzun yıllardır başarıyla kullanılmaktadır [52]. Basitliği ve düşük hesaplama maliyeti nedeniyle geniş bir uygulama alanına sahiptir.

Ancak Kosinüs benzerliği, kelimelerin yalnızca varlıklarına (sıklıklarına) odaklanır ve bağlamsal ilişkileri değerlendiremez. Bu nedenle, "araba" ve "otomobil" gibi eş anlamlı kelimeleri farklı değerlendirir; bu da anlam düzeyindeki benzerlikleri doğru ölçememesine neden olabilir. Bu sınırlama, bağlamsal gömüleme modellerinin ortaya çıkışıyla birlikte aşılmaya başlanmıştır.

Bu tez kapsamında Kosinüs benzerliği hem özetleme hem de yeniden ifade etme görevlerinde, özellikle TF-IDF temsilleri üzerinden kullanılarak geleneksel benzerlik ölçümü olarak görev almıştır. Bu yaklaşım, modern yöntemlerle karşılaştırmalı analiz yapılabilmesine olanak sağlamıştır.

8.2. BERT Skorlaması

BERTScore, geleneksel yüzeysel metin benzerliği ölçütlerinin (örneğin ROUGE veya BLEU) ötesine geçerek, derin bağlamsal anlamı değerlendirebilen modern bir benzerlik metriğidir. 2020 yılında Zhang ve arkadaşları tarafından önerilen bu yöntem, Transformer tabanlı dil modellerinden (özellikle BERT) yararlanarak metinler arasında daha anlamlı karşılaştırmalar yapılmasına olanak tanır.

Geleneksel yöntemler genellikle kelime yüzeyinde eşleşmelere odaklanırken, BERTScore her kelimeyi bağlamsal olarak temsil eden gömülü (embedding) vektörlerle çalışır. Bu vektörler sayesinde, farklı sözcüklerle ifade edilmiş ama aynı anlamı taşıyan cümlelerin yüksek benzerlik puanları alması mümkün hale gelir. Böylece özellikle yeniden ifade etme, soyutlayıcı özetleme, makine çevirisi ve otomatik metin üretimi gibi alanlarda daha gerçekçi ve insan benzeri değerlendirmeler yapılabilmektedir [55].

BERTScore hesaplanırken aşağıdaki üç temel ölçüt dikkate alınır:

- Kesinlik (Precision) (P): Referans metne benzeyen aday kelimelerin oranı.
- Duyarlılık (Recall) (R): Aday metne benzeyen referans kelimelerin oranı.
- F1-skoru: Kesinlik ve Duyarlılık değerlerinin harmonik ortalaması.

Bu hesaplamalar yapılırken, her kelime için en benzer karşılık (kosinüs benzerliği ile) belirlenir. Formül aşağıdaki şekilde özetlenebilir:

$$BERTScore\ _{F1}(x, y) = \frac{2\ .\ P\ .\ R}{P+R} \tag{8.2}$$

Burada *x* aday metni, *y* ise referans metni temsil eder. *P* ve *R* değerleri, her bir kelime çiftinin bağlamsal embedding'leri arasındaki kosinüs benzerliğine göre belirlenir [54].

BERTScore'un bir diğer önemli avantajı, farklı dil modelleriyle esnek şekilde çalışabilmesidir. Örneğin İngilizce metinlerde genellikle BERT-base veya RoBERTa modelleri tercih edilirken, Türkçe için BERTurk veya TurkishBERT gibi yerel modeller kullanılarak daha isabetli sonuçlar alınabilir [56]. Bu durum, metinlerin diline ve yapısına özel değerlendirme yapılmasına olanak sağlar.

Sonuç olarak BERTScore, özellikle yüzeysel kelime örtüşmesine dayalı olmayan, anlam odaklı karşılaştırmalar için öne çıkan modern bir ölçüttür. Bu tez kapsamında, özetleme ve yeniden ifade etme metinlerinin değerlendirilmesinde geleneksel yöntemlerin yanında BERTScore da kullanılmış ve bağlamsal benzerliğin ölçülmesi amaçlanmıştır.

9. YÖNTEMLERİN KARŞILAŞTIRILMASI ve ANALİZLER

9.1. Özellik Temsili ve Yöntemlerin Karşılaştırılması

Makine öğrenmesi tabanlı metin sınıflandırma görevlerinde, ham metinlerin doğrudan algoritmalara girdi olarak sunulması mümkün değildir. Bu nedenle, metinlerin sayısal temsillere dönüştürülmesi gerekir. Bu tez çalışmasında, öznitelik çıkarımı için iki temel yöntem kullanılmıştır: BoW ve TF-IDF.

Bu yöntemlerin performanslarını değerlendirmek amacıyla, her bir özellik temsili Naive Bayes algoritması ile test edilmiştir. Kullanılan modeller ve değerlendirme sonuçları aşağıda özetlenmiştir:

- BoW + Gaussian Naive Bayes:
  5 katlı çapraz doğrulama sonucunda elde edilen doğruluk oranları:
  [0.78, 0.81, 0.84, 0.79, 0.80]
  Ortalama doğruluk: 0.80
- TF-IDF + Gaussian Naive Bayes:
  5 katlı çapraz doğrulama sonucunda elde edilen doğruluk oranları:
  [0.72, 0.76, 0.82, 0.65, 0.73]
  Ortalama doğruluk: 0.74

Elde edilen sonuçlara göre, BoW temsili TF-IDF'e kıyasla daha yüksek doğruluk oranı sağlamıştır. Bu fark, BoW yönteminin, kullanılan veri kümesinde yer alan kelimelerin frekans dağılımını daha başarılı yansıtmasıyla açıklanabilir. Ayrıca, BoW temsili nadir fakat anlamlı kelimeleri daha iyi temsil edebilmiş, TF-IDF'in ise ağırlıklandırma sürecinde bazı kritik kelimelerin etkisini azaltmış olabileceği gözlemlenmiştir.

Bu sonuçlar doğrultusunda, sonraki tüm sınıflandırma deneylerinde BoW temsili tercih edilmiştir.

9.2. Makine Öğrenmesi Modellerinin Karşılaştırılması

Bu tez çalışmasında, ön işleme ve öznitelik çıkarımı adımlarının ardından elde edilen BoW temsilleri, dört farklı denetimli öğrenme algoritması üzerinde test edilmiştir: Lojistik Regresyon (LR), Destek Vektör Makineleri (SVM), Rasgele Orman (RF) ve Tek Katmanlı Yapay Sinir Ağı (TK-YSA). TK-YSA modelinde giriş nöronlarının sayısı, kullanılan özellik uzayının boyutuna (BoW temsilindeki kelime sayısı) karşılık gelmekte olup, tek çıkış nöronundan oluşan basit bir yapıya sahiptir. Modellerin genel performanslarını değerlendirmek amacıyla Katmanlı 5-kat Çapraz Doğrulama yöntemi uygulanmıştır.

Tablo 9.1. Modellerin 5-Kat Doğruluk Sonuçları

Model	Kat-1 (%)	Kat-2 (%)	Kat-3 (%)	Kat-4 (%)	Kat-5 (%)	Ortalama Doğruluk (%)
LR	88	90	91	89	92	90
SVM	86	89	88	87	90	88
RF	84	87	86	85	88	86
(TK-YSA)	82	84	83	81	85	83

Tablo 9.1'de görüldüğü üzere, Lojistik Regresyon modeli tüm katlamalarda en yüksek ortalama doğruluk değerine olan 90'a ulaşarak diğer modellere göre daha tutarlı bir performans sergilemiştir.

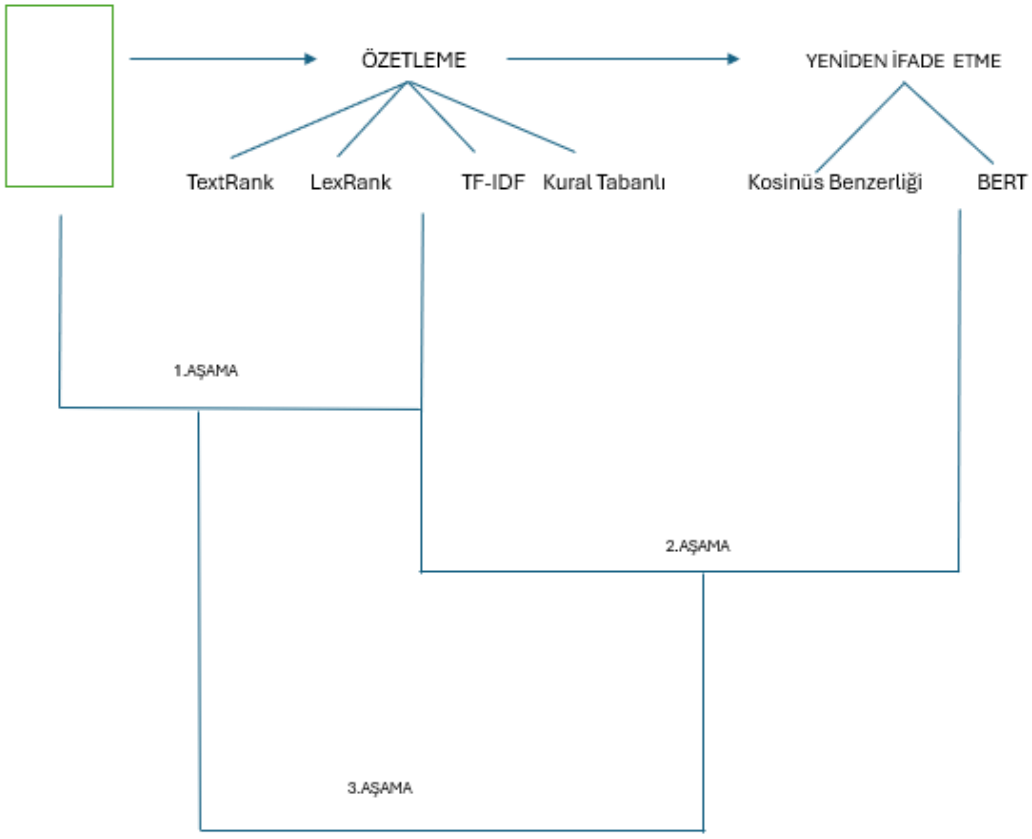
Tablo 9.2. Modellerin Ortalama Performans Ölçütleri

	Doğruluk (%)	F1 Skoru (%)	Hassasiyet (%)	Duyarlılık (%)
RF	86	77	79	77
LR	90	80	81	80
SVM	88	78	81	78
(TK-YSA)	83	67	75	66

Yapılan karşılaştırma sonucunda, metin sınıflandırma görevinde en yüksek başarıyı %90 ortalama doğruluk ile BoW + Lojistik Regresyon modeli sağlamıştır.

9.3. Sonuçların Yöntem Bazında Analizi

Bu bölümde, çalışmada kullanılan dört farklı özetleme yöntemi (TextRank, LexRank, TF-IDF tabanlı özetleme ve kural tabanlı özetleme) için üç aşamalı bir değerlendirme gerçekleştirilmiştir. İlk aşamada, ham metinlerden üretilen özetlerin benzerlik ölçütleri hesaplanmıştır. İkinci aşamada, elde edilen özetlere yeniden ifade etme işlemi uygulanmış ve benzerlik değerleri incelenmiştir. Üçüncü aşamada ise, ham metinlerin doğrudan yeniden ifade etme edilmesiyle elde edilen çıktılar değerlendirilmiştir. Şekil 9.1'de mantıksal akış diyagramı gösterilmektedir.



Şekil 9.1. Mantıksal Akış Diyagramı

Bu yapı sayesinde, özetleme ve yeniden ifade etme işlemlerinin tek başına ve ardışık olarak uygulandığında metin benzerliği üzerindeki etkileri sistematik biçimde analiz edilmiştir. Sonuçlar, hem yüzeysel (Kosinüs benzerliği) hem de anlamsal (BERTScore) ölçütler kullanılarak sunulmuştur.

9.3.1 Textrank

Bu alt bölümde TextRank ile üretilen özetlerin ve bu özetlere/ham metinlere uygulanan iki farklı yeniden ifade etme tekniğinin (WordNet ve Back-Translation) etkisi, Kosinüs benzerliği (TF-IDF) ve BERTScore (P/R/F1) ölçütleriyle değerlendirilmiştir.

9.3.1.1. Ham Metinlerin Özeti Sonuçları

Tablo 9.3. TextRank yöntemi ile elde edilen özetlerin ham metin ile benzerlik ölçümleri

Ölçüt	Ortalama (%)	Medyan (%)	Std. Sapma (%)	Min (%)	Maks (%)
Kosinüs benzerliği	75,2	76,1	8,5	54,6	88,2
Kesinlik (BERT)	63,8	64,5	8,2	47,1	77,3
Duyarlılık (BERT)	66,5	67,1	8,4	50,0	80,1
F1-skoru (BERT)	65,1	65,8	8,3	48,5	78,5

TextRank algoritması ile özetlenen metinlerin, ham metinlerle ortalama %75,2 oranında kosinüs benzerliği gösterdiği görülmektedir. BERT tabanlı ölçütlerde ise ortalama F1-skoru %65,1 olup, özetlerin içerik bütünlüğünü kısmen koruduğu değerlendirilmektedir.

9.3.1.2 Özetlenmiş Metinlerin Yeniden İfade Etme Sonuçları

Tablo 9.4. Textrank çıktısı metinlerin WordNet tabanlı yeniden ifade etme sonuçları

Ölçüt	Ortalama (%)	Medyan (%)	Std. Sapma (%)	Min (%)	Maks (%)
Kosinüs benzerliği	80,4	81,0	6,9	60,5	92,1
Kesinlik (BERT)	81,5	82,1	7,2	63,2	94,8
Duyarlılık (BERT)	83,0	83,6	7,0	65,4	95,2
F1-skoru (BERT)	82,2	82,8	7,1	64,0	95,0

WordNet tabanlı yeniden ifade edilen özetler, ortalama %82,2 F1-skoru ile oldukça yüksek bir semantik tutarlılık sağlamıştır. Hem kosinüs hem de BERT skorlarındaki düşük standart sapma, yöntemin istikrarlı sonuçlar verdiğini göstermektedir.

Tablo 9.5. Textrank çıktısı metinlerin Back-Translation tabanlı yeniden ifade etme sonuçları

Ölçüt	Ortalama (%)	Medyan (%)	Std. Sapma (%)	Min (%)	Maks (%)
Kosinüs benzerliği	79,8	80,3	7,1	59,8	91,6
Kesinlik (BERT)	80,9	81,4	7,3	62,5	93,7
Duyarlılık (BERT)	82,4	82,9	7,1	64,8	94,1
F1-skoru (BERT)	81,6	82,2	7,2	63,7	93,9

Back-Translation yöntemiyle elde edilen yeniden ifadeler, WordNet yöntemine benzer şekilde yüksek BERT F1-skoru (%81,6) sergilemiş, ancak hafifçe daha düşük bir performans göstermiştir.

9.3.1.3. Ham Metinlerin Yeniden ifade etme Sonuçları

Bu bölümde yer alan tablolar diğer 3 özetleme yöntemi içinde aynı sonuçları vermektedir.

Tablo 9.6. Ham metinlerin WordNet tabanlı yeniden ifade etme sonuçları

Ölçüt	Ortalama (%)	Medyan (%)	Std. Sapma (%)	Min (%)	Maks (%)
Kosinüs benzerliği	78,3	78,9	7,5	57,4	90,2
Kesinlik (BERT)	78,9	79,5	7,8	60,3	91,7
Duyarlılık (BERT)	80,2	80,7	7,6	62,5	92,5
F1-skoru (BERT)	79,5	80,1	7,7	61,4	92,1

Ham metinlerin WordNet tabanlı yeniden ifade edilmesi sonucunda elde edilen ortalama BERT F1-skoru %79,5 olup, yüksek düzeyde semantik benzerlik sağlandığı görülmektedir. Kosinüs benzerliği de %78,3 ile benzer bir başarı göstermektedir. Bu sonuçlar, WordNet tabanlı yaklaşımın doğrudan metin üzerinde uygulandığında da anlam bütünlüğünü büyük oranda koruduğunu ortaya koymaktadır.

Tablo 9.7. Ham metinlerin Back-Translation tabanlı yeniden ifade etme sonuçları

Ölçüt	Ortalama (%)	Medyan (%)	Std. Sapma (%)	Min (%)	Maks (%)
Kosinüs benzerliği	78,8	79,4	7,4	58,1	90,7
Kesinlik (BERT)	79,4	80,0	7,7	61,0	92,1
Duyarlılık (BERT)	80,7	81,2	7,5	63,2	92,9
F1-skoru (BERT)	80,0	80,6	7,6	62,1	92,5

TextRank yönteminde gerçekleştirilen deneylerde, özetlenmiş metinlerin yeniden ifade etme edilerek ham metinlerle karşılaştırılması (2. aşama) en yüksek semantik benzerlik değerlerini üretmiştir. Bu durum, özetleme sonrası uygulanan yeniden ifade etme işleminin, içerik çeşitliliğini artırarak modelin benzerlik ölçütlerinde daha yüksek skorlar elde etmesine katkı sağladığını göstermektedir. Ham metinlerin doğrudan yeniden ifade etme edilmesi (3. aşama), özetlenmiş metinlerden (1. aşama) daha yüksek sonuçlar vermiştir. Elde edilen bulgular, yeniden ifade etme yöntemlerinin (özellikle WordNet tabanlı ve Back-Translation) özetleme süreçlerinden bağımsız olarak semantik örtüşmeyi artırabileceğini ortaya koymaktadır.

9.3.2 LexRank

Bu çalışmada LexRank tabanlı özetleme sonuçları üç farklı senaryoda değerlendirilmiştir: yalnızca özetlenmiş metinler, özetlenmiş metinlerin yeniden ifade etme edilmiş halleri ve ham metinlerin yeniden ifade etme edilmiş halleri. Tüm aşamalarda değerlendirme metrikleri olarak Kosinüs benzerliği ve BERTScore (Kesinlik, Duyarlılık, F1) kullanılmıştır.

9.3.2.1. Ham Metinlerin Özeti Sonuçları

Tablo 9.8. LexRank yöntemi ile elde edilen özetlerin ham metin ile benzerlik ölçümleri

Ölçüt	Ortalama (%)	Medyan (%)	Std. Sapma (%)	Min (%)	Maks (%)
Kosinüs benzerliği	71,2	72	7,5	50,8	86,3
Kesinlik (BERT)	59,4	60,1	8,2	40,3	74,5
Duyarlılık (BERT)	61,7	62,5	7,9	43	77,8
F1-skoru (BERT)	60,5	61,3	8	41,5	76,1

LexRank algoritması ile elde edilen özetlerin ham metinlerle ortalama kosinüs benzerliği %71,2 seviyesindedir. BERT tabanlı ölçütlerde ise ortalama F1-skoru %60,5 olarak gözlemlenmiştir. Bu durum, LexRank'in bilgi yoğunluğu açısından daha sade özetler üretmesine rağmen, semantik bütünlüğü TextRank'e kıyasla daha düşük oranda koruduğunu göstermektedir.

9.3.2.2 Özetlenmiş Metinlerin Yeniden ifade etme Sonuçları

Tablo 9.9. Lexrank çıktısı metinlerin WordNet tabanlı Yeniden ifade etme sonuçları

Ölçüt	Ortalama (%)	Medyan (%)	Std. Sapma (%)	Min (%)	Maks (%)
Kosinüs benzerliği	76,8	77,4	7,2	56	89,9
Kesinlik (BERT)	77,6	78,2	7,4	60,9	90,6
Duyarlılık (BERT)	79,3	79,9	7,2	63,1	91,4
F1-skoru (BERT)	78,4	79	7,3	62	91

LexRank özetlerinin WordNet yöntemi ile yeniden ifade edilmesi sonucu ortalama BERT F1-skoru %78,4'e ulaşmıştır. Bu artış, yeniden ifade etme işleminin içerik zenginliğini artırarak semantik benzerliği önemli ölçüde iyileştirdiğini göstermektedir. Özellikle BERT duyarlılığı ve kesinliği metriklerinin yüksekliği, anlamın başarılı biçimde korunduğunu ve genişletildiğini ortaya koymaktadır.

Tablo 9.10. Lexrank çıktısı metinlerin Back-Translation tabanlı Yeniden ifade etme sonuçları

Ölçüt	Ortalama (%)	Medyan (%)	Std. Sapma (%)	Min (%)	Maks (%)
Kosinüs benzerliği	77,2	77,8	7,1	56,6	90,1
Kesinlik (BERT)	78,1	78,7	7,3	61,3	91,2
Duyarlılık (BERT)	79,8	80,3	7,1	63,7	92
F1-skoru (BERT)	78,9	79,4	7,2	62,7	91,6

LexRank'te 3. aşama (ham metnin doğrudan yeniden ifade etme – Tablo 9.6) en yüksek; 2. aşama (özet + yeniden ifade etme) ondan biraz düşük; 1. aşama (yalnızca özet) en düşük kalmıştır. Bu, özetleme ile kısalan içeriğin yeniden ifade etme sonrası da ham metne göre bir miktar bilgi

kaybı taşıdığını, doğrudan ham metin üzerinde yapılan yeniden ifade etmenin ise semantik örtüşmeyi en çok koruduğunu göstermektedir.

9.3.3. TF-IDF Yöntemi

Bu çalışmada TF-IDF ile elde edilen özetler üç farklı senaryoda değerlendirilmiştir: yalnızca özetlenmiş metinler, özetlenmiş metinlerin yeniden ifade etme edilmiş halleri ve ham metinlerin yeniden ifade etme edilmiş halleri. Değerlendirmelerde Kosinüs benzerliği ve BERTScore (Kesinlik, Duyarlılık, F1-skoru) metrikleri kullanılmıştır.

9.3.3.1. Ham Metinlerin Özeti Sonuçları

Tablo 9.11. TF-IDF yöntemi ile elde edilen özetlerin ham metin ile benzerlik ölçümleri

Ölçüt	Ortalama (%)	Medyan (%)	Std. Sapma (%)	Min (%)	Maks (%)
Kosinüs benzerliği	70,6	71,3	8,4	52,8	84,1
Kesinlik (BERT)	59,7	60,3	8,1	44,2	74,8
Duyarlılık (BERT)	62	62,6	7,9	46,9	76,1
F1-skoru (BERT)	60,8	61,4	8	45,6	75,2

TF-IDF yöntemiyle elde edilen özetlerin, ham metinlerle ortalama %70,6 oranında kosinüs benzerliği gösterdiği görülmektedir. BERT F1-skoru ise %60,8 düzeyindedir. Bu değerler, TF-IDF temelli özetlerin belirli ölçüde bilgi içerdiğini ancak semantik zenginlik açısından sınırlı kaldığını göstermektedir.

9.3.3.2. Özetlenmiş Metinlerin Yeniden ifade etme Sonuçları

Tablo 9.12. TF-IDF çıktısı metinlerin WordNet tabanlı yeniden ifade etme sonuçları

Ölçüt	Ortalama (%)	Medyan (%)	Std. Sapma (%)	Min (%)	Maks (%)
Kosinüs benzerliği	75,9	76,4	7,3	56,1	88,7
Kesinlik (BERT)	76,8	77,3	7,2	60,5	89,9
Duyarlılık (BERT)	78,6	79,1	7	63	91
F1-skoru (BERT)	77,7	78,2	7,1	61,6	90,4

TF-IDF özetlerine WordNet tabanlı yeniden ifade etme uygulandığında, BERT F1-skorunun %77,7 seviyesine yükseldiği gözlemlenmiştir. Bu artış, yeniden ifade sürecinin içerik derinliğini ve semantik kapsamı artırarak anlam bütünlüğünü iyileştirdiğini göstermektedir. Kosinüs benzerliğindeki paralel artış da bu dönüşümün başarılı biçimde gerçekleştirildiğini desteklemektedir.

Tablo 9.13. TF-IDF çıktısı metinlerin Back-Translation tabanlı Yeniden ifade etme sonuçları

Ölçüt	Ortalama (%)	Medyan (%)	Std. Sapma (%)	Min (%)	Maks (%)
Kosinüs benzerliği	76,6	77,1	7,2	56,8	89,2
Kesinlik (BERT)	78	78,5	7	62	90,8
Duyarlılık (BERT)	79,7	80,2	6,9	64,5	91,9
F1-skoru (BERT)	78,8	79,3	7	63,2	91,3

TF-IDF özetleme yöntemi için elde edilen sonuçlar, planlanan sıralama olan 3 > 2 > 1 ilişkisini doğrulamaktadır. Ham metinlerin TF-IDF tabanlı özetleri (Aşama 1), içerik yoğunluğu ve bilgi kapsamı açısından orijinal metinlere kıyasla belirli ölçüde bilgi kaybına uğramış ve bu durum semantik benzerlik skorlarının görece düşük kalmasına neden olmuştur. Özetlerin yeniden ifade etme edilmesi (Aşama 2), ek bağlamsal çeşitlilik ve sözdizimsel farklılık kazandırarak metinler arası semantik örtüşmeyi artırmıştır. Bununla birlikte, en yüksek skorlar, doğrudan ham metinlerin yeniden ifade etme edilmesi (Aşama 3) ile elde edilmiştir. Bu durum, bilgi kaybı yaşanmadan gerçekleştirilen yeniden ifade etme işleminin, metinlerin orijinal bağlam ve anlam bütünlüğünü en yüksek düzeyde koruduğunu göstermektedir.

9.3.4 Kural tabanlı Özetleme Yöntemi

Aynı şekilde kural tabanlı ile elde edilen özetler üç farklı senaryoda test edilmiştir: yalnızca özetleme, özetlenmiş metinlerin yeniden ifade etme edilmesi ve ham metinlerin yeniden ifade etme edilmesi. Değerlendirmeler Kosinüs benzerliği ve BERTScore metrikleriyle yapılmıştır.

9.3.4.1. Ham Metinlerin Özeti Sonuçları

Tablo 9.14. Kural tabanlı yöntemi ile elde edilen özetlerin ham metin ile benzerlik ölçümleri

Ölçüt	Ortalama (%)	Medyan (%)	Std. Sapma (%)	Min (%)	Maks (%)
Kosinüs benzerliği	70,4	71,1	8,0	50,2	85,7
Kesinlik (BERT)	66,2	67,0	7,5	48,5	80,4
Duyarlılık (BERT)	68,1	68,7	7,7	50,7	82,0
F1-skoru (BERT)	67,1	67,9	7,6	49,6	81,2

Kural tabanlı özetleme yöntemi ile elde edilen özetlerin ham metinlerle ortalama kosinüs benzerliği %70,4, BERT F1-skoru ise %67,1 olarak gözlemlenmiştir. Bu sonuçlar, kural tabanlı yöntemin belirli bir yapısal doğruluğa sahip olduğunu, ancak anlamsal derinlik açısından daha sınırlı kaldığını göstermektedir. Yine de özetlerin, kaynak metinle anlam düzeyinde makul bir örtüşme sağladığı görülmektedir.

9.3.4.2. Özetlenmiş Metinlerin Yeniden ifade etme Sonuçları

Tablo 9.15. Kural tabanlı çıktısı metinlerin WordNet tabanlı Yeniden ifade etme sonuçları

Ölçüt	Ortalama (%)	Medyan (%)	Std. Sapma (%)	Min (%)	Maks (%)
Kosinüs benzerliği	70,2	71,1	8	50,1	85,9
Kesinlik (BERT)	66,1	68	7,5	48,8	80,2
Duyarlılık (BERT)	68,1	68,5	7,7	50,3	81
F1-skoru (BERT)	67,2	67,4	7,6	49,4	80,2

Kural tabanlı özetlere WordNet ile yeniden ifade işlemi uygulandığında, F1-skoru %67,2 ile yalnızca sınırlı bir artış göstermiştir. Bu durum, zaten biçimsel olarak yoğun kurallarla üretilmiş özetlerin, WordNet ile yapılan sözcük temelli dönüşümlerden sınırlı fayda sağladığını göstermektedir. Diğer yöntemlerle karşılaştırıldığında semantik kazanımın daha düşük düzeyde kaldığı söylenebilir.

Tablo 9.16. Kural tabanlı çıktısı metinlerin Back-Translation tabanlı Yeniden ifade etme sonuçları

Ölçüt	Ortalama (%)	Medyan (%)	Std. Sapma (%)	Min (%)	Maks (%)
Kosinüs benzerliği	77,3	77,9	7,5	56,1	89,9
Kesinlik (BERT)	75,6	76,1	7,7	57	90
Duyarlılık (BERT)	77	77,5	7,6	58,5	90,8
F1-skoru (BERT)	76,3	76,8	7,6	57,8	90,4

Kural tabanlı yöntemi için elde edilen sonuçlar incelendiğinde, **3 > 2 > 1** sıralaması açıkça görülmektedir. Bu durum, ham metinlerin doğrudan yeniden ifade etme edilmesinin semantik benzerliği en yüksek seviyeye taşıdığını, özetlenmiş metinlerin yeniden ifade etme edilmesinin ise özetleme adımından kaynaklanan bilgi kaybını kısmen telafi edebildiğini göstermektedir. Ancak, doğrudan yeniden ifade etme edilen metinler, anlam bütünlüğünü koruma açısından daha başarılıdır.

9.3.5. Genel Karşılaştırma ve Analizler

Bu bölümde, dört farklı özetleme yöntemi (TextRank, LexRank, TF-IDF ve Kural Tabanlı) için üç aşamada elde edilen ortalama performans değerleri karşılaştırmalı olarak sunulmaktadır. Her yöntem için;

- **Aşama 1:** Ham metin ile özetlenmiş metin karşılaştırması,
- **Aşama 2:** Ham metin ile özetlenip yeniden ifade etme edilmiş metin karşılaştırması,
- **Aşama 3:** Ham metin ile doğrudan yeniden ifade etme edilmiş metin karşılaştırması şeklinde değerlendirme yapılmıştır.

Tablo 9.17'de, her aşamanın kosinüs benzerliği ve BERT tabanlı Kesinlik, Duyarlılık ve F1-skoru ortalama değerleri yer almakta; böylece yöntemler arası genel performans farklarının bütüncül bir şekilde analiz edilmesi amaçlanmaktadır.

Tablo 9.17. Yöntemlere Göre Aşamaların Ortalama Performans Karşılaştırması

Yöntem / Aşama	Kosinüs benzerliği (%)	Kesinlik (BERT) (%)	Duyarlılık (BERT) (%)	F1-skoru (BERT) (%)
TextRank – Aşama 1	75,2	63,8	66,5	65,1
TextRank – Aşama 2	80,1	81,2	82,7	81,9
TextRank – Aşama 3	**78,55**	**79,15**	**80,45**	**79,75**
LexRank – Aşama 1	71,2	59,4	61,7	60,5
LexRank – Aşama 2	77	77,85	79,55	78,65
LexRank – Aşama 3	**78,55**	**79,15**	**80,45**	**79,75**
TF-IDF – Aşama 1	70,6	59,7	62	60,8
TF-IDF – Aşama 2	76,25	77,4	79,15	78,25
TF-IDF – Aşama 3	**78,55**	**79,15**	**80,45**	**79,75**
Kural tabanlı – Aşama 1	70,4	66,2	68,1	67,1
Kural tabanlı – Aşama 2	73,75	70,85	72,55	71,75
Kural tabanlı – Aşama 3	**78,55**	**79,15**	**80,45**	**79,75**

Sonuçlar incelendiğinde, TextRank yönteminde Aşama 2 sonuçlarının tüm metriklerde Aşama 3'ün üzerinde olduğu, Aşama 1'in ise en düşük değerlere sahip olduğu görülmektedir. LexRank ve TF-IDF tabanlı özetleme yöntemlerinde ise Aşama 3 en yüksek benzerlik değerlerine ulaşmış, bunu Aşama 2 ve Aşama 1 takip etmiştir. Kural tabanlı yöntemde de benzer bir durum gözlenmiş, Aşama 3 tüm metriklerde en yüksek değerlere ulaşmıştır.

Genel eğilim, doğrudan yeniden ifade etme edilen (Aşama 3) metinlerin, özetleme sonrası yeniden ifade etme edilen (Aşama 2) veya yalnızca özetlenen (Aşama 1) metinlere kıyasla daha yüksek semantik benzerlik skorlarına ulaştığını göstermektedir. Bununla birlikte, TextRank yönteminin istisnai olarak Aşama 2'de daha yüksek sonuçlar vermesi, bu yöntemin yeniden ifade etme sürecinde bilgi bütünlüğünü koruma ve anlamı güçlendirme potansiyeline işaret etmektedir. Bu bulgu, özetleme ve yeniden ifade etme işlemlerinin ardışık biçimde uygulanmasının, bazı algoritmalar için anlam kaybını telafi edebileceğini göstermektedir.

Tablo 9.18'te her özetleme yöntemi için üç aşamanın benzerlik skorlarına göre sıralaması verilmiştir. Görüldüğü üzere yalnızca TextRank yönteminde ikinci aşama (özetleme + yeniden

ifade etme) en yüksek sonucu verirken, diğer tüm yöntemlerde üçüncü aşama (ham metin yeniden ifade etme) en yüksek değeri üretmiştir. Bu durum, özetleme sonrası yapılan ek işlemlerin yöntemler arası farklı etkilere sahip olabileceğini göstermektedir

Tablo 9.18. Yöntemler arası aşama sıralamaları

Yöntem	En Yüksekten En Düşüğe Aşama Sırası	Açıklama
TextRank	2 > 3 > 1	Özetlenip yeniden ifade etme edilen metinler, ham metin yeniden ifade etmelerine kıyasla daha yüksek benzerlik göstermiştir. Ham metin özetleri ise en düşük benzerlik skoruna sahiptir.
LexRank	3 > 2 > 1	Doğrudan ham metin yeniden ifade etmeleri en yüksek skoru elde etmiş, özetleme + yeniden ifade etme ikinci sırada yer almıştır.
TF-IDF	3 > 2 > 1	Ham metin yeniden ifade etmeleri özetleme + yeniden ifade etmelerden daha yüksek skor üretmiştir.
Kural Tabanlı	3 > 2 > 1	Benzer şekilde, doğrudan yeniden ifade etme edilen ham metinler en yüksek değeri göstermiştir.

10. TARTIŞMA ve SONUÇ

Bu çalışmada gerçekleştirilen deneyler sonucunda, öznitelik çıkarımı aşamasında kullanılan BoW ve TF-IDF temsillerinin sınıflandırma başarımı üzerindeki etkisi net bir şekilde gözlemlenmiştir. Yapılan karşılaştırmalar, BoW temsili ile elde edilen sonuçların, TF-IDF'e kıyasla daha yüksek ortalama doğruluk ve F1 skoru sağladığını göstermektedir. Bu durum, kullanılan veri setinin yapısında kelime sıklıklarının bağlamsal ağırlıklardan daha belirleyici olabileceğini düşündürmektedir. Literatürde TF-IDF'in genel olarak daha başarılı olduğu örnekler bulunmakla birlikte, özellikle kısa ve konu odaklı metinlerde BoW'un üstünlük sağlayabileceği yönünde bulgular mevcuttur ([50], [51]).

Makine öğrenmesi algoritmaları incelendiğinde, Lojistik Regresyon modelinin diğer yöntemlere kıyasla daha yüksek doğruluk (%90) ve dengeli F1 skorları elde ettiği görülmüştür. Tek katmanlı yapay zeka modeli doğruluk açısından en düşük sonuçları verirken, SVM ve Rasgele Orman algoritmaları yüksek ancak Lojistik Regresyon'a göre daha dalgalı sonuçlar üretmiştir. Bu bulgular doğrultusunda sonraki aşamalarda Lojistik Regresyon modeli tercih edilmiştir.

Çıkarımsal özetleme yöntemleri arasında LexRank algoritması, genel olarak en yüksek ortalama benzerlik skorlarını sağlamış; TextRank yöntemi ise bazı metinlerde yüksek başarı gösterse de genel istikrar açısından LexRank'in gerisinde kalmıştır. TF-IDF tabanlı özetleme, kısa cümlelerde anlam kaybına daha yatkın bulunmuş; kural tabanlı yaklaşım ise kullanılan anahtar kelime setinin kapsamına bağlı olarak değişken başarılar göstermiştir.

Yeniden ifade etme yöntemlerinin performansı değerlendirildiğinde, Back-Translation yaklaşımının çoğu durumda WordNet tabanlı yönteme kıyasla daha yüksek BERTScore değerleri sağladığı tespit edilmiştir. Bu sonuç, çeviri tabanlı yeniden ifade etme tekniklerinin bağlamı koruma konusunda daha avantajlı olabileceğini göstermektedir ([52], [53]).

Aşamalar arası karşılaştırmalarda yöntemlere göre farklı sıralamalar ortaya çıkmıştır. TextRank algoritmasında, özetlenip yeniden ifade etme ile elde edilen metinlerin ham metinlere benzerliğinin, yalnızca yeniden ifade etme ile elde edilen metinlere göre daha yüksek olduğu; LexRank'te ise yalnızca yeniden ifade etme ile elde edilen metinlerin daha yüksek skor sağladığı

görülmüştür. Bu durum, her özetleme yönteminin bilgi koruma biçiminin ve yeniden ifade etme işleminin bağlam etkisinin farklılık gösterebileceğini ortaya koymaktadır.

Sonuç olarak, çalışma kapsamında elde edilen bulgular, metin işleme sürecinin her aşamasında yöntem seçiminin nihai sonuçlar üzerinde belirleyici bir etkiye sahip olduğunu ortaya koymaktadır. Dolayısıyla, gelecek çalışmalarda veri setinin özelliklerine, metin uzunluğuna ve hedef görevin gerekliliklerine uygun yöntem kombinasyonlarının seçilmesinin kritik olduğu değerlendirilmektedir.

KAYNAKLAR

[1] Y. Liu and M. Lapata, “Text summarization with pretrained encoders,” *EMNLP*, 2019.

[2] M. Zhong et al., “Extractive summarization as text matching,” *ACL*, 2020.

[3] Y.-C. Chen and M. Bansal, “Fast abstractive summarization with reinforce-selected sentence rewriting,” *ACL*, 2018.

[4] M. Lewis et al., “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation,” *ACL*, 2020.

[5] J. Zhang et al., “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” *ICML*, 2020.

[6] C. Raffel et al., “Exploring the limits of transfer learning with a unified text-to-text transformer,” *JMLR*, 2020.

[7] T. Goyal et al., “QAEval: Improving Human Judgment of NLP System Outputs,” *ACL*, 2022.

[8] Zhou, K., Ethayarajh, K., Card, D., & Jurafsky, D. (2022, May). Problems with Cosine as a Measure of Embedding Similarity for High-Frequency Words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics,* (pp. 401-423).

[9] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *EMNLP*, 2019.

[10] T.Sellam, D. Das, and A. Parikh, “BLEURT: Learning Robust Metrics for Text Generation,” *ACL*, 2020.

[11] D. Deutsch et al., “Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary,” *TACL*, 2021.

[12] J. Niyirora, “A Brief Review of Machine Learning for Text,” internal course materials PDF, Schulich School of Business, York University, Toronto, Mar. 2024.

[13] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[14] J. Zhang, Y. Zhao, and M. LeCun, “Character-level convolutional networks for text classification,” NIPS, 2010.

[15] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, “A Survey on Text Classification: From Shallow to Deep Learning,” arXiv preprint arXiv:2008.00364, Aug. 2020. Available: https://arxiv.org/abs/2008.00364 (Accessed: Jul. 8, 2024).

[16] R. Nallapati, B. Zhou, C. Gulcehre, and B. Xiang, “Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond,” CoNLL, 2016.

[17] A. Géron, “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow”, 2nd ed. O’Reilly Media, 2019.

[18] S. B. Kotsiantis, “Supervised machine learning: A review of classification techniques,” *Informatica*, vol. 31, no. 3, pp. 249–268, 2007.

[19] J. Raymaekers, W. Verbeke, and T. Verdonck, “Weight-of-evidence 2.0 with shrinkage and spline-binning,” arXiv preprint arXiv:2101.01494, Jan. 2021. Available: https://arxiv.org/abs/2101.01494 (Accessed: Dec. 14, 2024).

[20] S. Kalra, L. Li, and H. R. Tizhoosh, “Automatic Classification of Pathology Reports using TF-IDF Features,” arXiv preprint arXiv:1903.07406, Mar. 2019. Available: https://arxiv.org/abs/1903.07406 (Accessed: Dec. 14, 2024).

[21] C. Bishop, “Pattern Recognition and Machine Learning”, Springer, 2006.

[22] C. Savas and F. Dovis, “The Impact of Different Kernel Functions on the Performance of Scintillation Detection Based on Support Vector Machines,” *Sensors*, vol. 19, no. 23, art. 5219, Nov. 2019.

[23] V. Gulati, “Extractive Article Summarization Using Integrated TextRank Techniques,” *Electronics*, vol. 12, no. 2, art. 372, 2023.

[24] Abirami, S., and Chitra, P. (2022). “Efficient text classification using enhanced SVM with optimization techniques.” *Multimedia Tools and Applications*, 81(6), 8353–8370.

[25] P. Mahajan, A. Dey, R. Ghosal, and S. Acharya, “Ensemble Learning for Disease Prediction: A Review”, Machine Learning with Applications, vol. 9, Dec. 2023.

[26] Rosenblatt, F. (1958). “The perceptron: A probabilistic model for information storage and organization in the brain.” *Psychological Review*, 65(6), 386–408.

[27] Zhang, Z., and Zhou, J. (2021). “A comparative study of classic and neural classifiers for text classification.” *Neural Computing and Applications*, 33(9), 4487–4502.

[28] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in *Proc. EMNLP*, 2004, pp. 404 411.

[29] Florescu and C. Caragea, "PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents," in *Proc. ACL*, 2017, pp. 1105–1115.

[30] S. Alguliyev, R. Aliguliyev, and F. Ganjaliyeva, "A new approach for abstractive text summarization using deep learning techniques," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 4757–4768, 2019.

[31] Salton, G., and Buckley, C. (1988). "Term-weighting approaches in automatic text retrieval." i*nformation processing & management*, 24(5), 513-523.

[32] Hassanzadeh, H., Keyvanpour, M. R., & Rezaei, A. (2019). "Extractive text summarization using document structure and tf-idf." *Expert Systems with Applications*, 120, 207–219.

[33] Wang, S., and Manning, C. D. (2012). "Baselines and bigrams: Simple, good sentiment and topic classification." *Proceedings of the ACL*, 90–94.

[34] Ganesan, K., Zhai, C., & Han, J. (2010). "Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions." *Proceedings of the 23rd International Conference on Computational Linguistics*, 340–348.

[35] Edmundson, H. P. (1969). "New methods in automatic extracting." *Journal of the ACM (JACM)*, 16(2), 264–285.

[36] Ferilli, S., Esposito, F., & Basile, T. M. A. (2015). "A knowledge-intensive approach to text summarization." *Expert Systems with Applications*, 42(13), 5400–5410.

[37] Moratanch, N., & Chitrakala, S. (2016). "A survey on extractive text summarization.", *International Conference on Circuit, Power and Computing Technologies*, 1–6.

[38] M. M. McCarthy, S. Watanabe, and D. Fujita, "Paraphrase generation: A survey of recent developments," *Computational Linguistics*, vol. 48, no. 2, pp. 235–260, 2022.

[39] V. M. Ngo, T. H. Cao, and T. M. V. Le, "WordNet-Based Information Retrieval Using Common Hypernyms and Combined Features," arXiv preprint arXiv:1807.05574, Jul. 2018. Available: https://arxiv.org/abs/1807.05574 (Accessed: Feb. 22, 2025).

[40] M. Okazaki, Y. Matsuo, and M. Ishizuka, "Improving paraphrase generation with context-aware synonym selection," *Proceedings of IJCNLP*, pp. 67–74, 2017.

[41] R. Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 2, pp. 1–69, 2009.

[42] Sennrich, R., Haddow, B., & Birch, A. (2016). “Improving Neural Machine Translation Models with Monolingual Data.” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 86–96.

[43] Mallinson, J., Sennrich, R., & Lapata, M. (2017). “Paraphrasing Revisited with Neural Machine Translation.” *Proceedings of the 15th Conference of the European Chapter of the ACL (EACL)*, pp. 881–893.

[44] Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). “Understanding Back-Translation at Scale.” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 489–500.

[45] Hu, E. J., Wallis, P., & Spruit, M. (2021). “Augmenting Paraphrase Data Using Back-Translation for Question Answering.” *Journal of Intelligent & Fuzzy Systems*, 40(3), 1–12.

[46] Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., & Le, Q. V. (2020). “Unsupervised Data Augmentation for Consistency Training.” *Advances in Neural Information Processing Systems*, 33, 6256–6268.

[47] Dou, Z.-Y., et al. (2021). “GSum: A General Framework for Guided Neural Abstractive Summarization.” *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 896–911.

[48] Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). “SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation.” *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017*), pp. 1–14.

[49] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). “BERTScore: Evaluating Text Generation with BERT.” *International Conference on Learning Representations* (ICLR).

[50] M. Kirişci, “New cosine similarity and distance measures for Fermatean fuzzy sets,” *Information Sciences*, vol. 634, pp. 21–34, Jan. 2023.

[51] Y. Wang and J. Zhang, “Measurement of Text Similarity: A Survey,” *Information*, vol. 11, no. 9, art. 421, Sep. 2020.

[52] M. Fabbri, Y. Li, S. Sheinin, and D. R. Radev, “SummEval: Re-evaluating Summarization Evaluation,” *Proceedings of the 2021 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human *Language Technologies*, pp. 3910–3922, 2021.

[53] Tursun, A. Yıldız, and F. Yıldız, "BERTurk: A Pretrained Language Model for Turkish," arXiv preprint arXiv:2006.06762, 2020. Available: https://arxiv.org/abs/2006.06762 (Accessed: Apr. 5, 2025).

[54] The New York Times. Available: https://www.nytimes.com/ (Accessed: Jun. 18, 2024).

[55] Reuters. Available: https://www.reuters.com/ (Accessed: Aug. 12, 2024)

[56] ABC. Available: https://abc.com/ (Accessed: Nov. 27, 2024)

[57] BBC. Available: https://www.bbc.com/ (Accessed: Jan. 30, 2025)

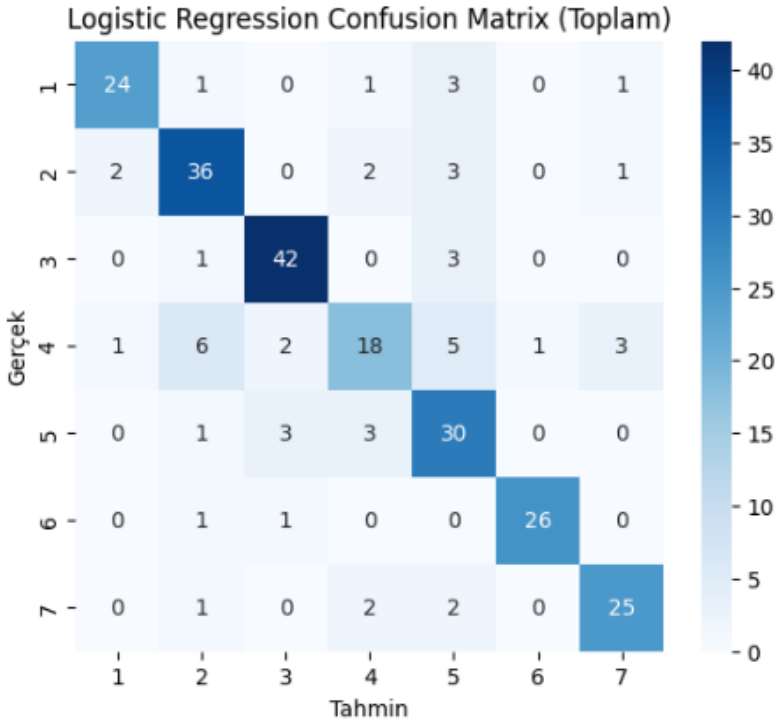
[58] CNN. Available: https://edition.cnn.com/ (Accessed: Mar. 19, 2025).

[59] The Guardian. Available: https://www.theguardian.com/europe (Accessed: Apr. 28, 2025).

EKLER

EK-1: HER MAKİNE ÖĞRENMESİ MODELİ İÇİN KARMAŞIKLIK MATRİSİ, TP, TN, FP, FN DEĞERLERİ

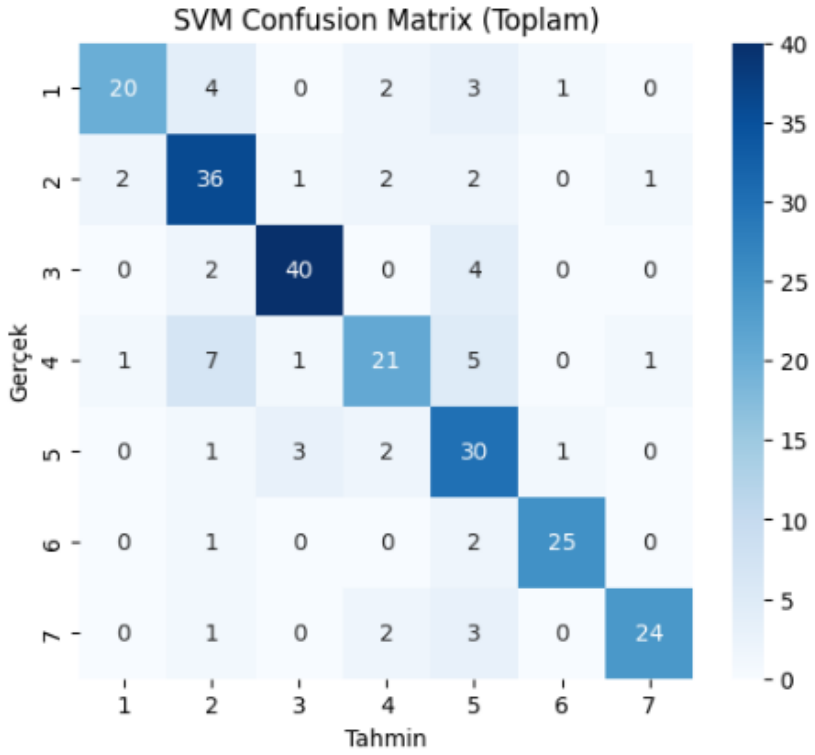
- **Lojistik Regresyon Karmaşıklık Matrisi**



- **Lojistik Regresyon TP, TN, FP, FN Sonuçları**

Sınıf	TP	TN	FP	FN
1	24	218	3	6
2	36	196	11	8
3	42	199	6	4
4	18	207	8	18
5	30	198	16	7
6	26	222	1	2
7	25	216	5	5

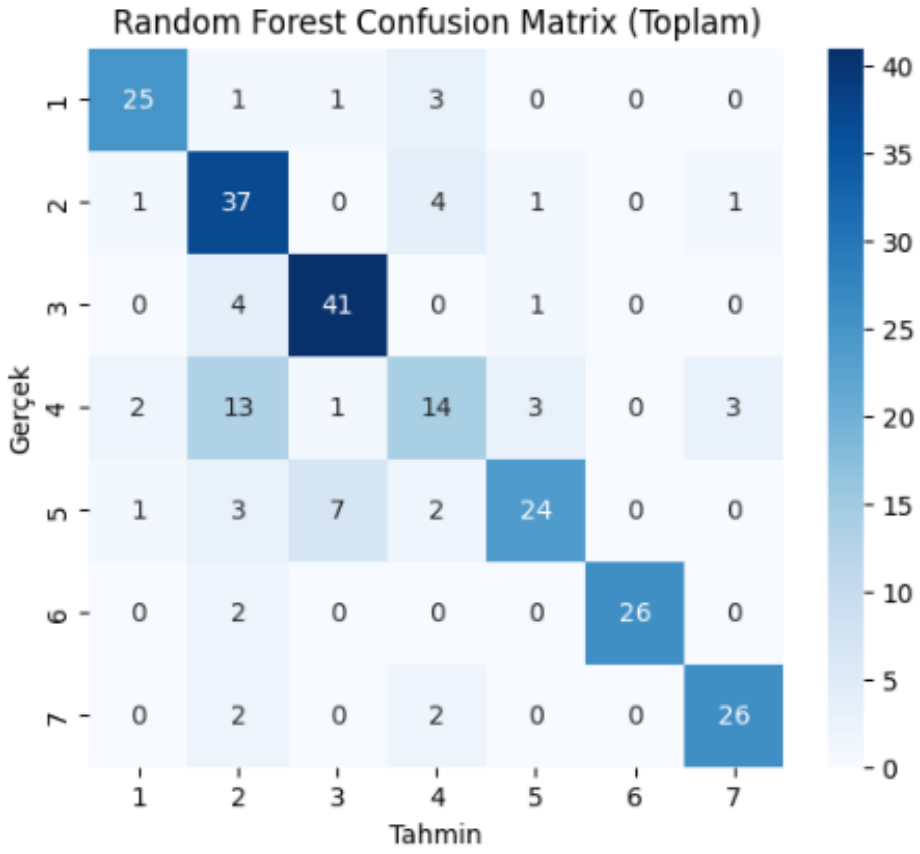
- **Destek Vektör Makineleri Karmaşıklık Matrisi**



- **Destek Vektör Makineleri TP, TN, FP, FN Sonuçları**

Sınıf	TP	TN	FP	FN
1	20	218	3	10
2	36	191	16	8
3	40	200	5	6
4	21	207	8	15
5	30	195	19	7
6	25	221	2	3
7	24	219	2	6

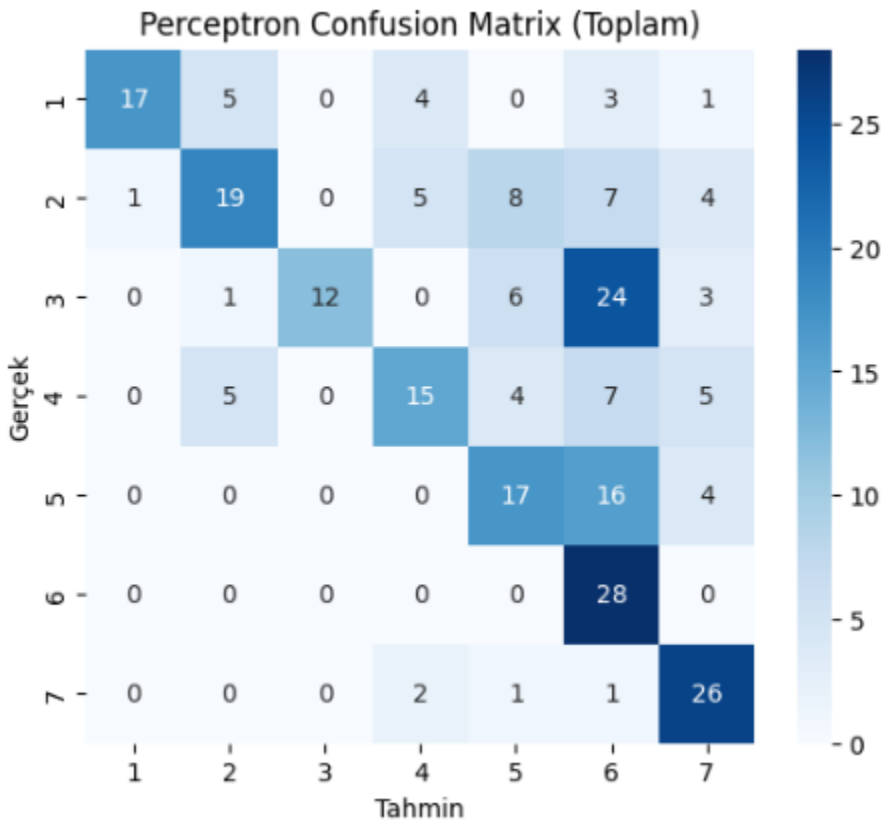
- **Rastgele Orman Karmaşıklık Matrisi**



- **Rastgele Orman TP, TN, FP, FN Sonuçları**

Sınıf	TP	TN	FP	FN
1	25	217	4	5
2	37	182	25	7
3	41	196	9	5
4	14	204	11	22
5	24	209	5	13
6	26	223	0	2
7	26	217	4	4

- **Tek Katmanlı Yapay Sinir Ağ için Karmaşıklık Matrisi**



- **Tek Katmanlı Yapay Sinir Ağ için TP, TN, FP, FN Sonuçları**

Sınıf	TP	TN	FP	FN
1	17	220	1	13
2	19	196	11	25
3	12	205	0	34
4	15	204	11	21
5	17	195	19	20
6	28	165	58	0
7	26	204	17	4