

**T.C
İNÖNÜ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**METİN ÇİZGELERİNDE ENTROPİ VE OPTİMİZASYON TABANLI
ÇIKARIMSAL METİN ÖZETLEME**

CENGİZ HARK

**DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

OCAK 2020

**T.C
İNÖNÜ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**METİN ÇİZGELERİNDE ENTROPİ VE OPTİMİZASYON TABANLI
ÇIKARIMSAL METİN ÖZETLEME**

CENGİZ HARK

**DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

OCAK 2020

Tezin Başlığı : Metin Çizgelerinde Entropi ve Optimizasyon Tabanlı Çıkarımsal Metin
Özetleme

Tezi Hazırlayan : Cengiz HARK

Sınav Tarihi : 28.01.2020

Yukarıda adı geçen tez jürimizce değerlendirilerek Bilgisayar Mühendisliği Ana Bilim
Dalında Doktora Tezi olarak kabul edilmiştir.

Sınav Jüri Üyeleri

Tez Danışmanı: Prof. Dr. Ali KARCI

İnönü Üniversitesi



Prof.Dr. A. Bedri ÖZER

Turgut Özal Üniversitesi



Prof. Dr. Davut HANBAY

İnönü Üniversitesi



Dr. Öğretim Üyesi A. ARİF AYDIN

İnönü Üniversitesi



Dr. Öğretim Üyesi İhsan TUĞAL

Muş Alparslan Üniversitesi



Prof. Dr. Kazım TÜRK

Enstitü Müdürü

ONUR SÖZÜ

Doktora Tezi olarak sunduğum “Metin Çizgelerinde Entropi Ve Optimizasyon Tabanlı Çıkarımsal Metin Özetleme” başlıklı bu çalışmanın bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın tarafımdan yazıldığını ve yararlandığım bütün kaynakların hem metin içinde hem de kaynakçada yöntemine uygun biçimde gösterilenlerden oluştuğunu belirtir, bunu onurumla doğrularım.

Cengiz HARK



ÖZET

Doktora Tezi

METİN ÇİZGELERİNDE ENTROPİ VE OPTİMİZASYON TABANLI ÇIKARIMSAL METİN ÖZETLEME

Cengiz Hark

İnönü Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

78 + xiv sayfa

2020

Danışman: Prof. Dr. Ali KARCI

Metin boyutlarında meydana gelen artış, otomatik belge özetleme araçlarına duyulan ihtiyacı arttırmaktadır. Tez çalışmasında, metin belgelerinin özetlenmesi için iki yeni metodoloji; Karcı Özetleme, Çok Amaçlı Modifiye Fidan Gelişim Algoritması (ÇA-FGA) sunulmaktadır. Benzersiz bir yaklaşımla, bir belge özetleme çalışmasında Karcı Entropi kullanılmış ve literatüre kazandırılmıştır. Önerilen sistem, eğitim verisi gerektirmemektedir. Optimizasyon teknikleri özetleme çalışmalarında sık kullanılan bir başka tekniktir. Ancak bu çalışmalarda, birbirleri ile çelişen farklı amaçlar genelde ağırlıklandırılarak veya birleştirilerek tek amaçlı probleme dönüştürülmektedir. Bu dönüşüm çözüm kalitesini doğrudan etkilemektedir. Tez çalışmasında Karcı Özetleme'ye ek olarak zengin bir özetin iki temel niteliği olan maksimum kapsam ve minimum fazlalığın optimizasyonun amaçları olarak modellendiği bir yaklaşım sunulmaktadır. Eklenen budama aşaması ile Fidan Gelişim Algoritması bir adım ileri taşınmış ve ÇA-FGA tanıtılmıştır. Sunulan yaklaşım kapsamında gerçek bir eşzamanlı optimizasyon uygulanmaktadır. Sunulan yaklaşımların sonuçları açık erişimli veri setleri (DUC 2002-2004) ile test edilmişlerdir. Yaklaşımların performansları, Recall-Oriented Understudy for Gisting Evaluation (ROUGE) ölçekleri kullanılarak incelenmiştir. Deneysel sonuçlar her iki yaklaşımında gelecek vaat eden teknikler olduğunu ortaya koymaktadır.

ANAHTAR KELİMELEER: Belge özetleme, Çıkarıcı özetleme, DUC, ROUGE, Çizge entropi, Karcı entropi, Fidan gelişim algoritması, Çok amaçlı optimizasyon, Pareto cephe.



ABSTRACT

Ph.D. Thesis

ENTROPY AND OPTIMIZATION BASED EXTRACTIVE TEXT SUMMARIZATION IN TEXT GRAPHS

Cengiz Hark

İnönü University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering

78 + xiv pages

2020

Supervisor: Prof. Dr. Ali KARCI

The increase in text sizes requires the increasing the need for automatic document summarizing tools. In the thesis, two new methodologies for summarizing text documents; Karcı Summarization, Multi-Objective Sapling Growing Up Algorithm (MO-SGA) are offered. Karcı Entropy was used in a document summarizing study with a unique approach and it was introduced to the literature. The proposed system does not require training data. Optimization techniques are another technique commonly used in summarizing studies. However, in these studies, the different contradictory objectives are often weighted or combined into a single-purpose problem. This conversion directly affects the quality of the solution. In addition to Karcı Summarization, the thesis presents an approach in which the two main characteristics of a rich summary, maximum coverage and minimum redundancy, are modeled as the objectives of optimization. With the added pruning step, the Sapling Growing Up Algorithm was moved one step further and MO-FGA was introduced. A simultaneously optimization is implemented within the scope of the presented approach. The results of the proposed approaches were tested with open access data sets (DUC 2002-2004). The performances of the approaches were examined using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) criteria. Experimental results reveal that there are promising techniques in both approaches.

KEYWORDS: Document summarization, Extractive summarization, DUC, ROUGE, Graph entropy, Karcı entropy, Sapling growing up algorithm, Multi-objective, Pareto front.



TEŞEKKÜR

Lisans ve Doktora öğrenimimin her aşamasında değerli katkılarını, yorumlarını, yapıcı eleştirilerini benden esirgemeyen, zamanını ve engin bilgisini benimle cömertçe paylaşan danışman hocam Prof. Dr. Ali KARCI'ya; gerek ders, gerekse tez döneminde bana kapıları sonuna dek açık olan, her türlü destek ve yardımı sunan Bilgisayar Mühendisliği Anabilim Dalı öğretim üyelerine, kendisiyle ortak çalışmalar yaptığım, fikir alışverişinde bulunduğum arkadaşım Taner UÇKAN'a teşekkürlerimi sunarım. Ayrıca yetişmemde sonsuz emeği olan altın kalpli anneme, tanıştığım ilk günden beri benim için her türlü özveriyi göstermekten imtina etmeyen eşime, bu süreçte kendilerini ihmal ettiğim varlık sebebim çocuklarım Eyüp Kerem ve Elif'e, tutunacak dallarım kardeşlerime duyduğum şükran ve minnet borcumu bu vesileyle belirtmek isterim.

teşekkür ederim.

İÇİNDEKİLER

ÖZET	v
ABSTRACT.....	vii
TEŞEKKÜR.....	ix
SİMGELER VE KISALTMALAR.....	xii
ŞEKİLLER LİSTESİ	xiii
1. GİRİŞ	1
1.1. Metin Özetlemeye Giriş	3
1.2. Amaç ve Kapsam	5
1.3. Tezde Yapılan Çalışmalar ve Literatüre Katkısı.....	7
1.4. Tezin Organizasyonu	9
1.5. Tezin Çıktıları	9
2. İLİŞKİLİ LİTERATÜR	11
3. ÇİZGE ENTROPİSİ	15
3.1. Entropi Türleri	16
3.1.1. Shannon Entropi.....	16
3.1.2. Karcı Entropi.....	17
3.2. Çizge Entropi ile Merkezilik.....	19
4. FİDAN GELİŞİM ALGORİTMASI.....	21
4.1. Fidanların Ekim Aşaması.....	21
4.2. Fidanların Büyüme Aşaması.....	23
4.2.1. Eşleşme	23
4.2.2. Dallanma	24
4.2.3. Aşılama	25
4.3. Algoritma	26
5. YÖNTEM VE UYGULAMALAR.....	28
5.1. Metin Önışleme ve Hazırlık.....	28
5.2. Metinsel Çizge Oluşumu.....	31
5.3. Kullanılan Veri Setleri	32
5.4. Değerlendirme Ölçütleri	33
5.5. Çizge Entropi ile Metin Özetleme Uygulaması	34
5.6. Karcı Özetleme: Karcı Entropi tabanlı Metin Özetleme amacı ile Basit ve Etkili bir Öneri	38
5.6.1 Deneyisel Sonuçlar.....	40

5.6.2	Karcı Özetlemenin Rekabetçi Yöntemler ile Karşılaştırılması	47
5.7.	ÇA-FGA : Çok amaçlı Fidan Gelişim Algoritması ile Çıkarıcı, Genel Metin Özetleme Önerisi	53
5.7.1	Benzerlik Tabanlı Cümle Puanlandırma	55
5.7.2	Çok Amaçlı Optimizasyon Problemi Olarak Modellenmesi	56
5.7.3	Önerilen Çok Amaçlı Fidan Gelişim Algoritması	57
5.7.4	DeneySEL Sonuçlar.....	63
6.	SONUÇ VE ÖNERİLER.....	68
7.	KAYNAKLAR	71
	ÖZGEÇMİŞ	77



SİMGELER VE KISALTMALAR

E	Entropi
H	Entropi
G	Çizge (Graph)
V	Düğüm (Vertice)
E	Kenar (Edge)
i	i 'inci düğüm
j	j 'inci düğüm
$e_{i,j}$	i ve j 'inci düğüm arasındaki kenar
$DDİ$	Doğal Dil İşleme
$w_{i,j}$	i ve j 'inci düğüm arasındaki ağırlık değeri
A_{ij}	Komşuluk matrisi
n	Düğüm sayısı
k	Düğüm noktasının derecesi - Boltzmann sabiti
w_{ij}	i ve j 'inci düğüm arasındaki ağırlık değeri
sp_{jk}	j ve k düğümleri arasındaki kısa yolların sayısı
$sp_{jk}(i)$	i düğümünden geçen j ve k düğümleri arasındaki kısa yolların sayısı
d_{ij}	i ve j düğümleri arasındaki mesafe
A	$n \times n$ boyutunda benzerlik matrisi
x_i	i düğümünün özvektör merkeziliği
λ	A 'nın en büyük özdeğeri
k_i	i 'inci düğümünün düğüm derecesi
cc	Kümelene katsayısı
T	Mutlak sıcaklık
TF-TBF	Terim Frekansı-Ters Belge Frekansı
dQ	Soğurulan ısı miktarı
dS	Isıdaki değişim
S	Entropi değeri
W	Olayın veya durumun oluşma olasılığı
p_i	i 'inci gözlenebilir durum
α	Alfa
$f(x)$	Herhangi bir fonksiyon
$g(x)$	Herhangi bir fonksiyon
$E(G)$	Çizgenin toplam entropisi
$E(v_i)$	i düğümünün entropisi
U	Evrensel küme
$s(A)$	A kümesinin eleman sayısı
$\in$	Elemanıdır
$\notin$	Elemanı değildir
$\vee$	Veya (maksimum)
$\wedge$	Ve (minimum)
$\emptyset$	Boş küme
∞	Sonsuz

ŞEKİLLER LİSTESİ

Şekil 4.1 .	Düzgün dağılım ile fidanların ekimi	22
Şekil 4.2 .	Örnek bir fidan üzerinde dallanma noktaları S1 fidan olsun.	24
Şekil 5.1 .	DUC-2002, “D065” belgesine ait metinsel çizge.	32
Şekil 5.2 .	Özetleme sisteminin ana hatları.	34
Şekil 5.3 .	Karcı özetleme modeline dair şematik yapı.....	39
Şekil 5.4 .	Karcı Özetleme ile elde edilen, Recall bazlı ROUGE metrik puanları üzerindeki α etkisi (DUC-2002, DUC-2004).....	42
Şekil 5.5 .	Pareto optimal küme.	58
Şekil 5.6 .	Rouge-2 (ortalama bazlı) puanları bakımından yöntemleri performans değerlerinin karşılaştırılması.....	64
Şekil 5.7 .	Rouge-2 (ortalama bazlı) puanları bakımından yöntemleri performans değerlerinin karşılaştırılması.....	67

ÇİZELGELER DİZİNİ

Çizelge 5.1.	Önerilen yaklaşım ile DUC-2002 veri setinde bulunan “d065” dosyasında gerçekleşen örnek değişimler.	30
Çizelge 5.2.	Veri Setlerinin Karakteristikleri.	33
Çizelge 5.3.	Çalışma kapsamında ve rekabetçi yöntemlerle oluşturulan 200 kelimelik özetler için rouge metrik değerleri.	37
Çizelge 5.4 .	Çalışma kapsamında ve rekabetçi yöntemlerle oluşturulan 400 kelimelik özetler için rouge metrik değerleri.	37
Çizelge 5.5.	Önerilen Karcı Özetleme yöntemi için farklı α değerlerine dayanan 100 kelimelik özetlerin ROUGE metrik değerleri.	44
Çizelge 5.6.	Önerilen Karcı Özetleme yöntemi için farklı α değerlerine dayanan 200 kelimelik özetlerin ROUGE metrik değerleri.	45
Çizelge 5.7.	Önerilen Karcı Özetleme yöntemi için farklı α değerlerine dayanan 400 kelimelik özetlerin ROUGE metrik değerleri.	46
Çizelge 5.8 .	Rekabetçi yöntemler ile Karcı Özetlemenin karşılaştırılması (100 kelimelik özetler).	48
Çizelge 5.9 .	Rekabetçi yöntemler ile Karcı Özetlemenin karşılaştırılması (200 kelimelik özetler).	49
Çizelge 5.10 .	Rekabetçi yöntemler ile Karcı Özetlemenin karşılaştırılması (400 kelimelik özetler).	52
Çizelge 5.11 .	Son Teknoloji özetleme çalışmalarının detaylı analizleri.	54
Çizelge 5.12 .	Önerilen ÇA-FGA ile rekabetçi yöntemlerin karşılaştırılması.	65
Çizelge 5.13 .	Önerilen ÇA-FGA ile rekabetçi yöntemlerin karşılaştırılması.	66
Çizelge 5.14 .	Önerilen yaklaşımın diğer modeller ile genel karşılaştırılması.	67

1. GİRİŞ

Günümüzde bilginin hızla artması ile araştırmacılar farklı bilgi kaynaklarından anlık olarak bilgiyi elde edebilmekte ve yorumlayabilmektedir. İnternet ortamı milyarlarca belge içermekte, bu belge miktarı her an üstel bir hız ile artmaya devam etmektedir [1].

Ham veri, bir konu veya kavram hakkında birbirleri ile olan ilişkileri tam anlamı ile bilinmeyen bilgi topluluklarıdır. Bir başka ifade ile sayısal olarak ifade edilebilen, taşınabilen dizelerdir. Sözü edilen ham verinin bilinen bir amaç doğrultusunda anlamlı birliktelikleri elde edilerek bilgiye ulaşılabilmektedir [2]. Devamlı olarak büyüyen veri miktarı yeni çözümler ve bakış açılarına olan ihtiyacı arttırmaktadır. İşlenmeyi bekleyen dijital verilerin büyük çoğunluğu belirli bir formata sahip olmayan ham metinlerden oluşmaktadır. Bahsedilen verilerin anlamlı ve faydalanılabilir veri kaynaklarına dönüşmesi ve bilgiye erişim zamanının kısaltılması için analiz edilmeleri gerekmektedir [2–4]. Büyük veri kümeleri üzerinde çalışmak yüksek zaman tüketimi ve yüksek hesaplama maliyetlerini beraberinde getirmektedir. Ancak veriyi sınıflandırarak, bölütleyerek, duygusal analizler yaparak veya özetleyerek bu işlemleri çok daha az maliyet ile gerçekleştirmek mümkün olabilmektedir. İnternetin günlük hayatın bir parçası olması ve çevrimiçi bilgi miktarının artması ile yüksek hacimli belgelerin düzenlenmesini amaçlayan yöntemlere olan talep artmaktadır [5, 6]. Ancak tüm bu metin düzenleme yöntemlerinin ortak amacı yüksek kalitede doküman koleksiyonlarının seçilmesi, iyi bir model sunulması ve uygun tekniklerinin belirlenmesidir [7]. Metin dokümanları, terimleri kelimelerden oluşan koleksiyonlardır. Bu metin koleksiyonlarını anlamak ve yorumlamak zordur. Bu nedenle, metin tabanlı koleksiyonlar makineler tarafından üzerinde çalışılabilir formatlara aktarılırlar [8].

Arama motorları, kullanıcılardan bir takım sorgular alarak en ilişkili belge veya web sayfalarını sıralayabilmektedir ancak en gelişmiş arama motorları bile kaynaklardan bilgi sentezleme yeteneğine sahip değildir. Sonuç olarak kontrol dışı artan bilgi bu bilginin yönetilebilmesi sorununu beraberinde getirmektedir. Bu sorun

ile başa çıkabilmek, veri kaynaklarına erişim yaparak verileri işleyen araçlar gerektirmektedir [9]. Doküman özetleme alanındaki en güncel gelişmelere rağmen problem tam olarak çözülememiştir. Otomatik doküman özetleme sistemleri, önemli bilgileri yoğunlaştırılmış bir formda elde etmeyi ve kullanıcılara sunmayı amaçlamaktadır [10]. Bilgiye erişim zamanının kısaltılması amacı ile verinin analiz edilerek bir takım özelliklerin elde edilmesi gerekmektedir [11]. Bu sorunlar otomatik özetleme sistemlerine olan ilgiyi arttırmıştır [9].

Bir olay veya durumun gerçekleşme ihtimali gözlemlendiğinde, ne oranda belirsizlik taşıdığına bakılmaksızın, olayın kaynağından edinilen bilginin ortalama içeriği hakkında bir fikir veya gözleme sahip olunma amacı taşınır. Entropi her kaynak için ortalama bilginin bir ölçümü olarak tanımlanır. Bilgi entropisi ise, bilginin stokastik bir veri kaynağı tarafından üretildiği ortalama orandır [12].

Bu tezde gerçekleştirilen metin özetleme çalışmaları kapsamında sunduğumuz ilk yaklaşım entropi bazlı bir metottur. Karcı Entropi [13–15] ile paragraf veya metin parçacıkları içerisinde etkili olan genelleyici ve en çok bilgi taşıyan cümleleri seçmek için Karcı Entropi bazlı bir çözüm önerilmektedir. Çalışmada Karcı Entropi kullanılmasının nedeni en popüler entropi yöntemi olan Shannon Entropi'den [12, 16] daha üstün ve yeni bir yaklaşım olmasıdır [12, 17, 18]. Dahası Karcı Entropi'de yer alan farklı α üs değerleri ile kelime öbeklerinin yerel ve global ağırlıklarını elde edebilmek mümkün olabilmektedir [1].

Geçmişte metin özetleme çalışmaları istatistiksel araçlar kullanılarak gerçekleştirilmiştir. Ancak istatistiksel araçlar oldukça zayıf performanslar sergilemektedirler. Doküman özetleme performansını arttırmak amacı ile son yıllarda araştırmacılar metin özetleme problemini bir optimizasyon problemi olarak değerlendirmektedirler [19–21]. Tez kapsamında tanıtılan ikinci yaklaşım ise optimizasyon temelli metin özetleme yaklaşımıdır. Özeti aranan metinden çıkarıcı özetleme yolu ile en uygun cümlelerin seçilmesi, problemin çözümünden elde edilecek maksimum faydanın sağlanması ve/veya minimum zararın edilmesi tamamen optimum seçimle mümkündür. Bu süreçlerin tümü optimizasyonun alanında değerlendirilir ve çözümlenmeye çalışılır.

Bu doktora tez çalışmasında, metin madenciliği yöntemlerinden biri olan ve her geçen gün daha fazla ihtiyaç duyulan otomatik metin özetlemeye odaklanılmıştır.

Araştırma kapsamında, yapılandırılmamış günlük konuşma dilinde ifade edilen metinlerin, normalize edilmelerine, matematik ve optimizasyon destekli yaklaşımlarla analiz edilmelerine yönelik geniş bir ölçekte araştırmalar yapılmıştır.

Bu amaç ile gerçekleştirilen deneysel süreçleri tamamlayıcı nitelikte literatürde bulunan çalışmalara yer verilmiş akabinde deneysel süreçler boyunca kullanılan bazı kavram ve konular sunulmuştur. Son olarak rekabetçi yöntemlere alternatif, farklı alt yapılara sahip entropi ve optimizasyon temelli iki yeni yöntem önerilmiştir. Elde edilen sonuçlar gerçekleştirilen her uygulama bazından ayrı ayrı rekabetçi yöntemlerle karşılaştırılmış, gerçekleştirilmesi planlanan gelecek çalışmalara yönelik bilgiler sunulmuştur.

Bu bölümde tezin amaç ve kapsamı genel olarak açıklanmış, tez kapsamında gerçekleştirilen çalışmalardan bahsedilerek literatüre olan katkısı açıklanmış, son olarak, tezin organizasyonuna ve tezin çıktılarına yer verilmiştir.

1.1. Metin Özetlemeye Giriş

Uzun metin belgelerinin insanlar tarafından özetlenmesi hem çok zor hem de gerçekten fazla zaman gerektirmektedir. İş dünyası, akademi, sağlık gibi birçok alanda özetleme oldukça gerekli görülmektedir. Otomatik metin özetleme problemine ilişkin olarak farklı tiplerde araştırmalar yürütülmektedir. Günümüz okuyucuları her zamankinden daha fazla, metinleri kısa özetlere dönüştürebilen etkili metin özetleme sistemlerine ihtiyaç duymaktadırlar. Sınırlı bir zaman dilimi içerisinde belirlenen bir konu kapsamında ve tatmin edici bir düzeyde okuyucunun gereksinimlerini karşılayan tek bir metin belgesi oluşturmak oldukça zor bir problemdir. Sözü edilen probleme bir çözüm sunması bakımından belge özetleme sistemleri kullanılmaktadır. Otomatik doküman özetleme işlemi belge/belgelerin önemli özelliklerini koruyarak daha özlü bir versiyonunu elde etme işlemidir. Otomatik çok belgeli özetleme sistemleri ile bilgiye daha hızlı bir bakışın mümkün olabilmesi için belgelerin ideale yakın özetlerinin oluşturulması amaçlanmaktadır [1, 11, 22–24]. 2019’da Joshi [25] otomatik metin özetlemeyi, son kullanıcılar için hızlı ve anlaşılabilir biçimde, uzun metin belgelerini sıkıştırarak temsil etmeyi amaçlayan DDİ dalı olarak tanımlamaktadır.

Özetlenecek doküman sayısına bağlı olarak özet, tek belgeli veya çok belgeli olabilir. Tek belgeli özetleme yalnızca bir adet belgenin daha kısa bir temsiline

yoğunlaşırken, çok belgeli özetleme bir doküman kümesinden özet elde etmeye odaklanmaktadır. Tek belgeli özetleme, çok belgeli özetleme sisteminin daha yalın bir kısmı olarak düşünülebilmektedir. Çok belgeli özetleme sistemi bir belge kümesinde yer alan bilgilerin ana hatlarını göstermek için kullanılabilir [26, 27]. Özetleme ile ilgili ilk çalışmalar tek belgeli özetleme kapsamında yapılmıştır. Devam eden çalışmalarda çok belgeli özetleme çalışmaları yapılmaya başlanmıştır ve aynı konu ile ilgili birden fazla metne uygulanan yöntemler geliştirilmiştir [9, 28–31].

Ayrıca yorumlayıcı ve çıkarıcı olmak üzere, farklı türde özetler üreten yaklaşımlar vardır. Yorumlayıcı özetler, özetlenecek metinde sunulan bilgilerin yeniden temsili ile elde edilir. Bu yüzden, elde edilen özetler girdi metinlerinden bağımsızdır. Bu özetleme biçimi, belgenin dilbilimsel ve kavramsal olarak anlaşılması ve maksimum kapsam ve tutarlılık esasına göre daha kısıtlı bir boyutta yeniden ifade edilmesini amaçlamaktadır. Öte yandan çıkarıcı özetleme, girdi dokümanlarının bir alt kümesini seçerek özet oluşturma yoluna gitmektedir. Bu yüzden çıkarıcı özetleme girdi dokümanlarına bağlı kalınarak yeniden ifade ve yorumlama olmaksızın gerçekleştirilmektedir. Çıkarıcı özetleme sistemleri, basitliği ve düşük maliyetinden ötürü daha popülerdir. Yorumlayıcı özetlemede metin içerisinde hangi cümlelerin nasıl sınırlandırılacağını ve yeniden yorumlanacağını kestirmek oldukça zordur. Bu durum önemli bir dezavantajdır [23, 32].

Oluşturulan özetlerin içeriklerine bağlı olarak, özetler sorgu bazlı (query based) veya genel özetler olarak sınıflandırılmaktadır. Sorgu bazlı özetleyiciler kullanıcıların sorgularına bağlı olarak özetleme gerçekleştirirken genel özetleyiciler herhangi bir dış bilgi ve yönlendirme olmaksızın kaynak dokümanın ana kavramlarını elde etmeye odaklanmaktadır [33]. Araştırmacılar tarafından gerçekleştirilen çalışmaların büyük çoğunluğu, genel özetleme bağlamında olmuştur, özet oluşturma amacı hakkında birkaç sınırlı varsayım ile gerçekleşen bu özetleme türünde genel konu içeriği korunarak, mümkün olduğunca fazla bilgi kapsamaya çalışılır. Genel özetlemeden farklı olarak bir özetin taşınması istenen bilgi içeriği kullanıcı tarafından özel olarak belirlenebilir. Bu şekilde sorgu odaklı özetleme kullanıcının ilgilendiği konuya odaklanır ve belirtilen konu ile ilgili metindeki bilgileri çıkarır [34, 35].

Ayrıca, doküman özetleme denetimli (supervised) denetimsiz (unsupervised) olmak üzere iki farklı şekilde gerçekleştirilebilir. Denetimli özetleme yöntemlerinde önemli cümleleri seçme görevi, özetlenecek doküman/dokümanlardaki tüm cümleleri özet ve özet olmayan cümlelere ayırarak ikili (binary) sınıflandırma problemi olarak

temsil edilir. Denetimsiz yöntemler genellikle önceden tanımlanmış bir dizi tanım ve özelliği kullanarak belgelerdeki cümleleri puanlandırırlar. Denetimli yöntemler büyük miktarda eğitim verisi gerektirirken denetimsiz öğrenme yöntemleri herhangi bir eğitim verisi gerektirmezler [1, 23, 26].

Bu tez kapsamında, metinlerden uygun özetler çıkarılması amacıyla genel, çıkarıcı ve çok belgeli yeni özetleme yöntemleri sunulmaktadır.

1.2.Amaç ve Kapsam

Bilgi iletişim teknolojilerinin gelişmesi, elektronik bilgi miktarının katlanarak artmasına neden olmaktadır. Metin koleksiyonlarına sağlanan erişim kolaylığı ise tüketicilerin önemli miktarda çevrimiçi bilgi ile karşı karşıya kalması sonucunu doğurmaktadır. Sayısal veri boyutlarının giderek büyüdüğü çağımızda, büyük veri akışı ile başa çıkabilmek için otomatik özetleme sistemleri kullanılmaktadır. Günümüz okuyucuları her zamankinden daha fazla, metinleri kısa özetlere dönüştürebilen etkili metin özetleme sistemlerine ihtiyaç duymaktadırlar [22, 36]. Sınırlı bir zaman dilimi içerisinde belirlenen bir konu kapsamında ve tatmin edici bir düzeyde okuyucunun gereksinimlerini karşılayan tek bir metin belgesi oluşturmak oldukça zor bir problemdir. Sözü edilen probleme bir çözüm sunması bakımından belge özetleme sistemleri kullanılmaktadır [23]. Yukarıda yer verilen tespitler, bu tez çalışması kapsamında, yüksek kaliteye sahip metin özetlerini elde etmek için yenilikçi ve sağlam metotlar tasarlamak noktasındaki motivasyonumuzu oluşturmuştur. Bu motivasyon ile yeni belge özetleme yöntemlerine odaklanılmıştır. Çalışmanın amacı, tüketicilerin çevrimiçi bilgi edinme süreçlerinde, bilgiye erişim sürelerini kısaltarak kalitesini arttırmak için çizge ve/veya sezgisel odaklı olmak üzere yeni metin özetleme süreçlerinin ortaya konmasıdır.

Bu yöntemlerin ilki, çizge teorisi ve entropi kavramlarının, bir arada kullanıldığı ve bu kavramların genel metin özetleme performansı üzerindeki toplam etkilerinin araştırıldığı bir çalışmadır. Özeti oluşturacak cümlelere yönelik tespitler, her bir düğümün, çizgeden birer birer çıkarılması ile toplam entropideki değişim hesaplanarak elde edilmiştir. Metinsel çizgeyi oluşturan düğümler tarafından taşınan bilginin miktarı sayısal olarak ölçülmüştür. Bu yaklaşım ile dokümanı temsil eden çizge üzerindeki küresel etki ölçülebilmektedir. Dahası cümlelerin çizgelere aktarımı aşamasında maksimum temsil hedeflenmiştir ve bu amaçla kendine özgü çalışma

prensibi ile benzersiz bir metin önişleme yöntemi sunulmuştur. Önerilen özetleme sistemi, gürültülü ve herhangi bir önişlem geçirmemiş olan belgeleri doğrudan özetlemek üzere tasarlanmıştır. Bu sebeple düzensiz verilere karşı oldukça duyarsızdır. Karcı Özetleme, her bir cümlelerin, genel metnin özünü, dışsal bilgilere ihtiyaç duymadan sayısal olarak ölçmeyi amaçlar.

Bildiğimiz kadarıyla, entropiye dayalı bu kadar ayrıntılı ve yüksek performanslı sonuçlar veren başka yayınlanmış bir yöntem mevcut değildir. Ayrıca, Karcı Entropi daha önce başka bir metin özetleme çalışmasında kullanılmamıştır. Bu nedenle, Karcı Özetleme, benzersiz ve denetimsiz olarak kabul edilir.

Tez çalışması kapsamında önerilen bir diğer metin özetleme çalışması ise, optimizasyon tabanlıdır. Bu çalışmada metin belgelerinden maksimum kapsam (maksimum coverage) ve boyut azaltma (redundancy reducing) değerleri temelinde çok amaçlı (multi-objective) bir fonksiyon kullanılarak çıkarıcı, genel bir özetleme sistemi tanıtılmaktadır. Doküman özetleme görevi bir optimizasyon problemi olarak temsil edilmiştir. Seçilen algoritma Fidan Gelişim Algoritmasıdır (Saplings Growing Up algorithm). Optimizasyon problemi çok amaçlı Fidan Gelişim Algoritması ile ele alınmıştır. Fidan Gelişim Algoritması, daha önce hiçbir özetleme çalışmasında kullanılmamıştır.

Ek olarak bu çalışmada özete seçilecek cümleler belirlenirken sadece dokümanın tamamı ile en yüksek kapsam değerine sahip cümlelere yer verilmemiştir. Özette yer alan cümle ve kavram çeşitliliğini genişlemek amacı ile fazlalık parametreside eş zamanlı olarak minimize edilmiştir. Bu duruma uygun bir modelleme yoluna gidilerek iki amaçlı bir fonksiyon oluşturulmuştur. Literatürde yer alan pek az çalışma gerçek bir şekilde eşzamanlı optimizasyon kullanarak özetleme gerçekleştirmektedir (Bkz: Çizelge 5.11). Çok amaçlı bir metin özetleme çalışmalarının birçoğu, farklı zamanlarda çalışan tek amaçlı bir kaç optimizasyon parametresinin/sonucunun matematiksel olarak bir araya getirilmesinden ibarettirler.

Önerilen yöntemler ve mevcut son teknoloji rekabetçi yöntemler, metin özetleme sistemlerinde sıkça kullanılan bir veri seti ile karşılaştırılmıştır. ROUGE puanları bakımından, önerilen modeller son teknoloji yöntemlerle karşılaştırılmış ve etkinlikleri doğrulanmıştır.

1.3. Tezde Yapılan Çalışmalar ve Literatüre Katkısı

Bu tez çalışması kapsamında doğal dil ile ifade edilmiş, herhangi bir yapısalılık taşımayan metinlerin özetlenebilmesi amacı ile çalışmalar yürütülmüştür. Yeni yaklaşımlar önerilmiştir. Tez kapsamında yapılan çalışmalar ve literatüre olan katkıları aşağıdaki gibidir;

Karcı Özetleme: Karcı Entropi kullanılarak otomatik metin özetlemesi için basit ve etkili bir yaklaşım:

- Cümleler arasındaki ilişkilerin, temsili metinsel çizgelere en doğru ve etkili bir biçimde aktarılabilmesi amacı ile benzersiz bir yaklaşım önerilmektedir. Bu yeni yaklaşım maksimum temsili hedeflemektedir. Bu yeni yöntem, bazı iyi bilinen metin normalleştirme yöntemlerinden (ör. Spacy, NLTK, OpenNLP, Stanford CoreNLP ve ClearNLP) farklı ve benzersiz bir yaklaşımdır.
- Çalışma kapsamında Çizge teorisi ve Karcı Entropi kavramları bir arada ve etkili bir biçimde kullanılmış ve çıkarıcı, genel bir metin özetleme sistemi modellenmiştir.
- Temsili metinsel çizgelere Karcı Entropi uygulanarak benzersiz ve matematiksel altyapıya sahip yeni bir denetimsiz model sunulmaktadır.
- Karcı Özetleme yaklaşımı herhangi bir eğitim verisi gerektirmeden her cümlenin, metnin tamamını ne oranda temsil ettiğine (sayısal olarak cümlelerin ağırlıklandırılması) yönelik bir ölçüm gerçekleştirmektedir.
- Bir metin özetleme çalışmasında ilk defa kullanılan Karcı entropi de bulunan α parametresinin yerel ve genel etkisine yönelik çıkarımlar gerçekleştirilmiş, bu sayede kelime öbeklerinin yerel ve genel ağırlıklarını elde edebilmek mümkün olmuştur. Karcı entropideki α değerinin 1 olduğu durumlarda Shannon entropisi ile aynı davranışı gösterdiği göz önüne alındığında, farklı α değerleri ile daha farklı duruma çözüm ürettiği için metin özetleme kapsamında Shannon entropiden çok yüksek değerler rapor edilmiştir.
- Yaklaşımımız yüksek seviyede uyarlanabilme özelliğine sahiptir. Ayrıca sistem gürültülü ve kusurlu metinlere karşı oldukça duyarsızdır.
- Shannon entropi ve Karcı entropinin metin özetleme kapsamındaki performans sonuçları karşılaştırılmış ve Karcı entropinin daha yüksek değerler rapor ettiği ortaya konulmuştur.

Hibrit Çok Amaçlı Fidan Gelişim Algoritması ile Çıkarıcı Metin Özetleme Yaklaşımı:

- Optimizasyon bazlı doküman özetleme yaklaşımları zaman tüketimi ve performans noktasında mevcut problemlerden ötürü daha fazla formülize edilmeyede ihtiyaç duymaktadırlar. Ayrıca, fazlalık azaltma, kapsam ve çeşitliliği en üst düzeye çıkarmak gibi ana hedefleri bakımından birbirlerinden ayrılırlar. Bu çalışmada kaliteli ve zengin içerikli özetler için birden fazla amaca odaklanılarak uygun ve yeterli formülasyon analizleri gerçekleştirilmiştir. Ek olarak kapsam ve çeşitlilik gibi farklı anlamsal amaçların eşzamanlı optimizasyonları ile ana metin içinden en uygun cümlelerin belirlendiği bir formülasyon kullanılmıştır.
- Özete seçilecek cümleler belirlenirken sadece dokümanın tamamı ile en yüksek kapsam değerine sahip cümlelere yer verilmemiştir. Özette yer alan cümle ve kavram çeşitliliğini genişletmek amacı ile fazlalık parametreside eş zamanlı olarak minimize edilmiştir. Literatürde yer alan pek az metin özetleme çalışmasında gerçek bir şekilde eşzamanlı optimizasyon kullanarak özetleme gerçekleştirilmektedir. Çok amaçlı bir metin özetleme çalışması sunduğunu iddia eden çalışmalar, farklı zamanlarda çalışan tek amaçlı optimizasyon problemidir veya birden fazla optimizasyon amacının matematiksel olarak bir araya getirilmesi ile elde edilen tek amaçlı optimizasyonlardır (Bkz: Çizelge 5.11).
- Ek olarak, çıkarıcı ve genel özetleme için yeni bir metodoloji olan Hibrit çok amaçlı Fidan gelişim algoritması tanıtılmıştır. Bu tez çalışması kapsamında ilk defa Fidan Gelişim Algoritmasının bir adımı olarak tanıtılan budama aşaması kullanılmıştır. Bu yeni aşamada Fidan gövdesi üzerindeki dallar budanarak çözüm kalitesi iyileştirilmektedir. Mevcut dallar (aday özette yer alan cümleler), ikili gruplandırılmış sonrasında cümlelerin orjinal belge içeriğini kapsama miktarına bağlı olarak güçsüz dallar budanmıştır. Önerilen bu aşamada özete ait ikili çözüm vektörüne ($x=(x_1,x_2,...,x_n)$) yeni dallar-cümleler eklenmemektedir.
- Sunulan yaklaşım eşzamanlı bir optimizasyon ile Pareto-optimal olarak adlandırılan en uygun çözümü içeren bir küme elde etmektedir.
- Önerilen her iki modelin performansı bu alanda sıklıkla kullanılan iki veri seti (DUC-2002, DUC-2004) ile doğrulanmıştır. Ayrıca Karcı özetleme çalışması

kapsamında α üs değerinin etkisi 100, 200 ve 400 kelimelik özetler için incelenmiştir.

1.4. Tezin Organizasyonu

Çalışmanın birinci bölümünde bilgi, özetleme gibi kavramlara ek olarak genel bilgilere yer verilmiş, tezin amaç ve kapsamı açıklanmış, gerçekleştirilen çalışmalardan bahsedilerek literatüre olan katkılarına ve tezin çıktılarına yer verilmiştir.

İkinci bölümde, gerçekleştirilen uygulamalar bazında konu ile ilgili literatürde bulunan çalışmalar bir bütünlük halinde sunulmuştur.

Üçüncü bölümde Entropi ve çizge entropi kavramlarından bahsedilerek çalışma çerçevesinde kullanılan entropi türlerine dair bilgiler sunulmuştur.

Dördüncü bölümde Fidan Gelişim Algoritması genel hatları ile özetlenmiştir.

Metin önışleme, temsili çizgeler, kullanılan veri setlerine ek olarak deneysel çalışmaların alt başlıklar halinde detaylıca anlatıldığı bölüm beşinci bölümdür. Bu bölümde çalışmalara ait sonuç ve bulgular çizelge ve grafiklerle izah edilmektedir. Ayrıca elde edilen sonuçlar literatürde bulunan rekabetçi çalışmalar ile karşılaştırılmıştır.

Sonuç ve Öneriler bölümünde, deneysel süreçler ile elde edilen bulgu ve yorumlara yer verilmiştir.

1.5. Tezin Çıktıları

Bu doktora tezi kapsamında gerçekleştirilen makaleler listesi aşağıda sunulmuştur.

- **C. Hark** and A. Karcı, “Karcı summarization: A simple and effective approach for automatic text summarization using Karcı entropy,” *Information Processing and Management*, vol. 57, no. 3, p. 102187, May 2020.
- **C. Hark, et al.** “Metin Özetlemesi için Düğüm Merkezliklerine Dayalı Denetimsiz Bir Yaklaşım”. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi , 8 (3) , 1109-1118 . DOI: 10.17798/bitlisfen.568883

- **C. Hark, et al.**, "Extractive Text Summarization via Graph Entropy," *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, Malatya, Turkey, 2019, pp. 1-5. doi: 10.1109/IDAP.2019.8875936
- **C. Hark, et al.**, "Graph-Based Suggestion For Text Summarization," *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Malatya, Turkey, 2018, pp. 1-6. doi: 10.1109/IDAP.2018.8620738,
- **C. Hark et al.**, "Doğal dil İşleme yaklaşımları ile yapısal olmayan dökümanların benzerliği," *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, Malatya, 2017, pp. 1-6. doi: 10.1109/IDAP.2017.8090306
- **C. Hark et al.**, "15 Temmuz Şehitler Köprüsü Kapsamında Hava Yol Durum Bilgileri Kullanılarak Trafik Kazalarının Nedenlerine Yönelik Birliktelik Çıkarımı," *2016 International Artificial Intelligence and Data Processing Symposium (IDAP)*, Malatya, 2016
- T. Uçkan, **C. Hark**, E. Seyyarer, A. Karcı (2019). Ağırlıklandırılmış Çizgelerde Tf-Idf ve Eigen Ayırışımı Kullanarak Metin Sınıflandırma. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi , 8 (4) , 1349-1362 . DOI: 10.17798/bitlisfen.531221
- T. Uçkan, **C. Hark**, E. Seyyarer, F. Ayata ve A. Karcı, "Çizge Benzerliği Yöntemi ile Doküman Sınıflandırma," *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Malatya, Turkey, 2018, pp. 1-4. doi: 10.1109/IDAP.2018.8620926
- T. Uçkan, **C. Hark**, A. Karcı, " Fp Growth Algoritması Ve Big Data Uygulamaları," *2016 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Malatya, Turkey, 2018, pp. 1-4.

2. İLİŞKİLİ LİTERATÜR

Bu bölümde, metin özetleme alanında yapılan çeşitli çizge, entropi ve optimizasyon tabanlı çalışmalar sunulmuştur.

Metin dokümanlarının özetlenmesi giderek popülerliği artan bir çalışma alanıdır. Günümüzde, otomatik metin özetleme problemine ilişkin olarak farklı tiplerde araştırmalar yürütülmektedir. DDİ kapsamında özetlemeye yönelik olarak güçlü çizge tabanlı yaklaşımlar öne sürülmüştür. [34, 37–41] Metin parçacığının frekans bilgisi, ana metin içerisindeki konumu, TF-TBF vb.. özellikler kullanılarak istatistik tabanlı yaklaşımlar gerçekleştirilmiştir. [42, 43]'de ise duygu ölçme gibi semantik teknikler kullanılmaktadır. [44], otomatik metin özetleme için bağlantı tahmini çalışmalarından ilham almıştır. Metinleri çizgeler ile temsil ettikten sonra, dokümanların yapıları doküman ilişkileri kullanılarak tanımlandı. Özetler, insanlar tarafından oluşturulan özetler ile karşılaştırılarak değerlendirildi. Bir başka makalede [45], anlamsal devamlılığı sağlayabilmek için yine çizge bazlı bir yaklaşım sunuldu. Çalışmada düğümler dokümanlara ait terimleri tutmakta, kenarlar ise bu düğümler arasındaki anlamsal ilişkileri yansıtmaktadır. Temel olarak çizgede yer alan tüm düğümlere ait çizge çapı hesabı gerçekleştirilmekte ve en kısa ve en uzun yollar en zayıf ve en güçlü bağlar olarak nitelendirilmektedir. [46]'da çizge yapıları ile dokümanlar tanımlanırken düğümler ve kenarlar yerel benzerliklerden yola çıkılarak oluşturulmaktadır. Ana dokümanlara ait özetleri elde etmek içinde Rastlantısal yürüyüş kuramı uygulanmaktadır. [47]'de ise biomedikal alan için bir özetleme sistemi önerilmektedir. Unified Medical Language (Birleşik Tıp Dili Sistemi) adı verilen bir sistem kullanılarak yarı sözlük tabanlı bir uygulama ile kavram ve ilişkilerden yola çıkılarak bir çizge elde edilmekte ve akabinde PageRank algoritması kullanılmaktadır. [48]'de Takviyeli rastgele yürüyüş (ing. Reinforced random walk) tabanlı yeni bir çizge önerisinde bulunmaktadır. Mihalcea [49], TextRank adını verdiği yaklaşımında çizge tabanlı cümle skrolama yöntemi ile metinleri özetleme yoluna gitmiştir. Benzer şekilde LexRank yaklaşımı ile Erkan [34] cümlelere puan vererek çıkarıcı özetleme gerçekleştirmiştir. Yine Parveen [50]'de Egraph'ı tanıtmıştır. Modern çizge tabanlı araştırmalar arasında sorgu odaklı çıkarıcı çalışmalarda bulunmaktadır [51]. Bu çalışmada önerilen metot, metinlerde yer alan konuları keşfetmek için bağlı çizgelerin kullanılmasıdır. [52]'de, bir belgedeki

merkezi cümlelerin çıkarılması için terimler ve cümleler arasındaki karşılıklı bilgi kullanılmış ve bu bağlamda bir belge özetleme çerçevesi önerilmiştir.

Osborne [53]'de dokümanlardan cümleleri seçebilmek için maksimum entropi sınıflandırıcısı kullanmıştır. [54]'de bazı entropi ölçümleri kullanılarak bir cümle sıralama tekniği sunulmuştur ancak yaklaşım konuya özel olmuştur kapsayıcılıktan uzak bu yaklaşım önceden belirlenmiş bir cümle kümesine atanarak gerçekleştirilebilmiştir. [55]'de, Ouyang ve Guo veri özetleme modeli sunmuştur ancak yaklaşım gerçek zamanlı veri aktarımı için kullanılmıştır. Metin özetleme yaklaşımı değildir. [56], metinlerden elde edilen özelliklerden yararlanarak gerçekleştirdiği çalışmasında kullandığı özelliklerden bir tanesi entropi bazlı bir özelliktir ve sınırlı düzeyde entropi kullanılmıştır. Kennedy [57], entropi temelli cümle seçim ölçüsü kullanarak benzersiz özet içerik birimlerinin artırılabilirliğini rapor etmiştir. Çalışmasında kütüphane bazlı cümle seçim stratejisi ve Shannon Entropi kullanmıştır. Özetteki belirsizliği ölçmek için entropi hesaplamıştır ve buna dayanarak özetle ne kadar bilgi bulunduğu karar vermiştir. Ancak ilgili çalışmada, verimlilik bakımından gelişim oldukça sınırlı düzeyde kalmıştır.

Çıkarıcı metin özetleme çalışmalarında etkili cümlelerin seçimleri geçmişte istatistiksel araçlar kullanılarak gerçekleştirilmekteydi. Ancak istatistiksel araçlar, cümle seçiminde oldukça zayıf performanslar sergilemektedirler. Doküman özetlemede performansı arttırmak amacı ile son 20 yılda birçok araştırmacı metin özetleme problemini bir optimizasyon problemi olarak değerlendirmektedir [19–21]. Çoklu doküman özetleme kapsamında farklı dokümanların bir araya getirilmesi, cümleler arasındaki ilişkiyi minimumda tutarak, maksimum kapsayıcılık hedefi ile bir özet elde edilmesi oldukça zor bir görevdir. Araştırmacılar, tek ve çok belgeli özetleme için Genetik algoritma, Parçacık sürü optimizasyonu gibi farklı optimizasyon algoritmalarına odaklanmışlardır [58]. Birçok araştırmacı tarafından yüksek kaliteli özetler elde edebilmek için kapsam ve/veya fazlalık gibi nitelikleri göz önünde bulunduran optimizasyon çalışmaları yürütülmektedir [59]. Kapsam, özetlenecek dokümanda bulunan önemli kavram ve konuların içerilmesi anlamına gelirken, fazlalık elde edilecek özetle benzer kavram ve konuların yer almamasını amaçlamaktadır. Sözü edilen amaçlar, doğrultusunda en uygun özeti bulmak NP-zor tipi bir problemdir. Bu nedenle, popülasyon bazlı [23, 60], sürü zekası [58, 61], yapay arı koloni [62, 63] karınca koloni ve guguk kuşu arama algoritması optimizasyon yaklaşımları dahil olmak üzere yakınsama çözümleri elde etmek için çeşitli yöntemler

kullanılmıştır. Tek-amaçlı optimizasyon yaklaşımlarında var olan bütün kriterler tek boyutlu bir fonksiyon elde edebilmek için ağırlıklandırılmaktadır. Ağırlıkların belirlenmesi tamamı ile kişisel bakış açıları ile gerçekleşmektedir. Bu durum optimizasyon sonucunu doğrudan ve oldukça etkilemektedir. Zhao&Tang [64], uzunluk, kapsam, bilgi minimum fazlalık elde etmek için sorgu tabanlı ve genetik algoritma temelli çoklu doküman özetleme yaklaşımı önermiştir. Kogilavani & Balasubramanie [65], Genetik algoritma kullanarak özellik bazlı çoklu, genel bir özetleyici sunmaktadır. Bir kümeleme algoritması ile benzer belgeler gruplanır, belirlenen kriterler ölçüsünce her kümeye ait önemli cümleler belirlenir. En iyi cümleyi seçmek için maksimum uzunluk, kapsam, bilgilendiricilik ve minimum fazlalık hedeflenmiştir. [66]'da Huang ve arkadaşları, bilgi kapsamı, önem, fazlalık ve metin uyum parametrelerini kullanarak çok-amaçlı bir optimizasyon problemi sunmuşlardır. [67]'de Aliguliyev Parçacık Sürü Optimizasyonu algoritmasını kullandığı çalışmada zengin özete ulaşabilmek için metinleri kümelendirmiş ve kümelerin birbirleri ile olan benzerliklerini optimize ederek genel, çoklu-doküman özetleme metot önermiştir. Nandhini & Balasundaram [68] metnin anlaşılması kolay bir özetini elde edebilmek için diferansiyel gelişim algoritmasını kullanmıştır. Mendoza et al.[69] çoklu dokümanlardan çıkarıcı bir özetleme sunmak için sunmuş oldukları bir algoritma ve ikili optimizasyon kullanarak kapsam ve fazlalık değişkenlerini optimize etmişlerdir. [26]'de bir diferansiyel gelişim algoritması ile çözülen boolean programlama problemine odaklanılarak belirlenen üç özellik optimize edilmiştir. Mirshojaei & Masoomi [70], çıkarıcı metin özetleme sisteminin performansını arttırmak için Guguk kuşu arama algoritması kullanmıştır. [37]'de yazarlar kitaplık bazlı genel ve tekli doküman özetleme yaklaşımı sunmuşlardır. Sistem özetle yer alan cümlelerin uyum ve kapsam değerlerini her cümleinin istatistiksel özelliklerini kullanarak elde eder. Umam et al. [71] kapsam, çeşitlilik ve uyum parametrelerini eş zamanlı optimize ederek yeni bir çoklu-doküman özetleme yaklaşımı sunmuşlardır. Saleh at al.[72] önceki çalışmalardan farklı olarak gerçek, çok amaçlı fonksiyon tanımlı gerçekleştirmiştir. Çıkarıcı metin özetleme amacı ile Non-dominated Sorting Genetic Algoritma-II (Baskın Olmayan Sıralı Genetik Algoritma) (NSGA-II) adını verdiği bir optimizasyon modeli sunmuştur. Sanchez-Gomez et al. [62] Çok-amaçlı optimizasyon yaklaşımlarının eksikliğinden bahsetmiştir ve çalışmada Çıkarıcı çoklu-doküman özetleme yaklaşımı sunmuştur. John et al. [23] popülasyon bazlı optimizasyon yaklaşımında genel, çıkarıcı bir

özetleme sunmaktadır. Belirlediği üç amaç fonksiyonuna göre ayrı ayrı optimizasyon gerçekleştirmektedir. Sonuçta elde ettiği üç ayrı aday özetle bulunan ortak cümlelere özetle yer vermektedir.

Bu bölümde yer verilen optimizasyon temelli metin özetleme yaklaşımlarından sadece Saleh at al.[72] ve John et al. [23], eşzamanlı bir biçimde parametreleri optimize ederek özetleme gerçekleştirmekte, birden fazla optimizasyon amacının matematiksel olarak bir araya getirerek tek amaçlı optimizasyona dönüştürmemektedir.



3. ÇİZGE ENTROPİSİ

Bilgi teorisinin tarihsel süreci ile birlikte entropiye dair birçok tanımlama yapılmıştır. Kaos, bilgi miktarı, enerjinin yayılımı, düzensizliğin ölçümü bu tanımlamalardan sadece bir kaçıdır. Entropi farklı bilim dallarına ait çeşitli tanımlamalara sahip olabilmektedir. Fiziğin temelini oluşturan entropi, bilgi teorisi ile birlikte mantık, istatistik, biyoloji gibi birçok farklı alanda çeşitli roller oynamaktadır. Termodinamiğe göre kapalı bir yapı içerisinde enerjinin aktarılma biçimi entropi ile ifade edilebilir. Bu durum dışarıdan bir müdahale olmaksızın, mevcut enerji ile sistemin sürekliliğinin söz konusu olamayacağı anlamındadır. Entropi kapalı sistemlerdeki düzensizliğe eğilimin ölçütü olarak tanımlanabilir zira kapalı bir sistemde entropi devamlı olarak artma eğilimi taşımaktadır. Kapalı bir sistemde bulunan enerji miktarındaki azalış, harici bir enerji girişiyle desteklenmediği sürece düzensizliğe, ölüme varma eğilimi taşımaktadır. Bu durumun sebeplerini anlayabilmek, ortadan kaldıracılabilmek amacı ile entropinin niçin devamlı olarak artış gösterdiğini anlamak oldukça önemlidir. Bir sistem hakkında çıkarımlar yapabilmenin yolu belirsizlikleri ortaya koyabilmektir. Düzensizlik belirsizliğe yol açmaktadır. Dolayısı ile düşük ihtimalli durumlar çok daha yoğun bilgi içerirler. Bilgi ve belirsizlik ters orantılı kavramlardır. Durumlara ait ihtimaller ortaya konabilirse, durumların niceliklerine ait olan entropi değerleri belirlenebilmektedir. Entropi bilgi niceliğini ifade etmek veya sisteme ait hata ve eksikliklerin ifadesi amacı ile kullanılan olasılık fonksiyonudur. Bir sistemin düzensizliği ile kast edilen durum, sisteme ait olan toplam enerji veya bilginin, sistemin bütünü oluşturan bütün üyeler arasındaki dağılımına ait olan durumdur. Sisteme dair düzensizliği ifade eden bu tanım, tez çalışmasına dair temel dayanak noktasını oluşturmuştur. Karcı Özetleme kapsamında, paragraf veya metin parçacıkları içerisinde en etkili cümleleri seçmek için Karcı Entropi bazlı bir çözüm önerilmektedir. Dahası Karcı Entropi’de yer alan farklı α üs değerleri ile kelime öbeklerinin local ve global ağırlıklarını elde edebilmek mümkün olabilmektedir [18, 73–81].

3.1. Entropi Türleri

Geçmişten günümüze araştırmacılar tarafından farklı varsayımlar ile değişen amaçlar için birçok entropi tanımlanmıştır. Bu tez çalışmamız kapsamında kullanılan entropi türlerine dair detaylara aşağıda yer verilmiştir.

3.1.1. Shannon Entropi

Shannon Entropi bilgi teorisinde önemli bir yere sahiptir. Birçok çalışmada belirsizliğin bir ölçüsü olarak isimlendirilmekte ve kullanılmaktadır. Sık kullanılan bir belirsizlik ölçüsüdür. Mesaj iletiminde, bilgi içeriğinin taşıdığı belirsizliğin toplam miktarını ölçmek için önerilmiş ve kullanılmıştır. Shannon entropi bilgisayar bilimlerinden sosyal bilimlere kadar çeşitli alanlarda sıklıkla kullanılmaktadır [82–84].

Shannon entropinin yöntemine ait adımlar;

$$[p_{ij}]_{m \times n} = \left[p_{ij} / \sum_{i=1}^m x_{ij} \right]_{m \times n} \quad i = 1, \dots, m \quad j = 1, \dots, n \quad (3.1)$$

Denklemden yer alan m alternatif sayısı, n değerlendirme kriteri sayısıdır. Birinci adımda, bağımsız ölçüm birimlerinden kaynaklanan düzensizlikler giderilerek normalizasyon sağlanmaktadır. İkinci adım entropi değerlerinin hesaplandığı adımdır.

$$e_j = -k \sum_{i=1}^m p_{ij} \cdot \ln p_{ij} \quad (j = 1, \dots, n) \quad (3.2)$$

Burada k Boltzman sabitidir. Kriterlere ait e_j değerleri entropi değerleridir ve kriterin karar problemi açısından önemine dair bilgi içermektedir (1'e yaklaştıkça önemi azalır). Üçüncü aşama kriterlere ait farklılık dereceleri olarak tanımlanan d_j değerlerinin hesaplandığı aşamadır. d_j , j kriterine ait ayırım gücüdür (yüksek d_j , değeri j kriterinin önemini yansıtır).

$$d_j = 1 - e^j \quad (3.3)$$

Dördüncü ve son aşamada kriterlere ait önem dereceleri hesaplanmaktadır.

$$w_j = \frac{d_j}{\sum_{j=1}^n d_j} \sum_{j=1}^n w_j = 1 \quad (j=1, \dots, n) \quad (3.4)$$

3.1.2. Karıcı Entropi

X 'in rastgele bir değişken olduğu varsayalım. İlgili olasılık fonksiyonunun ise $p(x = x_i)$ veya $p(a \leq x \leq b)$ olduğu varsayalım. X , ayrık, rastgele bir değişken ise, olasılık/kütle fonksiyonu, karşılık gelen noktanın mevcut olasılığıdır. X sürekli ve rastgele bir değerse, fonksiyonunun nokta değeri olasılık değildir. Bir aralıktaki olasılık yoğunluk fonksiyonunun söz konusu değerlerinin toplamı olasılığın değeridir.

$$p(a \leq x \leq b) = \int_a^b f(x) dx \quad (3.5)$$

$[a, b]$ aralığı $I_1 \cup I_2 \cup \dots \cup I_n = [a, a_1] \cup [a_1, a_2] \cup \dots [a_{n-2}, a_{n-1}] [a_{n-1}, b]$ gibi birçok aralığa bölünebilir. $1 \leq i \leq n$ ise bütün aralıklara karşılık gelen olasılıklar I_i için P_i dir. Shannon entropi;

$$H(p) = - \sum_{i=1}^n p_i \log p_i \quad (3.6)$$

Birçok araştırmacı türev adımlarının spesifik bir adıma kadar tamsayı olduğunu ve bu özel noktaya ulaşırken türev adımlarını gerçek sayı olarak varsaymıştır. Karıcı yeni bir kesir dereceli türev tanımlamıştır ve özelliklerini belirlemiştir [14, 17, 85, 86]. Karıcı tarafından kesir dereceli entropi elde etmek amacı ile belirlenen adımlar aşağıdaki gibidir.

$f(x): R \rightarrow R$, $\alpha \in R$ dir. Kesir dereceli türev;

$$f^{(\alpha)}(x) = \lim_{h \rightarrow 0} \frac{f^\alpha(x+h) - f^\alpha(x)}{(x+h)^\alpha - x^\alpha} \quad (3.7)$$

L'Hospital metodu uygulanan denklem aşağıdaki gibidir.

$$f^{(\alpha)}(x) = \lim_{h \rightarrow 0} L \left(\frac{f^\alpha(x+h) - f^\alpha(x)}{(x+h)^\alpha - x^\alpha} \right) \quad (3.8)$$

$$= \lim_{h \rightarrow 0} \frac{\frac{d(f^\alpha(x+h) - f^\alpha(x))}{dh}}{\frac{d((x+h)^\alpha - x^\alpha)}{dh}} = \left(\frac{f(x)}{x}\right)^{\alpha-1} f'(x)$$

Kesir dereceli entropi elde etmek için kesir dereceli türev aşağıdaki fonksiyona uygulanır.

$$f(t) = \sum_i p_i^{-t} \quad (3.9)$$

Son olarak aşağıdaki entropi formülü elde edilir.

$$E = \left| \frac{d^{(\alpha)} f(t)}{dt} \right| = \left| \frac{d^{(\alpha)}}{dt} \sum_{i=1}^n p_i^{-t} \right| = \sum_{i=1}^n \left| \left(\frac{p_i^{-t}}{t} \right)^{\alpha-1} (-1) p_i^{-1} \ln p_i \right| \quad (3.10)$$

$$E = \sum_{i=1}^n |(-p_i)^{\alpha-1} (-p_i) \ln p_i| = \sum_{i=1}^n |(-p_i)^\alpha \ln p_i| \quad (3.11)$$

Türev alındıktan sonra aşağıdaki denklem ile Shannon entropisi elde edilebilir [87].

$$H(p) = \lim_{t \rightarrow -1} \frac{d}{dt} \sum_i p_i^{-t} \quad (3.12)$$

Tsallis entropi'de benzer şekilde tanımlanabilir. D_q^t Jackson[75] türevi olarak kabul edilir.

$$H(p) = \lim_{t \rightarrow -1} D_q^t \sum_i p_i^{-t} \quad (3.13)$$

p_i , i'nci olayın meydana gelme ihtimalidir. Entropi değeri aşağıda belirtildiği gibi hesaplanır. Karcı Entropi;

$$D_t^{(\alpha)} \sum_i |p_i^{-t}| = \sum_i |(-p_i)^\alpha \log p_i| \quad (3.14)$$

Karcı çalışmasında bu yöntemi Standart Normal dağılım, Binom dağılım, Poisson dağılım ve Weibull dağılımına uyguladı. Elde edilen sonuçlar, (0,1) aralığında önerilen entropinin Shannon entropiden daha iyi olduğunu gösterdi [79, 80]. Daha önce hiçbir doküman özetleme çalışmasında Karcı Entropi kullanılmamıştır.

3.2. Çizge Entropi ile Merkezilik

Entropi bir sistemin belirsizliğini (düzensizliğini) ölçer. Çizgede entropi ise, çizgeye ait yapısal kararlılık anlamındadır. Yaygın olarak kullanılan bir graf entropi tanımı yoktur. İlişkisel yapıların entropilerini araştırmak için ayrık matematik, bilgisayar bilimleri, bilgi teorisi, istatistik, kimya gibi çok çeşitli alanlara ait problemlerde entropi kullanılmaktadır. Ancak bilgi teorisine göre, graf entropi, belirli değer çiftlerinin karıştırılabileceği bir bağlantı üzerinden işaretler ileterek elde edilebilecek bilgi oranının bir ölçüsüdür. İlk olarak 1970'lerde Körner tarafından tanıtılan bu ölçüm, o zamandan beri kullanılmaktadır [88]. Entropi çizgenin yapısal kararlılığını ve yapısal bilgi içeriğini yorumlamak için kullanılmaktadır. Entropi ölçümler yapısal bilgi içeriğinin miktarını araştırır.

N adet düğüme sahip ağırlıklı kenarları olan bir çizgedeki kenarların olasılıkları p_{ij} dir.

$$P_{ij} = \frac{w_{ij}}{\sum_{j \in E} w_{ij}} \quad (3.15)$$

Buradaki W_{ij} , i ve j düğümleri arasındaki ilişkiyi tutan kenarların ağırlığıdır. i düğümünün entropi değeri $E(V_i)$ dir.

$$E(V_i) = - \sum_{j=1}^n p_{ij} \log p_{ij}, \quad 0 \leq p_{ij} \leq 1 \quad (3.16)$$

$$E(V_i) = \sum_{j=1}^n |-p_{ij}^\alpha \log p_{ij}| \quad 0 < \alpha \quad (3.17)$$

$E(G)$ Tüm düğümlerin entropi değerleri toplamıdır ve aşağıdaki gibi hesaplanmaktadır.

$$E(G) = \sum_{i=1}^n E(V_i) \quad (3.18)$$

İlgili çizgelere Karcı Entropi uygulandı. Her bir düğüm ağdan sırayla çıkarılarak toplam Karcı Entropi hesaplandı. Çalışma kapsamında önerilen Karcı Özetleme süreci gereği değerler küçükten büyüğe sıralandı ve metinsel çizgeyi oluşturan düğümlere dair taşıdığı bilgi miktarı nicel olarak elde edildi. Böylece düğümün global çizge üzerindeki etkisi ölçüldü. Karcı Entropi hesaplanırken ilgili denklemde yer alan α üs değerinin etkisi çizge üzerinde yerel ve global etkiyi ölçecek şekilde gerçekleşmiştir. Küçük α değerleri yerel anlamda düğümlerin etkisini ortaya koyarken büyük α değerleri ise global ölçüde düğümleri incelemektedir[1, 12, 89].



4. FİDAN GELİŞİM ALGORİTMASI

Doğa, etkili fikirler için bir ilham kaynağı olabilmektedir. Doğal süreçlerin ilham verdiği birçok hesaplama modeli geliştirilmiştir. Karcı [90] , Fidan gelişim algoritmasında ağaçların ekilmesi ve büyümesi gibi bir doğal olaydan ilham almıştır. Fidan gelişim algoritması bir hesaplama modelidir. Bu algoritma ekim ve büyüme aşaması gibi iki ana aşama içermektedir. Fidan gelişim algoritması temel olarak fidanlardan oluşan aday çözümler arasında iteratif üretim ve değişim süreçleri olarak da tanımlanabilmektedir. Genel olarak Fidan gelişim algoritması arama uzayını seçimler ile yeni aday çözümler üreten değişim işlemlerinden oluşmaktadır. Ana değişim işlemleri eşleşme, dallanma ve aşılama işlemleridir. Ayrıca Fidan gelişim algoritması, bir optimizasyon probleminin amaç fonksiyonunda diferansiyel bilgiye ihtiyaç duymamaktadır. Bu özelliği ile Fidan gelişim algoritması çeşitli optimizasyon problemlerine kolaylıkla uygulanabilme özelliği taşımaktadır. Yöntemin ilk adımı fidanların ekimi aşamasıdır. Bu aşamada fidanlar kuzey, güney, doğu ve batı yönlerinden birbirlerine eşit mesafede olmalıdırlar. Sonrasında fidanlar eşleşir, dallanır ve aşılır [91, 92].

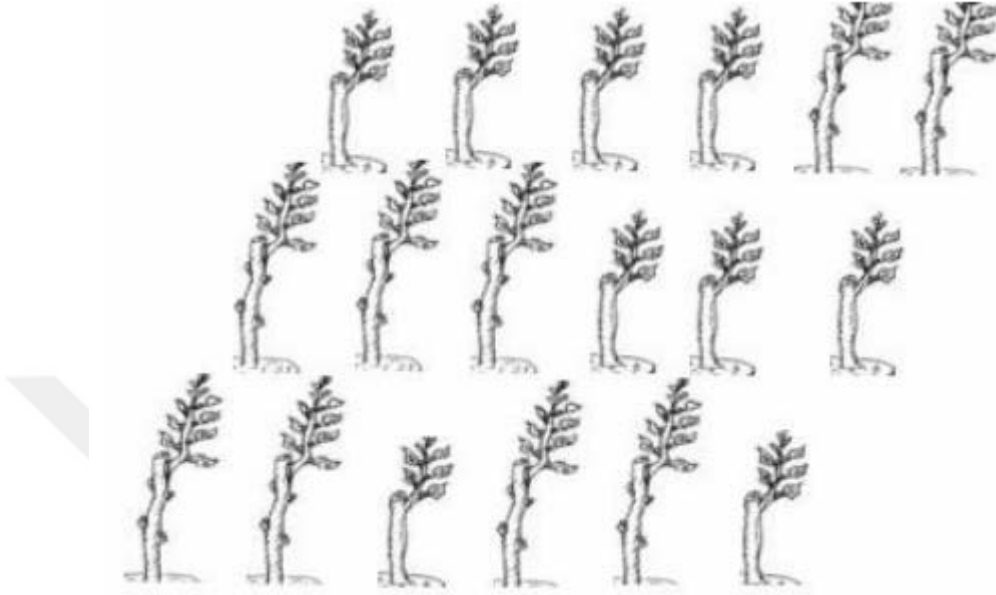
Fidan gelişim algoritması iki ana aşamadan oluşur. İlgili aşamalar sırasıyla Bölüm 4.1 ve Bölüm 4.2 altında detaylı olarak incelenmiştir.

4.1. Fidanların Ekim Aşaması

Çözüm uzayı bir fidan bahçesi olarak düşünülmekte ve bu nedenle bütün fidanlar bahçeye düzenli bir biçimde aktarılmaktadır. Bir çiftçi hızlı bir biçimde fidanları büyötmek için, fidanları eşit uzaklıkta ekmesi gerekmektedir. Fidanların büyümesini taklit ederek problemi çözmek için, üretilen başlangıç çözümleri arama uzayına düzenli bir şekilde dağıtılmalıdır. Her fidan dallardan oluşur ve başlangıçta fidanlar dal içermez sadece gözdeden oluşurlar. Bahçedeki başlangıç fidanları (arama uzayındaki başlangıç çözümleri) tek tiptirler ve düzenli bir forma sahiptirler.

Z çözüm uzayı olsun ve bu çözümler S eleman içersin. O halde $Z=\{z_0,z_1,...,z_{s-1}\}$ gibi ifade edilebilir. Çözüm uzayında yer alan çözümler ikili tabanda ifade ediliyor ise ve her fidanın iki adet dalı mevcut ise çözüm uzayı $Z=\{00,01,10,11\}$ şeklindedir. Her fidanda n adet dal varsa çözüm uzayının boyutu 2^n olarak hesaplanmaktadır. Bu durumda Z uzayının bazı $\{01,10\}$ dır. Diğer durumlar ise bu değerlerin lineer kombinasyonlarından elde edilmektedir. Fidan Gelişim Algoritmasının önemli bir

avantajı düzenli popülasyon sağlamasıdır. Bu durum mümkün olan çözüm aralığı oluşturulurken en küçük ve en büyük değerler ve bunlardan oluşan bir popülasyon oluşturması, bu sayede çözüm aralığını en iyi şekilde dağıtması ile açıklanabilir (Şekil 4.1).



Şekil 4.1 . Düzgün dağılım ile fidanların ekimi

Bu sayede en iyi çözüme yakınsamak daha az maliyet ile gerçekleşir ve iyi bir başarı sağlanır. Ayrıca Fidan gelişim algoritmasının bu başlangıç popülasyonu oluşturma yöntemi ile yerel minimum veya maksimumlara takılma durumu önlenir. Maksimum ve minimum değerler fidan gelişim algoritmasından tanımlanan alt sınır ve üst sınır değerleridir.

Başlangıçta S_0 ve S_1 iki fidan olsun. $S_0 = \{u_1, u_2, \dots, u_n\}$, $S_1 = \{l_1, l_2, \dots, l_n\}$

Burada n fidan boyutudur. $k=1$, $1 \leq i \leq n$ olması durumunda

u_i , karşılık gelen değişken için üst sınır değeridir.

l_i , karşılık gelen değişken için alt sınır değeridir.

Bölme faktörü $k=2$ için, eğer mümkünse S_0 ve S_1 iki eşit parçaya bölünerek devam edilir. Bu durumda dört yeni fidan türetilir ancak bu fidanların biri S_0 ile diğeri S_1 ile aynıdır. Kalanlar ile S_3 ve S_4 adında iki fidandır.

r rastgele bir sayıdır ($0 \leq r < 1$). k değeri artırılır ve bahçe dolana dek aynı yöntem uygulanmaya devam edilir.

Fidanların değerleri (çözümler) ikili sayılar olabilir.

$$S_1 = \{1, 1, \dots, 1\}$$

$$S_2 = \{0,0, \dots, 0\}$$

Bu durumda Karcı [90] sayısal değerler taşıyan fidanların üretilmesinden daha farklı bir yol önermekte S_3 ve S_4 fidanlarını mevcut fidanları bölmeleyerek elde etmektedir.

$$S_3 = \{1,1, \dots, 1, 0,0, \dots, 0\} \text{ ve } S_4 = \{0,0, \dots, 0, 1,1, \dots, 1\}$$

Bölme faktörü (k) arttırılır ve süreç bahçe dolana dek sürdürülür.

4.2. Fidanların Büyüme Aşaması

Bu aşama eşleşme dallanma ve aşılamanın birleşiminden oluşur. Şayet bahçe (başlangıç çözümü) G adet fidan içeriyorsa, yavru bahçeden ve mevcut bahçeden en iyi G adet fidan seçilerek devam edilmektedir. Bu nedenle jenerasyonların seçiminde stokastik bir süreç söz konusu değildir. Eşleşme, dallanma ve aşılama aşmaları ile en geniş ve zengin çözüm uzayına ulaşılır.

4.2.1. Eşleşme

Eşleşme operatörünün amacı mevcut bilgileri değiştirerek var olan fidanlardan yeni fidanlar türetmektir. Her fidan çifti için bir eşleşme faktörü vardır. Çiftler arasındaki mesafe fidan çiftlerinin eşleşme veya eşleşmemeleri noktasında belirleyici olmaktadır. S_1 ve S_2 fidanlar olsun,

$$S_1 = \{S_{1,1}, S_{1,2}, \dots, S_{1,i}, \dots, S_{1,n}\} \text{ ve } S_2 = \{S_{2,1}, S_{2,2}, \dots, S_{2,i}, \dots, S_{2,n}\}$$

S_1 ve S_2 fidanları arasındaki mesafe eşleşme sürecinin gerçekleşip gerçekleşmeyeceğini belirlemektedir. Bu durum tamamen çiftler arasındaki mesafeye bağlıdır. $P_m = (S_1, S_2)$ S_1 ve S_2 fidanlarının çiftleşme olasılığıdır ve aşağıdaki gibi hesaplanmaktadır.

$$P_m(S_1, S_2) = 1 - \frac{\left(\sum_{i=1}^n (S_{1,i}, S_{2,i})^2\right)^{\frac{1}{2}}}{R} \quad (4.1)$$

$$R = \left(\sum_{i=1}^n (u_i, l_i)^2\right)^{\frac{1}{2}} \quad (4.2)$$

u_i ve l_i , sırasıyla seçili olan fidan çiftleri arasındaki mesafe üst ve alt sınırlarıdır. Fidanların eşleşme olasılıkları çiftler arasındaki mesafeye bağlıdır. Eşleşme aşamasına ait adımlar aşağıdaki gibidir.

Eşleşme (S₁,S₂)

1 $j \leftarrow 1, \dots, n$

2 hesapla $R = (\sum_{i=1}^n (u_i, l_i)^2)^{\frac{1}{2}}$ ve $P_m(S_1, S_2) = 1 - \frac{(\sum_{i=1}^n (S_{1,i}, S_{2,i})^2)^{\frac{1}{2}}}{R}$

3 $i \leftarrow 1, \dots, n$

4 **if** $P_m(S_1, S_2) \geq \text{rastgele}[0,1)$ **then**

5 $S_1 \leftarrow S_1 - S_{1,i}$

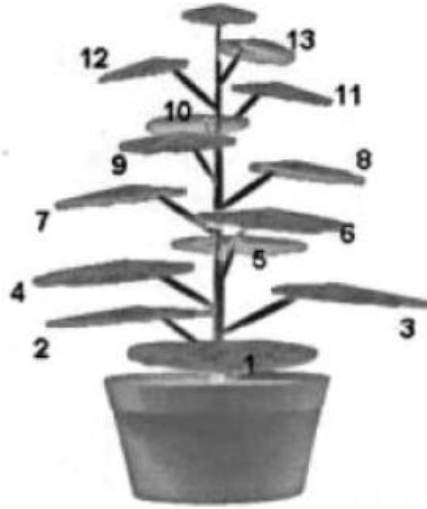
6 $S_2 \leftarrow S_2 - S_{2,i}$

7 $S_1 \leftarrow S_1 + S_{2,i}$

8 $S_2 \leftarrow S_2 + S_{1,i}$

4.2.2. Dallanma

Fidan gövdesi üzerinde bir dalın meydana gelebilmesi için daha önce aynı noktada bir dalın oluşmaması gerekmektedir. Şekil 4.2'de görüldüğü gibi 1 noktasında ilk dalın oluştuğunu varsayalım, 2 noktasında meydana gelecek olan dallanma olasılığı, 3 noktasında meydana gelecek dallanma olasılığından daha azdır. Bu mantık yerel çözümleri aramak için kullanılmaktadır. Bu adım mevcut çözüm üzerinde yerel değişikliklere neden olur.



Şekil 4.2 . Örnek bir fidan üzerinde dallanma noktaları S_1 fidan olsun.

$$S_1 = \{S_{1,1}, S_{1,2}, \dots, S_{1,i}, \dots, S_{1,n}\}$$

Eğer $S_{1,i}$ noktasında bir dal meydana geldi ise $S_{1,j}$ noktasında dal meydana gelme ihtimali iki şekilde hesaplanmaktadır. $S_{1,i}$ ve $S_{1,j}$ arasındaki mesafe, $|j-i|$ veya $|i-j|$ ile hesaplanabilmektedir. Eğer g_i bir dal ise $S_{1,j}$ noktasında bir dal meydana gelme olasılığı doğrusal durumda aşağıdaki gibi hesaplanmaktadır.

$$P(S_{1,j}|S_{1,i}) = 1 - \frac{1}{(|j-i|)^2} \quad (4.3)$$

Doğrusal olmayan durumda ise dal oluşma ihtimali aşağıdaki gibi hesaplanmaktadır.

$$P(S_{1,j}|S_{1,i}) = 1 - \frac{1}{e^{(|j-i|)^2}} \quad (4.4)$$

Aynı noktada iki dal meydana gelemeyeceği için, $i=j$ olması durumunda $P(S_{1,j}|S_{1,i}) = 0$ dır. Dallanma aşamasına ait adımlar aşağıdaki gibidir.

Dallanma (S_1, S_2)	
1	$i \leftarrow 1, \dots, n$
2	$j \leftarrow 1, \dots, n$
3	eğer dal yoksa
4	$P(S_{1,j}, S_{2,i})=1$ ve dallanma işlemi uygulanır
5	değilse
6	$P(S_{1,j} S_{1,i}) = 1 - \frac{1}{(j-i)^2}$ veya $P(S_{1,j} S_{1,i}) = 1 - \frac{1}{e^{(j-i)^2}}$
7	Eğer $P(S_{1,j} S_{2,i}) \geq \text{rastgele}[0,1]$
8	$S_{1,j}$ de bir dal meydana gelecek

4.2.3. Aşılama

Aşılama işlemi fidanların benzerlikleri durumunda iki farklı fidan arasında gerçekleşir. Fidanların benzerlikleri aşılamanın başarısını etkiler. Fidanların benzerlikleri ise iki şekilde hesaplanır:

S_1 ve S_2 fidanlar olsun,

$$S_1 = \{S_{1,1}, S_{1,2}, \dots, S_{1,i}, \dots, S_{1,n}\} \text{ ve } S_2 = \{S_{2,1}, S_{2,2}, \dots, S_{2,i}, \dots, S_{2,n}\}$$

$$\text{sim}(S_1, S_2) = \sum_{i=1}^n S_{1,i} \oplus S_{2,i} \quad (4.5)$$

Aşılama işlemi aşağıdaki gibi gerçekleştirilir. Eğer $\text{sim}(S_1, S_2) \geq \text{eşik değeri}$

$$S'_1 = \begin{cases} S_{1,i}, & \text{eğer } S_{1,i} = S_{2,i} \\ \text{rastgele}(1), & \text{eğer } S_{1,i} \neq S_{2,i} \end{cases}$$

$$S'_2 = \begin{cases} S_{2,i}, & \text{eğer } S_{2,i} = S_{1,i} \\ \text{rastgele}(1), & \text{eğer } S_{2,i} \neq S_{1,i} \end{cases}$$

Burada S'_1 ve S'_2 aşılama süreci sonrasında elde edilen değerlerdir. Fidanlar keyfi olarak aşılınmazlar. Aşılacak fidanlar benzerlikler üzerinde tanımlanacak eşitsizliği ($\text{sim}(S_1, S_2) \geq \text{eşik değeri}$) sağlamalıdır. Kullanılan eşik değeri problem çözücü tarafından belirlenir.

$$\text{sim}(S_1, S_2) = \sum_{i=1}^n \frac{|S_{1,i} - S_{2,i}|}{U_i - I_i} \quad (4.6)$$

Eğer $\text{sim}(S_1, S_2) \geq n * \epsilon$ ise S_1 ve S_2 aşağıdaki gibi aşılır. ϵ , kullanıcı tarafından tanımlanan bir sabittir ($0 < \epsilon < 1$).

Aşılama (S_1, S_2)	
1	$i \leftarrow 1, \dots, n$
2	$\text{sim}(S_1, S_2) = \sum_{i=1}^n S_{1,i} \oplus S_{2,i}$
3	Eğer $\text{sim}(S_1, S_2) \geq \text{rastgele}$
4	$S'_1 = \begin{cases} S_{1,i}, & \text{eğer } S_{1,i} = S_{2,i} \\ \text{rastgele}(1), & \text{eğer } S_{1,i} \neq S_{2,i} \end{cases}$
5	$S'_2 = \begin{cases} S_{2,i}, & \text{eğer } S_{2,i} = S_{1,i} \\ \text{rastgele}(1), & \text{eğer } S_{2,i} \neq S_{1,i} \end{cases}$

4.3. Algoritma

Fidan Gelişim Algoritmasına ait kabakodda görüldüğü gibi operatörler sırasıyla uygulanır. Elde edilen fidanlar geçici çözümler olarak kabul edilir. Sonraki nesiller geçici çözümlerden ve mevcut çözümlerden en iyi fidanlar seçilerek elde edilir.

Fidan Gelişim Algoritmasına ait kabakod.	
1.	$t \leftarrow 0$ // Başlangıç anı
2	$\text{fidanların_ekimi}(G(t))$ // Başlangıç popülasyonu
3	$\text{amaç_fonksiyonu}(G(t))$

4	While durdurma kriteri do
5	$G_1(t) \leftarrow \text{eşleşme_aşaması}(G(t))$
6	$G_2(t) \leftarrow \text{dallanma_aşaması}(G(t))$
7	$G_3(t) \leftarrow \text{aşılama_aşaması}(G(t))$
8	amaç_fonksiyonu ($(G_1(t) \cup G_2(t) \cup G_3(t))$)
9	$G(t+1) \leftarrow \text{seçim}(G_1(t) \cup G_2(t) \cup G_3(t) \cup G(t))$
10	$t \leftarrow t+1$

Başlangıç çözümü oluşturulurken, en küçük ve en büyük değerler ve bunlardan oluşan popülasyon oluşturulur. Bu sayede çözüm aralığı en iyi şekilde dağıtılır.

Dallanma aşaması, var olan dalların pozisyonlarına dayanarak, bir olasılık metodu kullanır ve yeni dallar üretir. O yerel bir arama operatörüdür.

Eşleşme işlemi, var olan fidanları eşleştirir ve yeni fidanlar üretir. İki fidandaki bilgiler değiştirilerek global bir arama gerçekleştirilir.

Aşılama işlemi, var olan benzer fidanlardan yeni ekimler gerçekleştirerek yeni fidanlar üretir.

5. YÖNTEM VE UYGULAMALAR

5.1. Metin Önişleme ve Hazırlık

Yoğun veri kümelerinin büyük kısmı yapılandırılmamıştır; ilgili veri kümelerinin pek azı yapılandırılmış formatlara sahip veri kümeleridir. Yapılandırılmış veri kümeleri çizelgeler halinde satır ve sütunlarla ifade edilebilen veya bir takım etiketlere sahip veri kümeleridir. Bu çalışmada üzerinde çalışılacak veriye belirli bir yapısalılık kazandırılması aşamaları oldukça karmaşık ve zaman alıcı süreçlerin geride bırakılması ile mümkün olmuştur. Belirli bir bütünlük taşımayan, buna karşın bir yapısalılık kazandırılması gereken veri kümeleri için bazı önişlemler zorunludur. Sisteme girdi olarak kullanılacak metinlerin üzerinde çalışılabilecek biçimlere aktarılmasının ve ayırt ediciliği olmayan gereksiz verilerden soyutlanmış olmasının sistemin başarımını arttıracakları öngörülmektedir zira bahsedilen yoğun kümeler doğal dil kullanılarak oluşturulmuşlardır. Metin özetleme süreçlerinde, üzerinde çalışılacak veriler herhangi bir önişlem sürecine katılmadan kullanılmamaktadır. Durak kelimeler (zamirler, edatlar, bağlaçlar, vb..) özetleme öncesinde veri setinden uzaklaştırılması gereken ve ayırt edici özelliği olmayan ifadelerdir. Bu nedenden ötürü verileri, üzerinde çalışılabilecek formatlara aktararak metin özetleme işleminin gerçekleştirilebilmesi için cümleler üzerinde temsil gücü olmayan durak kelimeler orijinal veri kümelerinden çıkarılmaktadır. Böylece sürece ait işlem yükü azaltılmakta ve kategorik anlam taşıyan kelimelerle devam edilmesi sağlanmaktadır [1, 4, 11, 24, 36, 93]. Çalışma kapsamında kullanılan DUC veri setinde çizgelerle ifade edilecek olan metinler .txt uzantısına sahip dosyalarda tutulmaktadır.

Sözü edilen normalizasyon adımları veri önişleme ve hazırlık aşamalarının ilk kısmını oluşturmaktadır. Buna ilaveten, yazım şekilleri bakımından birbirlerinden farklı olan ancak anlamsal olarak ortak bir kelime kökünden türeyen yakın anlamlı kelimelerin çizgeler oluşturulurken farklı kelimeler gibi değerlendirilmesi, cümleler arasındaki bağlantı ve ilişkilerin tespitini zorlaştırmaktadır. Sezgisel olarak sözü edilen problemin özetleme başarımı ciddi bir şekilde etkileyeceği öngörülerek Karıcı Özetleme modeli kapsamında metinsel çizge oluşturulması aşamasının öncesinde kullanılan bir metin işleme aracı geliştirilmiştir. Bu yazılım aracı .NET platformunda C# kullanılarak geliştirilmiştir. Önerilen metin dönüşümlerine ait adımlar aşağıdaki gibidir.

Gerçekleşen Dönüşümlere ait algoritmik adımlar

1	“text read”;	
2	sentences=[], words=[], alternative =[];	
3	best_alternative=0, output_text={}	
4	sentences []=text.Split(‘.’);	<i>Giriş metinlerinden cümle vektörlerini oluştur.</i>
5	for c to len(sentences) do	
6	sentences [c]=clear(c);	<i>Cümlelerin normalize aşamaları</i>
7	end for	
8	words[]=sentences.Split(‘ ‘);	<i>Cümle vektörlerinden kelime vektörlerine indirgeme.</i>
9	for k to len(words) do	<i>Alternatif kelime listesindeki en uygun kelime seçilir (n-grama göre).</i>
10	alternative[k]=alternative_search(words[k],words);	<i>Kelime vektörlerinde bulunan en uygun alternatifle değiştir.</i>
11	for u to alternative do	
12	best_alternative=best_alternative_search (k, alternative);	
13	end for	
14	words [k]=best_alternative;	<i>Kelime vektör boyutu kadar tekrar.</i>
15	end for	
16	for k to words do	
17	output_text+=k+” ”;	<i>Çıktı metnini oluştur</i>
18	end for	
19	return output_text;	

Çizelge 5.1, DUC-2002 Veri Seti’nde bulunan “d065” isimli örnek metin dokümanının önerilen kelime dönüşümlerinden sonra gerçekleştirilen cümle dönüşümlerini içermektedir. İlgili çizelge durak kelimeler çıkarıldıktan sonra elde edilen metinler üzerinde önerilen yazılım aracının işlevinin detaylandırılması amacı ile gerçekleştirilen dönüşüme dair kelime ve cümle bazında bazı örnekleri içermektedir. Yaklaşım, algoritması gereği dinamik bir formda çalışmaktadır. Değişikliğe gidilecek kelimeler her özetleme işleminde yeniden belirlenmektedir. Sadece özetlenecek metinlerin ortak olarak içerdikleri kelimeler üzerinde sadeleştirmeler yapılmaktadır. Metinler üzerinde herhangi bir kesişime sahip olmayan kelimeler için bir sadeleştirmeye gidilmemektedir. Bu dinamik çalışma prensibi ile her defasında özetlenecek metinlere bağlı olarak sadeleştirmeye gidilecek metin parçalarını yeniden belirlemektedir. Bu sayede metinler değişikçe farklı kelimeler sadeleştirilmektedir. Çizelge 5.1’de değişikliğe uğrayan veya silinen kelimeler kalın yazı tipi ile vurgulanmıştır. Program tarafından “the, and, from, until, as, his, he, of,

in” gibi kelimeler durak kelimeler oldukları için tamamen silinmiş iken yine aynı cümlede “official, serve, quayle ” kelimelerinin yalnızca taşıdığı yapım eki silinmiştir. Bazı kelimeler üzerinde hiçbir işlemin yapılmadığı görülmektedir. Yazılım aracının herhangi bir değişikliğe gitmediği kelimeler metinde yalnızca bir adet bulunan kelimelerdir. Bu kelimelerin metin içinde yer alan daha yalın bir biçimi mevcut değildir. Dolayısı ile anlamsal bakımından benzeştiği veya yakın anlamlı olduğu bir ifade ile değişimi mümkün olamamaktadır. Buna karşılık cümle bazında değişiklikleri içeren bir, altı numaralı cümlelerde yer alan “presidential” ve “spokesman”, gibi kelimeleri, tüm metinde bulunan anlamlı en yalın halleri ile yani “president” ve “spoke” kelimeleri ile değiştirilmiştir. Kelimeler yine metnin tamamında aranmış ve belirlenmiştir. Böylelikle sözü edilen ekli kelimelerin farklı anlamlar taşıyormuş gibi çizgeye aktarılmasının önüne geçilmektedir. Özetle değişikliğe uğrayan kelimenin yerini alacak olan en yakın ve anlamlı kelime yazılım aracı tarafından yine metin içinde aranmakta ve değişim yapılmaktadır.

Çizelge 5.1. Önerilen yaklaşım ile DUC-2002 veri setinde bulunan “d065” dosyasında gerçekleşen örnek değişimler.

		Önişlemeden Önce	Önişlemeden Sonra
Kelime-bazlı	1	presidential	president
	2	Quayle's, Division's	quayle, division
	3	Officials	official
	4	served	serve
	5	reporters	report
	6	spokesman	spoke
	7	advanced	advance
Cümle-bazlı	1	Republican vice presidential candidate Dan Quayle's official resume says he was the Indiana Consumer Protection Division's chief investigator a year before the unit was created, state records show.	republican vice president candidate dan quayle official resume indiana consumer protection division chief investigator unit created record show.
	2	The law creating the unit passed April 14, 1971, and took effect July 1 Officials said Quayle served as chief investigator from July until mid-September, when he became administrative assistant to then-Gov Edgar D Whitcomb Quayle's Senate biography, which his campaign is using as his official background, says he was chief investigator of the unit in	law creating unit passed april 14 1971 effect july 1 official quayle serve chief investigator july mid september administrative assistant then gov edgar whitcomb quayle senate biography campaign using official background chief investigator unit attorney general office 1970 1971.

Önişlemeden Önce		Önişlemeden Sonra
		<i>the Attorney General's office from 1970 to 1971</i>
3	<i>"He was assigned to our section in the spring of 1971," said Robert Smith, the unit's first director and now an attorney in Indianapolis.</i>	<i>he assigned section spring 1971 robert smith unit director attorney indianapolis</i>
4	<i>I didn't hold that job for two years</i> Quayle told <i>reporters</i> aboard <i>his</i> campaign plane.	<i>i didn hold job quayle told report aboard campaign plane.</i>
5	Earlier, <i>however</i> , Quayle <i>spokesman</i> David Prosperi called <i>the</i> revelation of <i>the</i> error <i>by</i> <i>The Plain Dealer</i> of Cleveland " cheap shot.	<i>earlier quayle spoke david prosperi called revelation error plain deal cleveland a cheap shot</i>
6	" Quayle <i>was</i> hired <i>as an entry-level</i> research assistant <i>by the state</i> Attorney <i>General's</i> office <i>in 1970 after he</i> returned <i>from six months of basic and advanced</i> training <i>for the</i> Indiana National Guard.	<i>quayle hired entry level research assistant attorney general office 1970 returned six months basic advance training indiana nation guard.</i>

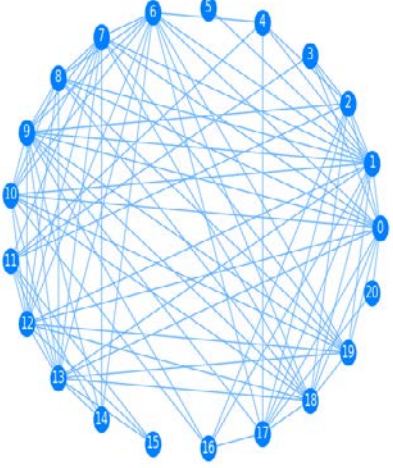
Kelimelerdeki değişiklikler, metindeki diğer kelimelerle ilişki ve n-gram kesişimleri dikkate alınarak gerçekleştirilir. Kısaca, bazı iyi bilinen Metin Normalleştirme yöntemlerine (ör. Spacy, NLTK, OpenNLP, Stanford CoreNLP ve ClearNLP) ek olarak farklı bir yaklaşım önerilmiştir. Önerilen yaklaşımda, ana metindeki kelimeler metinde bulunan en basit biçimlerine dönüştürülür. Tüm bu farklılıklar nedeniyle, önerilen metin işleme aracı standart araçlardan farklıdır. Amaç, metin sunumu adımı maksimum temsili sağlamaktır.

5.2. Metinsel Çizge Oluşumu

Veri önileme ve hazırlık aşamaları Bölüm 5.1’de anlatıldığı gibi gerçekleştirildikten hemen sonra önerilen Karcı Özetleme modeli, metin formatına sahip veriler üzerinde özetleme yapabilmek için metinleri temsil eden çizgeler oluşturmaktadır. Bu bölümde gerçekleştirilen işlemler Şekil 5.3’ de şematize edilen aşamalardan metinsel çizge aşamasına karşılık gelmektedir. Şekil 5.1, basit bir metin örneğini ve bu metin için oluşturulan çizgeyi içermektedir. Çizge ağırlıklı ve yönsüz bir çizgedir. Cümle-Kelime Çizgesi elde edilirken düğümlerin cümleler ile ayrıtların ise ortak kelime sayıları ile temsil edildiği bir dönüşüm önermekteyiz. Cümlelerin her birinin, diğer bütün cümlelerle mevcut kelime kesişim sayıları hesaplanarak cümle ikililerinin ilişki düzeyleri ortaya konmaktadır. Bu sayede cümleleri ve cümleler

arasındaki ilişkileri temsil gücü yüksek bir çizge ile ifade edebilmekteyiz. Şekil 5.1’de yer verilen çizge özetlenecek metinde yer alan tüm kelime öbek çiftleri arasındaki bağlantı ve ilişkilerin düzeyini ortaya koyan bir çizgedir.

(0)Republican vice presidential candidate Dan Quayle's official resume says he was the Indiana Consumer Protection Division's chief investigator a year before the unit was created, state records show.(1)The law creating the unit passed April 14,1971 and took effect July 1 Officials said Quayle served as chief investigator from July until mid-September, when he became administrative assistant to then-Gov Edgar D Whitcomb Quayle's Senate biography, which his campaign is using as his official background, says he was chief investigator of the unit in the Attorney General's office from 1970 to 1971.(2)"He was assigned to our section in the spring of 1971,"said Robert Smith, the unit's first director and now an attorney in Indianapolis.(3)He helped get the unit set up.(4)But it would not have been before 1971," Smith said Friday in a telephone interview.(5)"We formally began operations July 1.(6)"Quayle conceded Friday that the resume entry was "inaccurate" and said his staff had written the document.(7)I didn't hold that job for two years," Quayle told reporters aboard his campaign plane.(8)Earlier, however, Quayle spokesman David Prosseri called the revelation of the error by The Plain Dealer of Cleveland "a cheap shot.(9)"Quayle was hired as an entry-level research assistant by the state Attorney General's office in 1970 after he returned from six months of basic and advanced training for the Indiana National Guard.(10)Quayle, a 1969 graduate of DePauw University, was attending Indiana University Law School at the time.(11)M Stanton Evans, then an editor with the Indianapolis News,helped arrange his job interview with the state.(12)Evans worked for Eugene C Pulliam,Quayle's grandfather and publisher of the News.(13)"At no point in any of this did Dan Quayle ask any favors or use any influence,"said Evans, now with the National Journalism Center in Washington, DC "He wouldn't have been hired as a chief investigator.(14)He was hired in an entry-level job and was promoted to a whole bunch of jobs on his merits.(15)It's a nothing story,"Evans said.(16)Theodore Sendak, the Attorney General at the time, recently had surgery and is unable to comment, according to his son, Timothy Sendak.(17)Patricia Gifford, a Marion Superior Court judge who was then an assistant attorney general,said Friday that prior to the establishment of the Consumer Protection Division, "all complaints and inquiries were handled by me.(18)The legal assistant assigned to me and utilized as my chief investigator was Dan Quayle.(19)"Her statement was released by the Quayle campaign.(20)The Consumer Protection Division was created by the state legislature to resolve complaints over defective products, false claims in ads, faulty speedometer readings and mail-order fraud



Şekil 5.1 . DUC-2002, “D065” belgesine ait metinsel çizge.

Metinler içerisindeki her cümle için temsili çizgeye bir düğüm eklenmiştir. Düğümler arasındaki ayrıtlar için ise metinde bulunan kelime öbeklerine dair kesişen kelime sayıları dikkate alınarak ayrıt ağırlığı eklenmiştir. Şayet öbekler arasında kesişen kelime bulunmamakta ise ilgili düğümler arasına ayrıt eklenmemektedir. Gerçekleştirilen kelime dönüşümleri ile cümleler arasındaki ilişkinin anlamsal doğruluğu düzenlenmekte, düğümler anlamlı ilişkiler ile birbirlerine bağlanmaktadır. Böylelikle bütün cümleler, tüm kombinasyonları kapsayacak şekilde birbirleri ile olan ortak içerikleri bakımından ilişkilendirilmiş ve cümleler arasındaki ilişkiler çizgeye doğru bir şekilde aktarılabilmektedir.

5.3. Kullanılan Veri Setleri

Bu tez çalışması sınırlarında, önerilen yöntemlerin doğruluğunu test etmek amacı ile Document Understanding Conference (DUC) veri setinden faydalanılmıştır. Bu veri seti yorumlayıcı ve çıkarıcı özetlemeye yönelik dokümanlar bulundurmaktadır. Yapılan çalışmada çıkarıcı bir özetleme yöntemi önerildiği için çıkarıcı özetleme dosyalarından faydalanılmıştır. Ayrıca bu dosyalara ait 100, 200 ile 400 kelime sayısı sınır alınarak oluşturulan ve her birinde iki farklı değerlendirici tarafından oluşturulan özetler bulunmaktadır. Çizelge 5.2, veri setlerinin karakteristiklerine yönelik bilgiler içermektedir [94][95].

Çizelge 5.2. Veri Setlerinin Karakteristikleri.

Tanım	DUC-2002	DUC-2004
Küme Sayısı	59	50
Kümelerde ki Doküman Sayısı	~10	10
Doküman Sayısı	567	500
Özet Uzunluğu	200 ve 400 (kelime)	665 byte

5.4. Değerlendirme Ölçütleri

Bu tez çalışması sınırlarında, önerilen yöntemlerin performansını ölçmek için Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrikleri kullanılmıştır. Bu araç seti n-gram istatistiklerine dayanarak özetlerin değerlendirildiği bir performans metriğidir. Elde edilen özetler ile insanlar tarafından oluşturulan ideal özetler arasında meydana gelen kesişimlerin baz alındığı yaygın bir ölçüttür. Üretilen puanların yüksek olması, model özet ile paylaşılan ortak içeriklerin fazlalığı şeklinde yorumlanır. Rouge metrikleri sıfır ile bir arasında değerler almaktadır. ROUGE-N ile önerilen ve model özet arasında paylaşılan n-gram sayısı ölçülmektedir.

$$ROUGE\ N = \frac{\sum_{C \in Ref.\text{Özetler}} \sum_{gram_n \in S} Sayısı_{eşleşme}(gram_n)}{\sum_{C \in \{Ref.\text{Özetler}\}} \sum_{gram_n \in S} Sayısı(gram_n)} \quad (5.1)$$

Formüldeki n değeri, n-gram uzunluğudur. $Sayı_{eşleşme}(gram_n)$ değeri, referans özet ile kalitesi ölçülmek istenen taslak özetin maksimum kesişme değerini tumaktadır. ROUGE-L, iki alt kelime dizgesinde, en uzun ortak kelime dizisini hesaplamaktadır [96].

$$R_{lcs} = \frac{LCS(A, B)}{m} \quad (5.2)$$

$$P_{lcs} = \frac{LCS(A, B)}{m} \quad (5.3)$$

$$F_{lcs} = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2P_{lcs}} \quad (5.4)$$

Ayrıca bu çalışmada değişim ve varyasyon katsayısı dağılım istatistikleri kullanılmıştır.

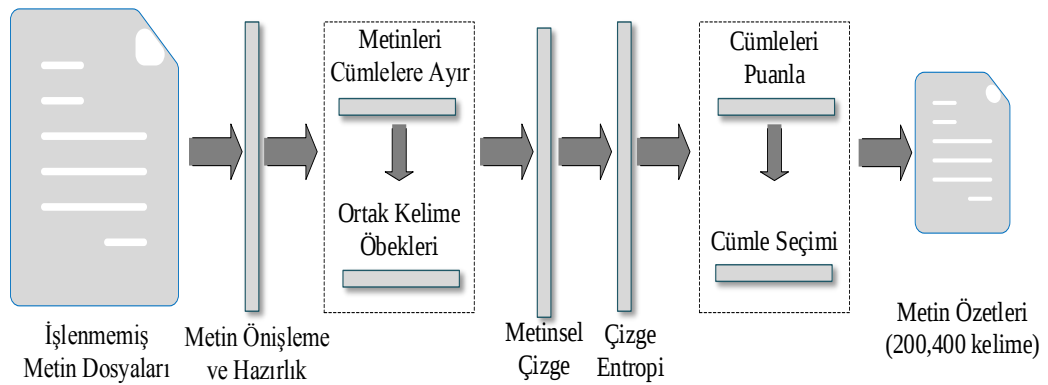
$$ROUGE_{yayılma_aralığı} = Rouge_{en_iyi} - Rouge_{en_kötü} \quad (5.5)$$

$$CV = \frac{ROUGE_{yayılma_aralığı}}{Rouge_{ortalama}} \cdot 100 \quad (5.6)$$

Deneyler boyunca pyrouge kullanılarak, ROUGE-1.5.5 uygulanmıştır (açık erişimli).

5.5. Çizge Entropi ile Metin Özetleme Uygulaması

Bu çalışmada çıkarıcı, genel ve denetimsiz bir özetleme sistemine odaklanılmaktadır. Özetlenmek istenen metinlerin çizgeler ile temsili sağlanmakta ve ardından metin dosyalarını temsil eden çizgeler üzerinde yapısal kararlılığı ve yapısal bilgi içeriğini yorumlamak için çizge entropi kullanılmaktadır. Çalışmamızda çıkarıcı metin özetleme çalışmaları kapsamında çizge ve entropi tabanlı bir yöntem kullanılmaktadır. Özetlenmek istenen metinlerin çizgeler ile temsili sağlanmaktadır. Çalışmada, metin dosyalarını temsil eden çizgeler üzerinde yapısal kararlılığı ve yapısal bilgi içeriğini yorumlamak için Shannon Entropi [82] kullanılmıştır.



Şekil 5.2 . Özetleme sisteminin ana hatları.

Bu çalışma kapsamında gerçekleştirilen entropi tabanlı metin özetleme çalışmasının doğruluğunu test etmek amacı ile DUC2002 veri seti kullanılmıştır. Kullanılan veri seti yorumlayıcı ve çıkarıcı özetlemeye yönelik metinler bulundurmaktadır. Yapılan çalışmada çıkarıcı bir özetleme gerçekleştirildiği için çıkarıcı özetleme metinleri kullanılmıştır. Kategorik anlam taşımayan ve durak kelimeler olarak adlandırılan ifadelerin metinlerden uzaklaştırılması suretiyle gerçekleştirilen normalizasyon adımları C# dili etiketleri kullanılarak gerçekleştirildi. Bu sayede kategorik anlamlar taşıyan ifadeler ile sürecin devamı hedeflendi. Böylelikle metinlere ait yüksek temsil gücüne sahip çizgeler elde edildi. Çizge entropi temelli elde ettiğimiz özetleme performans sonuçları önceki çalışmalar ile karşılaştırılmıştır. Kelime sıklık ve dağılımlarından elde edilen değerleri, cümlelerin önemlerini makine öğrenmesi teknikleri ile hesaplamak üzere kullanan Luhn [97] ön kabulsüz genel matematiksel yöntemleri kullanan Landauer [98], metinlerin ilişkilerini çizgelere aktararak iteratif ve denetimsiz bir uygulama gerçekleştiren TextRank [99] ve Google'ın PageRank'ından esinlenilen yine çizge bazlı LexRank [34] çalışmamızda ilgili çizgelelerde yer verdiğimiz rekabetçi yöntemlerdir. Daha önce gerçekleştirilen özetleme yaklaşımları ile uyumlu olması bakımından Çizelge 5.3 ve 5.4'te olduğu gibi 200 kelimelik ve 400 kelimelik özetler elde edilmiştir. Elde edilen çizgelere Shannon Entropi uygulandı. Çizgeyi oluşturan bütün düğümler sırasıyla çıkarılarak çizgeye ait toplam entropi hesaplanmıştır. Deneysel çalışmamızın süreci gereği değerler küçükten büyüğe sıralandı ve metinsel çizgeleri oluşturan düğümlere dair taşıdığı bilgi miktarı nicel olarak elde edildi. Böylece düğümün global çizge üzerindeki etkisi ölçüldü. Etkisi en büyük olan düğümleri temsil eden cümlelere önem sırasına göre özetle yer verilerek özetler elde edildi. Özetler Rouge-1, Rouge-2, Rouge-3, Rouge-4 performans metrikleri açısından değerlendirildi. Sözü edilen metrikler Recall, Precision ve F-Score bazlı değerlerdir. Çizelge 5.3'de görüldüğü gibi çalışma kapsamında 200 kelimelik özetlerde %95 güven aralığı için elde ettiğimiz tüm ROUGE türleri için (Recall, Precision ve F-Score) LSA özetleme yaklaşımı geride bırakılmıştır. Recall, Precision ve FScore bazlı elde ettiğimiz değerler ROUGE-2, ROUGE-3, ROUGE-4 metrik türleri için ise TextRank metin özetleme sistemini geride bırakmıştır. Çizelge 5.4'de 400 kelimelik özetler için elde edilen değerler yer almaktadır. İlgili çizelge göstermektedir ki, deneysel çalışma sürecinde elde ettiğimiz özetler tüm performans metrikleri ve metrik türleri için LSA metodundan daha yüksek değerler rapor etmektedir. TextRank metin özetleme

yaklaşımı ise Rouge-1 (Recall) hariç tüm metrikler açısından geride bırakılmıştır. Yine benzer şekilde Rouge-1 (Recall) metriği dışında bütün değerler açısından LexRank özetleme yönteminden daha yüksek değerler rapor edilmiştir. 200 ve 400 kelimelik özetler bazında ilgili çizelgelerde yer verilen rekabetçi yöntemlerin çoğunluğu geride bırakılmıştır. Kısmi olarak 400 kelimelik özetler, 200 kelimelik özetlere göre daha yüksek performans göstermiştir.



Çizelge 5.3. Çalışma kapsamında ve rekabetçi yöntemlerle oluşturulan 200 kelimelik özetler için rouge metrik değerleri.

200 kelimelik özetler						
ROUGE performans metriği						
Özetleme Yöntemleri						
	Luhn [97]	LSA [98]	TextRank [99]	LexRank [34]	Çizge Entropi	
Rouge-1	Recall	0.47503	0.37723	0.48417	0.48124	0.45989
	Precision	0.44489	0.37012	0.43685	0.45984	0.43035
	F-Score	0.45924	0.37342	0.45868	0.46997	0.44410
Rouge-2	Recall	0.20820	0.08950	0.16921	0.18507	0.17110
	Precision	0.19558	0.08910	0.15195	0.17566	0.15865
	F-Score	0.20160	0.08921	0.16000	0.18010	0.16446
Rouge-3	Recall	0.15318	0.04607	0.10382	0.12907	0.11577
	Precision	0.14392	0.04724	0.09316	0.12227	0.10720
	F-Score	0.14834	0.04660	0.09816	0.12546	0.11120
Rouge-4	Recall	0.13927	0.03300	0.08994	0.11500	0.10216
	Precision	0.13085	0.03468	0.08066	0.10910	0.09465
	F-Score	0.13486	0.03379	0.08501	0.11187	0.09814

Çizelge 5.4. Çalışma kapsamında ve rekabetçi yöntemlerle oluşturulan 400 kelimelik özetler için rouge metrik değerleri.

400 kelimelik özetler						
ROUGE performans metriği						
Özetleme Yöntemleri						
	Luhn [97]	LSA [98]	TextRank [99]	LexRank [34]	Çizge Entropi	
Rouge-1	Recall	0.57929	0.46361	0.57812	0.55018	0.55425
	Precision	0.57584	0.47293	0.52849	0.53431	0.55436
	F-Score	0.57746	0.46804	0.55179	0.54188	0.55409
Rouge-2	Recall	0.32775	0.16001	0.25654	0.23706	0.28035
	Precision	0.32546	0.16297	0.23379	0.23088	0.28132
	F-Score	0.32654	0.16141	0.24447	0.23382	0.28073
Rouge-3	Recall	0.27092	0.10774	0.17686	0.16341	0.21485
	Precision	0.26884	0.10998	0.16112	0.15975	0.21579
	F-Score	0.26983	0.10881	0.16851	0.16148	0.21524
Rouge-4	Recall	0.25511	0.09019	0.15543	0.14339	0.19590
	Precision	0.25312	0.09228	0.14158	0.14038	0.19675
	F-Score	0.25406	0.09119	0.14808	0.14179	0.19625

5.6. Karcı Özetleme: Karcı Entropi tabanlı Metin Özetleme amacı ile Basit ve Etkili bir Öneri

Metin özetleme için önerilen Karcı Özetleme modelinin süreçlerine dair blok diyagramı Şekil 5.3’de gösterilmektedir

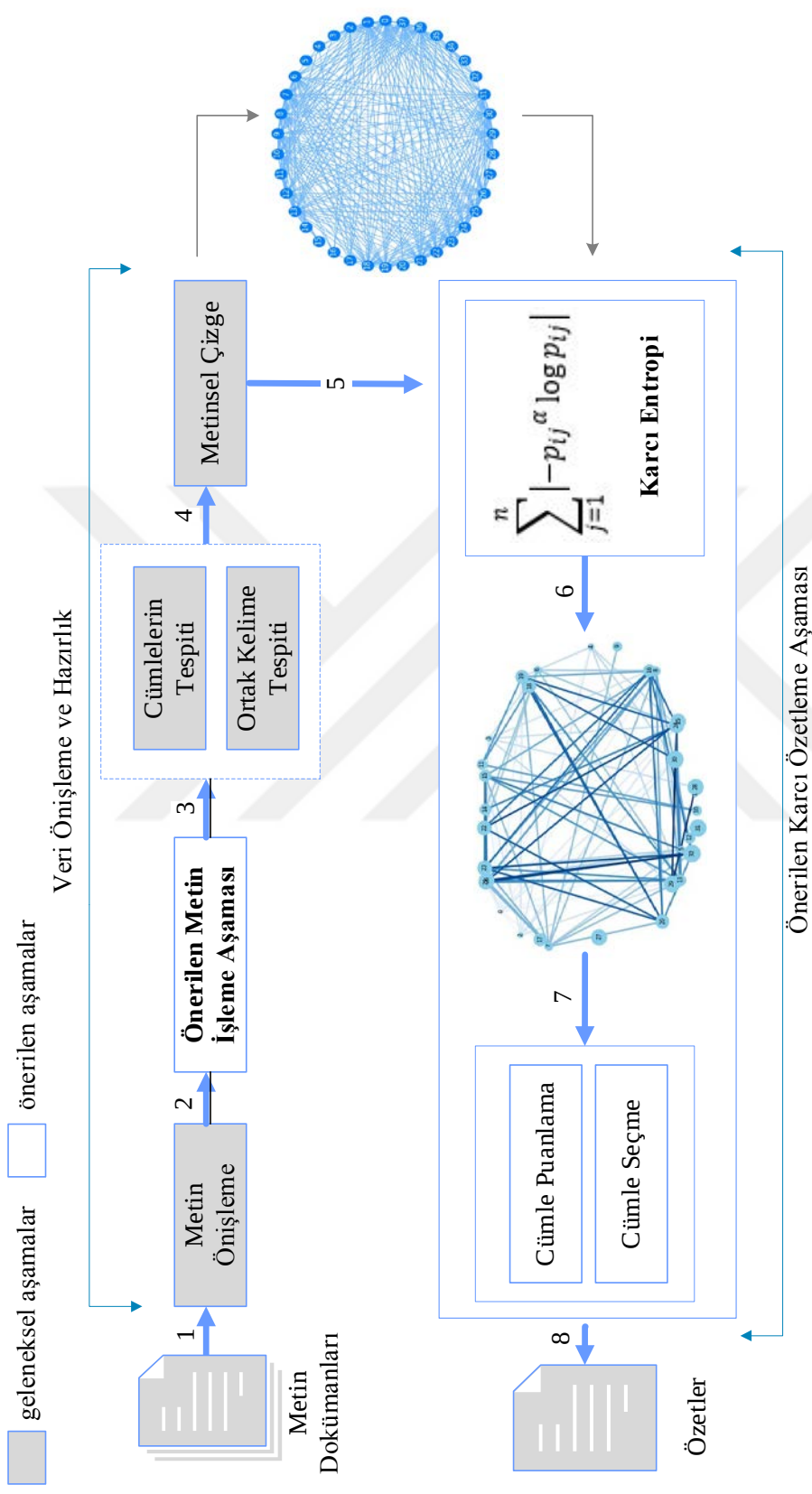
Bu çalışmada, verilen metinlerden uygun özetler çıkarılması amacıyla çizge tabanlı genel, çıkarıcı ve çok belgeli bir özetleme yöntemi sunulmuştur. Önerilen Karcı Özetleme yöntemi temelde iki ana aşamadan oluşmaktadır.

Veri Önileme ve Hazırlık Aşaması;

- İlk aşamada herhangi bir ayırt edicilik taşımayan durak kelimeler (zamirler, edatlar, bağlaçlar, vb..) veri setinden uzaklaştırılmıştır. Bu işlem Şekil 5.3’te belirtilen adımlardan ilkinin oluşturmaktadır.
- İkinci adım, geliştirilen metin işleme aracı kullanılarak birtakım önilemlerin yapıldığı aşamadır. Bu aşamada, yazım şekilleri bakımından birbirlerinden farklı olan ancak anlamsal olarak ortak bir kelime kökünden türeyen yakın anlamlı kelimelerin çizgeler oluşturulurken farklı kelimeler gibi işleme alınmasının önüne geçilmektedir.
- Veri önileme ve hazırlık aşamasının son aşaması olan dördüncü adımda kelime öbekleri arasındaki kelime ortaklıkları matematiksel olarak temsil edilmektedir. Cümle-Kelime Çizgesi elde edilirken düğümlerin cümleler ile ayrıtların ise ortak kelime sayıları ile temsil edildiği bir dönüşüm gerçekleştirilmektedir.

Önerilen Karcı Özetleme Aşaması;

- Beşinci adım, çalışma kapsamında Karcı Entropi’nin uygulandığı ana aşamadır. Önerilen yöntemin bu aşamasında farklı α değerleri için ($\alpha=0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3$) Karcı Entropi değerleri ayrı ayrı hesaplanmaktadır.
- Altıncı adım Karcı Entropi den alınan değerler doğrultusunda çizgede yer alan düğümlerin sıralandığı ve çizge üzerindeki ağırlık ve belirleyiciliklerinin ortaya konulduğu aşamadır.
- Karcı Entropi den alınan değerler doğrultusunda en ağırlıklı kelime öbekleri bir araya getirilerek özetin elde edildiği adımlar ise 7. ve 8. adımlardır.



Şekil 5.3 . Karıcı özetleme modeline dair şematik yapı

Etkili düğümleri ortaya çıkarmak için çizge entropi kullanıldı. Karcı Entropi daha önce herhangi bir metin özetleme çalışmasında kullanılmamıştır. Farklı α üslerinin Karcı Özetleme yaklaşımı üzerindeki etkilerinin detaylı olarak incelenebilmesi için tekrarlı Karcı Entropi uygulanmıştır. Aşağıdaki algoritmada gösterildiği gibi Karcı entropi, cümleyi temsil eden düğümlerin toplam çizge üzerindeki etkisini ölçer. Önerilen algoritmada, genel ağın nasıl etkilendiğini belirlemek için her bir düğüm sırasıyla çizgeden çıkarılır.

Önerilen Karcı Özetleme Algoritması

```

1  V : düğümlerin kümesi
2  E: kenarlar kümesi
3  G  $\leftarrow$  (V,E)
4  for each  $\alpha_i \in \{0.1, 0.2, \dots, 1.3\}$ 
5      for k=0 to len(V) do
6          çizgedeki kth düğümü çıkar
7          for i=0 to len(V)-1 do
8              for j=0 to len(V)-1 do
9                  Denklem (3.15) ile kenar olasılıklarını hesapla
10             end for
11             Denklem (3.17) ile kth düğüm için Karcı Entropi hesapla
12         end for
13         Denklem (3.18) ile çizge entropi hesapla
14         E(G) değerlerini sırala
15         en düşük Karcı Entropi değerinden başlayarak (200 ve 400 kelimelik
            özetler oluştur)
16 end for

```

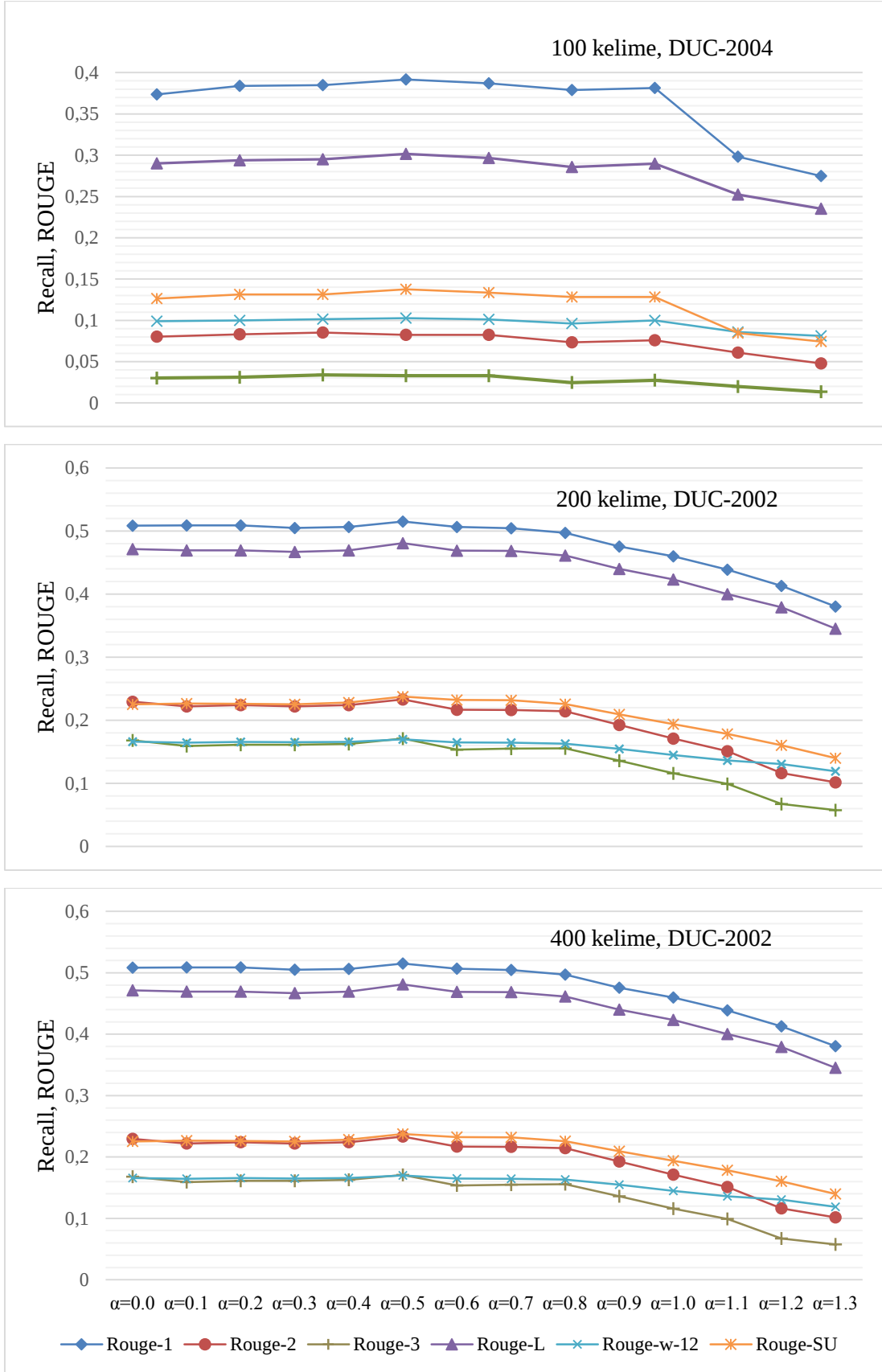
5.6.1 Deneysel Sonuçlar

Bütün deneysel süreçler Intel Xeon E32245 3,30GHz CPU ve 8Gb hafızaya sahip bir bilgisayar ile .Net ve Python kullanılarak gerçekleştirilmiştir.

Çalışmada kullanılan DUC veri setlerindeki metinler .txt formatına sahiptirler. Durak kelimeler gibi istenmeyen karakterlerin metinlerden uzaklaştırıldığı normalizasyon aşamaları C# dili kütüphane ve etiketleri kullanılarak gerçekleştirilmiştir. Böylece sürece ait işlem yükü azaltıldı ve kategorik anlam taşıyan kelimelerle Karcı özetleme sürecinin sürdürülmesi sağlandı.

Önişleme aşamasından sonra kelime öbekleri arasındaki ilişkilerin metinsel çizgelere en doğru bir biçimde aktarılabilmesini sağlayabilmek için çalışma kapsamında önermiş olduğumuz yazılım aracı kullanıldı. Yazılımın, basit ve etkili algortması sayesinde değişikliğe uğrayan kelimelerin yerine, yine özetlenecek metin içinden atamalar gerçekleştirilmiştir. Sonuç olarak metinlere ait yüksek temsil gücüne sahip çizgeler elde edildi.

İlgili çizgelere Karcı Entropi uygulandı. Her bir düğüm ağdan sırayla çıkarılarak toplam Karcı Entropi hesaplandı. Çalışma kapsamında önerilen Karcı Özetleme süreci gereği değerler küçükten büyüğe sıralandı ve metinsel çizgeyi oluşturan düğümlere dair taşıdığı bilgi miktarı nicel olarak elde edildi. Böylece düğümün çizgenin geneli üzerindeki etkisi ölçüldü. Karcı Entropi hesaplanırken ilgili denklemde yer alan α üs değerinin etkisi çizge üzerinde yerel ve genel etkiyi ölçecek şekilde gerçekleşmiştir. Küçük α değerleri yerel anlamda düğümlerin etkisini ortaya koyarken büyük α değerleri ise genel ölçüde düğümleri incelemektedir. Çalışmamızda α üs değeri 0.0 ile 1.3 arasında 0.1 değişim aralığı ile basit bir algoritma kullanılarak gerçekleştirilmiştir. α üs değerinin değişimi Karcı Özetleme üzerinde Recall-Oriented Rouge-1, Rouge-2, Rouge-3, Rouge-L, Rouge-W-1-2, Rouge-SU performans metriği açısından Şekil 5.4’de görüldüğü gibi bir değişim ortaya koymaktadır. Deneysel süreçte α üs değerinin etkisi ilgili şekilde görüldüğü üzere 200 ve 400 kelimelik özetler için ayrı ayrı gösterilmiştir. Duc-2004 veri seti ile 100 kelimelik özetler için gerçekleştirilen deneysel çalışmalarda $\alpha=0.1$ den itibaren artan Recall-Oriented Rouge değerleri elde edilmiştir. Genel olarak yüksek başarımlı değerleri $\alpha=0.4$, $\alpha=0.5$, $\alpha=0.6$ üs değerleri için rapor edilmiş olsa da, En yüksek değerler çoğunlukla 0.5 α değeri için elde edilmiştir. 200 kelimelik özetler için $\alpha=0.1$ artım miktarına karşılık değişken Recall-Oriented Rouge değerleri elde edilmiştir. Başlangıçtan itibaren kısmi artan bir grafik elde edilmiş, $\alpha=0.5$ değerini aldığı anda istisnasız maksimum ROUGE değerine ulaşıldı. Devam eden α değerleri için çalışılan örneklemin sonlarına doğru sürekli bir azalış sergilendi ve örneklem üzerinde $\alpha=1.3$ için istisnasız minimum değer aldı. Şekil 5.4’e göre 400 kelimelik özetler için de yine $\alpha=0.1$ den itibaren kısmi artan bir grafik elde edilmiş neredeyse maksimum Recall-Oriented Rouge değerlerinin tümü $\alpha=0.3$, $\alpha=0.4$, $\alpha=0.5$ üs değerlerinde kümelenmiştir. Örneklem üzerinde $\alpha=1.3$ için istisnasız minimum değer aldı.



Şekil 5.4 . Karıcı Özetleme ile elde edilen, Recall bazlı ROUGE metrik puanları üzerindeki α etkisi (DUC-2002, DUC-2004)

Gerçekleştirilen tüm testler Karcı Entropi α üs değeri 0.0 ile 1.3 arasında 0.1 değişim aralığı ile gerçekleştirilmiştir. Önerilen Karcı Özetleme performansı Çizelge 5.5, 5.6 ve 5.7’de yer verilen ROUGE performans metrikleri kullanılarak değerlendirildi. Kullanılan metrikler Rouge-1, Rouge-2, Rouge-3, Rouge-L, Rouge-W-1-2 ve Rouge-SU dur. Ayrıca her bir metrik türü Recall, Precission, F-measure odaklı ayrı ayrı rapor edilmiştir. Çizelge 5.5, 5.6 ve 5.7’nin ilk iki sütunu özetleme performans metrik türlerini gösterirken kalan sütunlar farklı α değerleri için Karcı Özetleme tarafından rapor edilen performans değerlendirme puanlarıdır. Çizelge 5.5’de görüldüğü gibi en yüksek Rouge-1, Rouge-2, Rouge-L, Rouge-W-1.2 and Rouge-SU (Recall bazlı) puanları sırasıyla 0.39183, 0.08541, 0.30148, 0.10278, ve 0.13771 dir. Çizelge 5.5’e göre $\alpha=0.5$ için birçok metrik türünde en iyi değerler rapor edilmiştir. Sonraki en yüksek sonuçlara sırasıya $\alpha=0.4$, $\alpha=0.6$ için ulaşılmıştır. Çizelge 5.6, tüm metrik türleri için en yüksek sonuçlara $\alpha=0.5$ değeri için ulaşıldığını göstermektedir. Elde edilen en yüksek Rouge-1, Rouge-2, Rouge-L, Rouge-W-1.2 ve Rouge-SU (Recall bazlı) skorları sırasıyla 0.51516, 0.23323, 0.48095, 0.16980, 0.23749 dir. Çizelge 5.6’ya ait bir diğer önemli gözlem, en yüksek ve en düşük değerlerin tümünün sırasıyla 0.5 ve 1.3 için elde edilmesidir.

Çizelge 5.5. Önerilen Karıcı Özetleme yöntemi için farklı α değerlerine dayanan 100 kelimelik özetlerin ROUGE metrik değerleri.

ROUGE değerlendirme kriterleri		100 Kelimelik Özetler (DUC-2004)									
		$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.9$	$\alpha = 1.0$ Shan.ent.	$\alpha = 1.1$	$\alpha = 1.3$
Rouge-1	Recall	0.37349	0.38398	0.38495	0.39183	0.38703	0.37894	0.38136	0.30470	0.29822	0.27471
	Prec.	0.33363	0.34213	0.34200	0.34511	0.34515	0.33676	0.32900	0.28414	0.27456	0.25828
	F-Sc.	0.35136	0.36081	0.36139	0.36605	0.36422	0.35597	0.35270	0.29337	0.28530	0.26549
Rouge-2	Recall	0.08028	0.08317	0.08541	0.08259	0.08259	0.07352	0.07589	0.06857	0.06097	0.04770
	Prec.	0.07196	0.07442	0.07677	0.07379	0.07432	0.06548	0.06544	0.06186	0.05416	0.04426
	F-Sc.	0.07565	0.07831	0.08067	0.07773	0.07808	0.06912	0.07015	0.06491	0.05726	0.04578
Rouge-L	Recall	0.29002	0.29383	0.29492	0.30148	0.29668	0.28576	0.28966	0.25982	0.25223	0.23524
	Prec.	0.25989	0.26225	0.26189	0.26557	0.26470	0.25387	0.25010	0.24260	0.23280	0.22187
	F-Sc.	0.27330	0.27636	0.27680	0.28166	0.27926	0.26838	0.26800	0.25032	0.24161	0.22772
Rouge-W-1.2	Recall	0.09896	0.09998	0.10132	0.10278	0.10111	0.09607	0.09989	0.08891	0.08599	0.08111
	Prec.	0.16237	0.16327	0.16447	0.16540	0.16486	0.15606	0.15764	0.15229	0.14572	0.13997
	F-Sc.	0.12271	0.12377	0.12515	0.12650	0.12515	0.11874	0.12206	0.11204	0.10796	0.10231
Rouge-SU	Recall	0.12644	0.13132	0.13140	0.13771	0.13349	0.12828	0.12814	0.09071	0.08466	0.07444
	Prec.	0.10061	0.10410	0.10387	0.10704	0.10613	0.10113	0.09540	0.07707	0.06984	0.06518
	F-Sc.	0.11067	0.11480	0.11496	0.11922	0.11739	0.11229	0.10866	0.08260	0.07591	0.06873

Çizelge 5.6. Önerilen Karıcı Özetleme yöntemi için farklı α değerlerine dayanan 200 kelimelik özetlerin ROUGE metrik değerleri.

ROUGE değerlendirme kriterleri		200 Kelimelik Özetler (DUC-2002)									
		$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.9$	$\alpha = 1.0$ Shan.ent.	$\alpha = 1.1$	$\alpha = 1.3$
Rouge-1	Recall	0.50889	0.50505	0.50645	0.51516	0.50669	0.50464	0.47555	0.45989	0.43878	0.38045
	Prec.	0.46700	0.46544	0.46733	0.47495	0.46456	0.46381	0.44236	0.43035	0.41569	0.36801
	F-Scor.	0.48674	0.48413	0.48592	0.49406	0.48453	0.48316	0.45806	0.44410	0.42667	0.37383
Rouge-2	Recall	0.22186	0.22197	0.22397	0.23323	0.21692	0.21641	0.19249	0.17110	0.15109	0.10178
	Prec.	0.20419	0.20640	0.20854	0.21692	0.20064	0.20084	0.17982	0.15865	0.14177	0.09754
	F-Scor.	0.21250	0.21376	0.21590	0.22470	0.20837	0.20824	0.18583	0.16446	0.14620	0.09950
Rouge-L	Recall	0.46931	0.46690	0.46929	0.48095	0.46896	0.46837	0.44011	0.42334	0.40001	0.34529
	Prec.	0.43072	0.43041	0.43319	0.44357	0.43015	0.43068	0.40929	0.39557	0.37889	0.33422
	F-Scor.	0.44890	0.44763	0.45035	0.46134	0.44855	0.44855	0.42387	0.40850	0.38893	0.33939
Rouge-W-1.2	Recall	0.16442	0.16512	0.16589	0.16980	0.16507	0.16444	0.15489	0.14480	0.13619	0.11899
	Prec.	0.23765	0.24009	0.24158	0.24716	0.23892	0.23870	0.22728	0.21334	0.20300	0.18117
	F-Scor.	0.19424	0.19554	0.19663	0.20123	0.19517	0.19465	0.18412	0.17232	0.16295	0.14354
Rouge-SU	Recall	0.22671	0.22540	0.22829	0.23749	0.23237	0.23201	0.20948	0.19373	0.17832	0.13983
	Prec.	0.19147	0.19227	0.19504	0.20258	0.19605	0.19702	0.18183	0.16858	0.15916	0.12960
	F-Scor.	0.20709	0.20701	0.21005	0.21835	0.21235	0.21273	0.19420	0.17942	0.16780	0.13404

Çizelge 5.7. Önerilen Karcı Özetleme yöntemi için farklı α değerlerine dayanan 400 kelimelik özetlerin ROUGE metrik değerleri

ROUGE değerlendirme kriterleri		400 Kelimelik Özetler (DUC-2002)									
		$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.9$	$\alpha = 1.0$ Shan.ent.	$\alpha = 1.1$	$\alpha = 1.3$
Rouge-1	Recall	0.57282	0.58503	0.58669	0.58352	0.57966	0.58085	0.56841	0.55425	0.55553	0.51024
	Prec.	0.57645	0.57977	0.57857	0.57553	0.57758	0.57210	0.56583	0.55436	0.55314	0.51656
	F-Scor.	0.57455	0.58228	0.58241	0.57929	0.57846	0.57626	0.56691	0.55409	0.55420	0.51325
Rouge-2	Recall	0.30322	0.31661	0.31804	0.31289	0.31001	0.31497	0.30003	0.28035	0.28048	0.23855
	Prec.	0.30537	0.31295	0.31369	0.30875	0.30963	0.30988	0.29905	0.28132	0.27894	0.23996
	F-Scor.	0.30425	0.31471	0.31576	0.31071	0.30975	0.31231	0.29944	0.28073	0.27964	0.23919
Rouge-L	Recall	0.54266	0.55430	0.55652	0.55405	0.55144	0.55433	0.54397	0.52518	0.52399	0.47627
	Prec.	0.54613	0.54918	0.54885	0.54639	0.54956	0.54591	0.54153	0.52519	0.52165	0.48199
	F-Scor.	0.54431	0.55162	0.55248	0.5500	0.55035	0.54991	0.54255	0.52498	0.52270	0.47899
Rouge-W-1.2	Recall	0.17150	0.17672	0.17653	0.17595	0.17496	0.17580	0.17160	0.16497	0.16434	0.14997
	Prec.	0.27205	0.27602	0.27456	0.27366	0.27511	0.27299	0.26951	0.26020	0.25800	0.23918
	F-Scor.	0.21033	0.21543	0.21480	0.21409	0.21380	0.21379	0.20958	0.20179	0.20071	0.18428
Rouge-SU	Recall	0.29520	0.30734	0.31021	0.31172	0.30401	0.30688	0.28785	0.27727	0.28101	0.24569
	Prec.	0.29960	0.30121	0.30195	0.30347	0.30245	0.29664	0.28525	0.27761	0.27816	0.25077
	F-Scor.	0.29723	0.30400	0.30564	0.30714	0.30291	0.30131	0.28617	0.27705	0.27932	0.24795

Çizelge 5.7’de görüldüğü gibi elde edilen en yüksek Rouge-1, Rouge-2, Rouge-L, Rouge-W-1-2 ve Rouge-SU (Recall bazlı) skorları sırasıyla 0.58669, 0.31804, 0.55652, 0.17672, 0.31172 dir. Çizelge 5.7 tüm metrik ve türleri için en yüksek sonuçların önemli bir kısmının $\alpha=0.4$ değeri için elde edildiğini göstermektedir. $\alpha=0.3$ ve $\alpha=0.5$ değerleri içinde bir kısım en yüksek değerler elde edilmiştir. En yüksek Rouge performance metrikleri $\alpha=0.3$, $\alpha=0.4$, $\alpha=0.5$ değerleri için rapor edilmektedir (Rouge-L, Precision dışında). Ancak benzer şekilde çalışılan örneklem içinde en düşük değerler Çizelge 5.5’de ve Çizelge 5.6’de’de olduğu gibi $\alpha=1.3$ için elde edilmiştir. 400 kelimelik özetlerde de en düşük değerlerin tümünün $\alpha=1.3$ için elde edilmiş olması dikkat çekici bir başka gözlemdir.

Karcı Entropi’de farklı α değerlerinin kullanılması, özetlenen metinde cümlelerin hem yerel hem de küresel ağırlıklarının elde edilmesini mümkün kılmıştır. α 'nın 1’e eşit olduğu bir durumda Karcı Entropi, Shannon entropisine eşittir [1, 12]. Shannon entropi ($\alpha = 1$) için elde edilen değerler ilgili çizelgede yer almaktadır.

5.6.2 Karcı Özetlemenin Rekabetçi Yöntemler ile Karşılaştırılması

Önerdiğimiz Karcı Özetleme yaklaşımı ile elde edilen Özetleme performans sonuçları önceki çalışmalar ile karşılaştırılmıştır. Daha önce gerçekleştirilen özetleme yaklaşımları ile uyumlu olması bakımından Çizelge 5.8, 5.9, 5.10’da görüldüğü üzere 100, 200 ve 400 kelimelik özetler elde edilmiştir.

Çizelge 5.8. Rekabetçi yöntemler ile Karıcı Özetlemenin karşılaştırılması (100 kelimelik özetler).

ROUGE değerlendirme kriterleri		100 kelimelik özetler (DUC-2004)							
		Metot						Kareci ($\alpha = 0.5$)	
		Luhn [97]	LSA [98, 100]	Textrank [101, 102]	Lexrank [34]	Random	SumBasic [103]		KL-Sum[104]
Rouge-1	Rec.	0.35768	0.27995	0.31301	0.39893	0.32206	0.33859	0.33778	0.39183
	Prec.	0.29675	0.25740	0.28574	0.33462	0.30458	0.31882	0.32224	0.34511
	F-Sc.	0.32333	0.26733	0.29759	0.36292	0.31255	0.32808	0.32924	0.36605
Rouge-2	Rec.	0.05205	0.03324	0.04145	0.07977	0.05439	0.05623	0.07030	0.08259
	Prec.	0.04292	0.03068	0.03862	0.06831	0.05170	0.05226	0.06883	0.07379
	F-Sc.	0.04691	0.03182	0.03983	0.07338	0.05292	0.05413	0.06946	0.07773
Rouge-L	Rec.	0.27711	0.21295	0.26099	0.30685	0.25785	0.24408	0.27448	0.30148
	Prec.	0.22885	0.19611	0.23881	0.25868	0.24444	0.23038	0.26256	0.26557
	F-Sc.	0.24987	0.20354	0.24845	0.27991	0.25053	0.23680	0.26789	0.28166
Rouge-W-1.2	Rec.	0.09350	0.07427	0.08836	0.10436	0.08957	0.08427	0.09315	0.10278
	Prec.	0.14127	0.12498	0.14797	0.16113	0.15517	0.14533	0.16340	0.16540
	F-Sc.	0.11212	0.09286	0.11011	0.12622	0.11334	0.10656	0.11841	0.12650
Rouge-SU	Rec.	0.11866	0.07285	0.09741	0.13998	0.09532	0.10063	0.10609	0.13771
	Prec.	0.08109	0.06058	0.08277	0.09683	0.08421	0.08946	0.09795	0.10704
	F-Sc.	0.09516	0.06534	0.08820	0.11328	0.08884	0.09437	0.10117	0.11922

Çizelge 5.9. Rekabetçi yöntemler ile Karcı Özetlemenin karşılaştırılması (200 kelimelik özetler).

ROUGE değerlendirme kriterleri		200 kelimelik özetler (DUC-2002)							
		Metot						Karcı ($\alpha = 0.5$)	
		Luhn [97]	LSA [98, 100]	Textrank [101, 102]	Lexrank [34]	Random	SumBasic [103]		KL-Sum[104]
Rouge-1	Rec.	0.47503	0.37723	0.48417	0.48124	0.40932	0.46128	0.37464	0.51516
	Prec.	0.44489	0.37012	0.43685	0.45984	0.38148	0.45114	0.36803	0.47495
	F-Sc0.	0.45924	0.37342	0.45868	0.46997	0.39470	0.45597	0.37104	0.49406
Rouge-2	Rec.	0.20820	0.08950	0.16921	0.18507	0.10562	0.15947	0.12113	0.23323
	Prec.	0.19558	0.08910	0.15195	0.17566	0.09868	0.15584	0.11938	0.21692
	F-Sc0.	0.20160	0.08921	0.16000	0.18010	0.10199	0.15757	0.12018	0.22470
Rouge-L	Rec.	0.44490	0.33477	0.46748	0.43813	0.36328	0.39661	0.34066	0.48095
	Prec.	0.41687	0.32892	0.42183	0.41867	0.33880	0.38768	0.33470	0.44357
	F-Sc0.	0.43022	0.33162	0.44289	0.42787	0.35043	0.39194	0.33741	0.46134
Rouge-W-1.2	Rec.	0.15858	0.11295	0.15655	0.15201	0.12324	0.13782	0.11839	0.16980
	Prec.	0.23417	0.17475	0.22316	0.22825	0.18074	0.21231	0.18321	0.24716
	F-Sc0.	0.18902	0.13715	0.18382	0.18237	0.14647	0.16710	0.14375	0.20123
Rouge-SU	Rec.	0.19901	0.12783	0.21541	0.20321	0.15112	0.17981	0.13361	0.23749
	Prec.	0.17444	0.12342	0.17468	0.18494	0.13124	0.17190	0.12913	0.20258
	F-Sc0.	0.18556	0.12527	0.19200	0.19309	0.14020	0.17548	0.13098	0.21835

Çizelge 5.9’da Luhn [97] kelime sıklıklarından ve dağılımlarından elde edilen istatistik bilgileri cümlelerin önemlerini makine öğrenmesi teknikleri ile hesaplamak üzere kullandı. En yüksek cümle skorlarına göre cümleleri elde ederek özeti oluşturdu. Landauer et al.[98][100] çalışmasında bilgi ön kabulü kullanmadan sadece genel matematiksel yöntemleri kullanarak güçlü etkiler sağlamıştır. Bunda nesne ve bağlamları kullanmıştır. Mihalcea [101][99] bir metnin bağlı yapısını çizgelere aktararak Textrank adını verdiği iteratif, çıkarıcı ve denetimsiz bir uygulama gerçekleştirmiştir. Sistemin önemli görülen özelliği herhangi bir ön tanım gerektirmemesidir. Google’ın PageRank’ını [105] temel alan bu iteratif çalışması sayesinde metin birimlerinin önemini temel alarak bu birimleri puanlamıştır. [34]’de Erkan ve ark. DDİ kapsamında metinsel birimlerin önemini hesaplamak için stokastik ve çizge bazlı bir yaklaşım önermişlerdir. LexRank adını verdikleri yaklaşımlarında cümlelerin temsili çizge üzerindeki önemini öz vektör merkeziliği (düğüm merkeziliği temelli) ölçüsünü baz alarak hesaplamışlardır. Yazarlar deneysel çalışmalarında LexRank’ın hem derece bazlı metotlardan hem de merkezilik bazlı metotlardan çoğu durumda daha iyi performans sergilediğini göstermişlerdir. Literatürde sıkça kullanılan temel bir yöntem olan SumBasic [103], etkili bir olasılık fonksiyonu kullanılarak özete dahil edilecek etkili cümlelerini seçer. [104], açgözlü yaklaşımla (greedy approach) belirli bir olasılık modeli kullanılarak cümlelerin özete eklendiği belge özetleme çalışması yürütmektedir. DUC-2004 veri kümesinden 100 kelimelik özetlerle yapılan deneylerde, önerilen algoritmanın ROUGE-1 ve ROUGE-2 değerlerine dayanan performans kriterleri için diğer rekabetçi yöntemleri aştığı görülebilir (ROUGE-1, Geri Çağırma dışında). ROUGE-L, ROUGE-SU ve ROUGE-W-1.2 metrik türleri için de birçok en iyi değer rapor edilmiştir. Her satır için en yüksek değerler, sonuç çizelgelerinde kalın yazı tipiyle vurgulanmıştır.

Luhn [97], LSA [98][100], TextRank [101][99] LexRank [34] ve önerilen Karıcı Özetleme’ın DUC-2002 veri seti üzerinde gerçekleştirilen 200 kelimelik özetleri üzerinden elde edilen performans verilerine göre, önerilen yöntem bütün performans metrikleri ve türleri bakımından daha iyi sonuçlar rapor etmiştir. İlgili çizelge üzerinde satırlarda yer alan en yüksek değerler kalın rakamlarla vurgulanmıştır.

Yöntemin uygulanabilirliğini vurgulamak adına deneyler DUC-2002 veri setinde yer alan 400 kelimelik özetler üzerinde tekrarlanmış ve Çizelge 5.10’da raporlanan sonuçlar elde edilmiştir. Çizelgede her satırda yer alan en yüksek değerler kalın rakamlarla vurgulanmıştır. Deneysel süreçte önerilen yöntem, bütün kategoriler için en iyi değerleri

elde etmese de genellikle en iyi puanları elde etmiştir. Önerilen yöntem, DUC-2002 ve DUC-2004 veri kümelerinden 100, 200 ve 400 kelimelik özetler temelinde diğer mevcut yöntemlerle birlikte incelenmiştir. Seçilen son teknoloji yöntemler arasında, önerilen Karıcı Özetleme yöntemine benzer ve karşılaştırılabilir olan matematiksel, istatistiksel ve grafik tabanlı yaklaşımlar bulunmaktadır.

Çizelge 5.8, 5.9 ve 5.10 da sunulan sonuçlar dikkate alındığında uygulama kapsamında sunulan yöntem;

- 100 kelimelik özetler üzerinde diğer yöntemleri önemli ölçüde geride bırakmıştır.
- 200 kelimelik özetler bazında istisnasız bütün rekabetçi yöntemler geride bırakılmıştır.
- 400 kelimelik özetler ile gerçekleştirilen deneysel süreçlerde, bütün karşılaştırmalarda ROUGE-1 türünde en yüksek değerler rapor edilirken ROUGE-L, ROUGE-SU metrik türlerinde en yüksek değerlerin çoğunluğu rapor edilmektedir. Luhn, Karıcı Özetlemenin ardından en iyi ikinci performansa ulaşmaktadır.

Bölüm 5.5.2’de sunulan ilgili çizelgeler, önerilen Karıcı Entropi tabanlı, benzersiz yöntemimizin DUC-2002 ve DUC-2004 veri kümeleri için önceki çalışmalardan daha iyi performans verdiğini göstermektedir.

Çizelge 5.10. Rekabetçi yöntemler ile Karıcı Özetlemenin karşılaştırılması (400 kelimelik özetler).

400 kelimelik özetler (DUC-2002)						
ROUGE değerlendirme kriterleri	Metot					
	Luhn[97]	LSA [98, 100]	Textrank[101, 102]	Lexrank[34]	Random	SumBasic [103] KL-Sum[104] Karıcı ($\alpha = 0.4$)
Rouge-1	Rec.	0.57929	0.46361	0.57812	0.52799	0.43746 0.58669
	Prec.	0.57584	0.47293	0.52849	0.50893	0.43254 0.57857
	F-Sc.	0.57746	0.46804	0.55179	0.51819	0.43490 0.58241
Rouge-2	Rec.	0.32775	0.16001	0.25654	0.20698	0.16256 0.31804
	Prec.	0.32546	0.16297	0.23379	0.19967	0.16017 0.31369
	F-Sc.	0.32654	0.16141	0.24447	0.20322	0.16132 0.31576
Rouge-L	Rec.	0.55118	0.42283	0.56505	0.51566	0.48516 0.55652
	Prec.	0.54784	0.43129	0.51644	0.50089	0.46755 0.54885
	F-Sc.	0.54940	0.42685	0.53926	0.50793	0.47611 0.55248
Rouge-W- 1.2	Rec.	0.17758	0.12558	0.17498	0.16162	0.14844 0.17653
	Prec.	0.27834	0.20148	0.25234	0.24770	0.22572 0.27456
	F-Sc.	0.21675	0.15464	0.20648	0.19550	0.17905 0.21480
Rouge-SU	Rec.	0.29304	0.19823	0.31213	0.27351	0.24097 0.31021
	Prec.	0.28977	0.20537	0.26032	0.25871	0.22412 0.30195
	F-Sc.	0.29118	0.20140	0.28307	0.26541	0.24052 0.30564

5.7. ÇA-FGA : Çok amaçlı Fidan Gelişim Algoritması ile Çıkarıcı, Genel Metin Özetleme Önerisi

Geçmişte, çıkarıcı metin özetleme çalışmaları çerçevesinde özetle yer alacak cümlelerin seçimleri istatistiksel araçlar kullanılarak gerçekleştirilmekteydi. Ancak istatistiksel teknikler ile çıkarıcı metin özetleme sistemleri beklenen başarı düzeylerine ulaşamamaktadır. Doküman özetlemede belge seçiminin performansını arttırmak amacı ile son 20 yılda birçok araştırmacı metin özetleme problemini bir optimizasyon problemi olarak değerlendirmektedirler.

Bu çalışmada metin dokümanlarından maksimum kapsam ve fazlalık değerleri temelinde çok amaçlı bir fonksiyon kullanılarak çıkarıcı, genel bir özetleme sistemi tanıtılmaktadır. Doküman özetleme bir optimizasyon problemi olarak temsil edilmiştir. Seçilen algoritma FGA'dır. Tanımlanan optimizasyon problemi, sunulan ÇA-FGA ile çözülmüştür.

Ek olarak bu çalışmada özete seçilecek cümleler belirlenirken sadece dokümanın tamamı ile en yüksek kapsam değerine sahip cümlelere yer verilmemiştir. Özetle yer alan cümle ve kavram çeşitliliğini genişlemek amacı ile fazlalık parametresi de eş zamanlı olarak minimize edilmiştir. Bu duruma uygun bir modelleme yoluna gidilerek iki amaçlı bir fonksiyon oluşturulmuştur. Literatürde yer alan pek az çalışma gerçek bir şekilde eşzamanlı optimizasyon kullanarak özetleme gerçekleştirmektedir (Bknz. Çizelge 5.11). Çok amaçlı bir metin özetleme çalışması sunduğunu iddia eden çalışmalar, farklı zamanlarda çalışan tek amaçlı bir kaç optimizasyon sonucunun matematiksel olarak bir araya getirilmesinden ibarettirler.

Çizelge 5.1.1. Son teknoloji özetleme çalışmalarının detaylı analizleri.

Yazar/Yazarlar	Yıl	Özet Türü	Tek/Çok amaçlı	Eş zamanlılık	Özet Tipi	Kullanılan Algoritma	Kullanılan Kriter/Amaç
Zhao&Tang 2010[64]	2010	Çoklu	-	Hayır	Sorgu tabanlı	GA	uzunluk, kapsam,bilgi ,fazlalık
Kogilavani-Balasubr. [65]	2010	Çoklu	Çok amaçlı	Hayır	Genel	GA	uzunluk, kapsam, fazlalık.
Huang et al [66]	2010	Çoklu	Tek amaçlı	Evet	Sorgu tabanlı	-	kapsam, önem, fazlalık, tutarlılığı
Aliguliyev [67]	2010	Çoklu	Tek amaçlı	Hayır	Genel	PSO	Bölütleme
Nandhini-Balasundar. [68]	2014	Tek	-	Hayır	Genel	DE	bilgi ve bağlılık
Mendoza et al. [69]	2014	Çoklu	Tek amaçlı	Hayır	Genel	CHC greedy	kapsam, fazlalık
Alguliyev et al. [26]	2015	Tek, çoklu	Tek amaçlı	Hayır	Genel	DE	ilgi, fazlalık ve uzunluk
Mirshojaei-Masoomi [70]	2015	Tek	Tek amaçlı	Hayır	Genel	CSOA	
Rautray et al. [19]	2015	Tek	Tek amaçlı	Hayır	Genel	DE, PSO	fazlalık, maximizing relevancy
Umam et al.[71]	2015	Çoklu	Tek amaçlı	Hayır	Genel	self-adap. DE (SaDE)	kapsam, çeşitlilik,tutarlılık
Saleh et al. [72]	2015	Çoklu	Çok amaçlı	Evet	Genel	(NSGA-II)	kapsam,fazlalık azaltma
Sanchez-Gomez et al. [62]	2017	Çoklu	Çok amaçlı	Evet	Genel	ABC	kapsam, fazlalık azaltma
John et al.[23]	2017	Çoklu	Çok amaçlı	Hayır	Genel	GA	kapsam, fazlalık azaltma

Çizelge 5.11, optimizasyon teknikleri kullanılarak sunulan yöntemlere ve özelliklerine etraflıca yer vermektedir. Çizelgede ilişkili çalışmalara, yayınların yıllarına, tek/çoklu doküman bilgisine, tek/çoklu amaç fonksiyon bilgisine, optimizasyonun gerçekleşme şekline, özet tipi bilgisine, uygulanan algoritma bilgisine ve optimize edilecek kriter bilgilerine yer verilmektedir. Çizelgede yer verilen ilişkili çalışmaların büyük bir kısmı DUC veri setini kullanmaktadırlar. ROUGE metrikleri ile yaklaşımlarını değerlendirmektedirler.

Yaygın olarak kapsam, fazlalık, cümle uzunluğu, uyum, önem kullanılan kriterler arasındadır.

5.7.1 Benzerlik Tabanlı Cümle Puanlandırma

Dokümanlar üzerinde çalışmak için, dokümanları oluşturan cümleler formülize edilmelidirler. Vektör uzay modeli (VUM) cümleleri temsil etmek için kullanılan en yaygın yöntemdir. Her bir terim, özellik vektör uzayındaki bir boyutu temsil eder. $D=\{d1, d2,...,dN\}$ doküman kümesinin $S=\{s1,s2,...,sn\}$ cümlelerine ayrıştığı varsayalım. Burada N dokümanların, n ise cümlelerin sayısıdır ve s_i , S deki i 'ninci cümledir. $T=\{t1, t2,...,tm\}$ kümesi S deki bütün terimleri temsil eder. Burada m , farklı terimlerin sayısıdır. Her s_i cümlesi m boyutlu vektör uzayında bir nokta olarak konumlanır $Si=[w_{i1},w_{i2},...,w_{im}]$. W_{ik} ağırlığı s_i cümlesindeki tk terimi ile ilişkilidir ve ters cümle frekansı kullanılarak hesaplanır. Terim frekansı ters cümle-frekans şeması (tf_isf) kullanılarak hesaplanabilir. Bu tf ve isf nin birleşimidir, burada tf , bir terimin bir cümlede kaç defa görüldüğünü tutar. isf , doküman koleksiyonunun kaç cümlesinin terimi içerdiği bilgisini tutar.

Yerel Terim ağırlıklandırma: $tf_{ik}=m_{ik}/m_i$, burada m_{ik} s_i cümlesinde tk teriminin meydana gelme sıklığıdır. m_i ise s_i cümlesindeki terim sayısıdır.

Genel Terim ağırlıklandırma: $isf=log(n/n_k)$ burada n dokümandaki toplam cümle sayısıdır. n_k , tk teriminin görüldüğü cümle sayısıdır.

$$w_{ik} = tf_{ik} \cdot \log(n/n_k), \quad (5.7)$$

Burada W_{ik} , s_i cümlesindeki tk teriminin ağırlığıdır.

Ana dokümanın ağırlık merkezi ve aday özet bir vektör olarak temsil edilebilir. D doküman koleksiyonunun genel içeriği T deki m terimin ortalama ağırlığı ile özetlenebilir. Bu bir ortalama vektör hesabı ile yapılabilir. $o=(o_1,o_2,...,o_m)$ için;

$$o_k = \frac{1}{n} \sum_{i=1}^n w_{ij} \quad k=1,2,\dots,m. \quad (5.8)$$

$S_i = [w_{i1}, w_{i2}, \dots, w_{im}]$ ve $S_j = [w_{j1}, w_{j2}, \dots, w_{jm}]$ cümleleri arasındaki kosinüs benzerliği:

$$sim_{cos}(s_i, s_j) = \frac{\sum_{k=1}^m w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^m w_{ik}^2 * \sum_{k=1}^m w_{jk}^2}} \quad i, j=1, \dots, n. \quad (5.9)$$

5.7.2 Çok Amaçlı Optimizasyon Problemi Olarak Modellenmesi

Bu bölümde çoklu-doküman özetleme problem için çok amaçlı bir optimizasyon problemi sunulmakta ve ardından bu optimizasyon probleminin çözülmesi için benimsenen stratejiler açıklanmaktadır. Bu çalışmada $S=\{s_1, s_2, \dots, s_n\}$ cümle kümesine ait olan özet, mümkün olduğunca geniş kapsam ve minimum fazlalığa sahip olmalıdır. Kapsam, üretilen özetin mümkün olduğu kadar tüm alt konuları içermesi anlamındadır. Ancak tüm alt konuların kapsanması amacı ile sadece en yüksek kapsam değerlerine sahip olan cümlelere odaklanılması alt konuların özetle tekrarlanması anlamına gelecektir. Çalışma kapsamında alt konuların minimum tekrarının sağlanması amacı ile ikinci bir amaca gerek duyulmuştur. Çünkü zengin bir özet mümkün olduğunca tüm alt konuları kapsamalıdır ve mümkün olduğunca az tekrar içermelidir. Bu nedenden ötürü sadece kapsama değil eşzamanlı bir şekilde minimum fazlalık değerine de dikkate alındı. Bu niteliklere sahip bir özet kaliteli ve zengin bir özet olacaktır.

Maksimum Kapsam: Özeti kapsam değeri, orjinal belgelerdeki içeriği ne oranda kapsadığı bilgisini yansıtır. Umam ve arkadaşlarına [71] göre kapsam hesaplaması, aday özet dokümanı ile ana doküman arasındaki içerik benzerlikleri hesaba katılarak yapılmalıdır. Özetteki her cümle ile orjinal dokümanın ana içeriği arasındaki benzerliği hesaplanarak, her bir özet cümlelerinin asıl belgedeki önemi öğrenilebilecektir. Radev ve arkadaşlarına [106] göre doküman terimlerinin ağırlıklarının ortalaması veya dokümanın ağırlık merkezi, doküman kümesinin ana içeriğini yansıtır. Özeti kapsam değerini hesaplamak için:

$$\text{özet}_{kapsam}(x) = \sum_{i=1}^n \frac{sim(s_i, o) \cdot x_i}{n} \quad (5.10)$$

Formülde $x_i \in \{0,1\}$, özete ait ikili çözüm vektörü $x=(x_1, x_2, \dots, x_n)$ dir.

Minimum fazlalık: Özeti minimum fazlalık değeri, özetde yer alan cümlelerin yüksek çeşitliliği ile ilgilidir. Minimum fazlalık, özeti oluşturan her cümle arasındaki benzerliğin hesaplanması ile elde edilir. Özete ait toplam cümle benzerliği ne kadar yüksekse, o kadar düşük çeşitlilik (diversity) söz konusudur. Kaliteli bir özet için yüksek çeşitlilik gereklidir [71]. Bir başka ifade ile düşük çeşitlilik zayıf bir özet sunma potansiyeli taşımaktadır çünkü fazlalık (redundant) bilgi içermektedir. İyi bir özet için özeti oluşturan cümleler arasındaki cümle benzerlik değerlerinin minimum olması gerekmektedir. Özeti fazlalık değerini hesaplamak için:

$$\text{özet}_{\text{fazlalık}}(x) = \frac{1}{\left(\sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sim}(s_i, s_j) * u_i * u_j\right) * \sum_{i=1}^n X_i} \quad (5.11)$$

Formülde $u_i, u_j \in \{0,1\}, i,j=1,\dots,n$

Maksimum Kapsam ve Minimum fazlalık gibi farklı amaçların eşzamanlı optimizasyonları ile çoklu-doküman metin özetleme problem formülasyonu aşağıdaki gibidir:

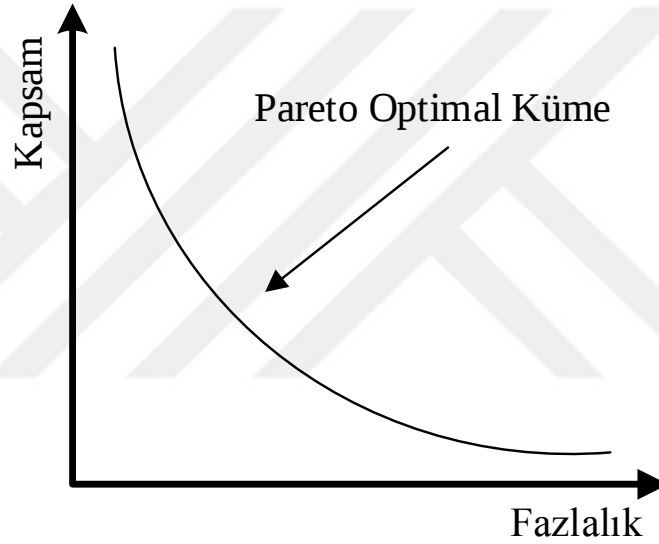
$$\begin{aligned} \text{maksimum } f(x) &= \{\text{özet}_{\text{kapsam}}(x), \text{özet}_{\text{fazlalık}}(x)\} \\ x_i &\in \{0,1\}, \quad i = 1, \dots, n \end{aligned} \quad (5.12)$$

5.7.3 Önerilen Çok Amaçlı Fidan Gelişim Algoritması

Özette yer alacak cümleler belirlenirken dokümanın tamamı ile en yüksek ilişki skorlarına (maksimum kapsam) sahip cümlelere özetde yer verilecektir. Dolayısı ile sadece en yaygın alt konuların içerilmesi eğilimi meydana gelecektir. Bu tarz bir özet gerçekten zengin bir özet olamayacaktır. Bu yüzden, bu çalışmada özeti oluşturacak cümlelerin minimum kavram örtüşmeleride (Minimum fazlalık) ilave bir amaç olarak düşünülmüştür. Uygun bir modelleme ile sistem başarımında olumsuz etkileri olan alt konu tekrarları minimuma indirgenmiştir. Çalışma kapsamında maksimum kapsam ve minimum fazlalık değerleri eş zamanlı bir biçimde optimize edilmiştir.

Optimizasyon problemini çözmek için FGA üzerinde uyarlamalar yapılarak problem çözülmüştür. Birçok durumda, çok kısıtlı karar problemlerinde tek bir en uygun çözüm bulunamaz. Bunun yerine, en az bir amaç için, diğer çözümlerden daha uygun değerlere sahip alternatif çözümler kümesi bulunur. Bu alternatif çözümler, Pareto Optimal çözümler

olarak bilinir [62, 63]. Bu yaklaşımda Pareto Optimum Küme'yi elde etmek için çözümler karşılaştırılır. Eğer bir çözüm, tüm amaçlar için diğer çözümleri bastırırsa, o çözüm diğerlerinden daha iyidir. Yani bu çözüm diğerlerinden daha optimum değerlere sahiptir. Yakın zamanda yapılan çalışmalar, pareto optimalin çok amaçlı optimizasyon problemlerinin çözümünde en etkili yöntem olduğunu ve diğerlerine üstünlüklerini göstermiştir [62, 72]. Çok amaçlı problemlerde, çoğu zaman amaçlar birbirleriyle çatışmakta ve bütün amaçları en iyileyen tek bir çözüm genelde bulunmamaktadır. Bu durumda, karar vericilerin çözümler üzerindeki tercihi önem kazanmakta olup, tüm çözüm uzayını temsil edebilecek uygun çözümlerin bulunmasında Pareto optimal başarıyla uygulanabilmektedir.



Şekil 5.5 . Pareto optimal küme.

Budama İşlemi : Budama adımı bu çalışma kapsamında ilk defa kullanılmaktadır. Bu aşamada fidan gövdesi üzerinde dallar budanarak çözüm kalitesi iyileştirilmektedir. Mevcut dallar (aday özette yer alan cümleler), ikili gruplandırılır. Grubu oluşturan çiftlerden zayıf olan dal budanır. Cümlelerin orjinal belge içeriğini kapsama miktarı bu aşamda belirleyici olmaktadır. Bu aşamada özete ait ikili (binary) çözüm vektörüne ($x=(x_1,x_2,...,x_n)$) yeni dallar (cümleler) eklenmemektedir. Budama aşması yerel bir arama işlemidir Dolayısı ile çözüm sayısında bir değişiklik meydana gelmemektedir.

Önerilen Budama operatörüne ait algoritma

1. summary_size=100 || 200 || 400
 2. **for** i=1 **to** len(cozum_matrisi) **do**
 3. binary_solutions =get_binary_solutions_indices(i th candidate_solutions)
 4. **While** candidate_solutions_length > summary_size
 5. j=random(binary_solutions)
 6. binary_solutions.remove(j)
 7. k=random(binary_solutions)
 8. binary_solutions.remove(k)
 9. **if** $sim(s_j, o) \geq sim(s_k, o)$
 10. candidate_solutions [k]=0
 11. **else**
 12. candidate_solutions [j]=0
 13. **end if**
 14. **end while**
 15. **end for**
-

Mutasyon işlemi : Mutasyon operatörü ile cümlelerin özette yer alıp almama koşuluna ilişkin değişiklikler yapılarak özet mutasyona uğratılır. Burada p_m ile mutasyon olasılığı kontrol edilir. Her cümle için benzersiz bir rastgele $r_i \sim [0,1]$ değişken üretilir. Eğer ($r_i \leq p_m$) koşulu sağlanırsa Denklem 5.13’da belirtilen ifadeye göre cümlelerin özette yer alıp almayacağına karar verilir.

$$sim(S_i, o) \geq \frac{1}{n} \sum_{j=1}^n sim(s_j, o) \quad (5.13)$$

Mutasyon operatörüne ait algoritma

1. $p_m=0.1$
 2. **for** j=1 **to** n **do**
 3. $r_i \sim [0,1]$
 4. **if** ($r_i \leq p_m$) **and** ($sim(S_i, o) \geq \frac{1}{n} \sum_{j=1}^n sim(s_j, o)$)
 5. $x_j=1$ //özetin ikili temsilindeki ilgili bit
 6. **else**
 7. $x_j=0$
-

-
8. **end if**
 9. **end for**
-

Onarım İşlemi: Oluşturulan özetin daha önceden belirlenen özet sınırları içerisinde kalmasını sağlamak için her çevrim sonunda kontrol edilmektedir. Denklem 5.14 ile kelime uzunluk sınırlamasının ihlal edilip edilmediği kontrol edilmektedir.

$$L - \varepsilon \leq \sum_{i=1}^n l_i \leq L + \varepsilon \quad (5.14)$$

Burada ε özet uzunluğunun tolerans katsayısıdır. l_i , ise s_i cümlesinin uzunluğudur. L istenilen özet boyutudur. Oluşturulan özet belirlenen sınırların altında bir boyuta sahipse önerilen yaklaşım kapsamında ilgili özet atılır. Eğer sınırların üzerinde bir boyut söz konusu ise özet atılmaz, bu durumda zengin bir özet için Onarım operatörü özeti modifiye eder. Özeti kalitesi noktasında burada esas belirleyici olan hangi cümle/cümlelerin çıkarılacak olduğudur. Çıkarılacak cümleler birbirlerine en benzer olan cümleler olmalıdırlar. Bir eşik değeri (threshold - δ) değeri belirlenir. Deneysel süreçler boyunca $\delta=0.3$ olarak seçilmiştir. Bu değerin üzerinde benzerlik taşıyan ikili cümle grupları özeten silinmeye aday cümlelerdir. Çünkü benzer cümlelerden birinin atılması özete ait kapsam değerini minimum miktarda düşürecektir (zengin bir özet maksimum kapsam değeri taşır, Bknz. Bölüm 5.6.2). S_i ve S_j cümleleri arasındaki benzerlik Denklem (5.15) teki gibi hesaplanmaktadır.

$$sim_{cos}(s_i, s_j) \geq \delta \quad (5.15)$$

Burada $i \neq j$, $S_i, S_j \in$ özet ve Denklem 5.15'de belirtilen koşulun sağlanması durumunda onarım operatörü çalıştırılır. Artık benzeyen cümlelerden hangisinin çıkarılacağı belirlenmelidir. $sim_{cos}(o^{sum}, o)$, içinde s_i cümlesinde bulunduğu özeti benzerliği, $sim_{cos}(o^{sum-s_i}, o)$ içinde s_i cümlesinin bulunmadığı özeti benzerliğidir. Böylelikle S_i cümlesinin özet üzerindeki etkisi belirlenir. Süreç aşağıdaki algoritmada belirtildiği gibi sürdürülür. S_i veya S_j cümlelerinin özet kalitesine olan etkisini vurgulamak için 10 katsayısı ile ağırlıklandırılmıştır.

Onarım operatörüne ait algoritma

1. **while** (özet_uzunluğu $> L + \varepsilon$)
 2. **for** i=0 to n **do**
 3. **for** j=1 to n **do**
-

```

4.   if  $Si, Sj \in \text{özet}$  and  $i \neq j$  and  $\text{sim}_{cos}(si, sj) \geq \delta$ 
5.        $\text{skor\_Si} = \text{sim}_{cos}(si, o) + ((\text{sim}_{cos}(o^{sum}, o) -$ 
         $\text{sim}_{cos}(o^{sum-si}, o)) * 10$ 
6.        $\text{score\_Sj} = \text{sim}_{cos}(sj, o) + ((\text{sim}_{cos}(o^{sum}, o) -$ 
         $\text{sim}_{cos}(o^{sum-sj}, o)) * 10$ 
7.       if  $\text{skor\_Si} \geq \text{skor\_Sj}$ 
8.           çıkar Sj
9.       else
10.          çıkar Si
11.       end if
12.   end for
13. end for
14. end

```

Bu tez çalışması kapsamında önerilerek ilk defa kullanılan çok amaçlı fidan gelişim algoritmasına ait adımlar aşağıdaki gibidir.

Çok amaçlı Fidan Gelişim Algoritmasına ait kabakod

```

1.    $t \leftarrow 0$  // başlangıç anı
2.   Fidanların_ekimi (G(t)) // bahçe (aday_çözümler)
3.   While  $t \leq \text{maksimum\_çevrim\_sayısı}$  do
4.        $G1(t) \leftarrow \text{eşleşme\_operatörü}(G(t))$ 
5.        $G2(t) \leftarrow \text{karıştır\&geliştir\_operatörü}(G1(t))$  // *2
6.        $G3(t) \leftarrow \text{budama\_operatörü}(G2(t))$ 
7.        $G4(t) \leftarrow \text{mutasyon\_operatörü}(G3(t))$ 
8.       amaç_fonksiyonu_hesapla((G4(t))
9.       pareto_ile_en_iyi_çözümleri_seç() // Pareto Front
10.       $G(t+1) \leftarrow \text{seçim}(G4(t)/2)$ 
11.       $t \leftarrow t+1$ 
12.  end while
13.  onarım_operatörü
14.  Özeti Oluştur

```

Öncelikle başlangıç durumu oluşturulur (bahçe). Başlangıç popülasyonu oluşturma adımı orjinal FGA’da önerilen biçimde gerçekleştirilir. Bu aşamada döküman koleksiyonundaki cümleler düzenli bir başlangıç popülasyonu oluşturarak seçilir (satır 2).

Oluşan başlangıç aday çözümleri FGA'nın doğası gereği rastgelelikten uzak ve oldukça muntazamdır.

Düzenli başlangıç çözümleri oluşturulduktan sonra ÇO-FGA'nın ana aşamaları belirlenen bir değer ölçüsüne tekrarlanır (3. ve 12. satırlar).

Eşleşme aşamasında, var olan geçici çözümler arasında (aday_çözümler) var olan bilgiler kullanılarak genel (global) bir arama biçiminde yeni fidanlar (aday_çözümler) üretilir (satır 4). Bu operatörde mevcut çözüm sayısı değişmez.

Önerilen ÇA-FGA'nda eşleşen fidanların çözüm kalitelerini iyileştirmek için mix&develop operatörü kullanılmaktadır (satır 6). Bu operatör S_i ve S_j fidanları arasındaki öklit uzaklığını kullanmaktadır. Bu durumda çözüm sayısı iki katına çıkmaktadır.

Sonraki adım budama (pruning) aşamasıdır (satır 6). Budama adımı bu çalışma kapsamında ilkdefa kullanılmaktadır. Bu aşamada fidan gövdesi üzerinde dallar budanarak çözüm kalitesi iyileştirilmektedir. Mevcut dallar (aday özette yer alan cümleler), ikili gruplandırılır. Sonrasında cümlelerin orjinal belge içeriğini kapsama miktarına bağlı olarak güçsüz dallar budanır. Bu aşamada özete ait ikili çözüm vektörüne ($x=(x_1,x_2,...,x_n)$) yeni dallar (cümleler) eklenmemektedir.

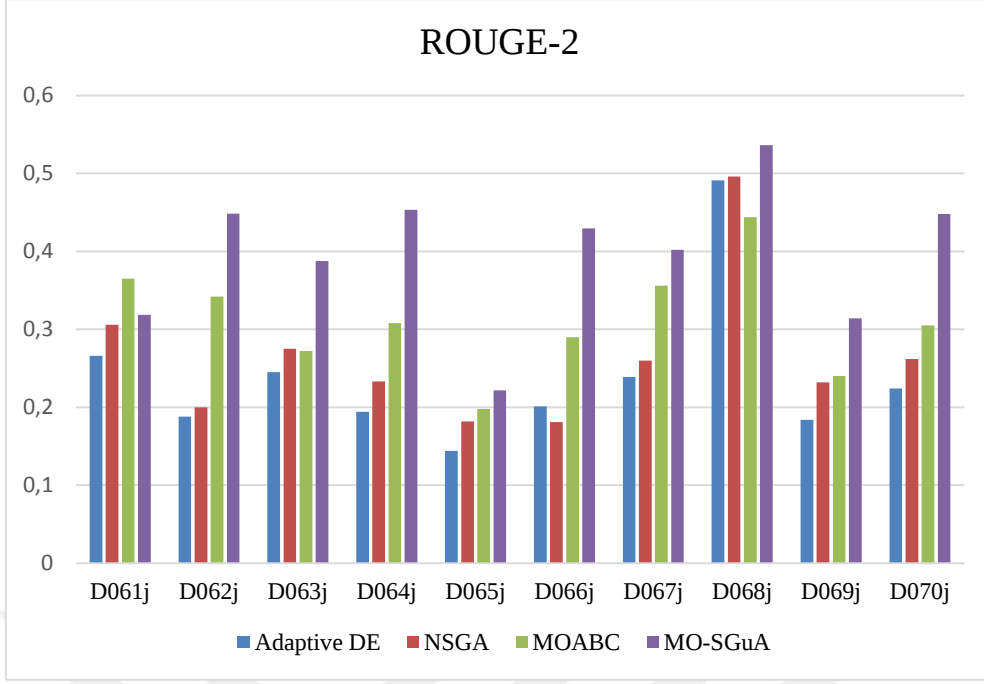
Mutasyon operatörü Bölüm 5'de belirtildiği gibi gerçekleştirilir (satır 7). Bu operatörde çözüm sayısı değişmez. Özete ait binary (ikili) çözüm vektörüne ($x=(x_1,x_2,...,x_n)$) cümle ekleyerek veya çıkararak, gelişim sağlanması amacı ile mutasyon uygulanır.

Yeni bireyler arasında, bir amaç değeri minimize olmadan (kötüleşme) diğer amaç değerinde bir artış (iyileşme) gerçekleşiyorsa yeni bireylerin seçimi gerçekleştirilecektir. Çözüm yalnızca orijinal olana hükmediyorsa, yani diğerlerini kötüleştirmeden bazı hedefleri iyileştirirse seçilecektir. Algoritmanın 9. satırında elde edilen çözümler içerisinde en iyi çözümler belirlenir ve seçilir. Bu işlem öncelikle pareto sets deki baskınlık ilişkilerine göre çözümleri sıralar ve seçer. Döngünün sonunda koloni boyutu orjinal boyuta döndürülür. Satır 13'de izin verilen kelime sınırlarının dışına bir taşma söz konusu ise Bölüm 6.3 de belirtildiği gibi Onarım operatörü uygulanır. Sonraki çevrim başlar.

5.7.4 Deneysel Sonuçlar

Deneysel süreçler boyunca maksimum 1000 döngü ve 20 tekrar gerçekleştirilmiştir. Önerilen ÇA-FGA performansı Çizelge 5.12 ve 5.13'te yer verilen ROUGE performans metrikleri kullanılarak değerlendirildi. İlgili çizelgeler her konu için $ROUGE_{ortalama}$, $ROUGE_{yayılma_aralığı}$ ve CV skorları için tüm konulara ait ortalama değerleri içermektedir (ROUGE-2, ROUGE-L). ÇA-FGA, oldukça iyi sayılabilecek performans verileri rapor etmektedir. Bu çalışma kapsamında önerilen ÇA-FGA ile elde edilen Özetleme performans sonuçları benzer rekabetçi yöntemler ile karşılaştırılmıştır. Daha önce gerçekleştirilen özetleme yaklaşımları ile uyumlu olması bakımından Çizelge 5.12 ve 5.13'de 200 kelimelik özetlere yer verilmiştir. Sonuçlar, benzer biçimde DUC-2002 veri setini kullanan Adaptive DE, NSGA-II ve MOABC ile karşılaştırılmıştır. NSGA-II ve MOABC literatürde var olan çok amaçlı iki yaklaşımdır. Literatürde başka çok amaçlı yöntemle rastlanmamıştır. Rekabetçi yöntemler ile uyumlu olacak biçimde ROUGE-2 ve ROUGE-L performans metrikleri kullanılmıştır. İlgili çizelgelerde ROUGE-2 ve ROUGE-L için $ROUGE_{ortalama}$, $ROUGE_{yayılma_aralığı}$ ve CV skorları yer almaktadır. Çizelge 5.12'de rapor edilen değerlere göre, ÇA-FGA diğer yöntemleri geride bırakmaktadır. Adaptive DE, NSGA-II ve MOABC bütün konular için ortalama, puan türünde geride bırakılmaktadır (MOABC'nin D061'i hariç). Çizelge 5.12'den yapılan bir başka gözlem ise Adaptive DE ve NSGA yaklaşımlarına göre on (10) dokümanın sekizinde (8) inde $ROUGE_{yayılma_aralığı}$ skorları bakımından daha iyi skorlar elde edildiğidir. Benzer şekilde CV puanları bakımından Adaptive DE ve NSGA dan çok daha iyi değerler alınarak tüm konularda daha üstün sonuçlar alınmıştır. Bu durum, önerilen yaklaşımın Adaptive DE ve NSGA yöntemlerine göre çok daha tutarlı ve kabul edilebilir olduğu ile açıklanabilir. ÇA-FGA ile tüm konular bazında en yüksek ortalama değerler rapor edilmiş, $ROUGE_{yayılma_aralığı}$ ve CV türlerinde ise en iyi ikinci değerler alınmıştır.

Şekil 5.6'da, DUC veri setinde bulunan konulara göre elde edilen ortalama ROUGE-2 puanları şematize edilmektedir. Şekilde görüldüğü üzere önerilen yaklaşım ile bütün konular bazında en yüksek değerler alınmıştır.



Şekil 5.6 . Rouge-2 ($ROUGE_{ortalama}$) puanları bakımından yöntemleri performans değerlerinin karşılaştırılması.

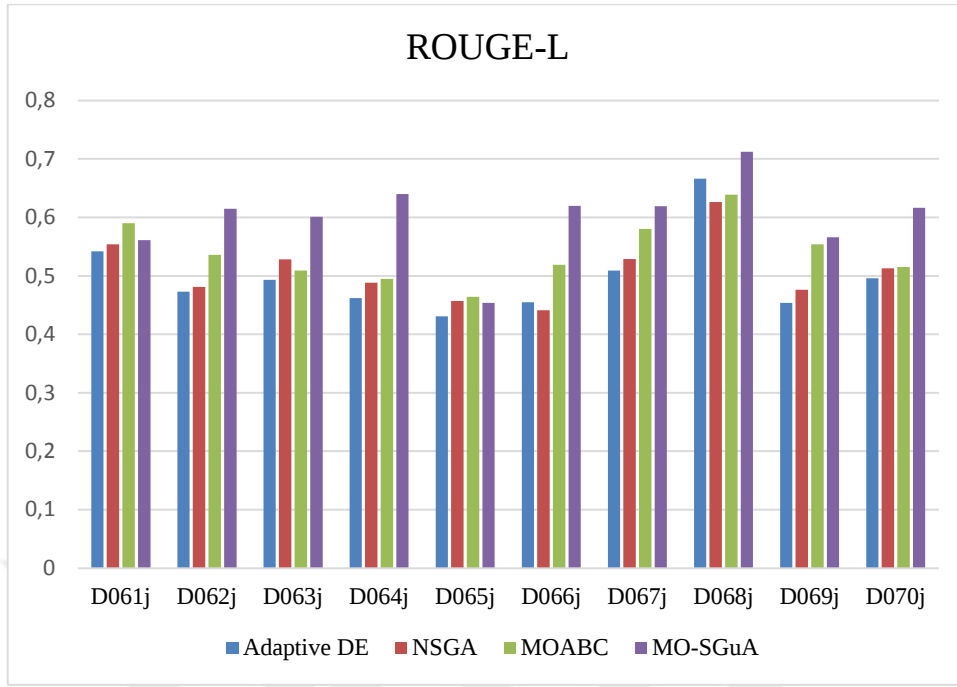
Çizelge 5.12’de, en yüksek on (10) $ROUGE_{ortalama}$ puanının sekiz (8) tanesinin ÇO-FGA na ait olduğu görülmektedir. $ROUGE_{yayılma_aralığı}$, CV skorları bakımından Adaptive DE ve NSGA yöntemlerinden çok daha sağlam ve tutarlı değerler alınmıştır. ROUGE-2’den sonra ROUGE-L puan türünden alınan sonuçlar önerilen yaklaşımın başarımını doğrulanmıştır.

Çizelge 5.12. Önerilen ÇA-FGA ile rekabetçi yöntemlerin karşılaştırılması.

200 kelimelik özetler (ROUGE-2)												
Doküman	Adaptive DE			NSGA-II			MOABC			Önerilen MO-SGuA		
	Ort.	Yayılma Aralığı	CV	Ort.	Yayılma Aralığı	CV	Ort.	Yayılma Aralığı	CV	Ort.	Yayılma Aralığı	CV
D061j	0.266	0.290	109.02	0.306	0.263	85.95	0.365	0.093	25.43	0.3186	0.1359	42.649
D062j	0.188	0.275	146.28	0.200	0.422	211.0	0.342	0.023	6.60	0.4483	0.1394	31.098
D063j	0.245	0.208	84.90	0.275	0.279	101.45	0.272	0.005	1.84	0.3876	0.6435	16.601
D064j	0.194	0.280	144.33	0.233	0.256	152.79	0.308	0.009	2.83	0.4531	0.1081	23.85
D065j	0.144	0.209	145.14	0.182	0.208	114.29	0.198	0.026	13.32	0.2215	0.1777	80.25
D066j	0.201	0.257	127.86	0.181	0.245	135.36	0.290	0.019	6.54	0.4294	0.1106	25.75
D067f	0.239	0.235	98.33	0.260	0.298	114.62	0.356	0.005	1.39	0.402	0.320	79.6
D068f	0.491	0.284	78.21	0.496	0.281	56.65	0.444	0.084	18.83	0.5362	0.2048	38.05
D069f	0.184	0.166	90.22	0.232	0.239	103.02	0.240	0.008	3.15	0.314	0.0854	27.2
D070f	0.224	0.260	116.07	0.262	0.215	82.03	0.305	0.002	0.76	0.4478	0.1407	31.436
Ortalama	0,238	0,256	114,04	0,263	0,280	115,72	0,312	0,027	8,07	0,395	0,2066	39,648

Çizelge 5.13. Önerilen ÇA-FGA ile rekabetçi yöntemlerin karşılaştırılması.

200 kelimelik özetler (ROUGE-L)									
Doküman	Adaptive DE			NSGA-II			MOABC		
	Ort.	Yayılma Aralığı	CV	Ort.	Yayılma Aralığı	CV	Ort.	Yayılma Aralığı	CV
D061j	0.542	0.208	38.38	0.554	0.205	37.00	0.590	0.056	9.46
D062j	0.473	0.239	50.53	0.481	0.306	63.62	0.536	0.017	3.26
D063j	0.493	0.156	31.64	0.528	0.171	32.39	0.509	0.050	9.74
D064j	0.462	0.235	50.87	0.488	0.287	58.81	0.495	0.011	2.19
D065j	0.431	0.141	32.71	0.457	0.174	38.07	0.464	0.057	12.29
D066j	0.455	0.196	43.08	0.441	0.149	33.79	0.519	0.007	1.36
D067f	0.509	0.232	45.58	0.529	0.244	46.12	0.580	0.012	2.15
D068f	0.666	0.226	33.93	0.626	0.226	36.10	0.639	0.071	11.09
D069f	0.454	0.135	29.74	0.476	0.191	40.13	0.554	0.010	1.81
D070f	0.496	0.173	34.88	0.513	0.158	30.80	0.515	0.005	0.90
Ortalama	0,498	0,194	39,13	0,509	0,211	41,68	0,540	0,03	5,425
							0,6003	0,1203	20,256



Şekil 5.7 . Rouge-L ($ROUGE_{ortalama}$) puanları bakımından yöntemleri performans değerlerinin karşılaştırılması.

Şekil 5.7’de DUC veri setinde bulunan konulara göre elde edilen ortalama ROUGE-L puanları şematize edilmektedir. Şekilde görüldüğü üzere önerilen yaklaşım, *D061j* dışında kalan tüm konularda en yüksek ROUGE-L değerlerini rapor etmektedir. Çizelge 5.14, ROUGE-2 ve ROUGE-L puanları için on (10) adet konuya dair $ROUGE_{ortalama}$ değişim miktarlarına ait yüzdelik değerleri içermektedir. Genel bir değerlendirme ile ÇA-FGA’ nın bütün ölçüm tiplerinde Adaptive DE ve NSGA-II yi geride bırakmaktadır. MOABC yaklaşımı ROUGE-2, $ROUGE_{ortalama}$ puan türünde %26,6 iyileşme sağlamaktadır. ROUGE-L, $ROUGE_{ortalama}$ puan türünde ise %11,17 lik bir iyileşme sağlanmaktadır. Ayrıca Çizelge 5.14 göstermektedir ki; ÇA-FGA, MOABC nin ardından Adaptive DE ve NSGA-II yi geride bırakarak en sağlam ve tutarlı değerleri elde etmektedir ($ROUGE_{ortalama}$ ve CV skorlarına göre).

Çizelge 5.14. Önerilen yaklaşımın diğer modeller ile genel karşılaştırılması.

Rekabetçi Yaklaşımlar	ÇA-FGA (%)					
	ROUGE-2			ROUGE-L		
	Avg.	Range	CV	Avg.	Range	CV
Adaptive DE	66,60	19,41	65,23	20,52	38,00	48,23
NSGA-II	50,68	-26,36	65,73	17,87	42,99	51,40
MOABC	26,87	-654,05	-391,36	11,15	-306,53	-273,38

6. SONUÇ VE ÖNERİLER

Çalışmada, metin dosyalarını temsil eden çizgeler üzerinde yapısal kararlılığı ve mevcut bilgi miktarını belirlemek amacı ile çizge entropi kullanılmaktadır. Çizgeyi oluşturan bütün düğümler sırasıyla çıkarılarak çizge ait toplam entropi hesaplanmaktadır. Çalışma kapsamında önerilen metin özetleme süreci gereği değerler küçükten büyüğe sıralanmıştır ve metinsel çizgeyi oluşturan düğümlere dair taşıdığı bilgi miktarı nicel olarak elde edilmiştir.

Birçok farklı çalışmada entropi türleri kullanılmasına rağmen giriş düzeyinde birkaç özetleme çalışması haricinde entropi tabanlı böylesine detaylı ve üstün sonuçların rapor edildiği çalışmalara rastlanmamıştır. Ayrıca sonuçlar göstermektedir ki entropi tabanlı metin özetleme yöntemi oldukça ümit verici bir tekniktir ve araştırmacıların dikkatini toplayacağı öngörülmektedir.

Kontrol dışı artan bilgiye erişim zamanının kısaltılması amacı ile verinin analiz edilerek bir takım özelliklerin elde edilmesi gerekmektedir. Bu zorlu görev otomatik özetleme sistemlerine olan ilgiyi her an arttırmaktadır. Bu nedenle tez kapsamında önerilen yaklaşımlar ve gerçekleştirilen uygulamalar çıkarıcı, denetimsiz, ve genel metin özetleme çerçevesinde gerçekleştirilmektedir.

Önerilen özetleme yaklaşımında, konuşma dilinin beraberinde getirdiği kurlsızlıklar önerilen yenilikçi bir yaklaşım ile giderilmiş ve geliştirilen yazılım aracı, kendine özgü algoritması ile metinleri özetleme için hazırlanmıştır. Bu yazılım aracı cümleler arasında anlamsal olarak ayırt edilebilir ve ölçülebilir ilişkilerin elde edilebilmesi bakımından oldukça olumlu etkiler göstermektedir.

Temsili çizgelerdeki ilk sonlu kümeyi cümleler temsil etmekte iken, diğer sonlu kümeyi ise, ortak kelime sayıları temsil etmiştir. Böylece metinlerin yönlendirilmemiş ve ağırlıklı çizgeler ile temsili sağlanmıştır.

Her bir düğüm ağdan sırayla çıkarılarak toplam Shannon ve Karcı Entropi değerleri hesaplandı. Çalışma kapsamında önerilen Karcı Özetleme süreci gereği değerler küçükten büyüğe sıralandı ve metinsel çizgeyi oluşturan düğümlere dair taşıdığı bilgi miktarı nicel olarak elde edildi. α üs değerinin değişimi Karcı Özetleme üzerinde Recall-Oriented Rouge-1, Rouge-2, Rouge-3, Rouge-L, Rouge-W-1-2, Rouge-SU performans metriği açısından etkisi ortaya konulmuştur. Deneysel süreçte α üs değerinin etkisi 100, 200 ve 400

kelimelik özetler için tekrarlanmıştır. DeneySEL süreçler boyunca aşağıdaki sonuçlar elde edilmiştir.

(1) 200 kelimelik özetler bazında elde edilen bütün performans metrik türlerinde rekabetçi yöntemler geride bırakılmıştır. 400 kelimelik özetler bazında ise rekabetçi yöntemlerin çoğunluğu geride bırakılmıştır. Kısmi olarak 400 kelimelik özetler, 200 kelimelik özetlere göre daha düşük performans göstermiştir.

(2) Karcı Entropi hesaplanırken ilgili denklemde yer alan α üs değeri nin özetleme performans değeri üzerinde doğrudan bir etkisi gerçekleşmiştir. Başlangıçtan itibaren kısmi artan bir grafik elde edilmiş, artan α değeri için çalışılan örneklemin sonlarına doğru sürekli bir azalış sergilendi ve örneklem üzerinde $\alpha=1.3$ için istisnasız minimum değeri rapor edildi. Bu basit ve etkili yaklaşım ile farklı α değeri için yerelden genele doğru farklı ROUGE değeri elde etmek mümkün olmuştur.

(3) Birçok farklı yaklaşımda Entropi türleri kullanılmasına rağmen giriş düzeyinde birkaç özetleme çalışması haricinde Entropi tabanlı böylesine detaylı ve üstün sonuçların rapor edildiği çalışmalara rastlanmamıştır. Daha önce hiçbir metin Özetleme çalışmasında Karcı Entropi kullanılmamıştır. Ayrıca sonuçlar göstermektedir ki Karcı Özetleme oldukça ümit verici bir tekniktir ve araştırmacıların dikkatini toplayacağı öngörülmektedir.

Tez kapsamında önerilen bir diğeri metin özetleme yaklaşımı ÇA-FGA olmuştur. Fidan gelişim algoritması ile defa bir metin özetleme çalışmasında kullanılmıştır. Ek olarak budama adımı bu çalışma kapsamında ilk defa fidan gelişim algoritmasına uyarlanmış ve algoritma bir adım ileri taşınmıştır. Bu aşamada fidan gövdesinde bulunan dalların belirlenen ölçütlere dayanılarak budanması prensibine dayanmaktadır. Bu sayede çözüm kalitesi iyileştirilebilmektedir. Mevcut dallar, ikili gruplandırılır. Sonrasında cümlelerin orjinal belge içeriğini kapsama miktarına bağlı olarak güçsüz dallar budanır.

İlgili çizelgelerde görüldüğü üzere literatürde sadece iki adet bulunan çok amaçlı metin özetleme yaklaşımı ile sunulan yöntem karşılaştırılmış ve ortalama puanları bakımından net bir üstünlük sağlanmıştır. Bu başarının fidan gelişim algoritmasının rastgelelikten uzak bir başlangıç çözümü oluşturmaya ve ilk defa tanıtılan budama aşaması ile ilişkili olduğunu düşünmekteyiz. Söz konusu uyarlamalar başarımlar üzerinde etkili olmuştur.

Optimizasyon bazlı doküman özetleme yaklaşımları zaman tüketimi ve performans noktasında bazı problemlerin giderilmesine ilaveten daha fazla formüle edilmeyede

ihtiyaç duymaktadırlar. Ayrıca, fazlalık azaltma, kapsam ve çeşitliliği en üst düzeye çıkarmak gibi ana hedefleri bakımından birbirlerinden ayrılırlar. Bu çalışmada kaliteli ve zengin içerikli özetler için birden fazla amaca odaklanılarak uygun ve yeterli formülasyon analizleri gerçekleştirilmiştir. Ek olarak kapsam ve çeşitlilik gibi farklı anlamsal amaçların eşzamanlı optimizasyonları ile ana metin içinden en uygun cümlelerin belirlendiği bir formülasyon kullanılmıştır.

Çok amaçlı optimizasyon problemlerinde araştırmacılar genellikle sabit ve değişken ağırlıklı amaç fonksiyonlarını kullanarak problemi tek amaçlı duruma getirerek çözümler elde etmişlerdir. Ancak bu yöntemler karar vericiye, problemi yetersiz bir şekilde modelleme, tek bir çözüm üretme, seçenekleri sınırlandırma gibi dezavantajlar getirmiştir. Bu çalışma kapsamında pareto optimal algoritması kullanılmıştır. Araştırmacılar için çok amaçlı optimizasyon problemlerinin çözümünde pareto optimal kullanmaları önerilmektedir.

7. KAYNAKLAR

- [1] C. Hark and A. Karci, *Karci summarization: A simple and effective approach for automatic text summarization using Karci entropy*, **Inf. Process. Manag.**, 57:3, (2020), doi: 10.1016/j.ipm.2019.102187.
- [2] Ş. Canberk, G. , Sağıroğlu, *Bilgi ve Bilgisayar Güvenliği : Casus Yazılımlar ve Korunma Yöntemleri*. Ankara: Grafiker Yayıncılık, 2006.
- [3] O. Durmaz, “*Metin Sınıflandırmada Boyut Azaltmanın Etkisi ve Özellik Seçimi*,” Gazi Üniversitesi, 2011.
- [4] C. Hark, A. Seyyarer, T. Uçkan, and A. Karci, *Doğal dil işleme yaklaşımları ile yapısal olmayan dökümanların benzerliği*, **IDAP 2017**, doi: 10.1109/IDAP.2017.8090306.
- [5] D. D. Lewis, *Naive (Bayes) at forty: The independence assumption in information retrieval*, **European conference on machine learning**, (1998) 4-15.
- [6] K. Aas and L. Eikvil, *Text categorisation: A survey. Technical report*, **Norwegian computing center**, (1999).
- [7] A. Pujari, *Data mining techniques*, 2001.
- [8] S. N. Bharath Bhushan and A. Danti, *Classification of text documents based on score level fusion approach*, **Pattern Recognit. Lett.**, 94 (2017), 118–126, doi: 10.1016/j.patrec.2017.05.003.
- [9] J. ge Yao, X. Wan, and J. Xiao, *Recent advances in document summarization*, **Knowl. Inf. Syst.**, (2017) 297–336, doi: 10.1007/s10115-017-1042-4.
- [10] L. Ermakova, J. V. Cossu, and J. Mothe, *A survey on evaluation of summarization methods*, **Inf. Process. Manag.**, 56:5 (2019) 1794–1814, doi: 10.1016/J.IPM.2019.04.001.
- [11] C. Hark, A. Seyyarer, T. Uçkan, and A. Karci, *Doğal dil işleme yaklaşımları ile yapısal olmayan dökümanların benzerliği*, **International Artificial Intelligence and Data Processing Symposium**, (2017) 1–6, doi: 10.1109/IDAP.2017.8090306.
- [12] İ. Tuğal and A. Karci, *Comparisons of Karci and Shannon entropies and their effects on centrality of social networks*, **Phys. A Stat. Mech. its Appl.**, (2019) 352–363, doi: 10.1016/J.PHYSA.2019.02.026.
- [13] A. Karci, *Fractional order entropy: New perspectives*, **Optik (Stuttg.)**, 127:20, (2016), 9172–9177, doi: 10.1016/j.ijleo.2016.06.119.
- [14] A. Karci, *New Approach for Fractional Order Derivatives: Fundamentals and Analytic Properties*, **Mathematics**, 4:2 (2016) 1–15.
- [15] D. Shi, J. Wang, D. Cheng, and J. Gao, *A global-local affinity matrix model via EigenGap for graph-based subspace clustering*, **Pattern Recognit. Lett.**, (2017) 67–72, doi: 10.1016/J.PATREC.2016.12.023.
- [16] C. E. Shannon, *A Mathematical Theory of Communication (Part I)*, **Bell Syst. Tech. J.**, 27:3, (1948) 379–423, doi: 10.1002/j.1538-7305.1948.tb01338.x.
- [17] A. Karci, *Notes on the published article ‘Fractional order entropy: New perspectives*, **Opt. - Int. J. Light Electron Opt.**, (2018) 107–108.
- [18] I. Tugal and A. Karci, *Centrality of Nodes with Karci Entropy*, **International Conference on Artificial Intelligence and Data Processing (IDAP)**, (2018) 1–7, doi: 10.1109/IDAP.2018.8620940.
- [19] R. Rautray and R. C. Balabantaray, *Cat swarm optimization based evolutionary framework for multi document summarization*, **Phys. A Stat. Mech. its Appl.**, (2017) 174–186, doi: <https://doi.org/10.1016/j.physa.2017.02.056>.
- [20] R. M. Alguliyev, R. M. Aliguliyev, N. R. Isazade, A. Abdi, and N. Idris, *COSUM: Text summarization based on clustering and optimization*, **Expert Syst.**, 36:1 (2019)

- doi: 10.1111/exsy.12340.
- [21] P. Verma and H. Om, *MCRM: Maximum coverage and relevancy with minimal redundancy based multi-document summarization*, **Expert Syst. Appl.**, (2019) 43–56, doi: 10.1016/j.eswa.2018.11.022.
 - [22] H. Grant, *Extractive Multi-document Summarization of News Articles*, **Linkiping University, Department of Computer and Information Science**, (2019).
 - [23] A. John, P. S. Premjith, and M. Wilscy, *Extractive multi-document summarization using population-based multicriteria optimization*, **Expert Syst. Appl.**, (2017) 385–397, doi: 10.1016/j.eswa.2017.05.075.
 - [24] C. Hark, T. Uçkan, and A. Seyyarer, Abubekir Karci, *Metin Özetleme İçin Çizge Tabanlı Bir Öneri*, **International Artificial Intelligence and Data Processing Symposium**, (2018) 1–6.
 - [25] A. Joshi, E. Fidalgo, E. Alegre, and L. Fernández-Robles, *SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders*, **Expert Syst. Appl.**, (2019) 200–215, doi: 10.1016/j.eswa.2019.03.045.
 - [26] R. M. Alguliyev, R. M. Aliguliyev, and N. R. Isazade, *An unsupervised approach to generating generic summaries of documents*, **Appl. Soft Comput. J.**, (2015) 236–250, doi: 10.1016/j.asoc.2015.04.050.
 - [27] R. M. Alguliyev, R. M. Aliguliyev, and N. R. Isazade, *A new similarity measure and mathematical model for text summarization*, **Probl. Inf. Technol.**, 6:1 (2015) 42–53.
 - [28] A. Joshi, E. Fidalgo, E. Alegre, and L. Fernández-Robles, *SummCoder: An Unsupervised Framework for Extractive Text Summarization Based on Deep Auto-encoders*, **Expert Syst. Appl.**, (2019), doi: 10.1016/j.eswa.2019.03.045.
 - [29] R. Mihalcea, *Language independent extractive summarization*, **ACL**, (2005) 49–52, doi: 10.3115/1225753.1225766.
 - [30] K. Sarkar, K. Saraf, and A. Ghosh, *Improving graph based multidocument text summarization using an enhanced sentence similarity measure*, **2015 IEEE 2nd Int. Conf. Recent Trends Inf. Syst. ReTIS 2015 - Proc.**, (2015) 359–365, doi: 10.1109/ReTIS.2015.7232905.
 - [31] J. Tan, X. Wan, and J. Xiao, *Abstractive Document Summarization with a Graph-Based Attentional Neural Model*, **In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics**, (2017) 1171–1181, doi: 10.18653/v1/P17-1108.
 - [32] I. Mani, *Advances in automatic text summarization*. MIT press, 1999.
 - [33] J. G. Carbonell and J. Goldstein, *The use of MMR, diversity-based reranking for reordering documents and producing summaries.*, **SIGIR**, (1998) 335–336.
 - [34] G. Erkan and D. R. Radev, *Lexrank: Graph-based lexical centrality as salience in text summarization*, **J. Artif. Intell. Res.**, (2004) 457–479.
 - [35] O. Kaynar, Y. Görmez, Y. E. Işık, and F. Demirkoparan, *Comparison of graph based document summarization method*, **International Conference on Computer Science and Engineering (UBMK)**, (2017) 598–603, doi: 10.1109/UBMK.2017.8093475.
 - [36] C. Hark, T. Uçkan, E. Seyyarer, and A. Karci, *An Unsupervised Approach Based on Node Centralization for Text Summarization*, **IDAP**, 8:3 (2019) 1109–1118.
 - [37] R. Rautray, R. C. Balabantaray, and A. Bhardwaj, *Document Summarization Using Sentence Features*, **Int. J. Inf. Retr. Res.**, 5:1 (2015) 36–47, doi: 10.4018/IJIRR.2015010103.
 - [38] C. Yadav, A. Sharan, M. Joshi, and C. S. Yadav, *Semantic graph based approach for text mining*, **International Journal of Information Retrieval Research**, 46:4 (2014) 46–70, doi: 10.1109/ICICICT.2014.6781348.

- [39] R. Shardan and U. Kulkarni, *Implementation and evaluation of evolutionary connectionist approaches to automated text summarization*, (2010).
- [40] P. G. Student and D. M. COE, *A Comparative Study Of Hindi Text Summarization Techniques: Genetic Algorithm And Neural Network*, (2015).
- [41] M. A. Fattah and F. Ren, GA, MR, FFNN, PNN and GMM based models for automatic text summarization, **Comput. Speech Lang.**, 23:1 (2009) 126–144.
- [42] M. A. Boudia, R. M. Hamou, and A. Amine, *A New Approach Based on the Detection of Opinion by SentiWordNet for Automatic Text Summaries by Extraction,* **Int. J. Inf. Retr. Res.**, 6:3 (2016) 19–36, doi: 10.4018/ijirr.2016070102.
- [43] C. S. Yadav and A. Sharan, *Hybrid Approach for Single Text Document Summarization Using Statistical and Sentiment Features*, **Int. J. Inf. Retr. Res.**, 5:4 (2015) 46–70, doi: 10.4018/ijirr.2015100104.
- [44] G. Salton and C. Buckley, *Term-weighting approaches in automatic text retrieval*, **Inf. Process. Manag.**, 24:5 (1988) 513–523, doi: 10.1016/0306-4573(88)90021-0.
- [45] O. Medelyan, *Computing Lexical Chains with Graph Clustering*, **In Proceedings of the ACL 2007 student research workshop**, (2007) 85–90.
- [46] Y.-N. Chen, Y. Huang, C.-F. Yeh, and L.-S. Lee, *Spoken Lecture Summarization by Random Walk over a Graph Constructed with Automatically Extracted Key Terms*, **In Twelfth Annual Conference of the International Speech Communication Association**, (2011).
- [47] L. Plaza, M. Stevenson, and A. Díaz, *Resolving ambiguity in biomedical text to improve summarization*, **Inf. Process. Manag.**, 48:4 (2012) 755–766, doi: 10.1016/j.ipm.2011.09.005.
- [48] S. Xiong and D. Ji, *Query-focused multi-document summarization using hypergraph-based ranking*, **Inf. Process. Manag.**, 2:4 (2016) 670–681, doi: 10.1016/j.ipm.2015.12.012.
- [49] R. Mihalcea and P. Tarau, *TextRank: Bringing Order into Texts*, **88AD**, (2004).
- [50] D. Parveen, H.-M. Ramsel, and M. Strube, *Topical coherence for graph-based extractive summarization*, **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, (2015) 1949–1954.
- [51] H. Van Lierde and T. W. S. Chow, *Query-oriented text summarization based on hypergraph transversals*, **Inf. Process. Manag.**, 56:4, (2019), 1317–1338, doi: 10.1016/J.IPM.2019.03.003.
- [52] Z. Zhang, S. S. Ge, and H. He, *Mutual-reinforcement document summarization using embedded graph based sentence clustering for storytelling*, **Inf. Process. Manag.**, 48:4, (2012), 767–778, doi: 10.1016/J.IPM.2011.12.006.
- [53] Miles Osborne, *Using Maximum Entropy for Sentence Extraction*, **Workshop on Automatic Summarization**, (2002), 1–8.
- [54] G. Ravindra, N. Balakrishnan, and K. R. Ramakrishnan, *Multi-document Automatic Text Summarization*, **In International Conference on Current Trends in Theory and Practice of Computer Science**, (2004), 289–300.
- [55] L. Ouyang and Q. P. Guo, *An entropy-based data summarization algorithm in data stream system*, **Pacific-Asia Workshop on Computational Intelligence and Industrial Application**, (2008), 2:872–876, doi: 10.1109/PACIIA.2008.132.
- [56] W. Luo, F. Zhuang, Q. He, and Z. Shi, *Effectively Leveraging Entropy and Relevance for Summarization*, **In Asia Information Retrieval Symposium**, (2010), 241–250, doi: 10.1007/978-3-642-17187-1_23.
- [57] A. Kennedy, T. Copeck, D. Inkpen, and S. Szpakowicz, *Entropy-based Sentence Selection with Roget’s Thesaurus*, **In TAC**, (2010).

- [58] R. M. Alguliev, R. M. Aliguliyev, M. S. Hajirahimova, and C. A. Mehdiyev, *MCMR: Maximum coverage and minimum redundant text summarization model*, **Expert Syst. Appl.**, 38:12, (2011), 14514–14522, doi: <https://doi.org/10.1016/j.eswa.2011.05.033>.
- [59] A. Qaroush, I. A. Farha, W. Ghanem, M. Washaha, and E. Maali, *An efficient single document Arabic text summarization using a combination of statistical and semantic features*, **J. King Saud Univ. - Comput. Inf. Sci.**, (2019), doi: <https://doi.org/10.1016/j.jksuci.2019.03.010>.
- [60] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, *Multiple documents summarization based on evolutionary optimization algorithm*, **Expert Syst. Appl.**, 40:5, (2013), 1675–1689, doi: <https://doi.org/10.1016/j.eswa.2012.09.014>.
- [61] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, *MR&MR-sum: Maximum relevance and minimum redundancy document summarization model*, **Int. J. Inf. Technol. Decis. Mak.**, 12:3, 2013, 361–393, doi: 10.1142/S0219622013500156.
- [62] J. M. Sanchez-Gomez, M. A. Vega-Rodriguez, and C. J. Pirez, *Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach*, **Knowledge-Based Syst.**, 159:2018, 1–8, doi: <https://doi.org/10.1016/j.knosys.2017.11.029>.
- [63] J. M. Sanchez-Gomez, M. A. Vega-Rodriguez, and C. J. Pirez, *Parallelizing a multi-objective optimization approach for extractive multi-document text summarization*, **J. Parallel Distrib. Comput.**, (2019), 166–179, doi: <https://doi.org/10.1016/j.jpdc.2019.09.001>.
- [64] X. Zhao and J. Tang, *Query-focused summarization based on genetic algorithm*, **Int. Conf. Meas. Technol. Mechatronics Autom.** (2010), 968–971, doi: 10.1109/ICMTMA.2010.429.
- [65] A. Kogilavani and P. Balasubramanie, *Clustering based optimal summary generation using Genetic Algorithm*, **Proc. 2010 Int. Conf. Commun. Comput. Intell. INCOCCI-2010**, (2010), 324–329.
- [66] L. Huang, Y. He, F. Wei, and W. Li, *Modeling Document Summarization as Multi-objective Optimization*, **Third International Symposium on Intelligent Information Technology and Security Informatics**, (2010), pp. 382–386, doi: 10.1109/IITSI.2010.80.
- [67] R. M. Aliguliyev, *Clustering techniques and discrete particle swarm optimization algorithm for multi-document summarization*, **Comput. Intell.**, 26:4, (2010), 420–448, doi: 10.1111/j.1467-8640.2010.00365.x.
- [68] K. Nandhini and S. R. Balasundaram, *Extracting easy to understand summary using differential evolution algorithm*, **Swarm Evol. Comput.**, (2014), 19–27, doi: 10.1016/j.swevo.2013.12.004.
- [69] M. Lozano, F. J. Rodríguez, and E. Herrera-Viedma, *A New Memetic Algorithm for Multi-Document Summarization based on CHC Algorithm and Greedy Search*, **Mex. Int. Conf. Artif. Intell.**, 8856:1, (2014), doi: 10.1007/978-3-319-13647-9.
- [70] S. H. Mirshojaei and B. Masoomi, *Text Summarization Using Cuckoo Search Optimization Algorithm*, **J. Comput. Robot.**, 8:2, (2015), 19–24.
- [71] A. goleman, daniel; boyatzis, Richard; Mckee, *Coverage, Diversity, and Coherence Optimization for Multi-Document Summarization*, **J. Chem. Inf. Model.**, 53:9, (2019), 1689–1699, doi: 10.1017/CBO9781107415324.004.
- [72] H. H. Saleh, N. J. Kadhim, and A. A. Bara’a, *A genetic based optimization model for extractive multi-document text summarization*, **Iraqi J. Sci.**, 56:2B, (2015), 1489–1498.
- [73] T. Carter. (2014).

- <https://www.csustan.edu/sites/default/files/CS/Carter/documents/info-lec.pdf> (online access on 16 Feb, 2019).
- [74] D. Bonchev, *Kolmogorov's information, Shannon's entropy, and topological complexity of molecules*, **Bulg. Chem. Commun.**, (1995) 567–582.
 - [75] S. Bouzebda and I. Elhattab, *New Kernel-type estimator of Shannon's entropy*, **Comptes Rendus Math.**, 352:1 (2014) 75–80, doi: 10.1016/j.crma.2013.11.011.
 - [76] T. Nie, Z. Guo, K. Zhao, and Z. M. Lu, *Using mapping entropy to identify node centrality in complex networks*, **Phys. A Stat. Mech. its Appl.**, (2016) 290–297, doi: 10.1016/j.physa.2016.02.009.
 - [77] C. E. Shannon, *Prediction and Entropy of Printed English*, **Bell Syst. Tech. J.**, 30:1 (1951) 50–64, doi: 10.1002/j.1538-7305.1951.tb01366.x.
 - [78] D. Hocaoglu, *Termodinamiğin İkinci Kanunu ve Entropi*, (2008).
 - [79] T. Balch, *Hierarchic social entropy: an information theoretic measure of robot group diversity*, **Auton. Robots**, 8:3 (2000) 209–237, doi: 10.1023/A:1008973424594.
 - [80] L. Fei and Y. Deng, *A new method to identify influential nodes based on relative entropy*, **Chaos, Solitons and Fractals**, (2017) 257–267, doi: 10.1016/j.chaos.2017.08.010.
 - [81] A. Mowshowitz and M. Dehmer, *Entropy and the complexity of graphs revisited*, **Entropy**, 14:3 (2012) 559–570, doi: 10.3390/e14030559.
 - [82] P. A. Bromiley, N. A. Thacker, and E. Bouhova-Thacker, *Shannon entropy, Renyi entropy and information*, **TINA Memos: Advanced Applied Statistics (2004-004)**, (2004).
 - [83] E. C. Kenley and Y.-R. Cho, *Entropy-Based Graph Clustering: Application to Biological and Social Networks*, **IEEE 11th International Conference on Data Mining**, (2011) 1116–1121, doi: 10.1109/ICDM.2011.64.
 - [84] A. Rényi, *On measures of entropy and information*, **Fourth Berkeley Symposium on Mathematical Statistics and Probability**, (1961) 547-1, doi: 10.1021/jp106846b.
 - [85] A. Karci, *Kesir Dereceli Türevin Yeni Yaklaşımının Özellikleri*, **J. Fac. Eng. Archit. Gazi Univ.**, 30:3 (2015) 487–501.
 - [86] A. Karci, *A New Approach for Fractional Order Derivative and Its Applications*, **Univers. J. Eng. Sci.**, 1:3 (2013) 110–117.
 - [87] M. R. Ubriaco, *Entropies based on fractional calculus*, **Phys. Lett. Sect. A Gen. At. Solid State Phys.**, 373:30 (2009) 2516–2519, doi: 10.1016/j.physleta.2009.05.026.
 - [88] J. Körner, *Coding of an information source having ambiguous alphabet and the entropy of graphs.*, **6th Prague Conf. Inf. theory**, (1973) 411–425.
 - [89] C. Hark, A. Seyyarer, T. Uçkan, and A. Karci, *Doğal dil işleme yaklaşımları ile yapısal olmayan dökümanların benzerliği*, **International Artificial Intelligence and Data Processing Symposium**, (2017), 1–6, doi: 10.1109/IDAP.2017.8090306.
 - [90] A. Karci, *Theory of Saplings Growing Up Algorithm*, **Adaptive and Natural Computing Algorithms**, (2007) 450–460.
 - [91] A. Karci and B. Alatas, *Thinking Capability of Saplings Growing Up Algorithm*, **Intelligent Data Engineering and Automated Learning**, (2006) 386–393.
 - [92] A. Karci, *Natural Inspired Computational Intelligence Method: Saplings Growing up Algorithm*, **IEEE International Conference on Computational Cybernetics**, (2007) 221–226, doi: 10.1109/ICCCYB.2007.4402038.
 - [93] P. Mehta and P. Majumder, *Effective aggregation of various summarization techniques*, **Inf. Process. Manag.**, 54:2 (2018) 145–158, doi: 10.1016/J.IPM.2017.11.002.

- [94] Document Understanding Conferences - Past Data.2019. <https://duc.nist.gov/data.html> (On-line acces on 23 Dec,2019)
- [95] P. Over, H. Dang, and D. Harman, *DUC in context*, **Inf. Process. Manag.**, 43:6 (2007) 1506–1520, doi: 10.1016/j.ipm.2007.01.019.
- [96] C.-Y. Lin, *Rouge: A package for automatic evaluation of summaries*, **Text Summ. Branches Out**, (2004).
- [97] H. P. Luhn, *The Automatic Creation of Literature Abstracts*, **IBM J. Res. Dev.**, 2:2, (1958) 159–165, doi: 10.1147/rd.22.0159.
- [98] T. K. Landauer and S. T. Dutnais, *A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge*, **Psychol. Rev.**, 104:2 (1997) 211–240.
- [99] R. Mihalcea and P. Tarau, *TextRank: Bringing Order into Texts*, **Proceedings of the ACL 2004 on Interactive poster and demonstration sessions** , (2004) , doi: 10.3115/1219044.1219064.
- [100] T. K. Landauer, P. W. Foltz, and D. Laham, *An introduction to latent semantic analysis*, **Discourse Process.**, 25:2 (1998) 259–284.
- [101] R. Mihalcea and P. Tarau, *A Language Independent Algorithm for Single and Multiple Document Summarization*, **2nd International Joint Conference on Natural Language Processing**, (2005) 19–24.
- [102] R. Mihalcea and P. Tarau, *TextRank: Bringing Order into Texts*, **In Proceedings of the 2004 conference on empirical methods in natural language processing**, (2004) 404–411, doi: 10.3115/1219044.1219064.
- [103] Lucy Vanderwende and Hisami Suzuki and Chris Brockett and Ani Nenkova, *Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion*, **Inf. Process. Manag.**, 43:6 (2007) 1606–1618.
- [104] A. Haghighi and L. Vanderwende, *Exploring content models for multi-document summarization*, (2009) 362, doi: 10.3115/1620754.1620807.
- [105] S. Brin and L. Page, *The anatomy of a large-scale hypertextual web search engine*, **Comput. networks ISDN Syst.**, 30:1–7 (1998) 107–117.
- [106] D. R. Radev, H. Jing, M. Styś, and D. Tam, *Centroid-based summarization of multiple documents*, **Inf. Process. Manag.**, 40:6 (2004) 919–938, doi: 10.1016/j.ipm.2003.10.006.

ÖZGEÇMİŞ

Ad Soyad : Cengiz HARK

Doğum Yeri ve Tarihi : Elazığ – 18.03.1984

Adres : Bitlis Eren Üniversitesi Enformatik Bölümü 13000
Merkez/BİTLİS

E-Posta : chark@beu.edu.tr, cengizhark@hotmail.com

Lisans : Fırat Üniversitesi Bilgisayar Mühendisliği (2008)

Yüksek Lisans : Fırat Üniversitesi Eğitim Bilimleri Enstitüsü Bilgisayar ve
Öğretim Teknolojileri

Mesleki Deneyimler : Workcube-İstanbul
Mühendis (2007-2009)

Bitlis Eren Üniversitesi
Öğretim Görevlisi (2009-Devam Ediyor)