YÖNETİM BİLİŞİM SİSTEMLERİ DERGİSİ
http://dergipark.ulakbim.gov.tr/ybs

Yayın Geliş Tarihi: 16.11.2017
Yayına Kabul Tarihi: 23.11.2017
Online Yayın Tarihi: 20.12.2017

Cilt:3, Sayı:2, Yıl:2017, Sayfa:62-75
ISSN: 2148-3752

GENETIC ALGORITHM BASED SENTENCE EXTRACTION FOR AUTOMATIC TEXT SUMMARIZATION

Oğuz KAYNAR*, Yunus Emre IŞIK, Yasin GÖRMEZ,

Ferhan DEMİRKOPARAN

Cumhuriyet University, Turkey

Abstract: With the development of the Internet, the amount of data in the digital environment is continuously increasing. Especially with web 2.0 technology, as a result of sites which users are able to add new content such as wikipedia, blogs and social media sites, the amount of information on the internet is increasing both in number and size. Accessing the required information in a medium where there are so many data is a serious problem. Today's information age make it necessary to use automatic text summarization systems in many areas about information retrieval in order to access the searched information. In this study, text summarization methods based on sentence extraction are discussed, firstly features to represent sentences in document is extracted and then the effectiveness of these attributes on summarization is tried to be determined by using genetic algorithm. The data set used in the study consists of 120 documents containing Turkish news texts and their summaries. 80 documents are trained with the help of genetic algorithm and the best weight values for the attributes are determined, then 40 test documents are summarized with these weights and the results are compared with the original summaries.

Keywords: Genetic Algorithm, Sentence Extraction, Statistic Method, Text Summarization.

OTOMATİK METİN ÖZETLEME İÇİN GENETİK ALGORİTMA TABANLI CÜMLE ÇIKARIMI

Özet: İnternetin gelişmesiyle beraber dijital ortamda bulunan veri miktarı sürekli artış göstermektedir. Özellikle web 2.0 teknolojisiyle birlikte wikipedia, blog, sosyal medya gibi, kullanıcıların yeni içerik ekleyebildiği sitelerin artması sonucunda internet ortamındaki bilgi miktarının hem sayısı hem de büyüklüğü sürekli artarak devasa boyutlara ulaşmıştır. Verilerin bu kadar çok olduğu bir ortamda istenilen bilgiye ulaşmak ciddi bir problemdir. Günümüz bilgi çağı, aranan bilgiye daha çabuk ve hızlı erişmek için otomatik metin özetleme sitemlerinin bilgi çıkarımı ile ilgili birçok alanda kullanımını zorunlu hale getirmektedir. Bu çalışmada cümle çıkarımına dayalı metin özetleme yöntemleri ele alınmış, ilk olarak doküman içerisinde yer alan cümleleri temsil edecek öznitelikler çıkarılmış, ardından bu

özniteliklerin özet oluşturmadaki etkinliği genetik algoritma yardımıyla belirlenmeye çalışılmıştır. Çalışmada kullanılan veri seti Türkçe haber metinleri ve bunların özetlerini içeren 120 dokumandan oluşmaktadır. 80 adet dokuman genetik algoritma yardımıyla eğitilerek, özniteliklere ilişkin en iyi ağırlık değerleri belirlenmiş, daha sonra bu ağırlıklar yardımıyla 40 adet test dokümanı özetlenmiş ve sonuçlar orijinal özetlerle karşılaştırılmıştır.

Anahtar Kelimeler: Genetik Algoritma, Cümle Çıkarımı, İstatistiksel Metotlar, Metin Özetleme

*Contact Author: okaynar@cumhuriyet.edu.tr, Cumhuriyet University, Sivas, Türkiye

GİRİŞ

Günümüz dünyasında teknolojinin ve internetin gelişmesiyle beraber dijital ortamda bulunan veri miktarı hızla artış göstermektedir. Bu artışın bir kısmını haber siteleri, bilimsel siteler, sosyal medya yazıları gibi metin içerikli veriler oluşturmaktadır. Ayrıca bu metin içerikli verilerin büyük bir kısmı da yapılandırılmamış haldedir. Verilerin bu kadar çok olduğu bir ortamda istenilen bilgiye ulaşmak ciddi bir problemdir. Günümüz bilgi çağı, aranan bilgiye daha çabuk ve hızlı erişmek için otomatik metin özetleme sitemlerinin bilgi çıkarımı ile ilgili birçok alanda kullanımını zorunlu hale getirmektedir. Özellikle basit, dilden bağımsız özetleme yöntemlerine olan talep her geçen gün artmaktadır. Özet çıkarma işlemini başarılı bir şekilde gerçekleştirmenin yolu, bu işlemin insanlar tarafından yapılmasıdır, ancak her haberin, makalenin ya da belgenin manuel olarak özetlenmesi oldukça zahmetli ve zaman alıcı işlemlerdir. Bu nedenlerden dolayı dokümanların içerdiği önemli bilgileri özetleyip çıkararak kullanıcıya sunabilecek özetleme sistemlerine ihtiyaç duyulmaktadır.

Otomatik doküman özetleme kullanıcıya metnin tamamını okumadan doküman hakkında temel bilgiyi vermeye çalışan ve doküman içerisindeki önemli bilgiyi tutarken boyutunu düşüren bir sistemdir (Lee vd., 2009). Bu sistem sürecinde özet dışında kalan parçalar, doküman açısından önemsiz ve temel konuyu yansıtmayan kısımlardır. Bundan dolayı bu parçaların özette yer almaması problem değildir. Ancak bu sistemin karşılaştığı bazı zorluklarda vardır. Bunlardan en belirgini doküman içerisindeki önemli bilgiyi içeren kısımların seçilmesidir (Moreno, 2014). Bu zorlukların aşılabilmesi için birçok farklı yöntem ve yaklaşım ortaya atılmıştır.

Otomatik özetleme sistemleri metni analiz etme ve özet oluşturma açısından çıkarıcı ve yorumlayıcı olarak ikiye ayrılmaktadır (Khan ve Salim, 2014). Yorumlayıcı özetleme dokümanın içeriğini anlayıp farklı kelimelerle daha az şekilde ifade edilmesidir. Bu işlem için öncelikle gelişmiş dilbilimsel yöntemler kullanılarak doküman incelenir ve konuyla ilgili kavramlar belirlenir. Daha sonra dokümandaki önemli bilgiler, belirlenen kavram çerçevesinde en iyi şekilde yeniden ifade edilerek özet metin oluşturulur. Bu nedenle orijinal dokümanda geçmeyen bazı cümle veya kelimeler özet içerisinde bulunabilir (Dalal ve Malik, 2013).

Çıkarıcı özetleme ise dokümandaki cümle, paragraf gibi metin parçalarının önemine göre seçilmesine dayanmaktadır. Bu parçaların öneminin belirlenmesi için doküman içerisindeki konumu, ipucu veya başlık kelimesi içerip içermediği gibi öznitelikleri istatistiksel bazı yöntemler ile tespit edilir. Bu öznitelikler kullanılarak her bir cümlenin skoru hesaplanır. Bu skor aynı zamanda o cümlenin dokümanı temsil etme skorudur. En son adımda ise en yüksek skora sahip cümle dokümanı en iyi temsil eden cümle olarak belirlenir ve özete eklenir (Grupta ve Lehal, 2010).

Çıkarıcı özetleme yaklaşımında oluşturulan özet cümleleri direkt orijinal kaynaktan alındığı için bir nevi dokümanın kısaltılmış halidir. Böylece özet içerisinde orijinal dokümandan farklı bir kelime bulunmaz. Ayrıca yorumlayıcı özetleme sistemindeki gibi gelişmiş dilbilimsel yöntemlere de ihtiyaç duymadığı için daha az karmaşıktır. Bundan dolayı literatürde çıkarıcı özetlemenin daha fazla kullanıldığı görülmektedir.

Otomatik Doküman özetleme konusunda ilk çalışma bu alanda öncü kabul edilen Luhn (1958) tarafından yapılmıştır. Çalışmada doküman özetleme için 2 aşamalı bir yöntem önermişlerdir. İlk aşamada kaynak doküman belirli ön işlemlerden geçirilir. Bu ön işlemler doküman içerisindeki konu ile ilgili bir anlam ifade etmeyen gereksiz kelimelerin (edat, zamir vs.) temizlenmesiyle başlar. Daha sonra geriye kalan kelimelerden 6 farklı harften az olanlar ve aynı ön eke sahip kelimeler aynı sözcük ailesine ait olarak kabul edilir. Son adımda ise seyrek olan kelimeler temizlenir. Böylece elde kalan kelimeler içerisinde frekansı en yüksek kelimeler anahtar kelimeler olarak belirlenir. Her bir cümleye içerdiği anahtar kelime sayısı ile orantılı olarak bir skor değeri atanmakta ve daha sonra bu skor değerine bağlı olarak cümleler sıralanmaktadır. Sıkıştırma oranına bağlı olarak özeti oluşturacak en iyi skora sahip n adet cümle seçilir.

Luhn'un çalışmasında cümlelere ait sadece bir tane öznitelik olan anahtar kelimelerin geçip geçmediği kullanılmıştır. Edmundson ise bu özetleme sistemini cümlenin farklı özniteliklerini kullanarak geliştirmiştir. Edmundson ve Wyllys (1961), tarafından yapılan çalışmada cümlenin dokümandaki yeri, uzunluğu ve başlıkta geçip geçmediği gibi farklı öznitelikler kullanılmıştır. Daha sonra Edmundson (1969) tarafından yapılan bir sonraki çalışmada ise bu özniteliklerin yanı sıra ipucu kelimeleri özniteliğini de eklemiştir. İpucu kelimeleri "sonuç olarak", "özellikle" gibi doküman içerisinde önem arz eden yerleri belirtmek için kullanılmaktadır.

Bu çalışmalar göstermiştir ki, cümlelere ait temel öznitelikler dokümanı hangi cümlenin daha iyi temsil ettiği hakkında bilgi vermektedir. Bunun sonucunda literatüre yeni bir özetleme yaklaşımı girmiştir. Özellik tabanlı özetleme yaklaşımı, cümlelerin bazı temel özniteliklerine göre dokümanı ne kadar temsil ettiğini ölçen ve en çok temsil eden cümlelerin birleştirilmesiyle özeti oluşturan çıkarıcı bir yöntemdir. Bu yöntemin en önemli noktalarından birisi, cümleleri en iyi şekilde temsil eden özniteliklerin belirlenmesidir. Literatür incelendiğinde, cümleleri temsil etmek üzere cümleye ait uzunluk ve lokasyon bilgisi (Fattah ve Ren, 2008), terim frekansları (Salton ve Buckley, 1988) , cümlelerin numerik veri içerip içermediği (Lin, 1999) ve cümle içerisinde özel isimlerin bulunup bulunmadığı (Ledneva vd., 2008) gibi bir çok farklı özniteliğin önerildiği görülmektedir

Özetleme işlemi sırasında cümleler için önerilen bu öz niteliklerden elde edilen skor değerleri toplanarak cümleye ait bir tek bir skor değeri belirlenir. Cümleler bu skor değerlerine göre sıralanarak özette yer alan önemli cümleler belirlenir. Ancak her bir öz niteliğin aynı önem derecesine sahip olduğunu varsayarak doğrudan öznitelik skorlarını toplamak çoğu durumda özetleme başarımının düşmesine neden olacaktır. Örneğin gazetecilikle ilgili yazının genelde ilk cümlesi ve başlığa olan benzerliği önemliyken (Brandow vd.,1995) bilimsel bir makalede tam aksine cümlenin en sonunda yer alan sonuç kısmı önem taşımaktadır. Bu durumda değişik konulardaki dokümanlar için aynı öznitelik farklı bir etkiye sahip olabilmektedir. Bu nedenle kullanılan öz niteliklerin önem derecesinin ne olacağını gösteren bir ağırlık atama yöntemine gereksinim duyulmaktadır. Ağırlıkların kullanıcı tarafından belirlenmesi objektif olmaktan uzak, çok sayıda deneme yanılma gerektirmesi ve konuyla ilgili alan bilgisi ihtiyacından dolayı dezavantajlara sahiptir. Ayrıca her bir özniteliğin etkisi için çözüm uzayındaki tüm olasılıkların manuel olarak denenmesi ve optimum ağırlıkların belirlenmesi son derece zor işlemdir. Bu nedenle çözüm uzayını en kısa sürede tarayarak ağırlıkların optimum değerlerini belirleyecek genetik algoritma, parçacık sürü algoritması, arı koloni vb. sezgiler algoritmalara ihtiyaç duyulmaktadır.

Çıkarıma dayalı özetleme yöntemleri incelendiğinde literatürde birçok çalışma yapılmıştır. Kupeic vd., (1995) doküman özetleme problemini bir sınıflandırma problemine çevirmiş ve eğitilebilir doküman özetleme sistemi önermiştir. Dokümana ait ipucu kelimesi, hangi paragrafta bulunduğu, tematik kelime sayısı gibi 5 farklı öznitelik Bayes yöntemiyle eğitilmiştir. Eğitilen makine %84 başarı oranı elde etmiştir.

Babar ve Patil (2015), cümle uzunluğu, benzerliği, pozisyonu, özel isim ve tematik kelime sayıları gibi farklı öznitelik kullanarak doküman özetleme yöntemi önermiştir. Çalışmada yöntem olarak ise bulanık mantık kullanılmıştır. Her bir öznitelik düşük, orta, yüksek olmak üzere 3 farklı şekilde bulanık mantık üyeliği ile temsil edilmiştir. Çıkış üyelikleri ise önemsiz, normal, önemli olarak cümleleri belirtmektedir. Bununla beraber cümlelerin gizli anlam skoru da belirlenerek her iki yöntem birleştirilmiş ve özet oluşturulmuştur.

Cığır vd., (2012) tarafından Türkçe dili üzerinde bir çalışma yapılmıştır. Cümlelere ait 5 temel öznitelik belirlenmiş ve oluşturulan fonksiyon ile cümleye ait skor hesaplanmıştır. 50 Haber ile eğitilen sistem, 65 haber ile test edilmiştir ve tüm öznitelikler kullanıldığında 0,338 skoru elde etmiştir.

Doküman özetleme sistemlerinde genel problemlerden birisi ağırlıkların doğru seçilmesidir. Literatürde ağırlıkların doğru seçilmesini sağlamak için sezgisel algoritmalarda başarılı bir şekilde kullanılmıştır.

Mesaj Anlama Konferansları (DUC, Document Understanding Conferences) metin özetleme sistemlerinin tanıtılması ve yol haritasının çizilmesi için yapılan doğal dil işleme konferanslarıdır. Bu konferanslarda en iyi özet değerlendirme paketi seçilen ROUGE (Lin, 2004) sistem tarafından oluşturulan özet ile orijinal özet arasındaki kaliteyi ölçmektedir. Binwahlan ve Salim (2009) hangi özniteliğin daha önemli olduğunu belirleme amacıyla parçacık sürü optimizasyonu kullanarak DUC-2002 verisetine uygulamışlardır.. Cümlelere ait 5 farklı öznitelik belirlenmiş ve DUC veri setinden 100 veri ile model eğitilmiştir. DUC, metin özeteleme sistemlerini ölçmek için yapılan konferanslardır. Yapılan testte model 0,430 Rouge-1 puanı ile MS Word Summarizer modülünden daha iyi sonuç elde etmiştir.

Uy vd., (2012) paragrafın konumu, cümlenin konumu, uzunluğu ve içerdiği kelimelerin TF/IDF skorları olmak üzere 4 farklı öznitelik kullanarak Vietnam dili üzerinde denemişlerdir. Yöntem olarak her bir özet için ayrı genetik programlama modeli oluşturulmuş ve en iyi başarıyı gösteren model, testte kullanılmıştır.

Suanmali ve Salim (2011), cümle uzunluğu, pozisyonu, terim frekansları, tematik ve özel isim içerip içermediği gibi 8 farklı özniteliği kullanarak genetik algoritma ile eğitmiştir. 100 DUC verisi ile Eğitilen algoritma sonucunda elde edilen ağırlıklar 62 DUC verisine uygulanmıştır. Sonuç olarak 0,45359 Rouge-1 skoru elde edilmiştir.

Bu çalışmada ise, çıkarıcı doküman özetleme yaklaşımlarından birisi olan özellik tabanlı yöntem ele alınmıştır. Özniteliklerin hangisinin daha önemli olduğu ise genetik algoritma yardımıyla belirlenmiştir. Çalışmanın yöntem kısmında genetik algoritma, doküman özetlemede kullanılan ön işlem basamakları, özniteliklerin çıkarılması ve genetik algoritma yardımıyla özniteliklere ait ağrılıkların belirlenmesi ile ilgili bilgiler verilmiştir. Uygulama kısmında elde edilen bulgular sunulmuş, elde edilen bu bulgular sonuç ve tartışma kısmında yorumlanarak önerilerde bulunulmuştur.

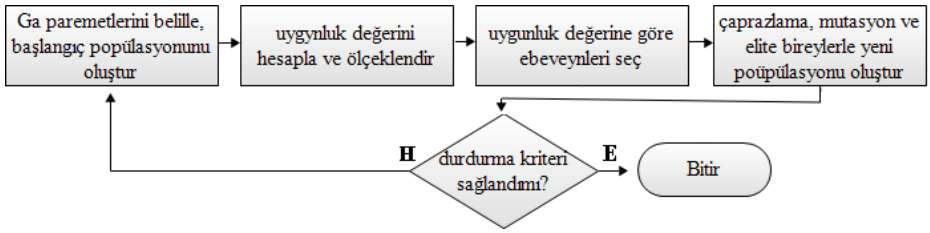
YÖNTEM

Genetik Algoritma

Genetik algoritma (GA) doğal seleksiyon, çaprazlama ve mutasyon tabanlı, biyolojiden ilham alan global arama optimizasyon tekniğidir (Holland, 1975). Bu yöntemde öncelikle aday çözümlerden oluşan bir popülasyon üretilir ve bu popülasyon belirlenen durdurma kriteri sağlanıncaya kadar seleksiyon, çaprazlama ve mutasyon adı verilen genetik işlemler aracılığı ile güncellenir. GA daha iyi çözümleri bulma sürecinde en iyi olanın hayatta kalması fikrini kullanır. GA tek bir çözümü kademeli olarak değiştirmektense bir çözüm popülasyonunu güncelleyerek arama yapması yönüyle geleneksel doğrusal olmayan optimizasyon tekniklerinden ayrılır. Klasik optimizasyon algoritmaları iterasyon noktalarının yerel özellikleri ile ilgilendiği için kolayca yerel ekstremum noktalarına takılabilirler. Bunun aksine GA sistematik aramaya ek olarak rasgele arama operatörü de kullandığından dolayı yerel

minimum veya maksimum noktasına takılması önlenmiş olur. Genetik algoritmanın çalışma mantığı şekil1 de gösterilmiştir.

Şekil 1. Genetik Algoritma Süreci



Genetik Algoritma ile Otomatik Doküman Özetleme

GA çalışma prensibi gereği uygunluk fonksiyonunu minimize veya maksimize etmeyi hedefler. Özellik tabanlı özetlemede ise hangi özniteliğin daha önemli olduğunu belirlemek için kullanılır. Ancak öncesinde özeti çıkartılacak doküman sistem tarafından daha iyi işlenebilmesi ve başarılı sonuç edilebilmesi maksadıyla birtakım ön işlemlerden geçirilir.

Şekil 2. GA ile Doküman Özetleme Sistemi

Dokümana uygulanacak ilk önişlem dokümanın gereksiz kelimelerden temizlenmesidir. Gereksiz kelimeler cümle içerisindeki geçen edat, bağlaç, zamir gibi doküman içerisinde sık geçmesine rağmen herhangi bir bilgi değeri bulunmayan kısımlardır. Bu kelimeler özetleme başarısını düşürdüğü gibi işlem süresini de uzatmaktadır. Doküman gereksiz kelimelerden temizlendikten sonra parçalama işlemine geçilir.

Çıkarıcı özetleme sistemlerinde önemli aşamalardan birisi dokümanlardaki hangi parçaların özeti oluşturacağının belirlenmelidir. Bu parçalar kelime öbeği, cümle, paragraf olabilir. Bu çalışmada ise cümleler seçilerek özet oluşturulması hedeflenmiştir. Bu nedenle çalışmamızda her bir doküman cümlelere bölünmüştür.

Bir sonraki adımda dokümanın başlık ve içerik kısımları belirlenmiştir. Bunlar dışında kalan yazar bilgisi, anahtar kelimeler, yayın tarihi gibi kısımlar doküman özetleme açısından herhangi bir öneme veya bilgiye sahip olmadığı için temizlenmiştir. Böylelikle doküman başlık ve içerikten oluşacak şekilde ham hale getirilmiştir.

Doküman gereksiz kelimelerden temizlendikten ve cümlelere ayrıldıktan sonra özette yer alacak cümlelerin seçilme işlemi başlar. Bu noktada dokümanı en iyi temsil eden, metnin ana konusu hakkında en çok bilgi içeren cümleler seçilmelidir. Bu amaçla dokümana ait 8 farklı öznitelik belirlenmiştir. Bu öznitelikler şunlardır:

1) *Cümle Uzunluğu:* Cümlenin uzunluğunun filtrelenmesiyle tarih belirten satırlar veya yazar isimleri gibi kısımların önüne geçilebilir. Bu işlem için bir eşik belirlenir ve bu eşiğin altında/üstünde olanlar özet dışında tutulur.

$$\text{Özellik}_{\text{Cümle Uzunluğu}}\,(S_i) = \frac{S_i\ \text{cümlesindeki Kelimelerin Toplam Sayısı}}{\text{En uzun Cümledeki Kelimelerin Toplam Sayısı}} \quad (1)$$

2) *Numerik Veri*: Cümle içerisinde bulunan nümerik veriler dokümanın genel konusuyla ilgili olabilecek önemli bir istatistiksel veriyi veya deneysel bir durumun sonucu hakkında bilgi verebilir. Bundan dolayı içerisinde nümerik verilerin bulunduğu cümlelerin seçilmesi özetin konu temsili açısından önem teşkil edebilir.

$$\text{Özellik}_{\text{Nümerik Veri}}\,(S_i) = \frac{S_i\ \text{cümlesindeki Toplam Nümerik Veri}}{S_i\ \text{cümlesindeki Toplam Kelime Sayısı}} \quad (2)$$

3) *Tematik Kelimeler*: Tematik veya konu ile ilişkin kelimeler, gereksiz kelimeler olan edat, bağlaç dışında dokümanda en fazla kelimeler olarak nitelendirilir. Dokümanda en fazla geçen belirli sayıdaki kelimeler tematik kelime olarak belirlenir. Bu kelimelerin konu ile daha fazla ilişkisi olduğu düşünülerek ilgili cümlede geçip geçmediği veya ne kadar geçtiğine bakılarak ilgili cümlenin özette olup olmaması gerektiği belirlenebilir.

$$\text{Özellik}_{\text{Tematik Kelime}}\,(S_i) = \frac{S_i\ \text{cümlesindeki Toplam Tematik Kelime Sayısı}}{S_i\ \text{cümlesindeki Toplam Kelime Sayısı}} \quad (3)$$

4) *Cümle Konumu:* Cümlelerin doküman içerisindeki bulundukları konumlar cümlelerin önemlerini belirtebilir. Örneğin, dokumanın ilk veya son cümlesi olması veya cümlenin doküman/paragrafta konumu ya da sırası o cümlenin önemli olup olmadığının göstergesi olarak kullanılabilir. Örneğin, dokümandaki 20 cümle olduğunu farz edersek ilk cümle konumu için 20/20, 5.cümle konumu için 15/20 olarak hesaplanabilir.

$$\text{Özellik}_{\text{Cümle Konumu}}\,(S_i) = \frac{\text{Toplam Cümle Sayısı} - S_i\ \text{Cümle Konumu}}{\text{Toplam Cümle Sayısı}} \quad (4)$$

5) *Başlık Kelimeleri:* Dokümanda yer alan cümlelerdeki kelimeler ile başlıktaki kelimelerin eşleşmesiyle hesaplanan ve cümlenin konu hakkında daha ilgili olup olmadığını gösterebilecek bir özniteliktir.

$$\text{Özellik}_{\text{Başlık Kelimeleri}}\,(S_i) = \frac{S_i\ \text{cümlesinde geçen Başlık Kelimeleri Toplamı}}{\text{Başlık Uzunluğu}} \quad (5)$$

6) *Özel İsim:* Maksimum miktarda özel isim içeren cümleler genellikle önemli cümleler olarak görülmektedir ve bir cümlede ne kadar çok özel isim var ise o oranda özet dokümanda bulunması gerekmektedir.

$$\text{Özellik}_{\text{Özel İsim}}\,(S_i) = \frac{S_i\ \text{özel isim Toplamı}}{S_i\ \text{cümle uzunluğu}} \quad (6)$$

7) *Cümle Benzerliği:* ***N*** cümleden oluşan bir dokümandaki her bir cümlenin S_i bir diğeriyle benzerliği hesaplanır ve N x N lik bir matris oluşturulur. Oluşturulan bu matrisin köşegenlerinde yer alan veriler her cümlenin kendisiyle benzerlik değeri olacağı için köşegenlerdeki değerlere 0 atanır. En son olarak ise her bir cümlenin diğer cümlelerle olan benzerlikleri toplanarak maksimum olan benzerliğe bölünerek cümle benzerliği özelliği hesaplanmış olur. Benzerlik ölçütü olarak kosinüs benzerliği kullanılmıştır.

$$\text{Özellik}_{\text{Cümle Benzerliği}}\,(S_i) = \frac{\text{Toplam}\left[(S_i, S_j)\right]}{\text{Maksimum}\left[(S_i, S_j)\right]}\ i = 1,2,3,\ldots N\ ;\ j = 1,2,3,\ldots N \quad (7)$$

8) *Terim Ağırlığı (TF/ISF)* **:** TF/ISF amacı bir kelimenin cümlede ne kadar önemli olduğunu göstermektedir. Bir cümledeki terimlerin var olup olmadığı o cümlenin önemini hesaplamada kullanılan yaygın yöntemlerden birsidir. İlgili cümledeki Terimlerin TF/ISF toplanarak cümlenin önem derecesi hesaplanabilir.

$$\text{Özellik}_{\text{Terim Ağırlığı}}\,(S_i) = \frac{S_i\ \text{cümlesinde}\ \frac{TF}{ISF}\ \text{toplamı}}{\text{Maximum}\ (\frac{TF}{ISF}\ \text{toplamı})} \quad (7)$$

Cümlelere ait öznitelikler hesaplandıktan sonra bu öznitelikler belirli bir fonksiyon ile birleştirilerek cümle skoru elde edilir. Klasik yaklaşımda oluşturulan fonksiyon ile öznitelik skorları denklem 8'deki gibi toplanır.

$$Skor(C_i) = f_{1(C_i)} + f_{2(C_i)}+, \dots, +f_{n(C_i)} \quad (8)$$

Daha sonra en yüksek skora sahip cümleler özet olarak seçilir. Klasik istatistiksel yöntemde bu fonksiyonda kullanılan her bir öznitelik eşit öneme sahip olarak kabul edilir. Ancak farklı konularda dokümanlarda hatta aynı konudaki farklı dokümanlar da bile aynı öznitelik daha fazla etkiye sahip olabilir. Bunun yanı sıra bazı özniteliklerin başarıya etkisi olumsuz olabilir ve bunların skor hesaplama fonksiyonunda çıkartılması gerekebilir.

Bunun için her bir özelliğe farklı bir ağırlık verilmesi gerekmektedir. Böylelikle önemsiz olan özniteliğin skora etkisi azaltılırken, önemli olanın etkisi arttırılabilir. ağırlıklandırılmış öznitelik skorları denklem 9'daki gibi hesaplanarak cümleye ait skor elde edilir.

$$Skor(C_i) = w_1 * f_{1(C_i)} + w_2 * f_{2(C_i)}+, \dots, +w_n * f_{n(C_i)} \quad (9)$$

GA hangi özelliğe daha fazla, hangi özelliğe daha az ağırlık verilmesi gerektiğini verimli bir şekilde hesaplayabilir. Bu nedenle öncelikle GA'nın ağırlıkları en verimli şekilde hesaplayabilmesi için uygunluk (fitness) fonksiyonu oluşturulması gerekmektedir. Uygunluk fonksiyonu, tüm popülasyon içerisindeki kromozomların hangisinin daha iyi olduğunu hesaplamaktadır.

Çalışmamızda oluşturulan uygunluk fonksiyonunda öncelikle her bir dokümandaki cümlenin skoru, cümleye ait özniteliklerin belirli ağırlıklarla çarpılmasıyla elde edilmiştir. Elde edilen bu skorlar azalan bir şekilde sıralanmıştır. Sonrasında ise en yüksek skora sahip belirlenen özet uzunluğu sayısı kadar cümle birleştirilerek sistem özeti oluşturulmuştur. Bu özet bir sonraki adımda orijinal özet ile Rouge-N ölçütüne göre karşılaştırılmıştır.

ROUGE ölçütleri ile dokümanın kalitesi hesaplanırken her iki özette geçen kelime ve kelime dizilerinin örtüşmesi Denklem 10'daki gibi hesaplar.

$$Rouge - N = \frac{N - Gram \text{ Örtüşen } Kelime\ Sayısı}{Orijinal\ Dokümandaki\ N - Gram\ Sayısı} \quad (10)$$

Buradaki N, cümledeki kelimeleri kaçarlı şekilde bölüneceğini belirtmektedir.

Oluşturulan fitness fonksiyonu GA ile eğitilerek en yüksek rouge skoru elde eden öznitelik ağırlıkları belirlenmiştir. Eğitim için oluşturulan GA modelinde popülasyon sayısı 40, jenerasyon sayısı ise 500 olarak belirlenmiştir. Bununla beraber çaprazlama oranı 0,9 iken Mutasyon oranı 0,1 seçilmiştir. Eğitim sonucunda elde edilen en iyi ağırlıklar doküman özetleme sisteminde kullanılacak ağırlıklar olarak belirlenir.

BULGULAR

Veriseti ve Değerlendirme

Çalışmada 120 adet Türkçe haberden oluşan veri seti (Özsoy vd., 2010) kullanılmıştır. Bu haberlerin 80 tanesi ve bu 80 taneye ait insanlar tarafından oluşturulmuş özetler eğitim için ayrılırken, 40 tane haber ise test sürecinde kullanılmıştır. Her bir doküman önceki kısımlardan belirtilen önişlemlerden geçmiş ve öznitelikleri çıkartılarak işlenebilir hale getirilmiştir. Ayrıca özetleri uzunluğu orijinal özetteki cümle miktarına eşitlenmiştir. Böylelikle insanlar tarafından oluşturulmuş özete en benzer şekilde otomatik özet oluşturulması amaçlanmıştır.

Çalışmada GA ile elde edilen optimum ağırlıklar test verileri üzerinde denenmiştir. Her bir ağırlık ilgili olduğu öznitelik ile çarpılarak test verisi içerisindeki her bir dokümanın cümle skorları hesaplanmıştır. Daha sonra bu cümle skorları sıralanarak en yüksek skora sahip olanlar seçilmiş ve sistem tarafından belirlenen özet olarak kabul edilmiştir. Bu özet ile ilgili dokümana ait insanlar tarafından oluşturulmuş orijinal özet Rouge-1 ve Rouge-2 ölçütleri kullanılarak değerlendirilmiştir.

Uygulama Sonuçları

GA ile eğitim verisi sonucunda elde edilen ağırlıklar daha sonra test verisi üzerinde kullanılmıştır. Tablo 1'de Eğitim verisi ve Test verisi üzerindeki Rouge ölçütü değerleri belirtilmektedir.

Tablo 1. Ağırlıklı Ortalama Rouge-1 ve Rouge-2 Değerleri

	Eğitim Verileri (80 Adet)			Test Verileri (40 Adet)		
	Recall	*Precision*	*F skoru*	*Recall*	*Precision*	*F skoru*
Rouge_1	0,867	0,830	0,847	0,828	0,828	0,828
Rouge_2	0,806	0,767	0,785	0,746	0,743	0,744

Elde edilen bu metriklerin yanında eğer her bir öznitelik için hiçbir ağırlık tanımlanmasaydı nasıl bir sonuç elde edileceği ölçülerek Tablo 2'de gösterilmiştir.

Tablo 2. Ağırlıklandırılmamış Rouge-1 ve Rouge-2 Değerleri

	Eğitim Verileri (80 Adet)			Test Verileri (40 Adet)		
	Recall	*Precision*	*F skoru*	*Recall*	*Precision*	*F skoru*
Rouge _1	0,753	0,609	0,666	0,775	0,686	0,725
Rouge _2	0,648	0,528	0,576	0,664	0,596	0,626

Böylece GA ile eğitilen ağırlıklandırma işleminin özet başarısına nasıl bir etki ettiği ölçülmüştür.

TARTIŞMA VE SONUÇ

Bu çalışmada Türkçe metinler için özellik tabanlı bir doküman özetleme sistemi önerilmiştir. Sistemde her bir dokümana ait cümleler ayrıştırılmış ve 8 farklı öznitelikleri çıkartılmıştır. Bu cümleler GA algoritma vasıtasıyla 80 veri ile eğitilerek her bir özniteliğin optimum ağırlığı bulunmuştur. Algoritma sonucunda elde edilen optimum ağırlıklar 40 test verisi üzerinde denenmiştir. Test aşaması sonucu göstermiştir ki GA'nın elde ettiği ağırlıklar eğitim verisi üzerinde 0,847 Rouge-1 skoru, 0,785 Rouge-2 skoru elde etmiştir. Test verileri üzerinde ise 0,828 Rouge-1 ve 0,744 Rouge-2 gibi yüksek skorlar elde edilmiştir.

Elde edilen ağırlıklandırılmış skorların yanı sıra sistem hiçbir ağırlıklandırma kullanılmadan sadece özniteliğin skorları toplanarak da test edilmiştir. Bu işlemin sonucunda ağırlıksız öznitelikler test verisi üzerinde 0,725 Rouge-1 ve 0,626 Rouge-2 Skoru elde etmiştir. Tüm bu sonuç değerleri incelendiğinde Genetik Algoritma vasıtasıyla ağırlıkları belirlenmiş olan sistem ciddi bir oranda başarı artışı göstermiştir.

TEŞEKKÜR

Bu çalışma, Cumhuriyet Üniversitesi Bilimsel Araştırma Projeleri (CÜBAP) tarafından İKT-112 proje numarası ile desteklenmiştir.

KAYNAKLAR

Babar, S. A., & Patil, P. D. (2015). Improving Performance of Text Summarization. Procedia Computer Science, 46, 354-363.

Binwahlan, M. S., Salim, N., & Suanmali, L. (2009, April). Swarm based text summarization. In Computer Science and Information Technology-Spring Conference, 2009. IACSITSC'09. International Association of (pp. 145-150). IEEE.

Brandow, R., Mitze, K., & Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. Information Processing & Management, 31(5), 675-685.

Cigir, C., Kutlu, M., & Cicekli, I. (2009, September). Generic text summarization for Turkish. In Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on (pp. 224-229). IEEE.

Dalal, V., & Malik, L. G. (2013, December). A survey of extractive and abstractive text summarization techniques. In Emerging Trends in Engineering and Technology (ICETET), 2013 6th International Conference on (pp. 109-110). IEEE.

Document understanding conferences (DUC) < *http://www-nlpir.nist.gov/projects/duc/index.html* >

Edmondson, H. P. (1969). New Methods in Automatic Extraction. Journal of the Association for Computing Machinery, vol. 16, no. 2, pp. 264–285, 1969.

Edmundson, H. P., & Wyllys, R. E. (1961). Automatic abstracting and indexing—survey and recommendations. Communications of the ACM, 4(5), 226-234.

Fattah, M. A., & Ren, F. (2008). Automatic text summarization. Gas, 692, 10785.

Gholamrezazadeh, S., Salehi, M. A., & Gholamzadeh, B. (2009). A comprehensive survey on text summarization systems. Proceedings of CSA, 9, 1-6.

Gupta, V., & Lehal, G. S. (2010). A survey of text summarization extractive techniques. Journal of emerging technologies in web intelligence, 2(3), 258-268.

Holland, John H. (1975) Adaptation in natural and artificial systems. An introductory analysis with application to biology, control, and artificial intelligence. Ann Arbor, MI: University of Michigan Press

Khan, A., & Salim, N. (2014). A review on abstractive summarization methods. Journal of Theoretical and Applied Information Technology, 59(1), 64-72.

Kupiec, J., Pedersen, J., & Chen, F. (1995, July). A trainable document summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 68-73). ACM.

Kutlu, M., Cığır, C., & Cicekli, I. (2010). Generic text summarization for Turkish. The Computer Journal, bxp124.

Ledeneva, Y., Gelbukh, A., & García-Hernández, R. A. (2008, February). Terms derived from frequent sequences for extractive text summarization. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 593-604). Springer Berlin Heidelberg.

Lee, J. H., Park, S., Ahn, C. M., & Kim, D. (2009). Automatic generic document summarization based on non-negative matrix factorization. Information Processing & Management, 45(1), 20-34.

Lin, C. Y. (1999, November). Training a selection function for extraction. In Proceedings of the eighth international conference on Information and knowledge management (pp. 55-62). ACM.

Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop (Vol. 8).

Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.

Ozsoy, M. G., Cicekli, I., & Alpaslan, F. N. (2010, August). Text summarization of turkish texts using latent semantic analysis. In Proceedings of the 23rd international conference on computational linguistics (pp. 869-876). Association for Computational Linguistics.

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

Suanmali, L., Salim, N., & Binwahlan, M. S. (2011). Genetic algorithm based sentence extraction for text summarization. International Journal of Innovative Computing, 1(1).

Torres-Moreno, J. M. (2014). Automatic text summarization. John Wiley & Sons.

Uy, N. Q., Anh, P. T., Doan, T. C., & Hoai, N. X. (2012, August). A study on the use of genetic programming for automatic text summarization. In Knowledge and Systems Engineering (KSE), 2012 Fourth International Conference on (pp. 93-98). IEEE.