Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 7, S. 8-23, 3, 2023
© Telif hakkı IJANSER'e aittir
Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 7, pp. 8-23, 3, 2023
Copyright © 2023 IJANSER
Research Article

Metin Özetleme Yöntemlerinin İncelenmesi ve Karşılaştırılması

Omar Alwandawybek[1*], Soydan Serttaş[1], Emre Güngör[2]

[1] *Kütahya Dumlupınar Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Kütahya, Türkiye*
[2] *Kütahya Sağlık Bilimleri Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, Kütahya, Türkiye*

[*] *(engomervanli@gmail.com) Başlıca yazarın mail adresi*

(Geliş Tarihi: 28 Mart 2023, Kabul Tarihi: 1 Nisan 2023)

(2nd International Conference on Engineering, Natural and Social Sciences ICENSOS 2023, April 4 - 6, 2023)

ATIF/REFERENCE: Alwandawybek, O., Serttaş, S. & Güngör, E. (2023). Metin Özetleme Yöntemlerinin İncelenmesi ve Karşılaştırılması. *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(3), 8-23.

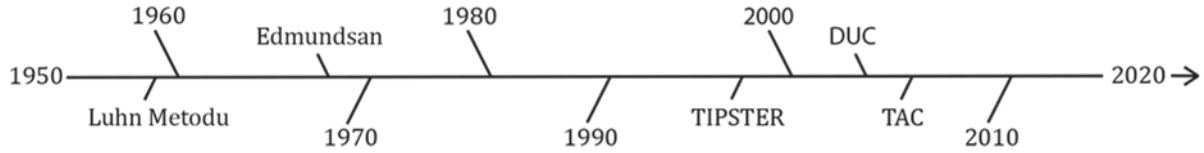
Özet – Günümüz dünyasında verilerin devasa boyutlara ulaşması ve gittikçe karmaşıklaşmasından dolayı verilerin işlenmesi ve yorumlanması zorlaşmaktadır. Bu işlemleri kolaylaştırmak ve veriler içerisinden istenilen bilgilere ulaşmak için bazı teknik ve algoritmaların kullanılması gerekmektedir. Yapay zeka çalışmalarının ilerlemesi doğal dil işleme alanında da yeni algoritmaları beraberinde getirmektedir. Otomatik metin özetleme konusu da bu algoritmaların kullanıldığı önemli alanlardan birisidir. Otomatik olarak metnin özetlenmesi, veri kümesindeki önemli özelliklerin çıkarımı ve yorumlanmasıyla elde edilen kısaltma işlemidir. Günlük konuşma dilinin yanısıra eğitim, hukuk, ticaret, teknik rapor, akademik makale gibi alanlarda oluşturulmuş metinlerin analiz edilmesi bir ihtiyaç haline gelmiştir. Bu çalışmada metin özetleme alanında kullanılan yöntemler ve tekniklerden bahsedilmiştir. Otomatik metin özetleme işleminde kullanılan iki ana yöntem olan çıkarıcı ve yorumlayıcı yöntemlerin yanı sıra; amaca göre incelenen belirtici ve bilgilendirici; belgenin türüne göre tekli ve çoklu doküman; makine öğrenmesi türüne göre denetimli, denetimsiz ve yarı denetimli; sonuç içeriğine göre alana ve konuya; son olarak da sorguya bağlı olup olmamasına göre yöntemler incelenmiştir. Ayrıca belirtilen yöntemler içerisinde kullanılan teknikler ve algoritmalar çalışmada yer almaktadır. Bu yöntemler belli başlıklar altında kategorize edilerek araştırmacıların metin özetleme yöntemleri hakkında detaylı bilgiye erişmeleri amaçlanmıştır. Çalışma otomatik metin özetleme yöntemlerinin avantaj ve dezavantajlarını belirttiğinden, amaca uygun algoritma ve tekniklerin seçimi konusunda araştırmacılara fayda sağlayacaktır.

Anahtar Kelimeler – Doğal Dil İşleme, Metin Özetleme, Özetleme Kategorizasyonu, Çıkarıcı Özetleme, Yorumlayıcı Özetleme

I. GİRİŞ

Bugün internette bulunan verilerin devasa boyutlara ulaşması ve gittikçe daha fazla büyümesinden dolayı bu verileri kullanışlı hale getirmek ve istenilenleri elde etmek için çalışmalar sürdürülmektedir. Veri madenciliği teknikleri, doğal dil işleme ve bu kapsamda bulunan otomatik özetleme sistemlerini kullanarak istenilen özet bilgiye ulaşmak hızlı bir hale gelmektedir. İnsan bilgisayar ilişkisinin giderek yoğunlaşmasından dolayı bu iletişimi kolaylaştırmak için metinden konuşma sentezleme gibi teknikler [1] günümüz ve gelecek için vazgeçilmez bir önem taşımaktadır. Metin özetleme metnin anlamını elde ederek daha kısa bir şekilde ifade etmeye çalışır. Otomatik metin özetleme tarihi 60 yıllık bir dönemi kapsamaktadır. Luhn, geliştirdiği yöntem ile terim sıklığını kullanarak metnin özetlemeye uygun olup

olmadığını belirlemiştir [2]. Bu yöntem bilgiyi taşıyan önemli kelimelerin metinde geçme sıklığına



Şekil 1. Metin özetleme sistemlerin tarihsel Süreci (Historical Process of text summarization systems)

dayalıdır. Edmundson yaptığı çalışmada konu başlıklarının, çalışmanın başlangıç ve son kısımlarında bulunan anahtar kelime ve ifadelerin büyük önem taşıdığını göstermiştir [3]. Otomatik metin özetleme alanında Kupiec ve arkadaşları tarafından yapılan çalışmada yapay zekâ yöntemleri ile birlikte hibrit sistemler çıkarıcı yöntemle istatistiksel bir sınıflandırma problemi olarak ele alınmıştır [4]. Çalışmalarında elle seçilmiş belge alıntıları içeren bir eğitim seti verildiğinde, belirli bir cümlenin bir alıntıda yer alma olasılığını tahmin eden bir sınıflandırma formulü kullanılmaktadır. Sonrasında yeniden çıkarıcı yöntemi kullanarak cümleler olasılığa göre sıralandırılmış ve en çok puan alan cümleler seçilerek özetler oluşturulmuştur. İnternet ağının gelişmesi ve kullanılan verinin artmasıyla birlikte cebirsel yöntemler metin özetleme alanında kullanılmaya ve geliştirilmeye başlanmıştır [5]. Mani ve arkadaşları tarafından yapılan çalışma, Savunma İleri Araştırma Projeleri İdaresi (DARPA) TIPSTER programının bir parçası olarak metin özetleme sistemlerinin değerlendirilmesini sağlayan bir metot olarak ortaya konmuştur. Bu metot üç testten oluşmaktadır. Birincisi, zaman ve doğruluk açısından bir özetleme koşulu testi, ikincisi farklı katılımcı sistemlerinin performansını karşılaştırmak için bir katılımcı teknoloji testi, üçüncüsü de bir tutarlılık testidir [6]. 2000 yılında DARPA tarafından yeni bir özetleme değerlendirme programı başlatılmış olup, Şekil 1'de görüldüğü gibi Belge Anlama Konferansları (DUC) [7] ve Metin Analizi Konferansları (TAC) [8] ile veri setlerinin hazırlanması metin özetleme sistemleri çalışmalarına ivme kazandırmıştır.

Rashmi Mishra ve arkadaşları tarafından yapılan çalışmada biomedikal alanlarda kullanılan karma doğal dil işlemeden bahsedilmiştir. Çalışmada veri madenciliği teknikleri kullanılarak tarama ve soyutlama işlemleri sistematik inceleme standartları uyarınca gerçekleştirilmiştir [9]. Ardından makine öğrenmesi yöntemleri ve derin öğrenme yöntemlerinin hızla gelişmesi ve Gigaword, New York Times, CNN/Daily Mail, NEWSROOM gibi veri tabanlarının ortaya çıkmasıyla metin özetleme yöntemleri gelişmesi hız kazanmıştır.

Metin özetleme sistemleri çıkarıcı özetleme (extractive) ve yorumlayarak özetleme (abstractive) olarak iki ana kısımdan oluşmaktadır. Ayrıca, amaca göre, belgenin türüne göre, makine öğrenmesi türüne göre, sonuç içeriğine göre ve sorguya bağlı olup olmamasına göre yöntemler incelenmiştir. Belirtilen yöntemler içerisinde kullanılan teknik ve algoritmalar da çalışmada yer almaktadır.

Bölüm 2'de metin özetleme çeşitlerinden bahsedilmiştir. Bölüm 3'te ise metin özetleme tekniklerinden bahsedilerek Bölüm 4'te karşılaştırma yapılmıştır. Bölüm 5'te sonuç ve öneriler sunulmuştur.

II. METİN ÖZETLEME ÇEŞİTLERİ (TEXT SUMMARİZATİON TYPES)

Doğal dil işleme (NLP) bölümlerinden biri olan metin özetleme, belgede anlatılan konuyu daha az sayıda kelime ile ifade ederek, konunun fikrini ve anlamını korumak şartıyla uzun olan belgeleri daha kısa hale getirme işlemidir. Başka bir ifade ile belirtilecek olursak, bir veya daha fazla metinden oluşan, orjinal metindeki önemli bilgileri ileten ve orjinal metnin yarısından uzun olmayan metine özetlenmiş metin denilir [10].

Metin özetleme tekniklerinin daha iyi anlaşılabilmesi için farklı kategorilere ayırarak tanımlanması gerekmektedir. Özetleme tekniklerinin birden fazla algoritma içermesi sebebiyle, karşılaştırma yapabilmek için her teknik incelenerek aşağıdaki gibi sınıflandırılmıştır.

A. Otomatik Metin Özetlemede İki Ana Yöntem (Two Main Methods for Automatic Text Summarization)

1) Çıkarıcı Özetleme (Cümle Seçerek Özetleme (Extractive Summarizing))

Çıkarıcı özetleme (cümle seçerek özetleme) yöntemi kelime veya cümlelerin bir alt kümesini seçerek çalışır. Alt küme özetin ne içereceğine karar verir yani daha fazla önem taşıyan cümleleri başlık yöntemi (title method), konum yöntemi (location method), terim sıklığı-ters belge sıklığı (Term Frequency-Inverse Document Frequency (TF-IDF)) ve anahtar kelime işaretleme yöntemi (Cue word method) gibi istatistiksel algoritmalarla metin işlenerek özet sunmaktadır. Var olan cümle yapısı bozulmadan cümleler seçilerek çıkarılmakta ve özet elde edilmektedir [11]. Çıkarıcı metin özetleme dört aşamadan oluşmaktadır:

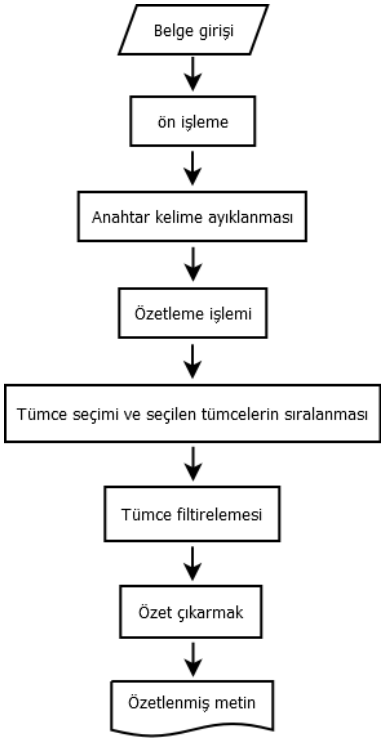
- Metnin ön işlemesi.
- Kelimelerin ve cümlelerin özelliklerinin çıkarılması.
- Cümle seçimi
- Cümlenin yapılması ve özet oluşturma.

Birinci aşamada ön işleme olarak, simgeleştirme (tokenizing), noktalama işaretlerini kaldırma ve sözcük köklendirme gibi özellik çıkarma yöntemleri kullanılır. İkinci aşamada, metindeki cümlelere [0,1] aralığında ağırlık değeri verilerek cümlelerin önemi tanımlanır. Üçüncü aşama olan cümle seçimi ve bağlanması, cümlelerin kelime sayısı açısından sıralanmasıyla oluşmaktadır. En az kelimeden oluşan cümle en yüksek sıraya konulur. En yüksek sıradaki cümleler eşik değerine göre alınmakta ve özet olarak kabul edilmektedir. Dördüncüsü ise özet oluşturma, orijinal belgedeki konum sırasına göre özete eklenen cümlelerdir [11].

Alhashemi'nin Şekil-2'de gösterilen çalışmasında çıkarıcı metin özetleme yöntemini uygulayarak özetlemek istenen metni dört aşamalı bir işleme tabi tuttuktan sonra anahtar kelimeler (insan tarafından belirtilen, girilen ya da seçilen) kelimeleri ayıklamak için aşağıdaki üç adım uygulamıştır [12]:

- Terim sıklığı-ters belge sıklığı
- Belge başlığı ve yazı tipini bulma
- Konuşma yaklaşımının parçasını bulma

Sistem sıkıştırılmış özeti yüksek kalitede sonuç verebilmektedir. Bu çalışmanın kullanıldığı alanlar, web arama motorları, metin sıkıştırma ve kelime işleme uygulamalarıdır.



Şekil 2. Çıkarıcı Metin Özetleme Akış Şeması (Flow Chart of Extractor Text Summarization) [12]

2) *Yorumlayıcı Özetleme – (Soyutlayıcı Özetleme (Abstractive Summarizing))*

Yorumlayıcı metin özetleme yöntemi metni anlamsal olarak özetleyerek insan tarafından yapılan özetlemeye en yakın olan özetlemeyi sunmaya çalışır.

Anlamsal özetleme işlemi, lexical chain, word net, çizge teorisi, kümeleme gibi yöntemler ile metin anlamını bozmadan yeni bir kelimi dizisi üretmektir. Yeni dizi üretmek demek var olan cümle yerine başka bir cümle üreterek bazen orijinal metinde olmayan kelimeleri içeren bir cümle sunmaktır. Şekil 3'te gösterildiği gibi belge girişi yapıldıktan sonra bir ön işlemeye tabi tutulur. Ön işlemede kullanılan sistem metni sistemin çalışacağı şekilde kümeleyerek yapılandırılmamış metni (veriyi) yapılandırılmış bir metin (veri) haline getirir. Bu yapılandırma işlemi için, noktalama işareti kaldırma (Stop word removal), kelime kökünü bulma (Stemming), simgeleştirme (tokenizing), segmentasyon (sentence segmentation) ve kelime frekansı (Word frequency) gibi teknikler (metodlar) kullanılır. Bu metodlar bir araya gelerek yapılandırama işlemi tamamlanmış olur.

Kelimelerin köklerini bulma işlemi yapıldıktan sonra kategorizelendirme işlemi yapılır. Bu işlem kelime kümeleme ya da kullanılan algoritmaya göre

kelime sınıflandırma işlemidir. Öznitelik belirleme işlemi kelimenin yapılan sınıflandırmaya göre belirlenmesi (adlandırılması) ile birlikte kelimeleri veri seti ile karşılaştırarak metni özetlenmesi için hazırlar. Başlık özelliği (Title feature), cümle uzunluğu (Sentence Length), terim ağırlığı (Term Weight), cümle konumu (Sentence Position), cümle cümleye benzerlik oranı (Sentence to Sentence Similarity), özel isim (Proper Noun), ve nesnel kelime (Thematic Word) gibi yöntemlerle öznitelik belirleme işlemi yapılır.

Şekil 3. Yorumlayıcı Metin Özetleme Akış Şeması (Interpretive Text Summarization Flow Chart)

Yorumlayıcı özetlemeye örnek verilecek olursa, "Annem yeni ocak, fırın ve davlumbaz aldı." ifadesi "Annem mutfak malzemeleri aldı." şeklinde özetlenmektedir. Bu tarz bir özetleme için geniş kelime dağarcığı gereklidir. Bu tarz yorumlayıcı oluşturmak için derin öğrenme teknikleri kullanılabilmektedir. ERHANDI ve ÇALLI, derin oto-kodlayıcılar ve LSTM ağlarını metin özetleme işlemi için kullanmışlardır [13].

Yorumlayıcı özetler, çıkarıcı özete göre daha düzgün olabilir ancak daha karmaşık bir sisteme sahiptir. Yorumlayıcı ve çıkarıcı özetleme, özet oluşturmak için ya istatistiksel ya da dilbilimsel yaklaşımları ya da her ikisinin birleşimini kullanmaktadır. Çıkarıcı özet oluşturmak daha kolaydır, ancak yorumlayıcı teknikle metin özetleme daha güçlüdür çünkü anlamsal olarak ilişkili ve üretilmesi zor olan özetler üretebilmektedir [14].

B. *Amaca Göre metin Özetleme (Summary Of Text By Purpose)*

1) Belirtici özetleme (Indicative summarizing)

Belirtici özetlemede hedef metnin içerisinde bulunan tekrar cümleler veya konu adının geçtiği cümleleri karşılaştırarak, konu adının en fazla geçtiği cümleleri ayırarak bir özet oluşturmaktır. Başka bir tanımla belirtici özetleme ana metinden konunun önemli sayılan kısımları bir araya getirilerek oluşturulan özetlerdir. Ayrıca ana metinde yer alan tüm bilgi özette yer almayabilir [15]. Belirtici özetleme belgeler arasındaki farkları vurgulayabilir ve araştırmacının uygun belgeyi bulmasına yardımcı olabilir. Belirtici özetin temel amacı, makale ya da özeti istenilen bir metnin içeriği hakkında ayrıntı vermeden makalenin içeriğini göstermektir. Kullanıcıya tam formu aktarmaya çalışır. Belirtici özetleme çoklu belge özetlerken belgeler arasındaki farklılıklar kısaca bir film fragmanı gibi bilgiyi kısaca tek belge halinde özet olarak sunar [16].

2) Bilgilendirici özetleme (Informative summarizing)

Belirtici özetlemeden daha detaylı bir şekilde özet üreterek okuyucu özeti okuduğunda ana metinde geçen bilgilerin çoğuyla ilgili bilgi elde edecektir. Fakat belirtici özetlere göre daha uzun olmaktadır. Bilgilendirici özetleme kaynak metnin içerisindeki tüm bilgiyi aktarmaya çalışarak bilimsel çalışmalarda tercih edilmektedir [15]. Bilgilendirici bir özet, bir tarayıcının genel bilgi ihtiyaçlarını karşılayan belgeler arasındaki ortaklıklardan sentezlenen bir özet şeklinde olabilir. Bilgilendirici bir özet, orijinal belgeyi temsil etmek (ve genellikle değiştirmek) içindir. Belirtici özetlemeden en önemli farkı bilgilendirici özetleme çoklu belge özetlerken belgeler arasındaki ortak bilgileri daha net ortaya koyar [16].

C. *Belge türüne göre metin özetleme (Summarizing text by document type)*

1) Tek belgeli özetleme (Single document summarizing)

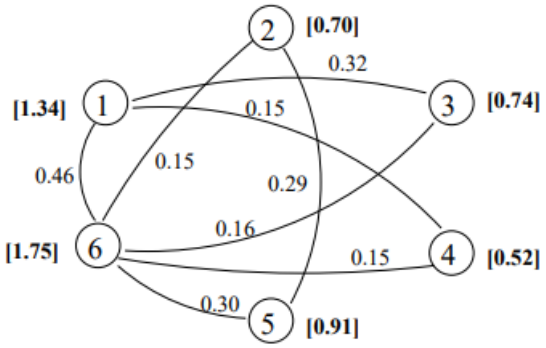
Tek belgeli özetleme işlemi metindeki cümleleri metnin genel olarak anlaşılması için önemlerine göre sıralar. Mihalcea ve Tarau çalışmalarında dilden bağımsız grafik tabanlı tek belgeli özetlemede metindeki her cümle için bir tepe noktası eklenerek bir grafik oluşturmuşlardır. Bu

grafiği oluşturmak için köşeler arasındaki kenarlar ve cümle ara bağlantıları kullanılır. Bu bağlantılar, içerik örtüşmesinin (benzerliğin) bir fonksiyonu olarak ölçüldüğü bir benzerlik ilişkisi kullanılarak tanımlanır. İki cümle arasındaki böyle bir ilişki "tavsiye" süreci olarak görülebilir. Bu tavsiye süreci metinde belirli kavramlara hitap eden bir cümle, okuyucuya metinde aynı kavramlara hitap eden diğer cümlelere atıfta bulunması için bir "tavsiye" verir. İki cümlenin örtüşmesi basitçe, sözcüksel temsilleri arasındaki ortak belirteçlerin sayısı olarak belirtilebilir. Tavsiye belirli bir sözdizimsel kategorisindeki sözcükleri sayan sözdizimsel filtrelerden geçirilir. Ayrıca, uzun cümleleri tercih etmekten kaçınmak için bir normalleştirme faktörü kullanılır. Bu normalleştirme faktörü, iki cümlenin içerik çakışmasının her bir cümlenin uzunluğuna bölünmesidir. Ortaya çıkan grafikte, metindeki çeşitli cümle çiftleri arasındaki bağlantıların gücünü gösteren kenarlar ağırlık değerleriyle orantılıdır. Grafik şu şekilde temsil edilebilir:

- basit yönsüz grafik.
- bir cümleden metinde takip eden cümlelere (ileriye yönelik) ayarlanan kenarların oryantasyonu ile yönlendirilmiş ağırlıklı grafik.
- bir cümleden metindeki önceki cümlelere (geriye doğru yönlendirilmiş) ayarlanan kenarların yönelimi ile yönlendirilmiş ağırlıklı grafik.

Sıralama algoritması grafik üzerinde çalıştırıldıktan sonra, cümleler puanlarına göre ters sıralanır ve en üst sıradaki cümleler tekli çıkarımsal özete dahil edilmek üzere seçilir.

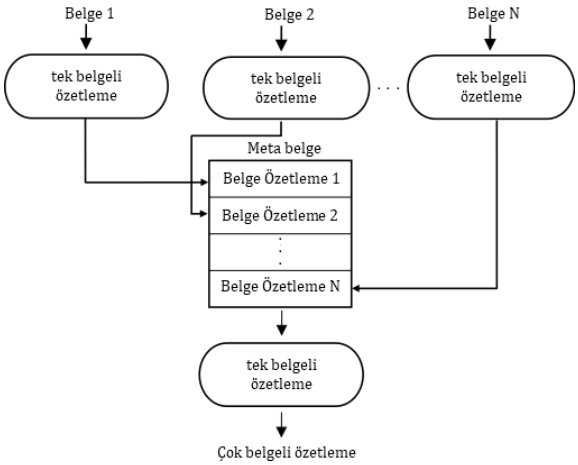
Şekil 4. altı cümlelik örnek bir metin için oluşturulmuş ağırlıklı bir grafik örneğini göstermektedir [17]. Cümlelerin benzerliklerini karşılaştırmak için belirli kavramlar kullanmışlardır. Uzun cümleleri seçmekten kaçınmak için bir normalleştirme işlemi uygulanmıştır. Tek doküman içerisinde bulunan cümleleri sıralama algoritmasını tersten uygulayarak en çok puan alan cümleyi özette en birinci yerleşerek tekrarlamadan kaçınmışlardır ve tekli çıkarıcı metin özetlemesini sunmuşlardır.



Şekil 4. Örnek bir metin üzerine kurulmuş cümle benzerliklerinin grafiği. Cümle önemini yansıtan puanlar, her cümlenin yanında parantez içinde gösterilmiştir

2) *Çok belgeli özetleme (Multi document summarizing)*

Çok belgeli özetleme bir meta özetleme prosedürü kullanılarak oluşturulur. İlk olarak, belirli bir belge kümesindeki her belge için, grafik tabanlı sıralama algoritmalarından (güvercin yuvası, kova, sayarak sıralama) biri seçilerek tek bir belge özeti oluşturulur. Ardından, aynı veya farklı bir sıralama algoritması kullanılarak bir özetin özeti üretilir. Şekil 5, bir N belge kümesiyle başlayan bir çoklu belge özeti oluşturmak için kullanılan meta özetleme sürecini göstermektedir.

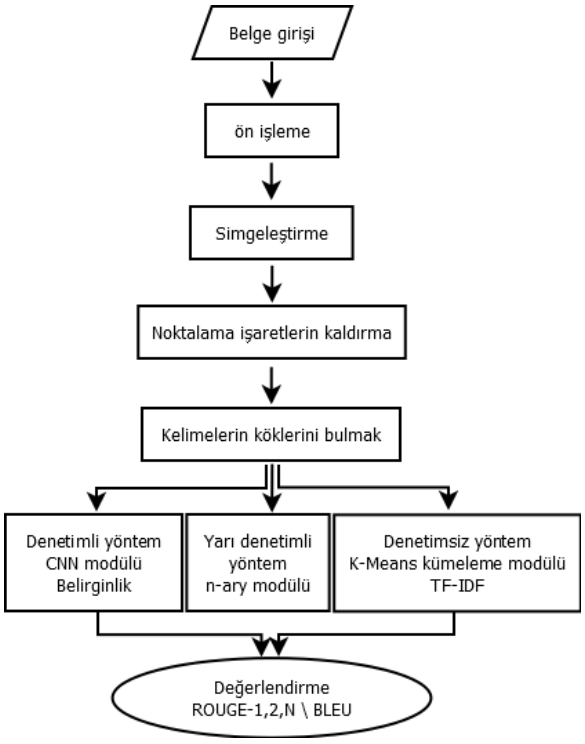


Şekil 5. Meta-özetleme kullanarak çok belgeli özetleme

Çok benzer içeriğe sahip cümlelerde tek belgeli özetlemenin aksine benzerlik oranına göre eşik değer belirlenir. Bu eşik değeri tekrar cümlelerden kaçınmak ve özetin aktardığı bilgi miktarını azaltmak için kullanılır. Grafik oluşturma aşamasında, benzerliği bu eşiği aşan cümleler (köşeler) arasına bağlantı (kenar) eklenmemektedir. Mihalcea ve Tarau çalışmalarında bir meta sistemi oluşturarak incelemişlerdir [17].

D. *Makine öğrenmeli yöntemler (Machine learning methods)*

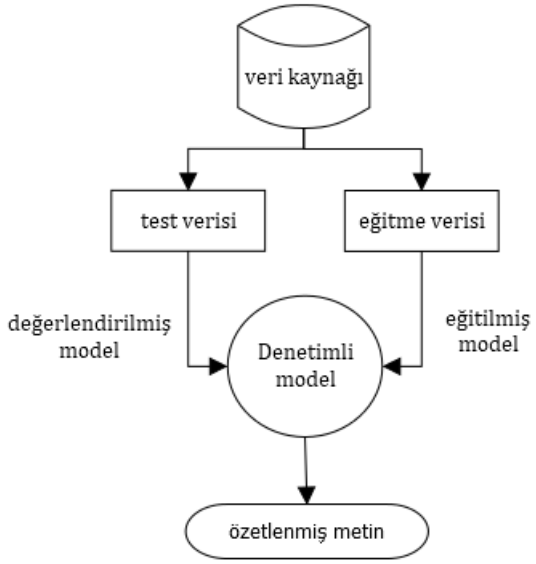
Yapay zekanın gelişmesi ile birlikte makine öğrenmesi kullanan metin özetleme yöntemleri gelişmeye başlamıştır. Makine öğrenmeli yöntemde önceden bahsi geçen sistemler gibi belge girişi yapıldıktan sonra ön işleme yapılır. Simgeleştirme, noktalama işaretlerini kaldırma, kelime köklerini bulma gibi işlemler yapıldıktan sonra makine öğrenmeli yöntemler (denetimli, denetimsiz ve yarı denetimli) uygulanarak özet sunulur. Çıkan özet ile insan tarafından yapılan özet karşılaştırarak ROUGE, BLEU gibi yöntemlerle başarım oranlarını kıyaslanır. Şekil 6' da gösterildiği gibi özetleme aşamaları kısaca gösterilmiştir.



Şekil 6. Denetimli, Denetimsiz ve Yarı Denetimli Özetleme Sistemlerin Akış Çizelgesi (Flowchart of Supervised, Unsupervised and Semi-Supervised Summarization Systems)

1) *Denetimli metin özetleme (Supervised text summarizing)*

Bu tür özetleme tekniğinde özetleme anahtar kelime çıkarma ya da cümle benzerliğini bulma gibi yöntemler kullanılır. Denetimli sistem eğiterek daha özelleştirilmiş bir sistem haline getirilebilir. Algoritmayı eğitme işlemi çok fazla eğitim verisi gerektirebilir. Denetimli yöntemler tipik olarak bu sorunu bir ikili sınıflandırma görevi olarak yeniden biçimlendirmektedir. Ancak bir model belirli bir ifadenin anahtar sözcük olup olmadığını belirlemek için açıklamalı veriler üzerinde eğitilir. Fazla veri ile eğitilmesine rağmen bazen istenilen sonuçları vermeyebilir [18]. Denetimli teknikte anahtar kelime çıkarma (keyphrase extraction), araştırma dokümanındaki bir cümlenin anahtar kelime olup olmadığını tahmin ederek bir sınıflandırma işlemine tabi tutmaktadır [19].



Şekil 7. Denetimli özetleme sisteminin akış şeması

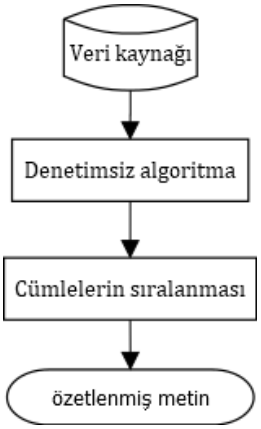
Şekil 7'de görüldüğü gibi Girdi veri kaynağı(seti), eğitim ve test veri seti olarak ikiye ayrılır. Eğitme (traning) veri kümesi, tahmin edilmesi veya sınıflandırılması gereken çıktı değişkenine sahiptir. Tüm modeller, eğitim veri kümesine göre eğitilerek tahmin veya sınıflandırma için test veri kümesine uygulanır [20].

2) *Denetimsiz metin özeleme (Unsupervised text summarizing)*

Denetimsiz metin özetleme sistemde kullanılan algoritmalar denetimsiz algoritmalardır. Eğitim verisi gerektirmeden özetleme işlemi yapılır. Denetimsiz özetlemenin amacı verinin modeline ve sunumuna göre bilgileri belirlemektir. Denetimsiz yöntemler denetimli yöntemlere kıyasla genellikle daha karmaşık matematiksel işlemleri barındırabilir. Denetimsiz yöntemlerin denetimli yöntemlere göre daha popüler olmasının ana sebebi denetimsiz algoritmaların eğitim gerektirmemesidir. Denetimsiz özetlemede anahtar kelime çıkarma (keyphrase extraction), çizge tabanlı sıralama (graph-based ranking) ve merkezilik ölçütleri yöntemleri iyi performans sergileyebilir. Thushara ve arkadaşları yaptıkları çalışmada anahtar kelime ve anahtar sözcükleri ayıklamak için TF-IDF

değerini kullanmışlardır [21]. Şekil 8'de denetimsiz metin özetleme sistemi açıklanmıştır.

Şekil 8'de gösterildiği gibi denetimsiz öğrenme algoritması uygulandıktan sonra, orijinal metin sıralamasına göre özeti barındıran cümleler sıralanır. Yeni veriler tanıtıldığında, verilerin sınıfını tanımak için önceden özellikleri bilinen veriler kullanılır. Denetimsiz sistemlerin denetimli sistemlerden ana farkı önceden işlenilmiş verilerin özellikleri keşfetmede kullanılmasıdır [21].



Şekil 8. Denetimsiz metin özetleme akış şeması

3) Yarı denetimli özetleme (Semi supervised summarizing)

Bu yöntem denetimli yöntemin eğitim verisi gerektirmesi, denetimsiz yöntemin de istenildiğinde müdahale edilememesi gibi dezavantajlarından dolayı ortaya çıkmıştır. Denetimli ve denetimsiz iki yöntemin özelliklerini barındıran özetleme yarı denetimli metin özetleme olarak adlandırılmıştır. Örnek vermek istersek D. Li ve arkadaşları çalışmalarında anlamsal (semantic) ağ hiper çizge tabanlı yarı denetimli (semi-supervised) sistemini önermişlerdir [22]. Çalışmalarında özetleme için elde edilen ifadeler arasında ikili ve çoklu (n-ary) ilişkiler hesaplanmıştır. Çizgedeki köşeler ifadeleri temsil ederken ağırlıklı hiper kenarlar anlamsal ilişkiyi temsil etmektedir. Algoritma anahtar kelimeyi bulmak yerine anahtar cümle bularak özetleme yapmaktadır. Hiper tümce köşe (hyper-edge) ağırlık değerini formül 1' den hesaplayabiliriz [22]:

$$w(e) = \frac{a}{|e|} \sum_{e_{ij} \subseteq e} w(e_{ij}) \qquad (1)$$

E. Sonuç içeriğine göre metin özetleme (Summarizing text by result content)

1) Alana yönelik metin özetleme (Domain summirizing)

Bir alanla ilgili birden fazla dokümandan oluşturulan özet türüdür. Bir alana yönelik sonuç getiren özetlerdir [15]. Farklı metin türlerini kapsar ve her türlü kullanıcı tarafından kullanılabilir. Bu tür sistemleri tasarlamaktaki amaç, özel bir alanla ilgili gerekli bilgileri içeren özet oluşturmaktır. Örneğin, tibbi ya da mühendislik alanında bilimsel ifadeleri içeren metinleri özetleme gibi özel özetleme gerektiren sistemlerde kullanılır [23].

2) Konuya ya da türe göre metin özetleme (Genre specific summirizing)

Konuya göre özetleme, sadece belirli bir konuya yönelik olarak yapılan özetleme çalışmalarıdır. Metnin içerisinde kullanıcının belirttiği konu çerçevesinde özet çıkarılır [15]. Sistem yalnızca gazete makaleleri, hikayeler, kılavuzlar vb. gibi özel metin türlerini girdi olarak kabul eder. Bu girdiler kullanılarak farklı şablon yapısına sahip olan metinlerin özeti çıkartılır [23].

F. Sorguya göre metin özetleme (Summarize text by query)

1) Sorgu tabanlı metin özetleme (Query based summarizing)

Sorgu tabanlı metin özetleme, bir kelime ya da bir soru ile bağdaşan metin özetleme türüdür. Yapılan sorgu ile uyuşan özetlerin oluşturulmasıdır. Genellikle arama motorlarında kullanılan özetleme biçimidir [15]. Sorgu tabanlı özetleme işleminden önce kullanıcının orijinal metnin konusunu bir sorgu biçiminde belirlemesi gerekir. Kullanıcı metin hakkında genel bilgilere sahip olduğundan dolayı genellikle bir sorunun cevabı olan belirli bilgileri arar. Böylece kullanıcı o özel bilgiyi sorgu şeklinde sorarak sistem sadece o bilgiyi metinden çıkarıp özet olarak sunar [23].

2) Genelleştirilmiş metin özetleme (Generalized summarizing)

Genelleştirilmiş metin özetleme, herhangi bir sorguya bağlı olmayan genel metin özetlemesidir [15]. Genelleştirilmiş metin özetleme sistemlerinde, kullanıcının metni önceden anlamadığı ve özetin farklı kullanıcılar tarafından kullanılabileceği

varsayılır. Bu nedenle tüm bilgiler aynı önem düzeyindedir [23].

III. METİN ÖZETLEMEDE KULLANILAN TEKNİKLER (TECHNİQUES USED İN TEXT SUMMARİZİNG)

A. *İstatiksel yaklaşımlar (Statistical approaches)*

İstatistiksel yaklaşımlar cümlenin başlık, yer ve terim sıklığı gibi istatistiksel özelliklerini kullanarak anahtar kelimelere ağırlık verir. Anahtar kelime ağırlıklarına göre cümlenin puanını hesaplar. En yüksek puan alan cümleyi özet için seçerek özet sunar [14].

1) *Terim frekansı-Ters belge frekansı (Term Frequency-Inverse document frequency)*

Terim frekansı-Ters belge frekansı (TF-IDF) ifadesi genellikle bilgi alma ve metin madenciliğinde bir ağırlıklandırma faktörü olarak kullanılır. Bir kelimenin koleksiyondaki veya derlemedeki bir belge için ne kadar önemli olduğunu yansıtan sayısal bir istatistiktir. TF-IDF, büyük ölçüde metin özetleme ve sınıflandırma uygulamasında kelimeleri filtrelemeyi durdurmak için kullanılır [24].

TF-IDF değeri bir kelimenin belgede görünme sayısıyla orantılı olarak artar. Bazı kelimelerin diğerlerinden daha yaygın olduğunu kontrol etmeye yardımcı olur ve tümcedeki kelimenin sıklığı ile dengelenir. TF-IDF ağırlıklandırma faktörü genellikle arama motorlarında bir kullanıcı tarafından istenilen belgenin alaka düzeyini puanlama ve sıralamada kullanılır [14].

Terim frekansı , ters belge frekansı ve TF-IDF değerini 2,3 ve 4 nolu formüllerden hesaplayabiliriz [11]:

$$TF = \frac{Toplam\ belgede\ bulunan\ kelimelerin\ görünümü}{Bir\ belgedeki\ toplam\ kelime\ sayısı} \quad (2)$$

$$IDF = log\frac{Toplam\ belge\ sayısı}{belge\ frekansı} \quad (3)$$

$$TF - IDF = TF \times IDF \quad (4)$$

2) *Anahtar kelime işaretleme yöntemi (Cue word method)*

Anahtar kelime işaretleme yönteminde bir kelime önemine göre ağırlıklandırılır. İşaretleme pozitif (olumlu) veya negatif (olumsuz) olarak iki farklı şekilde değer alabilir. Pozitif ve negatif gibi ipuçları içeren cümle özete dahil edilir. Anahtar kelime işaretleme yöntemi önemli cümleleri tanımlamak için "retorik" bir bağlam sağladığı varsayımına dayanır. Anahtar kelime işaretleme yöntemi kaynak soyutlama, dizi işaret ifadesi içeren cümleleri seçme işleminden oluşmaktadır [14].

3) *Başlık yöntemi (Title method)*

Bu yöntem başlıkta geçen cümlelerin daha önemli olarak kabul edildiğini ve özetlemeye dahil edilme olasılığının daha yüksek olduğunu belirtmektedir. Cümlelerin puanlaması cümle ile başlık arasında kaç kelime kullanıldığına göre hesaplanmaktadır. Belge herhangi bir başlık bilgisi içermiyorsa başlık yöntemi etkili olamayabilir [14].

4) *Konuma göre özetleme (Location method)*

Konuma göre özetlemede belgenin sonuç ya da giriş gibi konumsal olarak kelimenin görünüp görünmediğine göre ağırlık değerleri belirlenir. Belgedeki ilk ve son cümlelerinin daha önemli olduğu varsayılarak özetleme işlemi yapılmaktadır [14].

B. *Dilbilimsel yaklaşım (linguistic approach)*

Dilbilimsel yaklaşım dilin bilimsel olarak semantik ve pragmatik incelenmesidir. Semantik kavramı kelimelerden ve kavramlardan anlamın nasıl çıkarıldığını gösterir. Pragmatik kavramı ise anlamın bağlamdan nasıl çıkarıldığını içerir. Dilbilimsel yaklaşımlar kelimeler arasındaki bağlantıyı göz önünde bulundurur. Bu yaklaşım kelimeleri analiz ederek ana kavramı bulmaya çalışır. Dilbilimsel yaklaşım anlamsal işlemeyi içeren dilsel yöntemlere dayanarak yorumlayıcı metin özetleme işlemlerinde kullanılır [14].

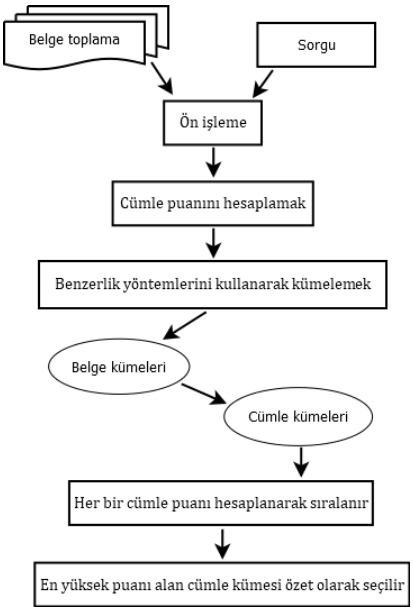
1) *Kümeleme yaklaşımı (Clustering)*

Kümeleme yaklaşımı, benzer verileri veya cümleleri gruplandırarak belge özetlemek için kullanılan bir yöntemdir. Özetleme cümle özelliklerine bağlı olmakla birlikte aynı zamanda cümle benzerlik oranına da bağlıdır. Örnek olarak Zhang ve Li çalışmalarında kelime formu ilişkisini cümleler arasındaki benzerlik oranını belirlemek için kullanmıştır [25]. Ayrıca kelime sırası ilişkisini kullanarak iki cümle arasındaki sıra benzerliğini tanımlamışlardır. Kümeleme işlemi için K-means algoritmasını seçmişlerdir.

Şekil 9'da gösterilen Kümeleme yaklaşımı içerisinde belgeler ve kelime sayısı belirleme sorguları kullanıcı tarafından parametre olarak girilir. Bu parametreler ön işleme öncesinde belirlenir. Ön işleme yapıldıktan sonra kümeleme

işlemi uygulanarak belgeler guruplandırılır ve cümle kümeleri oluşturulur. Oluşturulan cümle kümelerinden en yüksek puan alan küme en iyi cümle kümesi seçilerek özeti oluşturur. Deshpande ve Lobo çalışmalarında kosinus benzerlik yöntemini kullanarak sorgu tabanlı bir kümeleme algoritmasını önermişlerdir [26]. Bu kümeleme sisteminde vektör uzayı kullanılarak kullanıcı tarafından girilen sorgulara benzer olan cümleler bulunur.

Şekil 9'da genel bir kümeleme sistemi kullanılarak belgelerde bulunan cümlelerin kullanıcı sorgusu ile benzerliklerinin hesaplanması yapılır. Benzerlik değerlerine göre cümleler guruplandırılır. Cümle puanlaması kelime sıklığı ve cümle konumu özelliği kullanılarak hesaplanır. Bu puanlama yapıldıktan sonra en yüksek puana sahip olan cümle özet olarak seçilir ve özetleme işlemi tamamlanır.



Şekil 9. Kümeleme yaklaşımı akış şeması

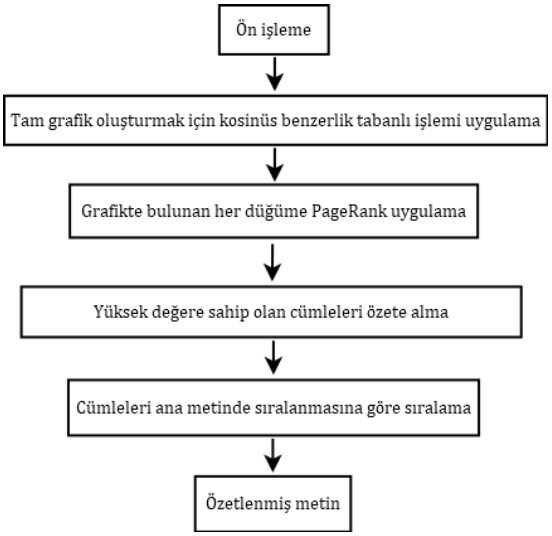
2) Çizge teorisi (Graph theory)

Bu yöntem algoritmalar vasıtasıyla metin özetleme için bir çizge oluşturur. Çizgede bulunan köşe noktasına göre puanlama yapılarak cümle seçilir. Çizge teorisi diğer yöntemlerin ortaya çıkaramadığı özellikleri ortaya çıkararak verilerden daha verimli faydalanma sağlamaktadır. TextRank ve ClusterRank algoritmaları çizge teorisinin metin özetleme alanında kullanımını daha detaylı göstermek için incelenmiştir.

a. TextRank algoritması (TextRank algorithm)

Çizge tabanlı TextRank algoritması iki cümle arasındaki benzerlik ilişkisini belirler. Denetimsiz bir algoritma olarak belgedeki en önemli anahtar kelimeleri seçer. Algoritma özel olarak oluşturulmuş bir çizge üzerinde PageRank algoritmasının varyasyonunu uygular. Çizgedeki öğeleri sıralayarak en önemli öğeler metni daha iyi tanımlayan öğeler olarak belirlenir. Bu yaklaşım, TextRank algoritmasının farklı dillerde kullanılmasını sağlar [27]. Şekil 10'da gösterilen TextRank algoritması her cümlenin çizgede bir düğümle temsil edildiği çizge tabanlı cümle çıkarma algoritmasıdır. Sözcüksel benzerliğe dayalı (lexical benzerlik) olarak iki cümle arasında yönsüz bir kenar oluşturulur. Örnek olarak eğer bir cümle Si bir dizi kelime olarak şu şekilde $S_i = W_1^i, W_2^i \ldots . W_{|S_i|}^i$ temsil ediliyorsa, iki cümle Si ve Sj arasındaki benzerlik formül (5)' ki gibi tanımlanır [28]:

$$Sim(S_i, S_j) = \frac{\left|\{w_k : w_k \in S_i \wedge w_k \in S_j\}\right|}{log(|S_i|) + log(|S_j|)} \quad (5)$$

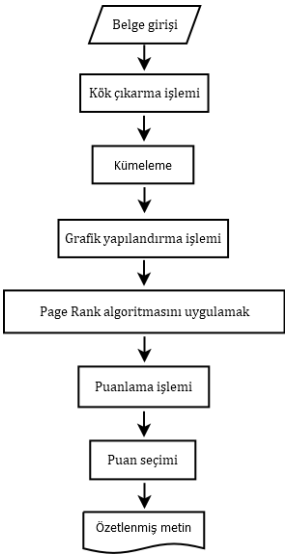


Şekil 10. TextRank algoritması akış şeması

Şekil 10'da gösterildiği gibi TextRank algoritmasını uygulamak için diğer algoritmalarda da olduğu gibi bir ön işleme yapılması gerekmektedir. Bu ön işleme çoklu belgeleri tekli belge haline getirerek metin olmayan içerikleri eleyerek sadece metin olan belgelere algoritmayı uygulamak üzere hazırlamaktır. Ardından oluşturulan çizgede bulunan düğümler üzerine PageRank algoritmasını uygular. En yüksek ilişki değerine sahip cümleler seçilerek özet oluşturulur. Bu özet oluşturulurken özet uzunluk kısıtlarına göre seçilen cümle sayısı belirlenir. Cümleleri orjinal metinde sıralanmasına göre düzenledikten sonra özetleme tamamlanmış olur.

b. ClusterRank algoritması (ClusterRank algorithm)

ClusterRank algoritması konuşma yazılarının çıkarımsal olarak özetlenmesi için tasarlanmış denetimsiz çizge tabanlı yöntemdir [28]. Ön işleme aşamasından sonra Porter'ın steming algoritması uygulanmıştır [29]. ClusterRank algoritması yine grafik tabanlı bir yöntem olan ve haber makalelerinden cümle çıkarmak için kullanılan TextRank algoritmanın uzantısıdır. Bu algoritmada ön işleme yapılarak noktalama işaretleri kaldırdıktan sonra Şekil 11'de gösterildiği gibi çalışır. Belge girişi ve kök çıkarma işlemi yapıldıktan sonra bir kümeleme işlemi yapılır. ClusterRank önce metni bir çizge düğümleri olarak temsil edilen kümelere ayırır. Daha sonra komşu kümelerin tüm cümleleri arasındaki benzerlik ölçülür. En yüksek benzerliğe sahip cümleler tek bir kümede birleştirilir. Ardından yüksek ilişki değerine sahip küme içerisindeki cümleler merkez tabanlı yaklaşımla puanlanır. Cümlelerin artıklık oranı yüksekse hatalı biçimlendirilmiş cümleler ölçülür. Belirli bir eşik değeri üzerinde benzerlik puanı olan cümleler seçilerek pencere boyutu eşiğine göre özete dahil edilir.

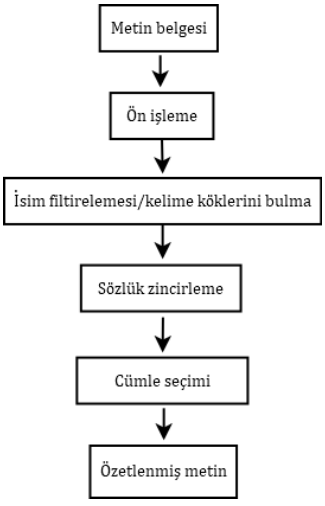


Şekil 11. ClusterRank algoritması akış şeması

3) Sözlük zinciri yaklaşımı (Lexical chain)

Sözlük zinciri yaklaşımı, sözcüklerin art arda bir araya gelerek anlamsal olarak bir bütün olmasına dayanır. Anlamsal bütünlük sözcüklerin tutarlılığının bir özelliğidir. Sözcüklerin tutarlılığı kavramı Halliday ve Hasan tarafından 1976 yılındaki çalışmalarında incelenmiştir [30]. Bu kavram 1991 yılında Jane ve Graeme [31] yaptıkları çalışmada metnin farklı bölümlerini bir bütün olarak işlev görecek şekilde, metinleri birbirine yapıştırmak için kullanılan bir araç olarak sunmuşlardır. Dilbilgisel uyumu (referans, ekleme, eksiltme ve bağdaşma), sözcüksel uyum (anlamsal olarak ilişkili kelimeler) kullanılarak elde edilir. Sözcük bağdaşıklığı, iki terim arasında ve ilgili sözcük dizileri arasında oluşur.

Sözcük zinciri, çoklu belgede de anlamsal olarak ilişkili kelime kümelerini birleştirerek çalışmaktadır. Hipernim/hiponim kavramları kelimelerin aynı sözlük zincirinde gruplandırılmasını sağlayan ilişkilerdir. Sözlük zinciri yaklaşımı, belgelerde bulunan gramer hatalarını düzeltmek için kullanılmaktadır. Sözlük zincirini işlemek için isimler ilişkilerine göre kümelenir. Her ismin bir gurup içerisinde yer alması gerekmektedir. Kelimeler en güçlü ve en uzun sözlük zincirini oluşturacak şekilde gruplandırılır [14].



Şekil 12. Sözlük Zinciri yaklaşımı akış şeması

Meru Brunn ve arkadaşları yaptıkları çalışmada Şekil 12'de gösterildiği gibi sözlük zinciri tabanlı bir sistem tasarlamışlardır [32]. Bu sistem 5 aşamadan oluşmaktadır. Birincisi, girilen metin belgelerinin bir ön işlemeden geçirildiği aşamadır. Ön işleme bölücü, etiketleyici ve ayrıştırıcı işlemlerinden oluşmaktadır. Bölücü cümleleri ifadeleri, işaretleri ve metinde olan farklı karekterleri birbirinden ayırarak etiketli kelimeler toplanır ve sözdizimsel yapılarına göre düzenlenir. Ardından isim filtreleme işlemi gerçekleştirilir. Bu aşamada ayrıştırılan isimler seçilerek ve kaldırılarak metin özetlemenin doğruluğu arttırılır. Bu isimler kaynak metinde etiketleyici tarafından belirlenen isimlerdir. Bu işlemin sebebi metinde tekrar eden

kelimeleri elemek ve özeti negatif bir şekilde etkilenmesinin önüne geçmektir. Sözlük zinciri (Lexical chainer) metnin farklı bölümlerini(segmentlerini) bir bütün olarak işlev görecek şekilde birbirine yapıştırmak için kullanılır. Sözlük zinciri dilbilgisel uyum ve sözcüksel uyum kullanılarak elde edilir. Cümle seçimi aşaması bölümlerin seçimi ve cümlenin seçimi olarak iki adımdan oluşmaktadır. Cümle bölümlerinin seçimi metin parçalarını seçmeyi amaçlar. Bölüm seçimi, bölümlerin formül 5'teki şekilde puan hesaplanmasına dayanır [32].

$$score(seg_j) = \sum_{i=1}^{m} \frac{score(chain\ M\ ember_i, seg_j)}{s_i} \quad (6)$$

Cümle çıkarma işlemi için en yüksek puanlara sahip ilk N bölüm(segment) seçilir. Cümle seçimi adımında her cümle toplanan sözcüksel bağdaşıklık puanlarının toplam sayısına göre sıralanır. Sıralama sürecinin amacı, her puanın önemini değerlendirmek ve tüm puanları her cümle için bir sıralamada birleştirmektir. Bu değerlendirmeyi gerçekleştirirken cümlelerin sözcüksel olarak uyumlu olması için gereken minimum bağlantı sayısını belirten bir eşik değer belirlenir. Bu prosedür uyarınca eşik değere göre cümleyle ilişkili sözcük bağdaşıklığı puanları toplanarak cümle seçimi gerçekleşir. Son olarak cümleler bir arada sıralanarak özet sunulmuş olur.

4) Kelime ağı yaklaşımı (Word Net)

Word Net İngilizce dili için kullanılabilen çevrimiçi bir sözlük veritabanıdır. İngilizce kelimeleri sys-nets adı verilen eşanlamlı kümeler halinde gruplandırılmaktadır. Ayrıca Word Net her bir sistem ağının kısa bir anlamını ve her bir sistem ağı arasındaki anlamsal ilişkiyi sağlamaktadır. Aynı zamanda birçok sistem tarafından kelimeler arasındaki ilişkiyi belirlemek için kullanılan sözlük zinciri ve çevrimiçi sözlük işlevi görmektedir. Eş anlamlı sözlüğü anlam benzerliğine göre gruplandırılmış kelimelerin bir listesini içeren kaynaktır . Sözcükler arasındaki anlamsal ilişkiler eşanlamlı kümeler ve eş anlamlı ağaçları ile temsil edilmektedir. Sözcük ağları bu ilişkilere göre sözcük zincirlerini oluşturmak için kullanılmaktadır. Word Net 118.000'den fazla farklı kelime formu içermektedir. Örnek olarak LexSum sözcük zincirini oluşturmak için Word Net kullanılmaktadır [14]. Meru Brunn ve arkadaşları da önceden bahsi geçen tasarladıkları sistemde Word Net'i kullanmışlardır [32].

IV. METİN ÖZETLEME YÖNTEMLERİNİN KARŞILAŞTIRILMASI (COMPARİSON OF TEXT SUMMARY METHODS)

Aşağıdaki tablolarda otomatik metin özetleme yöntemlerinin karşılaştırılması yapılmıştır.

Tablo 1. Çıkarıcı özetleme

Çıkarıcı	
Anlamı	İstatiksel özelliklere dayalı olarak önemli cümlelerin seçilmesinden oluşur
Avantajları	Semantik ifadelerle uğraşmadığı için hesaplaması kolay ve başarılıdır
Dezavantajları	Uzun metinlerde çok başarılı olmayabilir
Kullanım alanı	Kısa metin özetlemek içi idealdir
Örnek çalışma	Text Summarization Extraction System (TSES) Using Extracted Keywords [12]

Tablo 2. Yorumlayıcı özetleme

Yorumlayıcı	
Anlamı	Metni anlamsal olarak değiştirmeden yeni cümlelerle özetlemektir
Avantajları	Bilgi sıkıştırma oranının yüksek olmasından dolayı daha kısa özetler sunabilmesi
Dezavantajları	Uzun işlem süresi gerektirmektedir
Kullanım alanı	Haber özetleme gibi uzun metin özetlemesi için kullanılır
Örnek çalışma	Derin Öğrenme ile Metin Özetleme [13].

Tablo 3. Belirtici özetleme

Belirtici özetleme	
Anlamı	Metnin yalnızca ana fikrini kullanıcıya sunarak metnin okumaya değer olup olmadığına hızlıca karar vermek için kullanılabilir
Avantajları	Okuyucunun işine yarayan çalışmayı hızlıca keşfetmesini sağlayarak zaman kazandırır
Dezavantajları	Metnin ayrıntılarını sunmaz
Kullanım alanı	Hızlı sınıflandırma işlemlerinde kullanılır
Örnek çalışma	Applying Natural Language Generation to Indicative Summarization [16]

Tablo 4. Bilgilendirici özetleme

Bilgilendirici özetleme	
Anlamı	Ana metinden kısa bilgiler sunar
Avantajları	Ana metinden bir yedek oluşturmaya çalışır
Dezavantajları	Bazen önemli cümleleri kısaltarak hızlı genel bakış sağlamaz
Kullanım alanı	Daha önce belli olan konularda hızlıca kısa metinler sunmak için kullanılır
Örnek çalışma	Generating Indicative-Informative Summaries with SumUM [33]

Tablo 5. Tek belgeli özetleme

Tek belgeli özetleme	
Anlamı	Bir belge girişi yapılarak özetler
Avantajları	Donanımsal ve yazılımsal açıdan düşük maliyetlidir
Dezavantajları	Birden fazla belge özetlemez
Kullanım alanı	Kes-Yapıştır Metin Özetleme sistemlerinin yapısında
Örnek çalışma	Kes-Yapıştır Metin Özetleme [34]

Tablo 6. Çok belgeli özetleme

Çok belgeli özetleme	
Anlamı	Birden fazla belge girişi yapılabilir
Avantajları	Aynı konuyu içeren birden fazla belgeyi tek belgede özetleyebilir
Dezavantajları	Genellikle benzer cümlelerin oranı yüksek olması
Kullanım alanı	Birden fazla dökümanın işlenmesi gerektiği durumlarda
Örnek çalışma	Tek ve Çoklu Belge Özetleme için Dilden Bağımsız Bir Algoritma [17]

Tablo 7. Denetimli özetleme

Denetimli özetleme	
Anlamı	Belirli alana yönelik bilgi girişi yapılıp özet sunabilir
Avantajları	Uzmanlık alanına göre yönlendirilebilir
Dezavantajları	Eğitim verisi gerektirir. Bu da bazen maliyetli olabilir
Kullanım alanı	Web sistemlerinde
Örnek çalışma	Metin Özetleme için Denetimli Öğrenmeyi Geliştirmek için Etiketlenmemiş Verilerin Kullanımı [35]

Tablo 8. Denetimsiz özetleme

Denetimsiz özetleme	
Anlamı	Sisteme müdahele yapılmadan özetler
Avantajları	Eğitim verisi gerektirmez
Dezavantajları	Sisteme müdahele yapılmadığından dolayı istenilmeyen şekilde sonuçlanabilir
Kullanım alanı	Web sistemlerinde
Örnek çalışma	Denetimsiz derin öğrenmeyi kullanarak iki dilli otomatik metin özetleme [36]

Tablo 9. Yarı denetimli özetleme

Yarı denetimli özetleme	
Anlamı	Denetimli ve denetimsiz yöntemlerin özelliklerini barındırır
Avantajları	Uygulanması denetimli sistemlere göre daha kolaydır
Dezavantajları	Kullanıcıdan değer girişi gerektirmesi
Kullanım alanı	Web sistemlerinde
Örnek çalışma	Katibeh: Yeni yarı denetimli yaklaşımı kullanan farsça dili için bir haber özetleyicisi [37]

Tablo 10. Alana yönelik özetleme

Alana yönelik metin özetleme	
Anlamı	Sabit konulu alanda tanımlanabilecek metni özetler
Avantajları	Dayandığı bir özel alana bağımlı olarak çalışır
Dezavantajları	Belgenin konusu ile sınırlı sabit alana bağlıdır
Kullanım alanı	Patent(marka) arama sistemlerinde
Örnek çalışma	Using Genre-Specific Features for Patent Summaries [38]

Tablo 11. Konuya ya da türe göre özetleme

Konuya yada Türe göre özetleme	
Anlamı	Yalnızca özel metinler kabul eder
Avantajları	Farklı formatlarda belgeleri özetleme yeteneğine sahiptir
Dezavantajları	Belge uzunluğu konusunda sınırlıdır
Kullanım alanı	Belirli bir konuya yönelik (öykü, masal, haber vb.) gibi özetleme sistemlerinde
Örnek çalışma	Kısa Öyküleri Özetleme [39]

Tablo 12. Sorgu tabanlı özetleme

Sorgu Tabanlı özetleme	
Anlamı	Kullanıcı metnin konusunu sorgu şeklinde belirleyerek sistem sadece bu bilgiyi çıkarmaktadır
Avantajları	Belirlenilen bilgiler aranabilir ve kullanıcının ilgisini yansıtabilir
Dezavantajları	Veri setine bağımlı olması
Kullanım alanı	Arama motorları
Örnek çalışma	Yapay zeka teknikleri aracılığıyla Metin özetleme kullanarak uygun arama motoru tasarımı [40]

Tablo 13. Genelleştirilmiş özetleme

Genelleştirilmiş özetleme	
Anlamı	Kullanıcı türünden bağımsız olarak genelleştirilmiş özet sunar
Avantajları	Tüm bilgiler aynı önem düzeyindedir
Dezavantajları	Kullanıcı görüşü yerine yazarın görüşüne bağlı kalır
Kullanım alanı	Şikayetlerden bilgi çıkarma, adaletle ilgili kanuni alanlarda, resmi belgelerde
Örnek çalışma	Şikayetlerden bilgi çıkararak Chatbot tabanlı suç kaydı ve farkındalığı oluşturma modeli [41]

V. SONUÇ VE ÖNERİLER (CONCLUSION AND RECOMMENDATIONS)

Metin özetleme yöntemleri sürekli geliştiğinden dolayı sınıflandırma ihtiyacı oluşmaktadır. Çalışmada metin özetleme yöntemleri analiz edilerek kategorilere ayrılmıştır. Böylece geliştirilecek yeni yöntemler için kategorizelendirmede bir kaynak olarak sunulmuştur. Aynı zamanda metin özetleme alanına yönelen araştırmacıların çalışmalarını hızlandırarak yeni sistemlerin tasarlanabilmesine yol açacaktır.

Metin özetleme sistemlerinde çözüm bekleyen problemler bulunmaktadır. Örneğin Türkçe dili gibi bazı diller sondan eklemeli bir dil olduğundan dolayı özetlemede farklı zorluluklar bulunmaktadır.

Doğal dil işleme ve metin özetleme alanlarında Türkçe dili özelinde çalışmaların sayısının az olduğu gözlemlenmektedir. Bu çalışma Türkçe dili özelindeki özetleme sistemlerinin geliştirilmesi için de faydalı bir kaynak olacaktır.

KAYNAKLAR

[1] G. Yılmaz and M. Yağcı, "Türkçe Metinden Konuşma Sentezlemeye Yönelik Yapılan Çalışmaların İncelenmesi", *Mühendislik Bilimleri ve Tasarım Dergisi,* pp. 286-296, 2022.

[2] H. P. Luhn, "The Automatic Creation of Literature Abstracts", *IBM Journal, vol.2(2), pp. 159-165,* 1958.

[3] H. P. EDMUNDSON, "New Methods in Automatic Extracting", *the Association for Computing Machinery,* vol. 16, pp. 264-285, April 1969.

[4] J. Kupiec, J. pedersen and F. Chen, "A Trainable Document Summarizer", Dans les actes de 18th annual international ACM SIGIR conference on Research and development in information retrieval, *Xerox Palo Alto Research Cente,* 1995, pp. 68–73.

[5] K. Ježek and J. Steinberger, "Automatic Text Summarization (The state of the art 2007 and new challenges,)", *FIIT STU Bratislava, Ústav informatiky a softvérového inžinierstva,* In *Proceedings of Znalosti,* 2008, pp. 1-12.

[6] I. Mani, D. House, G. Klein, L. Hirschman, T. Firmin, & B. M. Sundheim, "The TIPSTER SUMMAC text summarization evaluation", In *Ninth Conference of the European Chapter of the Association for Computational Linguistics, 1999,* p. 77-85.

[7] P. Over, H. Dang, & D. Harman, "DUC in context", *Information Processing & Management*, vol. *43*(6), pp. 1506-1520, 2007.

[8] H. Ji, R. Grishman, H. T. Dang, K. Griffitt, & J. Ellis, "Overview of the TAC 2010 knowledge base population track", In *Third text analysis conference (TAC 2010) 2010, November, v*ol. 3(2), p. 3.

[9] R. Mishra, J. Bian, M. Fiszman, C. R. Weir, Jonnalagadda, S., Mostafa, J., & Del Fiol, G., "Text summarization in the biomedical domain: a systematic review of recent research", *Journal of biomedical informatics*, vol. *52*, pp. 457-467, 2014.

[10] D. Radev, E. Hovy, & K. McKeown, "Introduction to the special issue on summarization", *Computational linguistics*, vol. *28*(4), pp. 399-408, 2002.

[11] H. Christian, M. Pramodana Agus And D. Suhartono, "Single Document Automatic Text Summarization Using Term Frequency-Inverse Document Frequency (Tf-Idf)", İn *Comtech*, Indonesia, vol. 7(4), pp. 285-294, 2016.

[12] R. ALhashemi, "Text Summarization Extraction System (TSES)Using Extracted Keywords", *International Arab Journal of e-Technology, vol. 1(4),* pp. 164-168, 2010.

[13] B. Erhandı and F. Çallı, "Text Summarization with Deep Learning", in *3rd International Conference on Data Science and Applications (ICONDATA'20),* istanbul, 2020.

[14] N. Munot and S. S. Govilkar, "Comparative Study of Text Summarization Methods", *International Journal of Computer Applications, vol. 102(12),* pp. 33-37, 2014.

[15] C. Özkan, "*İnternet tabanlı Türkçe metinler için otomatik özetleme tekniği* ", Master's thesis, Maltepe Üniversitesi, Fen Bilimleri Enstitüsü, Turkey, 2019.

[16] J. L. Klavans, M. Y. Kan, & K. McKeown, "Domain-specific informative and indicative summarization for information retrieval", 2001

[17] R. Mihalcea, & P. Tarau, "A language independent algorithm for single and multiple document summarization", In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts, 2005.*

[18] K. S. Hasan, & V. Ng, "Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art", In *Coling 2010: Posters,* pp. 365-373, 2010.

[19] K. S. Hasan, & V. Ng, "Automatic keyphrase extraction: A survey of the state of the art", In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics vol. 1: Long Papers,* pp. 1262-1273, 2014.

[20] B. Mahesh, "Machine Learning Algorithms - A Review", *International Journal of Science and Research (IJSR), vol. 9,* pp. 381-386, 2018.

[21] M. G. Thushara, T. Mownika and R. Mangamuru, "A Comparative Study on different Keyword Extraction Algorithms", In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019*, pp. 969-973.

[22] D. Li, S. Li, W. Li, W. Wang and W. Qu, "A Semi-Supervised Key Phrase Extraction Approach: Learning from Title Phrases through a Document Semantic Network", in *Proceedings of the ACL 2010 Conference Short Papers*, Sweden, 2010, pp. 296-300.

[23] S. Gholamrezazadeh, M. A. Salehi, & B. Gholamzadeh, "A comprehensive survey on text summarization systems", *2nd International Conference on Computer Science and its Applications,* IEEE, 2009, pp. 1-6.

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

[25] P. Y. Zhang, & C. H. Li, "Automatic text summarization based on sentences clustering and extraction", In *2009 2nd IEEE international conference on computer science and information technology, IEEE, 2009,* pp. 167-170.

[26] A. R. Deshpande & L. M. R. J. Lobo, "Text summarization using clustering technique", *International Journal of Engineering Trends and Technology*, vol. *4*(8), pp. 3348-3351, 2013.

[27] R. Mihalcea, "Graph-based ranking algorithms for sentence extraction, applied to text summarization", In *Proceedings of the ACL interactive poster and demonstration sessions,* pp. 170-173, 2004.

[28] N. Garg, B. Favre, K. Reidhammer and D. Hakkani Tür, "Clusterrank: A Graph Based Method For Meetıng Summarızatıon", idiap reserch inistiute, 2009.

[29] M. Porter, 1980 "The Porter Stemming Algorithm", Accessible. [Online]. Available: http://www.tartarus.org/martin/PorterStemmer

[30] M. A. Halliday & R. Hasan, (1976), "Cohesion in English", Bath, England.

[31] J. Morris, & G. Hirst, "Lexical cohesion computed by thesaural relations as an indicator of the structure of text, "*Computational linguistics*, vol. *17*(1), pp. 21-48, 1991.

[32] M. Brunn, Y. Chali & C. J. Pinchak, "Text summarization using lexical chains", In *Proc. of Document Understanding Conference, 2001,* p. 29.

[33] H. Saggion, & G. Lapalme, Generating indicative-informative summaries with sumum. *Computational linguistics*, vol. *28*(4), pp. 497-526, 2002.

[34] H. Jing, "Cut-and-paste text summarization" ,*Columbia University*, 2002.

[35] M. R. Amini & P. Gallinari "The use of unlabeled data to improve supervised learning for text summarization", In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, 2002,* pp. 105-112.

[36] S. P. Singh, A. Kumar, A. Mangal & S. Singhal, "Bilingual automatic text summarization using unsupervised deep learning", In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), IEEE, 2016,* pp. 1195-1200.

[37] S. Farzi & S. Kianian, "Katibeh: A Persian news summarizer using the novel semi-supervised approach", *Digital Scholarship in the Humanities*, vol. *34*(2), pp. 277-289, 2019.

[38] J. Codina, N. Bouayad-Agha, A. Burga, G. Casamayor, S. Mille, A. Muller, H. Saggion and L. Wanner, "Using Genre-Specific Features for Patent Summaries", *Information Processing & Management*, vol. *53*(1), pp. 151-174, 2016.

[39] A. Kazantseva & S. Szpakowicz, "Summarizing short stories", *Computational Linguistics*, vol. *36*(1), pp. 71-109, 2010.

[40] K. Sekaran, P. Chandana, J. R. V. Jeny, M. N. Meqdad & S. Kadry, "Design of optimal search engine using text summarization through artificial intelligence techniques", *TELKOMNIKA (Telecommunication Computing Electronics and Control),* vol. *18*(3), pp. 1268-1274, 2020.

[41] S. Surana, J. Chekkala and P. Bihani, "Chatbot based Crime Registration and Crime Awareness System using a custom Named Entity Recognition Model for Extracting Information from Complaints", *International Research Journal of Engineering and Technology (IRJET),* vol. 7.529, pp. 3329-3336, 2021.