T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ÇIKARIMSAL ARAPÇA METİN ÖZETLEME: GENETİK ALGORİTMALAR İLE ÖZNİTELİK OPTİMİZASYONUNA DAYALI HİBRİT YAKLAŞIM

YÜKSEK LİSANS TEZİ

Hedil ELŞAVİ

Bilişim Sistemleri Mühendisliği Anabilim Dalı

EYLÜL 2025

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ÇIKARIMSAL ARAPÇA METİN ÖZETLEME: GENETİK ALGORİTMALAR İLE ÖZNİTELİK OPTİMİZASYONUNA DAYALI HİBRİT YAKLAŞIM

YÜKSEK LİSANS TEZİ

Hedil ELŞAVİ

Bilişim Sistemleri Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr.Öğr.Üyesi Tuğrul TAŞCI

EYLÜL 2025

Hedil ELŞAVİ tarafından hazırlanan "Çıkarımsal Arapça Metin Özetleme: Genetik Algoritmalar ile Öznitelik Optimizasyonuna Dayalı Hibrit Yaklaşım" adlı tez çalışması 02.09.2025 tarihinde aşağıdaki jüri tarafından oy birliği ile Sakarya Üniversitesi Fen Bilimleri Enstitüsü Bilişim Sistemleri Mühendisliği Anabilim Dalı'nda Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Jürisi

Jüri Başkanı :	**Doç.Dr. Selman HIZAL** Sakarya Ugulamalı Bilimler Üniversitesi	.....
Jüri Üyesi :	**Dr.Öğr. Üyesi Tuğrul TAŞÇI** (Danışman) Sakarya Üniversitesi	.....
Jüri Üyesi :	**Dr.Öğr.Üyesi Muhammed KOTAN** Sakarya Üniversitesi	.....

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Sakarya Üniversitesi Fen Bilimleri Enstitüsü Lisansüstü Eğitim-Öğretim Yönetmeliğine ve Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesine uygun olarak hazırlamış olduğum "Çıkarımsal Arapça Metin Özetleme: Genetik Algoritmalar ile Öznitelik Optimizasyonuna Dayalı Hibrit Yaklaşım" başlıklı tezin bana ait, özgün bir çalışma olduğunu; çalışmamın tüm aşamalarında yukarıda belirtilen yönetmelik ve yönergeye uygun davrandığımı, tezin içerdiği yenilik ve sonuçları başka bir yerden almadığımı, tezde kullandığım eserleri usulüne göre kaynak olarak gösterdiğimi, bu tezi başka bir bilim kuruluna akademik amaç ve unvan almak amacıyla vermediğimi ve 30.06.2026 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince Sakarya Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Enstitü tarafından belirlenmiş ölçütlere uygun rapor alındığını, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun ortaya çıkması halinde doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi beyan ederim.

(02/09/2025).

Hedil ELŞAVİ

TEŞEKKÜR

Bu tez çalışmasının yürütülmesi sürecinde bilgi ve tecrübesiyle beni yönlendiren, akademik desteğini ve sabrını esirgemeyen değerli danışmanım Tuğrul TAŞCI'ya en içten teşekkürlerimi sunarım.

Ayrıca, bu süreçte bana manevi ve bilimsel olarak büyük destek sağlayan, sistem mühendisliği alanında doktorasını tamamlamak üzere olan nişanlıma teşekkür ederim. Varlığı, motivasyonumu artırmış ve araştırma yolculuğumu daha anlamlı kılmıştır.

En derin şükranlarımı ise her zaman yanımda olan ve beni koşulsuz destekleyen aileme sunmak isterim. Özellikle sevgili anneme, her anımda gösterdiği sabır, sevgi ve desteğiyle bu tezin en büyük ilham kaynaklarından biri olmuştur.

Bu süreçte katkıda bulunan ve yanımda olan herkese gönülden teşekkür ederim.

Hedil ELŞAVİ

İÇİNDEKİLER

Sayfa

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ ..... **v**
TEŞEKKÜR ..... **vii**
İÇİNDEKİLER ..... **ix**
KISALTMALAR ..... **xi**
TABLO LİSTESİ ..... **xiii**
ŞEKİL LİSTESİ ..... **xv**
ÖZET ..... **xvii**
SUMMARY ..... **xix**
1. GİRİŞ **1**
1.1. Çalışmanın Gerekçesi 2
1.2. Çalışmanın Amacı 3
1.3. Çalışmanın Kapsamı ve Sınırlılıkları 4
1.4. Tezin Organizasyonu 4
2. OTOMATİK METİN ÖZETLEMEDE TEMEL KAVRAM VE YÖNTEMLER **7**
2.1. Metin Özetleme Yaklaşımlarının Kategorizasyonu 7
2.1.1. Dil-tabanlı Yaklaşımlar 7
2.1.2. Yöntem-tabanlı Yaklaşımlar 8
2.1.3. Kaynak-tabanlı Yaklaşımlar 9
2.1.4. Hedef-tabanlı Yaklaşımlar 9
2.1.5. İçerik-tabanlı Yaklaşımlar 9
2.1.6. Bağlam-tabanlı Yaklaşımlar 10
2.2. Metin Özetlemede Güncel Uygulamalar 10
2.3. Metin Özetlemede Arapça Dilinin Zorlukları 11
2.4. Değerlendirme Metrikleri 12
3. LİTERATÜR ARAŞTIRMASI **17**
3.1. Arapça Metin Özetleme Alanında Kullanılan Veri Setleri 17
3.2. Arapça Metin Özetleme Yaklaşımları 18
3.2.1. İstatistiksel yöntemler 18
3.2.2. Semantik-tabanlı yöntemler 19
3.2.3. Makine öğrenmesi tabanlı yöntemler 20
3.2.4. Graf- tabanlı yöntemler 21
3.2.5. Bulanık mantık tabanlı yöntemler 23
3.2.6. Meta-sezgisel yöntemler 24
3.3. Genetik Algoritma Tabanlı Özetleme Yöntemleri 25
3.4. Arapça ATS Yaklaşımlarının Karşılaştırmalı Değerlendirmesi 27
4. GENETİK ALGORİTMA TABANLI HİBRİT ARAPÇA METİN ÖZETLEME **29**
4.1. Arapça Metinler için Ön İşleme Adımları 31
4.1.1. Belgelerin içe aktarılması 32

4.1.2. Hareke ve noktalama işaretlerinin kaldırılması 32
4.1.3. Normalizasyon 32
4.1.4. Tokenizasyon ve stopword temizleme 33
4.1.5. Stemming ve lemmatization 35
4.2. Öznitelik Çıkarma Aşaması 36
4.2.1. Cümlenin paragraftaki konumu (Fsp) 37
4.2.2. Cümledeki sayısal veriler (Fnd) 37
4.2.3. Cümle uzunluğu (Fls) 38
4.2.4. Cümledeki anahtar kelimeler (Fkw) 38
4.2.5. Cümle benzerliği (Fss) 39
4.2.6. Cümlede geçen özel tabirler (Fne) 39
4.2.7. Cümlede geçen İngilizce-Arapça kelimeler (Feh) 39
4.2.8. Terim frekansı ve ters cümle frekansı 40
4.3. Özniteliklerin Vektör Gösterimi 41
4.4. Skor Hesaplama ve Cümlelerin Sıralanması 41
4.5. Ağırlık Optimizasyonu Yaklaşımı: Rastgele Arama 42
4.6. Skor Hesaplama ve Sıralama 44
4.7. Ağırlık Optimizasyonu 44
4.8. Cümle Seçiminde Genetik Algoritmaların Kullanım Süreci 44
5. BULGULAR VE DEĞERLENDİRME 47
5.1. Veri Seti 47
5.2. Gerçekleme Ortamı 48
5.3. Deneysel Senaryolar ve Uygulama Sonuçları 48
5.3.1. Senaryo A: benzerlik yöntemlerinin genetik algoritmadaki etkisi 49
5.3.2. Senaryo B: hibrit modelin performansının ölçülmesi 50
5.3.2.1. Senaryo B-1 Sıkıştırma oranının model performansına etkisi 51
5.3.2.2. Senaryo B-2: Genetik algoritma parametrelerinin etkisi 52
5.3.3. Senaryo C : Deneyler ve Karşılaştırmalı Değerlendirme 53
5.4. Değerlendirme 54
6. SONUÇ VE ÖNERİLER 55
KAYNAKLAR 57
EKLER 65
ÖZGEÇMİŞ 73

KISALTMALAR

ALPHA : Elitizm Katsayısı
F_eh : İngilizce-Arapça Terim Oranı
F_kw : Anahtar Kelime Yoğunluğu
F_ls : Cümle Uzunluğu Özelliği
F_nd : Sayısal Bilgi İçeriği
F_ne : Özel Varlık Varlığı
F_sp : Cümle Konumu Özelliği
F_ss : Cümleler Arası Anlamsal Benzerlik
F_ti : TF-ISF Tabanlı Terim Ağırlığı
GA : Genetik Algoritma
GEN : Nesil Sayısı
K : Turnuva Seçim Parametresi
ML : Makine Öğrenmesi (Machine Learning)
MUT : Mutasyon Oranı
NLP : Doğal Dil İşleme (Natural Language Processing)
POP : Popülasyon Büyüklüğü
ROUGE : Recall-Oriented Understudy for Gisting Evaluation
RS : Rastgele Arama (Random Search)
TF-ISF : Term Frequency-Inverse Sentence Frequency

TABLO LİSTESİ

Sayfa

Tablo 3.1. Arapça metin özetleme çalışmalarında kullanılan yaygın veri kümeleri. 17
Tablo 3.2. Arapça metin özetleme çalışmaları: kategorik özet tablo. 28
Tablo 4.1. Hareke ve noktalama işaretlerinin kaldırılması örneği. 33
Tablo 4.2. Tokenizasyon ve stopword temizliği örneği. 34
Tablo 4.3. Stemming ve lemmatization karşılaştırması. 35
Tablo 4.4. Özelliklerin işlevsel tanımları. 37
Tablo 5.1. EASC veri kümesinin genel özellikleri. 47
Tablo 5.2. EASC veri kümesine ait özgün metin ve insan özetleri. 48
Tablo 5.3. Ağırlık değerleri. 49
Tablo 5.4. Referans özetlere göre ROUGE skorlarının karşılaştırılması. 50
Tablo 5.5. Sıkıştırma oranının model performansına etkisi. 51
Tablo 5.6. GA parametreleri. 52
Tablo 5.7. GA parametre deney sonuçları. 52

ŞEKİL LİSTESİ

Sayfa

Şekil 4.1. Sistem mimarisi. 31
Şekil 4.2. Rastgele Arama ile Ağırlık Optimizasyonu..... 43
Şekil 5.1. Benzerlik yöntem karşılaştırması. 49
Şekil 5.2. Referans özetlere göre rouge skorlarının karşılaştırılması. 50
Şekil 5.3. Sıkıştırma oranının model performansına etkisi. 51
Şekil 5.4. Karşılaştırmalı değerlendirme. 53

ÇIKARIMSAL ARAPÇA METİN ÖZETLEME: GENETİK ALGORİTMALAR İLE ÖZNİTELİK OPTİMİZASYONUNA DAYALI HİBRİT YAKLAŞIM

ÖZET

Son yıllarda dijital devrim; çevrim içi platformlar, medya kuruluşları, akademik yayınlar, e-kitaplar ve sosyal medya aracılığıyla üretilen metinlerin hacminde benzeri görülmemiş bir artışa yol açmıştır. Bu durum, "bilgi fazlalığı" olgusunu ortaya çıkarmış ve bireylerin, öğrencilerin ve araştırmacıların en önemli ve bağlamsal olarak ilgili içerikleri seçip işlemesini giderek zorlaştırmıştır. Mevcut metinlerin boyutu, insanların bunları baştan sona okuma kapasitesini aşmakta ve bu nedenle temel bilgileri verimli bir şekilde çıkarabilen akıllı araçlara olan ihtiyaç artmaktadır. Bu bağlamda, otomatik metin özetleme, özgün metinlerin özünü ve bütünlüğünü koruyarak kısa özetler üretme ve büyük miktarda bilgiyi daha az zaman ve çabayla işleme olanağı sunan kritik bir çözüm olarak öne çıkmaktadır.

Arapça, yirmiden fazla ülkede resmi dil olarak kullanılan ve yüz milyonlarca insan tarafından konuşulan bir dil olup, Kur'an-ı Kerim'in dili ve zengin kültürel ile tarihsel mirasın taşıyıcısı olarak özel bir statüye sahiptir. Bu küresel öneme rağmen, Arapça; İngilizce ve Çince gibi dillere kıyasla doğal dil işleme alanında yeterince temsil edilmemektedir. Bunun başlıca nedenleri arasında kök-temelli morfolojik yapısı, sözdizimsel zenginliği, anlamsal çok anlamlılığı ve geniş bağlamsal çeşitliliği yer almaktadır. Ayrıca, geniş ölçekli açıklamalı derlemler, dijital sözlükler ve önceden eğitilmiş dil modelleri gibi kaynakların yetersizliği de bu zorlukları artırmaktadır. Bu durum, Arapça metinler için güvenilir ve etkili özetleme sistemlerinin geliştirilmesini hem teknik bir zorunluluk hem de kültürel bir gereklilik haline getirmektedir.

Bu tez kapsamında, tek belgeli Arapça metinler için extractive tabanlı bir özetleme modeli geliştirilmiştir. Önerilen sistem hem istatistiksel hem de anlamsal özelliklerin birlikte değerlendirildiği birleşik bir çerçeveye dayanmaktadır. Cümlelerin önemi; metin içindeki konumu, uzunluğu, anahtar kelime yoğunluğu, sayısal verilerin varlığı, diğer cümlelerle benzerliği, kişi ve yer adları gibi adlandırılmış varlıkların tespiti, Arapça–İngilizce karışık terimlerin varlığı ve TF–ISF yöntemiyle hesaplanan terim önemi gibi sekiz temel özelliğe göre belirlenmektedir. Her cümle normalize edilmiş bir özellik vektörü ile temsil edilmekte ve bu özelliklerin ağırlıkları, özetlerin bilgilendirici ve bütünlüklü olmasını sağlamak amacıyla Random Search algoritması ile optimize edilmektedir.

Cümle seçimini daha da iyileştirmek amacıyla, en etkili sezgisel yöntemlerden biri olan Genetik Algoritma (GA) kullanılmıştır. GA, bilgilendiricilik, kapsayıcılık ve mantıksal bütünlük arasında denge kurarak en temsili cümle kombinasyonlarını seçmektedir. Aday özetler kromozom olarak modellenmiş, uygunluk fonksiyonu aracılığıyla optimize edilmiş özellik puanları ve Cosine Similarity yöntemi üzerinden ölçülen anlamsal tutarlılık dikkate alınarak değerlendirilmiştir. Seçim, çaprazlama

(crossover) ve mutasyon (mutation) işlemleri aracılığıyla, algoritma nesiller boyunca özetlerin kalitesini kademeli olarak artırmıştır.

Geliştirilen model, Arapça özetleme alanında yaygın olarak kullanılan EASC (Essex Arabic Summaries Corpus) veri kümesi üzerinde test edilmiştir. Değerlendirme için ROUGE-1, ROUGE-2 ve ROUGE-L metrikleri kullanılmıştır. Deneysel analizler sonucunda, en yüksek performansın %35 sıkıştırma oranında elde edildiği görülmüş ve ROUGE-1 F skoru %70.63 olarak kaydedilmiştir. Bu sonuç, modelin bilgilendiricilik ile özlülük arasında dengeli ve etkili bir performans sergilediğini ortaya koymaktadır.

Sonuç olarak, istatistiksel ve anlamsal özelliklerin Genetik Algoritma ile birleştirilmesi ve cümleler arası benzerlik hesaplamasında Cosine Similarity yönteminin kullanılması, Arapça metin özetleme için güçlü ve pratik bir çerçeve sunmaktadır. Geliştirilen model, gereksiz tekrarları en aza indirirken özlü, bütünlüklü ve anlamsal açıdan zengin özetler üretme kapasitesiyle öne çıkmaktadır. Bu çalışma, Arapça doğal dil işleme alanında önemli bir katkı sunmakta ve bilgi fazlalığı sorununa karşı etkili bir çözüm geliştirmektedir.

EXTRACTİVE ARABİC TEXT SUMMARİZATİON: A HYBRİD APPROACH THROUGH GENETİC ALGORİTHM-BASED FEATURE OPTİMİZATİON

SUMMARY

In recent decades, the digital revolution has triggered an exponential growth in textual data across online platforms, news media, academic publications, e-books, and social networks. This unprecedented expansion of information has created what is widely known as information overload, a phenomenon that has made it increasingly difficult for individuals, students, and researchers to identify, process, and utilize the content most relevant to their needs. The overwhelming volume of documents far exceeds the natural cognitive capacity of human readers, making manual processing an impractical task in today's digital era. In this context, automatic text summarization has emerged as an indispensable solution, enabling the generation of condensed representations of large texts that preserve the essential meaning and coherence of the original while significantly reducing the time and effort required to engage with massive amounts of information.

Arabic, one of the six official languages of the United Nations and the native language of more than 400 million people, carries immense cultural and historical significance as the language of the Holy Qur'an and as a medium of communication across a wide range of countries and regions. Despite its global importance, Arabic remains underrepresented in the field of Natural Language Processing when compared to high-resource languages such as English and Chinese. This gap arises from a combination of linguistic and resource-related challenges. The Arabic language is characterized by a root-based morphological system in which a single root can generate dozens of derivatives with distinct syntactic and semantic functions, adding layers of complexity to preprocessing tasks such as tokenization, stemming, and lemmatization. For example, the root "k-t-b" produces words like kataba (wrote), kātib (writer), and maktūb (written), each with unique grammatical roles. In addition, Arabic exhibits rich inflection, extensive use of clitics, and flexible word order, all of which complicate syntactic analysis and semantic interpretation. The situation is further complicated by the coexistence of Modern Standard Arabic and numerous regional dialects, which differ considerably in vocabulary, orthography, and grammar. While Modern Standard Arabic is employed in formal contexts such as journalism and academic writing, dialectal Arabic dominates everyday communication and social media. This diglossic reality makes the design of unified systems particularly challenging. Compounding these linguistic complexities is the scarcity of resources such as large-scale annotated corpora, domain-specific lexicons, and robust pre-trained language models. Unlike English, which benefits from billions of tokens and mature tools such as WordNet and BERT-based architectures, Arabic NLP still suffers from limited resources, slowing research progress and reducing the effectiveness of developed systems. These challenges underscore the urgent necessity of building reliable Arabic text

summarization frameworks capable of handling the structural and semantic intricacies of the language.

The present thesis proposes a single-document extractive summarization model designed specifically for Arabic texts, integrating statistical and semantic features into a unified optimization framework driven by a Genetic Algorithm. The system begins with a rigorous preprocessing stage tailored to the complexities of Arabic. This stage includes the removal of diacritics, punctuation, and redundant spaces; the normalization of letter forms, such as unifying different representations of Alif; and the segmentation of documents into sentences and words. Stopwords that carry little semantic weight are filtered out, while stemming and lemmatization techniques reduce words to their root or lemma forms, improving consistency across the text. For instance, the word yudarrisūn (they teach) is normalized to yudarris (to teach), which ensures more reliable feature extraction and reduces lexical variation. These steps collectively reduce noise and enable a more accurate representation of the underlying semantics of the text.

Once preprocessing is complete, the system evaluates each sentence based on eight carefully selected features designed to capture structural, statistical, and semantic dimensions of importance. The position of the sentence in the document is a critical feature, as Arabic texts, particularly in the news domain, often place essential information at the beginning. Numerical data, including dates, percentages, and statistics, are also considered important because they provide concrete factual details. Sentence length is incorporated to maintain balance, avoiding both overly short sentences that lack informativeness and excessively long ones that may include redundancy. Keyword density highlights domain-relevant terms that are central to the text's topic. To enhance coverage and reduce redundancy, inter-sentence similarity is measured to ensure that selected sentences collectively represent diverse aspects of the document. Named entity recognition identifies references to people, places, and organizations that are frequently central to the text's meaning. The presence of Arabic–English mixed terms, common in contemporary technical and media contexts, is given additional importance. Finally, the TF–ISF (Term Frequency – Inverse Sentence Frequency) measure emphasizes words that are frequent within a single sentence but not across the entire document, ensuring that distinctive and informative content is prioritized. Each sentence is thus represented as a normalized feature vector, allowing for systematic evaluation.

The weights of these features are not assigned arbitrarily but optimized using a randomized search algorithm, enabling the system to adapt empirically to the data. Building on this representation, a Genetic Algorithm is employed to refine sentence selection and produce the final summary. Candidate summaries are represented as binary chromosomes, where each gene indicates whether a particular sentence is included or excluded. A fitness function evaluates these candidate solutions by combining the weighted sentence scores with semantic coherence, the latter measured using cosine similarity. Through iterative operations of selection, crossover, and mutation, the algorithm explores the solution space and incrementally improves the quality of summaries, ultimately converging toward an optimal balance of informativeness, coverage, and coherence.

The effectiveness of the proposed model was tested extensively using the Essex Arabic Summaries Corpus, a benchmark dataset consisting of 153 Arabic articles and 765 human-generated summaries spanning domains such as politics, economics,

education, sports, and culture. A range of experiments was conducted to analyze the system's performance under different compression rates and Genetic Algorithm configurations. At a compression rate of 20%, the generated summaries were found to be overly concise, with relatively high precision but low recall, leading to a significant omission of important information. This was reflected in the ROUGE-1 F1 score of only 57.7%, confirming that excessive compression compromises informativeness. Increasing the compression to 30% produced a substantial improvement, with ROUGE-1 F1 rising to 68.7% and ROUGE-2 F1 exceeding 47%, representing a more balanced trade-off between brevity and coverage. The highest performance was achieved at a 35% compression rate, where the system attained a ROUGE-1 F1 score of 70.63% alongside stable ROUGE-2 and ROUGE-L scores. This indicates that a moderate compression rate offers the best equilibrium between informativeness and conciseness. At 40%, summaries became longer and more redundant, improving recall but reducing precision and leading to a slight decline in overall F1 values, which demonstrates the importance of avoiding excessive expansion.

In addition to compression experiments, the sensitivity of the system to Genetic Algorithm parameters was thoroughly investigated. Several combinations of population size, number of generations, mutation rate, and tournament size were tested. Smaller populations and fewer generations often led to premature convergence, yielding suboptimal summaries. Larger populations improved diversity but increased computational cost. The optimal performance was obtained with a population size of 250, 70 generations, a mutation rate of 0.2, and a tournament size of 30, producing the highest ROUGE-1 F1 of 70.63%. These findings underscore the critical importance of careful parameter tuning to maximize the potential of evolutionary optimization techniques.

Beyond numerical performance, the system's outputs were compared to the five human-written reference summaries for each document. The model consistently produced summaries that aligned closely with human references, particularly matching the second reference summary in terms of informativeness and coherence. This highlights the ability of the system not only to capture surface-level features but also to identify the thematic essence of texts in a manner similar to human summarizers.

Taken together, these results demonstrate that the proposed model achieves reliable performance across diverse experimental conditions and produces summaries that are coherent, semantically meaningful, and practically useful. The robustness of the system is evidenced by its ability to adapt to parameter variations while maintaining stability in its outputs. Most importantly, the close alignment with human-generated summaries validates the model's effectiveness and confirms its potential as a competitive solution in the field of Arabic text summarization.

In conclusion, this research confirms that integrating statistical and semantic features with a Genetic Algorithm provides a strong and practical framework for automatic summarization of Arabic texts. The system is capable of generating concise yet informative summaries that address the pressing challenge of information overload while accommodating the linguistic complexities of Arabic. By optimizing feature weights empirically and employing evolutionary computation to refine sentence selection, the model contributes to advancing the state of Arabic Natural Language Processing. Furthermore, the study underscores the broader importance of developing intelligent summarization tools for Arabic, as such systems hold the potential to

enhance access to information for millions of Arabic speakers worldwide, bridging gaps in knowledge accessibility and strengthening the role of Arabic in the digital era.

1. GİRİŞ

Günümüzde internet ve dijital platformlar üzerinden üretilen ve paylaşılan metinsel verilerin boyutu hızla artmaktadır. Web siteleri, haber portalları, blog yazıları, sosyal medya gönderileri, kullanıcı yorumları, akademik yayınlar ve çevrimiçi kitaplar gibi çok çeşitli kaynaklardan elde edilen bu içerikler, bilgiye erişim sürecinde yeni zorlukları da beraberinde getirmiştir. Kullanıcıların, bu yalnızca ilgili ve önemli bilgilere ulaşabilmesi, yoğun zaman ve çaba gerektirmektedir. Özellikle bilimsel araştırmalarla ilgilenen bireylerin, yayınlanan çalışmaların hızına yetişememesi gibi durumlar bu sorunu daha da belirgin hâle getirmiştir. Bu nedenle içinde bulunduğumuz bu dijital çağda ihtiyaç duyulan bilgiyi filtreleyecek araçların geliştirilmesi zorunlu hâle gelmiştir.

Bu bağlamda, Otomatik Metin Özetleme (Automatic Text Summarization – ATS) sistemleri, uzun metinlerde yer alan temel bilgilerin hızlı, öz ve anlam bütünlüğü korunarak sunulmasına imkân sağlayarak, zaman tasarrufu ve bilgiye etkin erişim gibi önemli katkılar sunmaktadır [1]

Özetleme, "orijinal metnin temel bilgilerini içeren ve orijinal metnin yarısını aşmayan kısa bir versiyonu" olarak tanımlanmaktadır. Otomatik özetleme, en önemli bilgileri içeren, bağlamsal olarak tutarlı ve anlam açısından bütünlüklü bir özet üretmeyi amaçlamaktadır. Böylece uzun metinleri okumak için gereken süre azalırken, temel bilgilere hızlı ve verimli bir şekilde erişim sağlanmaktadır [2].

Otomatik özetleme sistemleri, insan müdahalesine gerek duymadan bilgi seçimi ve yoğunlaştırma işlemini gerçekleştirmeye çalışır. Bu sistemler genellikle üç aşamadan oluşur: analiz (girdi metnin incelenmesi ve özelliklerin çıkarılması), dönüşüm (çıkarılan bilgilerin özet formatına dönüştürülmesi), ve sentez (kullanıcı ihtiyaçlarına uygun bir özetin oluşturulması).

Literatürde bu amaçla geliştirilen yöntemler; istatistiksel teknikler, meta-sezgisel algoritmalar, hibrit yaklaşımlar, makine öğrenimi tabanlı sistemler, bulanık mantık modelleri ve grafik tabanlı algoritmalar gibi çeşitli kategorilere ayrılmaktadır. Son

yıllarda, özellikle meta-sezgisel algoritmaların, metin özetleme görevlerinde dikkat çekici sonuçlar verdiği görülmektedir.

Bununla birlikte, farklı diller arasında yapılan araştırmaların dağılımına bakıldığında, İngilizce dilinde yürütülen çalışmaların oldukça yoğun olduğu, ancak Arapça metin özetleme (ArTS) alanındaki çalışmaların hâlâ sınırlı kaldığı görülmektedir. Bu durum, Arapça'nın yapısal karmaşıklığı, morfolojik zenginliği, çok anlamlılık ve standart veri kümelerinin eksikliği gibi dilsel zorluklardan kaynaklanmaktadır [3], [4] Bu nedenle, Arapça için etkili ve uygulanabilir özetleme sistemlerinin geliştirilmesi hem bilimsel literatür hem de doğal dil işleme (NLP) topluluğu açısından önemli bir ihtiyaç hâline gelmiştir

1.1. Çalışmanın Gerekçesi

Arapça metin özetleme sistemleri, özellikle morfolojik karmaşıklık, çok anlamlılık ve lehçesel çeşitlilik gibi dilsel zorluklar nedeniyle istenilen başarı düzeyine ulaşamamaktadır. Latin dillerine kıyasla, Arapça dilinde geliştirilen doğal dil işleme uygulamaları hem yapısal hem de kaynak yetersizlikleri nedeniyle oldukça sınırlıdır. Özellikle açık erişimli standart veri kümelerinin eksikliği, model eğitimi ve sonuçların karşılaştırmalı olarak değerlendirilmesini güçleştirmektedir

Metin özetleme yöntemleri genel olarak abstractive ve extractive olmak üzere ikiye ayrılmaktadır. Bu çalışmada odak noktası extractive özetleme olup, bu bağlamda özellikle Arapça metinler üzerinde geliştirilen çıkarımsal yöntemler incelenmektedir

Günümüzde var olan Arapça çıkarımsal özetleme sistemleri çoğunlukla istatistiksel veya yüzeysel yöntemlere dayalıdır. Bu sistemler metnin içeriksel bütünlüğünü göz ardı edebilmekte ve cümle seçimini bağlamdan kopuk şekilde gerçekleştirmektedir. Bu durum, özellikle zamir kullanımı, anlam tutarlılığı ve bilgi yoğunluğu gibi kriterlerin sağlıklı şekilde değerlendirilememesine yol açmaktadır [5] Ayrıca, mevcut literatürde birçok model, cümlelerin önemini değerlendirirken sınırlı sayıda özellikle çalışmakta ve bu özelliklerin ağırlıklandırılması sabit ya da sezgisel yollarla yapılmaktadır. Bu da üretilen özetlerin referans özetlerle olan uyumunu azaltmakta ve otomatik özetleme sistemlerinin güvenilirliğini sorgulatmaktadır[6]. Bu bağlamda, bu çalışma, geniş kapsamlı uygulanabilirliğe ve yüksek performansa sahip çıkarım yöntemlerini kullanarak Arabic Text Summarization – ArTS sisteminin otomatik olarak oluşturulmasını hedeflemektedir. Bu sorunun çözümü için belirli

kriterlerin tanımlanması gerekmektedir. Bunlar arasında uygun bir durak kelime listesi seçimi, en uygun veri kümesinin belirlenmesi, ön işleme adımlarının optimize edilmesi, en iyi türetme tekniğinin seçilmesi, en verimli temel bileşenlerin belirlenmesi, özetleme sürecinde en önemli özelliklerin çıkarılması, sistemi değerlendirmek için uygun bir veri kümesinin seçilmesi ve özetin yeniden oluşturulma mekanizmasının belirlenmesi yer almaktadır.

Sonuç olarak, Arapça için güvenilir ve yüksek performanslı çıkarımsal özetleme yöntemlerinin geliştirilmesi önemli bir araştırma alanı olarak ortaya çıkmaktadır. Arapça, yaklaşık 280 milyon kişi tarafından anadil olarak konuşulmakta ve dünya genelinde 400 milyondan fazla kullanıcıya ulaşmaktadır. 20'den fazla ülkede resmi dil olması ve dijital ortamlarda üretilen içerik miktarının hızla artması Arapça metin özetleme çalışmalarının önemini daha da artırmaktadır. Ancak bu alanda hâlâ önemli araştırma boşlukları bulunmakta ve bu boşlukların giderilebilmesi için özelliklerin daha verimli şekilde kullanıldığı ve optimizasyon teknikleriyle desteklenen yeni yöntemlere ihtiyaç duyulmaktadır.

1.2. Çalışmanın Amacı

Bu çalışmanın amacı, Arapça metinler için özellik sıralamasına dayalı optimize edilmiş bir çıkarımsal özetleme yöntemi geliştirmektir. Araştırma, cümlelerin önemini belirleyen sekiz temel özelliğe dayanarak en uygun cümlelerin seçilmesine ve ağırlıklandırılmasına odaklanmaktadır. Bu özellikler; cümlenin metindeki konumu, sayısal verilerin varlığı, cümle uzunluğu, anahtar kelime içeriği, cümle benzerliği, adlandırılmış varlıkların varlığı ve terim sıklığı-ters cümle sıklığı (TF-ISF) olarak belirlenmiştir. Her cümle, normalize edilmiş bir özellik vektörü olarak temsil edilmekte ve önem derecesine göre sıralanmaktadır.

Özetleme doğruluğunu artırmak amacıyla, her özelliğe en uygun ağırlıkların belirlenmesini sağlayan genetik algoritma tabanlı bir optimizasyon yöntemi entegre edilmiştir. Böylece seçilen cümlelerin referans özetlerle daha yüksek uyumluluk göstermesi hedeflenmiştir. Bu çalışma, Arapça çıkarımsal özetleme sürecini iyileştirmeyi ve doğal dil işleme (NLP) alanında Arapça için yeni katkılar sağlamayı amaçlamaktadır.

Otomatik özetleme, çeşitli alanlarda vazgeçilmez bir araç haline gelmiştir ve araştırma ile bilgi geri çağırma verimliliğini artırarak, uzun metinlerin odaklanmış özetlerini

sunarak belgelerin tamamının okunmasında harcanan zamanı azaltır. Ayrıca, haberler ve makalelerin daha hızlı ve etkili bir şekilde analiz edilmesini sağlayarak, kullanıcıların olayları takip etmelerine ve bilinçli kararlar almalarına yardımcı olur [7] Etkileşimli yapay zeka sistemlerinin geliştirilmesinde de önemli bir rol oynar, metinleri anlayarak ve özetleyerek doğru yanıtlar veren akıllı asistanların performansını artırır [8] Ancak, bu alandaki büyük gelişmelere rağmen, otomatik özetleme alanında bazı büyük zorluklar bulunmaktadır. Bu zorluklar arasında, çıkarılan cümlelerin bağlamını koruma, çünkü doğrudan özetleme işlemi metnin genel bağlamını kaybetmeye yol açabilir [9], zamirler ve referans ifadeleriyle başa çıkma, çünkü cümlelerin ayrı ayrı çıkarılmasında zamirler net referanslara sahip olmayabilir [10], ve yüzeysel tekniklere dayanma, çünkü metnin derin anlamını anlamayan teknikler kullanmak hatalı çıkarımlara yol açabilir[11].

1.3. Çalışmanın Kapsamı ve Sınırlılıkları

Bu araştırmanın kapsamı, Arapça tek belge çıkarımsal (extractive) özetleme ile sınırlıdır. Literatürde Arapça için farklı veri setleri bulunmaktadır. Bu veri setleri genellikle haber makaleleri, Wikipedia içerikleri veya akademik metinler içermektedir. Bu çalışmada kullanılan EASC (Essex Arabic Summaries Corpus) veri kümesi ise Wikipedia ve Al-Rai ile Al-Watan gazetelerinden alınmış 153 belgeden oluşmakta, ayrıca her belge için uzmanlar tarafından hazırlanmış beş referans özet içermektedir. Bu yapısı sayesinde EASC, sistem performanslarının karşılaştırmalı olarak değerlendirilmesi için güçlü bir temel sunmaktadır. Bununla birlikte, Arapça dilinin morfolojik zenginliği, çok anlamlılık, lehçesel çeşitlilik ve yazım farklılıkları özetleme sürecini zorlaştıran önemli sınırlılıklar arasında yer almaktadır. Bu çalışmada, cümlelerin önemini belirleyen sekiz temel özellik tanımlanmış ve bu özelliklerin ağırlıkları genetik algoritma tabanlı bir optimizasyon yöntemiyle ayarlanarak çıkarımsal özetleme tekniğinin performansı artırılmaya çalışılmıştır. Çalışmanın kapsamı üretici (abstractive) özetleme tekniklerini içermemekte, yalnızca çıkarımsal yöntemlere odaklanmaktadır.

1.4. Tezin Organizasyonu

Bu çalışma altı ana bölümden oluşmaktadır. Birinci bölümde çalışmanın amacı, gerekçesi, kapsamı ve sınırlılıkları sunulmuştur. İkinci bölümde metin özetlemenin

kuramsal temelleri, sınıflandırmaları, uygulama alanları, yaklaşımlar ve değerlendirme metrikleri ele alınmıştır. Üçüncü bölümde, Arapça metin özetleme alanındaki mevcut çalışmalar, kullanılan veri kümeleri ve yöntemler incelenmiştir. Dördüncü bölümde önerilen hibrit sistemin uygulama süreci ayrıntılı olarak açıklanmış; ön işleme adımları, özellik çıkarımı, skor hesaplama, ağırlık optimizasyonu ve genetik algoritma ile cümle seçimi sunulmuştur. Beşinci bölümde deneysel senaryolar, uygulama sonuçları ve ROUGE metrikleriyle yapılan değerlendirmeler yer almaktadır. Altıncı ve son bölümde ise çalışmanın genel sonuçları özetlenmiş, tartışmalar yapılmış ve geleceğe yönelik öneriler sunulmuştur.

2. OTOMATİK METİN ÖZETLEMEDE TEMEL KAVRAM VE YÖNTEMLER

Metin Özetleme (Text Summarization-TS), Doğal Dil İşleme (NLP) alanında büyük metinleri özetleyerek temel bilgileri korumayı amaçlayan önemli bir çalışma alanı. Otomatik Metin Özetleme (ATS), bilgi erişimini iyileştirme, okuma süresini azaltma ve bilgi çıkarımını kolaylaştırma açısından kritik bir rol oynamaktadır. Bu nedenle hukuk, tıp, siyaset ve gazetecilik gibi birçok alanda kullanılmakta; haber analizleri, tıbbi kayıtların gözden geçirilmesi ve akademik çalışmaların özetlenmesi gibi uygulama alanlarına sahiptir [12][13] Dünya dillerinde bu tür uygulamalar yaygınlaşırken, özellikle Arapça dilindeki dijital içeriklerin hızla artması, çevrimiçi değerlendirmeler, makaleler, haberler ve araştırma raporlarının çoğalmasıyla birlikte özetleme tekniklerinin önemi daha da artmıştır.

Metin özetleme alanındaki bu gelişmeler ışığında, bu bölümde otomatik özetleme yaklaşımlarının temel kavramları, sınıflandırmaları, güncel uygulama alanları ve Arapça diline özgü zorluklar ele alınacak, ayrıca sistemlerin başarımını ölçmek için kullanılan değerlendirme metrikleri açıklanacaktır.

2.1. Metin Özetleme Yaklaşımlarının Kategorizasyonu

2.1.1. Dil-tabanlı Yaklaşımlar

- **Tek dilli (mono):** Hem giriş metni hem de oluşturulan özet aynı dilde yer alır. Bu tür sistemler, belirli bir dilin sözdizimsel ve anlamsal özelliklerine odaklanarak daha yüksek doğrulukta özetler üretmeyi amaçlar. Özellikle haber makaleleri, akademik yazılar ve tıbbi belgeler gibi belirli bir dilde yoğun veri üreten alanlarda yaygın olarak kullanılmaktadır.
- **Çok dilli (multi): F**arklı dillerde yazılmış birden fazla metni işleyerek tek bir özet sunmayı hedefler. Bu özet ya belirli bir dilde üretilebilir ya da çok dilli bir formatta sunulabilir. Çok dilli özetleme, özellikle uluslararası haber ajansları, çok uluslu şirketler ve çok dilli veri arşivleri gibi ortamlarda bilgi erişimini kolaylaştırmak amacıyla geliştirilmiştir.

- **Çapraz dilli (cross):** Giriş metni bir dilde olurken, oluşturulan özet farklı bir dilde üretilir. Bu yöntem, makine çevirisi ve doğal dil işleme tekniklerinin entegrasyonunu gerektirir. Çapraz dilli özetleme, farklı dillerde üretilmiş bilgilere erişimi artırmak ve çok dilli kullanıcı kitlelerine yönelik bilgi sunumunu kolaylaştırmak için önemli bir araç olarak kabul edilmektedir.

2.1.2. Yöntem-tabanlı Yaklaşımlar

- **Çıkarımsal özetleme (extractive summarization):** Bu yaklaşım, orijinal metinden en önemli cümleleri veya kelime öbeklerini seçerek özet oluşturur ve seçilen ifadelerin yapısı değiştirilmez. Seçim süreci; kelime sıklığı, cümle benzerliği, metindeki pozisyon gibi istatistiksel ve yapısal özelliklere dayanır. Bu süreçte genellikle kelime frekans analizi, cümle benzerlik matrisleri, PageRank ve TextRank gibi sıralama algoritmaları uygulanır [1].

Bu yaklaşımların avantajı; cümlelerin doğrudan metinden alınması nedeniyle anlam bozulmasının önlenmesidir. Bu nedenle çıkarımsal özetleme; haber metinleri, tıbbi kayıtlar ve hukuk belgeleri gibi bilgi kaybının tolere edilemeyeceği alanlarda yaygın şekilde kullanılır.

- **Soyutlayıcı özetleme (abstractive summarization):** Metnin temel fikrini kavrayarak yeniden ifade eder ve yeni, özgün cümleler üretir. Bu yaklaşım, doğal dilin derinlemesine anlaşılmasını ve yeniden üretimini gerektirir. Uygulanan modeller arasında Tekrarlayan Sinir Ağları (RNN), BiLSTM, Seq2Seq, Attention mekanizmaları ve Transformer tabanlı sistemler (ör. BERT, GPT, T5) yer alır [12], [14]. Soyut özetleme, daha doğal, akıcı ve insan benzeri özetler oluşturur. Bu yöntem; diyalog sistemleri, chatbotlar, haber ve akademik özetleme sistemleri gibi kullanıcı dostu etkileşim gerektiren ortamlarda öne çıkmaktadır.
- **Hibrit özetleme:** Çıkarımsal ve soyutlayıcı özetleme yaklaşımlarının bir araya getirilmesiyle oluşur. Bu tür yaklaşımlarda, öncelikle metinden önemli cümleler çıkarılır ardından bu cümleler soyutlayıcı modeller aracılığıyla yeniden ifade edilerek özet oluşturulur

Bu tür yaklaşımlarda, hem bilgi kaybını en aza indirir hem de dilsel akıcılığı artırır. Özellikle uzun ve karmaşık metinlerde, bilgilendirici ve anlam açısından zengin özetler elde etmek için etkili bir stratejidir. EA-LTS gibi bazı sistemler, grafik tabanlı

çıkarım tekniklerini RNN tabanlı soyutlayıcı mimarilerle birleştirerek yüksek performanslı hibrit modeller geliştirmiştir [15]

2.1.3. Kaynak-tabanlı Yaklaşımlar

Otomatik Metin Özetleme (ATS) sistemleri, özetlemek için kullanılan belge sayısına göre iki ana türe ayrılır:

- **Tek belge özetleme (single-document summarization):** Bu türde, sistem yalnızca tek bir kaynaktan (bir makale, bir haber metni gibi) bir özet üretir. Tek belge özetleme, özellikle bireysel haber yazıları veya akademik makaleler gibi tek kaynaklı içeriklerde yaygın olarak kullanılır (Gambhir & Gupta, 2017) [16].
- **Çoklu belge özetleme (multi-document summarization):** Bu yöntem, birden fazla belgeyi özetlemeyi amaçlar. Farklı belgelerden bilgi toplayarak bir bütün halinde kısa ve tutarlı bir özet üretir. Çoklu belge özetlemede tekrar eden bilgilerden kaçınmak, zamansal ilişkileri ve özne-nesne bütünlüğünü korumak gibi ek zorluklar vardır (Sahoo et al., 2016).[17]

2.1.4. Hedef-tabanlı Yaklaşımlar

Otomatik Metin Özetleme (ATS) sistemleri, üretilen özetin türüne göre iki ana sınıfa ayrılabilir:

- **Genel özet (Generic summarization):** Bu türde sistem, kaynak metindeki en önemli bilgileri öne çıkararak genel bir özet oluşturur. Kullanıcıya özgü bir ihtiyaç veya soru göz önüne alınmadan, metnin genel anlamını temsil etmeye çalışır. Haber metinleri veya akademik makaleler için yapılan klasik özetlemeler bu gruba girer (Gambhir & Gupta, 2017).[16]
- **Sorgu tabanlı özet (query-based summarization):** Bu tür özetleme, belirli bir kullanıcı sorgusuna veya bilgi ihtiyacına dayalı olarak yapılır. Özet yalnızca sorguyla en çok ilişkili olan içerikleri içerir. Bu yöntem, özellikle arama motorları, bilgi erişim sistemleri veya kişiselleştirilmiş özetleme uygulamalarında tercih edilir (Lloret, Plaza, & Aker, 2017).[18]

2.1.5. İçerik-tabanlı Yaklaşımlar

Otomatik Metin Özetleme (ATS) sistemleri, oluşturulan özetin amacı ve içeriğin sunumuna göre iki ana kategoriye ayrılabilir:

- **Bilgilendirici özet (informative summarization):** Bu özet türü, kaynak metindeki ana fikirleri ve önemli detayları kapsayacak şekilde hazırlanır. Amaç, kullanıcının kaynak metni okumadan konu hakkında yeterli bilgi edinmesini sağlamaktır. Özellikle haber özetleri, akademik makale özetleri gibi uygulamalarda yaygındır
- **İfade Edici özet (indicative summarization):** İfade edici özetler, metnin tüm içeriğini aktarmak yerine, temel konular, ana temalar veya yapısal bilgiler hakkında kısa bir fikir verir. Bu tür özetler, kullanıcının ilgili olup olmadığını anlamasına yardımcı olur, fakat tüm detayları vermez. Genellikle kataloglar, kütüphane özetleri veya içerik tanıtımları için kullanılır

2.1.6. Bağlam-tabanlı Yaklaşımlar

Otomatik Metin Özetleme (ATS) sistemleri, özetleme yapılacak metnin bağlamına göre üç kategoriye ayrılır:

- **Alan bağımlı özetleme (domain dependent summarization)**, belirli bir alan veya konuya özgü metinler için geliştirilir. Bu sistemler, tıp, hukuk veya haber gibi belirli bilgi alanlarındaki içeriklere odaklanır ve bu alanlara özgü terminoloji ve yapıları dikkate alır.
- **Alan bağımsız özetleme (domain ındependent summarization)**, herhangi bir özel alana bağlı kalmadan çalışır. Çeşitli konulardaki metinleri özetleyebilmek için tasarlanmıştır ve daha genel, esnek bir yapı sunar.
- **Tür özel özetleme (genre-specific summarization)**, metnin türüne (örneğin romanlar, teknik belgeler veya sosyal medya gönderileri) göre uyarlanır. Metnin yapısal ve dilsel özellikleri türüne göre farklılık gösterdiği için, özetleme stratejileri de buna göre optimize edilir.

2.2. Metin Özetlemede Güncel Uygulamalar

- **Gazetecilik:** Günlük olarak üretilen haber içeriklerinin hacmi nedeniyle özetleme, gazetecilikte kritik öneme sahiptir. Özellikle, haber başlıkları ve spot cümleleri oluşturmak için otomatik özetleme sistemleri kullanılmaktadır. En çok tercih edilen yöntem çıkarımsal özetlemedir çünkü cümleleri doğrudan metinden seçerek hızlı ve doğru özet üretimi sağlar.[19] Ayrıca, çoklu

kaynaklardan gelen içeriklerin özetlenmesi, bilgi çoğaltımı ve taraflılığı azaltmakta; böylece okuyucuya daha dengeli bir bakış açısı sunulmaktadır[20]

- **Hukuk:** Yasal belgeler genellikle uzun ve teknik içerikli olduğundan, özetleme hukuk uzmanlarına zaman kazandırır. Çıkarımsal özetleme; hukuki terminolojiyi bozmadan önemli hükümleri öne çıkarmada tercih edilir [21], [22]. Öte yandan, dava özetleri ve düzenleyici belgelerde soyut özetleme teknikleri, belgeyi daha kısa ve anlaşılır hale getirerek hukuk profesyonellerinin karar alma sürecini hızlandırabilir [23].
- **Bilimsel Araştırmalar:** Tıbbi ve bilimsel yayınların hacmindeki artış, özetlemeyi bu alanlarda vazgeçilmez hale getirmiştir. Tıbbi belgelerde, çıkarımsal özetleme yöntemleri genellikle klinik notlardaki tanı ve tedavi bilgilerini doğrudan yansıttığı için tercih edilmektedir[24].

Ancak, soyut özetleme modelleri, tıbbi terimleri sadeleştirerek, özellikle hasta bilgilendirme belgelerinde ve genel okuyucuya yönelik bilimsel içeriklerde anlaşılabilirlik sağlar. Bununla birlikte, bu tür uygulamaların başarısı; güvenilir, etiketlenmiş ve uzmanlaşmış veri kümelerine olan erişime doğrudan bağlıdır [25].

- **Ticari İçerikler:** Ticari içeriklerin özetlenmesi, ürün yorumları ve müşteri geri bildirimlerinin hızlı bir şekilde analiz edilmesini sağlayarak zaman kazandırır. Bu tür özetleme sistemleri, işletmelerin müşteri eğilimlerini daha iyi anlamasına ve stratejik kararlar almasına yardımcı olmaktadır.
- **Sosyal Medya** Sosyal medya içeriklerinin özetlenmesi ise farklı zorluklar barındırmaktadır. Sosyal medya dilinin gayri resmi yapısı, kısaltmaların yoğunluğu ve imla varyasyonları, özetleme sistemlerinin bağlamı anlama yeteneğini kritik hale getirmiştir. Bu bağlamda, derin öğrenme modelleri ile duygu analizi entegre edilerek kullanıcı eğilimleri ve yeni trendlerin ortaya çıkarılmasında etkili özetleme sistemleri geliştirilmiştir[26][20].

2.3. Metin Özetlemede Arapça Dilinin Zorlukları

Otomatik Metin Özetleme (ATS) sistemleri, metinlerdeki önemli bilgileri otomatik olarak seçmek ve bilgi yoğunluğunu korumak amacıyla çeşitli tekniklere dayanır. Bu alandaki yaklaşımlar; istatistiksel analizler, anlamsal modellemeler, grafik tabanlı yöntemler, makine öğrenimi algoritmaları ve derin öğrenme teknikleri gibi farklı

stratejiler kullanılarak geliştirilmiştir. Ayrıca, son yıllarda sezgisel algoritmalar ve hibrit modeller de özetleme performansını artırmak için yoğun şekilde araştırılmaktadır. Uygulanan yönteme göre, sistemlerin doğruluk, kapsam ve akıcılık düzeylerinde farklılıklar gözlemlenmektedir.

Özellikle Arapça dili bağlamında, özetleme sistemleri bazı dilsel zorluklarla karşılaşmaktadır. Arapça, kök-tabanlı sözcük türetme, çoklu çekim sistemleri ve zengin morfolojik yapısı ile dikkat çeker. Diglossia (Modern Standart Arapça ve bölgesel lehçeler) durumu, kelime dağarcığında ve sözdiziminde büyük çeşitlilik yaratır [20]. [21].

Ayrıca, sesli harflerin eksikliği (diakritics), yazım varyasyonları ve dilin bağlam duyarlılığı, kelime temelli modellerin performansını düşürebilir. Bu nedenle, yazım normalizasyonu ve lemmatizasyon tekniklerinin etkin kullanımı şarttır [25]. Arapça için altın standart (gold standard) özet veri kümelerinin eksikliği, model eğitimi ve değerlendirme süreçlerini zorlaştırmaktadır. Mevcut modellerde ROUGE gibi geleneksel ölçütler kullanılmakla birlikte, morfolojik özellikleri dikkate alan alternatif değerlendirme metriklerine ihtiyaç vardır. Buna rağmen, AraBERT, AraGPT ve AraT5 gibi önceden eğitilmiş derin öğrenme modelleri, bağlamsal temsil kabiliyetleri ile önemli başarılar sağlamıştır. Bu modeller, Arapça'nın sözdizimsel karmaşıklığını ve anlamsal derinliğini daha iyi kavrayarak özet kalitesini artırmaktadır [27]. [26]

2.4. Değerlendirme Metrikleri

Metin özetleme sistemlerinin performansını değerlendirmek, üretilen özetin ne kadar kaliteli ve bilgilendirici olduğunu anlamak açısından oldukça önemlidir. Bu amaçla, sistemin oluşturduğu otomatik özetler, insanlar tarafından hazırlanmış olan "altın standart" özetlerle karşılaştırılır. Elle yapılan değerlendirmeler güvenilir sonuçlar verse de, oldukça zaman alıcı ve maliyetlidir. Bu nedenle, otomatik değerlendirme metrikleri araştırmalarda yaygın şekilde tercih edilmektedir. Başlangıçta, makine çevirisi alanında geliştirilen BLEU gibi metrikler kullanılmış, ancak zamanla özetleme görevine özel olarak tasarlanan ROUGE metrikleri öne çıkmıştır. Bu bölümde, otomatik özetleme sistemlerinin başarısını ölçmekte yaygın olarak kullanılan Precision, Recall, F1-Score, ROUGE, BLEU gibi temel metriklerin yanı sıra METEOR ve BERTScore gibi daha yeni ve anlam tabanlı değerlendirme araçları açıklanmaktadır.

- **Precision (Kesinlik) ve Recall (Duyarlılık)**: Otomatik özetleme sistemlerinin performansını değerlendirmede sıkça kullanılan iki temel ölçüttür. Her ikisi de sistem özeti ile referans özet arasındaki örtüşmeleri sayısal olarak analiz eder, ancak odaklandıkları yönler farklıdır:
- **Precision (Kesinlik):** Sistem özetinde yer alan bilgilerin ne kadarının gerçekten referans özetle örtüştüğünü ölçer. Yani, sistemin doğru tahmin ettiği bilgilerin, ürettiği özet içindeki oranıdır.

$$\text{Precision} = \frac{\left|\text{gram}_{\text{sys}} \cap \text{gram}_{\text{ref}}\right|}{\left|\text{gram}_{\text{sys}}\right|} \tag{2.1}$$

- **Recall (Duyarlılık):** Referans özetinde bulunan bilgilerin ne kadarının sistem özeti tarafından yakalanabildiğini gösterir.

$$\text{Recall} = \frac{\left|\text{gram}_{\text{sys}} \cap \text{gram}_{\text{ref}}\right|}{\left|\text{gram}_{\text{ref}}\right|} \tag{2.2}$$

- **F1-Skoru (F-Measure):** Precision ve Recall değerlerinin harmonik ortalamasıdır. Dengeli bir performans değerlendirmesi sağlar:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.3}$$

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) otomatik özetleme sistemlerinin başarımını değerlendirmek için kullanılan en yaygın metriklerden biridir [28]Sistem özeti ile referans özet arasında geçen kelime ya da kelime grubu (n-gram) eşleşmelerine dayalı olarak benzerliği ölçer.

Sık kullanılan türler şunlardır:

- ROUGE-1: Unigram (tek kelime) eşleşmeleri
- ROUGE-2: Bigram (iki kelime) eşleşmeleri
- ROUGE-L: Cümle yapısına dayalı en uzun ortak alt dizi (LCS)

Bu metrikler, özellikle çıkarımsal özetlemede tercih edilir. ROUGE metrikleri genel olarak ROUGE-N başlığı altında ifade edilir.

$$\text{ROUGE-N} = \frac{\sum \text{Count}_{\text{match}}(n\text{-gram})}{\sum \text{Count}_{\text{ref}}(n\text{-gram})} \quad (2.4)$$

- **BLEU:** Otomatik olarak oluşturulan özetin, referans özetle kelime grubu (n-gram) düzeyinde ne kadar benzer olduğunu ölçen bir metriktir. Genellikle abstraktif özetleme sistemlerini değerlendirmek için kullanılır ve kesinlik (precision) odaklıdır.

Sistem özeti kısa olduğunda, haksız yere yüksek puan almaması için kısalık cezası (brevity penalty) uygulanır. BLEU skoru, farklı n-gram seviyelerindeki eşleşmelerin ortalamasına göre hesaplanır:

$$P_n = \frac{\sum_{C \in \text{kümeler}} \sum_{n\text{-gram} \in C} \text{Count}_{\text{clip}}(n\text{-gram})}{\sum_{C \in \text{Kümeler}} \sum_{n\text{-gram} \in C} \text{Count}(n\text{-gram})} \quad (2.5)$$

BLEU genellikle tam ifade benzerliğinin önemli olduğu durumlarda destekleyici bir ölçüt olarak kullanılır [29]

- **METEOR:** sistem özetindeki kelimeler ile referans özet arasındaki anlamsal benzerliği daha esnek biçimde ölçen bir metriktir. Hem kesinlik (precision) hem de duyarlılık (recall) değerlerini dikkate alır ve kelime kökleri, eşanlamlılar gibi eşleşmelere de puan verir. Bu yönüyle, ROUGE'dan farklı olarak kelimeler tam aynı olmasa da benzer anlam taşıyorsa sistemi ödüllendirir. Özellikle abstraktif özetleme değerlendirmelerinde tercih edilir [30]
- **BERTScore**: özet ve referans cümleler arasındaki anlamsal yakınlığı ölçmek için, kelimeleri vektörler olarak temsil eder. Bu temsil, önceden eğitilmiş BERT modeli ile yapılır .Kelimeler tam olarak eşleşmese bile anlam olarak benzer olduklarında yüksek puan verir. Bu nedenle, özellikle dil çeşitliliği olan özetlerde oldukça kullanışlıdır [31].

Her bir değerlendirme metriği, özetin kalitesini farklı bir boyuttan ele alır. Bu nedenle, sağlıklı ve güvenilir bir değerlendirme yapabilmek için genellikle birden fazla metriğin birlikte kullanılması tercih edilir. Tek bir ölçüt, özetin tüm yönlerini kapsayamayabilir. Örneğin, ROUGE metrikleri, özellikle çekme (extractive) özetlerde özetin referansla içerik açısından ne kadar örtüştüğünü ölçmede oldukça etkilidir. Ancak sistem özeti, referansla aynı bilgileri farklı kelimelerle ifade ettiğinde, ROUGE bu durumu yeterince yansıtamayabilir. Bu gibi durumlarda, METEOR veya

BERTScore gibi anlamsal benzerliği dikkate alan metrikler, daha doğru bir değerlendirme sağlar. Benzer şekilde, sadece recall değeri yüksek olan bir sistem, geniş kapsamlı olabilir; ancak precision düşükse, bu durum özetin gereksiz veya ilgisiz bilgiler içerdiğini gösterebilir. Bu durumda, F1 skoru, her iki değeri dengeleyerek daha adil bir sonuç sunar. Bu nedenlerle, akademik çalışmalarda sistem performansı raporlanırken genellikle birden fazla ROUGE metriği (ROUGE-1, ROUGE-2, ROUGE-L) birlikte sunulur ve ihtiyaç durumunda BLEU, METEOR veya BERTScore gibi ek metriklerle desteklenir. Farklı metriklerin birlikte değerlendirilmesi, sistemlerin güçlü ve zayıf yönlerini daha bütüncül şekilde ortaya koyar. Sonuç olarak, çok yönlü bir bakış açısı sunan bu yaklaşım, otomatik özetleme sistemlerinin başarısını ölçmede daha adil, kapsamlı ve güvenilir bir analiz imkânı sağlar.

3. LİTERATÜR ARAŞTIRMASI

Bu bölümde öncelikle metin özetleme alanındaki genel literatür ele alınmaktadır. Metin özetleme, İngilizce, Çince ve İspanyolca gibi çok konuşulan dillerde uzun süredir çalışılmakta ve bu diller için geniş veri setleri ile farklı yöntemler geliştirilmiştir. Bununla birlikte, düşük kaynaklı diller için yapılan çalışmalar sınırlı kalmıştır. Arapça da bu bağlamda morfolojik zenginliği ve kaynak kısıtlılığı nedeniyle özel zorluklar barındıran bir dildir. Bu nedenle, Arapça metin özetleme literatürü ayrı bir önem taşımakta ve bu çalışmada detaylı olarak incelenmektedir.

Bu çerçevede, bölümde öncelikle Arapça metin özetleme çalışmalarında kullanılan veri setleri tanıtılmakta, ardından istatistiksel, semantik, makine öğrenmesi, grafik tabanlı, bulanık mantık ve meta-sezgisel yaklaşımlar incelenmektedir. Devamında, genetik algoritma tabanlı yöntemler ayrıntılı olarak ele alınmakta ve son olarak mevcut yaklaşımların karşılaştırmalı bir değerlendirmesi sunulmaktadır.

3.1. Arapça Metin Özetleme Alanında Kullanılan Veri Setleri

Arapça metin özetleme üzerine yapılan çalışmaların büyük bir kısmı, kullanılan veri setlerinin kapsamı ve kalitesi ile doğrudan ilişkilidir. Bu nedenle, literatürde yer alan araştırmalarda kullanılan başlıca Arapça özetleme veri setlerinin kapsamını, kullanım türlerini ve içerik özelliklerini karşılaştırmalı olarak sunmak önemlidir. Aşağıdaki tabloda, bu çalışmalarda en sık başvurulan veri setleri özetlenmiştir

Tablo 3.1. Arapça metin özetleme çalışmalarında kullanılan yaygın veri kümeleri.

Veri Seti	Özetleme Türü	Açıklama
DUC2004	Tek belgeli - soyutlayıcı	100 Arapça haber makalesi ve İngilizce özetler.
EASC	Tek belgeli - çıkarımsal	150 makale ,765 insan yapımı özet.
KALIMAT	Tekli ve çoklu belge	20.291 tek belgeli ve 2.057 çok belgeli metin.
Gigaword 5	Kısa metin	başlık ve ilk satırlarla kısa özetler oluşturulmuş.
OSAC	Tek belgeli - soyutlayıcı	22.429 haber dokümanı içerir.
SANAD	Tek belgeli - çıkarımsal	190.000 haber makalesi.
RTAnews	Tek belgeli - çıkarımsal	23.837 makale; 40 kategoriye ayrılmış.

Tablo 3.1. (Devamı) Arapça metin özetleme çalışmalarında kullanılan yaygın veri kümeleri.

Veri Kümesi	Özetleme Türü	Açıklama
NADA	Tek belgeli - soyutlayıcı	OSAC ve DAA 13.066 makale
Multi-Document Summaries Corpora	Çok belgeli - çıkarımsal	30 makale
Xl_Sum	Tek belgeli - soyutlayıcı	46.897 Arapça makale + insan yapımı özet.
WikiHow	Tek belgeli - soyutlayıcı	770.000 makale- özet çifti; 29.229 Arapça metin.
AHS (Arabic Headline Summary)	Tek belgeli - soyutlayıcı	300.000 haber başlıklar özet
AMN (Arabic Mogalad Ndeef)	Tek belgeli - soyutlayıcı	265.000 haber ve özet.

3.2. Arapça Metin Özetleme Yaklaşımları

Arapça Otomatik Metin Özetleme (ArTS) alanında, farklı yöntemler geliştirilerek özetleme kalitesini artırmaya yönelik çeşitli yaklaşımlar benimsenmiştir. Bu yaklaşımlar, kullanılan tekniklere, dil işleme derinliğine ve özetin hedef niteliğine göre kategorilere ayrılmaktadır. Literatürde ArTS sistemleri genellikle istatistiksel yöntemler, anlamsal tabanlı yöntemler, makine öğrenimi tabanlı yaklaşımlar, grafik tabanlı yöntemler, bulanık mantık (fuzzy logic) yaklaşımları ve hibrit sistemler şeklinde sınıflandırılmaktadır

3.2.1. İstatistiksel yöntemler

Bu yöntemler, derin sözdizimsel ya da anlamsal analiz yapmadan yüzey düzeyindeki metriklere dayanır.

Douzidia ve Lapalme (2004)[32] tarafından geliştirilen, Arapça metin özetleme alanındaki ilk sistemlerden birini geliştirmiştir. DUC 2004'te sunulan bu sistem, önemli cümleleri seçmek için TF-IDF ağırlıklandırma ve cümle düzeyinde sezgisel kurallar kullanmıştır. Çalışma, bir Arapça özetleyicinin İngilizce sistemlerle karşılaştırmalı olarak değerlendirilmesi açısından öncü niteliktedir. Ayrıca, Arapça diline özgü istatistiksel çıkarımsal özetleme yöntemlerinin uygulanabilirliğini göstermiş ve dile özgü karşılaşılan zorlukları vurgulamıştır.

[33] çalışmalarında istatistiksel ve semantik modelleri birleştiren yeni bir Arapça metin özetleme algoritması geliştirdiler. İstatistiksel model, cümle özelliklerine göre aday cümleleri seçerken, semantik model bağlamsal anlamı korumak için Word2Vec kullandı. Sistem, EASC veri kümesi üzerinde değerlendirildi ve %75'lik bir özetleme hassasiyeti elde edilerek geleneksel yöntemlere kıyasla etkililiği gösterildi.

3.2.2. Semantik-tabanlı yöntemler

Arapça metin özetlemede kullanılan anlamsal yaklaşımlar, yalnızca görünen kelimelere değil, metnin derin anlamına odaklanır. Bu yöntemler, anlam sözlükleri, kelimeler arası ilişkiler ve anlamsal modelleme gibi dilsel araçları kullanır. Amaç, metnin özünü daha doğru şekilde yansıtan, bağlam açısından tutarlı ve anlamlı özetler üretmektir. Bu bağlamda, Arabic WordNet, anlamsal grafikler ve söylem çözümlemesi gibi kaynaklar sıkça başvurulan yöntemlerdendir.

Alami ve arkadaşları (2018)[34], hem sözcüksel örtüşmeleri hem de Arabic WordNet'ten türetilmiş anlamsal benzerlikleri gösteren iki katmanlı bir cümle grafiği oluşturarak istatistiksel ve anlamsal özellikleri birleştiren hibrit bir yöntem sunmuşlardır. Bu graf üzerinde, cümleleri puanlamak için değiştirilmiş bir PageRank algoritması uygulanmış; ardından özet içinde çeşitliliği sağlamak amacıyla Maximal Marginal Relevance (MMR) yöntemi kullanılmıştır. Önerilen sistem, EASC veri kümesi üzerinde geleneksel istatistiksel yaklaşımlara kıyasla daha yüksek ROUGE skorları elde ederek, çıkarımsal özetlemede anlamsal zenginleştirmenin önemini ortaya koymuştur.

Al-Khawaldeh ve Samawi (2015)[35], Arapça metinleri özetlemek için LCEAS adlı yeni bir yöntem önermiştir. Bu yöntem, metin içindeki benzer veya tekrar eden kelimeleri birbirine bağlayarak konunun daha iyi anlaşılmasını sağlar. Ayrıca, özetlenen cümlelerin yeni bilgi içerdiğinden ve tekrar etmediğinden emin olmak için "metinsel çıkarım" yöntemi kullanılmaktadır.

Lachkar ve Ouatik [36] Arapça metin özetlemesinde Gizli Anlamsal Analiz (Latent Semantic Analysis - LSA) yönteminin kullanımını araştırmışlardır. Bu yöntem, metindeki temel kavramsal konuları ortaya çıkarmak için kelimeler arasındaki eşgörünüm birlikte geçme desenlerini analiz eder. Daha sonra, bu kavramsal boyutlarla en yüksek ilişkiye sahip cümleler seçilir. Deneysel sonuçlar, LSA tabanlı özetleme yönteminin, sadece kelime sıklığına dayalı yöntemlere kıyasla önemli cümleleri daha doğru şekilde tespit ettiğini ve anlam çeşitliliği eş anlamlılık ve çok anlamlılık) gibi dilsel sorunları daha etkili şekilde ele aldığını göstermektedir.

Lagrini ve Redjimi (2021),[37] metinlerdeki söylem ilişkilerini belirlemek için RST ağacı kullanarak önemli cümleleri seçmiş, ardından konum ve uzunluğa dayalı istatistiksel puanlama ile özetlemeyi geliştirmiştir. EASC veri kümesindeki ROUGE

sonuçları, bu iki aşamalı yöntemin klasik istatistiksel yöntemlerden daha anlamlı özetler ürettiğini göstermiştir.

Elayeb ve arkadaşları (2020),[38] belge ile özeti arasındaki ilişkiyi "analogik orantı" mantığıyla modelleyen yeni bir yöntem önermiştir. Geliştirdikleri iki algoritmadan biri belge anahtar kelimelerinin özet cümlelerdeki varlığını kontrol ederken, diğeri bu kelimelerin frekansını dikkate almaktadır. Yapılan ROUGE ve BLEU değerlendirmelerinde bu yöntem, TextRank, LexRank ve LSA gibi geleneksel yöntemleri geride bırakarak anlam temelli özetlemede yenilikçi bir yaklaşım sunmuştur

3.2.3. Makine öğrenmesi tabanlı yöntemler

Makine öğrenimi yaklaşımları, özetleme görevini tahmine dayalı bir modelleme problemi olarak ele alır. Çıkarımsal özetlemede bu, genellikle bir cümlenin özet içinde yer alıp almaması gerektiğine karar vermek için sınıflandırıcı (classifier) veya sıralayıcı (ranker) modellerin eğitilmesi anlamına gelir. Son yıllarda bu yaklaşım, soyutlayıcı özetleme için derin öğrenme modellerini de kapsamaktadır. Arapça metin özetlemesi alanında, elle tanımlanmış özniteliklere dayalı klasik denetimli algoritmalardan, modern sinir ağlarına ve transformer tabanlı modellere doğru önemli bir ilerleme kaydedilmiştir.

Boudabous ve arkadaşları (2010),[39] Arapça tek belgeli metin özetleme görevini destek vektör makineleri (SVM) kullanarak ikili sınıflandırma problemi olarak ele almıştır. Her cümle için yapısal ve sözcüksel olmak üzere toplam 15 özellik çıkarılmış ve bu özellikler kullanılarak özetlemeye uygun cümleleri belirlemek amacıyla bir sınıflandırıcı eğitilmiştir. Bu çalışma, Arapça metin özetlemede makine öğrenimi kullanımının erken örneklerinden biri olup, özellik ağırlıklarının otomatik olarak öğretilmesinde etkili olduğunu göstermiş ve daha sonraki sınıflandırma tabanlı özetleme sistemleri için temel oluşturmuştur.

Belkebir ve Guessoum (2015), [40]Arapça metin özetleme alanında AdaBoost topluluk algoritmasını uygulamıştır. Sistem, her biri farklı bir cümle özelliğini değerlendiren zayıf sınıflandırıcılardan oluşmakta ve bu sınıflandırıcıların çıktıları birleştirilerek bir cümlenin özete dahil edilip edilmeyeceği tahmin edilmektedir. Arapça haber makalelerinden oluşan bir veri kümesi üzerinde yapılan testlerde, AdaBoost özetleyici; Naive Bayes ve SVM gibi diğer öğrenme tabanlı yöntemlere

kıyasla daha yüksek F-skora ulaşmıştır. Bu sonuçlar, boosting algoritmalarının çoklu cümle özelliklerini etkili bir şekilde değerlendirerek özetleme başarımını artırabileceğini ortaya koymaktadır.

2021 yılında Kahla ve arkadaşları, [41] Arapça özetleme için verinin yetersizliği sorununu çözmek amacıyla çok dilli transformer modelleriyle aktarım öğrenme (transfer learning) yöntemini kullanmışlardır. Çalışmalarında, insan eliyle özetlenmiş büyük bir Arapça haber metinleri veri kümesi sunulmuş ve önceden eğitilmiş modeller (mBERT, AraBERT ve mBART-50) bu veri üzerinde ince ayar (fine-tuning) ile yeniden eğitilmiştir. Özellikle yüksek kaynaklı dillerde özetleme görevleri için eğitilmiş modellerin, Arapçaya aktarılmasıyla çapraz dilli bilgi aktarımı sağlanmıştır. Bu yöntem, soyut özetleme performansını anlamlı şekilde artırmış; hem yüksek ROUGE skorları elde edilmiş hem de daha akıcı ve doğal özetler üretilmiştir. Bu çalışma, Arapça metin özetleme alanında transformer tabanlı mimarilerin başarısını göstermekte ve alanı İngilizce özetlemedeki başarı seviyelerine bir adım daha yaklaştırmaktır.

3.2.4. Graf- tabanlı yöntemler

Grafik tabanlı özetleme yöntemleri, cümleleri ya da paragrafları birer düğüm (node) olarak temsil eden ve aralarındaki benzerlik ya da anlamsal ilişkileri kenar (edge) olarak tanımlayan modellerdir. Bu yöntemlerde amaç, metnin en merkezi veya en bilgilendirici cümlelerini belirleyerek özet oluşturmaktır. TextRank ve LexRank gibi sıralama algoritmaları sıklıkla kullanılmaktadır. Arapça metin özetleme bağlamında, bu tür yaklaşımlar dilin morfolojik zenginliği ve sözcük çeşitliliği nedeniyle daha karmaşık hale gelmektedir. Bu nedenle, cümleler arasındaki anlamsal bağların doğru şekilde kurulması ve benzerlik matrislerinin etkili biçimde hesaplanması, Arapça için grafik tabanlı özetleme yöntemlerinin başarısında kritik rol oynamaktadır.

Al-Taani ve Al-Omour (2014) [42], tarafından yürütülen çalışmada, Arapça metinlerde çıkarımsal özetleme amacıyla cümlelerden oluşan bir grafik yapısı oluşturulmuştur. Bu yöntemde, her bir cümle bir düğüm (node) olarak tanımlanmakta ve cümleler arasındaki kenarlar (edges), ortak kökler veya içerik kelimeleri gibi benzerlik ölçütlerine dayalı olarak kurulmaktadır. Araştırmacılar, grafik bağlanabilirliğini optimize etmek amacıyla düğüm temsili için kelime, kök ya da n-gram gibi farklı birimleri denemişlerdir. Daha sonra, bu yapı üzerinde PageRank benzeri bir algoritma uygulanarak cümlelerin önem derecesi hesaplanmıştır. Sistem,

temel özetleme yöntemlerine kıyasla daha yüksek ROUGE skorları elde etmiş ve cümleler arası ilişkilere dayalı grafik yapısının, önemli bilgilerin belirlenmesinde etkili bir yaklaşım olduğunu ortaya koymuştur.

Ali (2019)[43] tarafından gerçekleştirilen çalışmada, Arapça da dahil olmak üzere çok dilli metin özetleme amacıyla konu modelleme ile grafik tabanlı sıralama birleştirilmiştir. Araştırmacı, metindeki gizli temaları belirlemek için Latent Dirichlet Allocation (LDA) yöntemini kullanmış ve her cümleyi, yedi farklı öznitelik ile konulara olan yakınlığına göre "önemli" veya "önemsiz" olarak sınıflandırmıştır. Sadece önemli cümlelerden oluşan bir grafik oluşturulmuş ve bu grafikteki kenarlar, cümleler arası anlamsal ya da yapısal ilişkileri temsil etmiştir. Daha sonra, bu grafikte değiştirilmiş bir PageRank algoritması uygulanarak cümlelerin nihai önem dereceleri hesaplanmıştır. TAC ve EASC veri kümelerinde yapılan değerlendirmelerde, konuya dayalı bu semantik model ile grafik sıralama kombinasyonunun, yalnızca grafik tabanlı yöntemlere kıyasla daha başarılı sonuçlar ürettiği gösterilmiştir.

Alselwi ve Taşcı (2024) [44] tarafından geliştirilen GEATS (Graph-based Extractive Arabic Text Summarization) yöntemi, geleneksel PageRank algoritmasını anlamsal gömme (semantic embedding) bilgisi ile zenginleştirerek yeni bir çıkarımsal özetleme tekniği sunmuştur. Bu yöntem kapsamında, her cümle düğümü, bir kelime gömme modelinden (örneğin Word2Vec) elde edilen vektör ile temsil edilmiş ve düğümler arasındaki kenar ağırlıkları, cümle vektörleri arasındaki kosinüs benzerlikleri kullanılarak hesaplanmıştır. Böylece sadece doğrudan kelime örtüşmeleri değil, aynı zamanda anlamsal yakınlık da grafiğe yansıtılmıştır. Yapılan değerlendirmelerde, GEATS yönteminin önceki en iyi çıkarımsal özetleme sistemine göre F-skorunda %7.5 oranında bir iyileşme sağladığı gösterilmiştir. Bu sonuçlar, grafik tabanlı özetleme yaklaşımlarına derin anlamsal benzerliklerin entegrasyonunun, Arapça metin özetleme kalitesini önemli ölçüde artırabileceğini ortaya koymaktadır.

Arapça metin özetleme çalışmaları kapsamında, temel grafik algoritmalarının (örneğin TextRank) Arapça'ya özgü dil özellikleri dikkate alınarak adapte edildiği görülmektedir. Bu uyarlamalarda, cümle grafiğinin yapısını güçlendirmek amacıyla benzerlik ölçütleri Arap morfolojisine uygun şekilde yeniden tanımlanmış; ayrıca durak kelime çıkarımı ve kök bulma (stemming) gibi ön işleme adımları entegre edilmiştir. Grafik tabanlı yöntemlerin, dil bağımsız yapıları sayesinde uygun ön işleme

süreçleriyle Arapça otomatik özetleme sistemlerinde de etkili bir performans sergilediği literatürde sıklıkla vurgulanmaktadır.

3.2.5. Bulanık mantık tabanlı yöntemler

Bulanık mantık tabanlı yöntemler, cümlelerin önem derecelerinin belirlenmesinde ortaya çıkan belirsizlik ve kesin olmayan değerlendirmeleri yönetmeyi amaçlamaktadır. Kesin eşik değerleri veya ikili kararlar yerine, üyelik dereceleri ve bulanık çıkarım kuralları kullanılarak bir cümlenin özet için ne kadar uygun olduğu belirlenir.

Arapça metin özetleme alanında, bulanık mantık teknikleri farklı kriterleri esnek bir şekilde birleştirerek cümle seçimini optimize etmekte kullanılmıştır. Bu yöntemler, özellikle özellik ağırlıklandırmasında hassasiyet gerektiren durumlarda önemli avantajlar sağlamış ve özetleme sistemlerinin performansında iyileşmeler elde edilmesine katkıda bulunmuştur.

Al-Taani ve Al-Sayadi (2024) [45], bulanık C-ortalamalar (Fuzzy C-Means) algoritması ile Latent Dirichlet Allocation (LDA) yöntemini birleştiren bir yaklaşım sunmuşlardır. Bu çalışmada, aynı konuya ait çoklu belgeler, öncelikle Fuzzy C-Means algoritması kullanılarak konu başlıklarına göre kümelenmiş; ardından, her bir küme içinde LDA uygulanarak belgelerdeki gizli konular belirlenmiştir. Bu süreç, özetlerde yer alacak en anlamlı cümlelerin seçilmesini sağlamıştır. TAC-2011 veri kümesi üzerinde gerçekleştirilen deneyler, önerilen yöntemin, karınca kolonisi ve diskriminant analiz algoritmalarını kullanan diğer Arapça özetleme yaklaşımlarıyla karşılaştırıldığında rekabetçi sonuçlar elde ettiğini göstermiştir.

Al-Qassem ve ark.. (2019) [46], cümleleri puanlamak için bulanık çıkarım sistemi kullanan çıkarımsal bir özetleyici önermiştir. Bu çalışmanın temel yeniliği, metindeki önemli isimleri belirlemek amacıyla geliştirilen özel bir isim çıkarım modülüdür. Dilbilgisel kurallara dayalı olarak seçilen bu önemli isimler, bulanık sistemin girdisi olarak kullanılmaktadır. Bulanık sistem, cümleleri isim yoğunluğu, cümle uzunluğu gibi çeşitli özellikler üzerinden keskin eşikler kullanmadan, esnek bulanık kurallarla değerlendirmektedir.

EASC veri kümesi üzerinde yapılan deneyler sonucunda, bu bulanık mantık tabanlı özetleyici, mevcut birçok ileri düzey Arapça özetleme sistemine kıyasla daha yüksek

başarı sağlamış ve özelliklerin ağırlıklandırılması ile dilsel belirsizliklerin daha iyi yönetilebileceğini göstermiştir.

3.2.6. Meta-sezgisel yöntemler

Meta-sezgisel algoritmalar, metin özetlemeyi bir optimizasyon problemi olarak ele alma yetenekleri sayesinde Arapça metin özetleme alanında geniş çapta araştırılmıştır. Bu yaklaşımlar, önceden tanımlanmış kurallara dayanmak yerine, genetik algoritmalar, parçacık sürüsü optimizasyonu, karınca kolonisi ve ateşböceği algoritmaları gibi doğadan ilham alan tekniklerle özet kalitesini maksimize eden cümle alt kümelerini ararlar. Kullanılan uygunluk fonksiyonları genellikle içerik kapsayıcılığı, bağlamsal alaka düzeyi veya anlamsal bütünlük gibi ölçütlere dayanır. Bu yöntemler, olası özetlerin geniş çözüm uzayında esnek gezinme imkânı sunar ve Arapçanın karmaşık morfolojik ve sözdizimsel yapısıyla etkili şekilde başa çıkabilir. Bu nedenle, geleneksel yöntemlere kıyasla daha anlamlı ve okunabilir özetler üretebilmektedirler.

Al-Abdallah ve Al-Taani (2017), de benzer şekilde sürü zekâsı tabanlı bir yaklaşım incelemiş ve tek belgeli Arapça metinler için otomatik özetleme amacıyla Parçacık Sürü Optimizasyonu (Particle Swarm Optimization – PSO) algoritması önermiştir. PSO çerçevesinde, her bir "parçacık", olası bir özet (bir dizi cümle) olarak temsil edilmekte ve bu parçacıklar uygunluk (fitness) geri bildirimi doğrultusunda çözüm uzayında hareket ederek en iyi özeti bulmaya çalışmaktadır. Bu yöntem, standart veri kümeleri üzerinde yapılan testlerde daha önceki sezgisel yöntemlere göre üstün sonuçlar elde etmiştir.

[47] Arapça metin özetleme için Ateşböceği Algoritması'nı (Firefly Algorithm) önermiştir. Bu yaklaşımda her cümle, belirli bir kalite fonksiyonu aracılığıyla "ışık yayan" bir ateşböceği olarak modellenmektedir. Yüksek kaliteli cümleler, diğer cümleleri kendilerine doğru çekerek özetin iteratif biçimde geliştirilmesini sağlar. Yapılan deneysel çalışmalar, bu ateşböceği tabanlı tekniğin, geleneksel yöntemlerle karşılaştırıldığında daha anlamlı ve tutarlı özetler ürettiğini ortaya koymuş, bu da doğadan esinlenen meta-sezgisel algoritmaların Arapça metin özetlemede etkili bir alternatif sunduğunu göstermiştir.

Al-Saleh ve Menai (2018), [49] çoklu belge özetleme odaklı çalışmalarında, Arapça metinlerden özet üretmek için Karınca Kolonisi Optimizasyonu (Ant Colony

Optimization- ACO) stratejisini kullanmışlardır. Yöntemlerinde, yapay karıncalar cümleler arasında gezinerek ve her cümlenin önemine orantılı olarak feromon izleri bırakarak özet oluşturmaktadır. Bilgi kapsamını optimize etmeye yönelik bu meta-sezgisel yaklaşım, TAC Multiling Arabic veri kümesinde üstün performans göstermiş; sistemleri, ROUGE değerlendirme sonuçlarında rakip yöntemler arasında en yüksek skorları elde etmiştir. Ayrıca, çok büyük belge koleksiyonlarının işlenmesi gibi önemli zorluklara da çözüm sunmayı hedeflemişlerdir

Baraka ve Al-Breem (2017),[50] çoklu Arapça belge özetlemesi için MapReduce mimarisi kullanan paralel bir Genetik Algoritma (GA) tabanlı yaklaşım önermişlerdir. Sistem, büyük ölçekli verileri işleyebilmek amacıyla GA işlemlerini dağıtık sistemler üzerinde çalıştırmakta; böylece özetleme sürecinde ölçeklenebilirlik, hız ve doğruluk sağlanmakta; aynı zamanda cümle tekrarları azaltılarak okunabilirlik ve bütünlük artırılmaktadır.

Bu çalışma ve benzeri araştırmalar, meta-sezgisel yöntemlerin Arapça metin özetlemede içerik seçiminde küresel arama tekniklerinden faydalanarak özetin doğruluğunu ve akıcılığını artırmada etkili bir çözüm sunduğunu ortaya koymaktadır. Bu çalışmada ise güçlü arama kabiliyeti, çeşitliliği koruma yeteneği ve optimizasyon problemlerinde yaygın başarı sağlaması gibi öne çıkan özelliklerinden dolayı genetik algoritmalar tercih edilmiş ve bu yöntem ayrıca detaylı olarak Bölüm 3.3'te ele alınmıştır.

3.3. Genetik Algoritma Tabanlı Özetleme Yöntemleri

Genetik Algoritmalar (GA), özellikle özellik seçimi ve cümle çıkarımı görevlerinde, metin özetleme alanında etkili optimizasyon araçları olarak ortaya çıkmıştır. Geleneksel yaklaşımların sabit ağırlıklandırma şemalarına ya da sezgisel kurallara dayanmasının aksine, GA yöntemleri, özellik ağırlıklarının ve cümle kombinasyonlarının uyarlanabilir bir şekilde öğrenilmesine olanak tanıyarak, oluşturulan özetlerin kalitesini ve bütünlüğünü artırır.

Bu yaklaşıma yönelik ilk çalışmalardan biri Silla Jr. ve arkadaşları [51] tarafından sunulmuştur. Bu çalışmada özetleme, çoklu betimleyici özelliklere dayanan bir sınıflandırma sistemi olarak ele alınmış ve GA'nın, Naïve Bayes sınıflandırıcısı için gerekli giriş özelliklerini azaltmadaki etkisi araştırılmıştır. Sonuçlar, GA temelli

seçimin performansı olumsuz etkilemediğini, aksine doğruluk kaybı olmadan verimliliği artırdığını göstermiştir.

GA, farklı diller üzerinde yapılan deneylerde de kullanılmıştır. Örneğin, Litvak ve arkadaşları [52], İngilizce ve İbranice haber makaleleri üzerinde 31 farklı cümle puanlama göstergesini uygulayarak, sırasıyla %59.21 ve %44.61 ROUGE-1 sonuçları elde etmiş, bu da GA'nın farklı dil bağlamlarına uyarlanabilirliğini ortaya koymuştur.

Suanmali ve arkadaşları [53], DUC 2002 veri kümesini kullanarak başlık kelimeleri, cümle uzunluğu ve konu kelimeleri gibi özellikleri birleştiren bir modeli test etmiş ve %49.80 doğruluk, %44.64 geri çağırma (recall) ve %46.62 F-skoru elde etmiştir.

Hintçe özetleme alanında da çeşitli çalışmalar, istatistiksel ve dilbilimsel özelliklerin birlikte kullanılmasının önemini ortaya koymuştur. Thaokar ve Malik [54], TF-ISF, cümle uzunluğu, sayısal veriler, dilbilgisel yapı (SOV) ve konu benzerliği gibi özelliklere dayalı bir GA modeli önermiştir. Kadam ve arkadaşları [55] ise konu kelimesi ve özel ad tanımaya odaklanmış, ancak sayısal sonuçlar sunmadan yalnızca ROUGE değerlendirmesini önermiştir.

İngilizce dilinde yapılan çalışmalar da oldukça kapsamlıdır. Vázquez ve arkadaşları [56], başlıkla benzerlik, uzunluk ve kapsamlılık gibi özelliklere odaklanarak ROUGE1'de %48.42ve ROUGE-2'de %22.47 sonuçları elde etmiştir.

Öte yandan, Fattah ve Ren [57], cümle özellikleri için optimal ağırlıkları belirlemek amacıyla GA'yı matematiksel regresyon ile birleştirmiştir. Modellerini 50 manuel olarak özetlenmiş belge üzerinde eğitmiş ve 100 dini makale üzerinde test ederek %30 sıkıştırma oranında ortalama %44.9 doğruluk elde etmiştir.

Arapça bağlamında da GA'nın potansiyelini gösteren çalışmalar mevcuttur. Tanfouri ve arkadaşları[58], EASC veri kümesinden belgeleri özetlemek için GA algoritmasını kullanarak ROUGE-1'de 0.41 ve ROUGE-2'de 0.30 skorlarına ulaşmıştır. Ancak aynı model çok dilli ortamlarda uygulandığında sonuçlar önemli ölçüde düşmüş; ROUGE1 için 0.16 ve ROUGE-2 için 0.029 elde edilmiştir. Bu durum, kaynakları sınırlı dillerde özetleme yapmanın zorluklarını yansıtmaktadır.

Tüm bu çalışmalar, genetik algoritmaların sadece etkili bir cümle seçimi aracı olmadığını, aynı zamanda farklı diller ve ortamlara yüksek uyum sağlayabilme esnekliğini gösterdiğini ortaya koymaktadır. Bu da GA yöntemlerini, özellikle

dikkatlice seçilmiş istatistiksel ve dilbilimsel özelliklerle birleştirildiğinde, otomatik metin özetleme için umut verici bir seçenek haline getirmektedir.

3.4. Arapça ATS Yaklaşımlarının Karşılaştırmalı Değerlendirmesi

Arapçanın morfolojik yapısı ve kaynak yetersizliği gibi zorluklara rağmen, çok sayıda çalışma özellikle çıkarımsal yöntemler üzerine yoğunlaşmıştır.

Denetimli öğrenme ve skor tabanlı yöntemleri kullanan Qaroush ve arkadaşları[59], EASC veri kümesinde Rouge-1: 0.643 ve Rouge-2: 0.617 skorları elde ederek istatistiksel ve anlamsal özniteliklerin birleşiminin gücünü göstermiştir. Aynı şekilde Abdulateef ve ark. [60], Word2Vec temelli vektörleştirme ve kümeleme tekniklerini bir araya getirerek çok belgeli özetlemede tekrarları azaltmayı hedeflemiş; Rouge-1: 0.664 ve Rouge-2: 0.559 skorlarına ulaşarak önemli başarı elde etmiştir. AraBERT modelini kümeleme algoritmaları ile birleştiren Nada ve ark. [61], Rouge-1: 0.54 ve Rouge-2: 0.51 sonuçlarıyla çıkarımsal yöntemlerde semantik temsillerin rolünü öne çıkarmıştır.

Elmadani ve arkadaşları [62], BERT modelini Arapça özetleme için ince ayar yaparak Rouge-1: 0.42, Rouge2: 0.2459 ve Rouge-L: 0.422 skorlarına ulaşmıştır. Çok dilli BERT kullanan Elayeb ve arkadaşları [38]ise iki farklı algoritma ile sırasıyla Rouge-1: 0.75 ve BLEU-1: 0.47; Rouge-1: 0.74 ve BLEU-1: 0.49 puanları elde etmiştir.

Soyutlayıcı özetleme yöntemleri de gelişim göstermektedir. Seq2Seq modelleri kullanan Wazery ve ark. [63] ile Suleiman ve Awajan [64] LSTM ve dikkat mekanizmaları ile yapılandırılmış encoder-decoder mimarileri kullanmış ve Rouge-1: 0.5149, Rouge-2: 0.12, Rouge-L: 0.343 ve BLEU: 0.41 gibi sonuçlar elde etmişlerdir. Bu tür mimariler, özetlerin akıcılığını ve bağlamsal tutarlılığını artırmada etkili olmuştur.

2023 yılında Al-Khassawneh ve Hanandeh, [13] üçgen tabanlı grafik puanlamasına dayalı yeni bir çıkarımsal yöntem önermiştir. Yöntemleri, cümlelere grafik yapısı ve anlamsal öneme göre ağırlık atamış ve EASC veri kümesi üzerinde ROUGE-2 F1 skorunda %61.7'ye ulaşmıştır.

Değerlendirme ölçütleri bakımından çalışmaların büyük çoğunluğu ROUGE-1, ROUGE-2, ROUGE-L ve BLEU metriklerini kullanmaktadır. Elbarougy ve arkadaşları [65] gibi bazı araştırmalar ise modifiye edilmiş PageRank algoritmalarıyla

çıkarımsal özetleme yaparken Rouge ve F-ölçütü skorlarına odaklanmıştır. Ayrıca, verisetleri açısından EASC, SANAD, AHS, AMN ve KALIMAT gibi kaynaklar öne çıkmakta; bazı araştırmalar kendi verisetlerini manuel olarak toplamıştır ([61], [46] [66]).

Tablo 3.2. Arapça metin özetleme çalışmaları: kategorik özet tablo.

Kaynak	Yıl	Yaklaşım Türü	Yöntem	Veri Seti	F measure	Tür	CR %
[67]	2013	Ontoloji	OSSAD	EASC	49.8	SD	40
[68]	2014	İstatistiksel	K-Means Algoritması	EASC	60.0	MD	—
[69]	2015	Leksikal	Leksikal Bağdaşıklık + Metinsel Çıkarım	Arabic textual entail.	69.98	SD	—
[70]	2016	Hibrit	Hibrit Yaklaşım	EASC	54.76	SD	40
[71]	2017	Meta-sezgisel	PSO	EASC	55.32	SD	40
[72]	2018	Meta-sezgisel	Genetik Algoritmalar (GA)	EASC	60.5	SD	40
[59]	2020	Denetimli + Skor	İstatistiksel + Anlamsal özellikler	EASC	64.3	SD	—
[73]	2020	Kümeleme	Word2Vec + Kümeleme	Çoklu belge	66.4	MD	—
[74]	2021	Soyutlayıcı (Seq2Seq)	Encoder–decoder (LSTM + Attention)	—	51.5	SD	—
[75]	2022	Statistical and Graph-Based	Statistical and Graph-Based	EASC	55.36	SD	45
[13]	2023	Graf-tabanlı	Üçgen tabanlı grafik puanlama	EASC	61.7	SD	—
[76]	2024	Graf-tabanlı	PageRank + Word Embedding	EASC	65.219	SD	40

4. GENETİK ALGORİTMA TABANLI HİBRİT ARAPÇA METİN ÖZETLEME

Bu bölümde, önerilen hibrit özetleme sisteminin metodolojisi ayrıntılı olarak sunulmaktadır. Öncelikle Arapça metinler için uygulanan ön işleme adımları açıklanmakta, ardından cümleleri temsil eden yapısal ve anlamsal öznitelikler tanıtılmaktadır. Daha sonra, bu özniteliklerin ağırlıklarının rastgele arama algoritması ile optimize edilmesi ve genetik algoritma aracılığıyla özet cümlelerinin seçilme süreci ele alınmaktadır. Son olarak, sistemin değerlendirilmesi ve kullanılan karşılaştırmalı yöntemlere ilişkin bilgiler verilmektedir.

Bu araştırma, Arapça metinler üzerinde çıkartımsal özetleme gerçekleştirmek amacıyla, istatistiksel ve anlamsal özelliklerin katkısını optimize eden melez (hibrit) bir sistem önermektedir. Sistem; veri hazırlama, özellik çıkarımı, ağırlık optimizasyonu ve genetik algoritma ile cümle seçimi gibi çok aşamalı bir iş akışına sahiptir.

Çalışmada kullanılan hibrit yapının genel mimarisi Şekil 4.1'de sunulmuştur

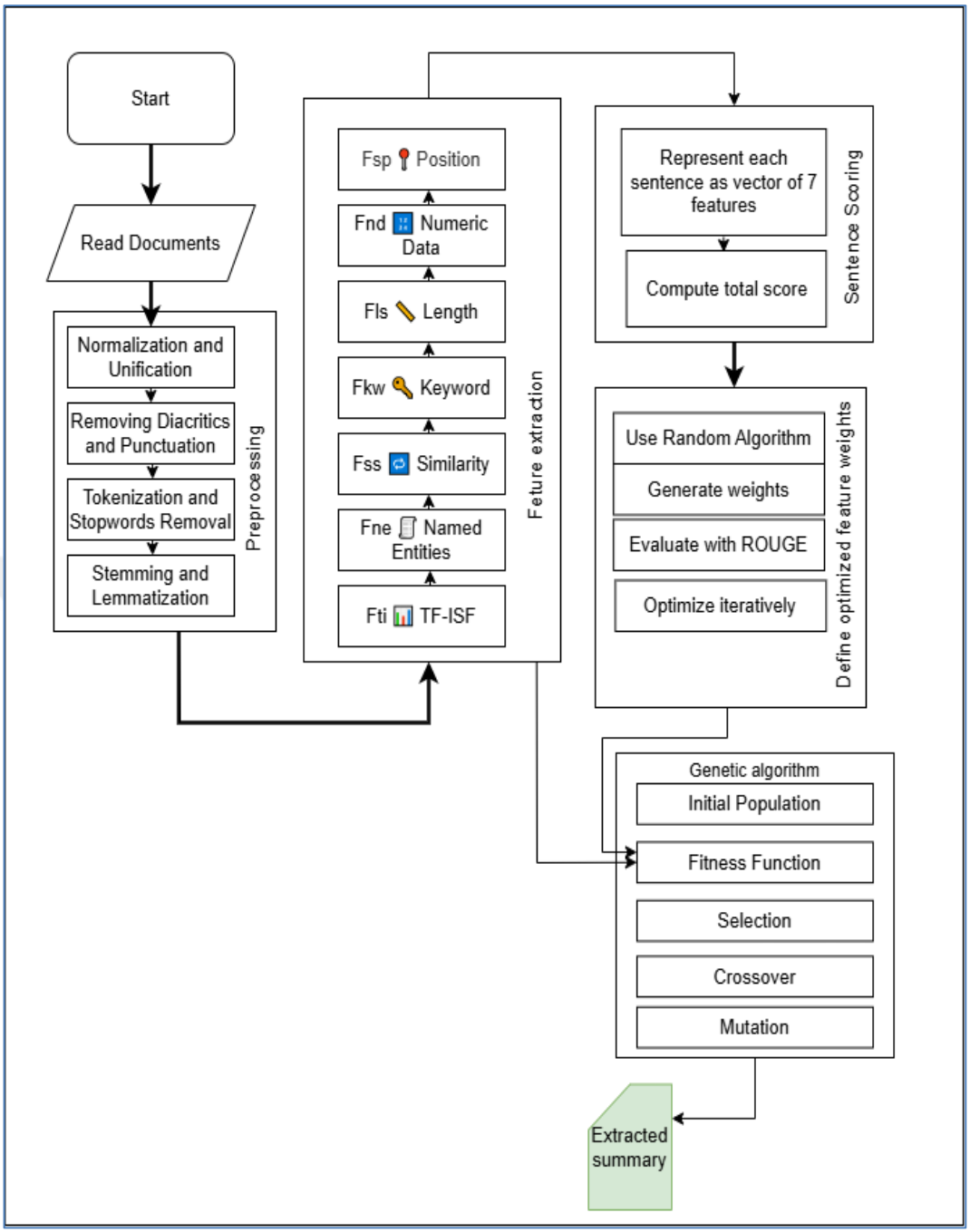
İlk aşamada, Arapça belgeler işlenmeden önce temizlenir; elif harfleri standartlaştırılır, harekeler, noktalama işaretleri, bağlantılar ve gereksiz boşluklar kaldırılır. Ardından, metinler NLTK gibi araçlar kullanılarak cümlelere ayrılır. Her cümle, yapısal ve anlamsal önemini yansıtan bir dizi özellikle temsil edilir. İstatistiksel özellikler arasında cümle konumu (F_sp), uzunluğu (F_nd), son cümle olup olmama (F_ls), anahtar kelime varlığı (F_kw) ve başlıkla benzerlik (F_ti) bulunurken; anlamsal özellikler olarak diğer cümlelerle benzerlik (F_ss), adlandırılmış varlıkların varlığı (F_ne) ve bağlaç veya güçlü noktalama işaretleri (F_eh) dikkate alınır.

Ağırlıkların optimizasyonu için ilk olarak Rastgele Arama (Random Search**)** algoritması kullanılır. Bu aşamada, özellikler için çok sayıda rastgele ağırlık =-

=kombinasyonu oluşturulur. Her bir ağırlık kombinasyonu için, metindeki her cümlenin puanı; o cümleye ait tüm özellik değerlerinin, ilgili ağırlıklarla çarpılıp bir araya getirilmesiyle hesaplanır. Bu işlem sonucunda elde edilen toplam puan, cümlenin özet açısından önem derecesini temsil eder. Elde edilen puanlara göre cümleler sıralanır ve en yüksek %30'luk kısmı özet olarak seçilir. Bu özet, referans özetle karşılaştırılır ve ROUGE-1, ROUGE-2 ve ROUGE-L metriklerine göre F1 puanı hesaplanır. Bu işlem, her ağırlık kombinasyonu için tekrarlanır ve en yüksek ROUGE puanını üreten ağırlık kombinasyonu sistem tarafından en iyi konfigürasyon olarak seçilir. Bu yaklaşım, ağırlıkların etkisinin doğrudan özet kalitesi üzerinden ölçülmesini sağlar.

Rastgele arama sonucunda elde edilen en iyi ağırlık kombinasyonu temel alınarak, ikinci aşamada Genetik Algoritma (GA) uygulanır. Bu algoritmada, her potansiyel özet bir "kromozom" olarak temsil edilir ve cümle indekslerinden oluşur. Başlangıç popülasyonu rastgele oluşturulur. Her kromozomun uygunluk değeri, hem toplam cümle skorları hem de özet içindeki cümleler arasındaki anlamsal bütünlük dikkate alınarak hesaplanır.GA süreci; turnuva seçimi (tournament selection), çaprazlama (crossover) ve mutasyon (mutation) adımlarını içerir. Her nesilde en iyi çözümler korunarak yeni nesiller üretilir.

Son aşamada sistem, Cosine Similarity ve AraBERT yöntemleri ile karşılaştırılmıştır. Elde edilen sonuçlar, Cosine Similarity'nin daha uygun olduğunu göstermiştir. Bu nedenle, önerilen sistemde temel yöntem olarak Cosine Similarity tercih edilmiştir.



Şekil 4.1. Sistem mimarisi.

4.1. Arapça Metinler için Ön İşleme Adımları

Arapça dili, zengin morfolojik yapısı ve karmaşık sözdizimsel esnekliği ile tanımlanan bir dildir. Bu nedenle, herhangi bir önişleme adımı uygulanmaksızın doğrudan bilgi erişimi amacıyla Arapça belgelerle çalışmak, metnin analizini daha karmaşık hale getirebilir ve kullanıcıya hatalı ya da yanıltıcı sonuçlar sunabilir. Bu sorunu önlemek adına, özetleme aşamasına geçmeden önce ve belge yükleme adımının hemen ardından

bazı dil işleme süreçlerinin gerçekleştirilmesi gereklidir. Bu noktada, belge sisteme alınır ve öznitelik çıkarımına hazırlık amacıyla işlenir.

4.1.1. Belgelerin içe aktarılması

Bu aşamada, özetleme işlemine tabi tutulacak Arapça belgeler sisteme yüklenir ve analiz için hazırlanır. Çalışmada, farklı konuları kapsayan 153 belgeden oluşan EASC veri kümesi kullanılmıştır. Süreç, belgelerin sisteme aktarılması ve her belgeden Arapça metnin çıkarılmasıyla başlar. Bu adım, tüm önişleme adımlarının temelini oluşturur ve verilerin tutarlı bir biçimde analiz edilmesini sağlar. Belgeler başarıyla yüklendikten sonra sistem, metinlerin standartlaştırılması ve birleştirilmesi aşamasına geçer.

4.1.2. Hareke ve noktalama işaretlerinin kaldırılması

Arapça metinlerde yer alan harekeler (tashkīl) ve noktalama işaretleri, dilin telaffuzuna ve anlamın inceliklerine katkı sağlasa da, otomatik özetleme sistemleri için fazladan gürültü oluşturabilir. Bu nedenle, önişleme aşamasında bu tür işaretlerin metinden çıkarılması, metin analizi süreçlerini sadeleştirerek daha tutarlı özniteliklerin çıkarılmasına olanak tanır. Hareke işaretleri örneğin fetha (ً), damma (ٌ), kesra (ٍ), şedde (ّ), sükûn (ْ), tenvin (ًٌٍ) vb. metindeki kelimelerin farklı biçimlerde temsil edilmesine neden olabilir. Benzer şekilde, noktalama işaretleri (., ،؟ ،؛!, vb.) ya da gereksiz boşluklar da cümle yapısında yapay ayrışmalara yol açabilir. Bu nedenle, özetleme sisteminin hem anlamsal hem de yapısal bütünlüğünü sağlamak adına bu işaretlerin temizlenmesi tercih edilir. Aşağıda verilen örnek tablo, hareke ve noktalama işaretlerinin kaldırılmasının metin üzerinde yarattığı değişimi göstermektedir:

4.1.3. Normalizasyon

Normalizasyon, Arapça metinlerin ön işleme sürecinde temel bir adımdır ve özetleme sisteminin genel performansını doğrudan etkileyen bir aşamadır. Bu işlem, metinleri daha tutarlı ve analiz edilebilir hâle getirmek amacıyla çeşitli dönüşümler uygulanarak dilin biçimsel çeşitliliğini sadeleştirmeyi amaçlar. Arapça'da, sesli harfleri ve gramer yapılarını göstermek için kullanılan harekeler (örneğin: Fetha ◌َ, Damma ◌ُ, Kesra ◌ِ, Sükun ◌ْ vb.) bulunur. Ancak bu işaretler, aynı kelimenin cümledeki konumuna göre değişebildiği ve farklı anlamlara yol açabildiği için, özetleme işlemine başlamadan önce metinlerden çıkarılmaları daha sağlıklı sonuçlar verir.

Normalizasyon sürecinde uygulanan işlemler arasında şunlar yer alır: metinlerdeki harekelerin (tashkil) kaldırılması, noktalama işaretlerinin ve web bağlantılarının silinmesi, yinelenen boşlukların temizlenmesi, çoklu nokta karakterlerinin sadeleştirilmesi ve farklı "elif" biçimlerinin (أ، إ، آ، ٱ) tek bir standart forma (ا) dönüştürülmesi. Bu işlemler, metinlerdeki biçimsel gürültüyü azaltarak, öznitelik çıkarımında daha tutarlı ve güvenilir sonuçların elde edilmesine katkı sağlar.

Aşağıda, harekelerin kaldırılmasının özetleme sürecine etkisini göstermek amacıyla basit bir örnek sunulmuştur:

Örnek metin (harekeli):
"إِنَّ اللهَ غَفُورٌ رَّحِيمٌ"

Normalizasyon uygulanmış metin (harekesiz):
"ان الله غفور رحيم"

Bu örnekte görüldüğü gibi, harekeler kaldırıldığında metin daha sade bir biçime kavuşur ve içerik olarak anlamını korur. Bu tür sadeleştirme işlemleri, özetleme sistemlerinde kullanılan istatistiksel ve anlamsal özniteliklerin daha kararlı çalışmasına katkı sağlar

Tablo 4.1. Hareke ve noktalama işaretlerinin kaldırılması örneği.

Orijinal Metin	**Temizlenmiş Metin**
.وَقَالَ الرَّسُولُ: يَا رَبِّ إِنَّ قَوْمِي اتَّخَذُوا هَذَا الْقُرْآنَ مَهْجُورًا	وقال الرسول يا رب ان قومي اتخذوا هذا القران مهجورا
.إِنَّ اللَّهَ غَفُورٌ رَّحِيمٌ	ان الله غفور رحيم
!فَصَبْرٌ جَمِيلٌ. وَاللَّهُ الْمُسْتَعَانُ عَلَى مَا تَصِفُونَ	فصبر جميل والله المستعان على ما تصفون

Bu örneklerde görüldüğü gibi hem harekeler hem de noktalama işaretleri kaldırıldığında metnin anlamı korunur, ancak metin daha sade ve işlemeye uygun bir forma ulaşır. Bu sadeleştirme işlemi, cümle skorlama ve öznitelik çıkarımı gibi sonraki adımlarda daha verimli sonuçlar elde edilmesini sağlar

4.1.4. Tokenizasyon ve stopword temizleme

Tokenizasyon (kelime ayırma), doğal dil işleme uygulamalarında temel bir önişleme adımıdır. Bu adımda, metin yapısı analiz edilerek cümleler içindeki kelimeler veya kelime grupları anlamlı birimlere (token'lara) ayrılır. Arapça gibi morfolojik açıdan zengin dillerde kelimeler genellikle birden fazla ek, bağlaç ya da zamir içerdiği için tokenizasyon işlemi oldukça önemlidir. Uygun şekilde tokenleştirilmemiş metinler,

hem öznitelik çıkarımı hem de cümle değerlendirme aşamalarında sistemin performansını düşürebilir.

Tokenleştirme işlemi, çoğunlukla boşluk, noktalama işaretleri ve sözdizimsel kurallar aracılığıyla gerçekleştirilir. Ancak Arapça'da bazı zamirler ve bağlaçlar kelimeye bitişik yazıldığından (örneğin: "وقاله" = "ve dedi ona"), dilbilgisel analizle desteklenen özel ayırıcılar kullanmak gerekebilir. Bu nedenle Arapça metinlerde tokenizasyon, hem sözdizimsel hem de morfolojik analizle desteklenmelidir.

Stopword temizliği ise tokenleştirme sonrasında gerçekleştirilen önemli bir adımdır. Stopword'ler; bağlaçlar, zamirler, edatlar, zaman zarfları gibi metin içinde yüksek sıklıkla geçen ancak anlam taşımayan kelimelerdir. Örnek olarak "bu", "ve", "için", "ile", "o zaman", "şöyle ki" gibi sözcükler gösterilebilir. Bu tür kelimeler, istatistiksel olarak metnin bilgi yoğunluğunu azaltır ve anahtar kavramların belirlenmesini zorlaştırır. Bu nedenle özetleme gibi anlam merkezli görevlerde stopword'lerin temizlenmesi özetin doğruluğu ve bilgi değeri açısından oldukça faydalıdır.

Stopword temizliği için genellikle hazır listeler kullanılır. Ancak Arapça gibi bağlam duyarlı ve lehçe çeşitliliği yüksek dillerde, bu listeler özelleştirilmeli ve uygulandığı bağlama göre ayarlanmalıdır. Aşağıdaki tablo, tokenizasyon işlemi ve ardından yapılan stopword temizliğinin bir örneğini göstermektedir:

Tablo 4.2. Tokenizasyon ve stopword temizliği örneği.

Orijinal Cümle	**Tokenize Edilmiş Sözcükler**	**Stopword'ler (Silinecek)**	**Temizlenmiş Tokenler**
وقال النبي في ذلك الوقت إن الصبر مفتاح الفرج.	وقال، النبي، في، ذلك، الوقت، إن، الصبر، مفتاح، الفرج	في، ذلك، إن	وقال، النبي، الصبر، مفتاح، الفرج
هذا هو الطريق الذي يؤدي إلى النجاح الحقيقي.	هذا، هو، الطريق، الذي، يؤدي، إلى، النجاح، الحقيقي	هذا، هو، الذي، إلى	الطريق، يؤدي، النجاح، الحقيقي
ما فائدة العلم إذا لم يعمل به الإنسان؟	ما، فائدة، العلم، إذا، لم، يعمل، به، الإنسان	ما، إذا، لم، به	فائدة، العلم، يعمل، الإنسان

Bu tablo sayesinde şu süreçler netleşmektedir:

- İlk sütunda, orijinal metin örnekleri yer almaktadır.
- İkinci sütunda, bu cümlelerin token'lara (kelimelere) ayrılmış hali görülmektedir.
- Üçüncü sütunda, çıkarılması gereken stopword'ler belirtilmiştir.

4.1.5. Stemming ve lemmatization

Stemming ve lemmatization, doğal dil işleme sistemlerinde sözcüklerin kök veya temel biçimlerine indirgenmesi için uygulanan önemli işlemlerdir. Özellikle Arapça gibi kök-temelli ve eklemeli bir dilde, kelimeler çok çeşitli biçimlerde üretilebildiği için bu tür işlemler; öznitelik çıkarımı, metin karşılaştırma ve anlam analizi gibi birçok NLP görevinin doğruluğunu doğrudan etkiler.

Stemming (kök bulma), bir kelimenin farklı eklerle türetilmiş biçimlerinden, sabit olan kök kısmını elde etme işlemidir. Bu işlem genellikle algoritmik bir yaklaşım izler ve kurallara dayalı olarak çalışır. Arapça'da bu işlem sırasında genellikle son harfler veya fiil kalıplarından hareketle kelimenin kök yapısı çıkarılır. Ancak stemming işlemi her zaman doğru kökü vermeyebilir çünkü bağlamı dikkate almaz; bu yüzden bazı durumlarda anlamlı olmayan veya eksik kökler elde edilebilir.

Öte yandan lemmatization (gövdeleme), sözcüğün sözlükteki gerçek kök biçimini (lemma) bulmayı hedefler. Bu işlem, kelimenin bağlamını, sözdizimsel rolünü ve dilbilgisel bilgisini dikkate alır. Arapça lemmatization, hem kök hem de kalıp analizi gerektirdiği için daha karmaşık ama daha doğru sonuçlar verir. Gövdeleme, özellikle anlam temelli özniteliklerin çıkarımı için oldukça önemlidir.

Aşağıdaki tablo, aynı kelime üzerinde yapılan stemming ve lemmatization işlemlerinin farklarını örneklerle göstermektedir:

Tablo 4.3. Stemming ve lemmatization karşılaştırması.

Orijinal Kelime	Anlamı	Stemming Sonucu	Lemmatization Sonucu	Açıklama	Orijinal Kelime
المدارس	Okullar	مدرس	مدرسة	"Stemming" kelimenin köküne indirgerken, "Lemmatization" sözlük biçimini verir.	المدارس
يُدرّسون	Onlar ders veriyorlar	درس	يُدرّسُونَ	Stem işlemi sadece kökü verir, lemma ise tam fiil biçimini korur.	يُدرّسون
المذاكرة	Ders çalışmak	ذكر	ذاكر	"Stem" kökü çıkarır ama bağlamı dikkate almaz; lemma, fiil anlamını korur.	المذاكرة
كتبتِ	Sen yazdın (kadın)	كتب	كتبْتِ	Lemma, sözcüğün zamir ve zaman bilgisini korur.	كتبتِ

Tabloda görüldüğü üzere, stemming işlemi daha basit ve hızlıdır; kelimeyi bir kurala göre indirger. Ancak bu işlem çoğu zaman anlam bilgisini yitirir ve özetleme sistemlerinde kelime benzerliklerini doğru şekilde yakalayamaz. Lemmatization ise daha karmaşık olsa da sözcüğün bağlamına uygun anlamlı bir biçim sağlar. Bu nedenle bu çalışmada her iki yöntem bir arada kullanılarak hem istatistiksel hem de anlamsal özniteliklerin daha doğru şekilde çıkarılması hedeflenmiştir.

Sonuç olarak, kök bulma ve gövdeleme işlemleri, Arapça dilinin yapısal zorlukları nedeniyle, otomatik özetleme sistemlerinde vazgeçilmez adımlar olarak değerlendirilir. Bu işlemler, aynı kavramı temsil eden fakat farklı biçimlerde geçen kelimeleri normalize ederek öznitelik vektörlerinin kalitesini artırır ve özetin bilgi yoğunluğunu optimize eder.

4.2. Öznitelik Çıkarma Aşaması

Öznitelik çıkarımı aşaması, önişlemden geçirilmiş ve temizlenmiş Arapça cümlelerin özetleme sistemine uygun biçimde analiz edildiği kritik bir adımdır. Bu aşamada, her bir cümle, 0 ile 1 arasında normalleştirilmiş değerler içeren sayısal bir öznitelik vektörü olarak temsil edilir. Bu vektör, cümlelerin içerik açısından önem düzeyini hesaplamak ve onları sıralamak amacıyla kullanılır. Sistemin bu süreçte dikkate aldığı her öznitelik, metnin hem yapısal hem de anlamsal boyutuna ışık tutmaktadır. Kullanılan öznitelikler; cümle konumu, sayı içerme durumu, cümle uzunluğu, anahtar kelime sıklığı, anlamsal benzerlik, özel ad (named entity) varlığı, TF-ISF skoru gibi çeşitli dilsel ve istatistiksel unsurları kapsamaktadır. Her bir öznitelik, daha sonra ağırlıklandırılarak, cümlenin genel önem puanına katkı sağlar. Böylece sistem, özetlemeye en uygun cümleleri belirlemek üzere nesnel ve çok boyutlu bir değerlendirme gerçekleştirmiş olur. Önceki çalışmalar bu özniteliklerin değerlendirme fonksiyonuna entegre edilmesinin, özellikle evrimsel algoritmalar veya melez yaklaşımlar çerçevesinde kullanıldığında, elde edilen özetlerin kalitesini önemli ölçüde artırdığını göstermiştir.[77][78][79]

Tablo 4.4. Özelliklerin işlevsel tanımları.

Feature (Özellik)	Açıklama
F_sp	Cümlenin paragraftaki konumunu belirler.
F_nd	Sayısal verileri tespit eder.
F_ls	Cümle uzunluğunu ölçer ve normalize eder.
F_ss	Diğer cümlelerle benzerliğini ölçerek merkezîliği belirler
F_kw	Anahtar kelimelerin varlığını kontrol eder.
F_ne	Kişi, yer, organizasyon gibi özel isimleri analiz eder.
F_eh	İngilizce kelimelerin varlığını kontrol eder.
F_ti	Cümlenin önemini TF-IDF ölçütüyle belirler.

4.2.1. Cümlenin paragraftaki konumu (Fsp)

Bir cümlenin paragraftaki konumunun onemini belirler. Örneğin, paragrafın başlangıcında yer alan cümleler, belgenin ana temasını taşıdığı için özet içinde bulunma olasılığı daha yüksektir. [78] Cümlenin paragraf içindeki konumu, aşağıdaki Denklem) ile hesaplanır:

$$s = \frac{\mathrm{n} - \mathrm{i}}{n^i} \tag{4.1}$$

Burada:

- n: Paragraftaki toplam cümle sayısı,
- i: 0 ile n arasında değişen cümle sırası,
- s_i: Paragraftaki i'nci cümleyi temsil eder.

4.2.2. Cümledeki sayısal veriler (Fnd)

Arapça metinlerde geçen sayısal veriler, zaman bilgisi ("2"), miktar ("3-4 مرات ") - (yüzde ifadeleri) ("%12 " hasta) . içerebilir.[80] Sayısal veri içeriği, aşağıdaki Denklem (2) ile hesaplanır:

$$\mathrm{nd_{si}} = \frac{\mathrm{n\_si}}{\mathrm{w\_si}} \tag{4.2}$$

Burada:

- n_si: Cümlede geçen toplam sayısal değer sayısı,
- w_si: Cümledeki toplam kelime sayısı,

- nd_{si}: Cümledeki sayısal veri yoğunluğu.

Bu şekil, her cümle için F_nd özniteliğinin hesaplanmasıyla elde edilen sonuçları göstermektedir.

F_nd, cümlede sayı içeren kelimelerin toplam kelime sayısına oranını ifade eder. Görüldüğü üzere, birinci cümlede herhangi bir sayısal veri bulunmadığı için değeri 0.00 olarak hesaplanmıştır. Diğer cümlelerde ise yer alan sayılar (örneğin, parasal ifadeler veya yüzdeler) doğrultusunda oranlar değişmektedir.
Bu özellik, cümlenin sayısal bilgi açısından zenginliğini gösteren önemli bir ölçüttür ve otomatik özetleme sürecinde cümlenin bilgilendiriciliğini değerlendirmeye katkı sağlar.

Bu çalışmada, sayısal veri oranını temsil eden F_nd özniteliğinin hesaplama yöntemi açıklanmaktadır.

Bu oran, cümlede yer alan sayısal bilginin yoğunluğunu nicel olarak ifade eder.

4.2.3. Cümle uzunluğu (Fls)

Cümlelerin aşırı kısa veya aşırı uzun olduğunda, özetleme için uygun olmayabilir. Cümle uzunluğu, aşağıdaki Denklem (3) ile hesaplanır: [81]

$$ln_{si} = \frac{w_{si}}{lg_{si}} \tag{4.3}$$

Burada:

- w_{si}: Cümledeki toplam kelime sayısı,
- lg_{si}: En uzun cümledeki toplam kelime sayısı,
- ln_{si}: Cümlenin uzunluk değeri.

Bu şekil, cümle uzunluğunu temsil eden F_ls özniteliğinin hesaplama sonuçlarını göstermektedir.

F_ls, her cümledeki kelime sayısının, aynı belgede bulunan en uzun cümleye oranı olarak hesaplanır.

4.2.4. Cümledeki anahtar kelimeler (Fkw)

Anahtar kelimeler, metnin ana temasını ve bilgi yoğunluğunu yansıtan önemli göstergelerdir. Otomatik özetleme sürecinde, anahtar kelime içeren cümleler

genellikle metnin odak noktalarını temsil ettiği için özetlerde öncelikli olarak seçilmektedir. Arapça metinlerde, bu kelimeler çoğunlukla konuyu doğrudan işaret eden ve yüksek bilgi değeri taşıyan sözcüklerden oluşmaktadır.

Bu çalışmada, Arapça belgelerden anahtar kelimeler (KWE) çıkarılmıştır. Bunun için çok dilli bir dil modeli (OpenAI GPT) kullanılmış, her cümle için en temsili kelimeler otomatik olarak belirlenmiştir. Elde edilen çıktılar daha sonra filtrelenerek yalnızca metinde gerçekten bulunan kelimeler dikkate alınmıştır. Böylece, özetleme sürecinde güvenilir bir anahtar kelime tabanı oluşturulmuştur.[82]

4.2.5. Cümle benzerliği (Fss)

Cümle benzerliği, her cümlenin metnin genel içeriğine olan yakınlığını ölçen bir özelliktir. Bunun için her cümle TF-IDF vektörüyle temsil edilir ve belgenin merkez vektörü, tüm cümlelerin ortalaması alınarak oluşturulur. Daha sonra her cümlenin bu merkeze olan benzerliği kosinüs benzerliği ile hesaplanır. [78]

$$F_{ss}(s_i) = \cos vi, c \tag{4.4}$$

Burada vivi, sisi cümlesinin TF-IDF vektörünü; cc ise belgenin merkez vektörünü ifade eder. Elde edilen değerler min–max yöntemiyle [0,1][0,1] aralığına normalize edilerek F_ss skorları elde edilir. Bu yöntem, metnin genel anlamını en iyi yansıtan cümleleri öne çıkarır.

4.2.6. Cümlede geçen özel tabirler (Fne)

Özetleme sürecinde cümlelerin içerdiği özel tabirler (Named Entities – NE), bilginin ayırt edici unsurlarını ortaya koyar. Kişi adları, kurumlar, yer isimleri, hastalık adları veya ürünler gibi tabirler, metnin bağlamını güçlendirir ve özetin daha temsili olmasına katkı sağlar. Arapça metinlerde özel tabirlerin doğru biçimde belirlenmesi, özellikle haber ve sağlık içeriklerinde, bilgi yoğunluğunu öne çıkarmak açısından önemlidir. Bu çalışmada özel tabirler; hastalıklar ("الربو" ,"السكري"), semptomlar ("الضعف" ,"الإرهاق"), yiyecek ve içecekler ("الجزر" ,"الفلفل الأسود") ve kişiler (" طبيب", "مريض"") gibi kategoriler altında değerlendirilmiştir. Bu tür tabirlerin yer aldığı cümleler, bağlama özgü bilgiler sundukları için özetleme sürecinde daha anlamlı kabul edilmiştir..[83]

4.2.7. Cümlede geçen İngilizce-Arapça kelimeler (Feh)

Arapça cümlelerde sıklıkla İngilizce kelimeler kullanılmaktadır. Örneğin:

"سرطان البروستات) cancer Prostate هو مرض خطير"

("prostate cancer هو مرض خطير "-" Prostat kanseri tehlikeli bir hastalıktır.")

Bu cümlede "Prostate cancer" kelimesi, İngilizce-Arapça birleşik bir terimdir ve Arapça kelime bankalarında yer almayabilir. Ancak, cümlenin önemini belirlemede oldukça faydalıdır.

Matematiksel olarak, Arapça bir cümlede geçen İngilizce kelimelerin oranı aşağıdaki

$$d_{\{eh\_s_i\}} = \frac{t_{\{eh\_s_i\}}}{l_{\{n\_s_i\}}} \tag{4.5}$$

Burda:

- $t_{\{eh\_s_i\}}$: Cümlede geçen İngilizce-Arapça kelimelerin toplam sayısı.
- $l_{\{n\_s_i\}}$: Cümlenin uzunluğu (toplam kelime sayısı).
- $d_{\{eh_s_i\}}$İngilizce-Arapça kelimelerin oranı.

4.2.8. Terim frekansı ve ters cümle frekansı

- TF-ISF, Terim Frekansı (TF) ve Ters Cümle Frekansı (ISF) anlamına gelir.
- TF (Term Frequency): Bir kelimenin belirli bir cümlede kaç kez geçtiğini ölçer. Ancak, tüm kelimelerin eşit derecede önemli olduğu varsayımı gerçekçi değildir.
- ISF (Inverse Sentence Frequency): Bir kelimenin metindeki önemini belirlemek için Bir kelimenin çok fazla cümlede geçmesi, onun özetleme açısından daha az kullanılır.[81]

Bu değerler aşağıdaki denklemlerle hesaplanır:

TF-ISF, bir kelimenin belirli bir cümledeki önemini hem yerel hem genel ölçekte değerlendirir. Hesaplama iki adımdan oluşur:

$$Tf\_i = \frac{fq\hat{}j\_i}{w_{sj}} \tag{4.6}$$

$$Isf_i = \log\left(\frac{tn_s}{ns\_ti}\right) \tag{4.7}$$

Burda:

- fq_i^j: i. kelimenin j. cümlede geçme sıklığı.
- w_{sj}: j. cümledeki toplam kelime sayısı.
- tn_s: Toplam cümle sayısı.
- ns_ti: i. kelimenin geçtiği cümle sayısı.
- Tf_i: i. kelimenin terim frekansı.
- Isf_i: i. kelimenin ters cümle frekansı.

Bu ölçümler, Arapça metinlerin özetlenmesi ve önemli bilgilerin belirlenmesi için oldukça önemlidir.

4.3. Özniteliklerin Vektör Gösterimi

Bu çalışmada geliştirilen otomatik özetleme sisteminde, her cümle, daha önce çıkarılmış olan sekiz özellikten oluşan bir vektörle temsil edilmiştir. Bu yaklaşımın temel amacı, metinsel ifadeleri sayısal bir forma dönüştürerek algoritmik değerlendirme ve sıralama süreçlerine uygun hale getirmektir. Özellik değerleri 0 ile 1 arasında normalize edilerek, farklı ölçeklerdeki değişkenlerin karşılaştırılabilirliği sağlanmıştır. Her cümle için elde edilen öznitelik vektörü şu genel formdadır:

$$\vec{F}\{si\} = [Fsp(si), Fnd(si), Fls(si), Fkw(si), Fss(si), Fne(si), Feh(si), Fti(si)] \quad (4.8)$$

Burada:

$\vec{F}\{si\}$: si cümlesinin özellik vektörüdür,

Her bir F bileşeni, ilgili özelliğin cümledeki normalleştirilmiş değerini gösterir

Bu yapı, her belgedeki tüm cümlelerin sistematik olarak karşılaştırılabilmesini ve sonraki aşamada ağırlıklandırılarak puanlanmasını mümkün kılar.

4.4. Skor Hesaplama ve Cümlelerin Sıralanması

Sistemin özetleme kararı verebilmesi için her cümleye sayısal bir önem puanı atanmaktadır. Bu puan, o cümleye ait özniteliklerin belirli ağırlıklarla değerlendirilmesi yoluyla hesaplanır. Her cümle, sekiz farklı özelliğe (konum, uzunluk, anahtar kelime yoğunluğu, benzerlik vb.) karşılık gelen sayısal bir vektör

olarak temsil edilir. Skorlama işlemi, bu vektördeki her özelliğin göreceli etkisini yansıtan ağırlıklarla birleştirilmesiyle gerçekleştirilir.

$$\text{Score(si)} = \sum_{k=1}^{n} W_k * F_{k(si)} \tag{4.9}$$

Burada:

si :i'nci cümle

$F_{k(si)}$: cümleye ait k'ncı özellik (örneğin anahtar kelime yoğunluğu, uzunluk, benzerlik vb.)

W_k: k'ncı özelliğin ağırlığıdır

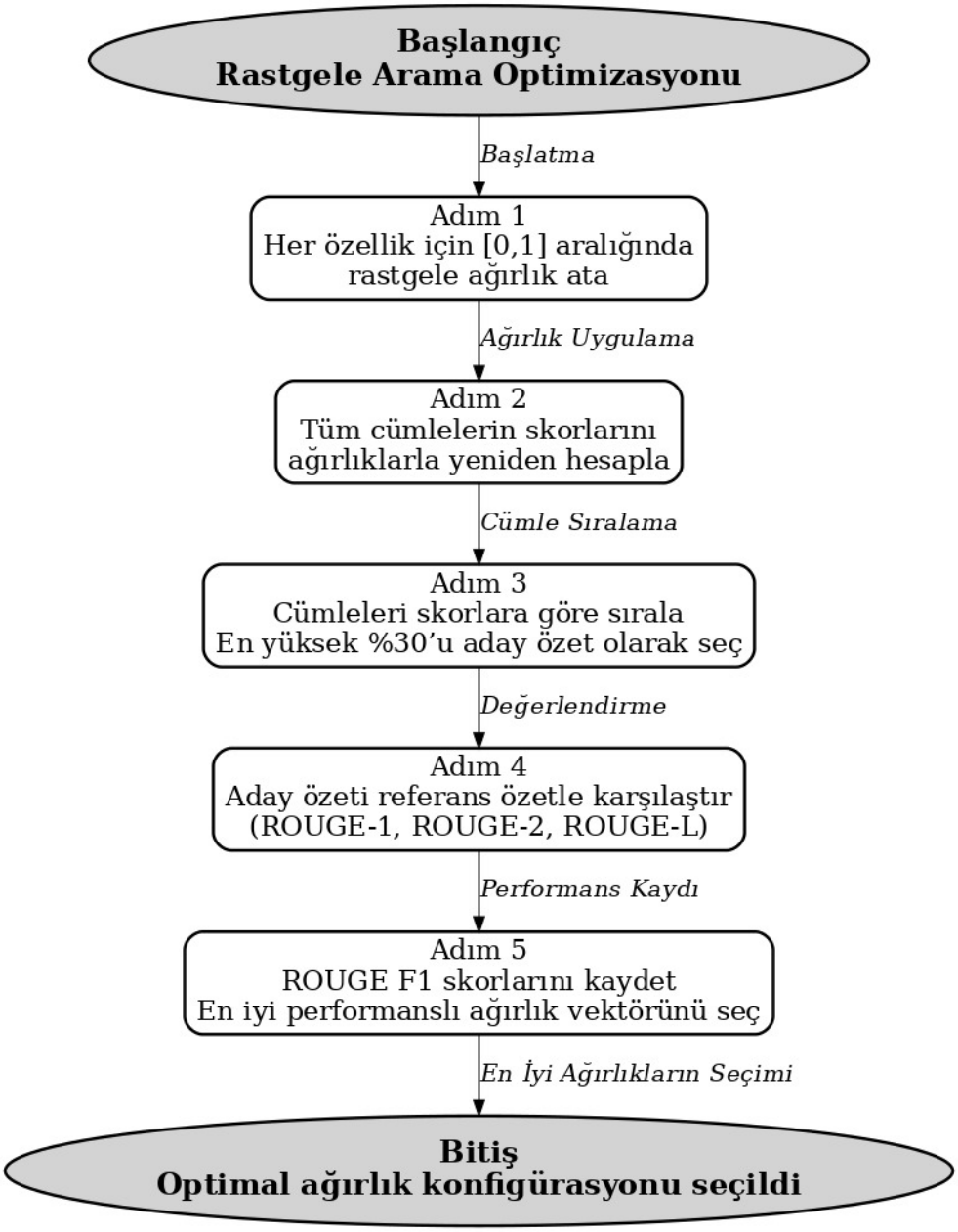
n: toplam özellik sayısıdır

Cümle skoru, özniteliklerin katkısına göre birleştirilmiş toplam bir değeri temsil eder. Başlangıçta tüm ağırlıklar eşit kabul edilerek temel bir sıralama yapılır. Bu ilk sıralama, sistemin hangi cümleleri öne çıkaracağını değerlendirmesi açısından başlangıç noktası olarak kullanılır. Ancak özniteliklerin etkisi sabit değildir; bu nedenle, ağırlıkların sistematik biçimde optimize edilmesi gerekmektedir. Bu amaçla bir sonraki aşamada rastgele arama algoritması uygulanır.

4.5. Ağırlık Optimizasyonu Yaklaşımı: Rastgele Arama

Cümlelerin skorlanmasında kullanılan öznitelik ağırlıkları, sistemin genel özetleme başarımını doğrudan etkileyen kritik bir faktördür. Sabit ya da sezgisel olarak belirlenen ağırlıklar, her belgeye veya bağlama uygun olmayabilir. Bu nedenle, bu çalışmada, ağırlıkların otomatik olarak optimize edilmesini sağlayan rastgeleleştirilmiş bir algoritma geliştirilmiştir.

Bu sürecin işleyiş adımları Şekil 4.2 de akış şeması şeklinde sunulmuştur.



Şekil 4.2. Rastgele Arama ile Ağırlık Optimizasyonu

Bu akış şeması, rastgele arama yaklaşımının temel adımlarını özetlemektedir. Her özellik için farklı ağırlıkların atanmasıyla başlayan süreçte, cümleler bu ağırlıklara göre yeniden puanlanmakta ve en yüksek skora sahip aday özetler seçilmektedir. Daha sonra, bu özetler insan yapımı referans özetlerle karşılaştırılarak ROUGE metrikleri üzerinden değerlendirilmekte ve en iyi performans sağlayan ağırlık vektörü belirlenmektedir. Böylece öznitelik ağırlıkları sezgisel olarak değil, deneysel verilere dayalı bir biçimde optimize edilmektedir.

4.6. Skor Hesaplama ve Sıralama

Cümlelerin özetleme açısından önem derecesini belirleyebilmek amacıyla, her bir cümle için bir toplam puan hesaplanmaktadır. Bu puan, cümleye ait özellik vektörünün belirli ağırlıklarla çarpılarak toplanmasıyla elde edilir. Böylece her cümle, sistem tarafından sayısal olarak değerlendirilmiş olur ve bu skorlar kullanılarak cümleler sıralanır.

Her bir cümlenin skorunun hesaplanmasında kullanılan temel denklem şu şekildedir: Bu yöntem, her cümleyi bir özellik vektörü olarak görür ve vektör ile ağırlıkların noktasal çarpımı (dot product) sonucunda elde edilen değeri, cümlenin özetleme açısından önem puanı olarak değerlendirir.

4.7. Ağırlık Optimizasyonu

Cümle skorlarının hesaplanmasında kullanılan öznitelik ağırlıkları (w_k) özetleme sisteminin genel başarımını doğrudan etkileyen temel faktörlerden biridir. Bu nedenle, sabit veya sezgisel şekilde belirlenmiş ağırlıklar yerine, en iyi sonucu veren ağırlık kombinasyonlarını otomatik olarak keşfetmek büyük önem taşımaktadır. Bu çalışmada, ağırlıkların en uygun şekilde belirlenebilmesi için rastgeleleştirilmiş bir optimizasyon algoritması uygulanmıştır. Bu yöntemde, sistem tarafından sekiz özniteliğe karşılık gelen ağırlıklar [0, 1] aralığında rastgele başlatılmış ve her bir ağırlık kombinasyonu ile cümle skoru hesaplanarak özetleme işlemi gerçekleştirilmiştir. Oluşturulan her özet, insan yapımı referans özetle karşılaştırılmış ve ROUGE-1, ROUGE-2 ve ROUGE-L metriklerine göre performansı değerlendirilmiştir. Her deneme sonucunda elde edilen ROUGE skorları kaydedilmiş ve en yüksek ROUGE-1 skorunu üreten ağırlık vektörü sistem için ideal konfigürasyon olarak belirlenmiştir. Bu işlem çok sayıda tekrar ile yürütülmüş ve sonuç olarak özniteliklerin etkisi yalnızca sezgiye değil, veri temelli deneysel sonuçlara dayanarak optimize edilmiştir.

4.8. Cümle Seçiminde Genetik Algoritmaların Kullanım Süreci

Rastgele arama yöntemiyle en uygun ağırlık vektörünün belirlenmesinin ardından, özetleme sisteminin bir sonraki aşamasında Genetik Algoritma (GA) uygulanmaktadır. Biyolojik evrim mekanizmalarından ilham alan bu optimizasyon

yaklaşımı, en bilgilendirici ve anlamlı cümle kombinasyonunu seçerek yüksek kaliteli özetler üretmeyi amaçlamaktadır.

- **Gösterim:** Sistemde her bir özet adayı, orijinal metinden seçilen cümlelerin indekslerinden oluşan bir kromozom ile temsil edilir. Bu yapı, özetin potansiyel bir çözüm olarak modellenmesini sağlar.
- **Başlangıç popülasyonu:** Algoritma, rastgele oluşturulmuş bir başlangıç popülasyonu ile başlatılır. Her birey, metinden seçilmiş sabit sayıda cümleden oluşur. Popülasyon büyüklüğü ve kromozom uzunluğu, sistemin parametrelerine bağlı olarak tanımlanır.
- **Uygunluk fonksiyonu:** Her kromozomun kalitesi, iki temel bileşen dikkate alınarak hesaplanır: (i) rastgele arama aşamasında elde edilen ağırlıklarla hesaplanan toplam cümle skoru ve (ii) özet içindeki cümleler arasında AraBERT dil modeli kullanılarak ölçülen anlamsal tutarlılık. Bu sayede seçilen özet hem bilgi açısından yoğun hem de içerik açısından bütünlüklü olur.
- **Evrimsel işlemler:** GA, çeşitli evrimsel adımlar içerir. Seçim aşamasında, turnuva seçimi yöntemiyle en yüksek uygunluk puanına sahip kromozomlar seçilir. Çaprazlama adımında, iki ebeveyn kromozomdan alınan cümleler birleştirilerek yeni bireyler oluşturulur. Mutasyon adımı ise, kromozomun bir veya daha fazla cümlesinin rastgele farklı cümlelerle değiştirilmesi yoluyla popülasyon çeşitliliği sağlar.
- **Evrimsel döngü:** Bu işlemler birden fazla nesil boyunca tekrar edilir. Her döngüde en iyi bireyler korunur ve yeni bireyler üretilerek bir sonraki nesil oluşturulur. Sürecin sonunda, en yüksek uygunluk puanına sahip kromozom, sistem tarafından nihai özet olarak seçilir.

Bu süreç sayesinde, genetik algoritma yalnızca en bilgilendirici cümleleri seçmekle kalmaz, aynı zamanda özetin bütünlüğünü ve anlamsal tutarlılığını da korur. Böylece, Arapça metinler için daha güvenilir ve yüksek kaliteli özetler elde edilmektedir.

5. BULGULAR VE DEĞERLENDİRME

Bu bölümde, geliştirilen Arapça metin özetleme sisteminin deneysel olarak değerlendirilmesi sunulmaktadır. İlk olarak, sistemin test edildiği veri seti tanıtılmış; ardından uygulamada kullanılan yazılım ortamı ve teknik araçlar açıklanmıştır. Devamında, optimize edilen öznitelik ağırlıklarına dayalı olarak yürütülen deneysel senaryolar ele alınmış ve sistemin farklı öznitelik yapılarına karşı gösterdiği performans ROUGE metrikleri üzerinden karşılaştırmalı olarak analiz edilmiştir. Amaç, önerilen yöntemin etkinliğini nesnel ölçütlerle ortaya koymaktır.

5.1. Veri Seti

Bu çalışmada kullanılan veri seti, Arapça metin özetleme yöntemlerinin değerlendirilmesi amacıyla Essex Arabic Summaries Corpus (EASC) olmuştur. EASC, 153 Arapça makaleden oluşmakta ve her bir makale için beş farklı insan tarafından hazırlanmış toplam 765 özet içermektedir. Makaleler Wikipedia, Alwatan gazetesi ve Alrai gazetesi gibi çeşitli kaynaklardan derlenmiştir. Veri kümesi; din, eğitim, bilim ve teknoloji, çevre, finans, turizm, sağlık, politika, spor, sanat ve müzik gibi farklı konuları kapsamaktadır. Bu çeşitlilik, özetleme yöntemlerinin farklı içerik türleri üzerindeki başarımını değerlendirmek için güçlü bir zemin sunmaktadır.

EASC veri kümesi, araştırma topluluğu tarafından yaygın olarak kullanılan ve Arapça metin özetleme çalışmalarında referans olarak kabul edilen bir kaynaktır. Veri kümesinin temel istatistiksel özentikleri aşağıdaki tabloda özetlenmiştir:

Tablo 5.1. EASC veri kümesinin genel özellikleri.

Veri Kümesinin	Belge Sayısı	Özet Sayısı	Konu Sayısı	Ortalama Cümle Sayısı
EASC	153	765	10	17

Veri kümesine ait örnek tablo aşağıda sunulmuştur. Bu tablo, özgün bir metin ile ona karşılık gelen insan özetlerini göstermektedir.

Tablo 5.2. EASC veri kümesine ait özgün metin ve insan özetleri.

İçerik Türü	Metin
Orijinal Belge	طرح مدير مدرسة حبيب بن زيد الأنصاري الابتدائية برنامجا هو الأول من نوعه على مستوى منطقة المدينة المنورة التعليمية, إن لم يكن الأول من نوعه على مستوى السعودية على حسب تصريحه, تمثل في إقامة البرنامج التأهيلي للطلاب المستجدين في المرحلة الابتدائية هذه الأيام ولمدة ثلاثة أسابيع, بدلا من إقامته في بداية العام الدراسي. وقال مدير المدرسة أحمد الجمعان عن التجربة التي لا تزال في بدايتها " لا يمكن خوض غمار هذه التجربة في ظل عدم مساندة وتعاون أولياء أمور التلاميذ, فـ90% من التلاميذ الجدد المسجلين انخرطوا في هذا البرنامج الذي نسعى من خلاله إلى كسر حاجز الخوف, وترغيب الطالب في المدرسة, كما نهدف إلى استثمار الوقت في الإجازة من خلال تفعيل دور الأسرة, إذ نوفر لكل طالب قرصا مدمجا (سي دي) لمنهج القراءة والكتابة, ومذكرة تحتوي على أنشطة تعليمية". وحضور التلميذ إلى المدرسة للمشاركة في هذا البرنامج غير مقيد بمدة زمنية معينة, ومن دون حقائب مدرسية, كما يترك فيه للطالب حرية التعرف على مرافق المدرسة ومعلمي الصفوف الأولية, كما يتضمن رحلة أسبوعية إلى مدينة ألعاب، وإجراء مسابقات يومية ترفيهية. يذكر أن عدد الطلاب المسجلين في البرنامج 75 طالبا.
İnsan Özeti 1	طرح مدير مدرسة حبيب بن زيد الانصاري الابتدائية برنامجا هو الاول من نوعه على مستوى منطقة المدينة المنورة التعليمية, ان لم يكن الاول من نوعه على مستوى السعودية على حسب تصريحه, تمثل في اقامة البرنامج التاهيلي للطلاب المستجدين في المرحلة الابتدائية هذه الايام .
İnsan Özeti 2	طرح مدير مدرسة حبيب بن زيد الانصاري الابتدائية برنامجا هو الاول من نوعه على مستوى منطقة المدينة المنورة التعليمية, ان لم يكن الاول من نوعه على مستوى السعودية على حسب تصريحه, تمثل في اقامة البرنامج التاهيلي للطلاب المستجدين في المرحلة الابتدائية هذه الايام .
İnsan Özeti 3	طرح مدير مدرسة حبيب بن زيد الانصاري الابتدائية برنامجا هو الاول من نوعه على مستوى منطقة المدينة المنورة التعليمية, ان لم يكن الاول من نوعه على مستوى السعودية على حسب تصريحه, تمثل في اقامة البرنامج التاهيلي للطلاب المستجدين في المرحلة الابتدائية هذه الايام .
İnsan Özeti 4	وقال مدير المدرسة احمد الجمعان عن التجربة التي لا تزال في بدايتها " لا يمكن خوض غمار هذه التجربة في ظل عدم مساندة وتعاون اولياء امور التلاميذ, فـ90% من التلاميذ الجدد المسجلين انخرطوا في هذا البرنامج الذي نسعى من خلاله الى كسرحاجز الخوف, وترغيب الطالب في المدرسة.
İnsan Özeti 5	وقال مدير المدرسة احمد الجمعان عن التجربة التي لا تزال في بدايتها " لا يمكن خوض غمار هذه التجربة في ظل عدم مساندة وتعاون اولياء امور التلاميذ, فـ90% من التلاميذ الجدد المسجلين انخرطوا في هذا البرنامج الذي نسعى من خلاله الى كسرحاجز الخوف, وترغيب الطالب في المدرسة.

5.2. Gerçekleme Ortamı

Bu çalışmada Arapça metin özetleme sistemi Python 3.10 programlama dili kullanılarak Anaconda dağıtımı üzerinden Jupyter Notebook ortamında geliştirilmiştir. Metin işleme ve analiz görevlerinde NumPy (v1.24.0), NLTK (v3.8) ve görselleştirme için Matplotlib (v3.7) kütüphanelerinden yararlanılmıştır.

Deneyler, macOS işletim sistemi yüklü bir MacBook Air üzerinde (Apple M1 işlemci, 8 GB RAM, 256 GB SSD) yürütülmüştür. Bu donanım, özetleme deneylerinin verimli bir şekilde gerçekleştirilmesine olanak sağlamıştır.

5.3. Deneysel Senaryolar ve Uygulama Sonuçları

Bu bölümde, geliştirilen çok özellikli Arapça metin özetleme sisteminin, farklı öznitelik grupları altında gösterdiği performans detaylı bir şekilde analiz edilmiştir. Yapılan deneyler, özetleme sisteminde kullanılan her bir özniteliğin özetleme kalitesi üzerindeki etkisini ortaya koymayı amaçlamaktadır. Özniteliklerin göreli önemini temsil eden ağırlıklar, Random Search algoritması kullanılarak optimize edilmiş ve bu süreç sonucunda elde edilen en iyi ağırlık kombinasyonu aşağıda sunulmuştur.

Tablo 5.3. Ağırlık değerleri.

F_sp	F_nd	F_ls	F_kw	F_ss	F_ne	F_eh	F_ti
0.9706	0.0491	0.9750	0.7005	0.5757	0.6298	0.5124	0.0658

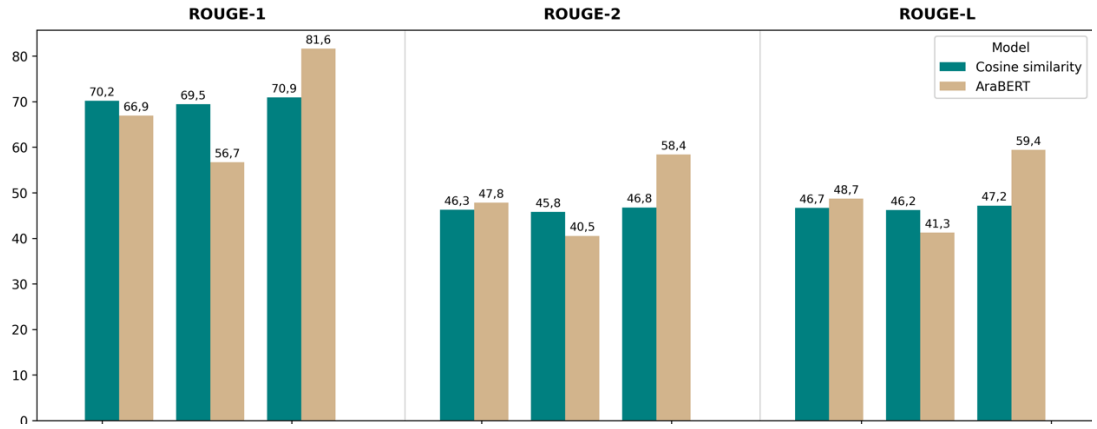
Bu ağırlık değerleri, her bir özniteliğin özet kalitesi üzerindeki etkisini yansıtmaktadır ve sistem tarafından en yüksek ROUGE puanlarını üreten konfigürasyon olarak belirlenmiştir. Ağırlıkların bu şekilde optimize edilmesi, özniteliklerin sadece sezgisel olarak değil, deneysel olarak da önem düzeylerinin ortaya konmasını sağlamıştır.

Devam eden alt bölümlerde, bu optimize edilmiş ağırlıklarla iki farklı deneysel senaryo değerlendirilmiştir:

5.3.1. Senaryo A: benzerlik yöntemlerinin genetik algoritmadaki etkisi

Bu çalışmada, genetik algoritmanın fitness fonksiyonu içerisinde cümleler arası benzerliği ölçmek amacıyla iki farklı yöntem değerlendirilmiştir: Cosine Similarity ve AraBERT. Benzerlik ölçümü, özet adaylarının tutarlılık (coherence) düzeyini belirlemede kritik bir rol oynamaktadır.

Şekil 5.1'te görüldüğü üzere, deneysel sonuçlar yöntemler arasında anlamlı farklılıklar olduğunu ortaya koymuştur. Cosine Similarity, özellikle ROUGE-1 metriğinde yaklaşık %70,2 oranında en yüksek değere ulaşarak kelime düzeyinde güçlü bir performans sergilemiştir. Buna karşın, AraBERT bağlamsal ilişkileri daha iyi yakalayarak ROUGE-2 ve ROUGE-L metriklerinde daha yüksek sonuçlar elde etmiştir. Bu bulgular, kelime tabanlı ve bağlamsal yaklaşımlar arasındaki tamamlayıcı farkları göstermektedir.



Şekil 5.1. Benzerlik yöntem karşılaştırması.

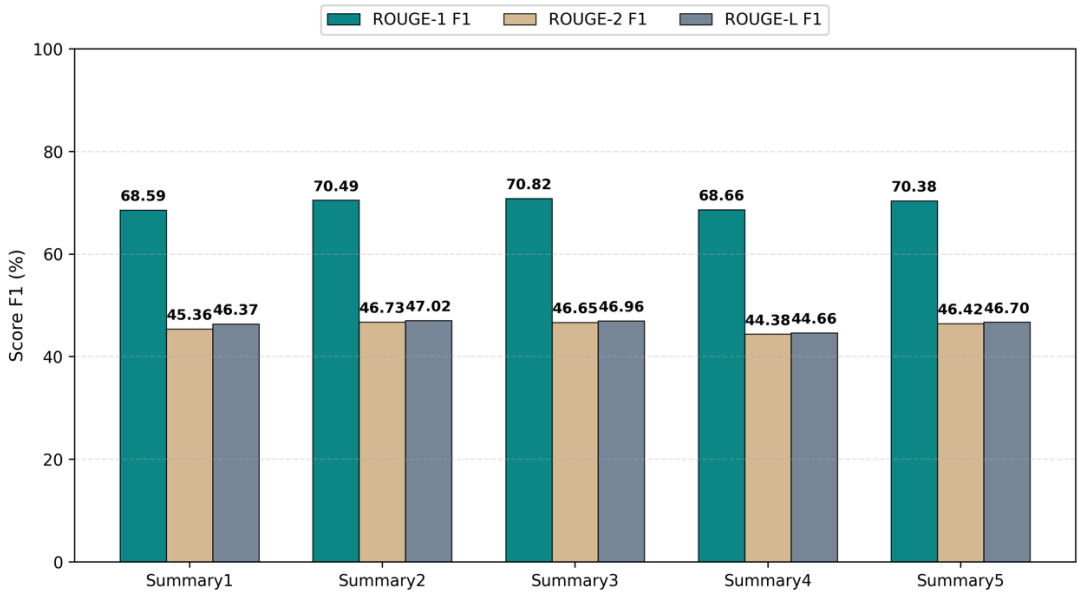
Elde edilen sonuçlara dayanarak, önerilen sistemde cümleler arası tutarlılık hesaplamasında Cosine Similarity temel yöntem olarak tercih edilmiştir. Böylece, hem doğruluk (accuracy) hem de hesaplama verimliliği dikkate alınarak özetlerin daha dengeli bir şekilde üretilmesi sağlanmıştır.

5.3.2. Senaryo B: hibrit modelin performansının ölçülmesi

Bu bulgular doğrultusunda, çalışmanın sonraki aşamalarında hibrit özellik seti tercih edilmiştir. Çalışmada kullanılan veri kümesi, her metin için beş farklı insan özeti (summary1 - summary5) içermektedir. Bu öznel referanslar üzerinden modelin çıktıları değerlendirilmiş ve ROUGE skorları hesaplanmıştır.

Tablo 5.4. Referans özetlere göre ROUGE skorlarının karşılaştırılması.

Summary	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1
Summary1	68.59	45.36	46.37
Summary2	70.49	46.73	47.02
Summary3	70.82	46.65	46.96
Summary4	68.66	44.38	44.66
Summary5	70.38	42.42	46.70



Şekil 5.2. Referans özetlere göre rouge skorlarının karşılaştırılması.

Tablo 5.5'te görüldüğü üzere, farklı insan özetleri arasında Summary2, ROUGE-1 (%70.49) ve ROUGE-2 (%46.73) metriklerinde yüksek değerler elde etmiştir. Bu nedenle, Summary2 modelin başarısını en iyi yansıtan referans özet olarak

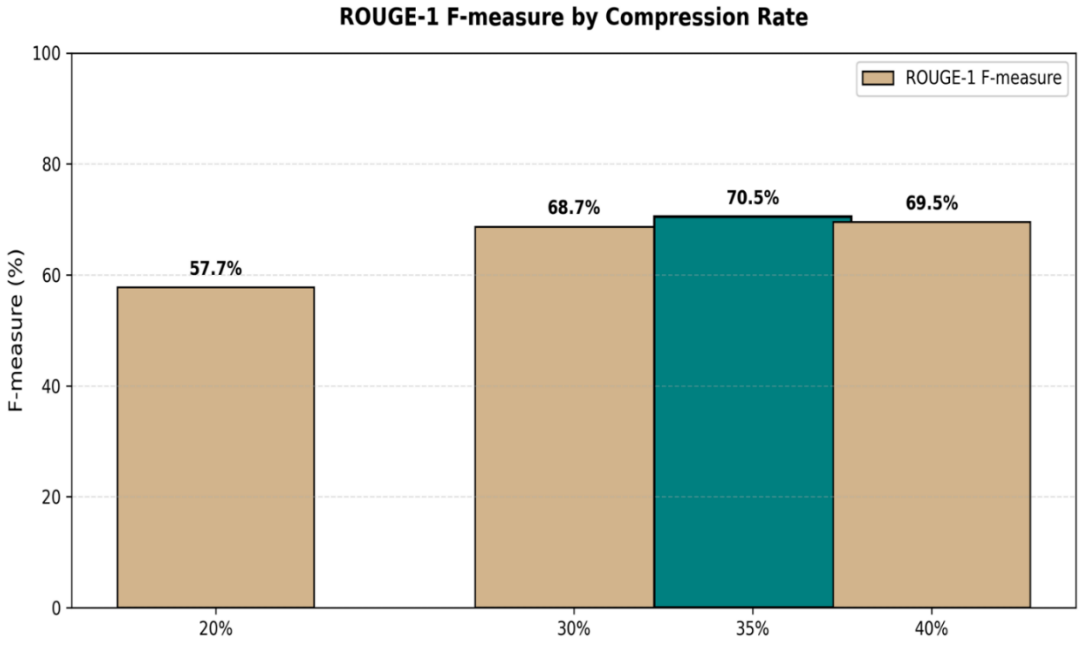
değerlendirilmiş ve sonraki analizlerde temel karşılaştırma ölçütü olarak tercih edilmiştir. Ayrıca, diğer özetlerde de benzer seviyelerde sonuçlar elde edilmesi, modelin genel olarak dengeli ve tutarlı bir performans sergilediğini göstermektedir.

5.3.2.1. Senaryo B-1 Sıkıştırma oranının model performansına etkisi

Bu bölümde geliştirilen modelin farklı sıkıştırma oranları altındaki performansı analiz edilmiştir. Tablo 5.7 ve Şekil 5.3'te sunulan sonuçlara göre, özetleme çalışmalarında yaygın olarak kabul edilen %20–%40 aralığı test edilmiştir.

Tablo 5.5. Sıkıştırma oranının model performansına etkisi.

Compression Rate	ROUGE-1	ROUGE-1	ROUGE-1	ROUGE-2	ROUGE-2	ROUGE-2	ROUGE-L	ROUGE-L	ROUGE-L
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
20%	57.695	76.598	46.275	36.126	47.964	28.975	37.163	49.339	29.807
30%	68.707	63.255	77.129	47.395	43.132	52.593	48.115	43.787	53.392
35%	70.528	69.308	71.999	47.664	46.773	48.590	47.883	46.988	48.813
40%	69.514	64.579	75.266	46.630	43.320	50.489	46.878	43.550	50.757



Şekil 5.3. Sıkıştırma oranının model performansına etkisi.

Sonuçlar, düşük sıkıştırma oranlarında (ör. %20) performansın sınırlı kaldığını, oran arttıkça ROUGE skorlarının belirgin şekilde yükseldiğini göstermektedir.

Özellikle %35 sıkıştırma oranı, en yüksek başarıya ulaşarak modelin en dengeli noktası olmuştur. %40 seviyesinde ise küçük bir düşüş gözlenmiş, ancak performans yine de yüksek seviyede kalmıştır.

Genel olarak, modelin en iyi özetleme performansı %35 civarında elde edilmiş ve bu oran, bilgilendiricilik ile özet uzunluğu arasında dengeli bir nokta olarak değerlendirilmiştir.

5.3.2.2. Senaryo B-2: Genetik algoritma parametrelerinin etkisi

Bu bölümde, modelin farklı genetik algoritma parametreleri altındaki performansı değerlendirilmiştir. Bu deneylerin temel amacı, modelin parametre hassasiyetini ortaya koymak ve en etkili yapılandırmayı belirlemektir.

Bu kapsamda, çeşitli genetik algoritma parametre kombinasyonları denenmiş ve en iyi sonuçları sağlayan yapılandırma aşağıda sunulmuştur:

Tablo 5.6. GA parametreleri.

POP_SIZE	GENERATIONS	MUTATION_RATE	TOURNAMENT_K
250	70	0.2	30

Bu konfigürasyon ile ROUGE-1 skoru 70.63% olarak elde edilmiştir. Detaylı sonuçlar Tablo 5.9'te sunulmuştur.

Tablo 5.7. GA parametre deney sonuçları.

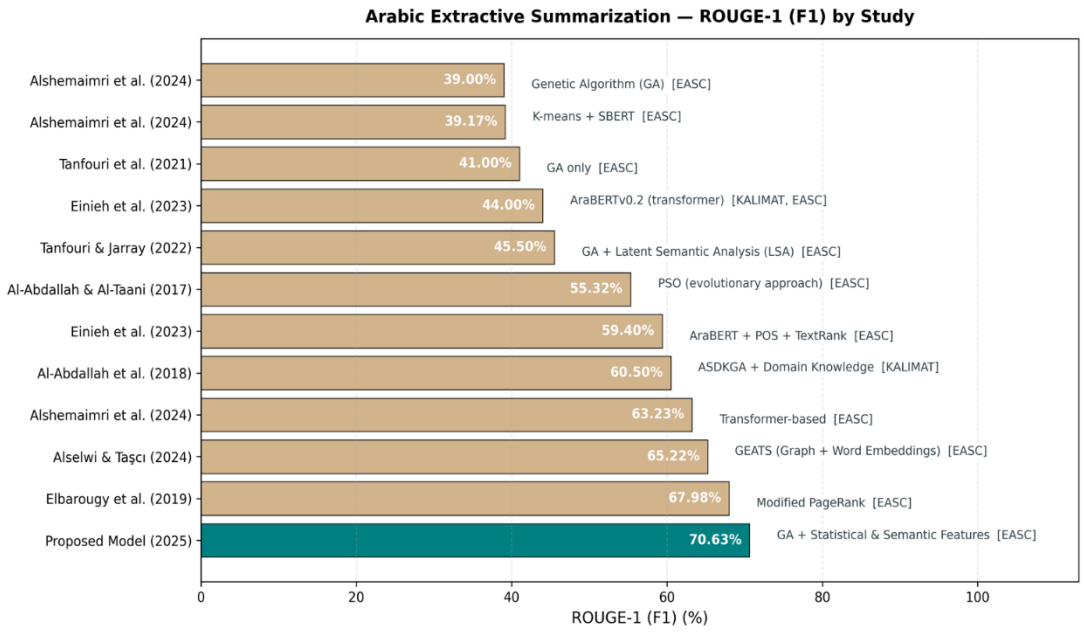
POP_SIZE	GENERATIONS	MUTATION_RATE	TOURNAMENT_K	ROUGE-1	ROUGE-2	ROUGE-L
50	10	0.1	8	69.91	45.80	46.03
100	30	0.1	15	70.34	46.51	46.94
120	30	0.1	15	70.59	47.35	47.75
200	30	0.1	15	69.56	45.22	45.42
200	70	0.1	30	70.00	45.88	46.26
250	**70**	**0.2**	**30**	**70.63**	**47.66**	**47.88**

Bu analizler, sistemin parametre ayarlarının çıktılar üzerinde doğrudan etkili olduğunu göstermektedir.

5.3.3. Senaryo C: Deneyler ve Karşılaştırmalı Değerlendirme

Önerilen modelin etkinliğini değerlendirmek amacıyla, ROUGE-1 F1 sonuçları literatürde raporlanan çeşitli Arapça metin özetleme yöntemleriyle karşılaştırılmıştır (Şekil 5.4). Grafik incelendiğinde, yalnızca Genetik Algoritma (Alshemaimri ve ark., 2024) [84]veya AraBERT (Einieh ve ark., 2023) tabanlı yaklaşımların daha düşük performans gösterdiği görülmektedir. Hibrit yaklaşımlar (ör. GA + LSA, Tanfouri ve Jarray, 2022)[58], [85] ve evrimsel algoritmalar (Al-Abdallah & Al-Taani, 2017) [86] bir miktar iyileşme sağlamış olsa da, sonuçlar %55–60 aralığında kalmıştır. Son yıllarda önerilen yöntemlerden transformer tabanlı yaklaşımlar (Alshemaimri ve ark., 2024) [84]%63.23'e ulaşırken, graf ve kelime gömme tabanlı yaklaşımlar (Alselwi & Taşçı, 2024)[44] %65.22 gibi daha yüksek sonuçlar üretmiştir. Modified PageRank (Elbarougy ve ark., 2019) [65]ise %67.98 değeri ile öne çıkmıştır.

Buna karşılık, bu tezde önerilen model %70.63 ROUGE-1 F1 değeri ile literatürdeki en yüksek performansa ulaşmıştır. Model, istatistiksel ve anlamsal özelliklerin genetik algoritma ile optimize edilmesi sayesinde, hem geleneksel meta-sezgisel yöntemlere hem de grafik/embedding tabanlı yöntemlere karşı rekabetçi bir üstünlük sergilemektedir. Ayrıca, yüksek hesaplama gücü veya büyük ölçekli eğitim verisi gerektirmeden düşük kaynaklı ortamlarda da uygulanabilir olması, önerilen yöntemin Arapça özetleme çalışmalarında güçlü bir alternatif sunduğunu göstermektedir.



Şekil 5.4. Karşılaştırmalı değerlendirme.

Bu çalışmada kullanılan stemming yöntemi, Arapça'nın karmaşık morfolojik yapısı nedeniyle her zaman doğru sonuçlar üretememiştir. Bu durum bazı kelimelerin yanlış köklenmesine yol açmış ve cümlelerin temsilinde kısmi hatalara neden olmuştur. Ayrıca deneyler yalnızca EASC veri kümesi üzerinde yürütülmüş ve değerlendirme ölçütü olarak ROUGE metrikleri kullanılmıştır. Bu sınırlılıklar, modelin genellenebilirliğini kısmen etkilese de gelecekte daha gelişmiş morfolojik analiz teknikleri ve farklı veri kümeleri ile yapılacak testler yöntemin kapsamını genişletebilir.

5.4. Değerlendirme

Yapılan tüm deneysel çalışmalar birlikte değerlendirildiğinde, geliştirilen hibrit modelin Arapça metin özetleme alanında yüksek bir performans sergilediği görülmektedir. Deneyler, cümleler arası benzerlik ölçümünün özetlerin tutarlılığını doğrudan etkilediğini ve farklı yöntemlerin sonuçlar üzerinde belirgin farklılıklar yarattığını ortaya koymuştur. Özellikle kelime düzeyinde benzerlik yaklaşımları hesaplama açısından daha verimli olurken, bağlamsal yöntemler içerik bütünlüğünü güçlendirmiştir.

Öte yandan, hibrit özellik setinin kullanılması ve genetik algoritma parametrelerinin optimize edilmesi sayesinde, model farklı sıkıştırma oranları altında dengeli ve tutarlı özetler üretmiştir. Bulgular,%35'lik sıkıştırma oranında bilgilendiricilik ile özet uzunluğu arasında en uygun dengenin sağlandığını, ayrıca parametrelerin doğru ayarlanmasıyla sistemin başarısının belirgin biçimde arttığını göstermektedir.

Sonuçlar, önerilen yaklaşımın yalnızca tekil senaryolarda değil, genel olarak tüm deneysel koşullar altında kararlı bir şekilde yüksek ROUGE skorları ürettiğini kanıtlamaktadır. Literatürde raporlanan yöntemlerle yapılan karşılaştırmalarda, geliştirilen modelin daha yüksek doğruluk sağladığı ve özellikle ROUGE-1 F1 metriğinde %70.63 değeri ile en iyi performansa ulaştığı görülmüştür. Bu durum, istatistiksel ve anlamsal özniteliklerin genetik algoritmalar aracılığıyla etkin bir biçimde bütünleştirilmesinin, Arapça özetleme çalışmalarında güçlü bir alternatif sunduğunu göstermektedir.

6. SONUÇ VE ÖNERİLER

Metin özetleme, doğal dil işleme alanında bilgi fazlalığına çözüm sunan temel araçlardan biridir. Bu yöntem, uzun metinlerdeki önemli içeriklere kısa sürede ve minimum çabayla erişim sağlaması açısından büyük önem taşımaktadır. Özellikle yapısal olarak zengin ve anlamsal açıdan karmaşık bir yapıya sahip olan Arapça gibi dillerde, etkili özetleme sistemlerinin geliştirilmesi daha da kritik hale gelmektedir.

Bu tez çalışmasında, extractive tabanlı bir Arapça metin özetleme yöntemi geliştirilmiştir. Önerilen modelde, cümlelerin yapısal ve anlamsal yönlerini temsil eden sekiz farklı öznitelik (cümle konumu, cümle uzunluğu, anahtar kelime yoğunluğu, anlamsal benzerlik, özel varlıklar, İngilizce kelime oranı, TF-ISF kelime ağırlığı ve başlık uyumu) kullanılmıştır. Bu özniteliklerin ağırlıkları Random Search algoritmasıyla optimize edilmiş, ardından Genetik Algoritma (GA)uygulanarak en bilgilendirici cümle kombinasyonları seçilmiştir. Ayrıca, cümleler arası tutarlılığın sağlanması için Cosine Similarity yöntemi esas alınmıştır.

Modelin başarısı, Arapça metin özetleme çalışmalarında yaygın olarak kullanılan EASC (Essex Arabic Summaries Corpus) veri kümesi üzerinde değerlendirilmiştir. Sistem, her metin için mevcut beş insan özeti üzerinden test edilmiş ve sonuçlar ROUGE-1, ROUGE-2 ve ROUGE-L metrikleriyle analiz edilmiştir. Deneysel analizler sonucunda, en yüksek performansın ROUGE-1 F skoru %70.63 ile elde edildiği görülmüştür.

Bu sonuçlar, önerilen extractive GA tabanlı yaklaşımın, bilgilendiricilik ile özet uzunluğu arasında etkili bir denge sağladığını ve Arapça metin özetleme alanında güçlü bir çözüm sunduğunu göstermektedir.

Bu çalışmada elde edilen sonuçlar, Arapça metin özetleme alanında önemli katkılar sunmakla birlikte, ilerleyen çalışmalarda geliştirilmesi gereken bazı yönler bulunmaktadır. Öncelikle, ön işleme ve normalizasyon adımlarının daha da iyileştirilmesi gerekmektedir. Özellikle hareke, yazım farklılıkları ve tokenizasyon zorlukları gibi Arapçaya özgü dilsel özelliklerle daha etkin bir şekilde başa çıkmak, özetlerin doğruluğunu artıracaktır.

Bunun yanı sıra, sadece ROUGE metrikleri ile sınırlı kalmayıp, BLEU, METEOR veya BERTScore gibi ek değerlendirme metriklerinin kullanılması, sistemin performansının daha kapsamlı bir şekilde analiz edilmesini sağlayacaktır.

Ayrıca, mevcut çalışmada kullanılan kök bulma (stemming) tekniklerinin geliştirilmesi ve daha doğru yöntemlerin uygulanması, cümle temsillerinin iyileştirilmesine katkı sağlayacaktır. Benzer şekilde, fitness fonksiyonuna sözdizimsel yapılar veya cümlenin konuya ilgisi gibi ek özelliklerin entegre edilmesi, özetlerin kalite ve bütünlüğünü artırabilir.

Son olarak, ileride hibrit sistemlerin geliştirilmesi hedeflenmektedir. Bu kapsamda, genetik algoritmanın diğer meta-sezgisel yöntemlerle veya grafik tabanlı yaklaşımlarla birleştirilmesi, özetleme performansını daha da yükseltecek ve literatürdeki mevcut yöntemlere karşı daha güçlü bir alternatif sunacaktır.

KAYNAKLAR

[1] D. R. Radev, E. Hovy, ve K. McKeown, "Introduction to the Special Issue on Summarization", *Computational Linguistics*, c. 28, sy 4, ss. 399-408, 2002, doi: 10.1162/089120102762671927.

[2] A. Abu-Errub, A. Odeh, Q. Shambour, ve H. Al-Haj, "Arabic roots extraction using morphological analysis", *IJCSI International Journal of Computer Science Issues*, c. 11, ss. 128-134, Oca. 2014.

[3] B. Talafha, A. Fadel, M. Al-Ayyoub, Y. Jararweh, M. AL-Smadi, ve P. Juola, *Team JUST at the MADAR Shared Task on Arabic Fine-Grained Dialect Identification*. 2019, s. 289. doi: 10.18653/v1/W19-4638.

[4] L. M. Al Qassem, D. Wang, Z. Al Mahmoud, H. Barada, A. Al-Rubaie, ve N. I. Almoosa, "Automatic Arabic Summarization: A survey of methodologies and systems", *Procedia Computer Science*, c. 117, ss. 10-18, Oca. 2017, doi: 10.1016/j.procs.2017.10.088.

[5] C. Sabty, F. Wasfalla, M. Islam, ve S. Abdennadher, "Data Augmentation Techniques on Arabic Data for Named Entity Recognition", *Procedia Computer Science*, c. 189, ss. 292-299, Oca. 2021, doi: 10.1016/j.procs.2021.05.092.

[6] T. Tran, T.-T. Do, I. Reid, ve G. Carneiro, "Bayesian Generative Active Deep Learning", içinde *International Conference on Machine Learning*, PMLR, May. 2019, ss. 6295-6304. Erişim: 18 Mart 2025. [Çevrimiçi]. Erişim adresi: https://proceedings.mlr.press/v97/tran19a.html

[7] "infoextr.pdf". Erişim: 17 Mart 2025. [Çevrimiçi]. Erişim adresi: https://www.cs.cmu.edu/~zechner/infoextr.pdf

[8] C.-Y. Lin ve E. Hovy, "From Single to Multi-document Summarization", içinde *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, ve D. Lin, Ed., Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Tem. 2002, ss. 457-464. doi: 10.3115/1073083.1073160.

[9] "(PDF) Word Sequence Models for Single Text Summarization". Erişim: 17 Mart 2025. [Çevrimiçi]. Erişim adresi: https://www.researchgate.net/publication/224384763_Word_Sequence_Models_for_Single_Text_Summarization

[10] "Narrative text classification for automatic key phrase extraction in web document corpora | Proceedings of the 7th annual ACM international workshop on Web information and data management". Erişim: 17 Mart 2025. [Çevrimiçi]. Erişim adresi: https://dl.acm.org/doi/abs/10.1145/1097047.1097059

[11] "Generic text summarization using local and global properties of sentences | IEEE Conference Publication | IEEE Xplore". Erişim: 17 Mart 2025. [Çevrimiçi]. Erişim adresi: https://ieeexplore.ieee.org/abstract/document/1241194

[12] "A Comprehensive Review of Arabic Text Summarization | IEEE Journals & Magazine | IEEE Xplore". Erişim: 15 Mart 2025. [Çevrimiçi]. Erişim adresi: https://ieeexplore.ieee.org/abstract/document/9745159

[13] Y. A. AL-Khassawneh ve E. S. Hanandeh, "Extractive Arabic Text Summarization-Graph-Based Approach", *Electronics*, c. 12, sy 2, Art. sy 2, Oca. 2023, doi: 10.3390/electronics12020437.

[14] M. Elmenshawy, T. Hamza, ve R. Deeb, "Automatic arabic text summarization (AATS): A survey", *Journal of Intelligent & Fuzzy Systems*, c. 43, ss. 1-16, Haz. 2022, doi: 10.3233/JIFS-213589.

[15] T. Cai, M. Shen, H. Peng, L. Jiang, ve Q. Dai, "Improving Transformer with Sequential Context Representations for Abstractive Text Summarization", içinde *Natural Language Processing and Chinese Computing*, J. Tang, M.-Y. Kan, D. Zhao, S. Li, ve H. Zan, Ed., Cham: Springer International Publishing, 2019, ss. 512-524. doi: 10.1007/978-3-030-32233-5_40.

[16] M. Gambhir ve V. Gupta, "Recent automatic text summarization techniques: a survey", *Artif Intell Rev*, c. 47, sy 1, ss. 1-66, Oca. 2017, doi: 10.1007/s10462-016-9475-9.

[17] "Aspect based multi-document summarization | IEEE Conference Publication | IEEE Xplore". Erişim: 18 Haziran 2025. [Çevrimiçi]. Erişim adresi: https://ieeexplore.ieee.org/document/7813838

[18] "The challenging task of summary evaluation: an overview | Request PDF", *ResearchGate*, doi: 10.1007/s10579-017-9399-2.

[19] H. N. Fejer ve N. Omar, "Automatic Arabic text summarization using clustering and keyphrase extraction", içinde *Proceedings of the 6th International Conference on Information Technology and Multimedia*, Kas. 2014, ss. 293-298. doi: 10.1109/ICIMU.2014.7066647.

[20] D. Miller, "Leveraging BERT for Extractive Text Summarization on Lectures", 07 Haziran 2019, *arXiv*: arXiv:1906.04165. doi: 10.48550/arXiv.1906.04165.

[21] A. B. Soliman, K. Eissa, ve S. R. El-Beltagy, "AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP", *Procedia Computer Science*, c. 117, ss. 256-265, Oca. 2017, doi: 10.1016/j.procs.2017.10.117.

[22] "Exploring Clustering for Multi-document Arabic Summarisation | SpringerLink". Erişim: 15 Mart 2025. [Çevrimiçi]. Erişim adresi: https://link.springer.com/chapter/10.1007/978-3-642-25631-8_50

[23] J. Pennington, R. Socher, ve C. Manning, "GloVe: Global Vectors for Word Representation", içinde *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, ve W. Daelemans, Ed., Doha, Qatar: Association for Computational Linguistics, Eki. 2014, ss. 1532-1543. doi: 10.3115/v1/D14-1162.

[24] M. Afsharizadeh, H. Ebrahimpour-Komleh, ve A. Bagheri, "Query-oriented text summarization using sentence extraction technique", içinde *2018 4th International Conference on Web Research (ICWR)*, Nis. 2018, ss. 128-132. doi: 10.1109/ICWR.2018.8387248.

[25] R. Sennrich, B. Haddow, ve A. Birch, "Neural Machine Translation of Rare Words with Subword Units", 10 Haziran 2016, *arXiv*: arXiv:1508.07909. doi: 10.48550/arXiv.1508.07909.

[26] M. Mohamed ve M. Oussalah, "SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis", *Information Processing & Management*, c. 56, sy 4, ss. 1356-1372, Tem. 2019, doi: 10.1016/j.ipm.2019.04.003.

[27] "Query-oriented text summarization using sentence extraction technique | IEEE Conference Publication | IEEE Xplore". Erişim: 19 Mart 2025. [Çevrimiçi]. Erişim adresi: https://ieeexplore.ieee.org/abstract/document/8387248

[28] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", içinde *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Tem. 2004, ss. 74-81. Erişim: 18 Haziran 2025. [Çevrimiçi]. Erişim adresi: https://aclanthology.org/W04-1013/

[29] K. Papineni, S. Roukos, T. Ward, ve W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation", içinde *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Philadelphia, Pennsylvania: Association for Computational Linguistics, 2001, s. 311. doi: 10.3115/1073083.1073135.

[30] S. Banerjee ve A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", içinde *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, ve C. Voss, Ed., Ann Arbor, Michigan: Association for Computational Linguistics, Haz. 2005, ss. 65-72. Erişim: 18 Haziran 2025. [Çevrimiçi]. Erişim adresi: https://aclanthology.org/W05-0909/

[31] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, ve Y. Artzi, "BERTScore: Evaluating Text Generation with BERT", 24 Şubat 2020, *arXiv*: arXiv:1904.09675. doi: 10.48550/arXiv.1904.09675.

[32] F. S. Douzidia ve G. Lapalme, "Lakhas, an Arabic summarization system".

[33] K. Omar ve M. Al-Shaar, "Method for Arabic text Summarization using statistical features and word2vector approach", içinde *Proceedings of the 2023 9th International Conference on Computer Technology Applications*, içinde ICCTA '23. New York, NY, USA: Association for Computing Machinery, Ağu. 2023, ss. 258-262. doi: 10.1145/3605423.3605444.

[34] N. Alami, Y. El Adlouni, N. En-nahnahi, ve M. Meknassi, "Using Statistical and Semantic Analysis for Arabic Text Summarization", içinde *International Conference on Information Technology and Communication Systems*, G. Noreddine ve J. Kacprzyk, Ed., Cham: Springer International Publishing, 2018, ss. 35-50. doi: 10.1007/978-3-319-64719-7_4.

[35] "(PDF) Lexical Cohesion and Entailment based Segmentation for Arabic Text Summarization (LCEAS)", ResearchGate. Erişim: 18 Haziran 2025. [Çevrimiçi]. Erişim adresi: https://www.researchgate.net/publication/274223839_Lexical_Cohesion_and_Entailment_based_Segmentation_for_Arabic_Text_Summarization_LCEAS

[36] H. Froud, A. Lachkar, ve S. A. Ouatik, "Arabic text summarization based on latent semantic analysis to enhance arabic documents clustering", 08 Şubat 2013, *arXiv*: arXiv:1302.1612. doi: 10.48550/arXiv.1302.1612.

[37] S. Lagrini ve M. Redjimi, "A New Approach for Arabic Text Summarization", içinde *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, M. Abbas ve A. A. Freihat, Ed., Trento, Italy: Association for Computational Linguistics, 2021, ss. 176-185. Erişim: 18 Haziran 2025. [Çevrimiçi]. Erişim adresi: https://aclanthology.org/2021.icnlsp-1.20/

[38] "Automatic Arabic Text Summarization Using Analogical Proportions | Cognitive Computation". Erişim: 14 Nisan 2025. [Çevrimiçi]. Erişim adresi: https://link.springer.com/article/10.1007/s12559-020-09748-y

[39] "(PDF) Digital Learning for Summarizing Arabic Documents", *ResearchGate*, doi: 10.1007/978-3-642-14770-8_10.

[40] R. Belkebir ve A. Guessoum, "A Supervised Approach to Arabic Text Summarization Using AdaBoost", içinde *New Contributions in Information Systems and Technologies*, A. Rocha, A. M. Correia, S. Costanzo, ve L. P. Reis, Ed., Cham: Springer International Publishing, 2015, ss. 227-236. doi: 10.1007/978-3-319-16486-1_23.

[41] M. Kahla, Z. G. Yang, ve A. Novák, "Cross-lingual Fine-tuning for Abstractive Arabic Text Summarization", içinde *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, R. Mitkov ve G. Angelova, Ed., Held Online: INCOMA Ltd., Eyl. 2021, ss. 655-663. Erişim: 18 Haziran 2025. [Çevrimiçi]. Erişim adresi: https://aclanthology.org/2021.ranlp-1.74/

[42] A. T. Al-Taani ve M. M. Al-Omour, "An Extractive Graph-based Arabic Text Summarization Approach".

[43] Z. Ali, "Multilingual Text Summarization based on LDA and Modified PageRank", Oca. 2020.

[44] G. Alselwi ve T. Taşcı, "Extractive Arabic Text Summarization Using PageRank and Word Embedding", *Arab J Sci Eng*, c. 49, sy 9, ss. 13115-13130, Eyl. 2024, doi: 10.1007/s13369-024-08890-1.

[45] A. T. Al-Taani ve S. H. Al-Sayadi, "Extractive text summarization of arabic multi-document using fuzzy C-means and Latent Dirichlet Allocation", *Int J Syst Assur Eng Manag*, c. 15, sy 2, ss. 713-726, Şub. 2024, doi: 10.1007/s13198-022-01783-2.

[46] L. A. Qassem, D. Wang, H. Barada, A. Al-Rubaie, ve N. Almoosa, "Automatic Arabic Text Summarization Based on Fuzzy Logic", içinde *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*, M. Abbas ve A. A. Freihat, Ed., Trento, Italy: Association for Computational Linguistics, Eyl. 2019, ss. 42-48. Erişim: 14 Nisan 2025. [Çevrimiçi]. Erişim adresi: https://aclanthology.org/W19-7406/

[47] M. Tomer ve M. Kumar, "Multi-document extractive text summarization based on firefly algorithm", *Journal of King Saud University - Computer and Information Sciences*, c. 34, sy 8, Part B, ss. 6057-6065, Eyl. 2022, doi: 10.1016/j.jksuci.2021.04.004.

[48] R. Z. Al-Abdallah ve A. T. Al-Taani, "Arabic Single-Document Text Summarization Using Particle Swarm Optimization Algorithm", *Procedia Computer Science*, c. 117, ss. 30-37, Oca. 2017, doi: 10.1016/j.procs.2017.10.091.

[49] A. Al-Saleh ve M. E. B. Menai, "Ant Colony System for Multi-Document Summarization", içinde *Proceedings of the 27th International Conference on Computational Linguistics*, E. M. Bender, L. Derczynski, ve P. Isabelle, Ed., Santa Fe, New Mexico, USA: Association for Computational Linguistics, Ağu. 2018, ss. 734-744. Erişim: 18 Haziran 2025. [Çevrimiçi]. Erişim adresi: https://aclanthology.org/C18-1062/

[50] R. S. Baraka ve S. N. Al Breem, "Automatic Arabic Text Summarization for Large Scale Multiple Documents Using Genetic Algorithm and MapReduce", içinde *2017 Palestinian International Conference on Information and Communication Technology (PICICT)*, May. 2017, ss. 40-45. doi: 10.1109/PICICT.2017.32.

[51] "Automatic Text Summarization with Genetic Algorithm-Based Attribute Selection", içinde *Lecture Notes in Computer Science*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, ss. 305-314. doi: 10.1007/978-3-540-30498-2_31.

[52] M. Litvak, M. Last, ve M. Friedman, "A new approach to improving multilingual summarization using a genetic algorithm: 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010", *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Conference Proceedings*, ss. 927-936, Oca. 2010.

[53] L. Suanmali, N. Salim, ve M. S. Binwahlan, "Fuzzy Genetic Semantic Based Text Summarization", içinde *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*, Ara. 2011, ss. 1184-1191. doi: 10.1109/DASC.2011.192.

[54] C. Thaokar ve L. Malik, "Test model for summarizing hindi text using extraction method", içinde *2013 IEEE Conference on Information & Communication Technologies*, Nis. 2013, ss. 1138-1143. doi: 10.1109/CICT.2013.6558271.

[55] D. Kadam, “A Comparative Study Of Hindi Text Summarization Techniques: Genetic Algorithm And Neural Network”, Oca. 2015, Erişim: 22 Temmuz 2025. [Çevrimiçi]. Erişim adresi: https://www.academia.edu/96741143/A_Comparative_Study_Of_Hindi_Text_Summarization_Techniques_Genetic_Algorithm_And_Neural_Network

[56] E. Vázquez, R. A. García Hernández, ve Y. Ledeneva, “Sentence features relevance for extractive text summarization using genetic algorithms”, *Journal of Intelligent & Fuzzy Systems*, c. 35, ss. 1-13, Haz. 2018, doi: 10.3233/JIFS-169594.

[57] M. A. Fattah ve F. Ren, “Automatic Text Summarization”, *International Journal of Electrical and Computer Engineering*, 2008.

[58] I. Tanfouri, G. Tlik, ve F. Jarray, “An automatic arabic text summarization system based on genetic algorithms”, *Procedia Computer Science*, c. 189, ss. 195-202, Oca. 2021, doi: 10.1016/j.procs.2021.05.083.

[59] A. Qaroush, I. Abu Farha, W. Ghanem, M. Washaha, ve E. Maali, “An efficient single document Arabic text summarization using a combination of statistical and semantic features”, *Journal of King Saud University - Computer and Information Sciences*, c. 33, sy 6, ss. 677-692, Tem. 2021, doi: 10.1016/j.jksuci.2019.03.010.

[60] “Multidocument Arabic Text Summarization Based on Clustering and Word2Vec to Reduce Redundancy”. Erişim: 14 Nisan 2025. [Çevrimiçi]. Erişim adresi: https://www.mdpi.com/2078-2489/11/2/59

[61] A. Abu Nada, E. Alajrami, A. Alsaqqa, ve S. Abu-Naser, *Arabic Text Summarization Using AraBERT Model Using Extractive Text Summarization Approach*. 2020.

[62] K. N. Elmadani, M. Elgezouli, ve A. Showk, “BERT Fine-tuning For Arabic Text Summarization”, 29 Mart 2020, *arXiv*: arXiv:2004.14135. doi: 10.48550/arXiv.2004.14135.

[63] “Abstractive Arabic Text Summarization Based on Deep Learning - Wazery - 2022 - Computational Intelligence and Neuroscience - Wiley Online Library”. Erişim: 14 Nisan 2025. [Çevrimiçi]. Erişim adresi: https://onlinelibrary.wiley.com/doi/full/10.1155/2022/1566890

[64] D. Suleiman ve A. Awajan, “Multilayer encoder and single-layer decoder for abstractive Arabic text summarization”, *Knowledge-Based Systems*, c. 237, s. 107791, Şub. 2022, doi: 10.1016/j.knosys.2021.107791.

[65] R. Elbarougy, G. Behery, ve A. El Khatib, “Extractive Arabic Text Summarization Using Modified PageRank Algorithm”, *Egyptian Informatics Journal*, c. 21, sy 2, ss. 73-81, Tem. 2020, doi: 10.1016/j.eij.2019.11.001.

[66] D. Suleiman ve A. Awajan, “DEEP LEARNING BASED ABSTRACTIVE ARABIC TEXT SUMMARIZATION USING TWO LAYERS ENCODER AND ONE LAYER DECODER”, . *Vol.*, sy 16, 2005.

[67] I. Imam, N. Nounou, A. Hamouda, ve H. Allah Abdul Khalek, “An Ontology-based Summarization System for Arabic Documents (OSSAD)”, *IJCA*, c. 74, sy 17, ss. 38-43, Tem. 2013, doi: 10.5120/12980-0237.

[68] S. A. Waheeb, “Multi-document text summarization using text clustering for Arabic Language”, masters, Universiti Utara Malaysia, 2014. Erişim: 19 Eylül 2025. [Çevrimiçi]. Erişim adresi: https://etd.uum.edu.my/4373/

[69] “(PDF) Lexical Cohesion and Entailment based Segmentation for Arabic Text Summarization (LCEAS)”. Erişim: 19 Eylül 2025. [Çevrimiçi]. Erişim adresi: https://www.researchgate.net/publication/274223839_Lexical_Cohesion_and_Entailment_based_Segmentation_for_Arabic_Text_Summarization_LCEAS

[70] Y. A. Jaradat ve A. T. Al-Taani, “Hybrid-based Arabic single-document text summarization approach using genatic algorithm”, içinde *2016 7th International Conference on Information and Communication Systems (ICICS)*, Nis. 2016, ss. 85-91. doi: 10.1109/IACS.2016.7476091.

[71] R. Z. Al-Abdallah ve A. T. Al-Taani, “Arabic Single-Document Text Summarization Using Particle Swarm Optimization Algorithm”, *Procedia Computer Science*, c. 117, ss. 30-37, 2017, doi: 10.1016/j.procs.2017.10.091.

[72] Q. A. Al-Radaideh ve D. Q. Bataineh, “A Hybrid Approach for Arabic Text Summarization Using Domain Knowledge and Genetic Algorithms”, *Cogn Comput*, c. 10, sy 4, ss. 651-669, Ağu. 2018, doi: 10.1007/s12559-018-9547-z.

[73] “Multidocument Arabic Text Summarization Based on Clustering and Word2Vec to Reduce Redundancy”. Erişim: 14 Nisan 2025. [Çevrimiçi]. Erişim adresi: https://www.mdpi.com/2078-2489/11/2/59

[74] “Abstractive Arabic Text Summarization Based on Deep Learning - Wazery - 2022 - Computational Intelligence and Neuroscience - Wiley Online Library”. Erişim: 14 Nisan 2025. [Çevrimiçi]. Erişim adresi: https://onlinelibrary.wiley.com/doi/full/10.1155/2022/1566890

[75] “Full Text PDF”. Erişim: 19 Eylül 2025. [Çevrimiçi]. Erişim adresi: https://etd.uum.edu.my/4373/1/s812273.pdf

[76] “Extractive Arabic Text Summarization Using PageRank and Word Embedding | Arabian Journal for Science and Engineering”. Erişim: 09 Nisan 2025. [Çevrimiçi]. Erişim adresi: https://link.springer.com/article/10.1007/s13369-024-08890-1

[77] H. P. Edmundson, “New Methods in Automatic Extracting”, *J. ACM*, c. 16, sy 2, ss. 264-285, Nis. 1969, doi: 10.1145/321510.321519.

[78] A. Jain, A. Arora, J. Morato, D. Yadav, ve K. V. Kumar, “Automatic Text Summarization for Hindi Using Real Coded Genetic Algorithm”, *Applied Sciences*, c. 12, sy 13, Art. sy 13, Oca. 2022, doi: 10.3390/app12136584.

[79] Y. A. AL-Khassawneh ve E. S. Hanandeh, “Extractive Arabic Text Summarization-Graph-Based Approach”, *Electronics*, c. 12, sy 2, Art. sy 2, Oca. 2023, doi: 10.3390/electronics12020437.

[80] V. C. da Silva, J. P. Papa, ve K. A. P. da Costa, “Extractive Text Summarization Using Generalized Additive Models with Interactions for Sentence Selection”, içinde *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2023, ss. 737-745. doi: 10.5220/0011664100003417.

[81] B. Mutlu, E. A. Sezer, ve M. A. Akcayol, “Multi-document extractive text summarization: A comparative assessment on features”, *Knowledge-Based Systems*, c. 183, s. 104848, Kas. 2019, doi: 10.1016/j.knosys.2019.07.019.

[82] R. Al-Hashemi, “Text Summarization Extraction System (TSES) Using Extracted Keywords”.

[83] V. Negi, “Text Summarization Using Machine Learning”, c. 13, 2023.

[84] B. Alshemaimri, I. Alrayes, T. Alothman, F. Almalik, ve M. Almotlaq, *Summarizing Arabic Articles using Large Language Models*. 2024, s. 23. doi: 10.5121/csit.2024.141002.

[85] I. Tanfouri ve F. Jarray, “Genetic Algorithm and Latent Semantic Analysis based Documents Summarization Technique”:, içinde *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Valletta, Malta: SCITEPRESS - Science and Technology Publications, 2022, ss. 223-227. doi: 10.5220/0011585700003335.

[86] R. Z. Al-Abdallah ve A. T. Al-Taani, “Arabic Single-Document Text Summarization Using Particle Swarm Optimization Algorithm”, *Procedia Computer Science*, c. 117, ss. 30-37, Oca. 2017, doi: 10.1016/j.procs.2017.10.091.

EKLER

EK A. Arapça Durdurma Kelime Listesi

EK B. Veri Setinden Örnek Metin ve İnsan Özetleri

EK C. Pseudocode: GA Tabanlı Hibrit Arapça Metin Özetleme

EK A.

Tablo A.1. Arapça durdurma kelime listesi.

‘	-	‘	-	ء	ءَ	آ	أ
ا	ا?	االا	االتى	آب	أبٌ	ابتدأ	أبدا
أبريل	أبو	ابين	اتخذ	اثر	اثنا	اثنان	اثني
اثنين	أجل	اجل	أجمع	أحد	احد	إحدى	أخٌ
أخبر	أخذ	آخر	اخرى	اخلولق	أخو	إذ	إذا
إذاً	اذا	آذار	إذما	إذن	أربع	أربعاء	أربعة
اربعة	أربعمائة	أربعمئة	اربعون	اربعين	ارتدّ	أرى	إزاء
استحال	أسكن	أصبح	اصبح	أصلا	آض	إضافي	أضحى
اضحى	اطار	أطعم	اعادة	أعطى	أعلم	اعلنت	أغسطس
أُفٍّ	أفٍّ	اف	أفريل	به أفعل	أقبل	أكتوبر	أكثر
اكثر	اكد	آل	أل	ألا	إلا	إلّا	الا
الاخيرة	الألاء	الألى	الآن	الان	الاول	الاولى	التي
التى	الثاني	الثانية	الحالي	الذاتي	الذي	الذى	الذين
السابق	ألف	الف	ألفى	اللاتي	اللتان	اللتيا	اللتين
اللذان	اللذين	اللواتي	الماضي	المقبل	الوقت	إلي	إلى
الي	الى	إلَيْكَ	إليكَ	إليكم	إليكما	إليكنّ	اليه
اليها	اليوم	أم	أما	أمّا	إما	إمّا	اما
أمام	امام	أمامك	أمامكَ	أمد	أمس	امس	أمسى
امسى	آمينَ	أن	أنَّ	إن	إنَّ	ان	أنا
آناء	أنبأ	انبرى	أنت	أنتِ	انت	أنتم	أنتما
أنتن	أنشأ	آنفا	أنفسكم	أنفسنا	أنفسهم	انقلب	أنه
إنه	انه	أنها	إنها	انها	أنّى	آه	آهٍ
آهِ	آهاً	أهلا	أو	او	أوت	أوشك	أول
اول	أولاء	أولالك	أولئك	أوّهْ	أي	أيّ	أى
إى	اي	اى	ى?ا	أيا	أيار	ايار	إياك
إياكم	إياكما	إياكن	ايام	ّأيّان	أيّان	إيانا	إياه
إياها	إياهم	إياهما	إياهن	إياي	أيضا	ايضا	أيلول
أين	إيهٍ	ب	باء	بات	باسم	بأن	بإن
بان	بخٍ	بد	بدلا	برس	بَسْ	بسّ	بسبب
بشكل	بضع	بطآن	بعد	بعدا	بعض	بعيدا	بغتة

Tablo A.2. Arapça durdurma kelime listesi.

بؤسا	بهذا	بها	به	بن	بلى	بَلْهَ	بل
تارة	تاء	ت	ة	بينما	بين	بيد	بئس
تحوّل	'تحت	تحت	تجاه	تبدّل	تانِك	تانِ	تاسع
تسعين	تسعون	تسعمئة	تسعمائة	تسعة	تسع	ترك	تخذ
تلقاء	تكون	تفعلين	تفعلون	تفعلان	تعلّم	تعسا	تشرين
ث	تينك	تَيْنِ	تِي	تِه	تموز	تم	تلك
ثلاثاء	ثلاث	ثانية	ثاني	ثان	ثامن	ثالث	ثاء
ثمّ	ثمَّ	ثم	ثلاثين	ثلاثون	ثلاثمئة	ثلاثمائة	ثلاثة
ثمنمئة	ثمّة	ثمانين	ثمانية	ثماني	ثمانون	ثمانمئة	ثمان
جنيه	جميع	جمعة	جلل	جعل	جدا	جانفي	ج
حار	حادي	حاء	ح	جيم	جير	جويلية	جوان
حدَث	حجا	حتى	حبيب	حبذا	حاي	حاليا	حاشا
حمو	حمدا	حمٌ	حقا	حسب	حزيران	حرى	حَذارِ
خاء	خ	حين	حيثما	حيث	حيَّ	حول	حوالى
خلال	خلافا	خلا	خَبَّر	خامس	خال	خاصة	خارج
خميس	خمسين	خمسون	خمسمئة	خمسمائة	خمسة	خمس	خلف
دونك	دون	دولار	دواليك	درى	درهم	دال	د
ذال	ذاك	ذات	ذا	ذ	دينار	ديك	ديسمبر
ذيت	ذِي	ذو	ذهب	ذِه	ذلك	ذانك	ذانِ
رُبَّ	رأى	راح	رابع	راء	ر	ذينك	ذَيْنِ
زعم	زاي	ز	ريث	ريال	رويدك	رزق	رجع
سبتمبر	سبت	سادس	سابع	ساء	س	زيارة	زود
ست	سبعين	سبعون	سبعمئة	سبعمائة	سبعة	سبع	سبحان
سرا	سحقا	ستين	ستون	ستمئة	ستمائة	ستكون	ستة
سوى	سوف	سنوات	سنتيم	سنة	سمعا	سقى	سرعان
شرع	شخصا	شتانَ	شَتَّانَ	شبه	شباط	ش	سين
صباحا	صباح	صار	صاد	ص	شين	شيكل	شمال
ض	صهْ	صهٍ	صفر	صراحة	صدقا	صبرا	صبر
طالما	طاق	طاء	ط	ضمن	ضد	ضحوة	ضاد
ظنَّ	ظلّ	ظل	ظاء	ظ	طَق	طفق	طرا

Tablo A.3. Arapça durdurma kelime listesi.

ع	عاد	عاشر	عام	عاما	عامة	عجبا	عدَّ
عدا	عدة	عدد	عَدَسْ	عدم	عسى	عشر	عشرة
عشرون	عشرين	عل	علَّ	علق	علم	علي	على
عليك	عليه	عليها	عن	عند	عندما	عنه	عنها
عوض	عيانا	عين	غ	غادر	غالبا	غدا	غداة
غير	غين	ف	فاء	فأن	فإن	فان	فانه
فبراير	فرادى	فضلا	فعل	فقد	فقط	فكان	فلان
فلس	فما	فهو	فهي	فهى	فو	فوق	في
فى	فيفري	فيه	فيها	ق	قاطبة	قاف	قال
قام	قبل	قد	قرش	قطّ	قلما	قليل	قوة
ك	كاد	كاف	كأن	كأنّ	كان	كانت	كانون
كأيّ	كأيّن	كثيرا	كِخ	كذا	كذلك	كرب	كسا
كل	كلا	كلًّا	كلتا	كلم	كلّما	كم	كما
كن	كى	كيت	كيف	كيفما	ل	لا	لات
لازال	لاسيما	سيما لا	لام	لأن	لايزال	لبيك	لدن
لدي	لدى	لديه	لذلك	لعل	لعلَّ	لعمر	لقاء
لك	لكن	لكنَّ	لكنه	للامم	لم	لما	لمّا
لماذا	لن	لنا	له	لها	لهذا	لهم	لو
لوكالة	لولا	لوما	لي	ليت	ليرة	ليس	ليسب
م	ما	أفعله ما	ماانفك	انفك ما	مابرح	برح ما	مادام
ماذا	مارس	مازال	مافتئ	ماي	مائة	مايزال	مايو
متى	مثل	مذ	مرة	مرّة	مساء	مع	معاذ
معظم	معه	معها	مقابل	مكانَك	مكانكم	مكانكما	مكانكنّ
مليار	مليم	مليون	مما	من	منذ	منه	منها
مه	مهما	مئة	مئتان	ميم	ن	نَّ	نا
نبًّا	نحن	نحو	نَخْ	نعم	نفس	نفسك	نفسه
نفسها	نفسي	نهاية	نوفمبر	نون	نيسان	نيف	ه
ها	هاء	هَاتانِ	هَاتِه	هَاتِي	هَاتَيْنِ	هاكَ	هبّ
هَجْ	هذا	هَذا	هَذانِ	هذه	هَذِه	هَذِي	هَذَيْنِ
هكذا	هل	هلّا	هللة	هلم	هم	هما	همزة

Tablo A.4. Arapça durdurma kelime listesi.

هن	هنا	هناك	هنالك	هو	هؤلاء	هَؤلاء	هي
هى	هيا	هيّا	هيهات	هَيْهات	ؤ	و	6و
وا	وأبو	واحد	واضاف	واضافت	واكد	والتي	والذي
وأن	وإن	وان	واهاً	واو	واوضح	وبين	وثي
وجد	وجود	وراءك	ورد	وُشْكَانَ	وعلى	وفي	وقال
وقالت	وقد	وقف	وكان	وكانت	وكل	ولا	ولايزال
ولكن	ولم	ولن	وله	وليس	وما	ومع	ومن
وهب	وهذا	وهو	وهي	وهى	وَيْ	ي	ى
ئ	ياء	يجري	يفعلان	يفعلون	يكون	يلي	يمكن
يمين	ين	يناير	ينبغي	يوان	يورو	يوليو	يوم
يونيو							

EK B:

Tür	Metin
Orijinal Metin (parça)	ارتفع الدولار أمس بعد أن أظهر تقرير انخفاضًا غير متوقع في العجز التجاري الأمريكي في شهر مارس. وتقلص العجز إلى 54.99 مليار دولار في مارس من الرقم المعدل في فبراير وهو 60.57 مليار دولار وجاء أقل بكثير من توقعات الاقتصاديين بعجز حجمه 61.5 مليار دولار. وعلى مدى السنوات الثلاث الماضية تقريبًا ظل العجز التجاري مصدر ضغط كبير على الدولار. وإذا واصل الأمريكيون شراء بضائع أجنبية بمعدل أسرع من قدرة الشركات الأمريكية على بيع بضائعها وخدماتها في الخارج فستظل حركة الدولار إلى الخارج قوية مما يفرض ضغوطًا على العملة. وهبط اليورو إلى أدنى مستويات الجلسة عند 1.2821 دولار بانخفاض حاد عن 1.2875 دولار قبل صدور البيانات بقليل وبانخفاض 0.4 % عن مستوياتها في أواخر معاملات نيويورك يوم أول من أمس. وأمام الين صعد الدولار إلى نحو 105.70 ينًا ارتفاعًا من 105.35 ينًا قبل صدور التقرير وبارتفاع 0.1 % عن أواخر معاملات يوم أول من أمس.
İnsan Özeti (Summary1)	ارتفع الدولار امس بعد ان اظهر تقرير انخفاضا غير متوقع في العجز التجاري الامريكي في شهر مارس. 5 مليار دولار. واذا واصل الامريكيون شراء بضائع اجنبية بمعدل اسرع من قدرة الشركات الامريكية على بيع بضائعها وخدماتها في الخارج فستظل حركة الدولار الى الخارج قوية مما يفرض ضغوطا على العملة.
İnsan Özeti (Summary2)	ارتفع الدولار امس بعد ان اظهر تقرير انخفاضا غير متوقع في العجز التجاري الامريكي في شهر مارس. وتقلص العجز الى 54.99 مليار دولار في مارس من الرقم المعدل في فبراير وهو 60.
İnsan Özeti (Summary3)	ارتفع الدولار امس بعد ان اظهر تقرير انخفاضا غير متوقع في العجز التجاري الامريكي في شهر مارس. وتقلص العجز الى 54.99 مليار دولار في مارس من الرقم المعدد في فبراير وهو 60.57 مليار دولار وجاء اقل بكثير من توقعات الاقتصاديين بعجز حجمه 61.
İnsan Özeti (Summary4)	ارتفع الدولار امس بعد ان اظهر تقرير انخفاضا غير متوقع في العجز التجاري الامريكي في شهر مارس. 5 مليار دولار. واذا واصل الامريكيون شراء بضائع اجنبية بمعدل اسرع من قدرة الشركات الامريكية على بيع بضائعها وخدماتها في الخارج فستظل حركة الدولار الى الخارج قوية مما يفرض ضغوطا على العملة.
İnsan Özeti (Summary5)	وتقلص العجز الى 54. وعلى مدى السنوات الثلاث الماضية تقريبا ظل العجز التجاري مصدر ضغط كبير على الدولار. واذا واصل الامريكيون شراء بضائع اجنبية بمعدل اسرع من قدرة الشركات الامريكية على بيع بضائعها وخدماتها في الخارج فستظل حركة الدولار الى الخارج قوية مما يفرض ضغوطا على العملة.

EK C.

```
INPUT: Raw Arabic documents D

# -------------------------
# Adım 1: Preprocessing
# -------------------------
HER belge d ∈ D için:
    sentences = split_sentences(d)
    clean_sentences[d] = []
    HER cümle s ∈ sentences için:
        s = normalize_and_clean(s)
        tokens = tokenize(s)
        tokens = remove_stopwords(tokens)
        tokens = stem_tokens(tokens)
        clean_sentences[d].append(join(tokens))

# -------------------------
# Adım 2: Feature Extraction
# -------------------------
HER belge d ∈ clean_sentences için:
    sentences = clean_sentences[d]

    F_sp = compute_sentence_position(sentences)
    F_nd = compute_numerical_ratio(sentences)
    F_ls = compute_length_ratio(sentences)
    F_kw = keyword_overlap(sentences, keywords[d])
    F_ne = entity_overlap(sentences, entities[d])
    F_eh = english_overlap(sentences, english_words[d])
    F_ti = tfidf_high_density(sentences)
    F_ss = arabert_similarity(sentences)

    feature_matrix[d] = combine_features([F_sp, F_nd, F_ls, F_kw, F_ne, F_eh, F_ti, F_ss])

# -------------------------
# Adım 3: Sentence Scoring
# -------------------------
HER belge d için:
    scores[d] = weighted_sum(feature_matrix[d], weights, mode)

# -------------------------
# Adım 4: Sentence Similarity
# -------------------------
HER belge d için:
    sim_matrix[d] = build_similarity_arabert(sentences)

# -------------------------
# Adım 5: Genetic Algorithm
# -------------------------
PARAMETERS: POP_SIZE, GENERATIONS, COMPRESSION_RATE, MUTATION_RATE, ELITISM, TOURNAMENT_K

HER belge d için:
    n = number of sentences
    L = round(COMPRESSION_RATE * n)

    population = init_population(n, L)

    GENERATIONS kez tekrarla:
        fitness_values = []
        HER chromosome c ∈ population için:
            f = fitness(c, scores[d], sim_matrix[d])
            fitness_values.append(f)

        elites = select_top(population, fitness_values, ELITISM)
        mating_pool = tournament_selection(population, fitness_values, TOURNAMENT_K)

        new_population = []
        HER (p1, p2) ∈ mating_pool için:
            child1 = crossover(p1, p2, L, n)
            child2 = crossover(p2, p1, L, n)
            child1 = mutate(child1, n, MUTATION_RATE)
            child2 = mutate(child2, n, MUTATION_RATE)
            new_population.add([child1, child2])

        population = replace_with_elites(new_population, elites)

    best = argmax_fitness(population, scores[d], sim_matrix[d])
    summary[d] = sentences_in(best)

# -------------------------
# Adım 6: Evaluation (ROUGE)
# -------------------------
HER reference_set R ∈ {summary1...summary5} için:
    HER belge i için:
        gen_summary = summary[i]
        ref_summary = R[i]

        scores_ROUGE[i] = {
            ROUGE-1: precision/recall/F1(ref_summary, gen_summary, n=1),
            ROUGE-2: precision/recall/F1(ref_summary, gen_summary, n=2),
            ROUGE-L: precision/recall/F1(ref_summary, gen_summary, LCS)
        }

    aggregate_scores(macro, micro)
    save_to_JSON(parameters, metrics)

OUTPUT: Best summaries per document + ROUGE metrics
```

ÖZGEÇMİŞ

Ad-Soyad : Hedil ELŞAVİ

ÖĞRENİM DURUMU:

- **Lisans** **:** 2022, Sakarya Uygulmalı Bilimler Üniversitesi, Teknolji Fakültesi, Elektrik ve Elektronik Mühendisliği
- **Yükseklisans** **:** Devam ediyor, Sakarya Üniversitesi, Bilişim Sistemleri Mühendisliği, Bilişim Sistemler Mühendisliği

MESLEKİ DENEYİM VE ÖDÜLLER:

- Coursera platformunda başarıyla tamamladığı projeler arasında *"Diabetic Retinopathy Detection with Artificial Intelligence"*, *"Facial Expression Classification Using Residual Neural Nets"* ve *"Asasiyetli Veri Analizi: Python ve Pandas ile"* yer almaktadır. Bu projeler, yapay zekâ, derin öğrenme ve veri analizi alanlarında uygulamalı deneyim ve beceriler kazandırmıştır.

TEZDEN TÜRETİLEN ESERLER:

- Elsavi, H. ve Taşçı, T. 2025. Arapça Metinler İçin İstatistiksel Özellik Tabanlı Bir Otomatik Özetleme Yaklaşımı, *I. Uluslararası Persepolis Bilimsel Araştırmalar ve İnovasyon Kongresi*, 10–11 Mayıs 2025, İran, Kongre Kitabı içinde, ss. 266–267. Erişim: https://www.wosconkongreleri.com/ (Sunum)

DİĞER ESERLER: