T.C.
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

GENETİK ALGORİTMA KULLANARAK CÜMLE SEÇME YAKLAŞIMI İLE OTOMATİK METİN ÖZETLEME

YÜKSEK LİSANS TEZİ

HAZIRLAYAN : Ercan BAYDAR
DANIŞMAN : Doç. Dr. Rıdvan SARAÇOĞLU

VAN-2018

T.C.
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

GENETİK ALGORİTMA KULLANARAK CÜMLE SEÇME YAKLAŞIMI İLE OTOMATİK METİN ÖZETLEME

YÜKSEK LİSANS TEZİ

HAZIRLAYAN: Ercan BAYDAR

VAN-2018

KABUL VE ONAY SAYFASI

Elektrik-Elektronik Mühendisliği Anabilim Dalı'nda Doç. Dr. Rıdvan SARAÇOĞLU danışmanlığında, Ercan BAYDAR tarafından sunulan "**Genetik Algoritma Kullanarak Cümle Seçme Yaklaşımı ile Otomatik Metin Özetleme**" isimli bu çalışma Lisansüstü Eğitim ve Öğretim Yönetmeliği'nin ilgili hükümleri gereğince .../.../2018 tarihinde aşağıdaki jüri tarafından oy birliği ile başarılı bulunmuş ve Yüksek Lisans tezi olarak kabul edilmiştir.

Başkan : Doç. Dr. Rıdvan SARAÇOĞLU İmza:

Üye : Yrd. Doç. Dr. ? İmza:

Üye : Yrd. Doç. Dr. ? İmza:

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun …./..../….. tarih ve ..... sayılı kararı ile onaylanmıştır.

İmza

……..……..

Enstitü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

Ercan BAYDAR

ÖZET

GENETİK ALGORİTMA KULLANARAK CÜMLE SEÇME YAKLAŞIMI İLE OTOMATİK METİN ÖZETLEME

BAYDAR, Ercan
Yüksek Lisans Tezi, Elektrik-Elektronik Mühendisliği Anabilim Dalı
Tez Danışmanı: Doç. Dr. Rıdvan SARAÇOĞLU
Mart 2018, 45 sayfa

Dijital ortamda bulunan bilgiler çok büyük boyutlara ulaşmış durumdadır. Bu yüzden elde edilmek istenen bilgiye en hızlı ve verimli şekilde ulaşmak zorlaşmıştır. Bu nedenle metinlerin özetinin elde edilmesi ihtiyacı ortaya çıkmıştır.

Bu çalışma, otomatik metin özetleme yaklaşımlarını, çalışmalarını, yazılımlarını, algoritmalarını ve metotlarını ortaya koyar. Otomatik metin özetleme genel olarak iki bölüme ayrılır. Bunlar çıkarımsal özetleme ve yorumlayarak özetlemedir. İlk olarak çıkarımsal özetlemede cümleler ve kelimeler bazı özelliklere göre puanlandırılır ve en yüksek puanlı cümleler seçilir. Kısaca metindeki en önemli cümleleri bulmaktır. İkinci olarak yorumlayarak özet çıkarmada ise cümleler ve kelimeler arasında anlamsal ilişkiler incelenir. İki yöntem arasındaki en önemli fark, yorumlayarak özet çıkarma yönteminde metinden bağımsız yeni kelimeler elde edilebilir.

Bu tez çalışmasında Türkçe metinler için çıkarımsal özetleme yöntemi ile başarının artırımı amaçlanmıştır. Üç farklı yöntem denenmiş ve ölçümler yapılmıştır. Bunlar sabit puanlı değerlendirme, sezgisel olarak belirlenen rastgele puan aralıklı değerlendirme ve genel aralıklı rastgele puanlama ile yapılan değerlendirmedir. Üç yöntem için de çıkarımsal özetleme yönteminde genetik algoritma kullanılmıştır ve cümleler buna göre puanlandırılmıştır. İlk yöntemde cümlelere puan verirken sezgisel yöntem uygulanmıştır. İkinci yöntemde sezgisel yönteme bir puan aralığı ilave edilerek sonuca ulaşılmaya çalışılmıştır. Üçüncü ve son yöntemde ise puan aralığı genişletilerek 1 ile 100 arasında bir tamsayı aralığı belirlenmiş olup cümleler puanlandırılmıştır. Yapılan ölçümlerde başarının arttığı tespit edilmiştir.

Anahtar kelimeler: Genetik Algoritma, Metin Madenciliği, Metin Özetleme

ABSTRACT

SINGLE DOCUMENT EXTRACTIVE AUTOMATIC TEXT SUMMARIZATION USING GENETIC ALGORITHMS

BAYDAR, Ercan
M. Sc. Thesis, Department of Electrical and Electronics Engineering
Supervisor: Assoc. Prof. Dr. Rıdvan SARAÇOĞLU
March 2018, 45 pages

Giant information have been available on the digital environments. For this reason it has hard to get the information what you want to achieve it, fast and most efficiently. For this reason, the need to obtain a summary of the texts has emerged.

This research presents automatic text summarizations' approaches, studies, software, algorithms, and methods. Automatic text summarization is generally divided into two systems. There are extraction based summarizations and abstraction based summarizations. Firstly, extractions summarization approach involves selecting sentences of high score from the document based on word and sentence features. In short, finding the most important sentences from text is performed. Secondly, it is examined the semantic relationships between words and sentences in abstraction approaches. The most important difference from the first approach, made semantic analysis of words and sentences afterward new words are obtained.

In this study, we focus on an extractive text summarization system and we purpose to increase the success about this issue. We have tried 3 different methods and made measurements. We applied genetic algorithm methods and scored the sentences in all three methods. We have tried to get the highest scores. In the first methods we made the scoring of sentences heuristic methods. Secondly, we have determined the range of points in the heuristic method. In the third method, we set the range from 1 to 100. We gave random integer points in this range and have tested the system. We have identified an increasing success in the measurements made in 3 methods.

Key words: Genetic algorithm, Text mining, Text summarization.

ÖN SÖZ

Bu tez çalışmamın tüm aşamasında, her türlü görüş ve yardımlarından dolayı başta danışmanım Sayın Doç. Dr. Rıdvan SARAÇOĞLU'na, hayatım boyunca benden desteklerini esirgemeyen sevgili eşim Canan BAYDAR'a ve varlığıyla ilham kaynağım olan biricik oğlum Yağız BAYDAR'a teşekkür ederim.

2018

Ercan BAYDAR

İÇİNDEKİLER

Sayfa

Sayfa

ÇİZELGELER LİSTESİ

Çizelge **Sayfa**

ŞEKİLLER LİSTESİ

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	**Açıklama**
GHz	GigaHertz
MB	Megabyte
GB	GigaByte

Kısaltmalar	**Açıklama**
DDİ	Doğal Dil İşleme
HTML	Hyper Text Markup Language
C3M	Cover Coefficient-Based Clustering Methodology
ARTEX	AnotheR TEXt Summarizer
FRESA	Framework for Evaluating Summaries Automatically
XML	Extensible Markup Language

1. GİRİŞ

Günümüz koşullarında elektronik ortamlardaki verilerin miktarı hızla artarken diğer taraftan bir veri yığınıyla karşı karşıya kalmış durumdayız. Üretilen bilginin boyutunu fazlalığı konusunda yapılan araştırmalarda karşımıza çıkan rakamlar nasıl bir tablo ile karşı karşıya oluğumuz konusunda ipucu vermektedir. Üretilen bilginin %90'dan fazlası sayısal ortama aktırılmış olup her bir insan için üretilen bilgi yaklaşık 800 MB' dir. WEB ortamında 550 milyar belge vardır. Sadece ABD'de günde 610 milyar e-Posta gönderilirken 2 milyar röntgen filmi çekilmekte ve bu veriler sayısal ortamlarda tutulmaktadır (Yılmaz, 2011). Özetlemek gerekirse dünya üzerinde sayısal ortamda bulunan bilgi miktarı devasa boyutlara ulaşmış durumdadır.

Saklanan verilerin türleri çeşitli olabilmektedir. Sayısal ortamlardaki veriler metin, ses, görüntü gibi türlerde saklanmaktadır. Metin verilerine erişim sağlayan insanlar elde ettikleri metinlerin gereksiz bölümlerine de erişim sağlarlar. İnsanlar elde etmek istedikleri verinin gereksiz verilerden arınmış olmasını isterler. Böylelikle gereksiz verileri okumamış ve zamanlarını boşa harcamamış olacaklardır. Akademik araştırma yapan bir araştırmacının istediği bilgiye erişmek için harcadığı zamana, okumak zorunda olduğu gereksiz bilgileri de eklersek, dünyada metinsel bilgiye erişmek isteyen insanların saf bilgiye ne kadar ihtiyaç duyduğu ortaya çıkacaktır. Başka bir örnek vermek gerekirse sınava hazırlanan bir öğrencinin kısıtlı bir zamanda okumak zorunda olduğu bir kitaba çalışma zamanı ile işine yarayacak özetine çalışacağı zaman aynı olmayacaktır.

Yukarıda anlatılan benzer durumlardan yola çıkarak zamanın verimli kullanılmasının bir zorunluluk haline geldiği sonucuna ulaşılır. Bu sorunsalı çözebilmek için araştırmacılar metin özetleme sistemlerini geliştirmeye çalışmaktadırlar.

Bu tezde metinlerin özetinin çıkarılması ile istenen bilgiye erişimin daha hızlı bir şekilde sağlanması amaçlanmıştır. Böylelikle zaman bakımından verim artmış olacak ve sayısal ortamlardaki gereksiz metin verileri azalmış olacaktır.

1.1. Metin Özetleme

Özet, bir metnin kendisinden uzun olmayan o metin hakkında gerekli bütün bilgileri veren başka bir metindir. Metin özetleme ise kısaca, işleme sokulan bir metnin işlem sonucunda o metne ait en gerekli ve önemli içeriğine sahip olduğu süreçtir. Sisteme bir belge verilip sonucunda o belgenin yarısından az olacak şekilde o belgeyi temsil edebilecek en önemli ve gerekli cümleleri içeren bir belge ortaya çıkacaktır. Buradaki temel konu "önem" veya "gereklilik" kelimeleridir. Bir kelimenin neden gerekli veya önemli olduğuna verilecek karar bizi başarıya ulaştıracak temel unsurdur.

Metnin içinde geçen herhangi bir kelimenin bir metin için önemli olup olmaması konu ile kurulan ilişki ile doğru orantılıdır. İfade edilen konu ile ne kadar ilişkili ise o kadar önemlidir veya kullanıcıya ne kadar yardımcı olacaksa o kadar gereklidir.

1.2. Metin Özetleme Çeşitleri

Metin özetleme sistemleri genel olarak sistemin verdiği sonuca yani çıktısına göre çıkarımsal (extractive) veya yoruma dayalı (abstractive) olmak üzere ikiye ayrılır. Yoruma dayalı özetleme sonucunda ortaya çıkan özet metin, özeti çıkarılacak olan metnin mantıklı bir yorumudur. Bir başka değişle yoruma dayalı özet çıkarma işleminde elde edilecek özet metin, özeti çıkarılmak istenen esas metinden farklı kelimeler içerebilir. Çıkarıma dayalı özetlemede ise özeti çıkarılacak metnin içinden o metni en yüksek derecede temsil ettiği düşünülen cümleler seçilip alt metin halinde sunulur. Bu ayırıma bir örnek vermek gerekirse yoruma dayalı özet çıkarma ile oluşturulan bir sistem için "Ahmet, tribündeki yerini alıp maç saatini heyecanla bekledi" cümlesi, "Ahmet maç izlemeye gitti" şeklinde bir alt metin olabilmektedir. Dikkat edilirse mantıklı bir yorumdan sonra Ahmet'in maç izlemeye gittiği sonucuna ulaşılır ve tamamen yorum yapılmıştır. Çıkarımsal özetlemede ise özeti çıkarılacak metinden birebir cümleler çeşitli yöntemlerle seçilir ve alt metin oluşturulur. Bu tezde çıkarıma dayalı özet çıkarma yöntemi kullanılmıştır.

Özetleme sistemleri sistemin girdisine göre de tekli doküman özetleme ve çoklu doküman özetleme olarak ikiye ayrılır. Girilen metnin kaynağına göre yapılan bu ayrımda tekli doküman özetleme sistemlerinde tek kaynaktan özet çıkarılırken çoklu

doküman özetleme yöntemlerinde birden fazla kaynaktan yararlanarak alt metinler elde edilir. Bu tezde tekli doküman özetleme yöntemi kullanılmıştır.

Bir başka ayrım genel ve sorgu tabanlı özetler arasındadır. Sorgu tabanlı özetlerde özeti çıkarılacak metinler sorguya bağlı içerikle sınırlandırırken genel amaçlı özetler belgenin içeriği hakkında genel bilgi verirler. Sorgu tabanlı özetler, kullanma kılavuzu gibi büyük veya konu çeşitliliği olan belgelerle ilgilidirler.

Metin özetleme sistemleri oluşturulurken metinde geçen cümlelerin özellikleri, kelimelerin özellikleri, dilin doğan yapısı önem teşkil eder. Özeti çıkarılacak metinlerin hangi konudan bahsettiği, cümlelerin ne anlama geldiğinin tespiti için metindeki kelimeler mercek altına alınır. Bunu yapabilmek için de o dilin yapısal ve/veya anlamsal özellikleri göz önünde bulundurulur. Yapısal olarak genelde dilin özellikleri de dikkate alınarak gövdeleme ya da köklerine ayırma işlemi uygulanır. İngilizce metinler için yapılan çalışmalarda çeşitli gövdeleyici (Porter-Stemmer) algoritmaları kullanılır. Türkçe gibi eklerin kelimenin anlamını değiştirdiği dillerde kök bulma zorlaşır. Türkçe metinler için araştırmacılar genelde Zemberek adında kelimeler için yapısal çözümleme yapan açık kaynak kodlu yazılım kullanmışlardır. Bu çalışmada da kelime çözümlemesi yapabilmek için Zemberek yazılımı kullanılmıştır.

1.3. Tezin Amacı

Bu tezde genetik algoritma yardımıyla çıkarımsal olarak Türkçe metinlerin başarılı özetinin çıkarılması amaçlanmıştır. Tek doküman kaynaklı ve çıkarımsal özetleme yöntemi kullanılmış olup sezgisel yöntemlerle elde edilen başarım genetik algoritma ile artırılmıştır. Bu amaç doğrultusunda literatürdeki Türkçe metinler için yapılan özetleme çalışmaları incelenmiş ve ortaya çıkan sonuçlar ile karşılaştırılmıştır.

Özetleme sistemlerinin genel amacı konu ile alakalı mümkün olduğu kadar fazla bilgi elde edip bu bilgiyi özete taşımaktır.

Yöntem olarak daha önce başka çalışmalarda kullanılan Türkçe veri seti kullanılarak literatürde bulunan sezgisel yöntemler, oluşturulan bir bilgisayar programı yardımı ile sonuçlandırılmıştır. Daha sonra bu sezgisel yöntemler geliştirilerek tekrar sonuçlandırılmıştır. Geliştirilen sezgisel yöntemlere genetik algoritma yöntemleri de eklenerek farklı sonuçlar elde edilmiştir. Sonuçlar karşılaştırıldığında geliştirilmiş

sezgisel yöntemin ve genetik algoritma ile güçlendirilen yöntemin literatürdeki sabit puanlı sezgisel yöntemden daha başarılı sonuçlar verdiği gözlemlenmiştir.

2. LİTERATÜR BİLGİSİ

Bu kısımda metin özetleme üzerine yapılan çalışmalar, elde edilen sonuçlar ortaya konulacaktır. Özetleme üzerine yapılan çalışmaların büyük bir çoğunluğu İngilizce dilinde yapılan çalışmalardır. Metin özetleme ile ilgili ilk çalışmalar bundan yaklaşık elli yıl önce İngilizce dili için yapılmıştır. Çeşitli problemlere yönelik birçok yeni yöntem bulunup geliştirilmesine rağmen, bugünkü araştırmalarda ve birçok uygulamada ilk ortaya çıkan çözüm yolları hala kullanılmaktadır. Bu yöntemlerin tercih edilmesinin nedeni verimliliğinin ve başarımın yüksek, maliyetlerin de düşük olmasıdır. Yoruma dayalı veya doğal dil işleme teknikleri gerektiren yöntemler, Türkçenin sondan eklemeli-kurallı yapısı nedeniyle, İngilizce için yıllardır geliştirilen yöntemlerin Türkçe için birebir uyarlanmasını mümkün kılmamaktadır.

Özetleme konusunda yapılan çalışmaları genel olarak 2 ana başlık altında toplayabiliriz. Birincisi cümle seçerek özetleme, ikincisi de yorumlayarak özetleme yöntemidir. Cümle seçerek özetlemede, elde edilen özet için, özeti çıkarılacak metinden çeşitli yöntemlerle seçilen önemli cümlelerin birleştirilmesiyle oluşturulan özet çıkarma sistemleridir. Çıkarıma dayalı özet olarak da adlandırılır. Yorumlayarak özet çıkarma yönteminde ise doğal dil işleme (DDİ) gerektiren yöntemler kullanılarak, sözcükleri derinlemesine inceleyip gövdeleme metotlarının da kullanıldığı bu tarz özetleme sistemlerinde cümle seçiminin aksine yepyeni sözcükler elde edilir. Örneğin “Selim menüyü okudu, yemeğini istedi ve yedi" cümlesi, "Selim karnını doyurdu" şeklinde özetlenebilir. Bu yöntemin kullanılabilmesi için sembolik kelimeler tablosu gereklidir (Uzundere ve ark., 2008).

Cümle seçerek özet çıkarma yönteminin odak noktası, önemli olduğu düşünülen cümlenin özet metinde yer almasıdır. Cümlenin önemli olup olmadığına nasıl karar verilecek? Bu sorunu cevabı cümle seçimi ile özet çıkarma yönteminin başarı oranını belirlemede önemlidir. 2000’li yıllardan önceki çalışmalarda, bazı özellikleri içeren cümlelerin önemli olduğu ve bu cümlelerin diğer cümlelerden puanlandırma yöntemiyle ayrılması gerektiğine karar verilmiştir. Bu yöntemlerde basit ama etkili çıkarımlar ön plandadır. Bir metinde neye vurgu yapılmak istenirse o kelime o metinde daha çok kullanılır. Örneğin bir bilişim makalesinde teknolojinin yenilikleri anlatılıyorsa

"teknoloji" kelimesi çok kullanılacaktır ve içinde teknoloji kelimesi geçen cümleler önemlidir ve özette yer almalıdır. Burada dikkat edilmesi gereken nokta günlük hayatta sıkça kullanılan "ve","veya","ama" gibi kelimelerin bir içerikten bahsetmediğinin göz ardı edilmemesidir. Dolayısıyla bu gibi kelimeler dikkate alınmaz. Kelime sıklığı yöntemi denilen bu yöntemi Luhn (1958) çalışmasında kullanmıştır. Bu düşünceye ek olarak Edmundson (1969) çalışmalarında metinlerde, "özetle", "sonuçta" gibi kelimelerden sonra aslında o metnin veya konunun kısa bir özetinin çıkarılmış olduğunu düşünerek, buna benzer kelimeleri içeren cümlelerin de özette yer alması gerektiğini düşünmüştür. Bunun gibi özel bazı kritik kelimelerin belirlenip o kelimeleri içeren cümlelere fazla puan vermek suretiyle gerçekleştirilen bu yönteme ipucu sözcük öbekleri yöntemi denilmektedir. Edmundson'un bir başka görüşü de metindeki giriş ve sonuç paragraflarının önemli olduğudur. Yazarlar bir metinde en önemli cümlelerini sonuç paragrafına veya giriş paragrafına koyar. Gelişme paragrafları ayrıntıya yer verdiğinden puanlandırmada giriş ve sonuç paragraflarındaki cümlelerden daha az önceliğe sahiptir. Cümle Konumu Özelliği yöntemi bu prensibe dayanır.

Metinlerin karakteristik özellikleri göz önüne alındığında başlık ve alt başlıkların metnin konusu hakkında bilgi verdiği sonucu ortaya çıkar. Buna dayanarak metinde geçen kelimelerin metnin başlık ya da alt başlıklarında geçen kelimelerden herhangi biri olup olmadığına bakılır. Eğer cümleler bu kelimeleri barındırıyorsa bu cümlelerin puanları daha fazla olacaktır. Bu yönteme Başlık Terimleri Yöntemi denir ve Edmundson (1969) çalışmalarında bu yöntemden yararlanmıştır.

Metindeki tarih, sayı, nümerik karakter bilgisi, konuya has özellikler, önemli kelime grupları, kelime ve cümle konumu, cümle uzunluğu, başlık bilgisi, pozitif ve negatif anlam taşıyan kelimeler, merkezilik bir cümlenin önemini tespit etmede kullanılan özelliklerdir. Bu özellikleri Yeh ve ark. (2005), Hernandez ve Ledeneva (2009) ve Quyang ve ark. (2010) çalışmalarında kullanmışlardır.

Altan'ın (2004) çalışması Türkçe metinler üzerine yapılan ilk çalışmadır. Veri seti olarak 50 adet finansal konularda yazılmış metinleri kullanmıştır. HTML etiketlerini kullanarak başlık ve paragrafları ayırıp kelime frekansı, cümlelerin konumlarını irdelemiştir. Başlık sözcükleri, olumlu ve olumsuz sözcükler de çalışmasında kullandığı parametreler olmuştur.

Kılcı ve ark. (2008) özetleme çalışmasında, Altan'ın (2004) çalışmasına ek olarak, tarih ifadeleri, özel isimler, anahtar söz öbekleri, noktalama işaretleri ve nümerik karakterleri de incelemiştir. Ağırlıkların önceden elle girildiği çalışmasında,10 adet metni veri seti olarak kullanmıştır.

Pembe (2011), internet aramalarında kullanılan sorgu tabanlı bir çalışma ile karşımıza çıkar. Altan'ın (2004) çalışmasındaki gibi HTML etiketleri yardımıyla metni bölüm ve alt bölümlere ayırır. Sorgu temelli çalışmasında cümleleri ve bölümleri puanlamıştır.

Çığır ve ark. (2009), çalışmalarında, puanlandırmayı normalize etmişlerdir. Özellikler sıfır ve bir aralığındadır. Bu çalışmada cümlelerin, metin içindeki ve merkeze bağlı konumları incelenmiştir, 120 adet metni veri seti olarak kullanmıştır.

Bahsedilen bu yöntemler dilin yapısal özelliklerini kullanan yöntemlerdir. Çoğu İngilizce metinler üzerine olan bu çalışmalar incelendiğinde, önemli cümlelerin tespit edilip puanlandırılmasına dayanan bu yöntemleri birleştiren hybrid (melez) yöntemlerin de önerildiği görülmektedir. Suanmali ve ark. (2009), Babar ve Patil (2014), Kyoomarsi ve ark. (2009), Binwahlan ve ark. (2010) bu özellikleri birleştirmek için bulanık mantık tekniklerini kullanır. Silla ve ark. (2004), Witte ve ark. (2007), Berker ve Güngör (2012), Filatova ve Hatzivassiloglou (2004), Mcdonald (2007) bu özelliklere verilen puanların sezgisel, genetik, sürü tabanlı ve diğer en uygun şekle sokma algoritmaları ile gerçekleşebileceğine dair çalışmalar yapmışlar. Alguliev ve ark. (2011) nonlineer programlama yöntemleriyle özetleme sorunlarına çözüm aramıştır.

Özetleme sorununu öğrenme problemi olarak gören Svore ve ark. (2007), Wong ve ark. (2008), makine öğrenmesi, bayes sınıflandırıcı, destek vektörü makineleri, yapay sinir ağları yöntemlerini kullanmışlardır.

Cümlelerin birbirlerine bazı özelliklerine göre benzediğinden hareketle K-ortalamalar yöntemini Agrawal ve Gupta (2014) kullanmıştır. Seinberger ve ark. (2005), çalışmalarında, cümlelerin seçiminin dışında, kelimeler arasındaki anlamsal ilişkiden yola çıkarak uyum tabanlı teknikleri kullanmıştır. Uyum tabanlı (choesion based) bu İngilizce özetleme çalışmalarında Wordnetten yararlanmışlardır. D'avanzo ve ark. (2004) varlık ismi tanıyan ve çoklu kelimeleri bulan, eğiticili öğrenmeye dayalı bir çalışma gerçekleştirmiştir.

Özetleme çalışmalarında kelimelerin anlamından yola çıkan çalışmalar incelendiğinde olay tabanlı event-based yöntemlere de literatürde rastlanmaktadır. Nedensellikten yola çıkılarak yapılan çalışmalarda Liu ve ark. (2007), Filotova ve Hatzivassiloglou (2004), metindeki olayları irdeler. Bir olay terim grafiği elde edilir. Aynı olayları belirten terimler gruplandırılır.

Belirlenen özelliklere ek olarak metinden çıkarılan birçok karakteristik özellik metin hakkında genel fikir sahibi olunmasını sağlar. Yapılan çalışmalarda kullanılan 13 farklı özelliği inceleyen bir çalışmada Uzundere ve ark. (2008), bu özelikleri şu şekilde sıralanmıştır:

Başlık: Özeti çıkarılacak metindeki cümlelerin kelimeleri, başlıkta veya alt başlıklarda geçen kelimelerden herhangi biri mi?

Yüksek Frekans: Metinde geçen bütün kelimelerin, metin içindeki tekrar sayısını tutan bir liste hazırlanır. Liste kelimelerin metinde geçen tekrar sayısına göre azalan yönde sıralanır. Bu listeden günlük hayatta sık kullanılan "ve", "veya", "ama", "ile", "için" vs. kelimeler dâhil edilmez. Sezgisel veya genetik algoritma yöntemleriyle belirlenen bir yüzdeye sahip kelimeler özete dâhil edilir.

Yer: Cümle bazlı özellik yönteminde, hangi cümlenin giriş, gelişme veya sonuç paragraflarından hangisinde olduğu bilgisine bakılır.

Anahtar Kelimeler: Kullanıcının da dâhil edildiği yöntemde kullanıcıdan, özeti çıkarılacak metnin içinde önemli olduğunu düşündüğü bir ya da birkaç kelime istenir ve bu kelimeleri içeren cümlelerin önemli olduğu kabul edilir. Örneğin spor konusundan bahseden bir metin için skor, sonuç gibi kelimeler önemli sayılacaktır.

Özel İsimler: Özel isimler çoğu zaman detaydan çok bilginin kendisidir. Özel isimlerin özette yer alması, önemlidir.

Pozitif Kelimeler: Metinde geçen "özetle", "sonuç olarak", "sonuçta", "neticede" vs. gibi kelimelerden sonra, ilgili metinle alakalı sıkıştırılmış bilgiler verileceği düşüncesinden yola çıkılarak, bu kelimeleri içeren cümlelere artı puan verilir.

Negatif Kelimeler: "Çünkü", "ancak"," öyleyse″ ve buna benzer kelimeler salt bilgiden çok detaya inileceğini gösterdiği için negatif puanlandırılır.

Eşdizimli Kelimeler: Anlamı pekiştiren sözcükler önemli kabul edilir.

Sayılar: Dikkat edilmesi gereken bazı noktalarla beraber içinde rakam bulunduran cümlelerin puanı yüksek olarak kabul edilir. Dikkat edilmesi gereken noktalardan bir tanesi de belgisiz sıfatların rakam olmadığıdır.

Çift Tırnak İşareti: Aktarım içeren kelime ve kelime grupları, özellikle haber metinleri için önem taşıdığından önemli olarak gösterilen bir diğer niteliktir.

Bitiş İşareti: Sonuna hangi noktalama işareti konulduğu cümlenin vurgulanıp vurgulanmaması adına önemlidir. Noktayla bitirilmiş bir cümle ile sonuna ünlem işareti konmuş bir cümle aynı önem derecesine sahip değildir.

Ortalama Uzunluk: Metinde yer alan cümlelerin sahip olduğu kelime sayısının ortalaması hesaplandıktan sonra bu ortalamaya yakın cümleler önemlidir.

Gün Ay: Gün, ay adı içeren cümleler salt bilgidir, özette yer alması için yüksek puanlandırılır.

Çalışmada metin sözcüklere, cümlelere, paragraflara ayrılmıştır. Ayırma işleminde zemberek kullanılmıştır. Ayrılan bu yapılar belirlenen 13 özelliğe göre incelenir. Çalışmada bu özellikler puanlandırılırken sezgisel yöntem kullanılmıştır. Cümlelerin ağırlıklandırılması Çizelge 2.1.'de gösterilmiştir.

Anahtar kelimelerden 2 adet ve rakam içeren 3 adet kelime bulundura bir cümle örnek olarak alınırsa, cümlenin toplam puanı: (2*8) + (3*3) = 25 olacaktır.

Her cümlenin puanı hesaplandıktan sonra, cümlenin, kullanıcının belirlediği bir orana göre (özetleme), özette geçecek olan cümleler en yüksek puandan başlanarak seçilmiş olur. Özetleme oranı örneğin yüzde 35 seçilirse en yüksek puanlı cümlelerin yüzde 35'i özette yer alır. Özetleme oranının değiştirilerek yapıldığı çalışmada bu değerlerin ortalaması alınarak, özetleme başarısı ortalama %55 olarak ölçülmüştür.

Türkçe internet sayfalarındaki metinlerin otomatik özetini çıkaran başka bir çalışmada Sami ve Diri (2010), cümle seçimini 12 farklı özelliğe göre belirlemiş, ağırlıklandırmaları sezgisel vermiştir. Web tabanlı bu sistemde başarı oranı % 59 olarak ölçülmüştür.

Çizelge 2.1. Kelimelere özelliklerine verilen puanlar (Uzundere ve ark., 2008)

Özellikler	Ağırlık Değerleri
Başlık	20
Yüksek Frekans	10
Giriş	20
Sonuç	2
Anahtar Kelimeler	8
Özel İsimler	3
Pozitif Kelimeler	15
Negatif Kelimeler	-20
Eşdizimli Kelimeler	4
Sayı	3
Çift tırnak İşareti	2
Bitiş İşareti	2
Ortalama uzunluk	10
Gün Ay	5

Çıkarımsal özetleme tekniği olarak metin ve cümle kümeleme tekniklerinin kombinasyonunun kullanıldığı başka bir çalışmada, benzer kelime ve dokümanlar kümelenmiş ve bu kümede bulunan en yüksek frekansa sahip cümle seçilmiştir (Dashpande ve Lobo, 2013).

Kelime çıkarımı yöntemlerini kullanan çalışmalar içinde dikkat çeken bir çalışmada Wadhvani ve Roy (2013), metinden elde edilecek bilgi iki aşamada elde edilir. İlk aşamada bazı kurallara göre etiketlenen cümlelerin ayrıştırılması yapılmıştır. İkinci aşamada sözcük öbekleri, kelime frekansı, konumu vs. yöntemlerle birlikte Modal Verb'ler kullanılmıştır. İngilizce metinlerin özetinin çıkarılması için gerçekleştirilen bu çalışmada bilgileri süzmek için cümlelerin zamanlarından yararlanılmıştır.

Güran ve ark. (2014) çalışmasında, çıkarıma dayalı Türkçe özetleme sisteminin kullanıldığını görmekteyiz. Cümle seçim metotlarına yeni bir kıstas eklenen bu çalışmada varlık ismi tanıma metodu ilk kez kullanılmıştır.

2014 yılında yayınlanan bir tezde Attokurov (2014), metin özetleme sistemlerinin asıl amacının bilgi tekrarını önlemek olduğunu öne sürmüştür. Tekrarın önlenmesi için ağaç budama algoritması (Optimal Tree Pruning algorithm) ile HAC

(Hierarchical Agglomerative Clustering) algoritması birlikte kullanılmıştır. HAC tekrarı ayıklamada, budama algoritması da özetlerdeki tekrarı azaltmada kullanılmıştır.

Kumar ve Chandrakala (2016) çalışmalarında, özetleme sorununa çözümler ararken optimizasyon algoritmalarıdan yararlanmıştır. Çalışmada, parçacık sürü optimizasyonu, yapay arı kolonisi algoritmaları, genetik algoritmalar ve karınca kolonisi optimizasyonu gibi optimizasyon algoritmalarının daha başarılı özetler çıkarmak için yardımcı olacağı vurgulanmıştır.

Anlamsal analiz yöntemlerini kullanan çalışmalar da mevcuttur. Sözcük-olay ilişkisinin iyi belirlenmesi ve terimlerin olay etiketinde gruplanması, metindeki anlamın ortaya çıkmasında önemli bir yere sahiptir. Bu tespitten sonra lineer cebir yöntemleri kullanılarak özet başarısı artırılmaya çalışılmıştır. Metin sözcüklere ve cümlelere ayrıldıktan sonra bir terim cümle matrisi elde edilip çarpanlarına ayrılır. Daha sonra sözcükler ve cümleler lineer bağımsız vektörlere dönüştürülür. Bu vektörler sözcük ve cümleleri anlamsal yönden gruplar. Önemli sözcükler ve dolayısıyla cümleler, anlamsal olarak gruplanan çarpan matrisleri ile seçilir. Bu işlemin ardından gizli anlamsal indexleme (Latent semantic analysis), olasılıksal gizli anlamsal analiz (Probablistic semantic analysis), negatif olmayan matris ayrışımı (Nonnegative matrix Factorizasyon) yöntemleri kullanılarak metinlerin anlamsal yönden analizi gerçekleştirilir. Gizli anlamsal analiz tekniğini kullanan çalışmalara örnek olarak Murray ve ark. (2005), Steinberger (2007), Yeh ve ark. (2005)' nın çalışmaları verilebilir. Negatif olmayan matris ayrışımı yöntemi uygulanan çalışmalar: Lee ve ark. (2009), Masheckin ve ark. (2011) ve Güran ve ark. (2011)'dir. Güran (2013) çalışmasında "Türkçe vikipedi" linklerinden sıralı kelimeler tespit etmiştir. Bhandari ve ark. (2007), Chris ve Ding (2005), alasılıksal gizli anlamsal analiz tekniklerini kullanmıştır.

Cümleler arasındaki ilişkiyi belirlemek için negatif olmayan matris ayrışımı (Non-negative Matrix Factorization) yöntemini, metnin önişleme kısmında kümeleme ve sınıflandırma yöntemlerini kullanıp birleştiren bir çalışmada kümelemenin özetleme kalitesini artırdığı tespit edilmiştir (Dmitry ve ark., 2013).

Anlamsal analiz yöntemi kullanan Gong ve Lui (2001), metni konulara ayırıp her konudan bir cümle seçilmesi mantığıyla hareket etmiştir. Buna karşılık Murray ve ark. (2005), her konudan birden çok önemli cümleyi seçer. Özsoy ve ark. (2005)

çalışmasında metindeki bütün konuları içeren cümleleri gürültüden arındırarak seçme yoluna gitmiştir.

Özsoy ve ark. (2005) gizli anlamsal analiz tekniğinden yola çıkarak yeni yöntemler kullandığı çalışmalarında, ”çapraz” ve “konu” yaklaşımlarını kullanmıştır. Çapraz yönteminin sonuçları daha başarılı çıkmıştır.

Sözcüklerin birbirine bağlılığının göstergesi olarak sözcük zincirlerini kullanan çalışmalar da literatürde mevcuttur. Berker (2011) çalışmasında, kelimelerin bağlılığından yararlanarak sözcük zincirleri oluşturmuş daha sonra bu derinlemesine bilgiyi daha üst seviye analizlerle birleştirmiştir. Bu farklı düzeydeki bilgileri daha sonra genetik algoritmalarla birleştirmiştir.

2010 yılında yapılan bir çalışmada Saggion ve ark. (2010), FRESA (Framework for Evaluating Summaries Automatically) adı verilen içerik tabanlı özet değerlendirme yazılımını geliştirmiştir. Çalışmada odak tabanlı, çoklu doküman girdili İngilizce özetleme yapabilen bir sistem ile genel, tek doküman girdili Fransızca ve İspanyolca özet çıkarabilen yöntem, Rouge değerlendirme paketi ile ölçülerek tanıtılmıştır.

Dokümanlardaki konuları indeksleyen belge vektörlerini (metin konusu- text topic) ve kelimeleri tutan kelime vektörlerini, cümlelerin birbirleriyle ilişkisini hesaplamak için kullanan ARTEX (AnotheR TEXt summarizer) adlı özetleme algoritmasının tanıtıldığı bir çalışmada cümleler önce kural tabanlı olmayan dilbilimsel önişlemden geçirilir. VSM (Vector Space Model) tabanlı çalışmada her cümlenin özelliği, normalize edilen puanlama ile matris formatında hesaplanır (Manuel ve Moreno, 2012). Dilbilimsel bilgi ve önceden hazırlanmış kaynak gerektirmemesi çalışmanın önemli yönlerindendir. Artex Fransızca ve İspanyolca metinler üzerinde denenmiş olup başarılı sonuçlar üretmiştir.

Özetleme uygulaması geliştirilen bir çalışmada, gizli anlamsal analiz tekniklerinden farklı uzay vektör modellerinin karşılaştırılması sunulmuştur (Padmakumar ve Eswaran, 2014). Ngram değerlendirme yöntemleriyle başarının ölçüldüğü projede, binary gösterim modelinin karmaşık terim frekansı gösterimlerine göre daha başarılı olduğu sonucuna ulaşılmıştır.

Anlam olarak birbirine bağlı kelimelerin istatistiksel dağılımlarına ve anlatımdaki kullanım şekillerine bakılarak dil işleme problemlerine çözüm getirilebilir.

Bu yargıdan yola çıkarak yapılan bir çalışmada (Gönenç (2012), önce kelimeler arası anlam ilişkisinin ölçülmesi için anlam bütünlüğü kullanılmıştır. Bilgi dağarcığını kullanmayı gerektiren bu yöntemde hem kelime ilişkilerinin elle girildiği anlam bilgi kümesi hem de kelimelerin metindeki konum dağılımından elde edilen bilgi dağarcığı kullanılmıştır. Daha sonra konularına bölümlenen metinler özet çıkarma problemlerinde kullanılmıştır. Kelime zinciri kaynaklı yöntemlerden daha başarılı sonuçların elde edildiği gözlemlenmiştir.

Çakır ve Çelebi (2011) çalışmalarında özetlemenin bütün dilleri kapsadığı bir yöntem sunmuşlardır. C3M (Cover Coefficient-Based Clustering Methodology) algoritması kullanılan çalışmada cümleler arası benzerlikleri kullanıp seçilebilecek cümleleri tespit etmişlerdir. Rouge değerlendirme paketi ile çalışmanın başarımını ölçtükleri sistem, başarılı sonuçlar üretmiştir.

Cümleler arası benzerlik yöntemleri genelde metinden önemli cümlelerin belirlenip seçildiği cümle seçerek özetleme yöntemlerinde kullanılmaktadır. Bağımsal dilbilgisini cümleler arası benzerlik bulma problemlerinde kullanan çoklu doküman özetleme çalışmasında, farklı bağımsal ağaç bazlı benzerlik tespiti yapan teknikler kullanılmıştır (Bilgin, 2014). Bağımsal ağaçlardaki bağlantıların cümlelerin önemli sözcüklerini tespit ettiği sonucuna varılmıştır.

Literatürde dikkat çeken nokta, çalışmaların çoğunun İngilizce dili üzerine olduğudur. İspanyolca, Çince, Fransızca dilleri için de metin özetleme sistemleri geliştirilmiştir. Arapça üzerine yapılan bir çalışmada Arapça özet çıkarma çalışmalarının azlığı nedeniyle Arapça özet standartlarının eksikliğinden bahsedilmektedir (Al-Saleh ve Menai, 2016). Bu olumsuz durumun yavaş yavaş değişmeye başladığı dile getirilmiş olup TAC 2011 MultiLing Pilot and ACL 2013 MultiLing Workshop çalıştaylarına Arapçanın dâhil edildiği vurgulanmıştır. Bu çalışmada ayrıca literatürdeki diğer Arapça özet çıkarma yöntemleri gruplanarak bir tablo halinde sunulmuştur.

Yapısal ve anlamsal özeliklerin bir arada kullanıldığı başka bir çalışmada, uzman kabiliyetine dayalı bulanık tabanlı bir yöntemi, otomatik olarak sezgisel bir algoritma ile birleştiren melez bir sistem önerilmiştir (Güran, 2013). Melez sistemle önerilen sistem her bir özelliğin ayrı ayrı kullanılması ile elde edilen sistemden daha

başarılı sonuçlar vermiştir. Çalışmada, terim cümle matrisinin oluşumu için kullanılan yeni ağırlıklandırma tekniklerinde, sözcüklerin metin içindeki konumuna dayalı ve sözcüklerin ait oldukları cümlelerin önemine dayalı olmak üzere iki yeni yöntem önerilmiştir. Çalışmada ayrıca bundan sonraki Türkçe metin özetleme çalışmalarında kullanılabilecek iki yeni veri seti üretilmiştir. Bunlardan ilki 130 haber dokümanından oluşan ve üç farklı kişinin okuyup özetini çıkardığı dokümandır. İkincisi ilkine göre daha kısa metinlerden oluşan 20 adet haber dokümanıdır.

2.1. Özetin Başarısının Ölçülmesi

Bir sistemin çıkardığı özetin başarılı olduğu nasıl anlaşılır? Başarıyı test etmek için öncelikle ideal özetin nasıl çıkarıldığına bakmak gerekir. Çünkü başarı testi için ideal özete sahip olunmalıdır.

İdeal özeti ortaya çıkarmak için araştırmacılar kendi çalışmalarında farklı yöntemleri kullanmışlardır. Örneğin Rath ve ark. (1961), değişik zamanlarda özet çıkarmak istedikleri kişilere bir makaledeki, makaleyi temsil edebilecek cümleleri seçmelerini istemiştir.

Başka bir çalışmada, bir program ara yüzü vasıtasıyla, çeşitli kişiler tarafından, özet çıkarılması istenen metinlerden, önemli olduğunu düşündükleri cümleleri seçmelerini istenerek veri seti elde edilmiştir (Güran, 2013). Başka bir çalışmada Morris ve ark. (1992), GMAT (Graduate Management Admission Test) okuduğunu anlama egzersizleri bölümündeki soruları cevaplamayı da kapsayan bir değerlendirme yöntemi önerir.

Başarı çeşitli yöntemlerle ölçüm yapılarak değerlendirilir. En yaygın yöntem, sistem tarafından oluşturulan özet ile ideal özetin çakışan cümle sayısının hesaplanmasıdır.

Diğer bir başarı ölçme yönteminde ise konu ile ilgili bilgi sahibi olan kişilerin değerlendirmeleri, özel alanlar belirlemek için kullanılır. Bu sebeple metin sınıflama (text categorization), bilgi çıkarımı (information retrievel) ve soru cevaplama (question answering) gibi teknikler kullanılmaktadır.

Literatürde en sık karşılaşılan yöntemler: Keskinlik (precision), anma (recall), F-ölçüm değeri (F-score); göreceli fayda değeri; kosinüs benzerliği, Ngram birliktelik istatistiği (ROUGE) değerleridir (Güran, 2013).

3. MATERYAL VE YÖNTEM

Bu bölümde bu tez için kullanılacak olan programlama ortamı, programlama dili ve algoritmalar hakkında bilgiler verilmiştir.

3.1. Kullanılan Yazılım Teknolojileri ve Ortamları

Bu tez süresince bir programlama dili ve geliştirme ortamı kullanılarak masaüstü uygulaması geliştirilmiştir. Bu kısımda geliştirilecek uygulamaya ilişkin kullanılan çeşitli yazılımlar ve donanımlar belirtilmiştir.

Uygulama geliştiriminde kullanılan donanımlar:

- Windows 7 işletim sistemli bir Dizüstü Bilgisayar
- i7 - 2,10 GHz işlemci
- 8 GB Ram

Uygulama geliştirme ortamları, yazılımlar, algoritmalar, veri seti:

- Microsoft Visual Studio 2010
- Zemberek
- NZemberek-Master
- Visual C#
- Genetik Algoritmalar
- Veri Seti

3.1.2. Zemberek

Zemberek, Türkçe ve diğer Türkî diller için yazılmış, biçimbirimsel çözümleme, yazım denetimi, sözcük üretme gibi temel DDİ (Doğal Dil İşleme) işlemlerini yapabilen açık kaynak kodlu DDİ kütüphanesidir. Zemberek'in açık kaynak kodlu olması ve

başarısının küçümsenmeyecek düzeyde olması nedeniyle, Türkçe doğal dil işleme yazılımları içinde önemli bir yere sahiptir. Araştırmacılar sözcüklerin kök ve gövdelerinin birleşimini veren çözümleyicilerden Zemberek'i sıklıkla kullanmışlardır.

Zemberek yapısal anlamda, dil yapı bilgisi ve derinlemesine dil işlemleri olmak üzere iki kısımdan oluşur. Esas kütüphane derinlemesine dil işlemleri için gerekli yazılımları içerir. Kök ve ek bilgilerini barındıran yapıları mevcuttur. Ek bilgileri için, Türkî dillere has ek bilgilerini tutan XML kütüphaneleri barındırır. Ekler özel durumları göz önünde bulunduran algoritmalarla uyumlu çalışır. Ses düşmesi, yumuşama, benzeşme gibi ses olaylarının üretimini yöneten birim de mevcuttur. Bunları sağlamak için graf ağaçları algoritmaları kullanılır. Zemberek saniyede 12000 sözcüğü çözümleme yeteneğine sahiptir. Türkçe özetleme çalışmalarında bir araç olarak kullanılan zemberek, çalışmalara yardımcı olması bakımından oldukça önemli bir yazılımdır.

İngilizce metinler için araştırmacılar Porter Stemmer algoritmasını kullanarak kelimeleri köklerine ayırırlar.

Bu tezde kelimelerin kökünü bulmak için Zemberek adlı yazılımdan yararlanılmıştır. Nzemberek-Master adlı yazılımı, C# programlama dili ile geliştirilmiş olup, yazılımda Türkçe ek almış kelimeleri eklerine ve köklerine ayırma işlemi olan çözümleme, Türkçe olup olmadığının belirlenmesi için kelime denetleme ve o kelimeden ekler yardımı ile türeyecek yeni kelimeler öneren modül bulunmaktadır. Şekil 3.1'de NZemberek-Master uygulamasının arayüzünün bir bölümü görülmektedir.

Şekil 3.1. NZemberek-Master uygulaması arayüzü.

3.1.3. Genetik algoritmalar

Genetik algoritmalar, doğada gözlemlenen evrimsel sürece benzer bir şekilde çalışan arama ve eniyileme yöntemidir. Karmaşık çok boyutlu arama uzayında en iyinin hayatta kalması ilkesine göre bütünsel en iyi çözümü arar (Emel ve Taşkın, 2002).

Genetik algoritmalar problemlerin çözümü için evrimsel süreci bilgisayar ortamında taklit ederler. Diğer eniyileme yöntemlerinde olduğu gibi çözüm için tek bir yapının geliştirilmesi yerine, böyle yapılardan meydana gelen bir küme oluştururlar. Problem için olası pek çok çözümü temsil eden bu küme genetik algoritma terminolojisinde nüfus adını alır. Nüfuslar vektör, kromozom veya birey adı verilen sayı dizilerinden oluşur. Birey içindeki her bir elemana gen adı verilir. Nüfustaki bireyler evrimsel süreç içinde genetik algoritma işlemcileri tarafından belirlenirler.

Problemin bireyler içindeki gösterimi problemden probleme değişiklik gösterir. Genetik algoritmaların problemin çözümündeki başarısına karar vermesindeki en önemli faktör, problemin çözümünü temsil eden bireylerin gösterimidir. Nüfus içindeki her bireyin problem için çözüm olup olmayacağına karar veren bir uygunluk fonksiyonu vardır. Uygunluk fonksiyonundan dönen değere göre yüksek değere sahip olan bireylere, nüfustaki diğer bireyler ile çoğalmaları için fırsat verilir. Bu bireyler çaprazlama işlemi sonunda çocuk adı verilen yeni bireyler üretirler. Çocuk kendisini meydana getiren ebeveynlerin (anne, baba) özelliklerini taşır. Yeni bireyler üretilirken düşük uygunluk değerine sahip bireyler daha az seçileceğinden bu bireyler bir süre sonra nüfus dışında bırakılırlar. Yeni nüfus, bir önceki nüfusta yer alan uygunluğu yüksek bireylerin bir araya gelip çoğalmalarıyla oluşur. Aynı zamanda bu nüfus önceki nüfusun uygunluğu yüksek bireylerinin sahip olduğu özelliklerin büyük bir kısmını içerir. Böylelikle, pek çok nesil aracılığıyla iyi özellikler nüfus içerisinde yayılırlar ve genetik işlemler aracılığıyla da diğer iyi özelliklerle birleşirler. Uygunluk değeri yüksek olan ne kadar çok birey bir araya gelip, yeni bireyler oluştururursa arama uzayı içerisinde o kadar iyi bir çalışma alanı elde edilir.

Bu tezde genetik algoritma yöntemleri kullanılmıştır.

3.1.4. Tez kapsamında kullanılan veri seti

Bu tezin uygulama aşaması için literatürdeki özetleme çalışmalarından birinde (Güran, 2013) kullanılan ve "Veri Seti 2" adındaki veri seti kullanılmıştır. Bu set 20 adet haber dokümanından oluşmaktadır. 30 kişiye bu 20 adet haber dokümanı verilmiş olup, her bir metinden o metnin özeti olabilecek cümleleri seçmeleri istenmiştir. Her bir haber metninin sahip olduğu cümle sayısının %35'i özet metinde yer almak üzere seçilmiştir.

3.2. Başarının Ölçümü

Özetleme sistemlerinin genel anlamda ikiye ayrıldığından bahsedilmişti. Çıkarımsal özet çıkarma yöntemlerinde sonuçta elde edilen özet metin içindeki önemli cümlelerin birebir aynısının seçilmesiyle elde edilir. Yorumlayarak özet çıkarma yöntemlerinde çıktı olarak elde edilen özet, metindeki cümlelerden farklı olabilir. Çıkarımsal özet çıkarma yöntemlerinde başarıyı tespit etmek için ideal diye tabir edilen ve metinden elde edilen cümlelerle test edilecek özetin cümlelerinin birebir eşleştirilmesine bakılır. Eşleşen cümle sayısı ile hesaplama yapılır. Ancak yorumlayarak özet çıkarma yöntemlerinde ideal özet ile test edilecek sistem özetinin birebir eşleşmesi mümkün değildir. Farklı teknikler ile başarı ölçülür.

Bu tezde çıkarımsal özet çıkarma yöntemi kullanıldığından bu yönteme ait sistem başarısının hesaplanabildiği yöntemler anlatılacaktır.

3.2.1. Keskinlik, anma, F-ölçüm değeri

KAF (keskinlik anımsama F skor değeri) yönteminde amaç ideal özet ve başarısı ölçülmek istenen özetin sahip olduğu ortak cümle sayısının bulunmasıdır. Aşağıda formülleri verilen yönteme göre, ideal özet ve başarısı ölçülmek istenen özetin sahip olduğu ortak cümle sayısını, başarısı ölçülmek istenen özetin sahip olduğu cümle sayısına bölerek elde edilen sayının bulunması ile formülize edilir. Buna kesinlik değeri (K) denir. İdeal özet ve başarısı ölçülmek istenen özetin sahip olduğu aynı cümle sayısını, ideal özetin cümle sayısına oranına anma değeri denir (A). Kesinlik ve anma

değerlerinin harmonik ortalaması F-ölçüm değeridir. Bu yöntemin sorunu, çıkarılan ideal özette önemli olduğu düşünülen cümlelerin seçiminin kişiden kişiye değişmesinin hesaba katılmamasıdır. Bu durumda aynı öneme sahip iki özetin başarımı farklı sonuçlar doğuracaktır (Bkz. Eş. 3.1).

$$K = \frac{|S \cap T|}{|S|} \qquad A = \frac{|S \cap T|}{|S|} \qquad F = \frac{2KA}{K+A} \tag{3.1}$$

3.2.2. Göreceli fayda

F-Ölçüm değeri yöntemindeki sorunu gidermek için bir çalışmada alternatif bir yöntem önerilmiştir (Dragomir ve ark,. 2000). Önerilen yöntemde cümleler özete katılma durumuna göre insan sezgilerine dayanılarak puanlandırılır ki buna fayda (utility) puanı denilir. Her cümle puanlandırılır. Ölçüm için bir veya birden fazla N adet kişiye n adet cümleyi puanlandırması istenir. Fayda puanına göre en yüksek puana sahip olan *e* adet cümleye "*e* boyutlu çıkarılmış cümle takımı" adı verilir.

$$Fayda = \frac{\sum_{j=1}^{n} \delta_j \sum_{i=1}^{N} u_{ij}}{2 \sum_{j=1}^{n} \in_j \sum_{i=1}^{N} u_{ij}} \tag{3.2}$$

Yukarıdaki eşitlikte u_{ij} , i. değerlendiricinin j. cümle için verdiği fayda puanını ifade ederken $\in_j$, bütün kişilerin verdikleri fayda değerlerinin toplamına göre en yüksek *e* cümle için 1 iken aksi durumda 0'dır. δ_j Sistem tarafından çıkarılan en yüksek skorlu *e* cümle için 1, diğer durumlar için 0'dır (Bkz. Eş. 3.2).

3.2.3. Kosinüs benzerliği

Başka bir ölçüm yöntemi olan kosinüs yönteminde, x_i ideal özette bulunan kelime sıklığı değeri, y_i ölçüm yapılacak özette bulunan kelime sıklık değeridir. (Bkz. Eş. 3.3). Kosinüs benzerliği yönteminin ayrıntıları Salton (1988) 'in çalışmasındadır.

$$\cos(X,Y) = \frac{\sum_i x_i y_j}{\sqrt{\sum_i (x_i)^2}\sqrt{\sum_i (y_i)^2}} \quad (3.3)$$

3.2.4. Ngram birliktelik istatistiği (ROUGE)

ROUGE (Recall- Oriented Understudy for Gisting Evaluation) yazılımı 2004'te Lin (2004) tarafından Perl programlama dili kullanılarak üretilmiştir. İlk olarak doküman anlama konferansında (DUC - Document Understanding Conference) kullanılmıştır (DUC Konferansı, 2012). Karşılaştırılacak olan iki dokümanın ortak kelime sayısına dayanan bir ölçüm yöntemidir.

ROUGE'nin beş farklı ölçüm şekli vardır: ROUGE-N, ROUGE-L, ROUGE-S, ROUGE-W, ROUGE-SU. ROUGE-N ölçülmesi istenen özet doküman ile ideal özet dokümanı arasındaki Ngram (N uzunluklu sıralı kelime grubu) değeridir Eş.3,4'teki gibi hesaplanır.

$$ROUGE - N = \frac{\sum_{S\in\{\text{Human Summeries Group}\}} \sum_{gram_{n\in S}} Calculate_{overlap}(gram_N)}{\sum_{S\in\{\text{Human Summeries Group}\}} \sum_{gram_{n\in S}} Calculate\,(gram_N)} \quad (3.4)$$

$Calculate_{overlap}(gram_N)$, ideal ve özeti sistem tarafından çıkarılmış özetlerin sahip olduğu ortak en fazla kelime grubu sayısıdır. $Calculate\,(gram_N)$ ise ideal özetteki toplam N uzunluktaki sıralı kelime grubu sayısıdır.

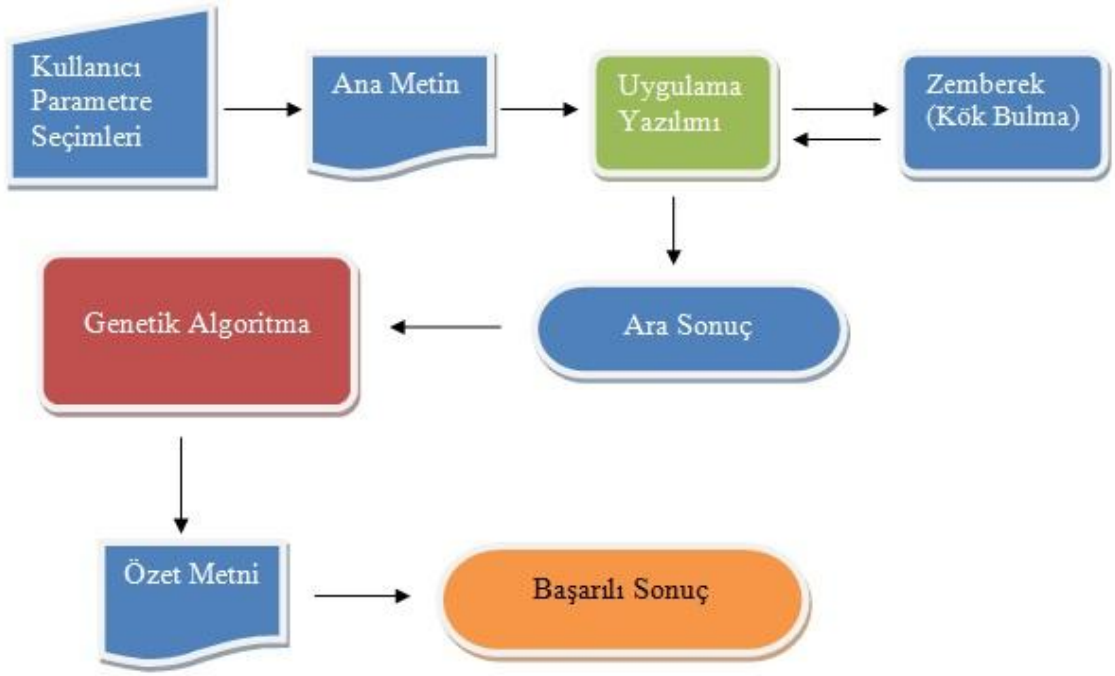
Bu tezde ideal özetle sistem özetinin cümleleri karşılaştırılarak başarı ölçülmüştür.

4. GENETİK ALGORİTMA İLE ÇIKARIMSAL METİN ÖZETLEME

4.1. Sistemin Açıklaması ve Tasarımı

Bu tezde çıkarıma dayalı özetleme yöntemi kullanılmıştır. Geliştirilen uygulama yazılımına bir metin belgesi girdi olarak alınıp işlemin sonucunda başarılı bir özet çıktı olarak elde edilir. Yapılan bu çalışmada genetik algoritma yönteminin yardımı ile başarı artırılmıştır.

Sistemin akış şeması Şekil 4.1'de görülmektedir. Buna göre öncelikle kullanıcı özeti seçeceği metni belirler. Daha sonra uygulama panelinden programın nasıl çalıştırılacağına dair parametreleri belirler ve uygula butonuna basar.



Şekil 4.1 Gerçekleştirilen uygulamanın akış şeması.

Özetle, sistem kullanıcının girdiği parametrelere göre öncelikle özeti çıkarılacak metnin kelimelerini köklerine ayırır bunu için NZemberek-Master adlı açık kaynak kodlu yazılımdan yararlanılmıştır. Sonraki aşamada belli özellikteki kelimelere puanlar verilmiştir. Daha sonra kelime puanları toplanarak cümle puanları hesaplanmıştır.

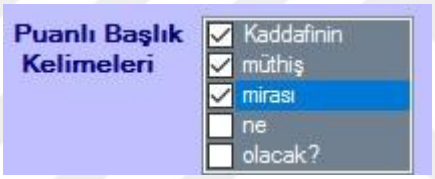
Azalan sırlama ile dizilen cümlelerden yine kullanıcının belirlediği bir özetleme yüzdesi ile en iyiler seçilmiştir. Bu en iyi cümleler özeti temsil eder. Başarı ölçümü ideal özetle karşılaştırılmıştır. İdeal özette var olan her bir cümle sistemin ortaya çıkardığı her cümle ile çakışıyorsa başarılı kabul edilmiştir. Bir başarı değeri elde edilmiştir. Buraya kadar bir bireyin başarısı hesaplanmış olur. Kullanıcının önceden belirlediği birey sayısı kadar birey oluşturulmuştur. Bu çalışmada varsayılan olarak 10 birey popülâsyona dâhil edilmiştir ve 10 adet başarılı birey bir sonraki nesil için girdi olarak alınmıştır. Parametrik ve varsayılan olarak 20 iterasyon olan genetik algoritma bitiş kriteri kadar genetik algoritma uygulanmış ve başarının artışı sağlanmıştır.

Sistemde kurgulanan yaklaşımda özeti çıkarılacak metindeki her cümleye özette yer alabilme yeteneğine göre puan verilmektedir. Buradaki temel amaç konu ile alakalı ve en fazla bilgiyi içeren kelimeleri tespit edip özete dâhil olmasını sağlamaktır. Puan verilirken her kelime ayrı ayrı ele alınıp kelimenin 12 farklı özelliğine göre önceden belirlenen puanlar verilmiştir. Cümlenin puanı, o cümleyi içeren kelimelerin puanları toplamına eşit olur. Kelimelere verilen puanlar o kelimenin özelliğine göre sezgisel belirlenmiştir. Puanlar, sezgisel olarak belirlenen bir aralık dâhilinde yazılım tarafından rastgele tamsayı değerleri verilerek elde edilmiştir. Kelimeler ele alınırken önce köklerine ayırma işlemi uygulanmıştır. Bu aşamada NZemberek-Master adlı yazılım kullanılmıştır. Bu çalışmada kelimelerin hangi özelliğine ne kadar puan verildiğinin bilgisi Çizelge 4.1'de gösterilmiştir.

Çizelge 4.1. Kelime özelliklerine verilen puan aralıkları

Özellikler	**Puan Aralıkları**
Başlık	30-20
Yüksek Frekans	10-20
Giriş	30-20
Sonuç	10-20
Anahtar Kelimeler	30-20
Özel İsimler	10-20
Pozitif Kelimeler	30-20
Negatif Kelimeler	0-10
Sayı	0-10
Çift Tırnak İşareti	0-10
Bitiş İşareti	0-10
Tarih	0-10

Başlık özelliğinde, metnin başlığında geçen her bir kelime metindeki kelimelerle karşılaştırılır aynı olan kelimelere sistem sıfır ile otuz arasında rastgele bir puan verir. Bu tez çalışmasındaki arayüzde başlıkta geçen kelimeler kullanıcının tercihine bırakılmıştır. Kullanıcı başlıkta geçen kelimelerden hangisine karar verirse metinde geçen o kelimelere puan verilir. Böyle bir ayırıma gidilmesinin nedeni başlıkta geçen her kelimenin özette yer alması adına önemli olup olmadığının sezgisel olarak kontrol edilmesidir. Örneğin bu veri setindeki sekizinci haberde başlık "Sigortacılıkta Türkiye için en büyük risk doğal afetler" şeklindedir. Kullanıcı arayüzünde, isteğe bağlı olarak "en" kelimesi başlık kelimesi olarak tercih edilmeyebilir. Bahsedilen "en" kelimesi herhangi bir konu hakkında bilgi vermediği için bu kısıtlamaya gidilmiştir. Uygulamada başlık kelimelerinin nasıl girileceği Şekil 4.2'de gösterilmiştir.



Şekil 4.2 Başlık kelimeleri seçim paneli.

Yüksek frekans özelliği her kelimenin metinde geçme sıklığını ifade eder. Bu şekilde her kelimenin geçtiği frekans listesi oluşturulur. Artan sıralama ile sıralanır. Bu listeye günlük hayatta çok kullanıldığı halde bilgi içermeyen "ve", "veya", "ya da" vb. ifadeler eklenmez. Liste oluşturulduktan sonra en yüksek frekansa sahip kelimelerin %10'u alınıp, puanlamaya dâhil edilir. Bu çalışmada frekans listesine dâhil edilmeyen kelime listesi "ve", "veya", "ile", "için", "gibi", "bu", "ama", "fakat", "çünkü", "bir", "de" sözcükleridir.

Giriş puanı, kelimelerin giriş paragrafına ait olma yetisine göre verilen puandır. Aynı şekilde sonuç puanları da kendi paragrafında olma durumuna göre puanlandırılır. Bir metinde genel itibariyle verilen bilgilerin bir özeti ya giriş ya da sonuç paragraflarına yerleştirilir.

Anahtar kelimeler özelliği kullanıcının uygulama başlatılmadan önce o metin hakkında önemli olduğunu düşündüğü kelimeleri girebileceği bir alandır. Örneğin bir

futbol maçı haberinde anahtar kelimeler futbol, maç, spor, skor, kart gibi kelimeler önceden girilerek bu kelimelerin geçtiği cümlelerin puanı artırılmaya çalışılır, böylelikle özette yer alma ihtimali artırılır. Anahtar kelime giriş paneli Şekil 4.3'te gösterilmiştir.

Şekil 4.3. Anahtar kelimeler giriş paneli.

Özel isimler özelliğinde metin içindeki her özel isimin puanı 10-ile 20 arasında herhangi bir tamsayı puanı alacaktır.

Pozitif ve negatif kelime özelliklerinde, örneğin bir metinde “özetle”, “sonuçta”, “neticede” gibi kelimelerden sonra o konunun kısa bir özeti anlatılacağından daha fazla puan verilmiştir. Aynı şekilde "ama", "ancak", "fakat" gibi kelimelerden sonra anlatılan konunun olumsuz veya olamayacağı yanlarından bahsedilir ki bu da özette yer almaması için daha az puan verilmiştir.

Sayı ve tarih ifadesi cümle içinde geçen sayı ve tarihlere pozitif ayrımcılık uygulamak, özette yer ama kapasitesini artırmak amacıyla olumlu anlamda puanlandırılır.

Çift tırnak ifadeleri bir konuşmanın alıntısı olduğundan veya vurgu yapmak kullanıldığından, önemlidir ve özet için güçlü aday konumundadır.

Bitiş işareti özelliğine verilen puanlar cümlenin bitişinde konulan ünlem, soru işareti, nokta gibi işaretlerin tespiti ile olur. Bir cümle ünlem işareti ile bitmişse bir konu vurgulanmak istenmiştir ve pozitif puanlandırılmıştır. Aynı şekilde soru işareti de nokta işaretine göre daha önemlidir ve o cümleyi herhangi bir cümle olmaktan çıkarır. Bu nedenle ünlem ve soru işareti ile bitmiş cümlelerin puanı artırılmıştır.

Kelimelerin puanlaması tamamlandıktan sonra bu puanlar toplanmıştır. Toplam puan, her bir cümlenin puanını ifade eder.

Bu işlemlerden sonra cümleler puanları ile birlikte azalan sıralama ile sıralanmıştır. Önceden belirlenen ve parametrik olan bir oran ile bu cümlelerden en yüksek puana sahip olanları seçilmiştir. Bu orana özetleme yüzdesi denir. Örneğin 100 cümle içeren bir metinde özetleme yüzdesi 40 ise, 40 adet en iyi puana sahip cümlenin seçimi anlamına gelir. Bu çalışmada özetleme yüzdesi varsayılan olarak %35'tir. Çalışmada kullanılan veri setinde, ideal özet olarak deneklere metinlerdeki cümlelerin %35'inin seçilmesi istendiğinden sistem de varsayılan olarak metnin %35'ini özet olarak kabul etmiştir. Uygulama arayüzünde özetleme yüzdesi kullanıcıya bırakılmıştır. Seçilen cümleler bir araya getirilip sistem özetini oluşturmuştur. Daha sonra bu cümleler önceden çeşitli insanlar tarafından çıkarılan özetteki cümleler ile birebir karşılaştırılmıştır. Sistemin ortaya çıkardığı özet ile ideal özetteki çakışan cümle sayısının oranı başarıyı vermiştir. Örneğin sisteme 10 cümlelik bir metin girdi olarak alındığını varsayımıyla, ideal özette yani insanlar tarafından oluşturulan özette bu cümlelerden üçüncü, beşinci ve yedinci cümle özette yer almış olsun. Geliştirilen yazılım da girdi olarak alınan metinden dördüncü, beşinci ve yedinci cümlenin özette yer almasına karar vermiş olsun. Senaryoda özet olarak 3 cümle seçilmiş durumdadır ve bunlardan beşinci ve yedinci cümleler çakışmaktadır. Yani sistem, ideal özetin üç cümlesinden iki cümlesini tahmin edebilmiştir. Bu durumda %66,6 gibi bir başarı elde edilmiş olur.

Bundan sonra devreye genetik algoritma girmiştir. Buraya kadar popülâsyondaki bir bireyin başarısı hesaplanmaktadır. Her bir 12 özelliği barındıran özellikler kombinasyonu bir bireyi temsil eder. 10 adet bireyin başarısı ölçüldükten sonra bu popülâsyon ilk nesil sayılır. İkinci nesilde işlem devem ederken birinci nesildeki başarılı bireyler girdi olarak alır.

Şekil 4.4. Genetik algoritma parametre giriş paneli.

Genetik algoritmada seçme yöntemi olarak rank tercih edilmiştir. Seçme işleminin ardından çaprazlama ve mutasyon işlemleri uygulanıp sistem başarısı test edilmiştir (Bkz. Şekil 4.4). Her bir nesilde en başarılı birey (bireyin genleri) bir sonraki nesle aktarılmıştır. 20 adet nesil çalıştırıldıktan sonra en başarılı özellikler kümesinin sahip olduğu birey, başarıyı yükseltmiş olan yeni özellik kümesidir.

Anlatılan işlemler C# programlama dili ile gerçekleştirilmiştir. Uygulama yazılımının kullanıcı arayüzü Şekil 4.5'te verilmiştir.

Şekil 4.5. Geliştirilen yazılımın kullanıcı arayüzü.

4.2. Sonuca Etki Eden Faktörler

Özetlemede sonuca birçok faktör etki etmektedir ve sonucun başarısını düşürmekte veya yükseltmektedir. Bu faktörlerden en önemlisi özeti çıkarılan metnin mutlak ve en başarılı özet olup olmadığını belirleyecek bir kriterin olmamasıdır. Sistem tarafından bir özet çıktısı üretilmekte ve sistemin çıkardığı özetin başarılı olup olmadığı test edilmektedir. Bunu gerçekleştirmek için sistemin çıkardığı özeti insanlar tarafından çıkarılan ve ideal diye tabir edilen özetle karşılaştırmak gerekmektedir. Bu noktada şu sorunun cevaplandırılması gerekmektedir: test etmek için kullanılan ve insanlar tarafından çıkarılan metin, ideal özeti ne kadar temsil ediyor? Her insan için bir metinde geçen cümlenin önemi farklı olmakla beraber bu söz konusu deneğin yetkinliğine hatta o anki psikolojik durumuna göre farklılık gösterebilir. Bu da çok

başarılı kurgulandığı iddia edilen özet sistemlerinin başarısında azımsanmayacak değişikliklere neden olabilir.

Özet çıkarma sistemleri bir uygulama yazılımı yardımı ile gerçekleştirilir. Bu tezde olduğu gibi çıkarıma dayalı özetleme sistemlerinde başlık kelimeleri, kullanıcının sistem başlatılmadan girdiği anahtar kelimeler gibi parametreler de özetin başarısına olumlu veya olumsuz katkı sunar. Örneğin bir futbol maçını anlatan metinde karşılaşmanın sonuç skoru ve istatistikler önemlidir. Kullanıcı bu konu için anahtar kelime olarak "skor", "futbol", "asist" gibi kelimeleri önceden uygulamaya dahil edip, içinde bu kelimelerin geçtiği cümlelerin özet yer alma ihtimalini arttırabilir bu da başarıyı arttırabilir. Aynı şekilde yine aynı metin ele alındığında kullanıcı futbol veya sporla alakası olmayan kelimeleri anahtar kelime olarak seçtiğinde, seçtiği kelimeleri içeren cümlelerin özette geçme olasılığını artırmış olacaktır böylelikle özetin başarısını düşürmüş olacaktır.

Çıkarımsal özetleme yöntemlerinde konu hakkında önemli olduğu tespit edilen kelimelerin tespit edilme yöntemleri de özetin başarısına yön verir. Bu kelimeler tespit edildikten sonra kelimelere verilecek puanın ne olacağına genelde sezgisel olarak karar verilir. Sezgisel verilen bu kararlar da sonucu etkiler. Sezgisel puanlama yapılan bir uygulamada genetik algoritma gibi yöntemler de sisteme dâhil edildiğinde sonuca etki edecek olan başka bir faktör daha ortaya çıkar. Genetik algoritma yöntemlerinde çaprazlama, seçme veya mutasyon yapabilmek için rastgele sayılara ihtiyaç duyulur. Uygulama yazılımının o an üreteceği rastgele sayılar da sonuca etki eder. Dahası en başarılı bireyin arandığı arama uzayının seçimi de yine sonucu değiştirebilir.

4.3. Teze Katkı Sunan Yenilikler

Çıkarımsal özet çıkarma metodunda özeti çıkarılacak metnin yapısal özelliklerine bakılır. Kelimelerin öneminin belirlenmesi, cümlelerin konumu, uzunluğu paragrafların, giriş, gelişme ve sonuç cümlelerinin tespiti gibi özellikler bunlardan bazılarıdır. Literatürde kullanılan özelliklerden biri de ortalama uzunluktur. Metinde yer alan cümlelerin kelime sayılarının ortalama uzunluğunu hesaplanır, ortalamaya en yakın cümleler özete dâhil edilmeye çalışılır. Bu tezde ortalama uzunluğa farklı bir bakış açısı

geliştirilmiştir. Bir örnekle açıklamak gerekirse; özeti çıkarılacak olan metinde çok uzun bir cümle olduğunu varsayılsın, yaklaşık bir sayfalık bir cümle olsun ve içinde metinde yer almasına gerek olmayan bilgiler olsun. Bu durumda örnek cümle özette yer alma ihtimali çok yüksektir çünkü fazla kelime barındırdığı için kelimelerin toplamı cümle puanını artırabilecektir. Bu çelişkinin önüne geçmek için bu tez kapsamında yapılan bir yenilik olarak ortalama uzunluk için geliştirilen yöntemde, her cümlenin kelime sayısına göre ortalaması alınır ve bu sayı kesim noktası olur. Ortalama alındıktan sonra o cümleye ait kelimelerden puanı en yüksek olana göre azalan sıralama ile kelimeler sıralandıktan sonra ortalama kesim sayısına kadar olan kelimelerin puanları toplanır. Bu toplam puan, cümlenin puanı olur. Her cümle için bu işlem yapılır. Böylece bir cümlenin gereksiz olduğu halde sadece uzun olduğu için özette yer alma ihtimali azaltışmış olur.

Sezgisel belirlenen 12 farklı özelliğe göre bir aralık kümesinden bahsedilmişti. Çizelge 4.1'de gösterildiği gibi, ilk çizelgedeki (Bkz. Çizelge 2.1) puanlamaya yine sezgisel olarak sabit değer yerine tamsayı olacak şekilde sayı aralığı verilmiştir. Bu aralıkta rastgele değerler üretilerek sonuca gidilmiştir ve sonuçta sistemin başarısının arttığı gözlemlenmiştir. Sezgisel yönteme ek olarak sıfır ile yüz (0-100) arasında tamamen rastgele tamsayı değerleri üretilerek sistemin başarısının test edilmesi ise bu tezde denenen farklı bir yöntemdir. Çizelge 4.1'deki bu aralık kümesi yerine her özelliğe sıfır ile yüz (0-100) arasında tamsayı değerleri uygulama yazılımı tarafından rastgele verilip sistemin başarısı test edilmiş ve başarının arttığı gözlemlenmiştir.

Başlık kelimelerini içeren cümlelerin önemli olduğu ve puanlamaya dâhil olduğu daha önce anlatılmıştı. Kurgulanan sistemde bir cümle, metnin başlığında geçen kelimelerden birini veya birkaçını ya da hepsini içeriyorsa o cümlenin özette yer alma ihtimali daha yüksek olur. Ancak başlık kelimeleri "ve", "veya", "gibi" vs. kelimeleri de içeriyorsa bu kelimeler de önemli hale gelmiş olacaktır. Bu sorunun önüne geçebilmek için başlıkta geçen ve kullanıcının sezgisel olarak sadece gerekli gördüğü kelimeleri seçmesi ve o kelimelerin puanlandırılması sağlanmıştır. Bunu için uygulama arayüzüne kullanıcının kelime seçimi yapabileceği bir panel eklenmiştir. Böylece özet için başlık kelimeleri daha anlamlı hale gelmiş olur.

5. DENEYSEL SONUÇLAR

Bu tez çalışmasında sistemin başarısının test edilmesi için kesinlik ölçüm değeri kullanılmıştır. 20 adet haber metni 30 adet erkek ve kadından oluşan kişilere, metinlerden önemli buldukları cümlelerden %35 oranında seçim yaparak özet çıkarmaları istenmiştir. Veri setinin detaylarına daha önce değinilmişti. Bu tez kapsamında bu veri seti kullanılmıştır. İnsanlar tarafından çıkarılan özet ideal özet olarak kabul edilip çıkarılan bu özetler, sistem tarafından çıkarılan özetle her bir cümlesi birebir eşleştirilerek karşılaştırılmıştır.

Tasarlanan sistemde cümlelere puanlar verilmektedir. En yüksek puana sahip cümleler azalan sıralama ile sıralanır ve ideal özetteki özetleme yüzdesi ne kadar ise o oranda bir kesme değeri belirlenir ve en yüksek puana sahip cümleler o kesme değerinden itibaren alınır. Elde edilen cümleler sistemin özetini ifade eder. Sistemin özeti de ideal özetteki cümlelerle karşılaştırılır ve başarı ölçülür.

Cümlelerin puanını elde etmek için birtakım özelliklere sahip olan kelimeler puanlandırılır ve kelimelerin puanlarının toplamı cümlenin puanını ifade eder. Hangi kelimelere ne kadar puan verileceğine ile ilgili yapılan çalışmalarda sezgisel yöntemler ön plana çıkmaktadır. Çizelgede gösterildiği gibi (Bkz. Çizelge 2.1) kelimelere puan verilirken sezgisel yöntemler kullanıldığı görülmektedir.

Sistemin başarısını etkileyen faktörlerden yukarıda bahsedilmişti. Buna göre test aşamasında uygulama arayüzündeki parametrik değerlerin nasıl girildiği de başarıyı etkileyecektir. Bu tez kapsamında ideal özeti çıkaran kişi, özetleme yüzdesi ve genetik algoritma parametreleri Çizelge 5.1'de verilmiştir. Sabit tamsayı değerli, sezgisel sınırlı aralıklı rastgele değerler ve genel aralıklı rastgele değerler olmak üzere üç ayrı değerlendirme yapılmış olup tüm bu değerlendirmedeki ortak parametrik değerler Çizelge 5.1'de gösterilmiştir. Test aşamasında girilen anahtar kelimelerin ne olduğu ayrıntılı çizelgelerde her metin için ayrı ayrı verilmiştir.

Çizelge 5.1. Uygulamaya arayüzüne girilen parametreler

PARAMETRE ADI	PARAMETRE DEĞERİ
İdeal Özet Çıkaran Kişi	Banu DİRİ
İdeal Özet Yüzdesi	% 35
Sistem Özet Yüzdesi	% 35
Popülasyondaki Birey Sayısı	10
Genetik Algoritma Mutasyon Parametresi	0,09
Genetik Algoritma Durdurma Kriteri	20 nesil

5.1. Sabit Puanlı Değerlendirme

Bu tez kapsamında çizelgedeki (Bkz. Çizelge 2.1) kelime özelliklerine verilen puanlar dâhilinde sistem, veri seti ile test edilmiştir.

Çizelge 5.2. Sabit puan aralıklı sonuçlar

METİN	BAŞARI (%)
1	50
2	50
3	40
4	66.667
5	33.333
6	0
7	50
8	60
9	25
10	50
11	25
12	50
13	50
14	60
15	0
16	25
17	60
18	33.333
19	50
20	66.667
ORTALAMA	42.25

Uygulamada veri setinde bulunan 1 kişinin çıkardığı 20 adet haber metnine ait özet, sistemin çıkardığı özetler ile karşılaştırılmıştır. Sonuçlar Çizelge 5.2'de listelenmiştir.

5.2. Sezgisel Olarak Belirlenen Rastgele Puan Aralıklı Değerlendirme

Çizelge 4.1'deki aralık kümesinde aynı aralıktaki özelikler gruplanarak Çizelge 5.3 elde edilmiştir ve aşağıda gösterilmektedir.

Çizelge 5.3. Özelliklere verilen puan aralıkları

Özellikler	Puan Aralıkları
Başlık, Giriş, Anahtar Kelimeler, Pozitif Kelimeler	30-20
Yüksek Frekans, Sonuç, Özel İsimler	10-20
Negatif Kelimeler, Sayı, Çift tırnak İşareti, Bitiş İşareti, Tarih	0-10

Belirlenen özelliklerdeki kelimelere sezgisel sabit tamsayı değerlerinin verildiği puanlamaya Çizelge 5.3'te gösterildiği üzere, yine sezgisel olarak bir aralık kümesi verilerek uygulama test edilmiştir. Dikkat edileceği üzere bu testin bir öncekinden farkı özellikleri belirlenen kelimelere puan verilirken sabit bir tamsayı değeri yerine değer aralığı verilip programın o değer aralığında rastgele tamsayı üreterek sonuca gitmesidir. Veri setinde bulunan 1 kişinin (B. D.) çıkardığı 20 adet haber metnine ait özet, sistemin çıkardığı özetler ile karşılaştırılmıştır.

Çizelge 5.4 incelendiğinde, her bir satır, verilen tamsayı aralıklarında genetik algoritma süreçlerinden geçen ve 20 nesil sonucunda ortaya çıkan en başarılı bireye ait kelime özelliklerini ve elde edilen başarıyı gösterir. Her satır, belirli bir yazara ait farklı birer haber metninin özetlerinin, sistem özetiyle karşılaştırılması ile oluşmuştur. Kelime özelliklerine sezgisel olarak belirlenen aralıklarda rastgele tamsayı değerleri verilerek gerçekleştirilen bu süreçte, sonuçları bir tabloda toparlanırsa Çizelge 5.5'teki gibi bir sonuç ortaya çıkar.

Çizelge 5.4. Sezgisel aralıklı değerlendirmenin ayrıntılı sonuçları

Metin Numarası	Seçilen Anahtar Kelimeler	Başlık Kelimeleri Puanları	Giriş Paragrafı Kelime Puanları	Pozitif Kelime Puanları	Anahtar Kelime Puanları	Sonuç Paragrafı Kelime Puanları	Kelime Frekansı Puanları	Özel İsimlerin Puanları	Negatif Kelime Puanları	Bitiş İşareti Puanları	Sayı Bilgisi Puanları	Tarih Bilgisi Puanları	Çift Tırnak İşareti Puanları	Başarı (%)
1	para, servet, nakit	25	25	24	22	16	16	14	3	4	3	8	3	**50**
2	milyar, Arap, dolar	20	21	21	24	18	14	13	2	5	7	8	8	**75**
3	Tarkan, acı, gün	26	21	27	24	12	13	17	3	6	8	7	8	**60**
4	Halis, Toprak, köşk	25	24	25	26	13	17	14	7	7	2	4	4	**66.667**
5	Arda, futbol, maç	23	22	25	26	16	15	14	4	4	8	3	3	**33.333**
6	şiddet, tecavüz	23	26	26	25	11	17	11	3	7	3	6	0	**25**
7	Van, yardım	23	22	25	25	13	12	16	6	5	5	4	7	**50**
8	düzenleme, afet, hasar	25	22	23	26	16	13	12	2	1	4	3	3	**60**
9	şike, rüşvet, suç	27	22	22	22	16	18	16	7	4	9	2	3	**50**
10	borç, kredi	25	27	25	24	12	19	18	4	5	9	5	2	**50**
11	devrik, cenaze	24	25	23	24	16	15	13	5	2	4	4	9	**25**
12	kanser, moral, tedavi	22	26	20	22	13	12	13	8	8	1	5	7	**50**
13	kitap, Apple	25	25	23	24	15	15	11	3	8	7	7	4	**50**
14	Hakkari, terör	24	22	22	26	15	15	12	3	5	6	6	0	**60**
15	adalet, mahkeme	22	24	29	28	16	16	17	5	7	7	0	5	**33.333**
16	voleybol, spor	22	22	20	24	15	12	13	5	3	6	7	5	**50**
17	fon, altın	22	23	24	23	13	10	13	4	4	3	7	4	**60**
18	araştırma, sağlık, şeker, su	22	29	27	22	19	15	16	4	4	4	5	7	**33.333**
19	öğrenci, kavga, yayın	22	26	27	22	14	14	12	6	3	6	1	3	**75**
20	eylem, PKK, terör	23	29	23	25	14	10	14	5	6	2	3	4	**66.667**

Çizelge 5.5. Sezgisel aralıklı puanlamanın genel sonuçları

METİN	BAŞARI (%)
1	50
2	75
3	60
4	66.667
5	33.333
6	25
7	50
8	60
9	50
10	50
11	25
12	50
13	50
14	60
15	33.333
16	50
17	60
18	33.333
19	75
20	66.667
ORTALAMA	51.166

Çizelge 5,4'te kelimeler sezgisel sabit tamsayı puanları alırken başarı ortalaması % 42.25 olmuştur. Kelime özelliklerine, sezgisel olarak bir aralık değeri belirlenip bu aralıkta rastgele tamsayı değerleri verildiğinde sistemin başarısı % 51.166 olmuştur. Başarıda bir artış söz konusudur.

5.2. Genel Aralıklı Rastgele Puanlama İle Yapılan Değerlendirme

Sezgisel aralıklı rastgele tamsayı değerleri verilerek elde edilen sonucun, sabit puanlamadan daha başarılı olduğu gözlemlenmiştir. Sistem genetik algoritmanın gücüyle belirlenen arama uzayında daha başarılı sonuç üretmiştir.

Bu yöntemde, aynı parametrik değerler ile arama uzayı genişletilerek başarının durumu test edilmiştir. Genetik algoritmaların sağladığı avantajların da desteği ile daha geniş bir aralıkta başarının artırımı sağlanmaya çalışılmıştır. Buna göre, belirlenen

kelime özelliklerine daha geniş bir puan aralığı ile rastgele tamsayı değerleri verilerek sonuca ulaşılmıştır. Puan aralığı 0 ile 100 arasında rastgele tamsayı değerleri olarak belirlenen test sonuçları Çizelge 5,6'da verilmiştir.

Sezgisel olarak belirlenen tamsayı aralı klı puanlama yönteminde olduğu gibi bu yöntemde de her satır ayrı bir metne ait sonuçları göstermektedir. Aynı genetik süreçten geçen bu yöntemin bir önceki yöntemden farkı ilk popülâsyonda kelime özelliklerine 0 (sıfır) ile 100 (yüz) arasında rastgele tamsayıların verilmesidir. Genel aralıklı rastgele puanlama yönteminde 20 adet test sonucunun ortalaması alındığında, sezgisel aralıklı puanlamadan daha başarılı sonuç elde edildiği gözlemlenmiştir. Genel ağırlıklı puanlama sonuçları Çizelge 5.7'de gösterilmektedir. Başarı ortalaması % 59.25'e yükselmiştir.

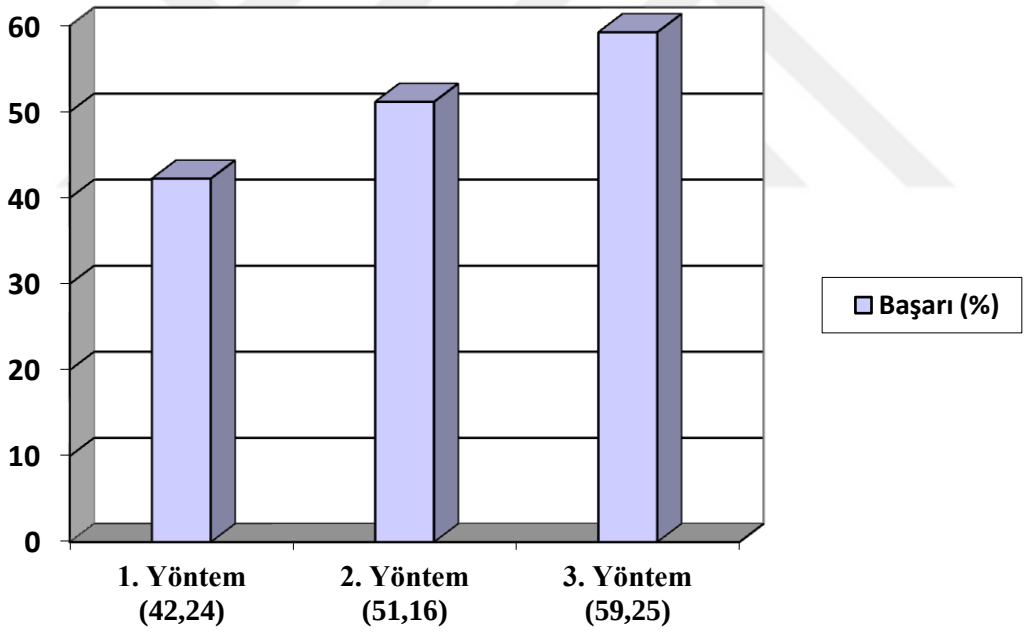
Çizelge 5.7. Genel aralıklı puanlamanın genel sonuçları

METİN	BAŞARI (%)
1	50
2	75
3	60
4	66.667
5	66.667
6	25
7	75
8	60
9	50
10	75
11	50
12	50
13	50
14	80
15	33.333
16	50
17	60
18	66.667
19	75
20	66.667
ORTALAMA	59.25

Çizelge 5.6. Genel ağırlıklı değerlendirmenin ayrıntılı sonuçları

Metin Numarası	Seçilen Anahtar Kelimeler	Başlık Kelimeleri Puanları	Giriş Paragrafı Kelime Puanları	Pozitif Kelime Puanları	Anahtar Kelime Puanları	Sonuç Paragrafı Kelime Puanları	Kelime Frekansı Puanları	Özel İsimlerin Puanları	Negatif Kelime Puanları	Bitiş İşareti Puanları	Sayı Bilgisi Puanları	Tarih Bilgisi Puanları	Çift Tırnak İşareti Puanları	Başarı (%)
1	para, servet, nakit	75	26	12	16	47	29	60	75	78	33	27	71	50
2	milyar, Arap, dolar	87	43	8	75	41	46	35	10	55	65	35	14	75
3	Tarkan, acı, gün	54	15	48	59	25	27	47	33	16	41	38	24	60
4	Halis, Toprak, köşk	60	82	10	74	38	40	4	75	62	23	42	42	66.667
5	Arda, futbol, maç	31	60	74	16	63	65	22	40	84	22	17	55	66.667
6	şiddet, tecavüz	12	99	67	61	63	77	39	37	41	35	35	56	25
7	Van, yardım	9	63	32	46	64	29	58	35	54	64	69	2	75
8	düzenleme, afet, hasar	48	37	65	17	76	44	40	60	40	69	65	17	60
9	şike, rüşvet, suç	32	53	42	48	21	51	73	57	81	33	24	62	50
10	borç, kredi	12	71	55	63	58	14	93	35	53	71	84	0	75
11	devrik, cenaze	58	18	46	53	56	81	42	73	34	33	30	87	50
12	kanser, moral, tedavi	19	60	34	29	29	47	85	64	65	31	41	24	50
13	kitap, Apple	41	81	39	57	26	51	68	61	69	51	37	44	50
14	Hakkari, terör	28	40	44	57	34	57	63	81	54	48	53	52	80
15	adalet, mahkeme	29	50	12	24	58	27	44	68	76	57	52	22	33.333
16	voleybol, spor	56	83	38	20	90	41	79	68	35	60	39	19	50
17	fon, altın	28	79	66	28	33	60	72	45	58	66	24	59	60
18	araştırma, sağlık, şeker, su	45	64	73	9	16	9	45	58	43	50	48	40	66.667
19	öğrenci, kavga, yayın	83	29	61	49	36	43	60	84	74	9	78	52	75
20	eylem, PKK, terör	74	44	13	53	77	44	45	26	17	79	61	17	66.667

Bu tezde 3 ayrı yöntem denenmiş olup başarıları test edilmiştir. Birincisi önceden belirlenen 12 adet özelliğe sahip kelimelere, puanlar verilirken daha önceki çalışmalarda kullanılan sezgisel yöntem denenmiştir ve sonuçta başarı % 42,24 olarak ölçülmüştür. İkinci yöntemde, sezgisel olarak belirlenen puanlar yerine yine sezgisel olarak belirlenen bir aralık dâhilinde, uygulama programının rastgele tamsayı puanları üretmesi ve sürece genetik algoritmanın dâhil edilmesiyle elde edilen sonuçta başarı % 51,16 düzeyine çıkarılmıştır. Üçüncü yöntem olarak programın belirli özellikteki kelimelere, bir önceki yöntemde belirlenen sezgisel aralık yerine gruplama yapmadan geniş bir aralıkta rastgele tamsayı puanları vermesi sağlanmıştır ve başarı % 59,25 olarak ölçülmüştür. Bu başarı değişim oranı aşağıda Şekil 5.1'de verilmiştir.



Şekil 5.1. Tez kapsamında test edilen 3 ayrı yöntemin başarı grafiği.

6. TARTIŞMA VE SONUÇ

Bu tez çalışmasında cümle seçerek otomatik metin özetleme konusuna değinilmiştir. Bu konuda yapılan önceki çalışmalarda kullanılan veri seti kullanılmıştır. Önceki çalışmalarda kullanılan, cümlelere puan vererek en yüksek puana sahip cümlelerin seçimine dayalı yöntem geliştirilmiştir. Bu bağlamda belirli özelliklere göre, kelimelere sezgisel olarak puan verilip elde edilen toplam puan cümle puanı olarak tanımlanmaktadır. Literatürdeki çalışmalarda sezgisel olarak değerlendirilen kelimelerin puanları, bu çalışmada ek olarak genetik algoritma süreçlerinden geçirilerek başarının artırımı sağlanmaya çalışılmıştır. Bunun için önceden belirlenen 12 ayrı özellikteki kelimelere puanlar verilmiş olup ölçümler yapılmıştır.

Başarıyı test etmek için bir uygulama programı gerçekleştirilmiştir. Program, daha önceki çalışmalarda kullanılmış ve başarısı test edilmiş bir veri setini kullanıp, sezgisel ve genetik algoritma yöntemlerini birlikte gerçekleştirip ortaya bir özet metin çıkarmaktadır. Veri setinde haber metinleri ile beraber bu metinlerin insanlar tarafından çıkarılan özetleri bulunmaktadır. Programın çıktısı sistem özeti, insanların çıkardığı özet ise ideal özet olarak bilinir. Program, sistem özeti ile ideal özeti literatürde bulunan başarı ölçüm yöntemlerinden biri ile karşılaştırmaktadır. Başarı ölçümü için sistem özeti ile ideal özetin çakışan cümle sayısının oranı esas alınmıştır.

Deneysel ölçümler sonrasında başarının arttığı gözlemlenmiştir. İlk yöntemde cümlelerdeki kelimelere sabit sezgisel puanlar verilmiş olup sonuç % 42.24 çıkmıştır. Kelimelerin özelliklerinin sezgisel bir puan aralığı dâhilinde rastgele sayılar verdiği ikinci yöntemde başarının % 51.16 olduğu gözlemlenmiştir. Buradaki başarı artış oranı % 21.07'dir. Üçüncü ve son yöntemde kelimelere herhangi bir sınıflandırma yapmadan sezgisel olarak geniş bir değer aralığında puanlar verilmiş ve başarı % 59.25 olarak test edilmiştir. Bu durumda ikinci yönteme göre % 17.39 oranında başarı artışı sağlanmıştır.

Bu tez kapsamında, cümle seçerek metin özetleme yöntemi kullanılarak cümle puanlandırmaya dayalı uygulamalardaki genel sorunlara değinmekte yarar vardır. Bunlardan birincisi sistemin çıkardığı özeti karşılaştıracak olan ideal özetin mutlak özet olup olmadığının belirlenmesi için bir standardın olmamasıdır. İnsanlar özet çıkarırken bilgi seviyesinden sosyo-ekonomik durumuna kadar bütün kişisel özellikleri o özetin

oluşumuna etki etmektedir. Her insan bir metinden farklı özetler çıkarabilir. Bu da sistemin başarısını etkiler. Bu sorunu azaltmak için daha fazla denekten elde edilecek özetle karşılaştırma yapılabilir. Bu noktada İngilizce dilince yapılan çalışmalarda daha fazla veri seti bulmak mümkündür. Ancak Türkçe için çalışılmış ve başarısı ölçülmüş çalışmalarda kullanılan veri setinin az olması nedeniyle özetleme konusunda özellikle Türkçe dili açısından çok hızlı bir ilerleme kaydedilememektedir.

Kelimelere puan verilirken kelimelerin kökleri ele alınmıştır. Türkçe dilindeki ek yapısı özellikle kelimenin anlamını değiştiren ekler açısından ele alındığında özetleme konusu için bir sorun teşkil etmektedir. Açık kaynak kodlu uygulama programları, sesteş kelimelerin türünün ne olduğunu belirleyebilmesi konusunda tam başarı sağlamış durumda değildir. Örneğin Zemberek açık kaynak kodlu program henüz tam anlamıyla kelimeleri yapısal ve anlamsal olarak irdeleyememektedir. Bu da Türkçe metinler için gerçekleştirilecek olan özetleme çalışmalarının ilerlemesi açısından bir sorun olarak karşımıza çıkmaktadır.

Bu tezde elde edilen başarı sistem özeti ile ideal özetin cümlelerinin birebir karşılaştırılması sonucu ortaya çıkan tabloyu yansıtmaktadır. Başarı ölçme yöntemi iki ayrı çıktının karşılaştırılmasından ibarettir. İleriye yönelik yapılacak olan çalışmalarda cümle seçerek metin özetleme sistemlerinin başarısının artırılmasına katkı sunmak amacıyla başarıyı ölçme konusunda bulanık mantık yöntemleri denenebilir.

KAYNAKLAR

Al-Saleh A.B., Menai,M.E.B., 2016. Automatic Arabic text summarization: a survey. ***Artificial Intelligence Review***, **45**(2):203-234.

Alguliev, R.M., Aliguliyev, R.M., Hajirahimove, M.S. Mehdiyev, C., 2011. MCMR: Maximum coverage and minimum redundant text summarization model, ***Expert Systems with Applications***, **38**(12):14514-14522.

Altan, Z., 2004. A Turkish Automatic Text Summarization System, ***IASTED International Conference on AIA***, Innsbruck, Austria.

Ayush A., Utsav G., 2014. Extraction based approach for text summarization using k-means clustering, ***International Journal of Scientific and Research Publications***, **4**(11):2-3.

Berker, M., 2011. ***Using Genetik Algorithms with Lexical Chains For Automatic Text Sumamrization*** (doktora tezi). Boğaziçi Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Berker, M. and Güngör, T., 2012. Using Genetic Algorithms with Lexical Chains for Automatic Text Summarization , ***4th International Conference on Agents and Artificial Intelligence***, Vilamoura, Portugal.

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***.

Bilgin, Ş.B., 2012. ***Multi-Document Sumamrization Suing Dependency Grammars***, (doktora tezi). Boğaziçi Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

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.

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.

Çakır M., Çelebi E., 2011. ***Kapsama Katsayısı Tabanlı Kümeleme İle Belge Özetleme, IEEE 19th Signal Processing and Communications Applications Conference*** (SIU 2011), Berlin, 186-189.

Çığır, C., Kutlu, M. and Cicekli, I., 2009. ***Generic Text Summarization for Turkish***, The Computer Journal, **53**(8):1315-1323.

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.

Dashpande, R.A., Lobo L. M. R. J., 2013. Text Summarization using Clustering Technique, ***International Journal of Engineering Trends and Technology (IJETT)***, **4**(8):3348-3351.

Dragomir, R., Hongyan J. and Malgorzata B., 2000. Centroid based summarization of multiple documents, ***In ANLP/NAACL Workshop on Automatic Summarization***, Seattle, USA.

Dmitry T., Mikhail P., and Igor M., 2013. Supervised and Unsupervised Text Classification via Generic Summarization, ***International Journal of Computer***

Information Systems and Industrial Management Applications, **5**(2013):509-515.

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

Filatova, E. and Hatzivassiloglou, V., 2004, A formal model for information selection in multi-sentence text extraction, ***20th International Conference on Computational Linguistics***, Ağustos 2004, Geneva. 23-27.

Gong, Y. and Liu, Xin., 2001. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis, ***In the proceeding of ACM SIGIR***, Eylül 2001, New Orleans, Louisiana.19-25.

Güran, A., Güler, N. and Bekar, E., 2013. Automatic summarization of Turkish documents using non-negative matrix factorization , ***Turkish Journal of Electrical Engineering and Computer Science***, **21**(5):1411-1424.

Güran, A., 2013, ***Otomatik Metin özetleme Sistemi*** (doktora tezi). Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Güran A., Arslan S. N., Kılıç E., Diri B., 2014 Sentence selection methods for text summarization, ***IEEE 22nd Signal Processing and Communications Applications Conference***. 23-25 Nisan 2014,Trabzon. 192-195.

Gönenç, E. 2012 ***Lexical Cohesion Analysis For Topic Segmentation,Sumamrization and Keyphrase Extraction*** **(**doktora tezi, yayınlanmamış). Bilkent Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

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 Şubat 2009, Mexico.

Kılcı, Y. and Diri,B., 2008. ***Turkish Text Summarization System***, Senior Project, Yıldız Technical University, Turkey.

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.

Lee, J.H., Sun P., Chan-Min A. and Daeho K., 2009. ***Automatic generic document summarization based on non-negative matrix factorization***, **45**(1): 20-34.

Lin, C.Y., 2004. ROUGE: a Package for Automatic Evaluation of Summaries, ***In Proceedings of the Workshop on Text Summarization Branches Out***, 25-26 Temmuz, İspanya.

Liu, Maofu., Li, W., Wu, M. and Lu, Q., 2007. Extractive Summarization Based on Event Term Clustering***, In Proceedings of the ACL***, Haziran 2007, Prag.

Luhn, H.R., 1958. ***The automatic creation of literature abstracts***, ***IBM Journal of research Development***, **2**(2):159–165.

Manuel, J. And Moreno,T., 2015. Artex is AnotheR TEXt summarizer, ***International Journal of Computational Linguistics and Applications,*** **7**(1):67-83.

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***, **37**(6):299-305.

Morris, Andrew H., George M.K., and Dennis A. A., 1992. The Effects and Limitations of Automated Text Condensing on Reading Comprehension Performance, ***Information Systems Research***, **3**(1):17-35.

McDonald, R., 2007. A study of global inference algorithms in multi- document summarization, ***29th European Conference on IR Research***, 2-5 April 2007, İtalya.

SAMi M.V., Banu D., 2010. Web tabanlı otomatik özet çıkarma sistemi, ***Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu***, 21-24 Haziran, Kayseri.

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***. Eylül 2005, Portekiz.

Akin, A.A,. 2016. Nzemberek-Master Türkçe Kelimeleri Köklerine ve Eklerine Ayırma Programı, ***https://github.com/espindl/nzemberek.*** Erişim tarihi: 01.05.2016.

Özsoy, M., Çiçekli, İ., Alpaslan, F.N., 2010. Text Summarization of Turkish Texts using Latent Semantic Analysis, ***In Proceedings of the 23rd International Conference on Computational Linguistics***, Ağustos 2010, Çin.

Padmakumar, A., Eswaran, D., 2014. **Extractive Text Summarization using Latent Semantic Analysis Natural Language Processing**. Indian Institute of Technology, Madras.

Pembe, C. 2011. ***Automated Query-Biased and Structure-Preserving Document Summarization for Web Search Tasks*** (doktora tezi), Boğaziçi Üniversitesi, Türkiye.

Raj Kumar V.S, Chandrakala.D., 2016. *A survey on text summarization using oprimization algorithm****. Elk Asia Pasific Journal Of Computer Science And Informatiaon Ssystems***, **2**(1):2454-3047.

Rath, G.J., Resnick, A., and Savage, T.R., 1961. The formation of abstracts by the selection of sentences. ***Advances in Automatic Text Summarization, MIT Press***, **12**(2):139-143.

S.A.Babar, Pallavi D.Patil., 2014. Improving Performance of Text Summarization, ***International Conference on Information and Communication Technologies (ICICT 2014).*** In IEEE international conference on fuzzy systems, 16–21 Temmuz.

Saggion, H., Cunha,I., Manuel,J. And Moreno,T. Sanhuan, E., 2010. Multilingual Summarization Evaluation without Human Models, ***Proceedings of the International Conference on Computational Linguistics.*** Pekin,Çin.

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

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**:305-314.

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***, Vancouver.

Steinberger, J., 2007. ***Text Summarization within the LSA Framework*** (doktora tezi), University of West Bohemia, Çek Cumhuriyeti.

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.

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 Haziran. Çek Cumhuriyeti.

Taşkın,Ç., Emel, G.G., 2002 Genetik Algoritmalar ve Uygulama Alanları, ***Uludağ Üniversitesi, İktisadi ve İdari Bilimler Fakültesi Dergisi,*** **21**(1):129-152.

Quyang, Y., Li, W., Lu, Q. and Zhang, R., 2010. A study on Position Information in Document Summarization, ***In Procceding of Coling***.Çin.

Ulukbek Attokurov., 2004. ***Multi Document Sumamrization Using Distortion-Rate Ratio*** (Doktora tezi). İstanbul Teknik Üniversitesi, İstanbul.

Uzundere E., Dedja E., Diri B., Amasyali M. F., 2008. ***Türkçe haber metinleri için otomatik özetleme***, ***Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu***.

Wadhvani, R., Roy D., 2013. Text Summarization Using Tense Recognizing Identifier, (IJCSIS) ***International Journal of Computer Science and Information Securit,*** **11**(6):20-24.

Witte, R., Krestel, R. and Bergler, S., 2007. Generating Update Summaries for DUC, ***In the Document Understanding Workshop***, 26-27 Nisan, NewYork, USA.

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***, Ağustos 2007, Manchester.

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**(1):75–95.

Yilmaz, B., 2011. Dijital Kütüphane Becerileri Konusunda Türkiye'de Durum: AccessIT Projesi Çerçevesinde Bir Değerlendirme, ***Türk Kütüphaneciliği***, **25**(1):117-123.

ÖZGEÇMİŞ

1984 yılında Van'da doğdu. İlk, orta ve lise öğrenimini Van'da tamamladı. 2003 yılında öğrenime başladığı Fırat Üniversitesi Bilgisayar Mühendisliği Bölümü'nden 2008 yılında mezun oldu.2008 - 2010 yılları arasında bir yazılım şirketinde 2 yıl çalıştı. Ardından askerlik görevini tamamladı.2011 yılında halen devam ettiği Şırnak Üniversitesi Bilgi İşlem Daire Başkanlığı'nda Bilgisayar Mühendisi olarak çalışmaya başladı. 2014 yılında Van Tapu ve Kadastro 15. Bölge Müdürlüğü'ne geçiş yaparak bu Kurumda Bilgisayar Mühendisi olarak çalışmaya başladı ve hala çalışmaya devam etmektedir. 2015 yılında Yüzüncü Yıl Üniversitesi Elektrik-Elektronik Mühendisliği Anabilim Dalında Tezli Yüksek Lisans eğitimine başladı. Evli ve bir çocuk babasıdır.

T.C
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
LİSANSÜSTÜ TEZ ORİJİNALLİK RAPORU

Tarih: 31./05./2018

Tez Başlığı / Konusu:
GENETİK ALGORİTMA KULLANARAK CÜMLE SEÇME YAKLAŞIMI İLE OTOMATİK METİN ÖZETLEME

Yukarıda başlığı/konusu belirlenen tez çalışmamın Kapak sayfası, Giriş, Ana bölümler ve Sonuç bölümlerinden oluşan toplam ....45..... sayfalık kısmına ilişkin, ...31./..05./..2018.. tarihinde şahsım/tez danışmanım tarafından Turnitin.....intihal tespit programından aşağıda belirtilen filtreleme uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % ....1..... (...BİR.....) dır.

Uygulanan filtreler aşağıda verilmiştir:
- Kabul ve onay sayfası hariç,
- Teşekkür hariç,
- İçindekiler hariç,
- Simge ve kısaltmalar hariç,
- Gereç ve yöntemler hariç,
- Kaynakça hariç,
- Alıntılar hariç,
- Tezden çıkan yayınlar hariç,
- 7 kelimeden daha az örtüşme içeren metin kısımları hariç (Limit inatch size to 7 words)

Van Yüzüncü Yıl Üniversitesi Lisansüstü Tez Orijinallik Raporu Alınması ve Kullanılmasına İlişkin Yönergeyi inceledim ve bu yönergede belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini bilgilerinize arz ederim.

Tarih ve İmza

Adı Soyadı: ERCAN BAYDAR
Öğrenci No: 149101203
Anabilim Dalı: Elektrik Elektronik Mühendisliği
Programı: Tezli Yüksek Lisans
Statüsü: Y. Lisans ☑ Doktora ☐

DANIŞMAN ONAYI
UYGUNDUR
Doç. Dr. Rıdvan SARAÇOĞLU

ENSTİTÜ ONAYI
UYGUNDUR
Prof.Dr.Sual ŞENSOY
Enstitü Müdürü
(Unvan, Ad Soyad, İmza)