T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

OTOMATİK METİN ÖZETLEME SİSTEMİ

AYSUN GÜRAN

DOKTORA TEZİ
MATEMATİK MÜHENDİSLİĞİ ANABİLİM DALI
MATEMATİK MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
YRD. DOÇ DR. NİLGÜN GÜLER BAYAZIT

İSTANBUL, 2013

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

OTOMATİK METİN ÖZETLEME SİSTEMİ

Aysun GÜRAN tarafından hazırlanan tez çalışması 26.02.2013 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Matematik Mühendisliği Anabilim Dalı'nda **DOKTORA TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Yrd. Doç. Dr. Nilgün GÜLER BAYAZIT

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Yrd. Doç. Dr. Nilgün GÜLER BAYAZIT

Yıldız Teknik Üniversitesi _____

Doç. Dr. Banu DİRİ

Yıldız Teknik Üniversitesi _____

Prof. Dr. Selim AKYOKUŞ

Doğuş Üniversitesi _____

Prof. Dr. Coşkun SÖNMEZ

Yıldız Teknik Üniversitesi _____

Doç. Dr. Olcay Taner YILDIZ

Işık Üniversitesi _____

Bu çalışma, Yıldız Teknik Üniversitesi Bilimsel Araştırma Projeleri Koordinatörlüğü' nün 2012-07-03-DOP01 numaralı projesi ve TÜBİTAK'ın Yurt İçi Doktora Burs Programı ile desteklenmiştir.

ÖNSÖZ

Teknolojinin ilerlemesi ile elektronik formda tutulan metinsel verilerin sayısı günden güne artmaktadır. Bu koşullar altında metinsel verilerden kullanışlı bilgiler elde etmenin ve bu süreçte az maliyetli ve yüksek başarımlı yöntemleri tercih etmenin önemi büyüktür. Bu önem kapsamında tez çalışmasında, "Otomatik Metin Özetleme" konusu incelenmiştir. Tezin temel amacı, metinlerdeki ana düşünceyi ve önemli bilgileri koruyacak şekilde verileri belirli özetleme oranları ile kısaltmak olmuştur. Tez çalışmasında bu amacı gerçekleştiren yöntemler incelenmiş ve çeşitli çözüm önerileri getirilerek araştırmacıların kullanımına sunulmuştur.

Çalışmam boyunca beni destekleyen tez yöneticim Yrd.Doç.Dr. Nilgün Güler Bayazıt'a, değerli yorumlarıyla desteklerini her zaman hissettiğim Prof.Dr.Selim Akyokuş, Doç.Dr. Banu Diri ve Doç. Dr. Olcay Taner Yıldız'a, yardımlarından ötürü sevgili çalışma arkadaşlarım Can Yalkın'a ve M. Zahid Gürbüz'e teşekkürlerimi sunarım. Ayrıca doktora süreci boyunca sağladığı destek nedeniyle TÜBİTAK kurumuna teşekkürü bir borç bilirim.

Bu tezi beni hayatım boyunca her zaman destekleyen sevgili eşim Barkan Güran'a, sevgili babalarım Harun Doğrusöz ve Ceyhan Güran'a, sevgili annelerim Meryem Doğrusöz ve Zeynep Güran'a, ablam Arife Karşıyakalılar'a, sevgili ağabeyim Güray Karşıyakalılar'a, biricik kardeşim Özge Doğrusöz'e ve son olarak ailemizin en küçük bireyi ve mutluluk kaynağımız sevgili Kaan Karşıyakalılar'a ithaf ediyorum. Sizler olmasaydınız bunu başaramazdım.

Şubat, 2013

Aysun GÜRAN

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ

A	Terim-doküman matrisi
$C_{önem}$	Bir terimin ait olduğu cümlenin önemi
$G(t_{ji})$	ji.terimin Global Ağırlığı
$L(t_{ji})$	ji.terimin Lokal Ağırlığı
M_i	Üçgensel Bulanık Sayı
N_{ipop}	Başlangıç popülasyonunun boyutu
N_{par}	Kromozom kodlamsında kullanılan başlangıç popülasyonu
N(A)	Boş uzay (null space)
$ö_{ij}$	Cümle önemini belirten özellikler
$P_{uygunluk}$	Genetik algoritmalarda kullanılan uygunluk fonksiyonu
R(A)	Satır uzayı (row space)
r	Bir matrisin rankı
S	TDA çarpanı olan köşegen matris
S_i	Sentetik boyut değeri
$tf(t_{ji})$	t_{ji} teriminin cümlede geçme sayısı
tf(maks)	İncelenen cümlede en çok geçen terim sayısı
$T_{Dağıtımsal}$	Terimin dağıtımsal özellikleri
$T_{Frekans}$	Terimin frekans özellikleri
U	TDA ile elde edilen sol çarpan matrisi
W_j	Melez sistemde kullanılan grup ağırlıkları
w_{ij}	Melez sistemde kullanılan grup içi özellik ağırlıkları
V	TDA ile elde edilen sağ çarpan matrisi
$Y_{Bulunma}$	Bölüm yoğunluğu
$Y_{ilk\text{-}Son}$	İlk ve son görünüm farkı yoğunluğu
Y_{pozVar}	Pozisyoların varyansı
$Yeni_{Ağırlık}$	Önerilen yeni ağırlık değeri
λ	Öz değer
σ	Tekil değer
$\mu_A(x)$	Üyelik fonksiyonu

KISALTMA LİSTESİ

AHS	Analitik Hiyerarşi Süreci
B	Birleşmiş Ağırlık
BAHS	Bulanık Analitik Hiyerarşi Süreci
BHÇGD	Biri Hariç Çapraz Geçerleme
DÖ	Dağıtımsal Özellikler
DUC	Document Understanding Conference
EVSD	Eğitim Ve Sınama Amaçlı kullanım Durumu
F	Frekansa Bağlı Ağırlık
GA	Genetik Algoritma
GAA	Gizli Anlamsal Analiz
GBAHS	Genelleştirilmiş Bulanık Analitik Hiyerarşi Süreci
GF	Göreceli Fayda
GK	Gerçek Kodlu
İ	İkili Ağırlık
İK	İkili Kodlu
JGAP	Java Genetik Algoritma Paketi
KAF	Keskinlik Anımsama F skor değeri
KCS	Kelime Cümle Skoru
L	Logaritmik Ağırlık
ODÖ	Ortalama Dağıtımsal Özellik
OMÖ	Otomatik Metin Özetleme
OKS-TDF	Ortalama Kelime Frekansı ve Ters Doküman Frekansı
ROUGE	Recall- Oriented Understudy for Gisting Evaluation
S	Sabit Ağırlık
T	Ters Doküman Sıklığı Bilgisi
TDA	Tekil Değer Ayrışımı

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

OTOMATİK METİN ÖZETLEME SİSTEMİ

Aysun GÜRAN

Matematik Mühendisliği Anabilim Dalı

Doktora Tezi

Tez Danışmanı: Yrd. Doç. Dr. Nilgün GÜLER BAYAZIT

Otomatik metin özetleme, bir bilgisayar programı aracılığı ile bir metnin özetlenmesi işlemidir. Bu işlem ile bilgisayara bir metin verilir ve bilgisayardan bu metne ait olan bir özet dokümanı alınır. Elde edilen özet dokümanı kullanıcıların inceledikleri metne ait olan ana temayı etkili bir şekilde anlamasını sağlar ve onların arama zamanını kısaltır.

Bir otomatik metin özetleme sistemi, çıkarıma ve yoruma dayalı olan özetleme görevlerini gerçekleştirebilir. Çıkarıma dayalı olan özetleme işlemi var olan cümleler arasından en önemli olanlarını seçmeye dayalı iken, yoruma dayalı olan özetleme işlemi yeni cümlelerin üretilme aşamalarını kapsamaktadır. Yoruma dayalı olan özetleme yaklaşımları dokümanların derinlemesine incelenmesini gerektirir. Yoruma dayalı olan özetleme yaklaşımlarının aksine, çıkarıma dayalı olan özetleme yaklaşımları daha pratiktir. Bu yaklaşımların çoğu incelenen dokümanları, dokümanlara ait olan cümlelerin önem derecelerinin cümle skoru fonksiyonlarıyla ifade edilmesini sağlayan bazı yapısal ve anlamsal özellikler ile temsil etmektedir.

Bu çalışma çıkarıma dayalı olan bir metin özetleme sistemi üzerinde yoğunlaşmıştır. Bu sistemde gizli anlamsal analiz temelli metin özetleme yöntemlerinde kullanılabilen yeni bir ağırlık değeri önerilmiştir. Önerilen yeni ağırlık değerine ait başarım sonucunun görülebilmesi için önerilen değer dört farklı gizli anlamsal analiz tabanlı yöntem üzerinde uygulanmış ve önerilen ağırlık değerinin tüm yöntem başarımlarını arttırdığı gösterilmiştir. Algoritmaların başarım analizleri insanlar tarafından oluşturulmuş olan dört farklı veri seti üzerinde analiz edilmiştir. Bu veri setlerinden ilk ikisi tez çalışması

için hazırlanan yeni Türkçe veri setleridir. Son iki veri seti ise sık kullanılan İngilizce veri setlerini içermektedir. Başarım ölçüm değeri olarak ilk üç veri seti için ideal ve otomatik özetler arasındaki çakışan cümle sayısına dayalı olan F-ölçüm skoru kullanılmıştır. Son veri seti için ise ideal ve otomatik olarak oluşturulmuş özetler arasındaki çakışan Ngram sayısına bağlı olan ROUGE değerlendirme paketi kullanılmıştır.

Tez çalışmasında ele alınan sistem aynı zamanda önemli cümle çıkarımı için yapısal ve anlamsal özelliklerin birleşimini sağlayan bir melez sistem önerisini de içermektedir. Önerilen sistem, içlerinden biri ilk kez tez çalışması kapsamında metin sınıflamadan metin özetlemeye adapte edilmiş olan, toplam on beş özelliği kapsamaktadır. Melez sistemde kullanılan özellikler iki farklı yaklaşım ile elde edilen ağırlıkların kullanılmasıyla birleştirilmiştir. Bu yaklaşımlardan ilki, özelliklerin ikili karşılaştırılmalarını içeren bir dizi uzman yargısına bağlı bir işlem olan bulanık analitik hiyerarşi sürecini kullanır. İkinci yaklaşım ise özellik ağırlıklarının otomatik olarak belirlenmesini sağlayan gerçek ve ikili kodlu genetik algoritmayı kullanmaktadır. Melez sisteminin başarım analizi Türkçe veri setleri üzerinde gerçekleştirilmiştir. Başarım ölçüm değeri olarak F-ölçüm skoru kullanılmıştır. Deneysel sonuçlar, özelliklerin birleştirilmesi suretiyle tüm özelliklerden yararlanılmasının, her bir özelliğin bireysel kullanımından daha iyi bir başarıma neden olduğunu göstermektedir.

Sonuç olarak bu tezde metin özetleme konusu ile ilgili bir çok yaklaşım önerilmiş ve araştırmacılar için kullanışlı sonuçlar elde edilmiştir. Bu tezin metin özetleme alanında hem Türkiye'de hem de Dünya'da yapılan çalışmalara katkıda bulunması dileğimizdir.

Anahtar Kelimeler: Türkçe metin özetleme, gizli anlamsal analiz, bulanık analitik hiyerarşi süreci, genetik algoritmalar

ABSTRACT

AUTOMATIC TEXT SUMMARIZATION SYSTEM

Aysun GÜRAN

Department of Mathematical Engineering

Phd. Thesis

Advisor: Assist. Prof. Dr. Nilgün GÜLER BAYAZIT

Automatic document summarization is a process where a computer summarizes a document. In this process, a document is entered into the computer and a summarized document is returned. The summarized document is extremely useful in allowing users to quickly understand the main theme of the whole document and effectively save their searching time.

ADS can perform extractive and abstractive summarization tasks. Extractive summarization techniques involve selecting the most important existing sentences, whereas abstractive summarization techniques involve generating novel sentences from given documents. The abstractive summarization approaches require a deeper understanding of the documents. In contrast to the abstractive summarization approaches, extractive summarization approaches are more practical. Most of them represent documents with some structural and semantic sentence features that indicate sentence importance using a sentence score function.

In this study, we focus on an extractive text summarization system. In this system we propose a new weighting scheme which can be used in Latent Semantic Analysis based text summarization methods. In order to see the performance of the proposed weighting scheme, we apply the new scheme on four different latent semantic analysis based summarization methods and we show that the proposed weighting factor makes improvements on all of the methods. The performance analysis of algorithms is

conducted on the human-generated extractive summary corpora that include four different data sets. The first two data sets are new Turkish data sets prepared for the thesis study. The last two data sets are the most common English data sets that are used in text summarization studies. As a performance measure, for the first three data sets, we use the F-measure score that determines the coverage between the manually and automatically generated summaries. For the last English data set, we supplemented the above metric with the ROUGE evaluation toolkit that is based on Ngram co-occurrence between the manually generated and automatically generated summaries.

The system also includes the proposal of a new hybrid system that combines structural and semantic sentence features used for important sentence extraction. The system employs fifteen features one of which is adapted from text categorization to text summarization for the first time. The features are combined by using weights calculated by two approaches. The first approach makes use of a fuzzy analytical hierarchical process which is a manual process that depends on a series of expert judgments based on pairwise comparisons of the features. The second approach makes use of the real and binary coded genetic algorithm for automatically determining the weights of the features. The performance analysis of hybrid system is conducted on the Turkish data sets. As a performance measure, we use the F-measure score that determines the coverage between the manually and automatically generated summaries. Experimental results show that exploiting all features by combining them resulted in a better performance than exploiting each feature individually.

Consequently, in this thesis many new approaches about text summarization subject have been proposed and useful results for researches have been produced. It is our wish that this thesis contributes to the studies about text summarization research areas in Turkey and the world.

Keywords: Turkish text summarization, latent semantic analysis, fuzzy analytical hierarchical process, genetic algorithm.

BÖLÜM 1

GİRİŞ

1.1 Literatür Özeti

İşletmeler ve kurumlar, kuruluş amaçlarına göre veritabanlarında çeşitli metinsel verileri depolamaktadırlar. Hızla büyüyen bilgi teknolojileri ışığında bu verilerden sadece en gerekli olan bilgileri elde edebilmenin önemi oldukça fazladır. Bu sebeple günümüzde uzun metinlerin analizini sağlayarak onları daha anlaşılır hale getiren "Metin Madenciliği" alanı önemli bir disiplin haline gelmiştir. Tez çalışmasında metin madenciliğinin bir alt konusu olan ve metinlerdeki önemli bilgilerin bir uzman gücüne gerek duyulmadan belirlenmesini sağlayan "Otomatik Metin Özetleme-OMÖ" konusu incelenecektir.

OMÖ, bir bilgisayar programı aracılığı ile bir metnin özetlenmesi işlemidir. Bu işlem ile bilgisayar programına giriş elemanı olarak bir metin verilir ve programdan çıkış elemanı olarak bu metne ait olan önemli bilgileri barındıran bir özet dokümanı alınır. Özet dokümanının bir program aracılığıyla otomatik olarak elde edilmesi kullanıcılara ciddi bir vakit kazandırmakta ve incelenen metinlerin ana temalarının daha etkili bir şekilde anlaşılmasını sağlamaktadır. OMÖ sistemlerinde ihtiyaca göre farklı türlerde özetler elde edilebilir: Bir OMÖ sisteminde elde edilecek olan özetler, "tek bir kaynaktan veya birden fazla kaynaktan çıkarılmış", "tek dil veya birden fazla dil ile yazılmış", "yoruma dayalı veya çıkarıma dayalı olan", "genel veya sorgu tabanlı", "gösterici veya bilgi verici" olabilir. Literatürde bu tarz özetlerin çıkarılmasını amaçlayan çok sayıda bilimsel çalışma mevcuttur. Bu bölümde metin özetleme konusu genel hatlarıyla tanıtılacak ve bu konu ile ilgili yapılan bilimsel çalışmalar detaylı bir şekilde ele alınacaktır.

1.1.1 Metin Özetleme Nedir?

Özet bir ya da birkaç dokümandan çıkarılmış olan, dokümana ait en gerekli bilgileri içeren ve dokümanın yarısından daha uzun olmayan metin parçasıdır. Özetleme işlemi ise kaynak dokümana ait en gerekli bilgileri içerecek şekilde yeni bir doküman yaratma işlemidir. Bu işlem bilgisayar tarafından belirli bir yazılım dili ile gerçekleştirildiğinde OMÖ ismini almaktadır. OMÖ işleminde geleneksel olarak sistem girdisi bir metindir. Özetleme işleminde ise girdi bir görüntü, video yada ses olabilmektedir.

Özetlenecek doküman sayısına göre tekli doküman özetleme (single document summarization) veya çoklu doküman özetleme (multiple document summarization) işlemlerinden bahsetmek mümkündür. Tekli doküman özetlemede bir tane kaynak doküman mevcutken, çoklu doküman özetlemede birbirleri ile ilgili olan birden fazla kaynaktan yararlanılmaktadır. Özeti çıkarılacak olan kaynak metin bir dil (monolingual) ile yada farklı dilleri (multilingual) içerecek şekilde yazılmış olabilir.

Özetleme sisteminin çıktısı yoruma dayalı (abstractive) yada çıkarıma dayalı olan (extractive) bir özet olabilir. Yoruma dayalı olan özetleme, özetlenecek metnin akıllıca yorumlanması ile yapılır. Bu özetlemede orijinal metindeki ifadeler akıllı bir şekilde kısaltılarak tekrar yazılmaya çalışılır. Örneğin, “Özge masaya oturdu, menüyü okudu, yemeğini yedi ve gitti” cümlesinin yorumlanmış hali “Özge restorana gitti” olacaktır. Birebir cümle seçimine dayalı olan özetlemede ise özetlenecek metindeki önemli cümleler, istatistiksel yöntemlerle, sezgisel çıkarımlarla veya bu yöntemlerin birleşimiyle seçilmektedir. Bu tarz özetlemede, özeti oluşturan cümleler yazı içinden olduğu gibi seçilmiş olan cümlelerdir.

Bir özet genel (generic) yada kullanıcıya yönelik (user-directed) olabilir. Bu iki kavram özetin etki alanı ile ilgilidir. Genel özet metnin ana temalarıyla ilgili olan ayrıntılı özettir. Kullanıcıya yönelik özet ise kullanıcının yazdığı sorgu ile ilgili olan özettir.

Çıktının stili dikkate alınarak bir özet gösterici (indicative) veya bilgi verici (informative) olabilir. Gösterici özette bir doküman içinde bahsi geçen genel başlıklar belirlenmektedir. Bilgi verici özette ise kullanıcının isteğine bağlı olan başlıklarla ilgili cümleler seçilmektedir.

Tez çalışmasında tek kaynaklı, çıkarıma dayalı ve genel özetlerin çıkarıldığı bir sistem üzerinde çalışılacaktır.

1.1.2 Metin Özetleme Yaklaşımları

Metin özetleme çalışmalarına 50 yıl kadar önce başlanmış olsa da, bu alandaki çalışmalar son zamanlarda gelişen teknoloji ve dil bilimsel çalışmaların ışığı altında popüler olmaya devam etmektedir. Bu bölümde metin özetleme alanında yapılmış olan akademik çalışmalar incelenecektir. Öncelikle İngilizce metinler üzerinde çalışan bilimsel yayınlar ele alınacak, daha sonra Türkçe metinleri kullanan çalışmalardan bahsedilecektir. İncelenen tüm çalışmalar yapılmış oldukları yıllara göre Ek A'da belirtilen çizelge ile sunulacaktır.

1.1.2.1 İngilizce Metinleri Kullanan Bilimsel Çalışmalar

Bir özetin otomatik olarak çıkarılması için özetin çıkarılacağı dokümana ait cümleler arasından en çok bilgiyi taşıyanlar tespit edilmelidir. Literatürde bu tespiti gerçekleştirmek için metinlerin yapısal veya anlamsal özelliklerini inceleyen çalışmalar mevcuttur. Bu yaklaşıma sahip olan ve İngilizce metinler üzerinde çalışan ilk yol gösterici çalışma Luhn [1] çalışmasıdır. Bu çalışmada cümleler terim frekanslarına göre puanlandırılmıştır. Çalışmaya göre bir metinde en sık geçen terimler günlük hayatta sık kullanılan terimlerdir ve bu terimler genelde içerik belirtmemektedir. Bu sebeple Luhn çalışmasında terimlerin yüksek sıklıktaki değerleri için bir kesme değeri belirlemiş ve bu değerin üzerinde olan terimlerin alınmamasını önermiştir. Çalışmada benzer şekilde bir alt kesme değeri de belirlenmiş ve bu değerin altındaki terimler de dikkate alınmamıştır. Kesme değerlerinin belirlenmesinin ardından, her cümle en az bir önemli terim ve dörtten fazla önemsiz terim içermeyecek şekilde parçalara bölünmüş ve her parçadaki önemli sözcük sayısının karesi parçadaki toplam sözcük sayısına bölünmüştür. Sonuçta en yüksek puana sahip parçanın puanı, cümlenin puanı olarak seçilmiş ve yüksek puanlı cümleler özete eklenmiştir. Edmunson [2], Luhn [1] çalışmasındaki kelime sıklığı bilgisine ek olarak "ipucu sözcük öbekleri", "başlık terimleri" ve "cümle konumu" gibi üç yeni özellikten bahsetmiştir. Edmunson sistemine göre belirtilen özelliklere ait tüm ağırlık değerleri eğitim ve test amaçlı kullanılan bir

veri seti üzerinde hesaplanarak lineer bir fonksiyonda parametrik hale dönüştürülmüştür. Sonuç olarak çalışmada ipucu, başlık ve konum yöntemlerinin birleşimi en yüksek başarım oranını vermiştir.

Literatürde bir cümlenin önemini tespit etmek adına kullanılmış olan yapısal özellikler: "cümle uzunluğu", "cümlenin göreceli uzunluğu", "ipucu sözcük öbekleri", "pozitif ve negatif sözcük öbekleri", "ünlem, soru işareti ya da tırnak işareti gibi vurgu belirten bazı noktalama işaretleri", "cümle konum bilgisi", "kelime konum bilgisi", "başlık kelimeleri", "tarih bilgisini belirten ifadeler", "konuya has sözcükler", "önemli Ngramlar", "metin içindeki isimler ya da nümerik karakterler", "cümlenin merkeziliği", gibi özelliklerdir. Bu özelliklerin kullanıldığı çalışmalar: Pollock ve Zamora [3], Kupiec vd. [4], Pardo vd. [5], Yeh vd. [6], Hernandez ve Ledeneva [7], Quyang vd. [8]; Radev vd. [9] çalışmalarıdır.

Yapısal özellikler farklı yöntemler ile birleştirilerek cümlelere puan verme işlemi bir melez sisteme göre verilebilir. Kaini ve Akbarzadeh [10], Suanmali vd. [11], Kyoomarsi vd. [12], Binwahlan vd. [13] çalışmaları yapısal özelliklerin birleşimini bulanık mantık tabanlı bir melez sistem ile gerçekleştirmişlerdir. Yapısal özelliklerin kullanıldığı sistemlerde cümleye puan verme işlemi genetik (genetic) ya da sürü tabanlı (swarm-based) algoritmalar gibi sezgisel (heuristic) yaklaşımlar kullanırak da yapılabilir. Belirtilen tarzdaki algoritmaların uygulanmış olduğu çalışmalar Silla vd. [14], Witte vd. [15], Binwahlan vd. [13], Berker ve Güngör [16] referansları ile belirtilen çalışmalardır.

Son zamanlarda metin özetleme problemi bir optimizasyon problemi olarak ele alınmıştır. Filatova vd. [17] metin özetlemeyi maksimum kavrama problemi (maximum coverage problem) olarak incelemiştir. McDonald [18] çalışması metin özetlemeyi sırt çantası (knapsack problem) problemini çözme amaçlı kullanılan dinamik programlama modeli olarak ele almış Alguliev vd. [19] ise metin özetlemeyi lineer olmayan tamsayılı programlama problemi olarak incelemiştir.

Literatürde metin özetleme işlemini makine öğrenmesi tekniklerini kullanılarak gerçekleştiren çalışmalar da mevcuttur: Copeck vd. [20], Svore vd. [21], Wong vd. [22], Hirao vd. [23], Lal ve Reuger [24]. Bu çalışmalarda cümlelerin önemini belirten özellikler bayes sınıflandırıcı, destek vektör makineleri (support vector machines),

yapay sinir ağları (artificial neural networks) gibi tekniklerle birlikte kullanılmıştır. Bazı çalışmalarda (Hernandez ve Ledeneva [7]), cümleler K-ortalamalar (K-Means) yöntemiyle gruplanmış ve metin özetlerinin çıkarılması için oluşturulan gruplar içindeki en önemli cümleler seçilmiştir.

Yapısal özelliklere dayalı olan yöntemler kelimeler arasındaki anlamsal ilişki durumunu incelememektedir. Bu bağlamda uyum (choesion) tabanlı yaklaşımlar önerilmiştir. Bu kategori altında gönderimsel ifadelerin (anaphoric expressions) analiz edildiği çalışmalar (Brandow vd. [25], Steinberger vd. [26], Orasan [27]) mevcuttur. Gönderimsel ifadeler daha önce bahsi geçen bir kelime yada söz öbeğine gönderme yapan zamirlerdir. Benzer bir yaklaşım ortak referans çözünürlüğü (co-reference resolution) yaklaşımıdır (Azzam vd. [28], Baldwin vd. [29], Branimir vd. [30]). Yine uyum tabanlı yaklaşımlarda, sözlüksel bağlantıların (lexical chains) işlendiği çalışmalar (Barzilay vd. [31], Karamüftüoğlu [32]) vardır. Bu çalışmalar kelimeler arasındaki uyumsal ilişkiyi WordNet eşanlamlılar sözlüğünü kullanarak kurmuşlardır. WordNet [33], 1985 yılında geliştirmeye başlanan ve 2006 yılında da son sürümü olan 3.0'ın yayınlandığı, İngilizce kelimelerden oluşan sözcük veritabanıdır. İçeriğinde 155,287 adet İngilizce kelime barındırır. Bu kelimeler kendi içlerinde eş anlamlılar olarak gruplanmıştır ve bu grupların her birine “eş anlamlılar grupları (synset)” ismi verilmiştir. Toplamda 117,659 adet eş anlamlılar grubu içeren bu veritabanının içinde 206,941 adet de kelime anlam çifti bulunmaktadır. Sözlüksel bağlantıların (lexical chains) işlendiği çalışmalarda kelimelerin WordNet’deki ilişki durumlarına bakılarak sözlüksel zincirler oluşturulur ve dokümandaki konu başlıkları belirlenir. Cümle seçimi güçlü sözlüksel zincir içeren cümlelerin belirlenmesi ile yapılmıştır. Kan ve McKeow [34] özetleme sistemlerinde varlık ismi tanıma (name entity recognation) ve çoklu kelimelerin (multiwords) tespiti işlemleri ile bilgi çıkarımı ve cümle çıkarımı tekniklerini birleştirmiştir. D’Avanzo [35], anahtar kelime çıkarımı, varlık ismi tanıma ve çoklu kelimelerin tespitini yapan bir sisteme sahiptir. Hovy ve Lin [36] kelimeler arasındaki anlamsal ilişkileri belirlemenin yanında doğal dil işleme tekniklerini kullanan ve farklı dilleri içeren bir sistem geliştirmişlerdir. Jin ve McKeow [37] yine doğal dil işleme teknikleri sayesinde cümle birleştirme ve düzenleme işlemlerini gerçekleştirmişlerdir.

Metin özetleme alanında yapılan çalışmalar incelenmeye devam edildiğinde doküman içindeki nedensellik olayını irdeleyen olay tabanlı (event-based) özetleme yöntemlerinin önerildiği görülür (Liu vd. [38], Filotova vd. [39]). Bu yöntemlerde öncelikle giriş elemanı olan metin cümle ve paragraflara ayrılır. Bu metin parçalarının kapsadığı olaylar bulunur. Bu işlem yapılırken olay-terim grafiği çıkarılır ve grafikte bir olayı belirten ilgili terimler belirli kümeler altında gruplanır. Özete eklenecek metin parçaları bu kümelere göre seçilir ve özet oluşturulur.

Literatürde kelimelerin birlikte geçme sıklıklarına göre gruplanmasıyla doküman içindeki gizli anlamsal yapıyı tespit etme olanağı sağlayan lineer cebir yöntemlerine dayalı olan özetleme yaklaşımları mevcuttur. Bu yöntemler, bir dokümanı terim-cümle matrisi şeklinde ifade ederek Gizli Anlamsal İndeksleme (Latent Semantic Analysis), Olasılıksal Gizli Anlamsal İndeksleme (Probabilistic Semantic Analysis) veya Negatif Olmayan Matris Ayrışımı (Nonnegative Matrix Factorization) gibi teknikleri kullanmaktadır. Bu tekniklerden gizli anlamsal analizi uygulayan çalışmalar: Gong ve Lui [40], Murray vd. [41], Steinberger vd. [42], Yeh vd. [6]; olasılıksal gizli anlamsal analizi uygulayan çalışma Bhandari vd. [43] ve negatif olmayan matris ayrışımını uygulayan çalışmalar: Lee vd. [44] ve Mashechkin vd. [45]'dir. Bu teknikler ile terim-cümle matrisleri çarpanlarına ayrıştırılır. Bu çarpan matrisleri ile terimler ve cümleler lineer bağımsız vektörler ile indekslenirler. Lineer bağımsız olan bu vektörler terim ve cümleleri anlamsal olarak kümelemektedirler. Bu tekniklerin uygulandığı metin özetleme aşamalarında önemli cümleler, terim ve cümlelerin anlamsal olarak kümelenmesini sağlayan çarpan matrisleri kullanılarak seçilmektedir.

1.1.2.2 Türkçe Metinleri Kullanan Bilimsel Çalışmalar

Türkçe metinler ile çalışan ilk özetleme metodu Altan [46] çalışmasına aittir. Bu çalışmada ekonomi alanına ait olan 50 doküman ile çalışılmıştır. Önerilen sistemde beş bölüm mevcuttur. İlk bölümde metinler HTML etiketleri yardımıyla başlık, cümle ve paragraflara ayrılmıştır. İkinci bölümde terim sıklığı bilgisi ve cümlenin konum bilgisi özellikleri belirlenmiştir. Üçüncü bölümde başlık terimleri incelenmiş, pozitif ve negatif cümle analizleri gerçekleştirilmiştir. Dördüncü bölümde kullanılan veri seti tanıtılmıştır. Beşinci bölümde ise özetleme sisteminin yapısı üzerinde durulmuştur.

Kılcı vd. [47] çalışması ile paragraf, cümle ve terimlerin yapısal özellikleri analiz edilmiştir. Çalışmada kullanılan yapısal özellikler: “Anahtar söz öbekleri”, “terim sıklığı”, “cümle konumu”, “başlık kelimeleri”, “pozitif ve negatif ipucu terimleri”, “bazı noktalama işaretlerinin varlığı”, “gün-ay isimleri”, “nümerik karakter varlığı”, “özel isim varlığı” özellikleridir. Çalışmada bu özellikler kullanılarak cümlelere bir skor değeri atanmıştır. Özellikler birleştirilirken her bir özelliğin katkısı manuel bir şekilde önceden belirlenmiştir. Özet çıkarımı en yüksek puana sahip olan cümlelerin seçilmesiyle gerçekleştirilmiştir. Çalışmada 10 dokümandan oluşsan bir veri seti kullanılmıştır.

Cığır vd. [48] yine yapısal özellikleri kullanan ve cümle seçimi için yapısal özelliklerin birleşimiyle her cümleye bir skor değeri veren bir sistem önermişlerdir. Skor fonksiyonu, “kelime sıklığı”, “başlığa benzerlik”, “anahtar söz öbekleri”, “merkezilik”, “cümle konumu” özelliklerini kullanmaktadır. Cümlelerin skorlanması bu özelliklerin 0-1 aralığında değerler almasıyla gerçekleştirilir. Her cümleye ait en uygun skor değeri, özellik ağırlıklarının 0.01 birim arttırılarak en yüksek sistem performansına ulaşıldığı andaki ağırlık değerlerinin dikkate alınmasıyla belirlenmiştir. Çalışmalarında iki Türkçe veri seti kullanmışlardır. Birinci veri seti 120 haber dokümanını ve bu haberlerin %40 sıkıştırma oranı ile hazırlanmış olan özetlerini içermekteyken, ikinci veri seti 100 Türkçe bilimsel yayını ve bu yayınları %5 sıkıştırma oranı ile oluşturulmuş özetlerini içermektedir.

Özsoy vd. [49] gizli anlamsal analiz tekniğini hazırlamış oldukları iki Türkçe veri seti üzerinde denemişlerdir. Hazırlanan iki veri seti de 50’şer metin içermektedir. İlk veri setinde daha uzun haber metinleri bulunmaktadır. Bu çalışmada gizli anlamsal analiz tabanlı iki yeni yaklaşım önerilmiştir: “Çapraz (Cross)” ve “Konu (Topic)”. Bu önerilerden “Çapraz” yaklaşımının çalışmada kıyaslanan diğer yöntemlerdan daha iyi bir sonuç verdiğini belirtmişlerdir.

Güran vd. [50] çalışmalarında negatif olmayan matris ayrışımı tekniğini, hazırlamış oldukları 100 Türkçe haber veri seti üzerinde uygulamışlardır. Çalışmalarında sistem performansını arttıran yeni bir ön işlem aşaması önermişlerdir. Bu aşama ile “Türkçe Vikipedi” link yapısı kullanılarak metin içindeki sıralı kelimeler tespit edilmiştir ve bu tespitin yöntem üzerindeki olumlu etkileri sergilenmiştir.

Pembe [51], doktora tezinde web aramaları için sorgu tabanlı ve yapısal özelliklere dayalı olan bir özetleme sistemi önermiştir. Önerilen sistem iki aşama içermektedir: İlk aşamada özetlenecek metinler incelenmiş ve HTML etiketleri kullanılarak hiyerarşik bir şekilde bölüm ve alt bölümlere ayrılmıştır. İkinci bölümde ise cümle seçimi sorgu tabanlı bir yapıda, cümle puanlama ve bölüm puanlama olmak üzere iki puanlamaya göre yapılmıştır. Tezde, önerilen sistemin Google özet çıkarımından daha iyi başarım gösterdiği belirtilmiştir.

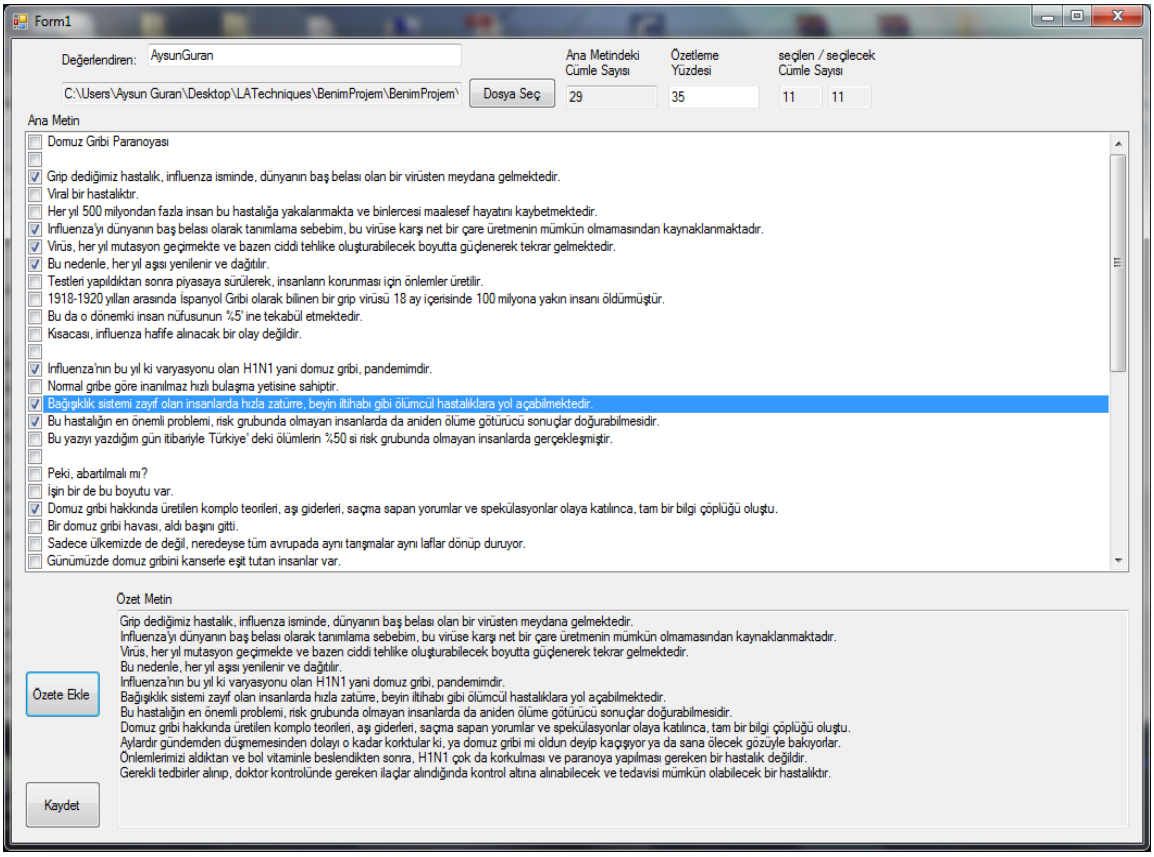
Türkçe metinlerin kullanılarak özetlendiği yukarıdaki çalışmalarda her ne kadar önerilen yöntemler bir veri setinde uygulanmış olsa da gerek veri sayılarının yetersizliği gerekse verilerdeki gürültülü yapı dolayısıyla pürüzler oluşturmaktadır. Bu nedenle tez çalışması kapsamında hem gürültüsüz hem de sayıca yeterli sayılabilecek iki veri seti hazırlanmıştır.

1.1.3 Tez Kapsamında Kullanılan Veri Setlerine ait İstatistikler

Metin özetleme alanında başarılı bir süreç geçirmenin en önemli koşulu, gürültüsüz ve geniş çaplı bir veri seti üzerinde çalışmaktır. Günümüzde İngilizce en sık kullanılan dil olduğundan, metin özetleme alanında çalışan kişiler tarafından oluşturulmuş olan geniş çaplı metin özetleme verilerine rastlamak mümkündür. Bu nedenle İngilizce veri setleri üzerinde gerçekleşen araştırma sayıları, Türkçe metinler üzerinde gerçekleşen araştırma sayılarından çok daha fazladır. Tez çalışması kapsamında, metin özetleme alanında yapılacak bilimsel araştırma sayılarını arttırmak adına iki yeni Türkçe veri seti hazırlanmıştır. İlk veri seti (VeriSeti-1) çeşitli haber sitelerinden toplanmış olan 130 haber dokümanını ve bu metinleri okuyan üç farklı kişi tarafından çıkarılmış olan 130 özet dokümanını içermektedir. Diğer veri seti (VeriSeti-2) ise ilk veri setine göre daha kısa olan 20 haber dokümanını ve bu dokümanlara ait otuz farklı kişinin hazırladığı 20'şer özet dokümanını içermektedir. Tez çalışması boyunca özetleyiciler tarafından oluşturulmuş olan özet dokümanlarına ideal özetler ismi verilecektir.

Veri setlerini oluşturan haber dokümanları dil bilgisi kurallarına uyumlu ve dokümanın her satırında bir cümle olacak şekilde oluşturulmuştur. Bu haber dokümanları Şekil 1.1 ile belirtilen bir arayüz ile özetleyicilere verilmiş ve özetleyicilerden doküman içerisinde

önemli olduklarını düşündükleri cümleleri seçmeleri istenmiştir. Böylelikle ideal özet dokümanları çıkarılmıştır.



Şekil 1. 1 Özet dokümanlarının çıkarılmasını sağlayan arayüz

Tez çalışması kapsamında kullanılan Türkçe ve İngilizce veri setleri ile ilgili ayrıntılı bilgiler aşağıdaki başlıklarla açıklanacaktır. (EK-B verisetlerine ait olan birer haber dokümanı ve bu haber dokümana ait olan özet dokümanını içermektedir.)

1.1.3.1 VeriSeti-1

Tez çalışması kapsamında önerilecek olan durumları test etmek üzere çeşitli gazetelerden toplanmış olan 130 haber metnini içeren bir veri seti hazırlanmıştır. Bu metinler, dilbilimsel kurallar açısından tüm kontrollerden geçmiştir ve dokümanların her satırına bir cümle gelecek şekilde oluşturulmuştur. Değerlendirme veri setine ait olan istatistikler Çizelge 1.1'de belirtildiği gibidir:

Çizelge 1. 1 Veri Seti-1 ile ilgili istatistikler

VeriSeti-1'e ait olan istatistikler	
Veri Setindeki Toplam Doküman Sayısı	130
Veri Setindeki Toplam Cümle Sayısı	2501
Ortalama Cümle Sayısı : (Toplam Cümle Sayısı/Toplam Doküman Sayısı)	19,24
Veri Seti İçindeki Dokümalardan en az Cümleye Sahip olan Dokümanın Cümle Sayısı	7
Veri Seti İçindeki Dokümalardan en fazla Cümleye Sahip olan Dokümanın Cümle Sayısı	63

Değerlendirme seti, üç farklı kişiye verilmiş ve kişilere özetleme oranı olarak herhangi bir kısıtlama getirilmeksizin, kişilerden her bir dokümana ait olan özetlerin çıkarılması istenmiştir. Kişiler bir dokümana ait olan özeti, o dokümandaki en önemli olan cümleleri seçerek oluşturmuşlardır. Böylece değerlendirme veri setine ait ideal özet grupları da oluşturulmuştur. Özet gruplarına ait olan istatistikler Çizelge 1.2 ile gösterilmiştir.

Çizelge 1. 2 VeriSeti-1'e ait olan özet dokümanlarının istatistikleri

Özetleyici	**Haber Veri Setindeki Cümle Sayısı**	**Haber Veri Setinden Çıkarılan Farklı Cümle Sayısı**	**Özetleme Oranı**
Özetleyici1	2501	850	%34
Özetleyici2	2501	923	%37
Özetleyici3	2501	652	%26
Ortalama Özetleme Oranı = %32.3			

Çizelge 1.2'deki verilere bakılarak üç özetleyicinin özet dokümanlarını, haber dokümanlarındaki cümlelerin ortalama olarak %32.3'sini seçerek oluşturduğu söylenebilir.

1.1.3.2 VeriSeti-2

Bu veri seti 20 haber dokümanını kapsayan bir değerlendirme setidir. Değerlendirme seti ile ilgili istatistikler Çizelge 1.3 ile belirtildiği gibidir:

Çizelge 1. 3 Veri Seti-2 ile ilgili istatistikler

VeriSeti-2'ye ait olan istatistikler	
Veri Setindeki Toplam Doküman Sayısı	20
Veri Setindeki Toplam Cümle Sayısı	201
Ortalama Cümle Sayısı : (Toplam Cümle Sayısı/Toplam Doküman Sayısı)	10,05
Veri Seti İçindeki Dokümalardan en az Cümleye Sahip olan Dokümanın Cümle Sayısı	7
Veri Seti İçindeki Dokümalardan en fazla Cümleye Sahip olan Dokümanın Cümle Sayısı	13

Bu veri setinin hazırlanmasındaki amaç literatürdeki yöntemlerin ve tez kapsamında sunulan önerilerin istikrarını göstermektir. Bu amaçla yirmi dokümanı içeren değerlendirme seti 15'i bay ve 15'i bayan olmak üzere toplam 30 farklı kişiye verilmiş ve kişilerden dokümanlara ait olan özetlerin %35'lik bir özetleme oranı ile çıkarılması istenmiştir. Bu oran VeriSeti-1'den elde edilen ortalama özetleme oranının yukarı yuvarlanması ile belirlenmiştir.

Çizelge 1.4 ve Çizelge 1.5 sırasıyla 20 dokümanlık veri setinde her bir dokümana ait olan cümlelerinin kaç bay ve kaç bayan özetleyici tarafından seçildiği bilgisini göstermektedir. Tabloların satırları dokümanları, sütunları ise dokümanın cümlelerini göstermektedir.

Çizelge 1. 4 Her bir cümlenin kaç bay özetleyici tarafından seçildiği bilgisi.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
1.txt	9	8	1	10	8	7	4	3	7	2	1		
2.txt	4	15	6	10	5	4	1	1	11	3			
3.txt	13	9	4	3	7	2	6	1					
4.txt	15	6	2	6	10	0	4	2					
5.txt	13	11	8	5	2	2	4						
6.txt	5	12	8	8	0	2	9	3	11	2			
7.txt	8	12	9	1	5	5	4	6	1	6	3		
8.txt	14	6	13	1	3	1	10	1	2	9			
9.txt	12	3	5	11	4	6	7	1	8	3			
10.txt	11	7	9	11	2	3	8	6	3	0	0		
11.txt	2	8	11	10	6	6	4	3	4	6			
12.txt	9	5	1	1	9	11	5	3	7	2	7		
13.txt	14	5	2	2	11	11	1	3	1	6	3	1	
14.txt	14	1	10	10	6	1	5	5	2	7	8	5	1
15.txt	11	4	8	6	1	3	6	1	5				
16.txt	15	10	7	3	5	1	8	7	0	4			
17.txt	7	7	2	1	1	13	7	10	3	5	4		
18.txt	14	0	12	4	7	4	1	0	3				
19.txt	14	7	2	8	3	1	2	10	7	1	0	5	
20.txt	15	0	1	10	3	8	1	7					

Çizelge 1. 5 Her bir cümlenin kaç bayan özetleyici tarafından seçildiği bilgisi.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
1.txt	6	10	4	12	8	5	5	4	3	2	1		
2.txt	4	12	8	10	4	5	1	1	11	4			
3.txt	14	9	2	3	9	2	3	3					
4.txt	12	8	6	7	8	0	4	0					
5.txt	13	10	11	5	0	2	4						
6.txt	3	13	3	11	3	4	7	4	10	2			
7.txt	8	10	11	1	8	5	4	5	2	1	5		
8.txt	13	8	11	0	2	1	12	5	3	5			
9.txt	9	4	3	11	10	2	9	1	7	4			
10.txt	8	3	13	13	3	5	6	6	3	0	0		
11.txt	6	9	11	7	8	4	4	3	3	5			
12.txt	8	8	1	2	10	10	1	4	4	1	11		
13.txt	12	9	2	4	11	7	1	3	5	5	1	0	
14.txt	14	3	11	11	3	2	4	7	0	5	5	7	3
15.txt	7	7	10	3	1	1	9	1	6				
16.txt	15	14	6	7	4	2	3	5	1	3			
17.txt	6	9	7	0	0	15	8	6	1	4	4		
18.txt	14	1	9	5	7	4	2	2	1				
19.txt	14	0	1	10	6	3	6	8	5	1	0	6	
20.txt	13	5	1	9	0	10	1	6					

Çizelgelerden görüldüğü gibi her bir dokümandan seçilen cümleleri belirleyen bay ve bayan özetleyici sayıları birbirleriyle oldukça paraleldir. Özetleyiciler genelde ilk sırada bulunan cümleleri seçmişlerdir.

1.1.3.3 VeriSeti-3

CAST veri seti Reuters haber ajansına ait olan haber dokümanlarını kapsamaktadır [52-43]. Bu veri setinde tüm haber dokümanları kelime tabanlı olarak etiketlenmiştir. Haberlere ait olan özetler cümle çıkarımına dayalı olan özetlerdir. Yani özetler orijinal haber dokümanlarından önemli görülen cümlelerin seçilmesiyle oluşturulmuştur. Özet dokümanları ayrı bir dosyada tutulmamış, orijinal haber dokümanlarında "önemli", "orta seviyede önemli" ve "önemsiz" isimleri ile etiketlendirilmiştir. Etiketleme işlemi bazı dokümanlar için birden fazla kişi tarafından yapılmıştır. Tez çalışması kapsamında etiketlenmesi ilk değerlendirici tarafından tam olarak yapılmış olan 92 haber metni üzerinde çalışılmıştır. Bu metinlere ait olan ideal özet grubu orijinal haber

dokümanında yanında “önemli”, “orta seviyede önemli” ve “önemsiz” etiketleri bulunan cümlelerin alınmasıya oluşturulmuştur.

92 haber dokümanı içeren değerlendirme verisetine ait olan istatistikler aşağıdaki çizelge ile belirtilmiştir.

Çizelge 1. 6 Veri Seti-3 ile ilgili istatistikler

VeriSeti-3’e ait olan istatistikler	
Veri Setindeki Toplam Doküman Sayısı	92
Veri Setindeki Toplam Cümle Sayısı	2154
Ortalama Cümle Sayısı : (Toplam Cümle Sayısı/Toplam Doküman Sayısı)	23,4
Veri Seti İçindeki Dokümalardan en az Cümleye Sahip olan Dokümanın Cümle Sayısı	7
Veri Seti İçindeki Dokümalardan en fazla Cümleye Sahip olan Dokümanın Cümle Sayısı	47

Oluşturulan ideal özet grubunda bulunan farklı cümle sayısı 672’dir. Bu durumda bu değerlendirme setine ait olan ideal özet grubu %31’lik bir özetleme oranı ile oluşturulmuştur.

1.1.3.4 VeriSeti-4

Bu veri seti 2002 yılında düzenlenen “Metin Anlama Konferansı” (Document Undertsanding Conference- DUC 2002) kapsamında oluşturulmuştur [53]. Veri seti farklı konuları içeren 567 adet haber dokümanını ve bu dokümanlara ait olan özet dokümanlarını kapsamaktadır. Özet dokümanlarının her biri 100’er kelime içermektedir. Çıkarılan özetler yoruma dayalı olan özetlerdir.

Tez çalışması kapsamında her bir haber metni her satırda bir cümle olacak şekilde parçalara ayrılmıştır. Bu işlem gerçekleştirilirken DUC 2002 konferansında önerilen kod kullanılmıştır [53]. Çıkarılan ideal özet gruplarına da aynı işlem uygulanmıştır. Bu şartlar altında elde edilen veri setine ait olan istatistikler aşağıdaki çizelgeden görülebilir:

Çizelge 1. 7 Veri Seti-4 ile ilgili istatistikler

VeriSeti-4'e ait olan istatistikler	
Veri Setindeki Toplam Doküman Sayısı	567
Veri Setindeki Toplam Cümle Sayısı	16432
Ortalama Cümle Sayısı : (Toplam Cümle Sayısı/Toplam Doküman Sayısı)	28,98
Veri Seti İçindeki Dokümalardan en az Cümleye Sahip olan Dokümanın Cümle Sayısı	5
Veri Seti İçindeki Dokümalardan en fazla Cümleye Sahip olan Dokümanın Cümle Sayısı	176

VeriSeti-4'e ait olan ilk ideal özet grubunda bulunan farklı cümle sayısı 3180'dir. Bu durumda bu değerlendirme setine ait olan ideal özet grubu %20'lik bir sıkıştı özetleme oranı ile oluşturulduğu söylenebilir.

1.1.4 Metin Özetlemede Değerlendirme Ölçütleri

Otomatik özetleme sistemlerinin kullanılan veri setleri üzerindeki başarımlarının değerlendirilmesi amacıyla kullanılan çok sayıda ölçüm yöntemleri bulunmaktadır. Bu değerlendirme yöntemleri görevden bağımsız (task-independent) ve görev tabanlı (task-based) yöntemler olmak üzere iki grup altında incelenebilir.

Görevden bağımsız yöntemler uzman görüşüyle oluşturulmuş olan özeti (ideal özet) temel alır. Bu bağlamda, bir özetin değerlendirilmesi otomatik sistem tarafından oluşturulan özet ile ideal özetin kıyaslanmasıyla yapılmaktadır.

Görev tabanlı yöntemler ise uzman değerlendirmesini özel bir alanı belirlemek adına kullanmaktadır. Bu amaçla görev tabanlı yöntemlerde metin sınıflama (text categorization), bilgi çıkarımı (information retrievel) ve soru cevaplama (question answering) gibi teknikler kullanılmaktadır.

Aşağıdaki başlıklarda görevden bağımsız ve görev tabanlı yöntemler anlatılmaktadır.

1.1.4.1 Görevden Bağımsız Yöntemler

Bu yöntemler otomatik sistem tarafından çıkarılan özet ile ideal özeti kıyaslar. Değerlendirme sırasında cümlelerin tümü bir bütün olarak düşünülüp iki doküman arasındaki çakışan cümle sayıları dikkate alınabileceği gibi cümleyi oluşturan

sözcüklerin çakışma oranı da kıyaslanabilmektedir. Bu durum görevden bağımsız yöntemlerin hem yoruma dayalı olan hem de yoruma dayalı olmayan özetleme sistemlerinde kullanılabileceğini göstermektedir. Akademik çalışmalar ile önerilmiş olan ve literatürde en çok kullanılan görevden bağımsız yöntemler: keskinlik (precision), anma (recall), F-ölçüm değeri (F-score) - (KAF); göreceli fayda değeri (GF); kosinüs benzerliği(KB); Ngram birliktelik istatistiği (ROUGE) değerleridir:

- **Keskinlik, Anma, F-ölçüm Değeri:**

Görevden bağımsız tekniklerden olan KAF ölçüm yöntemi otomatik sistem özetlerinde ve ideal özetlerde bulunan ortak cümlelerin bulunmasına dayanmaktadır. Keskinlik değeri (K), otomatik ve ideal sisteme ait olan özetlerin içerdiği ortak cümle sayısının otomatik sistemdeki cümle sayısına (S) oranıdır. Anma değeri (A), otomatik sistemin ve ideal sistemin içerdiği aynı cümle sayısının ideal sistemin toplam cümle sayısına (T) oranıdır. F-ölçüm değeri (F), keskinlik ve anma değerlerinin harmonik ortalaması olan birleşik bir ölçümdür. Bu değerler eşitlik (1.1) ile gösterilen şekilde formülüze edilir.

$$K = \frac{|S \cap T|}{|S|} \quad A = \frac{|S \cap T|}{|T|} \quad F = \frac{2KA}{K + A} \tag{1.1}$$

- **Göreceli Fayda :**

KAF ölçümündeki en büyük sorun, bir dokümandaki önemli cümlelerin belirlenmesinde insan seçimlerinin farklı olabileceğinin göz ardı edilmesidir. KAF'ı kullanmak eşit öneme sahip iki özetin farklı değerlendirilmesine yol açacaktır. İdeal özetin (1, 2) cümlelerini, A ve B gibi iki ayrı otomatik özetleme sisteminse sırasıyla (1, 2) ve (1, 3) cümlelerini içerdiğini düşünün. KAF ölçümüne göre A sistemi B sistemine göre daha değerli olacaktır. Ancak 2 ve 3 numaralı cümle aynı önem derecesine sahip olabilir. Bu durumda iki sistemin de eşit önemde olması gerekmektedir. İşte bu problemi ortadan kaldırmak için göreceli fayda (GF) ölçümü ortaya konmuştur. Bu ölçümde her cümle özete katılma durumuna göre skorlanır. Örneğin beş cümleden oluşan bir sistemde cümleler (1/5 2/4 3/4 4/1 5/2) şeklinde ifade edilebilir. Her bir sayı çiftindeki ikinci sayı cümlenin çıkarılacak özet için insan muhakemesine göre ne derece öneme sahip

olduğunu belirler. Bu sayı fayda (utility) adını alır. Bu sayı, doküman çeşidine, özet uzunluğuna ve muhakeme eden kişiye göre değişkenlik gösterir. Göreceli fayda ölçümüne (relative utility) göre önceki örnekte, (1, 2) sistemi (1, 3) sisteminden daha yüksek bir skora sahip olmayacaktır. Çünkü iki cümle grubu da 5+4 olmak üzere aynı fayda değerine sahiptir.

Göreceli fayda ölçümünü hesaplayabilmek için (N ≥ 1) kişinin n adet cümleye fayda skoru verdiği düşünülmektedir. Fayda skoruna göre en yüksek skora sahip olan e adet cümleye "e boyutlu çıkarılmış cümle takımı" adı verilir. Sistemin performans ölçümü aşağıdaki şekilde hesaplanır:

$$GF = \frac{\sum_{j=1}^{n} \delta_j \sum_{i=1}^{N} u_{ij}}{\sum_{j=1}^{n} \in_j \sum_{i=1}^{N} u_{ij}} \qquad (1.2)$$

Burada u_{ij}, i. değerlendiricinin j. cümle için vermiş olduğu fayda skorunu ifade etmekte; $\in_j$ değeri tüm kişilerin verdikleri fayda değerlerinin toplamına göre en yüksek e cümle için 1 iken aksi durumda 0'dır. δ_j ise sistem tarafından çıkarılan en yüksek skorlu e cümle için 1, diğer durumlar için 0'dır. Detaylar için [54]'a bakılabilir.

- **Kosinüs Benzerliği:**

Otomatik sistem özeti ile ideal özetin ne ölçüde benzediğini gösteren en temel içerik ölçüm metodunu kosinüs benzerliğidir. (1.3) eşitliği ile belirtilen formülde x_i ifadesi ideal özetteki kelimelerin frekans değerlerini belirtirken, y_i değerleri otomatik sistem özetindeki kelimelerin frekans değerlerini belirtmektedir. Ayrıntılar için [55] referansı incelenebilir.

$$\cos(X,Y) = \frac{\sum_i x_i . y_i}{\sqrt{\sum_i (x_i)^2} \sqrt{\sum_i (y_i)^2}} \qquad (1.3)$$

- **Ngram birliktelik istatistiği (ROUGE)**

ROUGE (Recall- Oriented Understudy for Gisting Evaluation) otomatik değerlendirme paketi Chin-Yew Lin tarafından oluşturulmuştur [56]. Perl dili ile tasarlanmış olan bu paket ilk olarak doküman anlama konferansında (DUC - Document Understanding Conference) kullanılmıştır [53]. Günümüzde metin özetleme alanında kullanılan en güncel özet değerlendirme paketidir. Bu paketin bir özetleme sistemini değerlendirme amacıyla kullanmış olduğu ölçüm, otomatik sisteme ve ideal özet grubuna ait olan özet dokümanlarında bulunan ortak kelime sayısına dayanmaktadır. ROUGE paketi toplam beş farklı ölçüm değerine sahiptir: ROUGE-N, ROUGE-L, ROUGE-S, ROUGE-W, ROUGE-SU.

ROUGE-N otomatik özet dokümanı ile ideal özet dokümanı arasındaki Ngram anımsama değeridir ve (1.4) eşitliği ile gösterildiği gibi hesaplanır.

$$\mathrm{ROUGE}-N=\frac{\sum_{S\in\{\text{İnsan Özetleri Grubu}\}}\sum_{gram_N\in S}Hesapla_{cakisan}(gram_N)}{\sum_{S\in\{\text{İnsan Özetleri Grubu}\}}\sum_{gram_N\in S}Hesapla(gram_N)} \qquad (1.4)$$

Burada $Hesapla_{cakisan}(gram_N)$ ideal ve otomatik sistem özetlerinin ortaklaşa sahip olduğu maksimum Ngram sayısı (N uzunluklu sıralı kelime grubu sayısı) ve $Hesapla(gram_N)$ ideal özetteki toplam Ngram sayısıdır.

ROUGE paketindeki ROUGE-S değeri ardışık sırada olmayan Ngramların çakışma oranını göstermektedir. Pakette yer alan diğer ölçüm değerleriyle ilgili ayrıntılar [56] referansında bulunabilir.

1.1.4.2 Görev Tabanlı Yöntemler

Görev tabanlı ölçüm yöntemleri özette bulunan cümleleri analiz etmemektedir. Bu yöntemler özetlemenin diğer alanlar (doküman sınıflama, bilgi çıkarımı, soru cevaplama sistemleri vs.) üzerinde etkilerini denetlerler.

Örneğin bir metin sınıflama sisteminde dokümanların orijinallerinin kullanması yerine dokümanlara ait olan özetler kullanılabilir. Burada çıkarılan özetin sistemi sınıflandıracak kadar bilgi içerip içermediğine bakılır. İlgi korelasyonu (Relevance correlation) [57] bir sistemde orijinal dokümanları kullanmak yerine özet dokümanlar kullanıldığına ilgi düşüşlerini değerlendirmek için kullanılan bir ölçümdür. Eğer bir özet aslını iyi temsil ediyorsa bilgi çıkarım makinesi indeksi (IR-machine index) güzel sonuçlar üretir. Başka bir örnek arama motorlarında sorgu tabanlı aramalardan verilebilir. D doküman grubunu içeren bir derlemde Q sorgusu yapıldığında arama motoru dokümanları sıralayacaktır. Dokümanlar yerine özetleri kullanıldığında sıralama değişebilir. Eğer özetler orijinal hallerini iyi temsil ediyorlarsa benzer bir sıralama gelecektir. Böylece bir özetleme sisteminin performansı metin sınıflama ya da sorgulama gibi farklı alanlarla da sınanmış olur.

Tez çalışması kapsamında görevden bağımsız yöntemlerden keskinlik, anımsama, F-ölçüm değeri ve Ngram istatistiğine dayalı olan ROUGE değerlendirme yöntemleri kullanılacaktır.

1.2 Tezin Amacı

Tez çalışmasında tek kaynaklı, tek dil ile yazılmış, çıkarıma dayalı olan ve genel özetlerin çıkarılacağı bir sistem üzerinde çalışılacaktır. Bu tarz yapıya sahip olan bir sistemin amacı özeti çıkarılacak olan metin içindeki en önemli cümlelerin belirlenmesi ve belirlenen cümlelerin var oldukları şekilde özet dokümanlarına eklenmesidir. Tez çalışmasında bu amaca ulaşmak için ilk aşamada literatürde bulunan yapısal ve anlamsal metin özetleme yöntemleri incelenmiştir. Sonrasında incelenen yöntemlerin başarımlarını arttırmak adına yenilikler önerilecektir. Ardından literatürde incelenen çalışmalar tarafından kullanılmış olan yapısal ve anlamsal özellikleri birleştiren yeni bir melez yaklaşım sunulacaktır. İncelenen yöntemler ve önerilen yeni durumlar, tez kapsamında hazırlanmış olan iki yeni Türkçe veri seti ve sık kullanılan iki İngilizce veri seti üzerinde uygulanacaktır. Yöntem ve yaklaşımlar java programlama dili kullanılarak Intel(R) Core(TM) i3-2367M - 1.40 GHz işlemcili ve 4GB belleğe sahip bir bilgisayar aracılığı ile test edilecektir.

Tez çalışmasında izlenecek adımlar aşağıdaki sırayla ele alınacaktır:

Bu bölüm ile OMÖ alanında bilinmesi gereken temel kavramlar tanıtılmış, ardından incelenen yöntemlerin başarım değerlendirme aşamalarından bahsedilmiş ve değerlendirmede kullanılacak veri setleri tanıtıldıktan sonra tez çalışmasının katkıları üzerinde durulmuştur.

İkinci bölümde metinler içindeki gizli anlamsal yapıyı ortaya çıkaran yöntemler açıklanmış ve bu yöntemlerin başarımlarını arttıran yeni bir öneri sunulmuştur. Önerilen yeni yaklaşıma ait başarım analizleri hem Türkçe, hem de İngilizce veri setleri üzerinde gerçekleştirilmiştir.

Üçüncü bölümde yapısal ve anlamsal özelliklerin bir arada kullanıldığı melez bir sistem önerilmiştir. Önerilen sistemde kullanılan yapısal ve anlamsal özellikler hem uzman gücüne dayalı olarak bulanık tabanlı bir yol ile hem de otomatik olarak sezgisel bir algoritma ile birleştirilmiştir. Sonuçta melez sistem ile elde edilen sonucun ele alınan her bir özellikten daha iyi başarım sonuçları verdiği gösterilmiştir. Literatürde İngilizce dokümanlar üzerinde çalışan ve cümle seçimi için kullanılan yapısal veya anlamsal özelliklerin birleşimini sağlayan melez sistem önerileri mevcuttur. Bu önerilerde özelliklerin birleşimi ile elde edilen yapıların sistem başarımları üzerindeki olumlu etkileri vurgulanmış ve bireysel özelliklerin katkıları üzerinde durulmuştur. Tez çalışması kapsamında hazırlanmış melez sistem Türkçe veri setleri üzerinde uygulanmıştır.

Nihayet son bölümde genel yorumlar üzerinde durularak tezin amaçları doğrultusunda hedeflenen adımlar özetlenmiştir.

1.3 Orjinal Katkı

Belirtilen tüm bu aşamalar eşliğinde tez çalışmasının katkıları şu şekilde sıralanabilir:

- Metin özetleme çalışmalarının Türkçe metinler üzerinde uygulanabilmesi için Türkçe haber dokümanlarından oluşan kapsamlı iki veri seti hazırlanmıştır.
- Gizli anlamsal analiz yönteminde kullanılmak üzere önerilen yeni bir ağırlıklandırma biçimi ile Gong ve Lui 2001, Murray vd. 2005, Steinberger vd. 2007 ve Özsoy vd. 2011 çalışmalarının başarımlarının arttığı gösterilmiştir.

- Metin özetleme çalışmalarında kullanılan yapısal ve anlamsal tüm özellikleri birleştirmek için bulanık tabanlı ve genetik algoritma tabanlı iki yeni ağırlıklandırma sistemi önerilmiş ve bu melez sistemler ile bulunan ağırlık değerlerinin kullanılmasıyla daha yüksek başarımlara ulaşıldığı gösterilmiştir.

BÖLÜM 2

TEKİL DEĞER AYRIŞIMINA DAYALI METİN ÖZETLEME YÖNTEMLERİ

Metin özetleme, kullanıcının ihtiyacı olan bilginin arındırılarak çıkarılması işlemidir. Literatürde doğru biçimde oluşturulmuş özetler çıkarmak için farklı yaklaşımlar mevcuttur. Bu yaklaşımların en yenilerinden biri eğiticisiz bir öğrenme modeline sahip olan ve Tekil Değer Ayrışımına (TDA) dayalı olan Gizli Anlamsal Analiz (GAA) yöntemidir. Bu bölümde GAA'ya dayalı farklı özet çıkarma yöntemleri incelenmiştir. İncelenen yöntemler "Yöntem1", "Yöntem2", "Yöntem3" ve "Yöntem4" isimleriyle gösterilmiş olan ve sırasıyla Gong ve Lui [40], Murray vd. [41], Steinberger [42], Özsoy vd. [49] çalışmalarına ait olan yöntemlerdir. Tez çalışmasında bu yöntemlerin başarımlarını yükselten yeni bir öneri sunulmuştur. Sunulan öneri, yöntemler üzerinde Türkçe ve İngilizce veri setleri kullanılarak test edilmiş ve başarım sonuçları F-skor ve ROUGE değerleri kullanılarak karşılaştırılmıştır.

2.1 Gizli Anlamsal Analiz

GAA metinsel verilerden anlamsal genellemeler çıkaran bir yöntemdir ve metinsel bilgi çıkarımı, metin bölütlemesi ve son zamanlarda metin özetleme gibi sistemlerde kullanılmaktadır. GAA sistem girdisi olarak bir metnin içeriğini almakta ve metin içindeki terimler ile cümleler arasındaki gizli anlamsal ilişkiyi ortaya çıkarmaktadır. Cümleler arasındaki anlamsal ilişkiler cümlelerin içerdiği ortak terimler ile kurulurken, terimler arasındaki anlamsal ilişkiler bu terimleri içeren cümlelerin kullanılmasıyla çıkarılmaktadır. Metin içindeki anlamsal yapı, cebirsel bir yöntem olan TDA'ya dayalı olarak belirlenmektedir. TDA bir metin içindeki gizli anlamsal ilişkiyi çıkarmanın yanında

metin içindeki gürültüyü de azaltmaktadır. Bu durum metin analizinde sistem başarımını arttırmaktadır.

GAA'yı metin özetleme alanında kullanan ilk çalışma Gong ve Liu tarafından 2001 yılında yapılmıştır[40]. Bu çalışmadan görülebileceği gibi GAA'nın sistem girdisi mxn boyutuna sahip olan ve bir terim-cümle matrisi olan $A_i = [a_{1i}, a_{2i}, ..., a_{ni}]$ matrisidir. Bu matrisin satır değerleri dokümanı oluşturan terimleri içerirken, sütun değerleri dokümana ait olan cümleleri içermektedir. Matrisin hücre değerleri olan a_{ji} değerleri terimlerin önemini belirten bazı ifadelerle hesaplanmaktadır. Gong ve Liu, sistemlerinde a_{ji} değerlerini (2.1) eşitliği ile ifade edildiği gibi lokal $L(t_{ji})$ ve global $G(t_{ji})$ ağırlık değeri adı verilen değerlerin çarpımıyla elde etmişlerdir. Tez çalışmasında bu ağırlık sistemi terimlerin frekans bilgisi ile ilgili olduğundan $T_{frekans}$ şeklinde ifade edilmiştir.

$$a_{ji} = T_{frekans} = L(t_{ji}) \times G(t_{ji}) \quad (2.1)$$

Çalışmada lokal ağırlık değeri için dört farklı hesaplama kullanılmıştır. Bu ifadeler:

- Frekansa Bağlı Ağırlık (F): $L(t_{ji}) = tf(t_{ji})$, burada $tf(t_{ji})$ ifadesi t_{ji} teriminin cümlede geçme sayısıdır.
- İkili ağırlık (İ): Eğer t_{ji} terimi cümlede geçiyorsa $L(t_{ji}) = 1$, aksi durumda $L(t_{ji}) = 0$ olarak alınır.
- Birleşmiş Ağırlık (B): $L(t_{ji}) = 0.5 + 0.5 \times \frac{tf(t_{ji})}{tf(maks)}$, burada $tf(maks)$ incelenen cümlede en çok geçen kelime sayısıdır.
- Logaritmik Ağırlık (L): $L(t_{ji}) = \log(1 + tf(t_{ji}))$ şeklinde hesaplanır.

Global ağırlık değeri , $G(t_{ji})$, için ise aşağıdaki iki ifade kullanılmıştır:

- Sabit Ağırlık (S) : Her terim için $G(t_{ji}) = 1$ alınır.
- Ters Doküman Sıklığı Bilgisi (T): $G(t_{ji}) = \log(\frac{N}{n_i}) + 1$, burada N dokümandaki toplam cümle sayıdır. n_i ise i.terimi içeren cümle sayısıdır.

GAA'de sistem girdisi olan A matrisi oluşturulduktan sonra, matrisin TDA'sı gerçekleştirilir. TDA ile A matrisi (2.2)'de gösterildiği gibi 3 ayrı çarpana ayrılmaktadır.

$$A = USV^T \tag{2.2}$$

Burada $U = \lfloor u_{ij} \rfloor$ matrisi kolonları sol tekil vektörler olarak adlandırılan, mxn boyutlu birim dikey matristir. S matrisi köşegenlerinde negatif olmayan tekil değerlerin büyükten küçüğe sıralanmış bir şekide dizildiği $n \times n$ boyutlu köşegen matristir ve son olarak $V = \lfloor v_{ij} \rfloor$ matrisi kolonları sağ tekil vektörler olarak adlandırılan $n \times n$ boyutlu birim dikey matristir.

A matrisinin rankı r iken S ve A matrisleri sırasıyla aşağıdaki koşulları sağlamaktadır:

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r \succ \sigma_{r+1} = ... = \sigma_n = 0 \tag{2.3}$$

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T \tag{2.4}$$

$$A = \sigma_1 u_1 v_1^T + ... + \sigma_k u_k v_k^T + ... + \sigma_r u_r v_r^T \tag{2.5}$$

$k < n$ olmak üzere $U_k = (u_1, u_2, ..., u_k)$, S_k ve $V_k = (v_1, v_2, ..., v_k)$ matrisleri ele alındığında $A_k = U_k S_k V_k^T$ çarpımına tekil değer ayrışımının indirgenmiş formu denir ve bu form bir çok uygulamada kullanılır.

GAA ile ilgili verilen bu genel bilgilerin ardından Bölüm 2.1.1'de TDA ile ilgili bir teorem ve ispatı incelenmiş [58] ve TDA aşamaları özetlenmiştir. Bölüm 2.1.2'de ise TDA ayrışımının metinler üzerinde kullanımı bir örnek ile anlatılmıştır.

2.1.1 TDA ile İlgili Bir Teorem ve İspatı

Eğer A matrisi $m \times n$ boyutlu bir matris ise A tekil değer ayrışımına sahiptir [58].

İspat:

$A^T A$, $n \times n$ boyutlu simetrik bir matristir. Bu yüzden öz değerleri reeldir ve A matrisini köşegenleştirebilen bir dik V matrisine sahiptir. Dahası tüm öz değerleri pozitiftir. λ, $A^T A$'nın özdeğeri ve x λ'ya ait olan bir öz vektör olsun:

$$\|Ax\|^2 = x^T A^T Ax = \lambda x^T x = \lambda \|x\|^2 \tag{2.6}$$

Bu yüzden,

$$\lambda = \frac{\|Ax\|^2}{\|x\|^2} \geq 0 \qquad (2.7)$$

$A^T A$ matrisinden elde edilen öz değerleri büyükten küçüğe sıralayarak $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq 0$ bu değerlere karşılık gelen öz vektörlerin oluşturduğu matrisi V matrisi olarak tanımlayalım. Buna göre tekil değerler, $\sigma_j = \sqrt{\lambda_j}$ $j = 1,2,...,n$ şeklinde tanımlanırlar.

A matrisinin rankı r ise, $A^T A$ matrisinin de rankı r'dir. $A^T A$ simetrik bir matris olduğundan rankı sıfırdan farklı özdeğer sayısı kadardır. Bu yüzden $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r \succ 0$ ve $\lambda_{r+1} = \lambda_{r+2} = ... = \lambda_n = 0$ sonucuna ulaşılır. Benzer ilişki tekil değerler için de geçerlidir: $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r \succ 0$ ve $\sigma_{r+1} = \sigma_{r+2} = ... = \sigma_n = 0$ dır.

Şimdi $V_1 = (v_1,...,v_r)$ ve $V_2 = (v_{r+1},...,v_n)$ olduğunu varsayalım. Aynı zamanda

$$S_1 = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & . & \\ & & & \sigma_r \end{bmatrix} \qquad (2.8)$$

olsun. S_1, $r \times r$ boyutlu köşegen matristir ve köşegen elemanları sıfırdan farklı tekil değerleri içerir. $m \times n$ boyutlu S matrisi ise aşağıdaki şekilde ifade edilsin:

$$S = \begin{bmatrix} S_1 & 0 \\ 0 & 0 \end{bmatrix} \qquad (2.9)$$

V_2'nin kolon vektörleri $A^T A$ matrisinin $\lambda = 0$ özdeğerlerine ait olan özvektörleridir. Bu yüzden $A^T A v_j = 0$, $j = r+1,...,n$ 'dir. Sonuç olarak denilebilirki V_2'nin kolon vektörleri $N(A^T A) = N(A)$ nın birim dikey taban vektörleridir. Bu yüzden $V_2 = 0$'dır ve V birim dikey matris olduğundan aşağıdaki koşulları sağlar:

$$I = VV^T = V_1 V_1^T + V_2 V_2^T \qquad (2.10)$$

$$A = AI = AV_1 V_1^T + AV_2 V_2^T = AV_1 V_1^T \qquad (2.11)$$

Şu ana kadar V ve S matrisleri ile ilgili açıklamalar yapılmıştır. İspatı tamamlamak adına $m \times m$ boyutlu birim dikey U matrisinin de aşağıdaki ifadeleri sağladığını göstermemiz gerekmektedir:

$$A = USV^T \tag{2.12}$$

$$AV = US \tag{2.13}$$

(2.13) denkleminde her iki taraftaki ilk r kolon dikkate alındığında $Av_j = \sigma_j u_j$, $(j = 1,...,r)$ ifadesinden $u_j = \frac{1}{\sigma_j} Av_j$ ifadesi elde edilir. Aynı zamanda $U_1 = (u_1,...,u_r)$ matrisi kullanılarak $A_1 V_1 = U_1 S_1$ eşitliği yazılabilir.

U_1 matrisinin kolon vektörleri aşağıdaki koşul sağlandığı müddetçe birim dikeydir:

$$u_i^T u_j = (\frac{1}{\sigma_i} v_i^T A^T)(\frac{1}{\sigma_j} Av_j) = \frac{1}{\sigma_i \sigma_j} v_i^T (A^T A v_j) = \frac{1}{\sigma_i \sigma_j} v_i^T v_j = \delta_{ij} \qquad 1 \le i, j \le r \tag{2.14}$$

Burada her bir u_j, $1 \le i \le r$, A vektörünün kolon uzayıdır. Kolon uzayının boyutu r'dir. Dolayısıyla $u_1,...,u_r$ $R(A)$'nın birim dikey tabanlarıdır. $R(A)^{\perp} = N(A^T)$ vektör uzayı $m \times r$ boyutuna sahiptir.

$\{u_{r+1}, u_{r+2},..., u_m\}$, $N(A^T)$'nin birim dikey taban vektörleridir. Bu sete $U_2 = (u_{r+1}, u_{r+2},..., u_m)$ diyelim. $U = (U_1 \quad U_2)$ matrisi, $\{u_1, u_2,..., u_m\}$ vektör seti R^m'in birim dikey taban vektörleri olduğu için dikey bir matristir. Sonuç itibariyle gösterilebilir ki:

$$USV^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} S_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = U_1 S_1 V_1^T = AV_1 V_1^T = A \tag{2.15}$$

Böylelikle ispat tamamlanmıştır.

2.1.1.1 Tekil Değer Ayrışımı için İşlem Basamakları

U, S ve V matrislerini elde etmek için,

1) $A^T A$ matrisinin özdeğerleri bulunur ve azalan sırada dizilir.

2) $A^T A$ matrisinin sıfırdan farklı özdeğerlerinin sayısı ($r = rank(A^T A)$) bulunur.

3) $A^T A$ matrisinin özdeğerlerine karşılık gelen özvektörler bulunur ve öz vektörler azalan sırada dizilmiş olan özdeğerlere göre sıralanarak V matrisinin sütunlarını oluşturur.

4)Köşegen S matrisi asıl köşegenine azalan sırada yerleştirilen $\sigma_i = \sqrt{\lambda_i}$ tekil değerleri ile elde edilir.

5) AA^T matrisinin özdeğerlerine karşılık gelen özvektörler bulunur ve sütun-vektörlerden oluşan U matrisi elde edilir veya U matrisinin elemanları $u_i = \sigma_i^{-1} A v_i \quad i = 1,\ldots, rank(A^T A)_i$ bağıntısından ele edilir.

Örnek: Verilen A matrisinin TDA değerlerini bulalım:

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

1)Önce $A^T A$ matrisini bulalım:

$$A^T A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Bu matrisin özdeğerleri $\lambda_1 = 3, \lambda_2 = 1$'dir.

2)Matrise ait sıfırdan farklı özdeğer sayısı: $rank(A^T A) = r = 2$'dir.

$A^T A$ matrisinin $\lambda_1 = 3$ve $\lambda_2 = 1$özdeğerlerine karşılık gelen diklik koşulunu sağlayan özvektörleri, $(A^T A)v = \lambda_i v$ eşitliğinden normalleştirilmiş $v_i = \dfrac{v}{\|v\|}$ vektörleridir:

$$v_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \text{ ve } v_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$$

Bu iki sütun vektöründen elde edilen V matrisi ise $V = [v_1, v_2] = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$

şeklinde bulunur.

3)Tekil değerleri içeren S matrisi, $S = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & \sqrt{1} \end{bmatrix}$ şeklindedir. Burada köşegen elemanları özdeğerlerin kareköklerinden, geri kalan elemanlar ise 0 değerinden oluşmaktadır.

4) U matrisinin ilk sütun vektörü $u_1 = \sigma_1^{-1} A v_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix}$

şeklindedir. İkinci sütun vektörü ise $u_2 = \sigma_2^{-1} A v_2 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0 \\ -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$'dir.

5)Sonuç olarak A matrisinin TDA bileşenleri :

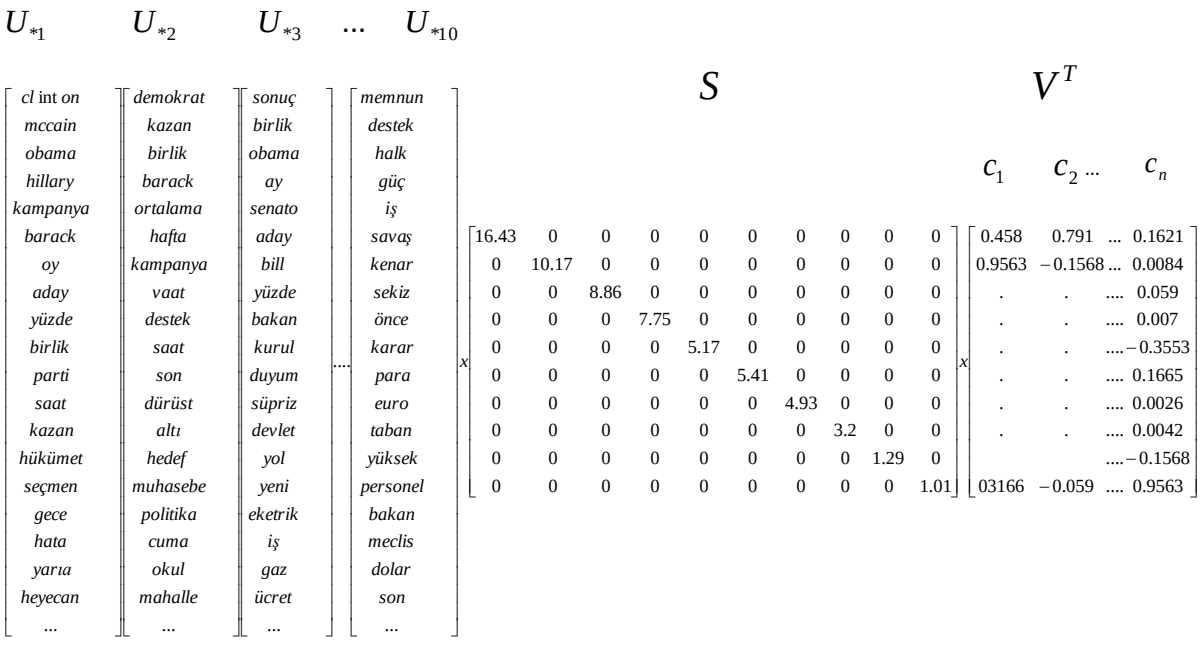
$$A = \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & \sqrt{1} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$ şeklinde bulunmuş olur.

2.1.2 TDA'nın Dil Bilimsel Yorumu

TDA'nın metin özetleme alanındaki kullanım amacı bir dokümandaki terimleri ve cümleleri ayrışım sonucu elde edilen çarpan matrislerinin sütun ve satırları ile indekslemektir. Terim-cümle matrisinin ayrışımı sonucu elde edilen U çarpan matrisi sütun değerleri ile dokümandaki terimleri birlikte geçme sıklıklarına göre indekslemektedir. Birlikte geçen kelimelerin bir özvektör altında indekslenmesi, öz vektörlerin doküman içerisinde bahsi geçen konular hakkında bilgi verdiği sonucunu ortaya çıkarmaktadır. Bu yüzden literatürde öz vektörlere dokümanda bahsi geçen "konular" denilmektedir [45]. Terim-cümle matrisinin ayrışımı sonucu elde edilen S çarpan matrisindeki tekil değerler büyükten küçüğe doğru sıralanmıştır. En büyük tekil değere karşılık gelen öz vektör dokümanda en çok bilgiyi taşıyan öz vektördür. Ayrışım

sonucunda elde edilen V^T matrisinin satırları her bir öz vektörü sütunları ise her bir cümleyi ifade etmektedir. S çarpan matrisindeki tekil değerlerin büyükten küçüğe doğru sıralanmasından ötürü V^T matrisinde ilk sırada bulunan öz vektör en çok bilgiyi indeksleyen "konu" vektörüdür ve metin özetleme aşamasında, amaç en çok bilgiyi içeren cümleleri seçmek olduğundan, V^T matrisi kullanılmaktadır.

Konunun daha iyi anlaşılabilmesi için, Şekil 2.1 yardımıyla, bir dokümana ait olan TDA ile elde edilmiş çarpan matrislerinin anlamsal yapısını gösteren bir örnek sunulmuştur.



Şekil 2. 1 Terim ve cümlelerin indekslenme durumu

Bu örnekte r=10 alınmıştır. $U_{*1}, U_{*2,}...,U_{*10}$ değerleri dokümandaki terimleri indeksleyen konu vektörlerini ifade etmektedir. Bu matrise göre U ile belirtilen vektörlerin önem derecesi S matrisinde köşegen üzerinde bulunan sayılar ile orantılıdır. Buna göre U_1 en önemli konu vektörünü U_{10} ise önem derecesi en az olan konu vektörünü belirtmektedir. Bu konu vektörleri V^T matrisi baz alındığında bu matrisin satırları ile indekslenir. Metin özetleme aşamasında kullanılan V^T matrisinin kolonları dokümandaki cümleleri ifade etmektedir. V^T matrisinin hücre değerleri incelenen cümlenin incelenen konu ile ilgili olma durumunu yansıtmaktadır. Örneğe bakılacak V^T matrisinin ilk satırı ile indekslenen en önemli konu vektörünün en çok bu satırdaki en büyük değere (0.791) sahip olan ikinci cümle ile temsil edildiği görülür.

Bu örnek ile TDA sonucu elde edilen çarpan matrislerinin dil bilimsel anlamı daha anlaşılır bir halde gösterilmiştir.

2.2 GAA Temelli Metin Özetleme Yöntemleri

Bu bölümde, literatürde karşılaşılan GAA temelli dört yöntem incelenmiştir. İncelenen bu yöntemler Çizelge 2.1'de görülen şekilde kategorize edilmiştir.

Çizelge 2. 1 GAA temelli metin özetleme yöntemleri

Yöntem Adı	**Çalışmanın Yapılma Zamanı**	**Yöntemin Dayandığı Temel Prensibi**
Yöntem1	2001 – Gong ve Lui [40]	Doküman içinde bahsi geçen her farklı konudan bir cümle seçilmesi
Yöntem2	2005 – Murray vd. [41]	Doküman içinde bahsi geçen her farklı konudan birden fazla cümle seçilmesi
Yöntem 3	2007 – Steinberger [42]	Doküman içinde bahsi geçen tüm farklı konuları içeren cümlelerin seçilmesi
Yöntem 4	2011 – Özsoy vd. 2011 [49]	Doküman içinde bahsi geçen tüm farklı konuları içeren cümlelerin gürültüden arındırıldıktan sonra seçilmesi

Bu çalışmaların tamamında amaç her bir haber dokümanı içindeki en önemli cümleleri seçmektir. Bu amaçla yöntemler terim-cümle matrisini oluşturduktan sonra matris üzerinde TDA uygulamış ve elde edilen çarpan matrislerinden V^T matrisini kullanmıştır. Çünkü Bölüm 2.1.2'de de belirtildiği gibi V^T matrisinin satırları doküman içeriğini oluşturan önemli konuları ve sütunları incelenen dokümanı oluşturan cümleleri ifade etmektedir.

Yöntemler önerilen farklı cümle seçim kriterleriyle birbirinden farklılaşmaktadırlar. Bölüm 2.2.1-2.2.4'te bu farklılıklar tüm ayrıntıları ile ele alınmıştır.

2.2.1 Yöntem1- Gong ve Lui [40] Yaklaşımı

Gong ve Liu [40] çalışması TDA'yı metin özetlemeye uyarlayan ilk çalışmadır. Bu çalışmada özete eklenecek olan en önemli s adet cümleyi belirlemek adına TDA uygulandıktan sonra aşağıdaki yaklaşım $k=1$ başlangıç koşulu ile s kere uygulanmıştır:

- V^T matrisindeki en büyük indeks değerine sahip olan $k.$ sağ tekil vektör seçilir. Bu vektör matrisin $k.$ satırında yer alan ve metin içindeki en önemli $k.$ konunun

indekslenmiş olduğu vektördür. Özeti oluşturmak için, bu satırdaki en büyük indeks değerine sahip olan cümle seçilir ve bir sonraki seçim için k bir arttırılır.

- k istenen s sayısına ulaştığında işlem durdurulur, aksi durumda ilk işlem tekrarlanır.

Yaklaşımı daha iyi anlayabilmek için, V^T matrisinin Çizelge 2.2'deki gibi olduğu kabul edilsin. Bu durumda öncelikle ilk satırda bulunan "Konu1" vektöründeki en büyük indeks değeri olan 0.791 sayısına ulaşılmalıdır. Daha sonra bu indeks değerine karşılık gelen cümle tespit edilmelidir. Çizelgeden görülebileceği gibi bu cümle "Cümle2"'dir. Bir sonraki seçim aynı matık ile V^T matrisinin ikinci satırından yapılır. İşlem istenen sayıda cümle özete eklendiği anda durdurulur.

Çizelge 2. 2 Yöntem1 için verilen bir örnek

V^T	Cümle1	**Cümle2**	Cümle3	Cümle4	Cümle5
Konu1	0.458	**0.791**	0.132	0.120	0.332
Konu2	0.246	0.573	0.642	0.246	-0.563
Konu3	0.731	-0.135	0.332	0.156	-0.166
Konu4	0.128	0.434	-0.111	0.265	0.783

Bu yaklaşımda özete s adet cümle eklenmek istediğinde ilk s konu vektörü dikkate alınmaktadır. Bu durumda özette bahsi geçecek olan önemli konu sayısı, özeti oluşturacak cümle sayısı ile aynı olmaktadır. Dolayısıyla özete eklenecek olan cümle sayısı arttıkça, özete önem derecesi daha düşük olan cümleler eklenmiş olur. Murray vd. [41], çalışmalarında bu durum üzerinde yoğunlaşmış ve ana dokümandaki önemli konu sayısından bağımsız bir özetleme sistemi oluşturmaya çalışmışlardır.

2.2.2 Yöntem2 – Murray vd. [41] Yaklaşımı

Murray vd. [41] çalışmalarında V^T matrisini elde ettikten sonra bu matrisin her bir satırından bir cümle seçmek yerine birden fazla cümle seçmeyi önermiştir. İlgili satırdan (konudan) kaç adet cümle seçileceği S matrisinde ilgili öz değerin geri kalan öz değerlerin toplamının yüzde kaçını belirttiğine bağlı olarak değişmektedir. Bu yöntemle oluşturulan özetler ana dokümandaki herbir önemli konuya ait birden fazla cümleyi barındırmaktadırlar.

V^T matrisi Çizelge 2.3'deki matris olsun. [41] çalışmasında ilk satırdan en yüksek indeks değerine sahip olan birden fazla cümle seçilir.

Çizelge 2. 3 Yöntem2 için verilen bir örnek

V^T	Cümle1	Cümle2	Cümle3	Cümle4	Cümle5
Konu1	0.458	**0.791**	0.132	0.120	0.332
Konu2	0.246	0.573	0.642	0.246	-0.563
Konu3	0.731	-0.135	0.332	0.156	-0.166
Konu4	0.128	0.434	-0.111	0.265	0.783

2.2.3 Yöntem3 – Steinberger [42] Yaklaşımı

Gong ve Liu [40] yaklaşımında, k değerinin artması önem derecesi daha düşük olan cümlenin seçilmesi anlamına gelmektedir. Steinberger [42] doktora çalışmasında bu açığı giderdiğini belirtmiştir. Bu çalışma her bir GAA boyutunun istatistiksel belirginliğinin o boyuta (özvektöre) ait olan öz değerin karesi ile ölçüldüğü [76] gerçeğini baz alarak cümlelerin seçim kriterini değiştirmiştir. Öncelikle B matrisini tanımlamıştır:

$$B = S^2 V^T \tag{2.16}$$

Daha sonra cümlelerin önem derecesini belirten S_k değerini aşağıdaki şekilde ifade ederek en yüksek S_k değerine sahip olan istenen adet cümleyi özete eklemiştir.

$$S_k = \sqrt{\sum_{i=1}^{r} b_{ik}^2} \tag{2.17}$$

S_k değerine göre cümlelere verilen ağırlık değerleri Çizelge 2.4'deki gibi olsun. Bu durumda en yüksek S_k değerine sahip olan üçüncü cümle, "Cümle3" özete eklenecek olan ilk cümledir.

Çizelge 2. 4 Yöntem3 için verilen bir örnek [40]

V^T	Cümle1	Cümle2	**Cümle3**	Cümle4
Konu1	0.846	0.455	0.562	0.378
Konu2	0.344	0.235	0.632	0.186
Konu3	0.231	0.432	0.735	0.248
Konu4	0.210	0.342	0.857	0.545
S_k değeri	0.432	0.543	**0.723**	0.235

Yapılan değişiklik ile tüm önemli konular ile ilgili bilgiler içeren cümleler eklenmiştir. Steinberger [42], Gong ve Liu [40] yaklaşımı üzerinde gerçekleştirmiş olduğu bu modifikasyonun özetleme sisteminin başarımını yükselttiğini göstermiştir.

2.2.4 Yöntem4 – Özsoy vd. [49] Yaklaşımı

Özsoy vd. [49] yaptıkları çalışma ile Steinberger çalışmasına bir ön işlem aşaması eklemiş ve bu ön işlem aşamasının, uzun metinlerin özetlenmesinde, sistem başarımını arttırdığını göstermişlerdir. Eklenen ön işlem aşamasında V^T matrisinin her satırının ortalaması bulunmuş ve o satıra ait hücre değerleri ortalamadan küçük olduğunda sıfırlanmıştır. S_k değerleri sıfırlama işlemi gerçekleştirildikten sonra her satırdaki hücre değerlerinin toplamıyla bulunmuştur. Bu değişiklik ile aslında önemli konulara ait olan cümle seçiminde cümlelerdeki gürültülerin giderildiğini belirtmişlerdir. Çalışmada yapılan bu değişikliğin uzun dokümanlar üzerinde etkili olduğu belirtilmiştir.

Çizelge 2.10 ile gösterilen V^T matrisi ele alındığında Özsoy diğerleri önerdikleri ön işlem aşamasında ilk önce her satırın ortalamasını bulmuşlardır. Daha sonra her satırda ortalamadan küçük olan değerleri tespit etmişler (ilk satır için 2.41 ve 0.110) ve bu değerleri sıfır kabul ederek, Çizelge 2.6 ile gösterilen şekilde S_k değerlerini hesaplamışlar ve özete en yüksek S_k değerine sahip olan cümleleri eklemişlerdir. Örneğe göre özete eklenecek ilk cümle ikinci cümle, “Cümle2”dir.

Çizelge 2. 5 Yöntem4 için verilen bir örnek [49]

V^T	Cümle1	**Cümle2**	Cümle3	Cümle4	Ortalama
Konu1	0.557	0.691	**0.241**	**0.110**	0.399
Konu2	**0.345**	0.674	0.742	**0.212**	0.493
Konu3	0.732	**0.232**	0.435	**0.157**	0.389
Konu4	**0.628**	**0.436**	0.738	0.865	0.678
Konu5	0.557	0.691	**0.241**	0.710	0.549

Çizelge 2. 6 Yöntem4 ile uygulanan ön işlem aşamasında cümle seçimi

V^T	Cümle1	**Cümle2**	Cümle3	Cümle4
Konu1	0.557	0.691	**0**	**0**
Konu2	**0**	0.674	0.742	**0**
Konu3	0.732	**0**	0.435	**0**
Konu4	**0**	**0**	0.738	0.865
Konu5	0.557	0.691	**0**	0.710
S_k değeri	1.846	**2.056**	1.960	1.575

2.3 GAA Temelli Yöntemlerin Başarımlarını Arttırmak için Önerilen Sistem

GAA temelli metin özetleme yöntemlerinin en önemli aşaması terim-cümle matrisi olan $A=[a_{1i},a_{2i},...,a_{ni}]$ matrisinin yaratılma aşamasıdır. Bu aşama ile matrisin hücre değerleri olan a_{ji} değerleri Bölüm 2.1'de belirtilen ve $T_{frekans}$ ifadesi ile gösterilen, lokal ve global ağırlık değerlerinin çarpımıyla elde edilmektedir. Şu ana kadar literatürde bulunan ve tez çalışmasında Yöntem1, Yöntem2, Yöntem3 ve Yöntem4 olarak isimlendirilen GAA temelli çalışmaların tamamında terim frekansına dayalı olan ağırlık değerleri kullanılmıştır. Benzer ağırlıklar sadece metin özetleme sistemlerinde değil, metin sınıflama alanında da en sık kullanılan ağırlıklardır.

Xue ve Zhou [59] metin sınıflama sistemlerinde bir terimin önem derecesinin belirlenmesi adına frekans bilgisinin tek başına yeterli olmadığını belirtmişlerdir. Bu amaçla çalışmalarında, "Dağıtımsal Özellikler –DÖ" başlığı altında yeni ağırlık değerlerinin kullanılmasını önermişlerdir. DÖ'ler, "Bölüm Yoğunluğu- $Y_{\text{Bölüm}}$", "İlk ve Son Görünüm Farkı Yoğunluğu- $Y_{\text{İlk-Son}}$" ve "Pozisyonlarının Varyansı- Y_{PozVar}" olmak üzere toplam üç adet ölçümü içermektedir. Önerilen DÖ'ler, terimlerin incelenen metin parçaları içindeki dağılımı ile ilgili olup terimlerin yoğunluğu (compactness) hakkında bilgi vermektedir. Çalışmada bir terimin dağılımının, dokümanın belirli bir yerine özgü olması durumda o terimin yoğun bir terim olduğu söylenmiştir [59]. Örneğin A ve B gibi iki doküman olsun. Bu dokümanlardan A dokümanının eğitim sistemlerinde kullanılan taşınabilir bilgisayarlar ile ilgili olduğu, B dokümanının ise taşınabilir bilgisayarların donanımsal parçaları ile ilgili olduğu varsayılsın. Aynı zamanda "taşınabilir bilgisayar" teriminin bu iki dokümanda geçme sıklıklarının eşit ve dokümanlardaki yoğunluklarının farklı olduğu varsayılsın (A dokümanı içinde terimin

dokümanın her yerine yayılmış, B dokümandaki içinde ise terimin dokümanın belirli yerlerine özgü olduğu düşünülsün). Bu durumda terim frekansı bilgisine göre bu terimin iki doküman bazındaki önemi eşit çıkmaktadır. DÖ'lerin kullanımı sayesinde bu önemin eşit olmadığı gösterilebilmektedir.

Terimlerin önem dercelerinin farklı ek bilgilerin kullanımı ile tespit edilebileceğini belirten bir başka bir çalışma Ko vd. [60] çalışmasıdır. Bu çalışmada metin sınıflama sistemleri için terim frekans bilgisine ek olarak, terimlerin ait oldukları cümlelerin önem derecelerini yansıtan bir hesaplama kullanılmış ve bu hesaplamanın metin sınıflama sistemlerinin başarımlarını arttırdığı gösterilmiştir.

Tez çalışmasında, Xue ve Zhou [59] ve Kov d. [60] çalışmalarından esinlenerek, metin özetlemede terim-cümle matrisinin oluşum aşamasında Bölüm 2.1'de bahsi geçen $T_{frekans}$ ağırlık değeri, iki yeni ağırlık değeri ile birlikte kullanılmıştır. Önerilen yeni ağırlıklandırma, terim frekansına dayalı olan $T_{frekans}$ ağırlık değerinin dışında terimin metin içinde bulunma yerine dayalı olan DÖ değerlerinin ortalamasını (ODÖ) $, T_{Dagitimsal} = (\frac{Y_{Bölüm} + Y_{ilk-son} + Y_{pozVar}}{3})$, ve terimin ait olduğu cümlenin önemini yansıtan $\mathrm{C}_{\ddot{o}\mathrm{nem}} = \sum_{i=1}^{13} \ddot{o}_i$ ifadesini içermektedir. İncelenen terimin ait olduğu cümlenin önemin belirlenmesi için on üç adet cümle özelliği incelenmiş ve cümlenin bu özelliklere göre sahip olduğu toplam puan ($\mathrm{C}_{\ddot{o}\mathrm{nem}} = \sum_{i=1}^{13} \ddot{o}_i$) ifadesiyle belirtilmiştir. Sonuç olarak, tez çalışması kapsamında önerilen yeni ağırlık değeri kullanılarak terim-cümle matrisinin a_{ji} değerleri (2.18) eşitliği ile belirtildiği gibi hesaplanmıştır.

$$a_{ji} = \mathrm{Yeni}_{\text{A ğ ı r}} = \{T_{frekans}\} \times \{T_{Dagitimsal}\} \times \{C_{önem}\} \qquad (2.18)$$

Bu yeni hesap Yöntem1, Yöntem2, Yöntem3 ve Yöntem4 olarak isimlendirilen tüm yöntemlerin, tüm veri setleri üzerindeki başarımlarını yükseltmiştir.

Önerilen yeni ağırlık değerinin içeriğini oluşturan $T_{frekans}$, ODÖ'yü içeren $T_{Dagitimsal}$ ve cümle önemini belirten $C_{önem}$ ifadeleri sırasıyla Bölüm 2.3.1, Bölüm 2.3.2 ve Bölüm 2.3.3'de açıklanacaktır.

2.3.1 Terim Frekansına Dayalı Ağırlıklandırma

Terim frekansına dayalı olan $T_{frekans}$ ağırlıklandırma sistemleri Bölüm 2.1'de belirtilen terimlerin cümleler içindeki geçme sayısına bağlı olan özelliklerdir. Bu özellikler (2.1) eşitliği ile gösterilen lokal ve global ağırlık değeri adı verilen değerlerin çarpımıyla elde edilmektedir. Tez çalışması kapsamında Gong Lui [40] çalışmasında kullanılan ve Bölüm 2.1'de açıklanan dört lokal değer ve iki global değer kullanılmıştır.

2.3.2 Terimin Bulunma Yerine Dayalı DÖ'ler ile Oluşturulan Ağırlıklandırma

Xue ve Zhou [59] çalışmasında önerilen DÖ'ler, terimlerin incelenen metin parçaları içindeki dağılımları ile ilgili olup terimlerin bulunma yoğunluğu hakkında bilgi vermektedir.

Bir terimin yoğunluğu, o terimin doküman içindeki dağılımına göre belirlenen bir özelliktir ve bu dağılım incelenen terimin cümleler içindeki frekansını tutan bir vektörle gösterilir. Bu vektör $dizi(t,d) = \{c_0, c_1, ..., c_{n-1}\}$ ifadesi ile gösterilsin. Burada c_i değeri, terim t'nin i indeksli cümle içindeki frekans bilgisidir. Bu vektör kullanılarak dokümanı oluşturan tüm terimlerin yoğunlukları hesaplanabilir.

DÖ'ler terimlere ait dağılımsal vektörlerinin kullanılması ile elde edilen "Bölüm Yoğunluğu- $Y_{\text{Bölüm}}$", "İlk ve Son Görünüm Yoğunluğu- $Y_{\text{İlk-Son}}$" ve "Pozisyonlarının Varyansı- Y_{PozVar}" olmak üzere toplam üç adet ölçümü içermektedir. Bu değerleri daha iyi anlayabilmek için örnek olarak alınan "tencere" kelimesinin bir doküman içindeki dağılımsal vektötürünün $dizi(tencere,d) = \{2,1,0,0,1,0,0,3,0,1\}$ olduğunu varsayalım. Bu vekörün ilk elemanı olan 2, "tencere" kelimesinin dokümanın ilk cümlesinde iki kez geçtiği bilgisini, ikinci elemanı olan 1 ise kelimenin dokümanın ikinci cümlesinde bir kez geçtiği bilgisini tutmaktadır. Diğer elemanlar için de benzer açıklamalar geçerlidir. Artık bu üç özellik kolaylıkla anlaşılabilir:

- **Bölüm Yoğunluğu- $Y_{\text{Bölüm}}$:** Bir terimin bölüm yoğunluğu $Y_{Bölüm} = \sum_{i=0}^{n-1} c_i \succ 0 \,?\, 1 : 0$

formülü ile hesaplanmaktadır. Burada c_i değeri, terim t'nin i indeksli cümle içindeki geçme sıklığı bilgisidir. Bu formüle göre "tencere" kelimesinin $Y_{Bölüm}$ değeri şu şekilde hesaplanır: $Y_{Bölüm}(\text{tencere}, d) = 1+1+0+0+1+0+0+1+0+1=5$.

Bu ölçüm bir terimin yoğun bir terim olup olmadığını ölçmektedir. Eğer bir terim bir dokümanın farklı yerlerinde dağılmış olarak bulunuyorsa, bu o terimin az yoğun bir terim olduğunu göstermektedir [51].

- **İlk ve Son Görünüm Yoğunluğu-** $Y_{\text{İlk-Son}}$: Bu özellik ile bir terim yoğunlunun

hesaplanması için terimin ilk ve son görünümleri arasındaki fark alınmıştır. Eğer terim az yoğun bir terim ise terimin ilk ve son kullanımı arasındaki fark fazladır. İlk ve son konum farkı özelliği aşağıdaki şekilde hesaplanır:

$$Y_{İlk-Son}(t,d) = Son_{Gorunum}(t,d) - İlk_{Gorunum}(t,d) \quad (2.19)$$

$$İlk_{Gorunum}(t,d) = \min_{i\in\{0..n-1\}} c_i \succ 0 ? i : n \quad (2.20)$$

$$Son_{Gorunum}(t,d) = \max_{i\in\{0..n-1\}} c_i \succ 0 ? i : -1 \quad (2.21)$$

Yine "tencere" kelimesi baz alındığında

- $İlk_{Gorunum}(tencere, d) = \min\{0,1,10,10,4,10,7,10,9\} = 0$ olarak bulunurken,

- $Son_{Gorunum}(tencere, d) = \max\{0,1,-1,-1,4,-1,-1,7,-1,9\} = 9$ olarak bulunur ve nihayet ilk ve son görünüm yoğunluğu;

- $Y_{İlk-Son}(tencere, d) = Son_{Gorunum}(tencere, d) - İlk_{Gorunum}(tencere, d) = 9 - 0 = 9$ olarak bulunur.

- **Pozisyonların Varyansı:** Bu özellikte incelenen terimin yoğunluk hesabı için

terimin tüm görünüşlerinin varyansları kullanılmıştır. İlk önce tüm görünüşlerinin ortalaması alınmıştır. Daha sonra her bir görünüşün ortalama görünüşten farkı alınmıştır ve elde edilen ortalama konum bilgisi pozisyonların varyansı olarak isimlendirilmiştir:

$$Y_{PozVar}(t,d) = \frac{\sum_{i=0}^{n-1} c_i * |i - merkez(t,d)|}{say(t,d)} \quad (2.22)$$

$$say(t,d) = \sum_{i=0}^{n-1} c_i \quad ve \quad merkez(t,d) = \frac{\sum_{i=0}^{n-1} c_i * i}{say(t,d)} \quad (2.23)$$

Yine "tencere" kelimesi örnek olarak alındığında:

- $say(tencere,d) = 2+1+1+3+1 = 8$ olarak bulunurken

- $merkez(tencere,d) = (2\times 0+1\times 1+1\times 4+3\times 7+1\times 9)/8 = 4.375$ olarak hesaplanır ve nihayet pozisyonların varyansı;

- $Y_{PozVar}(tencere,d) = (2\times 4.375+1\times 3.375+1\times 0.375+3\times 2.625+1\times 4.625)/8 = 3.125$ olarak bulunur.

Sonuç olarak tez çalışması kapsamında, Xue ve Zhou [59] çalışmasında ait olan ODÖ'sü $,T_{Dagitimsal} = (\frac{Y_{Bölüm} + Y_{ilk-son} + Y_{pozVar}}{3})$, (2.18) eşitliği ile belirtilen ağırlık değerinde bir çarpan olarak kullanılmıştır.

2.3.3 Terimin Bulunduğu Cümle Önemine Dayalı Ağırlıklandırma

Bu bölümde (2.18) eşitliği ile gösterilen ve yeni ağırlıklandırma sisteminde bir çarpan olarak kullanılan $C_{önem}$ ifadesi anlatılacaktır. Bu ifade, bir terimin ait olduğu cümlenin önemini belirtmektedir. Bir cümlenin doküman içindeki önemi, cümlenin bilgi içerme miktarına bağlıdır. Bu durumu irdelemek için cümlenin doküman içindeki yapısal on üç özelliği incelenmiş ve cümleye her bir özellik ile ilgili ayrı ayrı puanlar verilmiştir. Daha sonra tüm puanlar 0-1 aralığına getirilerek normalize edilmiş ve toplanarak incelenen cümleye atanmıştır ($C_{önem} = \sum_{i=1}^{13} ö_i$). Cümle önemini belirten özellikler aşağıdaki alt başlıklarda anlatılmıştır:

•$ö_1$-Cümle Konumu:

Edmundson [2] tarafından ilk kez ortaya atılan bu özellik doküman içindeki belli cümlelerin konu belirtme olasılıklarının yüksek olması temeline dayanılarak uygulanmıştır. Tez çalışmasında dokümanı oluşturan her bir cümleye cümlenin konumuna göre (2.24) eşitliğindeki formüle göre bir skor değeri verilmiştir:

$$Skor_{ö_1}(C_i) = \frac{N - P_i}{N} \tag{2.24}$$

Bu formülde N dokümandaki toplam cümle sayısı iken P_i değeri cümlenin doküman içindeki kaçıncı cümle olduğu bilgisidir.

•$ö_2$ - İlk Cümleye olan Benzerlik:

Bu özellik dokümandaki her bir cümleye cümlenin dokümandaki ilk cümleye olan benzerliğine göre bir skor değeri vermektedir. Cümleler arasındaki benzerlik kosinüs benzerliğine göre hesaplanmaktadır.

$$Skor_{ö_2}(C_i) = \text{kosinüs}(C_i, C_1) = \frac{çarp(C_i, C_1)}{\|C_i\|\|C_1\|} \tag{2.25}$$

Burada iki vektör arasındaki kosinüs bağlantısı için iki vektörün çarpımının iki vektörün boylarının çarpımına oranı alınmıştır.

•$ö_3$ - Son Cümleye olan Benzerlik:

Bu özellik dokümandaki her bir cümleye cümlenin dokümandaki son cümleye olan benzerliğine göre bir skor değeri vermektedir. Cümleler arasındaki benzerlik kosinüs benzerliğine göre yapılmaktadır.

$$Skor_{ö_3}(C_i) = \text{kosinüs}(C_i, C_{son}) = \frac{çarp(C_i, C_{son})}{\|C_i\|\|C_{son}\|} \tag{2.26}$$

•$ö_4$ - Başlığa olan Benzerlik:

Bu özellik dokümandaki her bir cümleye cümlenin dokümandaki başlığa olan benzerliğine göre bir skor değeri vermektedir. Cümleler arasındaki benzerlik kosinüs benzerliğine göre yapılmaktadır.

$$Skor_{ö_4}(C_i) = \text{kosinüs}(C_i, C_{baslik}) = \frac{çarp(C_i, C_{baslik})}{\|C_i\|\|C_{baslik}\|} \quad (2.27)$$

•$ö_5$ – Cümle Uzunluğu:

Uzun cümleler kısa cümlelere göre daha çok bilgi içerirler. Bu nedenle dokümandaki her bir cümleye, cümlenin sahip olduğu kelime sayısına göre bir skor değeri verilmektedir.

• $ö_6$ - Kelime Sıklığı Bilgisi:

Bu özeliği kullanmak için metin içerisinde yer alan her terimin frekansı hesaplandıktan sonra, bir frekans listesi oluşturulur. Yüksek frekanstan düşük frekansa doğru sıralanmış olan bu liste, metindeki her bir kelimeyi ve geçme sıklığını temsil eder. Bu listeye "ve, veya, ile, için, vd. gibi" tek başına bir anlamı olmayan kelimeler dahil edilmemiştir. Liste oluşturulduktan sonra en yüksek frekansa sahip kelimeler (listenin %10'u) özetleme işleminde dikkate alınmaktadır. Cümlenin, bu en yüksek frekanslı kelimeleri içerip içermediği incelenir. Eğer cümle bu kelimeleri içeriyorsa içerdiği yüksek kelime frekansları toplanarak cümleye bir skor değeri atanmaktadır.

•$ö_7$ -Kelime Cümle Skoru Bilgisi:

[13] referansından alınan bu özelliğe göre, eğer tj terimini içeren cümle sayısı >= 1/2LS ise kelime cümle skoru (KCS) (2.28) eşitliği ile belirtilen şekilde hesaplanır.

$$KCS = 0.1 + \frac{\sum_{t_j} \in S_i W_{ij}}{HTFS} \quad (2.28)$$

Burada;

0.1: Terimin önemli olmaması durumunda cümleye verilecek minimum skor

W_{ij}: S_i cümlesine ait t_{ij} teriminin terim ağırlığıdır (TS-TTS). Burada TS-TTS, terim sıklığı – ters terim sıklığı ifadesinin kısaltılmış halidir. Terim sıklığı (ts), t_{ij} teriminin S_i cümlesindeki frekansı iken ters terim sıklığı (tts) aşağıdaki şekilde hesaplanır:

$$tts = 1 - \lfloor \log(cf(t_{ij}) + 1) / \log(n+1) \rfloor \quad ise \quad TS - TTS = ts_{ij} * tts \quad (2.29)$$

Yukarıdaki ifadede c_f değeri t_{ij} terimini içeren cümle sayısını, n ise metindeki toplam cümle sayısını göstermektedir.

LS: Özet uzunluğudur ve HTFS değeri doküman içindeki maksimum TS-TTS değeridir.

$$W_{ij} = s_{ij} * tts = ts(t_{ij}, s_i)\left[1 - \frac{\log\left(cf\left(t_{ij}\right)+1\right)}{\log(n+1)}\right] \quad (2.30)$$

• ö$_8$ -Ortalama Kelime Frekansı ve Ters Doküman Frekansı-(OKF-TDF):

Bu özellik terim sıklığı bilgisini sadece doküman bazında değil, veri seti içindeki tüm dokümanlar bazında incelemektedir ve üç farklı kabule dayanmaktadır [16]: i) Bir terimin önemi o terimin bulunduğu dokümandaki frekans bilgisiyle doğru orantılıdır. ii) Bir dokümanın uzunluğu bir terimin önem derecesini etkilemez. iii) Bir terim bir veri seti grubunda az sayıda geçiyorsa, o terim daha önemli bir terimdir. Bu kabuller altında $ts(d,t)$, bir terimin incelenen dokümandaki geçme sıklığını; $\max - ts$, bir dokümanda en çok geçen terimin geçme sıklığını; d, veri setinde bulunan toplam doküman sayısını ve $ds(v,t)$, terimin veri seti (v) içindeki kaç dokümanda geçtiği bilgisini ifade etmek üzere (OKF-TDF) aşağıdaki şekilde hesaplanır:

$$OKF - TDF = \frac{ts(d,t)}{\max - ts} \times \log\left(\frac{d}{ds(v,t)}\right) \quad (2.31)$$

•ö$_9$ – Sayısal Karakter İçerme Durumu:

Bu özelik ile cümlelere içerdikleri toplam nümerik karakter sayısı kadar puan verilir.

•ö$_{10}$ - “?” ve “!” İçerme Durumu:

Bir cümlenin ünlem işareti veya soru işareti ile bitmesi diğer cümlelere göre daha önemli olduğunun bir işaretidir. Buna göre eğer cümle soru işareti ya da ünlem işareti içeriyorsa cümleye bir puan verilir.

•ö$_{11}$ – Pozitif Kelimeleri İçerme Durumu:

Bu özellik ile cümlelerin “özetle”, “sonuçta”, “neticede” gibi toparlayıcı kelimeleri içerip içermediği incelenir. Buna göre cümlelere içerdikleri toplam pozitif kelime sayısı kadar puan verilir.

•ö$_{12}$- İsim Soylu Kelimeleri İçerme Durumu:

Metinlerde yer alan isimler, metinin içeriği hakkında bilgi vermektedir. Bu yüzden metin özetleme sistemi isimlerin geçtiği cümlelere sahip olduğu isim sayısı kadar puan vermektedir. Metinler içindeki terimlerin isim olup olmadığı Türkçe dokümanları içeren

veri setlerinde Zemberek yazılımı [61] kullanılarak tespit edilmiştir. İngilizce dokümanları içeren veri setlerinde ise Stanford Üniversitesinin tasarladığı söz dizimsel analiz aracı kullanılmıştır [62]. Aynı zamanda İngilizce terimlerin köklere ayrılması için "Porter Stemmer [63]" kök bulma kodu kullanılmıştır.

•$ö_{13}$-Merkezilik Özelliği:

Bir cümlenin merkezilik özelliği [13] referanslı çalışmada kullanılan bir özelliktir ve toplam üç özelliği barındırmaktadır. Bu özellikler: toplam benzerlik ($\sum_{j=1}^{n-1} benzerlik(C_i, C_j)$), yakınlık sayısı ($\sum_{j=1}^{n-1} yakinlik(C_i, C_j)$) ve sahip olunan Ngram sayısıdır ($\sum_{j=1}^{n-1} Ngram(C_i, C_j)$).

$i \neq j \quad ve \quad benzerlik(C_i, C_j) \geq \phi$ koşulları altında (ϕ=0.03) merkezilik özelliği (2.32) eşitliği ile belirtilen şekilde hesaplanmaktadır:

$$Skor(S_i)_{ö10} = \frac{\sum_{j=1}^{n-1} benzerlik(C_i, C_j) + \sum_{j=1}^{n-1} yakinlik(C_i, C_j) + \sum_{j=1}^{n-1} Ngram(C_i, C_j)}{n-1} \quad (2.32)$$

Aşağıdaki başlıklarda toplam benzerlik, yakınlık sayısı ve sahip olunan Ngram sayısı özellikleri anlatılacaktır:

-Toplam Benzerlik:

Toplam benzerlik dokümanda bulunan bir cümlenin diğer tüm cümlelere olan kosinüs benzerliklerinin toplamıdır.

$$benzerlik(C_i, C_j) = \text{kosinüs}(C_i, C_j) = \frac{çarp(C_i, C_j)}{\|C_i\| \|C_j\|} \quad (2.33)$$

-Yakınlık Sayısı:

Yakınlık sayısı i ve j birbirinden farklı iken aşağıdaki şekilde bulunmaktadır:

$$yakinlik(C_i, C_j) = \frac{C_i(yakinlari) \cap C_j(yakinlari)}{C_i(yakinlari) \cup C_j(yakinlari)} \quad (2.34)$$

Burada $C_i(yakinlari)$ ifadesi C_i cümlesine benzerlik değeri belirli bir sınırın üzerinde olan cümleler listesidir. Benzerlik ölçümündeki sınır değeri deneyimsel olarak belirlenmiş ve 0.03 alınmasının uygun olduğu görülmüştür.

-Sahip Olunan Ortak Ngram Sayısı:

Ortak Ngram sayısı i ve j birbirinden farklı iken aşağıdaki şekilde bulunmaktadır:

$$Ngram(C_i, C_j) = \frac{C_i(Ngram) \cap C_j(Ngram)}{C_i(Ngram) \cup C_j(Ngram)} \quad (2.35)$$

Burada $C_i(Ngram)$ ifadesi C_i cümlesinin içerdiği Ngram listesini barındırmaktadır.

2.4 Sonuçlar

Bu bölümde Yöntem1, Yöntem2, Yöntem3 ve Yöntem4 çalışmaları Bölüm 1.3.3'de belirtilmiş olan toplam dört veri seti (VeriSeti-1, VeriSeti-2, VeriSeti-3 ve VeriSeti-4) üzerinde test edilmiştir. Ayrıca tez çalışması kapsamında önerilen yeni ağırlıklandırma sistemi tüm yöntemler üzerinde denenmiş ve önerilen sistemin yöntem başarımlarını arttırdığı gözlemlenmiştir. Yöntem uygulamalarına geçmeden önce dokümanlar ön işlem aşamalarından geçirilmiştir. Türkçe dokümanlara ait olan kelimeler Zemberek [61] ile İngilizce dokümanlara ait kelimeler ise Porter Stemmer [63] algoritması ile köklerine ayrılmış ve tüm veri setlerindeki dokümanlar durak kelimelerinden arındırılmıştır. Durak kelimeleri dokümanlarda sık geçen ve bilgi taşıyıcılığı önemsenmeyen bazı bağlaç, edat ve zamirleri içeren kelimelerdir.

Uygulamalar öncelikle hazırlanan iki faklı Türkçe veri seti (VeriSeti-1 ve VeriSeti-2) üzerinde denenmiştir. Bu veri setlerinden ilki (VeriSeti-1) Bölüm 1.3.3.1'de anlatılan toplam 130 haber dokümanını ve bu haber dokümanlarının üçer kişi tarafından çıkarılmış özetlerini içermektedir. VeriSeti-1'de bulunan haber metinlerinin uzunluğu Veri Seti-2'de bulunan haber metinlerin uzunluğundan daha fazladır. Böylelikle önerilen yeni ağırlık biçiminin, hem uzun hem de kısa haber metinleri üzerindeki etkileri incelenmiştir.

Çizelge 2.7 ile Çizelge 2.9 arasındaki çizelgeler yöntem başarımlarını (sırasıyla birinci, ikinci ve üçüncü özetleyici esas alınarak) sergilemektedir. Bu değerler, yöntemlerin çıkartmış oldukları özetler ile ideal özetler arasındaki çakışan cümle sayısına göre belirlenen ve Bölüm 1.3.4.1'de ayrıntı ile ele alınmış olan F-ölçüm değerine göre elde edilmişlerdir.

Çizelgelerde kırmızı şeridin üst kısmı, yöntemlerin terim-cümle matrisleri yaratılırken, sadece terim frekansına dayalı olan $T_{frekans}$ ağırlık değerlerinin kullanılmasıyla elde edilen performans sonuçlarını göstermektedir. Burada kolon başlıkları Bölüm 2.1'de belirtilen lokal ve global ağırlık değeri seçeneklerini göstermektedir. Örneğin, "İS" ifadesi lokal ağırlık değeri olarak "İkili", global ağırlık değeri olarak ise "Sabit" ağırlık değerinin kullanıldığını göstermektedir.

Kırmızı şeridin altındaki kısım, tez çalışmasında önerilen ve (2.18) eşitliği ile belirtilen $\{T_{frekans}\}\times\{T_{Dagitimsal}\}\times\{C_{önem}\}$ ağırlık değerlerinin bir arada kullanılması ile elde edilen yöntem başarım sonuçlarını ifade etmektedir. Belirtilen çizelgelerde kolon başlıkları, kullanılan $T_{frekans}$ çeşidini ve çarpan olarak eklenilen yeni iki ifadeyi içermektedir. Örneğin "LT-ODÖ-C" ifadesi terim ağırlıklandırma aşamasında "Logaritmik" lokal ağırlığının, "Ters Doküman Sıklığı" global ağırlığının, dağıtımsal özellik ortalamasının (ODÖ) ve cümle öneminin (C) kullanıldığını göstermektedir.

Çizelgelerin sonunda bulunan "Ortalama" kolonu ise çizelgenin her bir satırının ortalamasını göstermektedir. Yani "Ortalama" kolonu, yöntemlerin farklı ağırlık değerlerinin kullanılması ile elde edilen ortalama başarımlarını göstermektedir.

Çizelge 2. 7 Önerilen ağırlığın yöntemler üzerindeki etkisi (İlk özetleyici)

YÖNTEMLER	iS	İT	BS	BT	LS	LT	FS	FT	Ortalama
Yöntem1	0,4897	0,4836	0,5013	0,4964	0,4937	0,482	0,5002	0,478	0,490613
Yöntem2	0,4423	0,4396	0,4956	0,4622	0,4265	0,4093	0,4476	0,4268	0,443738
Yöntem3	0,5405	0,501	0,5321	0,4999	0,5488	0,5111	0,5425	0,5046	0,522563
Yöntem4	0,53	0,4816	0,53	0,4936	0,5484	0,4901	0,5558	0,4973	0,51585
YÖNTEMLER	**İS-ODÖ-C**	**İT-ODÖ-C**	**BS-ODÖ-C**	**BT-ODÖ-C**	**LS-ODÖ-C**	**LT-ODÖ-C**	**FS-ODÖ-C**	**FT-ODÖ-C**	**Ortalama**
Yöntem1	0,4959	0,5239	0,5325	0,5135	0,5222	0,5338	0,5362	0,5141	0,521513
Yöntem2	0,4925	0,4932	0,512	0,5102	0,5026	0,5198	0,5136	0,5051	0,506125
Yöntem3	0,5425	0,5528	0,5425	0,5456	0,5447	0,5521	0,5398	0,5502	0,546275
Yöntem4	0,5383	0,5346	0,5432	0,5346	0,5331	0,5397	0,526	0,5389	0,53605

Çizelge 2. 8 Önerilen ağırlığın yöntemler üzerindeki etkisi (İkinci özetleyici)

YÖNTEMLER	iS	İT	BS	BT	LS	LT	FS	FT	Ortalama
Yöntem1	0,4562	0,4502	0,442	0,4429	0,4512	0,4276	0,4497	0,4287	0,4436
Yöntem2	0,4397	0,4452	0,4564	0,4459	0,4319	0,42	0,4286	0,4338	0,4377
Yöntem3	0,4909	0,4676	0,4758	0,4579	0,4881	0,4668	0,485	0,4635	0,4745
Yöntem4	0,4851	0,4587	0,48	0,453	0,4555	0,445	0,4692	0,4532	0,4625
YÖNTEMLER	**İS-ODÖ-C**	**İT-ODÖ-C**	**BS-ODÖ-C**	**BT-ODÖ-C**	**LS-ODÖ-C**	**LT-ODÖ-C**	**FS-ODÖ-C**	**FT-ODÖ-C**	**Ortalama**
Yöntem1	0,4638	0,4909	0,4794	0,4684	0,4654	0,4829	0,486	0,4926	0,4787
Yöntem2	0,4715	0,4649	0,482	0,4853	0,4687	0,4803	0,4673	0,4687	0,4736
Yöntem3	0,4931	0,498	0,4896	0,4806	0,4847	0,4876	0,4779	0,4894	0,4876
Yöntem4	0,4889	0,49	0,489	0,4796	0,488	0,4823	0,488	0,4912	0,4871

Çizelge 2. 9 Önerilen ağırlığın yöntemler üzerindeki etkisi (Üçüncü özetleyici)

YÖNTEMLER	iS	İT	BS	BT	LS	LT	FS	FT	Ortalama
Yöntem1	0,3839	0,3817	0,3583	0,3468	0,3913	0,3866	0,3868	0,3835	0,377363
Yöntem2	0,3386	0,3709	0,3777	0,3823	0,3408	0,345	0,3402	0,3452	0,355088
Yöntem3	0,4069	0,3932	0,4016	0,3784	0,4191	0,3906	0,4058	0,3959	0,398938
Yöntem4	0,4	0,3722	0,3962	0,3758	0,3923	0,3718	0,3937	0,3803	0,385288
YÖNTEMLER	**İS-ODÖ-C**	**İT-ODÖ-C**	**BS-ODÖ-C**	**BT-ODÖ-C**	**LS-ODÖ-C**	**LT-ODÖ-C**	**FS-ODÖ-C**	**FT-ODÖ-C**	**Ortalama**
Yöntem1	0,4516	0,4429	0,4615	0,451	0,4536	0,45	0,4602	0,4452	0,452
Yöntem2	0,391	0,4045	0,4307	0,4428	0,3883	0,4151	0,3972	0,4043	0,409238
Yöntem3	0,443	0,448	0,4644	0,4559	0,4484	0,4424	0,4431	0,4414	0,448325
Yöntem4	0,4433	0,4492	0,47	0,4598	0,4452	0,4437	0,4249	0,4391	0,4469

Yukarıda belirtilen üç çizelgeden de görüldüğü gibi tez çalışmasında önerilen ağırlık değerinin kullanılması, hemen hemen her farklı ağırlık değeri kombinasyonu için, tüm

yöntemlerin başarımını arttırmıştır. GAA temelli metin özetleme yöntemlerinde terimlerin frekans bilgisine ek olarak, terimlerin bulunma yerleri ve ait oldukları cümlelerin önem değerlerinin eklenmesi olumlu bir etki göstermiştir. Bu olumlu etki çizelgelerin son kolonundaki ortalama değerlerden de açıkça görülmektedir. Yine "Ortalama" kolonuna göre uzun haber metinlerini içeren bu veri seti üzerinde en başarılı olan yöntem Steinberger [42] çalışmasına ait olan Yöntem3'tür. Yöntemlerdeki başarı sıralaması genelde "Yöntem3" > "Yöntem4" > "Yöntem1" > "Yöntem2" şeklindedir.

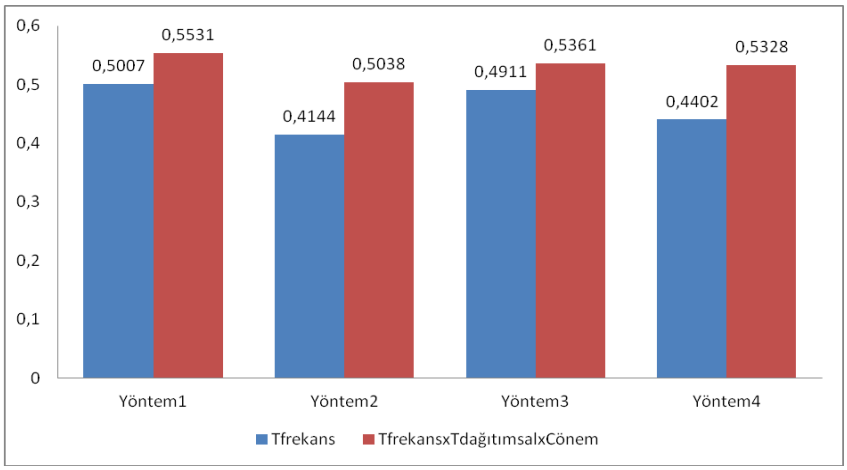
Çizelge 2.10 yöntemlerin VeriSeti-2 üzerindeki sonuçlarını göstermektedir. Bölüm 1.3.3.2'de anlatıldığı gibi bu veri seti toplam 20 adet haber dokümanını ve her bir dokümanın 30 kişi tarafından çıkarılmış olan özetlerini içermektedir. Bu veri setine ait olan haber dokümanları daha kısa haberleri içermektedir. Bu veri setindeki amaç yeni ağırlıklandırmanın kısa dokümanlar üzerindeki etkisini incelemektir.

Bu çizelgede kırmızı şeridin sol tarafı, yöntemlerin terim-cümle matrisleri yaratılırken, sadece terim frekansına dayalı olan $T_{frekans}$ ağırlık değerleri içinden lokal ağırlık değerleri olarak "Logaritmik Ağırlık-(L)" değerinin ve global ağırlık değerleri olarak "Ters Doküman Sıklığı- (T)" değerinin kullanılmasıyla elde edilen başarım sonuçlarını göstermektedir. Sağ tarafı ise "LT-ODÖ-C" ağırlık değerinin kullanılması ile elde edilen başarım sonuçlarını göstermektedir. Sonuçlardan görüleceği gibi tezde önerilen yeni ağırlıklandırma kısa dokümanlar üzerinde de etkili olmuş, hemen hemen her özetleyici için sistem başarımının arttığı görülmüştür.

Çizelge 2. 10 Önerilen ağırlık değerinin VeriSeti-2 üzerindeki etkisi

	Yöntem1	Yöntem2	Yöntem3	Yöntem4	Yöntem1	Yöntem2	Yöntem3	Yöntem4
Bay1	0,5533	0,4658	0,5450	0,5092	0,5383	0,5658	0,5925	0,6050
Bay2	0,5075	0,4450	0,5242	0,4617	0,6383	0,5367	0,5883	0,5717
Bay3	0,5033	0,3558	0,5033	0,4867	0,6133	0,5658	0,6175	0,6300
Bay4	0,5350	0,3992	0,5142	0,4117	0,5908	0,5025	0,5758	0,5758
Bay5	0,5075	0,4533	0,5367	0,4950	0,5967	0,5092	0,5633	0,5717
Bay6	0,4517	0,3558	0,4225	0,3225	0,5117	0,4992	0,4950	0,5158
Bay7	0,6467	0,5158	0,6425	0,4767	0,6133	0,5342	0,5633	0,5758
Bay8	0,4575	0,3992	0,4575	0,4075	0,5008	0,4550	0,4675	0,4425
Bay9	0,4975	0,3450	0,5183	0,4283	0,5450	0,5033	0,5908	0,5742
Bay10	0,5133	0,4158	0,4842	0,3658	0,5108	0,4592	0,5233	0,5192
Bay11	0,4308	0,4408	0,3558	0,4075	0,4617	0,4492	0,4408	0,4533
Bay12	0,5183	0,3908	0,4600	0,4200	0,5508	0,4842	0,5133	0,5383
Bay13	0,4408	0,3408	0,4158	0,3242	0,4950	0,4050	0,4717	0,4800
Bay14	0,5658	0,4408	0,5617	0,4783	0,5992	0,5592	0,5633	0,5342
Bay15	0,4783	0,4183	0,4575	0,4158	0,4933	0,4767	0,5075	0,4658
Bayan1	0,4325	0,3033	0,4033	0,3408	0,6025	0,4175	0,5092	0,5050
Bayan2	0,5142	0,4000	0,5558	0,4725	0,6117	0,5800	0,5800	0,5508
Bayan3	0,3933	0,3250	0,3600	0,3892	0,5325	0,4658	0,4742	0,4992
Bayan4	0,5142	0,4617	0,5350	0,5033	0,6117	0,5200	0,5242	0,5200
Bayan5	0,4850	0,4450	0,5183	0,4592	0,4800	0,5033	0,4992	0,5242
Bayan6	0,4533	0,3925	0,4242	0,4242	0,5550	0,4842	0,5258	0,5092
Bayan7	0,5408	0,4758	0,5367	0,4817	0,4575	0,4950	0,4742	0,4742
Bayan8	0,6350	0,5158	0,6433	0,5117	0,5933	0,5075	0,5783	0,5617
Bayan9	0,4700	0,4325	0,4825	0,4825	0,5692	0,5233	0,5792	0,5417
Bayan10	0,5308	0,5492	0,5142	0,5617	0,4975	0,5033	0,4992	0,4992
Bayan11	0,4975	0,3725	0,5183	0,4725	0,6408	0,5950	0,6075	0,5908
Bayan12	0,4683	0,3542	0,4225	0,3975	0,5575	0,5450	0,5408	0,5783
Bayan13	0,4933	0,4117	0,4350	0,3842	0,4575	0,4158	0,4800	0,4758
Bayan14	0,4850	0,3642	0,4475	0,4200	0,5508	0,5217	0,5608	0,5483
Bayan15	0,4992	0,4475	0,5367	0,4933	0,6175	0,5325	0,5758	0,5508
Ortalama	**0,5007**	**0,4144**	**0,4911**	**0,4402**	**0,5531**	**0,5038**	**0,5361**	**0,5328**

Otuz kişinin başarım ortalaması alındığında önerilen ağırlık değerinin kullanılmasının başarım üzerindeki olumlu etkileri daha rahat görülebilir.



Şekil 2. 2 Sunulan önerinin VeriSeti-2 üzerindeki etkisi

Şekilden 2.2'den de görüleceği gibi önerilen yeni ağırlık değerinin kısa haber dokümanlarından oluşan VeriSeti-2 üzerinde kullanılması tüm yöntemlerin ortalamadaki başarımlarını arttırmıştır. Bu veri setin üzerindeki en yüksek başarım 0.5531 ile Yöntem1'e aittir.

Yöntem performansları Türkçe veri setlerinden sonra yine Bölüm 1.3.3.3 ve Bölüm 1.3.3.4'de belirtilen iki İngilizce veri seti (VeriSeti-3 ve VeriSeti-4) üzerinde denenmiştir. Çizelde 2.11 yöntem başarımlarının VeriSeti-3 üzerindeki etkilerini göstermektedir. VeriSeti-3 haber portallarından toplanmış 92 adet haber metnini ve bu haber metinlerine ait 92 adet yoruma dayalı olmayan haber özetlerini içermektedir.

Çizelge 2. 11 Önerilen ağırlık değerinin VeriSeti-3 üzerindeki etkisi

YÖNTEMLER	iS	İT	BS	BT	LS	LT	FS	FT	Ortalama
Yöntem1	0,3477	0,3607	0,342	0,3484	0,3598	0,3404	0,3479	0,3429	0,3487
Yöntem2	0,3728	0,3683	0,3457	0,3483	0,3135	0,3469	0,3045	0,3299	0,3412
Yöntem3	0,3671	0,3941	0,3073	0,3647	0,31	0,3617	0,2419	0,3368	0,3355
Yöntem4	0,3817	0,3833	0,3075	0,3593	0,3029	0,3469	0,2405	0,3183	0,3301
YÖNTEMLER	**İS-ODÖ-C**	**İT-ODÖ-C**	**BS-ODÖ-C**	**BT-ODÖ-C**	**LS-ODÖ-C**	**LT-ODÖ-C**	**FS-ODÖ-C**	**FT-ODÖ-C**	**Ortalama**
Yöntem1	0,4175	0,4225	0,4256	0,4058	0,433	0,4421	0,4164	0,4338	0,4246
Yöntem2	0,381	0,414	0,3301	0,3845	0,3469	0,4069	0,3303	0,3452	0,3674
Yöntem3	0,407	0,4065	0,3375	0,3905	0,3147	0,3752	0,2674	0,3194	0,3523
Yöntem4	0,3971	0,4035	0,3353	0,3908	0,3079	0,3738	0,2682	0,3197	0,3495

Çizelge 2.11'de görüleceği gibi yeni ağırlıklandırma sistemi İngilizce veri seti üzerinde denendiğinde tüm yöntemlerin başarımını arttırmışır. "Ortalama" kolonu altındaki değere bakıldığında yöntemlerın bu veri üzerindeki performans sıralaması Yöntem1 > Yöntem2 > Yöntem3 > Yöntem4 şeklindedir.

Çizelge 2.12 yöntemlerin VeriSeti-4 üzerinde sadece $T_{frekans}$ ağırlık değerinin kullanılması sonucu elde edilen başarım değerlerini göstermektedir. Çizelge2.13 ise tez çalışmasında önerilen $\{T_{frekans}\}\times\{T_{Dagitimsal}\}\times\{C_{önem}\}$ ağırlık değerinin kullanılması sonucu elde edilen yöntem başarımlarını göstermektedir. 567 adet haber dokümanından oluşan VeriSeti-4'de, özetler yoruma dayalı olduğundan yöntem başarımları çakışan cümle sayısına göre değil, çakışan Ngram sayısına göre (ROUGE-N [56]) belirlenmektedir.

Çizelgelerden görüldüğü gibi tezde önerilen yeni ağırlık değeri, yoruma dayalı olan uzun metin özetlerinde de olumlu etkilere sahip olmuştur.

Çizelge 2. 12 Terim frekans bilgisinin VeriSeti-4 üzerindeki etkisi

	ROUGE-1			ROUGE-2			ROUGE-L			ROUGE-S4		
DUC-2002-Tfrekans	R	P	F	R	P	F	R	P	F	R	P	F
Yöntem1	0,3529	0,3485	0,3495	0,1639	0,1615	0,1622	0,3280	0,3238	0,3248	0,1270	0,1252	0,1257
Yöntem2	0,3643	0,3581	0,3600	0,1724	0,1690	0,1702	0,3390	0,3330	0,3349	0,1353	0,1325	0,1334
Yöntem3	0,3392	0,3438	0,3403	0,1489	0,1510	0,1494	0,3154	0,3195	0,3163	0,1159	0,1177	0,1164
Yöntem4	0,3620	0,3586	0,3593	0,1718	0,1702	0,1705	0,3377	0,3345	0,3351	0,1348	0,1333	0,1337

Çizelge 2. 13 Önerilen ağırlık değerinin VeriSeti-4 üzerindeki etkisi

	ROUGE-1			ROUGE-2			ROUGE-L			ROUGE-S4		
DUC-2002 Tfrekans-Tdağıtımsal-Cönem	R	P	R	R	P	F	R	P	F	R	P	F
Yöntem1	0,4080	0,4040	0,4047	0,2046	0,2028	0,2030	0,3823	0,3785	0,3791	0,1610	0,1594	0,1596
Yöntem2	0,4167	0,4115	0,4128	0,2121	0,2096	0,2101	0,3923	0,3874	0,3886	0,1688	0,1664	0,1671
Yöntem3	0,3973	0,3966	0,3955	0,1949	0,1943	0,1939	0,3729	0,3721	0,3712	0,1558	0,1551	0,1549
Yöntem4	0,4110	0,4052	0,4067	0,2087	0,2056	0,2065	0,3879	0,3823	0,3838	0,1663	0,1635	0,1643

ROUGE değerlerine göre yöntemlerin başarım değerlerine bakıldığında yöntemlerin en yüksek başarım sonuçlarına ROUGE-1 değerlerinde ulaştığı görülmektedir.

Sistemlerin VeriSeti-1 üzerindeki çalışma süreleri incelenecek olursa $T_{frekans}$'a dayalı gizli anlamsal analiz uygulamaları 130 doküman için yaklaşık 0,13 dakika; $\{T_{frekans}\}\times\{T_{Dagitimsal}\}\times\{C_{önem}\}$ dayalı gizli anlamsal analiz uygulamaları yaklaşık 2,34 dakika sürmektedir. VeriSeti-2 üzerindeki çalışma süreleri incelenecek olursa $T_{frekans}$'a dayalı gizli anlamsal analiz uygulamaları 20 doküman için yaklaşık 0,05 dakika; $\{T_{frekans}\}\times\{T_{Dagitimsal}\}\times\{C_{önem}\}$ dayalı gizli anlamsal analiz uygulamaları yaklaşık 0,18 dakika sürmektedir. VeriSeti-3 üzerindeki çalışma süreleri incelenecek olursa $T_{frekans}$'a dayalı gizli anlamsal analiz uygulamaları 92 doküman için yaklaşık 0,12 dakika; $\{T_{frekans}\}\times\{T_{Dagitimsal}\}\times\{C_{önem}\}$ dayalı gizli anlamsal analiz uygulamaları yaklaşık 1,52 dakika sürmektedir. Son olarak VeriSeti-3 üzerindeki çalışma süreleri incelenecek olursa $T_{frekans}$'a dayalı gizli anlamsal analiz uygulamaları 567 doküman için yaklaşık 0,98 dakika; $\{T_{frekans}\}\times\{T_{Dagitimsal}\}\times\{C_{önem}\}$ dayalı gizli anlamsal analiz uygulamaları yaklaşık 12.9 dakika sürmektedir.

Sonuç olarak tez çalışmasında önerilen yeni ağırlıklandırma ile yöntemlerin birbirleri ile kıyaslanma durumundan bağımsız olarak toplam dört veri seti üzerinde denenmiş ve yeni ağırlıklandırmanın yöntem başarımlarını her koşulda arttığı gözlemlenmiştir. Yöntemler hem matematiksel olarak ayrıntılı bir şekilde ele alınmış hem de çalışma süreleri ile birlikte elde edilen başarım sonuçları araştırmacıların kullanımına sunulmuştur.

BÖLÜM 3

METİN ÖZETLEMEDE YENİ BİR MELEZ YAKLAŞIM

Cümle çıkarımına dayalı olan metin özetleme sistemlerinde metni oluşturan cümleler daha önce belirlenmiş olan yapısal veya anlamsal özelliklere göre incelenmektedirler. Literatürde İngilizce dokümanlar üzerinde çalışan ve cümle seçimi için kullanılan yapısal ve anlamsal özelliklerin birleşimini sağlayan melez sistem önerileri mevcuttur. Bu önerilerde özelliklerin birleşimi ile elde edilen yapıların sistem başarımları üzerindeki olumlu etkileri vurgulanmış ve bireysel özelliklerin katkıları üzerinde durulmuştur [10, 11, 13, 16, 64]. Tasarlanan melez sistemler özelliklerin birleşim aşamalarında genellikle makine öğrenmesine dayalı yöntemleri [64] veya sezgisel [16] yada bulanık tabanlı yaklaşımları [10, 11, 13] kullanmışlardır.

Bu bölümde özelliklerin birleşimini sağlayan bir melez sistemin Türkçe metinler üzerindeki etkilerini görmek adına literatürde önemli cümlelerin tespitini sağlamak için kullanılan hemen hemen tüm özellikleri içeren geniş çaplı bir melez sistem önerisi sunulmuştur. Önerilen melez sistem, özelliklerin birleşim aşamasında iki farklı yöntemi içermektedir. Bu yöntemlerden ilki özelliklerin uzman görüşlerine göre katkılarını göz önüne alan Bulanık Analitik Hiyerarşi Sürecini (BAHS), ikincisi ise özellikleri ait oldukları gruplara göre genetik algoritmalar (GA) aracılığıyla birleştiren otomatik bir süreci kapsamaktadır.

Bu bölüm ile öncelikle melez sistemin yapısını oluşturan yapısal ve anlamsal özellikler anlatılacak ve bu özelliklerin birleşimi ayrıntılı bir şekilde ele alınacaktır.

3.1 Melez Yaklaşımı Oluşturan Yapısal ve Anlamsal Özellikler

Melez sistem kapsamına giren yapısal ve anlamsal özellikler benzerliklerine göre Çizelge 3.1'de belirtilen şekilde gruplandırılmışlardır. Melez sistem, toplamda beş ana grup altında birleşen on beş farklı özelliği içermektedir. Bu özellikler Çizelge 3.1'den görülebilir.

Melez sistemin amacı belirtilen özellikleri uygun bir cümle skoru fonksiyonuna göre birleştirmek ve her cümleye bu skor fonksiyonuna göre bir puan vermektir. Böylelikle puanları belli olan cümleler, özetleme aşamasında puanlarına göre büyükten küçüğe dizilmekte ve en yüksek puanlı istenen sayıda cümle özete eklenerek özet dokümanları oluşturulmaktadır.

Çizelge 3. 1 Melez Sistemin Yapısını Oluşturan Yapısal ve Anlamsal Özellikler

HEDEF	ANA ÖZELLİKLER	ALT ÖZELLİKLER
ÖZELLİKLERİN OPTİMAL AĞIRLIK DEĞERLERİNİ ELDE ETMEK	G_1: KONUM BİLGİSİ	$ö_{11}$: Cümle Konumu
		$ö_{12}$: Kelimelerin Dağıtımsal Özelliği
	G_2: TEMEL CÜMLELERE BENZERLİK	$ö_{21}$: İlk Cümleye olan Benzerlik
		$ö_{22}$: Son Cümleye olan Benzerlik
		$ö_{23}$: Başlığa olan Benzerlik
	G_3: KELİME SIKLIĞI BİLGİSİ	$ö_{31}$: Cümle Uzunluğu
		$ö_{32}$: Kelime Sıklığı Bilgisi
		$ö_{33}$: Kelime Cümle Skoru Bilgisi
		$ö_{34}$: Ortalama Kelime Frekansı ve Ters Doküman Frekansı
	G_4: TEMASAL ÖZELLİKLER	$ö_{41}$: Sayısal Karakter İçerme Durumu
		$ö_{42}$: "?" ve "!" İçerme Durumu
		$ö_{43}$: Pozitif Kelimeleri İçerme Durumu
		$ö_{44}$: İsim Soylu Kelimeleri İçerme Durumu
	G_5: ANLAMSAL ÖZELLİKLER	$ö_{51}$: Gizli Anlamsal Analize Dayalı Anlamsal Özellik
		$ö_{52}$: Merkez Olma Durumu

Melez sistemde, Türkçe dokümanlara ait olan kelimeler Zemberek [62] ile köklerine ayrılmış ve tüm veri setlerindeki dokümanlar durak kelimelerinden arındırılımıştır.

Çizelge 3.1'de görüldüğü gibi ana gruplardan olan "KONUM BİLGİSİ" dokümanı oluşturan cümle ve kelimelerin doküman içinde bulunma yerlerini kapsayan bir özelliktir. Konum bilgisi iki özelliği barındırmaktadır:

* $ö_{11}$-Cümle Konumu: Bölüm 2.3.3'de açıklanan ve (2.24) eşitliği ile belirtilen ifadeye göre hesaplanan cümle skor değeridir.

* $ö_{12}$- Kelimelerin Dağıtımsal Özelliği: Bölüm 2.3.3'de açıklanan ve (2.18) eşitliğinde bir çarpan olarak kullanılan $T_{Dağıtımsal}$ ifadesi ile belirtilen terim özellikleri bulunduktan sonra, dokümana ait olan her bir cümlenin terimlerine ait $T_{Dağıtımsal}$ değerlerinin toplamı cümelere ağırlık değerleri olarak verilmiştir.

Ana gruplardan olan "TEMEL CÜMLELERE BENZERLİK" dokümanı oluşturan cümlelerin dokümandaki ilk cümleye, son cümleye ve başlığa olan benzerliklerine göre hesaplanan özellikleri içeren bir gruptur. Toplam üç özelliği barındırmaktadır:

* $ö_{21}$ - İlk Cümleye olan Benzerlik: Bölüm 2.3.3'de açıklanan ve (2.25) eşitliği ile belirtilen ifadeye göre hesaplanan cümle skor değeridir.

* $ö_{22}$ - Son Cümleye olan Benzerlik: Bölüm 2.3.3'de açıklanan ve (2.26) eşitliği ile belirtilen ifadeye göre hesaplanan cümle skor değeridir.

* $ö_{23}$ - Başlığa olan Benzerlik: Bölüm 2.3.3'de açıklanan ve (2.27) eşitliği ile belirtilen ifadeye göre hesaplanan cümle skor değeridir.

Ana gruplardan olan "KELİME SIKLIĞI BİLGİSİ" dokümanı oluşturan metin parçalarının içindeki kelimelerin doküman içinde bulunma sıklıklarını kapsayan bir özelliktir. Çizelge 3.1'den görüleceği gibi kelime sıklığı bilgisi dört özelliği barındırmaktadır:

* $ö_{31}$ – Cümle Uzunluğu: Bölüm 2.3.3'de açıklanan cümlelere sahip oldukları kelime sayılarına göre puan veren cümle skorudur.

* $ö_{32}$ - Kelime Sıklığı Bilgisi: Bölüm 2.3.3'de açıklanan cümlelere sahip oldukları kelime frekanslarının toplamına göre puan veren cümle skorudur.

* $ö_{33}$ -Kelime Cümle Skoru Bilgisi: Bölüm 2.3.3'de açıklanan ve (2.28) eşitliği ile belirtilen ifadeye göre hesaplanan cümle skor değeridir.

* $ö_{34}$ -Ortalama Kelime Frekansı ve Ters Doküman Frekansı: Bölüm 2.3.3'de açıklanan ve (2.31) eşitliği ile belirtilen ifadeye göre hesaplanan cümle skor değeridir.

Ana gruplardan olan "TEMASAL ÖZELLİKLER" dokümanı oluşturan metin parçalarının içindeki kelimelerin doküman içinde bulunma sıklıklarını kapsayan bir özelliktir. Çizelge 3.1'den görüleceği gibi temasal özellikler dört özelliği barındırmaktadır:

* $ö_{41}$ – Sayısal Karakter İçerme Durumu: Bölüm 2.3.3'de açıklanan cümlelere sahip oldukları sayısal karakterlerin toplamına göre puan veren cümle skorudur.

* $ö_{42}$ – "?" ve "!" İçerme Durumu: Bölüm 2.3.3'de açıklanan cümlelere sahip oldukları "?" ve "!" toplamına göre puan veren cümle skorudur.

* $ö_{43}$ – Pozitif Kelimeleri İçerme Durumu: Bölüm 2.3.3'de açıklanan cümlelere sahip oldukları pozitif kelimelerin frekans toplamına göre puan veren cümle skorudur.

* $ö_{44}$ – İsim Soylu Kelimeleri İçerme Durumu: Bölüm 2.3.3'de açıklanan cümlelere sahip oldukları isim soylu kelimelerin frekans toplamına göre puan veren cümle skorudur.

Ana gruplardan olan "ANLAMSAL ÖZELLİKLER" dokümanı oluşturan metin parçalarının içindeki kelimelerin anlamsal yapısını çıkaran özellikleri kapsamaktadır. Çizelge 3.1'den görüleceği gibi anlamsal özellikler iki özelliği barındırmaktadır:

* $ö_{51}$ – Gizli Anlamsal Analize Dayalı Anlamsal Özellik:

Bölüm 2.2.3'de açıklanan ve (2.17) eşitliği ile belirtilen ifadeye göre hesaplanan cümle skor değeridir.

* $ö_{52}$ – Merkez Olma Durumu:

Bölüm 2.3.3'de açıklanan ve (2.32) eşitliği ile belirtilen ifadeye göre hesaplanan cümle skor değeridir.

Özelliklerin tanıtımının ardından, özellik birleşimi için izlenen adımlar Bölüm 3.2 ile anlatılacaktır.

3.2 Melez Sistemin Yapısını Oluşturan Özellik Birleşim Yöntemleri

Bu bölümde melez sistemin yapısını oluşturan yapısal ve anlamsal özelliklerin uzman gücüne dayalı bir sistem olan BAHS ve otomatik bir sistem olan GA ile birleştirilme aşamalarından bahsedilecektir. Birleşim işlemi her iki yönteme göre de, hem ana gruplara verilen ağırlık değerleri hem de grup içindeki özelliklere verilen ağırlık değerleri kullanılarak eşitlik (3.1) ile belirten şekilde yapılacaktır.

$$W_1\sum_{i=1}^{2} w_{1i}\ddot{o}_{1i} + W_2\sum_{i=1}^{3} w_{2i}\ddot{o}_{2i} + W_3\sum_{i=1}^{4} w_{3i}\ddot{o}_{3i} + W_4\sum_{i=1}^{4} w_{4i}\ddot{o}_{4i} + W_5\sum_{i=1}^{2} w_{5i}\ddot{o}_{5i} \qquad (3.1)$$

Bu eşitlikte j=1,2,3,4,5 olmak üzere W_j'ler ana grup ağırlıklarını, w_{ji}'ler ise grup içindeki özellik ağırlıklarını belirtmektedir. Amaç bu ağırlık değerlerinin özelliklerin bireysel başarımlarına göre mi yoksa bu başarımlardan bağımsız olarak mı belirlendiğini tespit etmektir.

3.2.1 BAHS ile Özelliklerin Birleşim Aşaması

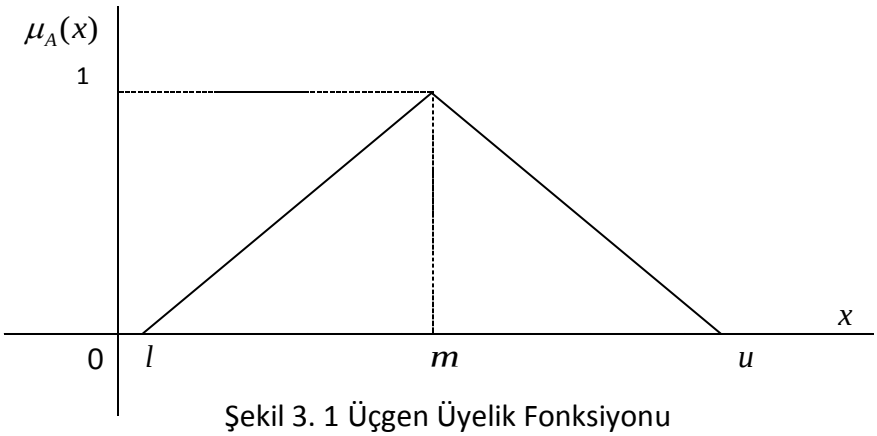
Analitik Hiyerarşi Süreci (AHS), karar hiyerarşisinin tanımlanabilmesi durumunda kullanılan, kararı etkileyen faktörler açısından karar noktalarının yüzde dağılımlarını veren bir karar verme ve tahminleme yöntemidir. İlk olarak 1968 yılında Myers ve Alpert ikilisi tarafından ortaya atılmış ve 1977'de Saaty tarafından geliştirilerek karar verme problemlerinin çözümünde kullanılmıştır.

Karar verme sürecindeki belirsizlik durumunlarında Saaty (1977)'nin geliştirdiği klasik AHS tekniği [65] ile bulanık küme teorisi bütünleştirilir ve süreç bulanık analitik hiyerarşi süreci (BAHS) olarak isimlendirilir. BAHS'de karar verici genellikle kesin değerler içeren değerlendirmeler yapmak yerine, aralıklı değerlendirmeler yapmayı tercih etmektedir. Dolayısıyla bulanık yaklaşım ile karar verme süreci daha hassas bir şekilde tanımlanır.

Analitik hiyarerşi süreci [66] çalışmasıyla sınıflayıcı birleşiminde, [67] çalışmasıyla diz üstü bilgisayar seçiminde ve [68] çalışmasıyla veri tabanı yönetimi projesinde kullanılmıştır.

Bir karar verme probleminin BAHS ile çözümlenebilmesi için öncelikle karar verme problemi hiyerarşik bir yapıda tanımlanmalı, daha sonra karar vericiler tarafından kararı etkileyen nesneler arasında ikili karşılaştırma matrisleri oluşturulmalıdır. İkili karşılaştırma matrisleri, üçgensel bulanık sayılardan oluşan karşılaştırma sayılarını içermektedir. Bir üçgensel bulanık sayı üç elemandan oluşur: $M = (l, m, u)$. Bu ifadede l ve u üçgensel sayının alt ve üst sınırlarını, m ise üçgensel bulanık sayının tepe noktasını ifade etmektedir. Üçgen bulanık sayılar için, üyelik fonksiyonu (3.2) eşitliği ile ve grafik ifadesi ise Şekil 3.1 ile gösterilmiştir.

$$\mu_A(x) = \begin{cases} 0, & \mathrm{x} < \mathrm{l} \\ \dfrac{x-l}{m-l}, & \mathrm{l} \le \mathrm{x} \le \mathrm{m} \\ \dfrac{u-x}{u-m}, & \mathrm{m} \le \mathrm{x} \le \mathrm{u} \\ 0, & \mathrm{x} > \mathrm{u} \end{cases} \qquad (3.2)$$



Şekil 3. 1 Üçgen Üyelik Fonksiyonu

Bir bulanık sayı, üyelik derecesi 0 ile 1 arasında değişen konveks bir bulanık kümedir. Üçgen bir bulanık sayının üyelik fonksiyonu aşağıda verilen özellikleri taşımalıdır:

a) $x \in (-\infty, l] \cup [u, +\infty)$ ise $\mu_A(x) = 0$ olur.
b) $\mu_A(x)$, $[l, m]$ aralığında artan $[m, u]$ aralığında azalandır.
c) $x = m$ olduğunda $\mu_A(x) = 1$'dir.

Üçgensel bulanık sayılar ile farklı işlemler gerçekleştirilebilmektedir. $M_1 = (l_1, m_1, u_1)$ ve $M_2 = (l_2, m_2, u_2)$ üçgensel bulanık sayılar olmak üzere, bu sayılar arasında yapılacak aritmetik işlemler (3.3) ile belirtilen eşitlik grubuyla özetlenmektedir.

$$M_1(+)M_2 = (l_1, m_1, u_1) + (l_2, m_2, u_2) = (l_1 + l_2, m_1 + m_2, u_1 + u_2) \quad (3.3)$$
$$M_1(-)M_2 = (l_1, m_1, u_1) - (l_2, m_2, u_2) = (l_1 - l_2, m_1 - m_2, u_1 - u_2)$$
$$M_1(\times)M_2 = (l_1, m_1, u_1) \times (l_2, m_2, u_2) = (l_1 \times l_2, m_1 \times m_2, u_1 \times u_2)$$
$$M_1(/)M_2 = (l_1, m_1, u_1)/(l_2, m_2, u_2) = (l_1 / u_2, m_1 / m_2, u_1 / l_2)$$
$$M_1^{-1} = (l_1, m_1, u_1)^{-1} = (1/u_1, 1/m_1, 1/l_1)$$

BAHS'in bir aşaması olan üçgensel bulanık sayılardan oluşan ikili karşılaştırma matrisi aşağıda belirtilen şekilde ifade edilmektedir:

$$A = \begin{bmatrix} (1,1,1) & (l_{12},u_{12},m_{12}) & \dots & (l_{1n},u_{1n},m_{1n}) \\ (l_{21},u_{21},m_{21}) & (1,1,1) & \dots & (l_{2n},u_{2n},m_{2n}) \\ . & . & \dots & . \\ . & . & \dots & . \\ (l_{n1},u_{n1},m_{n1}) & (l_{n2},u_{n2},m_{n2}) & \dots & (1,1,1) \end{bmatrix}$$

Karşılaştırmalar, karşılaştırma matrisinin köşegeninin üstünde kalan değerler (a_{ij}) için yapılır. Köşegenin altıda kalan değerler için ise (3.4) ile belirtilen formülü kullanmak yeterli olacaktır.

$a_{ij} = (l_{ij}, m_{ij}, u_{ij})$ iken a_{ij}'nin bulanık ölçek eşleniği $a_{ij}^{-1} = (1/u_{ji}, 1/m_{ji}, 1/l_{ji})$ (3.4)

BAHS'de karşılaştırmalar gerçekleştirilirken kullanılan ölçek çesitleri uygulamaya ve uzman görüşüne göre değişmektedir. Yaygın olarak kullanılan ölçek çeşidi bulanık üçgensel sayılardan oluşan Çizelge 3.2'de verilen ölçektir.

Çizelge 3. 2 Bulanık Analitik Hiyerarşi Süreci Önem Ölçeği

Sözel Önem	**Bulanık Ölçek**	**Bulanık Ölçek Eşleniği**
Eşit Derecede Önemli	(1, 1, 1)	(1, 1, 1)
	(1/2, 3/4, 1)	(1, 4/3, 2)
Biraz daha fazla önemli	(2/3, 1, 3/2)	(2/3, 1, 3/2)
	(1, 3/2, 2)	(1/2, 2/3, 1)
Kuvvetli derecede önemli	(3/2,2,5/2)	(2/5,1/2,2/3)
	(2, 5/2, 3)	(1/3, 2/5, 1/2)
Çok kuvvetli derecede önemli	(5/2,3,7/2)	(2/7,1/3,2/5)
	(3, 7/2, 4)	(1/4, 2/7, 1/3)
Tamamıyla önemli	(7/2,4,9/2)	(2/9,1/4,2/7)

Çizelgede "Sözel Önem" kolonu altındaki boş satırlar ara önem değerlerini belirtmektedirler. Literatürde yer alan çeşitli yazarlar tarafından ortaya konmuş olan bir çok BAHS yöntemi bulunmaktadır. Bunlardan biri Chang tarafından 1996 yılında önerilmiş olan "Genelleştirilmiş Bulanık Analitik Hiyerarşi Süreci (GBAHS)" [69-54] yöntemidir.

3.2.1.1 GBAHS Yöntemi

BAHS'nin uygulandığı birçok problemde Chang tarafından önerilen GBAHS yöntemi [69] kullanılmıştır. Bu yöntemin en avantajlı yanı hesap gereksiniminin az olması ve klasik AHS'nin adımlarını izleyerek ilave işlem gerektirmemesidir.

GBAHS yöntemine göre, $X = (x_1, x_2, ..., x_n)$ nesneler kümesini ve $U = (u_1, u_2, ..., u_n)$ bir hedef kümesini göstermek üzere, bu kümelere ait olan herbir nesne ele alınarak, herbir amaç için genişletilmiş analiz tekniği uygulanır. Bu durumda m adet boyut değeri ortaya çıkmaktadır. Bu boyut değeleri aşağıda belirtilen sembollerle ifade edilmektedir:

$$M_{gi}^{1}, M_{gi}^{2}, ..., M_{gi}^{m} \quad i = 1,2,...,n \tag{3.5}$$

Burada verilen tüm M_{gi}^{j} $j = 1,2,...,m$ değerleri üçgensel bulanık sayılardır. Genelleştirilmiş analiz yönteminin tüm aşamaları aşağıda belirtilmiştir.

1.Aşama: Problem hiyerarşik bir yapıya oturtulur ve üçgensel bulanık sayıların kullanılmasıyla hiyerarşik yapı içerisindeki nesneler arasında ikili karşılaştırma matrisleri oluşturulur. Karşılaştırma matrislerinin tutarlı olması için her bir matrisin tutarlılık değeri belirli tutarlılık oranlarından küçük olmalıdır[65].

2.Aşama: Bulanık sayıların karşılaştırılmasına zemin hazırlayan, S_i bulanık sentetik boyut değerleri aşağıdaki formüller yardımıyla hesaplanmaktadır.

$$S_i = \sum_{j=1}^{m} M_{gi}^{j} \otimes [\sum_{i=1}^{n} \sum_{j=1}^{m} M_{gi}^{j}]^{-1} \tag{3.6}$$

$$\sum_{j=1}^{m} M_{gi}^{j} = (\sum_{j=1}^{m} l_j, \sum_{j=1}^{m} m_j, \sum_{j=1}^{m} u_j) \tag{3.7}$$

$$[\sum_{i=1}^{n}\sum_{j=1}^{m} M_{gi}^{j}] = (\sum_{j=1}^{m} l_j, \sum_{j=1}^{m} m_j, \sum_{j=1}^{m} u_j) \quad (3.8)$$

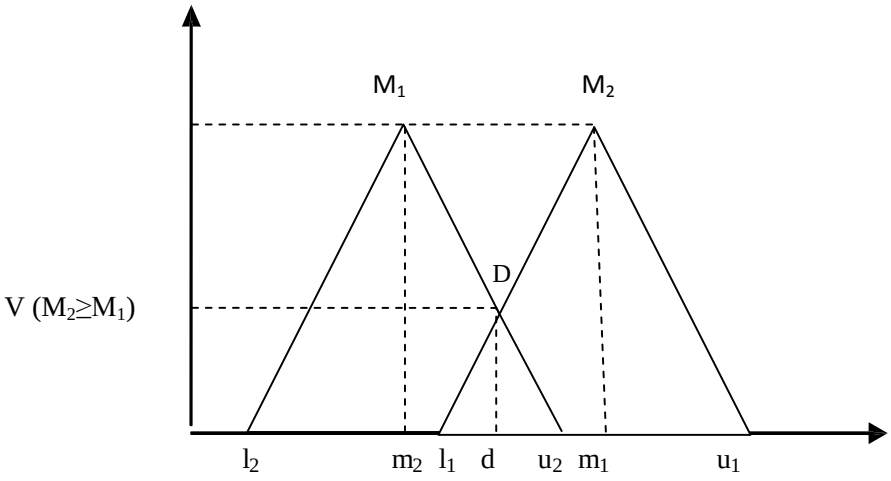
$$[\sum_{i=1}^{n}\sum_{j=1}^{m} M_{gi}^{j}]^{-1} = (\frac{1}{\sum_{i=1}^{n} u_i}, \frac{1}{\sum_{i=1}^{n} m_i}, \frac{1}{\sum_{i=1}^{n} l_i}) \quad (3.9)$$

3.Aşama: $M_1 = (l_1, m_1, u_1)$ ve $M_2 = (l_2, m_2, u_2)$ üçgensel bulanık sayılar olmak üzere $M_1 = (l_1, m_1, u_1)$ ≥ $M_2 = (l_2, m_2, u_2)$ ifadesinin olabilirlik derecesi hesaplanır. Bu durum aşağıda gösterilmektedir:

$$V(M_2 \geq M_1) = H(M_2 \cap M_1) = \sup_{y \geq x}[\min(\mu M_1(x), \mu M_2(y))] = \mu_{M_2}(d) \quad (3.10)$$

$M_1 = (l_1, m_1, u_1)$ ve $M_2 = (l_2, m_2, u_2)$ üçgensel ve dışbükey bulanık sayılar olmak üzere bu sayıların kesişiminin üyelik fonksiyonu (3.11) ve grafik ifadesi Şekil 3.2'de gösterilmektedir.

$$\mu_{M_2}(d) = \begin{cases} 1 & ,m_2 \geq m_1 \\ 0 & ,l_1 \geq u_2 \\ \dfrac{(l_1 - u_2)}{(m_2 - u_2) - m_1 - l_1)} & ,aksidurumlar \end{cases} \quad (3.11)$$



Şekil 3. 2 M_1 ve M_2 bulanık sayılarının kesişimi

Şekilden de görüldüğü gibi $V(M_2 \geq M_1)$ değeri, μ_{M_2} ve μ_{M_1} üyelik fonksiyonlarının kesişim noktasının ordinatıdır. M_1 ve M_2 bulanık sayıları arasında bir karşılaştırma yapılabilmesi için $V(M_2 \geq M_1)$ ve $V(M_1 \geq M_2)$ değerlerinin her ikisine de gereksinim duyulmaktadır.

4.Aşama: Bir konveks bulanık sayının k tane konveks bulanık sayıdan $M_i, i = 1,2,...,k$ daha büyük olmasının olabilirlik derecesi aşağıdaki gibi tanımlanır:

$$V(M \geq M_1, M_2, \ldots, M_k) = V[(M \geq M_1) \text{ve} (M \geq M_2) \text{ ve } \ldots \text{ve} (M \geq M_k)] = \min V(M \geq M_i) \quad (3.12)$$

i=1, 2,...,k ; k=1,2,...,n ;$k \neq i$ için $d'(A_i) = \min V(S_i \geq S_k)$ olduğu varsayıldığında ağırlık vektörü (3.13) eşitliği ile ifade edilmektedir.

$$W' = (d'(A_1), d'(A_2), ..., d'(A_n))^T \quad (3.13)$$

Burada $A_i = (1,2,...,n)$ n sayısı kadardır.

5.Aşama: (3.13) ile verilen ağırlık vektörü normalize edilerek, normalize edilmiş ağırlık vektörüne ulaşılmaktadır. Normalize edilmiş ağırlık vektörü W, bulanık olmayan bir vektördür ve (3.14) eşitliği ile gösterilmektedir.

$$W = (d(A_1), d(A_2), ..., d(A_n))^T \quad (3.14)$$

3.2.1.2 GBAHS'nın Cümle Puanlarının Hesaplanması Amacıyla Kullanımı

Bölüm 3.2.1.1'de belirtimiş olan GBAHS aşamaları metin özetleme problemi üzerinde aşağıda belirtilen sıra ile uygulanmıştır:

1.Aşama: GBAHS'ye göre bir problemin ele alınabilmesi için ilk etapta problemin amacı belirlenmeli ve problem hiyerarşik bir yapıya oturtulmalıdır. Çizelge 3.1'de belirtildiği gibi metin özetleme probleminin amacı, özetleri oluşturacak olan cümlelerin belirlenmesi adına, kullanılan özelliklerin optimal ağırlık değerlerini elde etmektir. Bu amaçla ilk etapta metin özetleme probleminde kullanılan yapısal ve anlamsal özellikler Çizelge 3.1 ile gösterilen şekilde bir amaç altında hiyerarşik bir yapıya oturtulmuştur. Daha sonra dil bilimi ile

uğraşan uzmanlar tarafından üzerinde mutabık kılınmış ve Çizelge 3.2'deki önem ölçeğini baz alan ikili karşılaştırma matrisleri oluşturulmuştur. İkili karşılaştırma matrislerinde ilk önce gruplar kendi aralarında kıyaslanmış, daha sonra her bir grup içindeki özelliklerin kendi aralarındaki ikili kıyaslamaları yapılmıştır. Uzman görüşleri altında oluşturulan ikili kıyaslama matrisleri Çizelge 3.3 ile başlayıp Çizelge 3.8'e kadar devam eden çizelgeler ile belirtilmiştir.

Çizelge 3. 3 Ana Grupların ikili karşılaştırma matrisi

	G_1	G_2	G_3	G_4	G_5
G_1	(1, 1, 1)	(2/3, 1, 3/2)	(1/2, 3/4, 1)	(3/2, 2, 5/2)	(1/2, 3/4, 1)
G_2	(2/3, 1, 3/2)	(1, 1, 1)	(1, 4/3, 2)	(2/3, 1, 3/2)	(2/3, 1, 3/2)
G_3	(1, 4/3, 2)	(1/2, 3/4, 1)	(1, 1, 1)	(2/3, 1, 3/2)	(1, 4/3, 2)
G_4	(2/5, 1/2, 2/3)	(2/3, 1, 3/2)	(2/3, 1, 3/2)	(1, 1, 1)	(2/3, 1, 3/2)
G_5	(1, 4/3, 2)	(2/3, 1, 3/2)	(1/2, 3/4, 1)	(2/3, 1, 3/2)	(1, 1, 1)

Çizelge 3. 4 G_1 altındaki özelliklerin ikili karşılaştırma matrisi

	$ö_{11}$	$ö_{12}$
$ö_{11}$	(1, 1, 1)	(1/2, 3/4, 1)
$ö_{12}$	(1, 4/3, 2)	(1, 1, 1)

Çizelge 3. 5 G_2 altındaki özelliklerin ikili karşılaştırma matrisi

	$ö_{21}$	$ö_{22}$	$ö_{23}$
$ö_{21}$	(1, 1, 1)	(2/3, 1, 3/2)	(1, 4/3, 2)
$ö_{22}$	(2/3, 1, 3/2)	(1, 1, 1)	(1, 4/3, 2)
$ö_{23}$	(1/2, 3/4, 1)	(1/2, 3/4, 1)	(1, 1, 1)

Çizelge 3. 6 G_3 altındaki özelliklerin ikili karşılaştırma matrisi

	$ö_{31}$	$ö_{32}$	$ö_{33}$	$ö_{34}$
$ö_{31}$	(1, 1, 1)	(2/3, 1, 3/2)	(1, 4/3, 2)	(1/2, 3/4, 1)
$ö_{32}$	(2/3, 1, 3/2)	(1, 1, 1)	(2/3, 1, 3/2)	(1, 4/3, 2)
$ö_{33}$	(1/2, 3/4, 1)	(2/3, 1, 3/2)	(1, 1, 1)	(2/3, 1, 3/2)
$ö_{34}$	(1, 4/3, 2)	(1/2, 3/4, 1)	(2/3, 1, 3/2)	(1, 1, 1)

Çizelge 3. 7 G_4 altındaki özelliklerin ikili karşılaştırma matrisi

	ö41	ö42	ö43	ö44
ö41	(1, 1, 1)	(1/2, 3/4, 1)	(2/3, 1, 3/2)	(2/5, 1/2, 2/3)
ö42	(1, 4/3, 2)	(1, 1, 1)	(2/3, 1, 3/2)	(2/3, 1, 3/2)
ö43	(2/3, 1, 3/2)	(2/3, 1, 3/2)	(1, 1, 1)	(1, 4/3, 2)
ö44	(3/2, 2, 5/2)	(2/3, 1, 3/2)	(1/2, 3/4, 1)	(1, 1, 1)

Çizelge 3. 8 G_5 altındaki özelliklerin ikili karşılaştırma matrisi

	ö51	ö52
ö51	(1, 1, 1)	(1, 4/3, 2)
ö52	(1/2, 3/4, 1)	(1, 1, 1)

2.Aşama: Bu aşamada Çizelge 3.3 ile başlayıp Çizelge 3.8'e kadar uzanan tüm ikili kıyaslama matrislerine ait olan özelliklerin bulanık sentetik boyut değerleri bulunmuştur.

Bölüm 3.1.1.1'de belirtilen ikinci aşamada bahsi geçen formüllerin daha iyi anlaşılabilmesi için, formüller sadece Çizelge 3.3 ile belirtilen ana grupların ikili karşılaştırma matrisi üzerinde uygulanmıştır. Bu durumda her gruba ait olan ve (3.6) eşitliği ile gösterilen bulanık sentetik boyut değerleri şu şekilde bulunur:

$$S_{G_1} = (4.1666 \quad 5.5000 \quad 7.0000) \times \left(\frac{1}{35.1667} \quad \frac{1}{25.8333} \quad \frac{1}{19.5667}\right) = (0.1185 \quad 0.2129 \quad 0.3578)$$

$$S_{G_2} = (4.0000 \quad 5.3333 \quad 7.5000) \times \left(\frac{1}{35.1667} \quad \frac{1}{25.8333} \quad \frac{1}{19.5667}\right) = (0.1137 \quad 0.2065 \quad 0.3833)$$

$$S_{G_3} = (4.1666 \quad 5.4167 \quad 7.5000) \times \left(\frac{1}{35.1667} \quad \frac{1}{25.8333} \quad \frac{1}{19.5667}\right) = (0.1185 \quad 0.2097 \quad 0.3833)$$

$$S_{G_4} = (3.4000 \quad 4.5000 \quad 6.1667) \times \left(\frac{1}{35.1667} \quad \frac{1}{25.8333} \quad \frac{1}{19.5667}\right) = (0.0967 \quad 0.1742 \quad 0.3152)$$

$$S_{G_5} = (3.8333 \quad 5.0833 \quad 7.0000) \times \left(\frac{1}{35.1667} \quad \frac{1}{25.8333} \quad \frac{1}{19.5667}\right) = (0.1090 \quad 0.1968 \quad 0.3578)$$

3.Aşama: Bulanık sentetik değerlerinin elde edilmesinden sonra (3.10) eşitliği kullanılarak, elde edilen sentetik değerler arasında karşılaştırma işlemleri yapılır. Verilen örneğin devamı olması açısından ana grupların ikili karşılaştırma

matrisinin (Çizelge 3.3) baz alınmasıyla elde edilen sentetik değerlerin karşılaştırılma sonuçları Çizelge 3.9 ile belirtilmiştir.

Çizelge 3. 9 Bulanık sentetik değerleri arasındaki karşılaştırma sonuçları

$V(S_{G_1} \succ S_{G_2})$=1	$V(S_{G_2} \succ S_{G_1})$=0.976	$V(S_{G_3} \succ S_{G_1})$=0.988	$V(S_{G_4} \succ S_{G_1})$=0.835	$V(S_{G_5} \succ S_{G_1})$=0.936
$V(S_{G_1} \succ S_{G_3})$=1	$V(S_{G_2} \succ S_{G_3})$=0.988	$V(S_{G_3} \succ S_{G_2})$=1	$V(S_{G_4} \succ S_{G_2})$=0.862	$V(S_{G_5} \succ S_{G_2})$=0.961
$V(S_{G_1} \succ S_{G_4})$=1	$V(S_{G_2} \succ S_{G_4})$=1	$V(S_{G_3} \succ S_{G_4})$=1	$V(S_{G_4} \succ S_{G_3})$=0.847	$V(S_{G_5} \succ S_{G_3})$=0.948
$V(S_{G_1} \succ S_{G_5})$=1	$V(S_{G_2} \succ S_{G_5})$=1	$V(S_{G_3} \succ S_{G_5})$=1	$V(S_{G_4} \succ S_{G_5})$=0.901	$V(S_{G_5} \succ S_{G_4})$=1.0

Bu karşılaştırma işlemleri Çizelge 3.4 - 3.8 arasındaki çizelgeler ile belirtilen tüm karşılaştırma matrisleri için de bulunmuştur.

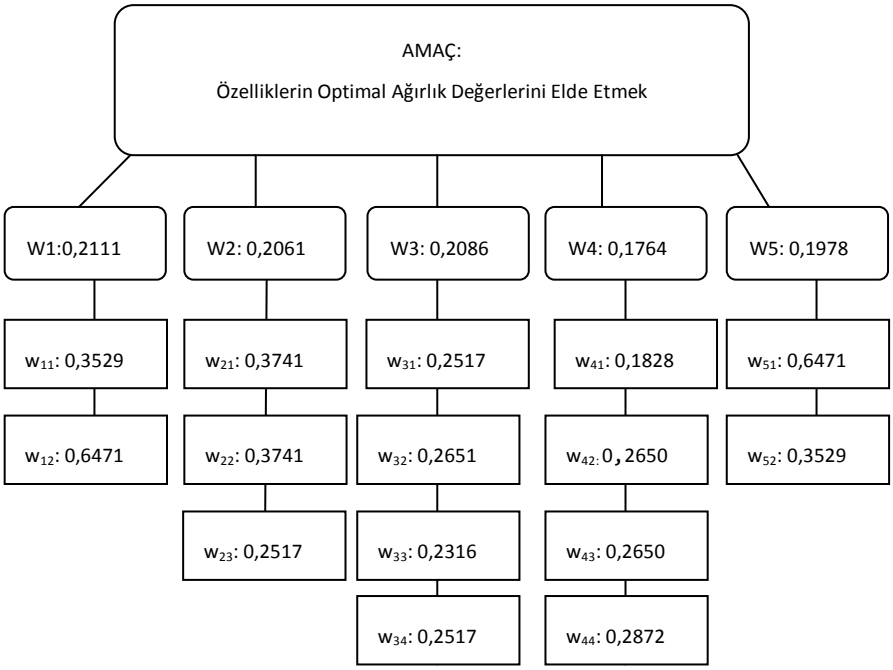
4.Aşama: Bulanık sentetik değerlerin kıyaslanma işlemi sona erdikten sonra ağırlık değerlerinin buluması için k=1,2,...,n ;$k \neq i$ için $d'(A_i) = \min V(S_i \geq S_k)$ koşulları altında (3.13) eşitliği kullanılarak normalize edilmemiş olan W' ağırlık değerleri elde edilir. Yine verilen örneğin devamı olması açısından ana grupların ikili karşılaştırma matrisinin (Çizelge 3.3) baz alınmasıyla elde edilen normalize edilmemiş olan W' ağırlık değerleri aşağıdaki şekilde bulunmuştur:

W' = (1.0000 0.9762 0.9880 0.8355 0.9368)

5.Aşama: Nihayet elde edilen ağırlık değerleri (3.14) eşitliği ile normalize edilmiş ve melez sistemde kullanılacak olan ağırlık değerlerinin son hali elde edilmiştir. Ana grupların ikili karşılaştırma matrisinin (Çizelge 3.3) baz alınmasıyla elde edilen normalize edilmiş olan W ağırlık değerleri aşağıdaki şekilde bulunmuştur:

W = (0.2111 0.2061 0.2086 0.1764 0.1978)

Yukarıdaki aşamalarda bahsi geçen işlemlerin tamamı, ana grupların altındaki özelliklerin arasındaki ikili kıyaslama matrislerini gösteren Çizelge 3.4 - Çizelge 3.8 arasındaki matrislere uygulandıktan sonra elde edilen tüm ağırlık değerleri Şekil 3.3 ile belirtilen hiyerarşik yapıdan görülebilmektedir:



Şekil 3. 3 GBAHS ile elde edilen gruplar arası ve grup içi özellik ağırlıkları

BAHS metin özetlemede kullanılan özelliklerin sahip olması gereken ağırlık değerlerini uzman kişilerin görüşüne göre ikili karşılaştırma matrislerinin kurulmasıyla tespit eden bir yöntemdir. İkili karşılaştırma matrisindeki değerler bireysel özelliklerin birbirlerinden ne kadar daha önemli olduğunu göstermektedir. BAHS yönteminin avantajı ağırlık değerlerinin tespiti için kullanılan veri setininin eğitilmesine ihtiyaç duyulmamasıdır. Yani BAHS sayesinde eğitime gerek olmaksızın veri setindeki dokümanlar için genel ağırlık değerleri tespit edilir. Dolayısıyla genel hatlarıyla yapısal olarak birbirlerine yakın dokümanlar içeren veri setleri üzerinde iyi sonuçlar verebilir. BAHS'nin dezavantajı ise sisteme yeni bir özellik eklendiğinde veya çıkarıldığında ikili karşılaştıma matrislerinin tekrar oluşturulması gerekliliğidir.

Bir sonraki bölümde özelliklerin ağırlık değerleri otomatik olarak GA ile buldurulmaya çalışılmıştır.

3.2.2 Genetik Algoritmalar ile Özelliklerin Birleşimi

Genetik algoritmalar (GA), doğada gözlemlenen evrimsel sürece benzer bir şekilde çalışan arama ve en iyileme yöntemidir. GA, ilk kez Michigan Üniversitesi'nde John

Holland tarafından çalışılmıştır. Holland 1975 yılında yaptığı çalışmaları “Adaptation in Natural and Artificial Systems [70]” adlı kitabında sunmuştur.

GA, bir problemin çözümü için olası çözümleri içeren ve “popülasyon” ismiyle anılan bir kümeyi içermektedir. Popülasyonlar içindeki olası çözümler “kromozom” veya “birey” adı verilen sayı dizilerinden oluşur. Kromozomlar “gen” adı verilen parametrelerden oluşmaktadır. Popülasyon içindeki her kromozomun problem için çözüm olup olmayacağına karar veren bir “uygunluk fonksiyonu” vardır. Uygunluk fonksiyonundan dönen değere göre yüksek değere sahip olan kromozomlara, popülasyondaki diğer kromozomlar ile çoğalmaları için fırsat verilir. Bu çoğalma durumu, “çaprazlama” işlemi yoluyla gerçekleştirilir. Yeni bireyler üretilirken düşük uygunluk değerine sahip bireyler daha az seçilmektedir. Dolayısıyla bu bireyler bir süre sonra popülasyon dışında bırakılırlar. Yeni popülasyon, bir önceki popülsyonda yer alan uygunluğu yüksek bireylerin bir araya gelip çoğalmalarıyla oluşur. Böylelikle, pek çok nesil aracılığıyla iyi özellikler nüfus içerisinde yayılırlar ve genetik işlemler aracılığıyla da diğer iyi özelliklerle birleşirler.

Tipik bir genetik algoritma, genetik algoritmalara özel süreçleri içeren bazı adımlardan oluşur. Örnek bir genetik algoritmanın çalışma adımları aşağıdaki çizelgede verilmiştir [71].

Çizelge 3. 10 Genetik algoritmanın çalışma adımları

Adım	Yapılan İşlem
1	Gösterim yönteminin belirlenmesi
2	Başlangıç nüfusunun oluşturulması
3	Başlangıç nüfusundaki her bireyin başarımının amaç fonksiyonuna göre hesaplanması
4	Yeni neslin oluşturulmasında kullanılacak bireylerin seçilmesi
5	Seçilmiş bireylere genetik işlemlerin uygulanarak yeni neslin elde edilmesi
6	Yeni neslin bireylerinin performanslarının uygunluk fonksiyonuna göre hesaplanması
7	Bitiş koşulu sağlanmamışsa 4. adıma dönülmesi
8	Bitiş koşulu sağlanmışsa en iyi bireyin sonuç olarak dönülmesi

Çizelge 3.10’da görüldüğü gibi herhangi bir problemin GA ile çözülebilmesi için, öncelikle, aday çözümlerin uygun şekilde belirlenmesi gerekmektedir. Aday çözümler,

çoğu zaman çözümün her bir elemanının 1 veya 0 değeri alabildiği ikili değerlerin kullanıldığı sabit uzunluklu kromozomlar olarak kodlanır. Uygulamaya bağlı olarak, bazı problemler için çözüm parametrelerinin daha uygun ifade edilebilmesi nedeniyle tam sayı veya gerçel sayıların kullanılması da mümkündür.

Gösterim yönteminin belirlenmesinin ardından kodlanmış bireylerden oluşan bir başlangıç popülasyonu oluşturulur. Başlangıç popülasyonunu oluşturan kromozomlar rasgele veya çözülecek probleme özgü bilgiler kullanılarak kodlanırlar. Her ne kadar nüfus büyüklüğü problemin yapısına göre değişebilse de, belirlenen başlangıç nüfusu olası çözümlerin önemli bir bölümünü kapsayacak şekilde farklı çözümler barındıracak kadar geniş olmalıdır. Başlangıç popülasyonu oluşturulduktan sonra varolan çözüm alternatiflerinin performansı bir uygunluk fonksiyonu kullanılarak ölçülür. Hesaplanan uygunluk değeri, bireyin amaç fonksiyonuna göre değerini, dolayısıyla da çözüme yakınlığını göstermektedir.

Genetik algoritmaların, çözüm uzayında en iyi sonucu arama süreci üç önemli genetik işlemi kapsamaktadır: “Seçim”, “Çaprazlama” ve “Mutasyon”. Bu genetik işlemler kullanılarak başlangıçta rasgele oluşturulmuş çözüm adayları ardışık işlemlerle en iyi veya en iyiye yakın çözümleri içeren popülasyona evirilmeye çalışılmaktadır. Bu dönüşüm aşamasında “Seçim” mevcut nüfusdaki hangi bireylerin genetik işlemlere tabi tutularak yeni nesillerin oluşturulmasında kullanılacağını belirler. Rulet tekerlek, truva seçimi [72] sıklıkla kullanılan seçim yöntemleridir. Seçim algoritması sonucunda belirlenen, mevcut nüfus içinde en iyi çözüme daha yakın olması muhtemel adaylar, çaprazlama işlemi ile birleştirilerek yeni bireyler oluşturulur. Bu şekilde, çözüm uzayının araştırılması süreci rasgele olmayıp, mevcut genetik bilgi doğrultusunda en iyi sonuca götürecek şekilde yönlendirilmiş olur. Seçim işleminde olduğu gibi çaprazlama işleminde de alternatif yöntemler mevcuttur. En çok kullanılan yöntem tek noktadan çaprazlama olsa da, iki noktadan çaprazlama, çok noktadan çaprazlama ve homojen çaprazlama yöntemleri de kullanılabilmektedir [73]. Noktalı çaprazlama işleminde çaprazlama için seçilen iki çözüm adayının karşılıklı değerlerinin (genlerinin) çaprazlama noktası/noktalarına göre değiş tokuşu yapılır. Homojen çaprazlamada ise yeni bireyin genleri rasgele olarak çaprazlamaya katılan bireylerin genlerinden gelir. Çaprazlama işlemi sonucunda her iki atanın kısmi özelliklerine sahip yeni bir birey elde edilmiş olur.

Çaprazlama önceden belirlenmiş bir olasılık yüzdesine bağlı olarak rasgele olarak yapılır. Mutasyon işleminde seçilen bir bireyin genetik bilgisi, kullanılan kodlama sistemine uygun şekilde rasgele olarak değiştirilir. Örnek olarak, ikili kodlama sistemi ile kodlanmış bir bireyin 1 olan bir geninin 0'a veya 0 olan bir geninin 1'e dönüştürülmesi verilebilir. Mutasyon ile var olan populasyona yeni genetik bilgi eklenmiş olur. Mutasyon işlemi, genetik algoritmalara yerel alt optimumlarda takılıp kalmama ve yeni ve daha önceden fark edilmemiş çözümlere ulaşabilme özelliği kazandırmaktadır. Mutasyon da çaprazlama gibi önceden belirlenmiş bir olasılık yüzdesine göre rasgele olarak yapılır. Mutasyonlar optimal çözüme ulaşma açısından faydalı da olabilirler faydasız da olabilirler. Bu nedenle, populasyonun genetik bilgisinin bir anda çok fazla değişmemesi için mutasyon olasılığı düşük tutulur.

Genetik algoritmanın çalışma döngüsünün sona erip, çözüme ulaşılabilmesi için önceden belirlenmiş bir çıkış koşulunun sağlanmış olması gerekir. Çıkış koşulunun sağlanması üç durumdan birinin gerçekleşmesi ile olabilir [74]: "Tatmin edici bir sonuca ulaşılmış olması", "Genetik algoritmanın belli bir çözüme yakınsaması", "Önceden belirlenmiş en fazla nesil sayısına ulaşılması".

Genetik algoritma belli bir çözüme yakınsadığında nesli oluşturan tüm bireyler benzerdir ve genetik algoritmanın adımlarının tekrarlanması sonucu değiştirmez. Genetik algoritma, sonuca ulaşmadığı, yakınsamadığı veya en fazla nesil sayısına ulaşmadığı durumlarda seçim, çaprazlama ve mutasyon işlemlerini kullanarak yeni nesillerin oluşturulmasına devam eder.

3.2.2.1 GA'nın Cümle Puanlarının Hesaplanması Amacıyla Kullanımı

Bu bölümde eşitlik (3.1) ile gösterilen cümle skor fonksiyonundaki W_j ve w_{ji} ağırlık değerleri, *JGAP* java paketi [75] kullanılarak tespit edilmiştir. Bu tespit sırasında hem gerçek kodlu GA hem de ikili kodlu GA'lar kullanılmıştır.

GA'nın çalışmaya başlaması için, N_{ipop} kromozomdan oluşan bir başlangıç popülasyonuna ihtiyaç vardır. Popülasyonun kromozomları $N_{par}xN_{ipop}$ boyutlu bir matris ile temsil edilebilir. Gerçek kodlu GA'da kromozomlar (3.14) eşitliği ile belirtilen şekilde rasgele üretilirler.

$$IPOP=(P_H-P_L) \text{ x rasgele}\{N_{ipop}, N_{par}\}+P_L \quad (3.14)$$

Burada P_H parametrelerin üst sınır değerini, P_L parametrelerin alt sınır değerini ve rasgele$\{N_{ipop}, N_{par}\}$, N_{ipop}xN_{par} boyutlu matris formunda 0-1 arasında üretilen rasgele sayıları belirtmektedir. Tez çalışmasında kromozom kodlaması için kullanılan parametre sayısı Şekil 3.4 ile görüleceği gibi 20'dir. Her bir parametre için P_L =0 ve P_H=1 olarak belirlenmiştir. Başlangıç popülasyon sayısı N_{ipop} = 100 olarak belirlenmiştir.

G_1	G_2	G_3	G_4	G_5	$ö_{11}$	$ö_{12}$	$ö_{21}$	$ö_{22}$	$ö_{23}$	$ö_{31}$	$ö_{32}$	$ö_{33}$	$ö_{34}$	$ö_{41}$	$ö_{42}$	$ö_{43}$	$ö_{44}$	$ö_{51}$	$ö_{52}$

Şekil 3. 4 Gerçek kodlu GA'da kullanılan kromozom yapısı

Kromozomda görülen ilk beş parametrenin grup ağırlıklarını (W_j), kalan kısımların ise her bir grup altındaki özellik ağırlıklarını (w_{ji}) gösterdiği kabul edilmiştir.

Sistemde ikili kodlu GA kullanımı gerçekleştirildiğinde ise kromozom yapısı Şekil 3.5 ile görüleceği gibi 15 parametreden oluşan ve her bir parametrenin 1 yada 0 değerini aldığı bir yapıya sahip olmuştur. Bu yapıya sahip olunma durumunda eşitlik (3.1) ile belirtilen cümle skor fonksiyonundaki W_j grup ağırlıkları 1 olarak alınmıştır.

$ö_{11}$	$ö_{12}$	$ö_{21}$	$ö_{22}$	$ö_{23}$	$ö_{31}$	$ö_{32}$	$ö_{33}$	$ö_{34}$	$ö_{41}$	$ö_{42}$	$ö_{43}$	$ö_{44}$	$ö_{51}$	$ö_{52}$

Şekil 3. 5 İkili kodlu GA'da kullanılan kromozom yapısı

GA'nın metin özetlemedeki kullanım amacı sistem özetleri ile ideal özetler arasındaki en yüksek çakışma oranını elde etmektir. Bu amaçla GA yapısında yer alacak olan uygunluk fonksiyonu, S sistem özeti ve T ideal özet olmak üzere aşağıdaki gibi belirlenmiştir.

$$P_{Uygunluk} = \frac{|S \cap T|}{|S|} \quad (3.15)$$

Bu aşamadan sonra gelecek nesilde hangi kromozomların yer alacağını belirleme adına kromozomlar uygunluk değerine göre büyükten küçüğe doğru sıralanır. Gerçek kodlu GA'da gelecek iterasyonda kullanılmak üzere N_{pop} kadar kromozom tutulur ve geri kalanı dikkate alınmaz. Doğal seçim işlemi çaprazlama ve mutasyon ile en iyi uygunluk

değerine sahip olan kromozom bulunana kadar devam ettirilir. Tez çalışması kapsamında çaprazlama oranı 0,80 olarak ve mutasyon oranı 0,10 olarak alınmıştır.

Genetik algoritmanın çalışma döngüsünün sona erip, çözüme ulaşılabilmesi için önceden belirlenmiş bir çıkış koşulunun sağlanmış olması gerekir. Tez çalışmasında çıkış koşulu önceden belirlenen en fazla popülasyon sayısına ulaşılması şeklinde belirlenmiştir. Tez çalışmasında önceden belirlenen en fazla popülasyon sayısı 100 olarak alınmıştır.

3.3 Melez Sistem Sonuçlarının Yorumlanması

Literatürde İngilizce dokümanlar üzerinde çalışan ve cümle seçimi için kullanılan yapısal ve anlamsal özelliklerin birleşimini sağlayan melez sistem önerileri mevcuttur. Bu önerilerde özelliklerin birleşimi ile elde edilen yapıların sistem başarımları üzerindeki olumlu etkileri vurgulanmış ve bireysel özelliklerin özetleme üzerindeki etkileri incelenmiştir [10, 11, 13, 16, 64]. Tez çalışmasının bu bölümü ile çalışma kapsamında hazırlanmış olan GA ve BAHS tabanlı melez sistemin etkisi, 130 Türkçe haber dokümanından oluşan ve üç özetleyici tarafından oluşturulmuş olan ideal özet dokümanlarını içeren VeriSeti-1 üzerinde test edilmiştir. Ayrıca melez sistemin istikrarı, yine 20 adet Türkçe haber dokümanını ve toplamda otuz kişi tarafından oluşturulmuş olan ideal özet dokümanlarını içeren VeriSeti-2 vasıtasıyla sınanmıştır.

Çizelge 3.11, tez çalışması kapsamında incelenen on beş özelliğin, VeriSeti-1 üzerindeki F-ölçüm değerine göre hesaplanmış olan bireysel başarımlarını göstermektedir. Bu çizelgede koyu çizgilerle çevrelenmiş kısım GA ve BAHS yardımıyla özelliklerin beş grup altında birleşimini sağlayan melez sistemin başarım oranlarını içermektedir.

GA tabanlı birleşim sisteminde hem gerçek kodlu (GK) hem de ikili kodlu (İK) genetik algoritma kullanılmıştır. Uygulamalar sırasında GA'lar iki farklı durum altında ele alınmıştır. Bu durumlardan ilki, sistemde özetlenecek olan dokümanların hem eğitim hem de sınama amaçlı kullanıldığı durumdur (EVSD). İkinci durum ise, veri seti üzerinde biri hariç çapraz geçerlemenin kullanıldığı durumudur (BHÇGD). BHÇGD'ye göre, incelenen sistemde N adet doküman varken sistem N-1 eğitim kümesi üzerinde eğitilir ve dışarıda kalan bir örnek üzerinde sınanır. Bu işlem her örnek bir kez sınama

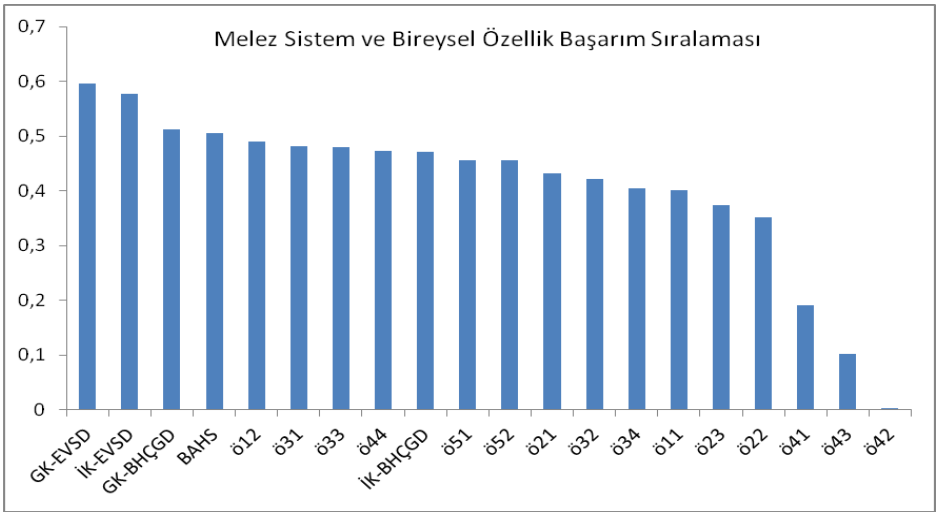
amaçlı kullanılacak şekilde tekrarlanır. Yani sistem N kez çalıştırılmış olur ve sistem başarımı denemelerin başarım ortalaması alınarak belirlenir. Bu yolla verinin etkin bir şekilde kullanılması mümkün kılınmıştır. Çizelgenin en son kolonunda görülen GK-EVSD ifadesi gerçek kodlu kromozom yapısına sahip eğitim ve sınama durumunda kullanılan GA'yı; GK-BHÇGD ifadesi gerçek kodlu kromozom yapısına sahip biri hariç çapraz geçerleme durumunda kullanılan GA'yı; İK-EVSD ifadesi ikili kodlu kromozom yapısına sahip eğitim ve sınama durumunda kullanılan GA'yı; İK-BHÇG ifadesi ikili kodlu kromozom yapısına sahip biri hariç çapraz geçerleme durumunda kullanılan GA'yı belirtmektedir.

Çizelge 3. 11 Melez sistemin ve bireysel özelliklerin VeriSeti-1'deki başarımları

	Özellikler	**Özetleyici1**	**Özetleyici2**	**Özetleyici3**	**Başarım Ortalaması**
G1	**ö11**	0,4225	0,4140	0,3663	0,4009
	ö12	0,5533	0,4937	0,4253	0,4908
G2	**ö21**	0,4412	0,4588	0,3944	0,4315
	ö22	0,3705	0,3970	0,2887	0,3521
	ö23	0,3938	0,3903	0,3363	0,3735
G3	**ö31**	0,5398	0,4850	0,4177	0,4808
	ö32	0,4509	0,4233	0,3914	0,4219
	ö33	0,5409	0,4864	0,4109	0,4794
	ö34	0,4445	0,4040	0,3673	0,4053
G4	**ö41**	0,1821	0,1981	0,1940	0,1914
	ö42	0,0023	0,0031	0,0035	0,0030
	ö43	0,1029	0,0941	0,1110	0,1027
	ö44	0,5321	0,4811	0,4059	0,4730
G5	**ö51**	0,5111	0,4668	0,3906	0,4562
	ö52	0,5084	0,4692	0,3907	0,4561
Melez Sistem	**GK-EVSD**	0,6342	0,5987	0,5571	0,5967
	İK-EVSD	0,6186	0,5789	0,5331	0,5768
	GK-BHÇGD	0,5648 S.Sapma: 0,1749	0,5040 S.Sapma: 0,2068	0,4732 S.Sapma: 0,2205	0,5114
	İK-BHÇGD	0.53707 S.Sapma: 0.1827	0.4456 S.Sapma: 0.2055	0,4312 S.Sapma: 0.2209	0,4713
	BAHS	0,5615	0,4910	0,4621	0,5048

Çizelge 3.13'ün en son kolonu, üç özetleyici temel alındığında elde edilen ortalama başarım sonuçlarını göstermektedir. Çizelgede görülen ortalama başarım sonuçlarına

göre, İK-BHÇGD hariç diğer (GK-EVSD, GK-BHÇGD ve İK-EVSD) durumları ile cümle çıkarımına dayalı bir metin özetleme sistemi için tasarlanan bir melez yapının, sistem üzerinde olumlu bir etkiye sahip olduğu görülmüştür. Bu durum Şekil 3.6 ile daha net bir şekilde anlaşılabilir:



Şekil 3. 6 Melez sistem ve bireysel özelliklerin başarım sıralaması

Bu şekle göre, melez sistemde en yüksek başarım değerine sahip olan durum GK-EVSD ve İK-EVSD durumlarıdır. GA'lar, tüm veri seti baz alındığında, ideal özetlere benzeme oranı yüksek olan özet dokümanlarına ulaşmayı sağlayan ağırlık değerlerini otomatik olarak bulma konusunda oldukça başarılıdır. Ancak bu uygulamada GK-BHÇGD ve İK-BHÇGD'nin yansıttığı sonuçlar daha gerçekçidir. Çünkü bu durumda veri seti çapraz geçerleme ile daha etkin bir şekilde kullanılmıştır. Bu süreçte başarım sonuçlarının standart sapma değerleri Çizelge 3.11'de görüldüğü gibi üç özetleyici için GK-BHÇGD durumunda sırasıyla 0,1749; 0,2068; 0,2205 ve İK-BHÇGD durumda 0,1827; 0,2055; 0,2209 değerlerine sahiptir. Yüksek sayılabilecek bu standart sapma değerlerine sahip olma durumu, GA'lar ile elde edilen ağırlık değerlerinin çok fazla genelleştirilemeyeceği sonucunu ortaya çıkarmıştır. Ayrıca GA'ların kullanımı sırasında GK'lı kromozom yapılarının kullanımı, İK'lı kromozom yapılarının kullanımlarına göre daha iyi sonuçlar üretmektedir.

Uygulamalar neticesinde uzmanların görüşlerine göre belirlenmiş ağırlıkları içeren BAHS tabanlı bir melez sistemin, bireysel özelliklerin başarımlarından daha yüksek sonuçlar verdiği görülmüştür. Eğiticisiz bir öğrenme yapısına sahip olan BAHS tabanlı

melez sistemin başarımı, GK-BHÇGD başarımına oldukça yakın sonuçlar üretmiş ve ve İK-BHÇGD başarımını geçmiştir. Bu sonuçlara göre BAHS tabanlı bir melez sistemin metin özetleme sisteminde kullanılabileceği açıktır. Özetleme sisteminin başarımı insanlar tarafından oluşturulan ideal özetler ile otomatik özetlerin kıyaslaması yoluyla yapılıyorsa, bir melez sistemde ağırlıklı olarak hangi özelliklerin kullanılması gerektiği uzman görüşleri altında belirlenebilir. Ancak BAHS'nin bu veri seti üzerinde başarılı olmasının en önemli etkeni, veri setinde yapısal olarak benzer dokümanların bulunuyor oluşudur (dokümanların tamamı çeşitli haber portallarından toplanmış haber dokümanlarını içermektedir). BAHS'nin dezavantajı ancak yapısal olarak benzer şekilde hazırlanmış dokümanlar üzerinde başarı elde edebilmesidir.

Şekil 3.6 incelenmeye devam edildiğinde, VeriSeti-1'i oluşturan özetleyicilerin en çok özellik $ö_{12}$ ile ifade edilen "Kelimelerin Dağıtımsal Özelliği"nin yansıttığı durumu dikkate alarak özet çıkardıkları görülür. Bu özellik Bölüm 2'de $T_{frekans}$ gösterimi ile belirtilen ve metin sınıflama probleminde sistem başarısını arttırdığı [60] nolu referans ile belirtilen çalışma ile gösterilmiş olan özelliktir. Tez çalışmasıyla bu özelliğin kullanımı ilk kez metin özetlemeye uyarlanmıştır. Sonuçlardan görüldüğü gibi VeriSeti-1 üzerinde, kelimelerin dağıtımsal özelliği 0,490 F-ölçüm değeri ile en yüksek başarıma sahip olan özellik olmuştur. Kelimelerin dağıtımsal özelliği dışında VeriSeti-1 üzerinde, özetleyicilerin en çok dikkat ettikleri diğer durumlar $ö_{31}$-cümle uzunluğu , $ö_{33}$-kelime cümle skoru bilgisi, $ö_{44}$-isim soylu kelimeleri içerme durumu, ve $ö_{51}$- anlamsal özellik gibi özelliklerdir. Özet çıkarma işlemi sırasında en az dikkate alınan durum ise $ö_{42}$ ile ifade edilen "?" ve "!" içerme durumu olmuştur.

Çizelge 3.12 melez sistemin, VeriSeti-2 üzerindeki F-ölçüm değerleri ile hesaplanmış başarım oranlarını göstermektedir. Çizelge 3.13 ise melez sistemi oluşturan bireysel özelliklerin başarım oranlarını içermektedir. Her iki çizelgede ortalama satırlarına bakıldığında, VeriSeti-1'de görüldüğü gibi, melez yapının sistem üzerindeki olumlu etkileri görülebilir. Çizelge 3.12'ye göre F-ölçüm değerleri BAHS tabanlı melez sistem için 0,552; GK-EVSD için 0,650; İK-EVSD için 0,631; GK-BHÇGD 0,566 için ve İK-BHÇGD 0,560'dır. BAHS yine BHÇGD ile kullanılan GA'lara yakın sonuçlar üretmiştir. Ayrıca, GK'lı kromozom yapılarının kullanımı, İK'lı kromozom yapılarının kullanımlarına göre daha iyi sonuçlar üretmiştir. Bu değerler otuz kişiye ait olan başarımların ortalamasını

yansıtmaktadır. Dolayısıyla bu analizler ile tez çalışmasında incelenen melez sistemlerin başarımlarının tesedüf olmadığı gösterilmiştir.

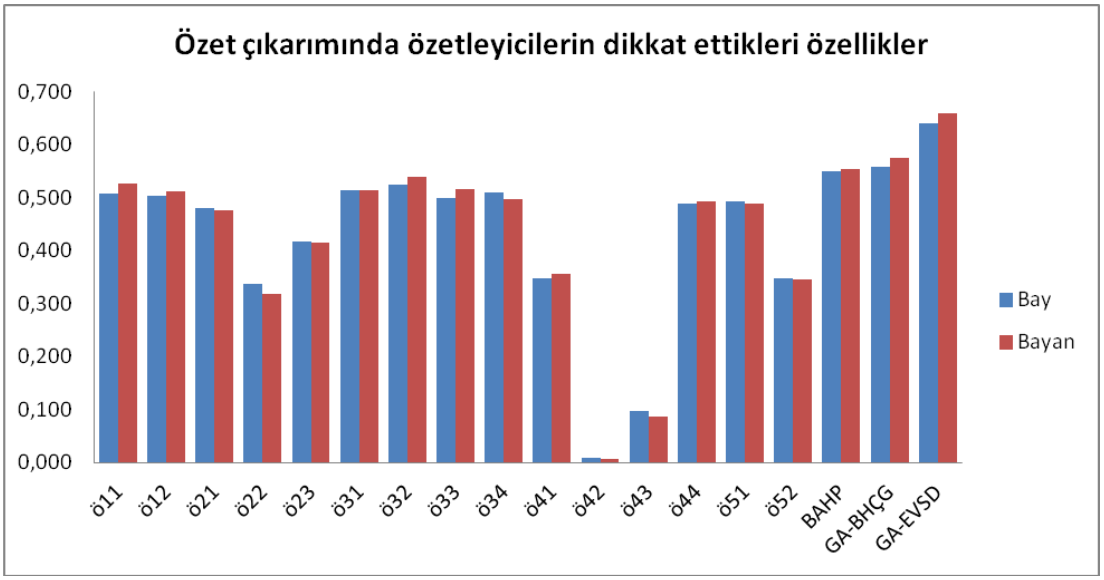
Çizelge 3. 12 Melez sistemin VeriSeti-2'deki başarımları

	Melez Sistem				
	BAHP	**İK-EVSD**	**GK-EVSD**	**GK-BHÇGD**	**İK-BHÇGD**
Bay1	0,597	0,638	0,638	0,576	0,500
Bay2	0,643	0,722	0,748	0,680	0,693
Bay3	0,597	0,643	0,668	0,586	0,562
Bay4	0,576	0,643	0,669	0,564	0,500
Bay5	0,618	0,665	0,678	0,620	0,603
Bay6	0,503	0,588	0,601	0,531	0,478
Bay7	0,593	0,669	0,686	0,649	0,672
Bay8	0,476	0,574	0,599	0,497	0,524
Bay9	0,574	0,620	0,633	0,603	0,533
Bay10	0,553	0,609	0,609	0,565	0,667
Bay11	0,445	0,552	0,555	0,403	0,458
Bay12	0,53	0,602	0,653	0,534	0,567
Bay13	0,447	0,553	0,562	0,447	0,445
Bay14	0,63	0,672	0,697	0,612	0,612
Bay15	0,47	0,587	0,612	0,489	0,470
Bayan1	0,534	0,607	0,626	0,534	0,522
Bayan2	0,597	0,684	0,687	0,648	0,624
Bayan3	0,533	0,688	0,713	0,615	0,655
Bayan4	0,616	0,670	0,667	0,637	0,653
Bayan5	0,512	0,637	0,658	0,608	0,495
Bayan6	0,518	0,603	0,636	0,528	0,551
Bayan7	0,487	0,578	0,628	0,484	0,522
Bayan8	0,562	0,673	0,673	0,628	0,670
Bayan9	0,588	0,668	0,699	0,638	0,667
Bayan10	0,512	0,599	0,628	0,512	0,499
Bayan11	0,649	0,744	0,744	0,708	0,676
Bayan12	0,562	0,630	0,648	0,513	0,568
Bayan13	0,472	0,551	0,57	0,422	0,408
Bayan14	0,578	0,612	0,633	0,573	0,488
Bayan15	0,601	0,659	0,684	0,580	0,526
Ortalama	0,552	0,631	0,650	0,566	0,560

Çizelge 3.13 ile bireysel özelliklerin bay ve bayan özetleyiciler bazında başarım ortalamaları temel alınarak, özetleyicilerin özet çıkarımı sırasında dikkat ettikleri özelliklerin aynı olup olmadığı gözlemlenmiştir. Bu durum Şekil 3.7 yardımıyla incelenmiştir.

Çizelge 3. 13 Bireysel özelliklerin VeriSeti-2 üzerindeki başarımları

	G_1		G_2			G_3				G_4				G_5	
	$ö_{11}$	$ö_{12}$	$ö_{21}$	$ö_{22}$	$ö_{23}$	$ö_{31}$	$ö_{32}$	$ö_{33}$	$ö_{34}$	$ö_{41}$	$ö_{42}$	$ö_{43}$	$ö_{44}$	$ö_{51}$	$ö_{52}$
Bay1	0,480	0,580	0,530	0,291	0,468	0,584	0,493	0,572	0,480	0,418	0,000	0,079	0,599	0,545	0,410
Bay2	0,623	0,538	0,551	0,289	0,488	0,566	0,603	0,558	0,569	0,438	0,000	0,096	0,510	0,524	0,356
Bay3	0,603	0,518	0,493	0,320	0,393	0,530	0,582	0,522	0,561	0,373	0,013	0,092	0,474	0,503	0,303
Bay4	0,535	0,538	0,441	0,406	0,391	0,553	0,585	0,512	0,568	0,281	0,013	0,133	0,495	0,514	0,341
Bay5	0,493	0,551	0,516	0,275	0,374	0,566	0,530	0,583	0,493	0,395	0,013	0,117	0,560	0,537	0,353
Bay6	0,537	0,470	0,534	0,349	0,472	0,424	0,553	0,428	0,541	0,320	0,013	0,121	0,410	0,423	0,372
Bay7	0,362	0,601	0,462	0,377	0,412	0,638	0,445	0,576	0,362	0,395	0,000	0,117	0,558	0,643	0,459
Bay8	0,468	0,480	0,349	0,288	0,353	0,474	0,443	0,487	0,438	0,316	0,000	0,096	0,473	0,458	0,374
Bay9	0,509	0,516	0,516	0,306	0,408	0,523	0,559	0,502	0,518	0,345	0,013	0,108	0,508	0,518	0,298
Bay10	0,563	0,486	0,476	0,308	0,401	0,482	0,497	0,478	0,576	0,343	0,000	0,038	0,484	0,484	0,288
Bay11	0,476	0,378	0,491	0,406	0,449	0,381	0,484	0,343	0,484	0,308	0,013	0,083	0,325	0,356	0,306
Bay12	0,468	0,463	0,484	0,445	0,405	0,474	0,488	0,453	0,522	0,299	0,013	0,083	0,491	0,460	0,318
Bay13	0,470	0,447	0,399	0,350	0,374	0,458	0,487	0,408	0,491	0,249	0,013	0,121	0,389	0,416	0,324
Bay14	0,528	0,563	0,503	0,288	0,424	0,578	0,591	0,583	0,495	0,408	0,000	0,079	0,568	0,562	0,370
Bay15	0,505	0,433	0,470	0,358	0,445	0,474	0,509	0,474	0,534	0,336	0,013	0,079	0,499	0,458	0,328
Bayan1	0,608	0,447	0,547	0,289	0,476	0,430	0,520	0,447	0,558	0,306	0,013	0,079	0,416	0,403	0,255
Bayan2	0,547	0,580	0,451	0,273	0,388	0,558	0,563	0,574	0,522	0,390	0,000	0,067	0,527	0,556	0,343
Bayan3	0,648	0,412	0,495	0,277	0,412	0,387	0,594	0,441	0,623	0,268	0,013	0,096	0,389	0,360	0,249
Bayan4	0,576	0,512	0,499	0,298	0,453	0,539	0,530	0,518	0,509	0,295	0,013	0,071	0,529	0,535	0,352
Bayan5	0,451	0,512	0,395	0,333	0,420	0,553	0,418	0,508	0,376	0,424	0,000	0,083	0,493	0,518	0,327
Bayan6	0,537	0,488	0,499	0,339	0,399	0,441	0,583	0,483	0,528	0,314	0,000	0,079	0,498	0,424	0,316
Bayan7	0,368	0,512	0,402	0,441	0,293	0,520	0,368	0,495	0,327	0,349	0,000	0,083	0,491	0,537	0,437
Bayan8	0,474	0,553	0,458	0,277	0,391	0,648	0,541	0,573	0,491	0,403	0,013	0,104	0,621	0,643	0,402
Bayan9	0,526	0,529	0,523	0,273	0,457	0,522	0,551	0,551	0,443	0,388	0,000	0,067	0,491	0,483	0,380
Bayan10	0,433	0,474	0,428	0,314	0,383	0,556	0,441	0,502	0,387	0,420	0,000	0,096	0,513	0,514	0,406
Bayan11	0,669	0,545	0,508	0,254	0,483	0,548	0,703	0,589	0,632	0,333	0,013	0,104	0,533	0,518	0,318
Bayan12	0,590	0,516	0,470	0,343	0,370	0,474	0,594	0,508	0,603	0,422	0,000	0,079	0,448	0,423	0,324
Bayan13	0,426	0,505	0,463	0,420	0,384	0,474	0,522	0,466	0,455	0,301	0,013	0,108	0,381	0,435	0,373
Bayan14	0,573	0,561	0,540	0,277	0,444	0,499	0,569	0,545	0,544	0,323	0,000	0,079	0,510	0,448	0,341
Bayan15	0,480	0,538	0,469	0,362	0,473	0,563	0,576	0,551	0,443	0,418	0,013	0,104	0,558	0,537	0,368
Ortalama	0,517	0,508	0,479	0,327	0,416	0,514	0,531	0,507	0,502	0,353	0,007	0,091	0,491	0,491	0,346



Şekil 3. 7 VeriSeti-2'de özetleyicilerin dikkat ettikleri özellikler

Bu analize göre cinsiyet ayrımı farketmeksizin bay ve bayan özetleyicilerin dikkat ettikleri özelliklerin hemen hemen aynı olduğu gözlemlenmiştir.

Çizelge 3.13'ün en sonunda görülen ortalama satırı sıralandığında özetleyicilerin sırasıyla $ö_{32}$-"kelime sıklığı bilgisi"; $ö_{11}$-"cümle konumu"; $ö_{31}$-" cümle uzunluğu"; $ö_{12}$-"kelimelerin dağıtımsal özelliği"; $ö_{33}$-" kelime cümle skoru bilgisi"; $ö_{34}$-"ortalama kelime frekansı ve ters doküman frekansı"; $ö_{44}$-" isim soylu kelimeleri içerme durumu"; "$ö_{51}$-gizli anlamsal analize dayalı anlamsal özellik"; $ö_{21}$-" ilk cümleye olan benzerlik"; $ö_{23}$-"başlığa olan benzerlik"; $ö_{41}$-"sayısal karakter içerme durumu"; $ö_{52}$-" merkez olma durumu"; $ö_{22}$- "son cümleye olan benzerlik"; $ö_{43}$-"pozitif kelimeleri içerme durumu" ve $ö_{42}$-"?" ve "!" içerme durumu özelliklerine dikkat ettikleri görülmüştür. Buna göre, tez çalışması kapsamında ilk kez kullanılan $ö_{12}$-"kelimelerin dağıtımsal özelliği" yine başarım olarak kısa dokümanlar üzerinde de üst sıralarda yer almıştır.

Melez sistemin VeriSeti-1 üzerindeki çalışma süreleri incelenecek olursa 15 özelliğin tespit edilme süresi üç özetleyici bazında ortalamada 130 doküman için toplam 2,92 dakika sürmektedir. Bu süre üzerine BAHS için uzmanların matrisleri hazırlama süresinin; EVSD'ler için genetik algoritmanın ağırlık değerlerini bulma süresi olan 4,22 dakikanın; ve BHÇGD için EVSD'lerin ağırlık değerlerini bulma süresinin çaprazlama sayısı ile çarpımı kadar dakikanın eklenmesi gerekmektedir.

Melez sistemin VeriSeti-2 üzerindeki çalışma süreleri incelenecek olursa 15 özelliğin tespit edilme süresi 20 doküman için toplam 0,12 dk sürmektedir. Bu süre üzerine BAHS için uzmanların matrisleri hazırlama süresinin; EVSD'ler için genetik algoritmanın ağırlık değerlerini bulma süresi olan 0,32 dakikanın; ve BHÇGD için EVSDlerin ağırlık değerlerini bulma süresinin çaprazlama sayısı ile çarpımı kadar dakikanın eklenmesi gerekmektedir.

Sonuç olarak tez çalışmasının bu bölümünde bireysel özelliklerin birleşimine dayalı olan bir melez sistemin Türkçe veri setleri üzerindeki etkileri incelenmiş ve melez sistemin olumlu etkilere sahip olduğu gösterilmiştir. Ayrıca özellik birleşim aşamasında, uzman gücü ile oluşturulan ikili kıyaslama matrisilerine dayalı BAHS'nin metin özetleme probleminde kullanılabileceği gözler önüne serilmiştir. GA tabanlı melez sistemde gerçek kodlu kromozom yapılarının kullanımının, ikili kodlu GA'lara göre daha iyi sonuçlar verdiği gösterilmiştir. Son olarak "kelimelerin dağıtımsal özelliği"nin metin özetleme üzerinde kullanılabilir bir özellik olduğu ifade edilmiştir.

BÖLÜM 4

SONUÇ VE ÖNERİLER

Bu tez çalışması, cümle seçimine dayalı olan otomatik metin özetleme konusunu incelemiştir. Günümüzde İngilizce en sık kullanılan dil olduğundan, İngilizce dokümanlar üzerinde çalışan bilimsel çalışmalara sıkça rastlanmaktadır. Ancak maalesef, Türkçe dokümanlar üzerinde çalışmış olan yeterli sayıda bilimsel yayın mevcut değildir. Bu durumun en temel nedeni bu alanda üzerinde çalışılabilecek geniş kapsamlı veri setlerinin bulunmayışıdır. Tez çalışması ile metin özetleme probleminde kullanılabilecek Türkçe dokümanlardan oluşan iki veri seti hazırlanmıştır. Bu veri setlerinden ilki çeşitli haber sitelerinden toplanmış olan 130 haber dokümanını ve bu haber dokümanlarına ait olan üçer kişi tarafından oluşturulmuş olan özet dokümanlarını kapsamaktadır. İkinci veri seti ise ilk veri setine göre daha kısa olan 20 haber dokümanını ve bu haberlere ait olan otuz kişi tarafından oluşturulmuş özet dokümanlarını barındırmaktadır. İngilizce dokümanları kapsayan veri setleri dahi, genelde bir yada iki özetleyici tarafından oluşturulmuş özet gruplarına sahiptir. Bu bağlamda tez çalışması kapsamında hazırlanan veri setlerinin yöntemlerin istikrarının ölçülmesi açısından önem teşkil ettiği söylenebilir.

Tez çalışmasının ilk bölümü ile metin özetlemenin genel tanıtımı yapılmış ve literatürde metin özetleme alanında yapılan eğiticisiz ve eğiticili öğrenme yöntemlerine sahip olan bilimsel çalışmalar incelenmiştir. İncelenen bilimsel çalışmalar yapılmış oldukları yıllara göre bir çizelge ile sıralanmıştır. Bu sıralamaların ardından, tez çalışması kapsamında kullanılan çıkarıma (VeriSeti-1;VeriSeti-2;VeriSeti-3) ve yoruma dayalı (VeriSeti-4) özet

dokümanlarını içeren veri setlerinin tanıtımı yapılmıştır. Son olarak metin özetleme sistemlerinin başarım değerlendirme süreçleri üzerinde durulmuştur.

Tez çalışmasının ikinci bölümünde eğiticisiz bir öğrenme modeline sahip olan gizli anlamsal analiz yöntemi ele alınmıştır. Bu yöntem tekil değer ayrışımına dayalı olup cümleleri oluşturan terimleri bir takım dönüşüm matrisleri ile anlamsal olarak kümlemekte ve bu kümelenmiş yapılar ile metnin içindeki gizli anlamsal yapıyı ortaya çıkartmaktadır. Tez çalışmasında literatürde gizli anlamsal analiz temeline sahip olan ve değişik cümle seçim kriterleriyle birbirlerinden farklılaşan dört bilimsel çalışma incelenmiştir [40, 41, 42, 49]. Bu çalışmaların tümü, metin özetleme işleminde ilk aşama olarak, incelenen dokümanları terim-cümle matrislerine dönüştürmüşlerdir. Bu dönüşümü dokümanları oluşturan terimlerin frekans bilgilerini kullanarak gerçekleştirmişlerdir. Tez çalışması kapsamında terim-cümle matrisi oluşturulurken terim frekanslarının ($T_{frekans}$) yanında, terimlerin yoğunluklarının ortalamasını içeren ($T_{Dağıtımsal}$) özelliğini ve terimleri içeren cümlelerin önem derecelerini on üç farklı yapısal özelliğin toplamı ile gösteren ($C_{önem}$) ifadesini içeren yeni bir ağırlık değeri önerilmiştir. Önerilen yeni ağırlık değeri ($Yeni_{Ağırlık}$), tez çalışması kapsamında incelenen dört yöntemin sahip olduğu sistem başarımlarını arttırmıştır. Bu başarım artışları hem çıkarıma dayalı özetlerin bulunduğu iki Türkçe (VeriSeti-1; VeriSeti-2)ve bir İngilizce veri seti (VeriSeti-3) üzerinde hem de yoruma dayalı özetleri içeren bir İngilizce veri seti (VeriSeti-4) üzerinde elde edilmiştir.

Çıkarıma dayalı olan metin özetleme sistemlerinde dokümanlardaki cümlelerin önem derecelerini belirleyen bazı yapısal ve anlamsal cümle özellikleri kullanılmaktadır. Literatürde İngilizce dokümanlar üzerinde çalışan ve cümle seçimi için kullanılan yapısal veya anlamsal özelliklerin birleşimini sağlayan melez sistem önerileri mevcuttur. Bu önerilerde özelliklerin birleşimi ile elde edilen yapıların sistem başarımları üzerindeki olumlu etkileri vurgulanmış ve bireysel özelliklerin katkıları üzerinde durulmuştur [10, 11, 13, 16, 64]. Tez çalışması kapsamında, neredeyse literatürde kullanılan tüm özellikleri barındıran, geniş kapsamlı bir melez sistem tasarlanmıştır ve tasarlanan melez sistemin Türkçe metinler üzerideki etkisi ayrıntılı bir şekilde analiz edilmiştir. Melez sistem ile cümle önemini yansıtan özellikler iki farklı yöntemin kullanılmasıyla birleştirilmiştir. Bu yöntemlerden ilki uzman gücüne dayalı olan bulanık analitik

hiyerarşi süreci yöntemdir. İkincisi ise eğiticili bir öğrenme yapısına sahip olan genetik algoritma tabanlı otomatik bir birleşim yöntemidir.

Bulanık tabanlı birleşim yöntemi özelliklerin uzman görüşlerine göre kıyaslanmalarını sağlayan ikili kıyas matrislerine dayalıdır. Analitik hiyarerşi süreci, [66] çalışmasıyla sınıflayıcı birleşiminde, [67] çalışmasıyla diz üstü bilgisayar seçiminde ve [68] çalışmasıyla veri tabanı yönetimi projesinde kullanılmıştır. Tez çalışması ile bu süreç ilk kez metin özetleme problemi üzerinde, yapısal ve anlamsal özelliklerin sahip olması gereken uygun ağırlık değerlerinin tespit edilmesi için kullanılmıştır. Bulanık melez sistem, uzmanlar tarafından ortak bir görüşe göre oluşturulmuş toplam beş ana grubun ikili kıyaslanma değerlerini içeren bir ana kıyas matrisini ve bu matris dışında ana grupları oluşturan özelliklerin ikili kıyas değerlerini içeren beş farklı kıyas matrisini içermektedir. Bulanık hiyerarşi süreci uzmanların oluşturduğu bu kıyas matrislerini kullanılarak, yapısal ve anlamsal özelliklere ait olan ağırlık değerlerini belirlemektedir. Sonuçta özelliklerin birleşimi için tüm dokümanlar üzerinde kullanılabilecek genel ağırlık değerlerinin elde edilmesini sağlar. Uygulamalar sonucunda tez çalışması kapsamında önerilmiş olan bulanık analitik hiyerarşi sürecine dayalı bir melez sistemin Türkçe veri setlerinden (VeriSeti-1 ve VeriSeti-2) oluşan dokümanlar üzerinde olumlu sonuçlar verdiği görülmüştür. Veri setlerinin eğitilmesini gerektirmeyen ve genel ağırlık değerlerinin tespit edilmesini sağlayan bu sürecin, metin özetlemede başarılı olabilmesi için veri setlerini oluşturan dokümanların yapısal benzerliklerinin çok farklı olmaması gerekmektedir.

Tez çalışmasında tasarlanan melez sitemde bulanık analitik hiyerarşi sürecine dayalı birleşim yaklaşımıyla birlikte, özellik ağırlıklarını otomatik olarak tespit eden genetik algoritma tabanlı bir birleşim yaklaşımının etkileri de incelenmiştir. Bu birleşim yaklaşımında hem gerçek kodlu hem de ikili kodlu genetik algoritma kullanılmıştır. Gerçek kodlu genetik algoritmada sistemin sahip olduğu kromozomlar 0-1 aralığındaki ondalık sayıları içeren toplam 20 adet parametre içermektedir. Bu parametrelerden ilk beşi melez sistemi oluşturan ana grup ağırlıklarını, kalan on beşi ise bu gruplar altında toplanan bireysel özellik ağırlıklarını temsil etmektedir. İkili kodlu genetik algoritmada ise kromozomlar 0 ve 1 değerlerinden oluşan toplam on beş parametreye sahiptir. Gerçek yada ikili kodlu genetik algoritma yapıları iki farklı durum altında kullanılmıştır.

Bu durumlardan ilki, sistemde özetlenecek olan dokümanların hem eğitim hem de sınama amaçlı kullanıldığı durumdur. İkinci durum ise, veri seti üzerinde biri hariç çapraz geçerlemenin kullanıldığı durumudur. Biri hariç çapraz geçerlemeye göre, incelenen sistemde N adet doküman varken sistem N-1 eğitim kümesi üzerinde eğitilir ve dışarıda kalan bir örnek üzerinde sınanır. Bu işlem her örnek bir kez sınama amaçlı kullanılacak şekilde tekrarlanır. Yani sistem N kez çalıştırılmış olur ve sistem başarımı denemelerin başarımının ortalaması alınarak belirlenir. Bu yolla verinin etkin bir şekilde kullanılması mümkün kılınmıştır.

Melez sistem uygulamalarının sonucunda genetik algoritma tabanlı birleşim yöntemlerinin veri setleri üzerinde olumlu etkilere sahip olduğu sonucu ortaya çıkmıştır. İkili kodlu genetik algoritmalar yerine, bulanık analitik hiyerarşi sürecindeki gibi ondalıklı sayılardan oluşan bir ağırlık değeri yapısının kullanılması daha yüksek sonuçlar üretmiştir. Aynı zamanda uygulamalar neticesinde genetik algoritmaların, tüm veri seti baz alındığında ideal özetlere benzeme oranı yüksek olan özet dokümanlarına ulaşmayı sağlayan ağırlık değerlerini otomatik olarak bulma konusunda oldukça başarılı olduğu söylenebilir. Ancak biri hariç çapraz geçerlemenin kullanılma durumunda veri daha etkin bir şekilde kullanıldığından, çalışma süresini uzatmasına rağmen, bu durumun kullanılmasının daha sağlıklı olduğu tespiti yapılmıştır.

Tez çalışması kapsamında ilk bölüm ile incelenen eğiticisiz ve ikinci bölüm ile incelenen eğiticili öğrenme teknikleri kıyaslanacak olursa çalışma süresi bakımından eğiticisiz öğrenme tekniklerinin daha pratik çözümler ürettiği söylenebilir. Üstelik önerilen yeni ağırlık değeri yaklaşımı ile eğiticisiz sistemin elde ettiği başarım oranları, genetik algoritmalar ile elde edilen başarım sonuçlarına çok yakın değerlerin üretilmesi sağlanmıştır.

Tez çalışmasının amacı özellikle metin özetleme problemindeki yöntemleri analiz etmek ve yöntem başarımlarını arttıracak yeni öneriler sunabilmekti. Yapılan uygulamalar ile hem eğiticisiz hem de eğiticili öğrenmeye dayalı yöntemler ayrıntılı bir şekilde analiz edilmiş ve bir takım yenilikler önerilerek sistem başarımlarının artması sağlanmıştır. Ayrıca tez çalışmasıyla, cümle önemlerini belirten özellikler ayrıntılı bir şekilde incelenmiş ve cümle birleşimini sağlayan bir melez sistemin farklı koşullar

altındaki etkileri gözler önüne serilmiştir. Sonuçta bu alanda çalışacak araştırmacılar için kullanışlı sonuçlar elde edilmiştir.

Bu tezin metin özetleme alanında yapılan diğer çalışmalara katkıda bulunması dileğimizdir.

KAYNAKLAR

[1] Luhn, H.R., (1958). “The automatic creation of literature abstracts.”, IBM Journal of Research Development, 2(2):159–165.

[2] Edmundson, H.P., (1969). “New methods in automatic extracting.”, Journal of the Association for Computing Machinery, 16(2): 264–285.

[3] Pollock, J. J. and Zamora, A., (1999), “Automatic Abstrating Research at Chemical Abstracts.”, In Inderjeet Mani and Mark Marbury, editors, Advances in Automatic Text Summarization, MIT Press.

[4] Kupiec, J., Jan O. P., and Francine C., (1995). “A trainable document summarizer.”, In Research and Development in Information Retrieval, 09-13 July 1995, Washington, United States.

[5] Pardo, T., Rino, L. and Nunes, M., (2003), “GistSumm: A summarization tool based on a new extractive method.”, In 6th Workshop on Computational Processing of the Portuguese Language, 2003,São Carlos, Brazil.

[6] Yeh, J.Y., Ke, H.R. , Yang, W.P. and Meng, I.H., (2005). “Text Summarization using a Trainable Summarizer and Latent Semantic Analysis”, Journal of Information Processing and Management, 41:75–95, 2005.

[7] Hernandez, R.A.G and Ledeneva, Y., (2009). “Word Sequence Models for Single Text Summarization”, In Proceeding of Second International confererence on Advances in Computer-Human Interactions, 1-7 Feb. 2009, Mexico.

[8] Quyang, Y., Li, W., Lu, Q. and Zhang, R., (2010). “A study on Position Information in Document Summarization”, In Procceding of Coling 2010, Poster Volume, August 2010, Beijing.

[9] Radev, R., Blair-Goldensohn, S. and Zhang, Z., (2001). “Experiments in Single and Multi-Docuemtn Summarization using MEAD”, In First Document Understanding Conference, New Orleans, LA.

[10] Kiani, A. and Akbarzadeh, M.R., (2006). “Automatic text summarization using: Hybrid fuzzy GA–GP.”, In IEEE international conference on fuzzy systems, 16–21 July. Vancouver, Canada.

[11] Suanmali, L., Salim, N. and Binwahlan, M.S., (2009). “Fuzzy Logic Based Method for Improving Text Summarization”, International Journal of Computer Science and Information Security, 2(1): 65-70.

[12] Kyoomarsi, F., Khosravi, H., Eslami, E. and Dehkordy, P.K, (2009). “Optimizing Machine Learning Approach Based on Fuzzy Logic in Text Summarization”, International Journal of Hybrid Information Technology, 2(2): 105-116.

[13] Binwahlan, M.S., Salim N. and Suanmali, L., (2010). “Fuzzy Swarm Diversity Hybrid Model for Text Summarization”, Information Processing and Management, 46(5):571-588.

[14] Silla C.N., Pappa, G.L., Freitas, A.A. and Celso A.A., (2004). “Automatic Text Summarization with Genetic Algorithm-Based Attribute Selection”, 9th Ibero-American Conference on AL, Lecture Notes in Computer Science, 3315 (2004):305–314.

[15] Witte, R., Krestel, R. and Bergler, S., (2007). “Generating Update Summaries for DUC 2007.” In the Document Understanding Workshop, 26-27 April 2007, Rochester, NewYork USA.

[16] Berker, M. and Güngör, T., (2012), “Using Genetic Algorithms with Lexical Chains for Automatic Text Summarization”, 4th International Conference on Agents and Artificial Intelligence, February 2012, p.595-600, Vilamoura, Portugal.

[17] Filatova, E. and Hatzivassiloglou, V., (2004), “A formal model for information selection in multi-sentence text extraction”, 20th International Conference on Computational Linguistics, August 2004, Geneva, Switzerland.

[18] McDonald, R., (2007). “A study of global inference algorithms in multi-document summarization”, 29th European Conference on IR Research, 2-5 April 2007, Rome, Italy.

[19] Alguliev, R.M., Aliguliyev, R.M. , Hajirahimove, M.S. and Mehdiyev, C., (2011), “MCMR: Maximum coverage and minimum redundant text summarization model”, Expert Systems with Applications, 38(2011): 14514-14522.

[20] Copeck, T., Szpakowicz, S. and Japkowic, N., (2002). “Learning How Best to Summarize”, In Workshop on Text Summarization, 2002, Philadelphia.

[21] Svore, K., Vanderwende, L. and Burges, C., (2007). “Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources”, In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 28–30 June 2007, Czech Republic.

[22] Wong, K.F., Wu, M. and Li, W., (2008). “Extractive Summarization Using Supervised and Semi-Supervised Learning”, In Proceedings of the 22nd ınternational Conference on Computational Linguistics, August 2008, Manchester.

[23] Hirao, T., Sasaki, Y. and Isozaki, H., (2002). “NTT's Text Summarization system for DUC-2002”, In Workshop on Text Summarizatio), Philadelphia.

[24] Lal, P. and Rueger, S., (2002). “Extract-based Summarization with Simplication”, In Workshop on Text Summarization, Philadelphia.

[25] Brandow, R., Mitze, K. and Rau, L.F., (1995). “Automatic condensation of electronic publications by sentence selection”, Information Processing Management, 31(5): 675-685.

[26] Steinberger, J., Massimo, P. and Sanchez-Graillet, A., (2005). “Improving the LSA based Summarization with Anaphora Resolution”, In Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, Vancouver.

[27] Orasan, C., (2007). “Pronomial Anaphora Resolution for Text Summarizaiton”, In Proceedings of the Recent Advances in Natural Language Processing, 2007, Borovets, Bulgaria.

[28] Azzam, S., Humphrey, K. and Gaizauskas. R., (1999). “Using coreference chains for text summarisation”, In A. Bagga, B. Baldwin, and S. Shelton, editors, Coreference and Its Applications, June 1999., University of Maryland, College Park, Maryland, USA.

[29] Baldwin, B. and Morton, T.S., (1998). “Dynamic coreference-based summarization”, In Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing, June 1998, Granada, Spain.

[30] Branimir, B. and Christopher K., (1997). “Salience-based content characterization of text documents”, In Proceedings of the Workshop on Intelligent Scalable Text Summarization, 1997, Madrid, Spain.

[31] Barzilay, R. and Elhadad, M., (1999). Using Lexical Chains for Text Summarization, In Inderjeet Mani and Mark Marbury, editors, Advances in Automatic Text Summarization, 1999, MIT Press.

[32] Karamuftuoglu, M., (2002). “An approach to summarizaion based on lexical bonds”, In Workshop on Text Summarization, 2002, Philadelphia.

[33] WordNet sözlüğü sayfası, http://wordnet.princeton.edu/, 1 Haziran 2009.

[34] Kan, M. Y. and McKeown, K., (1999). Information extraction and summarization: Domainn independence through focus types. Technical report, Computer Science Department, Columbia University.

[35] D'Avanzo, E., Magnini, B. and Vallin, A., (2004)., “Keyphrase Extraction for Summarization Purposes: The LAKE System at DUC-2004”, In the Document Understanding Workshop ,Boston, USA.

[36] Hovy, E. and Lin, C. Y., (1999), Automated Text Summarization in SUMMARIST, In Inderjeet Mani and Mark Marbury, editors, Advances in Automatic Text Summarization, MIT Press.

[37] Jing, H. and McKeown, K. R., (2000). “Cut and paste based text summarization”, In: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics, April 29-May 04, 2000, Seattle, Washington.

[38] Liu, M., Li, W., Wu, M. and Lu, Q., (2007). “Extractive Summarization Based on Event Term Clustering”, In Proceedings of the ACL, June 2007, Prague.

[39] Filatova E. and Hatzivassiloglou, V., (2004). “Event-based Extractive summarization”, In Proceedings of ACL 2004 Workshop on Summarization, July 2004, Barcelona, Spain.

[40] Gong, Y. and Liu, X., (2001). “Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis”, In the proceeding of ACM SIGIR, September 2001, New Orleans, Louisiana, United States.

[41] Murray, G., Renals, S. and Carletta, J., (2005). “Extractive summarization of meeting recordings”, In Proceedings of the 9th European Conference on Speech Communication and Technology, September 2005, Lisbon, Portugal.

[42] Steinberger, J., (2007). Text Summarization within the LSA Framework, PhD Thesis, University of West Bohemia in Pilsen, Czech Republic, January 2007.

[43] Bhandari, H., Shimbo, M., Ito, T., Matsumoto ,Y. and Bhandari, Y., (2007). “Generic Text Summarization Using Probabilistic Latent Semantic Indexing”, 3rd International Joint Conference on Natural Language Processing.

[44] Lee, J.H., Sun P., Chan-Min A. and Daeho K., (2009). “Automatic generic document summarization based on non-negative matrix factorization”, Inf. Process. Manage, 45(1): 20–34.

[45] Mashechkin, I.V., Petrovskiy, M.I., Popov, D. S. and Tsarev, D.V., (2011). “Automatic text summarization using latent semantic analysis”, Programming and Computer Software, 2011: 299-305.

[46] Altan, Z., (2004). "A Turkish Automatic Text Summarization System", IASTED International Conference on AIA, 16-18 February 2004, Innsbruck, Austria.

[47] Kılcı, Y. and Diri,B., (2008). Turkish Text Summarization System, Senior Project, Yıldız Technical University, Turkey.

[48] Cığır, C., Kutlu, M. and Cicekli, I., (2009). “Generic Text Summarization for Turkish”, The Computer Journal, 53(8):1315-1323.

[49] Özsoy, M., Çiçekli, İ. and Alpaslan, F.N., (2010). "Text Summarization of Turkish Texts using Latent Semantic Analysis", In Proceedings of the 23rd International Conference on Computational Linguistics, August 2010, Beijing, China.

[50] Güran, A., Güler, N. and Bekar, E., (2011). "Automatic summarization of Turkish documents using non-negative matrix factorization ", INISTA 2011, Istanbul, Türkiye.

[51] Pembe, C. (2011). Automated Query-Biased and Structure-Preserving Document Summarization for Web Search Tasks, PhD Thesis, Boğaziçi University, Turkey.

[52] CAST projesi, http://clg.wlv.ac.uk/projects/CAST/corpus/index.php, 10 Haziran 2012.

[53] DUC konferansı, http://duc.nist.gov/, 10 Haziran 2012.

[54] Dragomir, R., Hongyan J. and Malgorzata B., (2000). "Centroid based summarization of multiple documents", In ANLP/NAACL Workshop on Automatic Summarization, Seattle, USA.

[55] Salton, G., (1988). Automatic text processing: Automa ic Text Processing: The Transformation Analysis and Retrieval of Information by Computer, Addison-Wesley Publishing Company.

[56] Lin, C.Y., (2004). "ROUGE: a Package for Automatic Evaluation of Summaries", In Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, 25 – 26 July 2004, Spain.

[57] Radev, D., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Qi, H.,Celebi, A. and Liu, D., Drabek, E., (2003). "Evaluation Challenges in Large-scale Document Summarization", In Proceeding of the 41st meeting of the Association for Computational Linguistics, 2003, Sapporo, Japan.

[58] Johnson, L.W., Riess, R.D. and Arnold, J.T., (1997), Introduction to linear algebra, Addison-Wesley press.

[59] Xue, X.B. and Zhou, Z.H., (2009). "Distributional Features for Text Categorization," IEEE Trans. Knowledge and Data Eng., 21(3): 428–444.

[60] Ko, Y., Park, J. and Seo, J., (2002). "Automatic text categorization using the importance of sentences", In the proceedings of the 19th international conference on Computational linguistics, August 24-September 01, 2002, Taipei, Taiwan.

[61] Zemberek Projesi Geliştirme Sayfaları, https://zemberek.dev.java.net/, 1 Haziran 2009.

[62] Stanford Üniversitesi söz dizimsel analiz aracı, http://nlp.stanford.edu/software/tagger.shtml, 1 Haziran 2010.

[63] Porter Stemmer kök ayırma kodu, http://tartarus.org/martin/PorterStemmer/, 1 Mayıs 2010.

[64] Svore, K., Vanderwende, L. and Burges, C., (2007)., "Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources", In Proceedings of the Joint Conference on Empirical Methods in Natural LanguageProcessing and Computational Natural Language Learning, 2007, Prague, Czech Republic.

[65] Saaty, T.L., (1980), The Analytic Hierarchy Process, McGraw-Hill, New York.

[66] Felföldi, L. and Kocsor, A., (2004). "AHP-Based Classifier Combination", 4th International Workshop of PRIS, 2004, Porto.

[67] Srichetta, P. and Thurachon, W., (2012)., "Applying Fuzzy Analytic Hierarchy Process to Evaluate and Select Product of Notebook Computers", International Journal of Modeling and Optimization, 2(2): 168-173.

[68] Catak, F.O., Karabas, S. and Yildirim, S., (2012). "Fuzzy Analytic Hierarchy Based DBMS Selection In Turkish National Identity Card Management Project", International Journal of Information Sciences and Techniques, 2 (4): 29-38.

[69] Chang, D.Y., (1996). "Applications of the extent analysis method on fuzzy AHP", European Journal of Operational Research, 95: 649-655.

[70] Holland, J. H. (1975). Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI.

[71] Genel, H., (2004), Genetik Algoritmalar ile Portföy Optimizasyonu, Yüksek Lisans Tezi, Ankara Üniversitesi, Türkiye.

[72] Goldberg D.E., (1989)., Genetic algorithms in search, optimization, and machine learning, Addison-Wesley, Reading, MA.

[73] Beasley, D., Bull, D.R. and Martin, R.R., (1993)., "An Overview of Genetic Algortihms: Part 2, Research Topics", University Computing, 15(4): 170-181.

[74] Beasley, D., Bull, D.R. and Martin, R.R., (1993)., "An Overview of Genetic Algortihms: Part 1, Fundamentals", University Computing, 15(2): 58-69.

[75] Genetik algoritmalar Java paketi, http://jgap.sourceforge.net/, 1 Haziran 2012.

[76] Chris, H. and Ding, Q., (2005)., "A probabilistic model for latent semantic indexing", In Journal of the American Society for Information Science and Technology, 56(6): 597–608.

EK-A

YILLARA GÖRE ARTAN SIRADA DİZİLMİŞ LİTERATÜR ÇALIŞMASI

Referans-Yıl	Giriş Elemanı	Alan	Kullanılan Özellikler	Çıktı
Luhn [1], 1958	Tekli doküman özetleme	Teknik makaleler	•Terim filtreleme ve terim frekansı kullanılmıştır (düşük frekanslı terimler dikkate alınmamıştır.) •Cümleler içerdikleri önemli terimlere göre ağırlıklandırılmıştır. •Cümle segmentasyonu ve cümle seçimi gerçekleştirilmiştir.	Cümle seçimine dayalı özet
Edmundson [2], 1969	Tekli doküman özetleme	Belirli bir alandaki makaleler	•Bu teknikte kullanılan yapısal özellikler: *kelime sıklığı*, *ipucu söz öbekleri*, *başlık kelimeleri* ve *cümle konum* bilgisidir. •Külliyat (corpus) tabanlı bir metot kullanmıştır.	Cümle seçimine dayalı özet
Pollock ve Zamora [3], 1975	Tekli doküman özetleme	Kimya alanındaki makaleler	Çalışmada kullanılan ana ifadeler: •İpucu söz öbekleri •Terim sıklığı bilgisi •Cümle seçimi ve cümle düzenlemedir.	Yoruma dayalı olan gösterici cümleler (indicative abstracts)
Brandow vd. [25], 1995	Tekli doküman özetleme	Haber metinleri	Çalışmada kullanılan ana ifadeler: •Terim - cümle ağırlıklandırma •Artgönderim olmayan çözümleme (non-anaphora resolution) •Dokümandaki ilk cümle özete direkt eklenmiştir.	Cümle seçimine dayalı özet

Referans-Yıl	Giriş Elemanı	Alan	Kullanılan Özellikler	Çıktı
Kupiec vd.[4], 1995	Tekli doküman özetleme	Haber metinleri	•Kullandıkları özellikler; "Cümle uzunluğu kesme özelliği", "Belirli sözcük öbeği", "Paragraf Özelliği", "Konuya Has sözcük özelliği" ve "Büyük harfli sözcük özelliği"dir. •Bu özelliklerin birleştirilmesi için Bayes sınıflandırıcısı kullanır.	Cümle seçimine dayalı özet
Barzilay vd. [31], 1997	Tekli doküman özetleme	Bilinmiyor	•Terimleri gruplayarak sözlüksel zincirler (lexical chains) oluşturup dokümandaki konu başlıklarını belirlemişlerdir. •Cümle seçimi güçlü sözlüksel zincir içeren cümlelerin belirlenmesi ile yapılmıştır. •Artgönderim olmayan çözümleme (non-anaphora resolution) kullanılmıştır.	Cümle seçimine dayalı özet
Hovy ve Lin[36], 1998	Tekli doküman özetleme	Haber metinleri	•Farklı dilleri içeren bir sistemdir. •Doğal dil işleme teknikleri ile sembolik kavram (sembolic-concept) seviyesi bilgisi birleştirilmiştir. •Özetleme aşamasında konu belirleme, yorumlama ve üretim olmak üzere üç aşama kullanılmıştır.	Cümle seçimine dayalı özet Ve Yoruma dayalı özet
Branimir ve Christopher[30], 1997	Tekli doküman özetleme	Alandan bağımsız metinler	•Dokümanın içeriğinin belirlenmesi için söz öbekleri tespiti yapılmıştır. •Söz öbekleri tespiti ile dokümandaki içerik belirleme işlemleri gerçekleştirilmiş ve özetler içerik belirleyen söz öbeklerine göre çıkarılmıştır.	Özetler önemli söz öbeklerini içermektedir.

Referans-Yıl	Giriş Elemanı	Alan	Kullanılan Özellikler	Çıktı
Baldwin vd. [29], 1998	Kullanıcıya yönelik sorgu tabanlı tekli doküman özetleme	Alandan bağımsız metinler	•Özet çıkarma işlemi kullanıcı tarafından yazılan sorgudaki söz öbeklerinin tamamını kapsayan özetlerin çıkarımıyla gerçekleştirilmiştir. •Sorguda bulunan söz öbeklerini referans eden kısaltmalar ile söz öbekleri eşleştirilmiştir. •Bu işlem gerçekleştirilirken cümlelere içerdikleri söz öbeklerine göre 1-7 arasında bir skor değeri verilmiştir. Özetleme işlemi için cümleler skor değerlerine göre seçilmişlerdir.	Cümle seçimine dayalı özet
Azzam vd. [28], 1999	Tekli doküman özetleme	Alandan bağımsız metinler	•Özetlenecek metinler öncelikle dil bilimsel işlemlerden geçmiş, daha sonra metinler ayrıştırılmış ve önemli referans zincirleri tespit edilmiştir. •Özetler güçlü referans zincirlerine sahip olan cümlelerin seçilmesiyle oluşturulmuştur.	Cümle seçimine dayalı özet
Kan ve McKeow [34], 1999	Tekli doküman özetleme	Alandan bağımsız	•Bilgi çıkarımı (information extraction) ve cümle çıkarımı tekniklerini birleştirmiştir. •Dokümanın konusu, çalışmada dokümandaki konu foci olarak isimlendirilmiştir, varlik ismi tanıma(name entity recognation) ve çoklu kelimelerin (multiwords) kullanılmasıyla belirlenmiştir.	Cümle seçimine dayalı özet
Jin ve McKeow [37], 2001	Tekli doküman özetleme	Alandan bağımsız	•İpucu söz öbekleri, tf x idf skor, cümle konum bilgisi, sözlüksel tutarlılık (lexical coherence) özelliklerini kullanarak cümleleri tespit etmişlerdir.	Yoruma dayalı olan özet

Referans-Yıl	Giriş Elemanı	Alan	Kullanılan Özellikler	Çıktı
Radev vd. [9], 2001	Çoklu doküman özetleme	Haber metinleri	•Merkezcil skorlama (centroid score), konum bilgisi, ve ilk cümle ile olan ortak kelime sayısına göre cümle seçim algoritması önermişlerdir. •Birbirlerine çok benzeyen cümleleri dikkate almamışlardır. •Dokümanlar arası yapısal ilişkileri (Cross-document structural relationships) kullanmışlardır.	Cümle seçimine dayalı özet
Copeck vd. [20], 2002.	Tekli doküman özetleme	Biyografi Özeti	•Makine öğrenmesi teknikleri kullanılmıştır. •Anahtar kelimeler çıkarılmış ve sıralanmıştır. •Doküman aynı konu ile ilgili cümleleri içerecek şekilde bölümlere ayrılmıştır.	Cümle seçimine dayalı özet
Gong ve Lui[40], 2002	Tekli doküman özetleme	Alana bağımlı haber metinleri	•Gizli anlamsal indeksleme tabanlı cümle seçim metodu önermişlerdir. •Ayrıca kelime sıklığı bilgisi kullanılarak incelenen cümle ve tüm dokümanın benzerliğine bakılmış, dokümana en çok benzeyen cümleler özete dahil edilmiştir.	Cümle seçimine dayalı özet
Hirao vd. [23], 2002	Tekli doküman özetleme	Bilinmiyor	•Destek vektör makinelerini kullanarak cümleleri özet ile ilgili ya da ilgili olmayan etiketlerine göre sınıflamıştır. •Sınıflama işleminde cümle konumu, uzunluğu, ağırlığı, başlığa olan benzerliği, önemli fiil ve edatları içerme durumu gibi özellikler kullanılmıştır.	Cümle seçimine dayalı özet

Referans-Yıl	Giriş Elemanı	Alan	Kullanılan Özellikler	Çıktı
Karamuftuoglu [32], 2002	Tekli doküman özetleme	Bilinmiyor	•Çıkarım-azaltım-organization paradigmasına dayalıdır. •Sözlüksel bağlantılar ve bağlar (lexical links and bonds) metodunu metodunu kullanmışlardır. •Herhangi iki cümle arasında ortak geçen kelime varsa bu kelimler sözlüksel bağlantı adını almaktadır. Eğer birden fazla sözlüksel bağlantı varsa o zaman cümleler arasında sözlüksel bağ vardır. •Cümle seçimi destek vektör makineleri kullanılarak gerçekleştirilmiştir.	Cümle seçimine dayalı özet
Lal ve Reuger [24], 2002	Tekli doküman özetleme	Bilinmiyor	•GATE[1] çerçevesini (framework) kullanmışlardır. •Cümle çıkarımı için Bayes sınıflandırıcısını kullanmışlardır. •GATE ANNIE[2] modülünü kullanarak art gönderim çözümlemesi yapmışlardır.	Cümle seçimine dayalı özet
Pardo vd. [5], 2003	Tekli doküman özetleme	Alandan bağımsız	•Kaynak dokümanın en önemli parçalarına öz (gist) adını vermişlerdir. •İstatistiksel metotlar kullanılarak anahtar kelimeler tespit edilmiş ve öz cümleler çıkarılmıştır.	Cümle seçimine dayalı özet
D'Avanzo [35], 2004	Tekli doküman özetleme	Alana bağımlı haber metinleri	•Anahtar kelime çıkarımı •Eğiticili öğrenmeye dayalıdır •Varlik ismi tanıma (name entity recognation) ve çoklu kelimelerin (multiwords) tespiti yapılmıştır.	Çok kısa cümle seçimine dayalı özet

[1] General architecture for Text Engineering, University of Sheffield

[2] A Nearly New Information Extraction System

Referans-Yıl	Giriş Elemanı	Alan	Kullanılan Özellikler	Çıktı
Silla vd. [14], 2004	Tekli doküman özetleme	Alana bağımlı haber metinleri	•Cümleleri özete eklenmek üzere önemli ve önemsiz şeklinde etiketlemeye çalışmışlardır. •Makine öğrenmesi tekniklerini kullanmışlardır. •Cümle önemini belirten 7 adet özellik kullanmışlardır. •Eğitim için Naive Bayes ve C4.5 karar ağaçlarını kullanmışlardır.	Cümle seçimine dayalı özet
Filatova ve Hatzivassiloglou [17], 2004	Tekli doküman özetleme ve sorgu tabanlı özetleme	Alandan bağımsız metinler	•Verileri iki boyutlu bir sistemle göstermişlerdir. Bu sistemde birinci boyut metin parçalarını ikinci boyut ise metindeki konseptleri belirtmektedir. •İçerdikleri bilgi çakışması en az olan metin parçalarının seçildiği bir özetleme sistemi tasarımı yapılmıştır.	Metin birimi seçimine dayalı özet
Filatova ve Hatzivassiloglou [39], 2004	Çoklu Doküman özetleme	Haber metinleri	•Olay tabanlı özetleme sistemi tasarımını gerçekleştirmişlerdir. •Özetleri doküman içinde tanımlamış oldukları atomik olaylar arasındaki ilişkileri belirleyerek çıkartmışlardır. •Greedy algoritması kullanımıştır.	Metin birimi seçimine dayalı özet
Altan [46], 2004	Tekli doküman özetleme		•Ekonomi alanına ait olan 50 doküman ile çalışılmıştır. •Metinler HTML etiketleri yardımıyla başlık, cümle ve paragraflara ayrılmıştır. •Terim sıklığı bilgisi ve cümlenin konum bilgisi özellikleri belirlenmiştir. • Başlık terimleri incelenmiş, pozitif ve negatif cümle analizleri gerçekleştirilmiştir.	Cümle seçimine dayalı özet

Referans-Yıl	Giriş Elemanı	Alan	Kullanılan Özellikler	Çıktı
Steinberger vd. [26], 2005	Tekli doküman özetleme	Alana bağımlı haber metinleri	•Gizli anlamsal indeksleme temelli özetleme sistemlerine ek bilgi olarak art gönderim çözümlemesi konulmuştur.	Cümle seçimine dayalı özet
Yeh vd. [6], 2005	Tekli doküman özetleme	Siyasi makaleler	•Cümle konumu, pozitif ve negatif anahtar sözcükler, merkeziyet, başlığa benzerlik özelliklerinin genetik algoritmalar ile birleşimi kullanılmıştır. •İkinci bir öneri olarak Gizli anlamsal indeksleme tabanlı bir yaklaşıma cümleler bir çizge üzerine oturtulmuş, cümlelerin benzedikleri cümle sayılarına göre özet metinler çıkarılmıştır.	Cümle seçimine dayalı özet
Murray vd. [41], 2005	Tekli doküman özetleme	İş görüşmeleri verisi	•Gizli anlamsal indeksleme tabanlı olan, önemli konulara ait cümlelerin belirlenmesini sağlayan yeni bir öneri getirmişlerdir.	Cümle seçimine dayalı özet
Kiani ve Akbarzadeh [10], 2006	Tekli doküman özetleme	Alana bağımlı haber metinleri	•Genetik algoritma ve genetik programlamayı birleştirerek bulanık kümelerin optimize edilmesini sağlamışlardır. •Sonuçlarını Microsoft ve Copernic özetleme sistemi ile kıyaslamışlardır.	Cümle seçimine dayalı özet
Witte vd. [15], 2007	Tekli ve Çoklu doküman özetleme	Alana bağımlı haber metinleri	•Sezgisel tabanlı bir sistem kullanılmıştır. •Bulanık Uyum Kümeleme Çizge (Fuzzy Coference Cluster Graph) yapısı kullanılarak farklı çeşitteki özetler yaratılmıştır.	Cümle seçimine dayalı özet
Svore vd. [21], 2007	Tekli doküman özetleme	Alanı belirli makaleler	•Cümle özelliklerini birleştirmek için yapay sinir ağlarını kullanmışlardır. •Dokümanı en iyi ifade eden 3 cümle seçilmiş ve özete eklenmiştir.	Cümle seçimine dayalı özet

Referans-Yıl	Giriş Elemanı	Alan	Kullanılan Özellikler	Çıktı
Liu vd. [38], 2007	Tekli doküman özetleme	Alanı belirli makaleler	•Olay tabanlı özetleme (event-based summarization) sistemini önermişlerdir. •Olay-terim grafiği çıkarılmış ve grafikte ilgili terimler kümeler altında gruplanmıştır. •Önerilen sistemin klasik sayfa sıralama (PageRank) tekniklerinden daha iyi performans gösterdiği belirtilmiştir.	Cümle seçimine dayalı özet
Bhandari vd. [43], 2007	Tekli doküman özetleme	Haber metinleri	•Olasılıksal gizli anlamsal indeksleme metodunu kullanarak tekli doküman özetleme sitemi geliştirmişlerdir. •Geliştirilen sistem 4 farklı cümle seçim kriterine sahiptir.	Cümle seçimine dayalı özet
Orasan [27], 2007	Tekli doküman özetleme	Akademik makaleler	•Art gönderim çözümlemesinin metin özetleme üzerindeki etkilerini incelemiştir. •Daha önce önerilmiş olan 3 farklı art gönderim çözümlemesinden bahsedilmiş, kelime tabanlı özetleme yaklaşımlarına art gönderim analizi çalışmasını ekleyerek yeni yaklaşımlarının özetleme performansını arttırdıklarını belirtmişlerdir.	Cümle seçimine dayalı özet
Steinberger [42], 2007	Tekli Doküman Özetleme	Haber metinleri	•Önerilen cümle seçimi cümlelerin doküman içerisinde bahsi geçen tüm konu başlıkları ile ilgili olan cümleleri seçme prensibine dayalıdır. Yeni cümle seçimi gizli anlamsal analiz yöntemine dayalı olan diğer çalışma başarımlarını geçmiştir.	Cümle seçimine dayalı özet

Referans-Yıl	Giriş Elemanı	Alan	Kullanılan Özellikler	Çıktı
McDonald [18], 2007	Çoklu doküman özetleme		• DUC konferansına ait 2005 yılı veri setlerini kullanarak çoklu metin özetleme sistemi önermişlerdir. •Çalışmalarında hem genel hem de kullanıcı sorgularına cevap veren özetler çıkarılmıştır. • İlk etapta greedy yöntemini, daha sonra dinamik programlamayı ve son olarak tam sayılı lineer programlamayı metin özetleme sistemlerine uyarlamışlardır.	Cümle seçimine dayalı özet
Wong vd. [22], 2008	Tekli doküman özetleme	Alanı belirli makaleler	•Eğiticili ve yarı-eğiticili teknikleri kullanarak toplam dört ana başlık altında birleştirilmiş farklı özellikleri kullanan bir sistem tasarlamışlardır. •Ana özellikler: yapısal (surface), içerik (concent), olay (event), ilgi (relevance) tabanlı özelliklerdir. •Eğitim aşamasında olasılık tabanlı karar destek makinelerini kullanmışlardır.	Cümle seçimine dayalı özet
Kılcı ve Diri [47], 2008	Tekli doküman özetleme	Haber metinleri	•Paragraf, cümle ve terimlerin yapısal özellikleri analiz edilmiştir. •Çalışmada kullanılan yapısal özellikler: “Anahtar söz öbekleri”, “terim sıklığı”, “cümle konumu”, “başlık kelimeleri”, “pozitif ve negatif ipucu terimleri”, “bazı noktalama işaretlerinin varlığı”, “gün-ay isimleri”, “nümerik karakter varlığı”, “özel isim varlığı” özellikleridir. Özellikler kullanılarak cümlelere bir skor değeri atanmıştır.	Cümle seçimine dayalı özet

Referans-Yıl	Giriş Elemanı	Alan	Kullanılan Özellikler	Çıktı
Cığır vd. [48], 2009	Tekli doküman özetleme	Haber metinleri	•Yapısal özellikleri kullanan ve cümle seçimi için yapısal özelliklerin birleşimiyle her cümleye bir skor değeri veren bir sistem önermişlerdir. •Skor fonksiyonu, "kelime sıklığı", "başlığa benzerlik", "anahtar söz öbekleri", "merkezilik", "cümle konumu" özelliklerini kullanmaktadır. •Cümlelerin skorlanması bu özelliklerin 0-1 aralığında değerler almasıyla gerçekleştirilir. •Özet çıkarımı en yüksek puana sahip olan cümlelerin seçilmesiyle gerçekleştirilmiştir.	Cümle seçimine dayalı özet
Hernandez ve Ledeneva[7], 2009	Tekli doküman özetleme	Haber metinleri	•Terim ağırlığı, terim sıklığı, ters doküman sıklığı özellikleri kullanılmıştır. •Belirtilen özellikler ile K-ortalamalar yöntemi kullanılmış ve cümleler benzer gruplar altında toplanmıştır. •Gruplar oluşturulduktan sonra grup içindeki en iyi cümle belirlenmiştir.	Cümle seçimine dayalı özet
Lee vd. [44], 2009	Tekli doküman özetleme	Haber metinleri	•Negatif olmayan matris ayrışım metodunu metin özetlemeye uyarlayan ilk çalışmadır. •Yeni bir cümle seçim kriteri önermişlerdir. •Performans sonuçlarını gizli anlamsal analiz algoritması ile kıyaslamışlar ve kendi performanslarının daha yüksek olduğunu belirtmişlerdir.	Cümle seçimine dayalı özet

Referans-Yıl	Giriş Elemanı	Alan	Kullanılan Özellikler	Çıktı
Suanmali vd. [11], 2009	Tekli doküman özetleme	Haber metinleri	•Öncelikle özetlenecek metinler ön işlem aşamalarından geçirilmiştir. •Daha sonra cümlelerin önemini belirten 8 özellik bulanık mantık kurallarına göre birleştirilmiş ve her cümleye bir skor değeri verilmiştir. •Sonuçlar Microsoft ve DUC 2003 konferansından temel aldıkları bir çalışma ile kıyaslanmış ve önerdikleri sistemin daha iyi bir performans gösterdiğini belirtmişlerdir.	Cümle seçimine dayalı özet
Kyoomarsi vd. [12], 2009	Tekli doküman özetleme	TOEFL metinleri	•Metinler ön işlem aşamalarından geçirilmiştir. •Cümle önemini belirten 12 adet özellik kullanılmıştır. •Özelliklerden ikili (binary) olanlara dokunulmamış, sürekli (continues) olanlar ayrıklaştırılmıştır. •Veri seti eğitim ve test kümelerine ayrılarak belirlenen özellikler Naive bayes ve C4.5 karar ağacına göre değerlendirilmiştir. •Bu işlemlerin dışında özelliklerin değerleri bulanık mantık kullanılarak tekrar değerlendirilmiştir.	Cümle seçimine dayalı özet
Özsoy vd. [49], 2010	Tekli doküman özetleme	Haber metinleri	•Gizli anlamsal analiz tabanlı iki yeni yaklaşım önerilmiştir: “Çapraz (Cross)” ve “Konu (Topic)”. Bu önerilerden “Çapraz” metodunun daha iyi bir sonuç verdiğini belirtmişlerdir.	Cümle seçimine dayalı özet

Referans-Yıl	Giriş Elemanı	Alan	Kullanılan Özellikler	Çıktı
Quyang vd. [8], 2010	Tekli ve Çoklu doküman özetleme	Haber metinleri	•Kelime pozisyon bilgisini kullanarak yeni bir özetleme sistemi önermişlerdir. •Çalışmalarını farklı özetleme alanlarında kullanarak kelime pozisyon bilgisinin cümle pozisyon bilgisinden daha iyi sonuç verdiğini göstermişlerdir.	Cümle seçimine dayalı özet
Binwahlan vd. [13], 2010	Tekli doküman özetleme	Haber metinleri	•Benzer cümleleri bulup, bunlar arasından en bilgi içerici olan cümleleri belirleyen bir melez sistem tasarımı yapmışlardır. •Özellik birleşiminde bulanık mantık ve sürü tabanlı (swarm-based) bir yaklaşım kullanılmıştır.	Cümle seçimine dayalı özet
Berker ve Güngör [16], 2010	Tekli doküman özetleme	Haber metinleri	•11 adet cümle önemini belirten özellik kullanmışlardır. Bu özellikler içinde sözlüksel bağlantılar da mevcuttur. •Özellikleri birleştirmek için genetik algoritmaları tercih etmişlerdir. •Özelliklerin birleşimi ile elde edilen sistem performansının, özelliklerin ayrı ayrı ele alınması ile elde edilen performanslardan daha iyi neticeler verdiği sonucuna ulaşmışlardır.	Cümle seçimine dayalı özet
Pembe [51], 2011	Tekli doküman özetleme	Web metinleri	•Doktora tezinde web aramaları için sorgu tabanlı ve yapısal özelliklere dayalı olan bir özetleme sistemi önermişlerdir. •Tezde, önerilen sistemin Google özet çıkarımından daha iyi performans gösterdiği belirtilmiştir.	Cümle seçimine dayalı özet

Referans-Yıl	Giriş Elemanı	Alan	Kullanılan Özellikler	Çıktı
Güran vd. [50], 2011	Tekli doküman özetleme	Haber metinleri	•Negatif olmayan matris ayrışımı tekniğini hazırlamış oldukları 100 Türkçe haber veri seti üzerinde uygulamışlardır. Çalışmalarında sistem performansını arttıran yeni bir ön işlem aşaması önermişlerdir.	Cümle seçimine dayalı özet
Alguliev vd. [19], 2011	Çoklu doküman özetleme	Haber metinleri	•Metin özetleme problemi bir tamsayılı lineer programlama problemi olarak düşünülmüştür. •Cümleye ait üç özellik optimize edilmeye çalışılmıştır: ilgililik (relevance), gereksizlik (redundancy) ve uzunluk (length).	Cümle seçimine dayalı özet
Mashechkin vd.[45], 2011	Tekli doküman özetleme	Haber metinleri	• DUC konferansı için hazırlanan 2001 ve 2002 veri setleri kullanaılarak negatif olmayan matris ayrışımına dayalı olan yeni bir cümle seçimi tekniği önerisi sunmuşlardır. •Sunulan öneri tekil değer ayrışımı yapısına benzetilmeye çalışılmıştır. •Önerilen sistem daha önceki negatif olmayan matris ayrışım tekniklerinin başarım oranlarından daha yüksek başarım sonuçlarına sahiptir.	Cümle seçimine dayalı özet

EK-B

VERİ SETLERİNE AİT ÖRNEK DOKÜMANLAR

VeriSeti-1'e ait olan bir doküman örneği

Bebekler duygusal nedenlerle ağlar mı?

Yeni anne babaları en zorlayıcı bebek davranışı ağlamadır. Anne babalar bebekleri ağladığında fazlasıyla endişelenip büyük bir panik yaşarlar.

Bebekleri ağladığında anne babalarda oluşan paniğin ilk nedeni anne babaların, bebeğin ne istediğini anlayıp sorunu çözebilecekleri konusunda kendilerine yeterince güvenmemeleri, diğer nedeni ise bebeğin ağlamasından "kızdı, üzüldü, korktu" gibi duygusal anlamlar çıkarmalarıdır. Oysaki bebekler neredeyse 6 aylık olana kadar genel bir memnuniyet ve genel bir sıkıntı hali dışında bir duygu yaşamazlar.Yani bebekler derin duygusal nedenlerden ötürü ağlamazlar. Bebekler açlık, altının ıslanması, yorgunluk, sindirim sistemindeki sorunlar gibi fiziksel nedenlerle ya da sizi yanında istediği veya görsel, işitsel, dokunsal uyaran aradığı için ağlarlar. Erken dönemdeki bu ağlamaların hiç birinde "kızmak, korkmak, üzülmek, acı çekmek" gibi duygusal anlamlar yoktur.

Ağlama bir bebeğin ebeveynle iletişim kurma biçimidir. Henüz konuşarak kendini anlatma becerisine sahip olmadığı için istek ve ihtiyaçlarını size ağlayarak anlatmaktadır. Çoğu anne baba böyle bir durum karşısında kendisini üzgün, bir şeyler yapmak zorunda ama yapamadığı için yetersiz, çaresiz ve kötü hisseder. Hele bir de ağlayan kendi bebeği olduğunda durum iyice zorlaşır.

İşte bu noktada anne babaların bir bebeğin ağlaması ile daha büyük birinin ağlaması arasındaki farkı anlaması çok önemlidir. Bir yetişkin ağladığında bunun sebebi kendisini

suçlu, utanmış ya da üzgün hissetmesi olabilir. Oysa ki bebeklerde bu türden bir değerlendirme yapmayı sağlayacak, beynin sofistike mekanizmaları henüz gelişmemiştir.O yüzden bebeğinizin ağlıyor olmasını onun üzgün, kızgın, ya da korkmuş hissetmesine bağlamak mümkün değildir. Bebekler istek ve ihtiyaçlarını anlatmak için ağlamaya mecburdur. Açlık, tuvaletini yapma ya da sosyalleşme ihtiyacı gibi çok temel, basit şeyleri anlatma yöntemidir. Bebeklerin ağlamasına duygusal atıflarda bulunma eğilimi bebeğin ağlamasına doğru şekilde cevap verme becerisini büyük oranda düşürür. Bebeğin ağlamasını duyduğunuzda sorunun ne olduğuna odaklanmak yerine "zavallıcık ne kadar acı çekiyor, çok üzüldü, ya da korktu" diye anlamlar çıkartmaya başladığınızda ağlama nedenini bulamaz ve gerekli sakinleştirici adımları atamazsınız.

Bebek ağladığında anne babaların hissettiği gerginlik ve çaresizlik hissi çoğu zaman bebeğin daha fazla ağlamasına ya da ağlamasının daha uzun sürmesine neden olur.

Bebekler anne babaların davranış tonuna çok duyarlıdırlar. Mesela, anne kendini stresli hissettiği zamanlarda bebeğini kucağına alıp emzirmeye ya da uyutmaya çalıştığında bebeğinden negatif bir tepki alma ihtimali çok yüksektir.Anne gerginken bebeğin iyi bir şekilde emip huzurlu bir şekilde uykuya dalması pek mümkün olmaz. O nedenle bebeğinizle ilgilenirken stres düzeyini düşürmelisiniz. Bunun ilk adımı sizi gereğinden fazla endişeye sokan bebeğin ağlamasına duygusal anlamlar yükleme eğiliminden kurtulmaktır. Ağlamanın altında yatan nedeni bulma konusunda kendinize olan güveni korursanız sakinliğinizi de koruyabilirsiniz. Ağlayan bebeğe müdahale eden kişinin sakin olabilmesi bebeği sakinleştirmede en önemli unsurdur. Kişisel özellikleriniz ve bebek bakımı konusundaki bilgi düzeyiniz ne olursa olsun bebeğinizin istek ve ihtiyaçlarını en iyi şekilde karşılayabilecek kişi sizsiniz.

Bebeğiniz ağladığında sakin bir şekilde onu rahatlatacak adımları uyguluyorsanız, susup susmamasının aslında önemi yoktur. Bebeğiniz ağladığında önemli olan sizin özgüveninizi, sakinliğinizi koruyarak güven hissinin bebeğinize de geçmesini sağlamanız ve bebeğinizin yanında olmanızdır.

VeriSeti-1'e ait olan bir özet dokümanı örneği

Bebekleri ağladığında anne babalarda oluşan paniğin ilk nedeni anne babaların, bebeğin ne istediğini anlayıp sorunu çözebilecekleri konusunda kendilerine yeterince güvenmemeleri, diğer nedeni ise bebeğin ağlamasından "kızdı, üzüldü, korktu" gibi duygusal anlamlar çıkarmalarıdır.

Bebekler açlık, altının ıslanması, yorgunluk, sindirim sistemindeki sorunlar gibi fiziksel nedenlerle ya da sizi yanında istediği veya görsel, işitsel, dokunsal uyaran aradığı için ağlarlar.

Erken dönemdeki bu ağlamaların hiç birinde "kızmak, korkmak, üzülmek, acı çekmek" gibi duygusal anlamlar yoktur.

Bebeğin ağlamasını duyduğunuzda sorunun ne olduğuna odaklanmak yerine "zavallıcık ne kadar acı çekiyor, çok üzüldü, ya da korktu" diye anlamlar çıkartmaya başladığınızda ağlama nedenini bulamaz ve gerekli sakinleştirici adımları atamazsınız.

Bebek ağladığında anne babaların hissettiği gerginlik ve çaresizlik hissi çoğu zaman bebeğin daha fazla ağlamasına ya da ağlamasının daha uzun sürmesine neden olur.

O nedenle bebeğinizle ilgilenirken stres düzeyini düşürmelisiniz.

Ağlayan bebeğe müdahale eden kişinin sakin olabilmesi bebeği sakinleştirmede en önemli unsurdur.

Bebeğiniz ağladığında önemli olan sizin özgüveninizi, sakinliğinizi koruyarak güven hissinin bebeğinize de geçmesini sağlamanız ve bebeğinizin yanında olmanızdır.

VeriSeti-2'ye ait olan bir doküman örneği

Tarkan'ın acı günü

Türk Pop Müziği sanatçısı Tarkan Tevetoğlu, annesi Neşe Tevetoğlu'nun yaklaşık 50 yıllık arkadaşı alzheimer hastası 76 yaşındaki Rukiye Güler'in durumunun ağırlaşması üzerine geçen Salı, annesi, kardeşleri Handan ve Hakan Tevetoğlu ile birlikte Güler'in Ortakent-Yahşi Beldesi'ndeki evine geldi.

Fenalaşıp dün hayatını kaybeden Güler için bugün Ortakent Merkez Camii'nde cenaze töreni düzenlendi. Güler yakınlarının gözyaşları arasında öğlen kılınan cenaze namazının ardından, Ortakent Mezarlığı'na götürüldü. Camideki törene katılmayan Süper Star Tarkan, mezarlıkta Güler için dua etti. Güler, toprağa verilirken duygu dolu anlar yaşayan ünlü sanatçı, "Rukiye teyze soyadı gibi hep gülerdi, çok neşeliydi, bizi çocuklarından ayırmazdı. Çok yakın aile dostumuz olduğu için ağırlaştığını duyunca buraya geldik" dedi. Tarkan'ın annesi Neşe Tevetoğlu'nun cenazede ayakta durmakta güçlük çektiği ve gözyaşı döktüğü görüldü. Rukiye Güler'in 30 yıllık arkadaşı 60 yaşındaki Alime Kurt, "Rukiye hanımın 20 gündür çocuklarıyla birlikte yanındayız, durumu çok ağırdı, Tarkan'ı ve annesi Neşe'yi çok severdi. Tarkan ve ailesi Rukiye hanım ölmeden önce buraya gelip iki gün dizinin dibinde oturdular" diye konuştu.

Tarkan ve ailesinin akşam saatlerinde Bodrum'dan ayrılacağı öğrenildi.

VeriSeti-2'ye ait olan bir özet dokümanı örneği

Türk Pop Müziği sanatçısı Tarkan Tevetoğlu, annesi Neşe Tevetoğlu'nun yaklaşık 50 yıllık arkadaşı alzheimer hastası 76 yaşındaki Rukiye Güler'in durumunun ağırlaşması üzerine geçen Salı, annesi, kardeşleri Handan ve Hakan Tevetoğlu ile birlikte Güler'in Ortakent-Yahşi Beldesi'ndeki evine geldi. Güler, toprağa verilirken duygu dolu anlar yaşayan ünlü sanatçı, "Rukiye teyze soyadı gibi hep gülerdi, çok neşeliydi, bizi çocuklarından ayırmazdı. Çok yakın aile dostumuz olduğu için ağırlaştığını duyunca buraya geldik" dedi. Rukiye Güler'in 30 yıllık arkadaşı 60 yaşındaki Alime Kurt, "Rukiye hanımın 20 gündür çocuklarıyla birlikte yanındayız, durumu çok ağırdı, Tarkan'ı ve annesi Neşe'yi çok severdi. Tarkan ve ailesi Rukiye hanım ölmeden önce buraya gelip iki gün dizinin dibinde oturdular" diye konuştu.

VeriSeti-3'e ait olan bir doküman örneği

CRICKET-INDIA SET FOR VICTORY AFTER WEST INDIES CRASH .

India were poised to claim only their third win in 31 tests in the Caribbean after their seam bowlers routed the West Indies batsmen during the third day of the third test on Sunday .

The West Indies were bowled out for a meagre 140 in their second innings , their lowest score ever against India , to leave the tourists needing 120 to win in their second innings with two full days left . At the close , India , were two without loss .

Navjot Sidhu was not out two and Vangipurrapu Laxman was still to get off the mark .

Abey Kuruvilla , a tall , gangling fast bowler in only his third test , was India's hero with career-best figures of five for 68 off 21 overs. Kuruvilla was given admirable support from Venkatesh Prasad , whose three for 39 off 18 overs gave him match figures of eight for 121 , the best match haul by an Indian in six tests at the Kensington Oval .

Dodda Ganesh chipped in with two for 28 off six overs as only stand-in West Indies captain Brian Lara , with 45 , made more than 25 runs .

The West Indies' total , eleven runs shortof the 151 they managed at Madras during the 1978-79 series , would have been even lower butfor a last wicket stand of 33 runs between fast bowlers Mervyn Dillon (21) and Curtly Ambrose (18 not out) . Earlier in the day the West Indies had done well in reducing India from 249 for three overnight to 319 all out , capturing the last four wickets for 29 runs and restricting India's first innings lead to 21 runs . India lost the wickets of Saurav Ganguly (22) , Rahul Dravid (78) and Nayan Mongia (one) in the first session for the addition of only 41 runs as the West Indies bowlers dominated . Dravid , who was bowled by fast bowler Ian Bishop , batted for six hours and 12 minutes , faced 243 balls and struck eight fours .

In the third over after lunch , former captain Mohammad Azharuddin , who struggled for two hours in scoring 17 , was caught by wicketkeeper Courtney Browne off fast bowler Franklyn Rose .

In the next over , Curtly Ambrose bowled Kuruvilla for nought to leave India on 296 for eight . Rose wrappedup the innings to end with 4-77 as the home team appeared to be back in with a shout .

In fact , the West Indies' problems were just begining . They quickly lost opener Stuart Williams (0) and first innings century-maker Shivnarine Chanderpaul (3) in reaching on 31 for two at tea .

After the break , the wickets of Sherwin Campbell (18) , Carl Hooper (4) , Lara , Courtney Browne (1) , Roland Holder (13) and Bishop (6) tumbled in quick succession as the West Indies collapsed to 95 for eight . Prasad captured the key wicket of Lara , who after facing 67 balls in 104 minutes , edged an outswinger to Azharuddin at second slip .

VeriSeti-3'e ait olan bir özet dokümanı örneği

India were poised to claim only their third win in 31 tests in the Caribbean after their seam bowlers routed the West Indies batsmen during the third day of the third test on Sunday .

The West Indies were bowled out for a meagre 140 in their second innings , their lowest score ever against India , to leave the tourists needing 120 to win in their second innings with two full days left .

At the close , India , were two without loss .

The West Indies' total , eleven runs shortof the 151 they managed at Madras during the 1978-79 series , would have been even lower butfor a last wicket stand of 33 runs between fast bowlers Mervyn Dillon (21) and Curtly Ambrose (18 not out) .

Earlier in the day the West Indies had done well in reducing India from 249 for three overnight to 319 all out , capturing the last four wickets for 29 runs and restricting India's first innings lead to 21 runs .

VeriSeti-4'e ait olan bir doküman örneği

hurricane gilbert heads toward dominican coast...

hurricane gilbert swept toward the dominican republic sunday, and the civil defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. the storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.

``there is no need for alarm,'' civil defense director eugenio cabral said in a television alert shortly before midnight saturday. cabral said residents of the province of barahona should closely follow gilbert's movement. an estimated 100,000 people live in the province, including 70,000 in the city of barahona, about 125 miles west of santo domingo.

tropical storm gilbert formed in the eastern caribbean and strengthened into a hurricane saturday night. the national hurricane center in miami reported its position at 2 a.m. sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of ponce, puerto rico, and 200 miles southeast of santo domingo. the national weather service in san juan, puerto rico, said gilbert was moving westward at 15 mph with a ``broad area of cloudiness and heavy weather'' rotating around the center of the storm.

the weather service issued a flash flood watch for puerto rico and the virgin islands until at least 6 p.m. sunday. strong winds associated with the gilbert brought coastal flooding, strong southeast winds and up to 12 feet feet to puerto rico's south coast. there were no reports of casualties. san juan, on the north coast, had heavy rains and gusts saturday, but they subsided during the night.

on saturday, hurricane florence was downgraded to a tropical storm and its remnants pushed inland from the u.s. gulf coast. residents returned home, happy to find little damage from 80 mph winds and sheets of rain. florence, the sixth named storm of the 1988 atlantic storm season, was the second hurricane. the first, debby, reached minimal hurricane strength briefly before hitting the mexican coast last month.

VeriSeti-4'e ait olan bir özet dokümanı örneği

tropical storm gilbert in the eastern caribbean strengthened into a hurricane saturday night.

the national hurricane center in miami reported its position at 2 a.m. sunday to be about 140 miles south of puerto rico and 200 miles southeast of santo domingo.

ıt is moving westward at 15mph with a broad area of cloudiness and heavy weather with sustained winds of 75mph gusting to 92mph.

the dominican republic's civil defense alerted that country's heavily populated south coast and the national weather service in san juan, puerto rico issued a flood watch for puerto rico and the virgin ıslands until at least 6 p.m. sunday.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Aysun GÜRAN

Doğum Tarihi ve Yeri : 17 Haziran 1983 - Kadiköy

Yabancı Dili : İngilizce

E-posta : aysunguran@gmail.com

ÖĞRENİM DURUMU

Derece	**Alan**	**Okul/Üniversite**	**Mezuniyet Yılı**
Y. Lisans	Matematik Müh.	Yıldız Teknik Üniv.	2007
Lisans	Matematik Müh.	Yıldız Teknik Üniv.	2005
Lise	Fen-Matematik	Kadir Has Anadolu	2001

İŞ TECRÜBESİ

Yıl	**Firma/Kurum**	**Görevi**
2005- …	Doğuş Üniversitesi/Bilgisayar Müh.	Araştırma Görevlisi

YAYINLARI

Makale

1. Güran A., Güler Bayazıt, N, Gürbüz, M.Z., (2012),"Efficient feature integration with Wikipedia based semantic feature extraction for Turkish text summarization", Turkish Journal of Electrical Engineering and Computer Science (SCI-E),(baskıda).

Bildiri

1. Güran, A., Güler, N., Bekar, E., (2011). "LSA-based Turkish Text Summarization with Consecutive Words Detection", CSIE 2011, Changchun, Çin.
2. Güran, A., Güler, N., Bekar, E., (2011). "Automatic summarization of Turkish documents using non-negative matrix factorization ", INISTA 2011, Istanbul, Türkiye.
3. Güran, A., Bekar, E., Akyokuş, S. (2010). "A Comparison of Feature and Semantic-Based Summarization Algorithms for Turkish", INISTA 2010, Kapadokya.
4. Gürbüz, Z., Akyokuş, S., Emiroğlu, İ., Güran, A., (2010). "An efficient Algorithm for 3D Rectangular Box Packing", AAS 2009, Ohrid, Makedonya.
5. Güran, A., Akyokuş, S., Güler, N., Gürbüz, Z., (2009). "Turkish Text Categorization Using N-Gram Words", INISTA 2009, Trabzon.

Kitap İçi Bölüm

1. Güran, A., Güler Bayazıt, N. (2012)., "A New Preprocessing Phase for LSA-Based Turkish Text Summarization", Recent Advances in Computer Science and Information Engineering Lecture Notes in Electrical Engineering, 2012, 124: 305-310.

Proje

1. Tezim Yıldız Teknik Üniversitesi “2012-07-03-DOP01” kodlu BAP projesi tarafından desteklenmiştir.

ÖDÜLLER ve BURSLAR

1. Matematik Mühendisliği Lisans Bölüm Birinciliği (3.69/4.0)

2.Yüksek Lisans ve Doktora Tübitak Bursu