Fırat Üniversitesi Müh. Bil. Dergisi
35(1), 71-79, 2023

Araştırma Makalesi
https://doi.org/10.35234/fumbd.1155617

Çizgeler Üzerinde Farklı Ağırlıklandırma Yöntemleri Ve Merkezilik Ölçütleri İle Çıkarımsal Metin Özetleme

Abdulsamet AYDIN[1], Taner UÇKAN[2*]
[1] Yapay Zekâ ve Robotik Anabilim Dalı, Fen Bilimleri Enstitüsü, Van Yüzüncü Yıl Üniversitesi, Van, Türkiye
[2] Bilgisayar Mühendisliği, Mühendislik Fakültesi, Van Yüzüncü Yıl Üniversitesi, Van, Türkiye
[1] a.sametaydn@gmail.com, [*2] taneruckan@yyu.edu.tr

(Geliş/Received: 04/08/2022; Kabul/Accepted: 11/12/2022)

Öz: Çıkarıma dayalı metin özetleme konusunda birçok farklı yaklaşım vardır. Bu çalışmada Kosinüs Benzerliği, Jaccard Benzerliği, Levenshtein Benzerliği ve Pearson Korelasyon Katsayısı ölçütleri kullanarak ağırlıklı çizgeler oluşturulmuştur. Bu çizgelerdeki düğümler ile temsil edilen cümleler arasından en değerli olanları belirlemek amacı ile Arasındalık Merkeziliği, Yakınlık Merkeziliği, Derece Merkeziliği ve Özvektör Merkeziliği ölçümleri kullanılmıştır. Çıkarımsal metin özetlemede kullanılan yaklaşımların faklı kombinasyonları ile her bir metinden 16 adet 200 ve 400 kelimelik özetler oluşturularak en başarılı özetlerin hangi yaklaşımlar ile elde edildiğinin tespit edilmesi hedeflenmiştir. Çalışma, Document Understanding Conference (DUC-2002) veri seti üzerinde gerçekleştirilmiştir. ROUGE değerlendirme metrikleri ile performansı hesaplanmış ve elde edilen sonuçlar ayrıntılı olarak karşılaştırılmıştır. En başarılı sonuçlar, sırasıyla 200 kelimelik özetlerde Jaccard Benzerliği ve Yakınlık merkeziliği yaklaşımı ile 0.46091 ve 400 kelimelik özetlerde ise Kosinüs Benzerliği ve Özvektör Merkeziliği yaklaşımı ile 0.52485 F-Skor değerleri ile elde edilmiştir.

Anahtar kelimeler: Çıkarımsal Metin Özetleme, Düğüm Merkeziliği, Benzerlik Yöntemleri

Extractive Text Summarization Using Different Weighting Methods And Centrality Measures On Graphs

Abstract: There are many different approaches to extractive text summarization. In this study, weighted graphs were created using Cosine Similarity, Jaccard Similarity, Levenshtein Similarity and Pearson Correlation Coefficient measures. Betweenness Centrality, Closeness Centrality, Degree Centrality and Eigenvalue Vector Centrality measurements were used to determine the most valuable nodes among the nodes representing the sentences in these graphs. With the different combinations of approaches used in inferential text summarization, it is aimed to determine which approaches the most successful summaries are obtained by creating 16 pieces of 200 and 400 word summaries from each text. The study was carried out on the Document Understanding Conference (DUC-2002) dataset. Its performance was calculated using ROUGE evaluation metrics and the results were compared in detail. The most successful results were obtained with the Jaccard Similarity and Closeness centrality approach in the 200-word abstracts, 0.46091 and the Cosine Similarity and Eigenvector Centrality approach in the 400-word abstracts, with an F-Score of 0.52485, respectively.

Key words: Extractive Text Summarization, Node Centrality, Similarity Methods

1. Giriş

Teknolojinin gelişmesi ile birlikte akademik ve iş dünyasında internet ve bilgisayar kullanımı her geçen gün artmaktadır. Dijital ortamlardaki metinsel verilerin katlanarak büyümesi, faydalı bilgilere erişimi oldukça zorlaştırmıştır. Bilgiye erişim sırasında aranılan bilgi ile ilgili binlerce doküman elde edilebilir. Her bir dokümanı okumak ve gerçekten de aranılan bilginin dokümanda bulunup bulunamadığını tespit etmek zahmetli bir iştir. Bu nedenlerle bilgiye erişimde zamandan tasarruf etmek amacıyla dokümanları özetleyerek önemli bilgileri kullanıcıya sunan yöntemler geliştirilmektedir.

Metin özetleme bir dokümanı genel hatlarıyla anlatan kısaltılmış halidir. Metin özetlemede çıkarıcı ve yorumlayıcı yöntemler kullanılır. Yorumlayıcı metin özetleme dilbilimsel yöntemler kullanılarak dokümandaki cümlelerden aynı anlamı ifade eden yeni cümleler üretme süreci olarak tanımlanabilir. Bu sebeple özeti çıkarılan dokümanda bulunmayan yeni cümleler veya yeni kelimeler özette yer alabilir [1]. Çıkarıma dayalı özetleme, cümle bazında önemli içeriği bulmaya dayanır. Frekans ve benzerlik gibi özellikler kullanılarak önemli cümlelere karar verilir. İlk çalışmalar 1950 yıllarında Luhn tarafından yapılmıştır [2]. Bu çalışmalarda genel olarak kelime frekansları kullanılmıştır. Yapılan araştırmalar bir metinde çok fazla geçen kelimeler ile edatlar, bağlaçlar metin

[*] Sorumlu yazar: taneruckan@yyu.edu.tr. Yazarların ORCID Numarası: [1] 0000-0002-5329-4407, [2] 0000-0001-5385-6775

içeriği hakkında pek fazla bilgi vermediğini ortaya çıkarmıştır [3]. Çıkarıma dayalı metin özetlemede önemli olan metin hakkında en fazla bilgi içeren cümlelerin seçilerek özete konulmasıdır.

Otomatik metin özetleme işleminin yapılabilmesi için öncelikle kelimelerin bilgisayarlar tarafından anlaşılır hale getirilmesi gerekmektedir. Bu nedenle TF-IDF, Word2Vec, GloVe, BERT ve FastText gibi algoritmalar kullanılır [4-6]. Böylece metin içerisindeki cümleler vektör olarak temsil edilebilmektedir. Bu çalışmada cümleler, TF-IDF algoritması kullanılarak vektör haline getirilmiştir. Bu vektörler, Kosinüs Benzerliği, Jaccard Benzerliği, Levenshtein Benzerliği ve Pearson Korelasyon Katsayısı benzerlik yöntemleri kullanılarak ayrı ayrı benzerlik matrisleri oluşturulmuş ve çizgelere dönüştürülmüştür. Çizge olarak temsil edilen metindeki cümleler (düğümler) Arasındalık Merkeziliği, Yakınlık Merkeziliği, Derece Merkeziliği ve Özvektör Merkeziliği yaklaşımları kullanılarak en değerli cümlelerin seçilmesi ile özetler oluşturulmuştur. Bu çalışmada benzerlik ve merkezilik yöntemleri birlikte kullanılarak elde edilen özetlerin performansları karşılaştırılarak en başarılı özetlenin hangi yaklaşımlar ile oluşturulduğu tespit edilmiştir.

2. Literatür

Otomatik metin özetleme konusunda çıkarımsal, soyutlayıcı ve hibrit yaklaşımlar bulunmaktadır. Bilimsel makale, hukuki metin, biyomedikal doküman vb. alanlarda otomatik metin özetleme uygulamaları yaygın olarak kullanılmaktadır. Çizge tabanlı çıkarımsal metin özetleme uygulamalarında çizgedeki en değerli düğümler tespit edilmeye çalışılarak cümle bazında önemli içeriğin özette yer alması sağlanır [7].

İlk çalışmalar 1950 yıllarında Luhn tarafından yapılmıştır [2]. Bu çalışmalarda genel olarak kelime frekansları kullanılmıştır. Yapılan araştırmalar bir metinde çok fazla geçen kelimeler, edatlar ve bağlaçların metnin içeriği hakkında pek fazla bilgi vermediğini ortaya çıkarmıştır [3]. Çıkarıma dayalı metin özetlemede önemli olan metin hakkında en fazla bilgi içeren cümlelerin seçilerek özete konulmasıdır. Çizge tabanlı özetlemede ise cümleler çizge olarak temsil edilir ve düğümlerin ilişkilerine göre cümleler puanlanarak sıralanır. En yüksek puana sahip düğümler tarafından temsil edilen cümleler, metindeki sıralarına göre birleştirilerek özet oluşturulur.

Belwal ve diğ. [8], cümlelerin birbirleri ile olan benzerliklerinin yanı sıra cümleler ile dokümanın bütünü arasındaki benzerliği de dikkate alan çizge tabanlı bir yaklaşım önermiştir. Bu yaklaşımda cümlelerin birbirlerine benzerlikleri ve cümlelerin dokümanın konusuna olan benzerlikleri birlikte değerlendirilerek ağırlıklı çizgeler elde edilmiştir. Çizgedeki en yüksek değere sahip olan düğümler seçilerek temsil ettikleri cümleler ile özet oluşturulmuştur.

Joshi ve diğ. [9], konu bilgisi, anlamsal içerik, önemli anahtar kelimeler ve konum özelliklerini hesaplayarak her bir cümle için sıralama yapmışlardır. Her özellik için hesaplanan sıralama değerleri birleştirilerek dokümandaki her cümlenin nihai puanları ortaya çıkarılmıştır. En yüksek puana sahip olan cümlelere özette yer verilmiştir.

Azadani ve diğ. [10], biyomedikal alanına özgü bilgiden ve sık öğe kümesi madenciliği adı verilen bir veri madenciliği tekniğinden yararlanarak çizge tabanlı özetleme yöntemi önermiştir. Çalışmalarında cümlelerin benzerliklerini tespit etmek amacıyla Jaccard Benzerliği yaklaşımından yararlanmışlardır.

Metin özetleme alanında yapılan çalışmalar göstermiştir ki, bir cümle değerlendirilirken cümlenin belgedeki konumu, uzunluğu ve anahtar sözcüklere yer verilip verilmediği, cümledeki kelimelerin başlıkta yer alıp almadığı ve terim frekansı gibi kıstaslar değerlendirmeye dahil edilerek özetlemenin başarısı arttırılmıştır [11-13]. Mihalcea ve Tarau, PageRank algoritmasını TextRank adını verdikleri bir metin özetleme yöntemine uyarlamışlardır. Bu yöntemde her bir cümle bir düğüme atanmıştır. Düğümler arasındaki bağlantılar ise benzerlik fonksiyonları ile sağlanmıştır. En yüksek değere sahip cümleler dokümanı en iyi özetleyen cümleler olarak belirlenmiştir [14].

Çizge tabanlı çıkarımsal metin özetleme üzerine yapılan bir çalışmada, 4 farklı ilişkilendirme yönteminin "TextRank" algoritması üzerindeki etkisi incelenmiştir. Çalışmada DUC ve CAST veri setinden yararlanılmıştır. Bu çalışmaya ilave olarak hiyerarşik birleştirici kümeleme ve "TextRank" yöntemleri ile bir sistem geliştirilmiştir. Önerilen yöntemde cümleler belli bir ölçüte göre kümelenmektedir. Kümelerden cümle seçebilmek için "TextRank" algoritması uygulanmıştır. DUC 2002 veri seti ile yapılan çalışmalarda önerilen sistemin daha performanslı çalıştığı gözlemlenmiştir. CAST veri seti ile yapılan çalışmalar ise önerilen yöntemin 2 ilişkilendirme yöntemden daha performanslı olduğunu, diğer 2 yöntemle ise arasında az bir fark olduğu gözlemlenmiştir [15].

Kupeic ve diğ. [16], cümleleri puanlamak için cümlenin konumunun ilk 10 paragrafta olup olmadığı, büyük harfli kelimeleri barındırıp barındırmadığı, tematik kelime bulunup bulunmadığı gibi farklı özellikler belirlemişlerdir. Bu özellikler Bayes sınıflandırma algoritması ile eğitilerek bir sınıflandırıcı elde edilmiştir. Oluşturulan bu sınıflayıcı yeni bir dokümandaki hangi cümlenin önemli olduğunu tespit etmek için kullanılmıştır.

Kaynar ve diğ. [17], metin özetlemede cümle seçimine, benzerlik yöntemlerinin etkisini incelemiştir. Uygulamalarında TextRank algoritması ile birlikte Kosinüs, Jaccard ve Levenshtein benzerlik yöntemlerini kullanmışlardır. Levenshtein mesafesi benzerlik yöntemi olarak kullanıldığında diğer benzerlik yöntemlerine göre daha iyi sonuç verdiği gözlemlenmiştir. Cengiz ve diğ. [18], metin çizgeleri kullanarak cümlenin önemine göre doğrusal ağırlıklandırma ile bir metin özetleme yöntemi önermektedir.

3. Materyal ve Metot

Bu çalışmada farklı benzerlik ölçütleri kullanılarak oluşturulan çizgelerin merkezilik değerlerine göre çıkarımsal özet elde edilmiştir. Jaccard Benzerliği, Kosinüs Benzerliği, Pearson Korelasyon Katsayısı ve Levenshtein Benzerliği ölçütleri kullanılmıştır. Bu benzerlik yöntemleri ile cümlelerin birbirlerine olan benzerlikleri hesaplanarak benzerlik matrisinde tutulur.

3.1. Benzerlik ölçütleri

3.1.1. Kosinüs benzerliği

Cümlelerin birer vektör olarak ifade edildiği bu yaklaşımda, iki vektör arasındaki açının kosinüs değeri cümleler arasındaki ilişkiyi ifade etmektedir [19].

$$\cos(\theta) = \frac{A.B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

A ve B vektörleri arasındaki açının kosinüsü ne kadar az ise bu iki vektör birbirlerine o kadar benzer demektir.

3.1.2. Jaccard benzerliği

Jaccard benzerliği iki kümenin kesişimi ile birleşimlerinin birbirine bölünmesi şeklinde ifade edilir. Eğer benzerlik ölçütü 0 ise hiç benzemiyor, 1 ise tam benziyor demektir [20].

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

3.1.3. Pearson korelasyon katsayısı

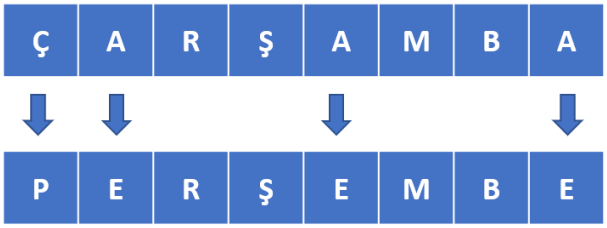
Pearson Korelasyon Katsayısı, iki sayısal ölçüm arasındaki doğrusal ilişkiyi ölçer. Pearson Korelasyon Katsayısı, +1 ile -1'lik arasında bir değer alarak ilişkinin şiddetini ve yönünü belirlemek için kullanılan bir istatistik yöntemidir. +1 değeri cümlelerin mükemmel ilişkili olduğunu, -1 ise cümlelerin birbirleri ile ilişkili olmadığı anlamına gelmektedir [21].

$$Pearson(A, B) = \frac{\sum AB - \frac{\sum A \sum B}{N}}{\sqrt{\left(\sum A^2 - \frac{(\sum A)^2}{N}\right)\left(\sum B^2 - \frac{(\sum B)^2}{N}\right)}} \quad (3)$$

3.1.4. Levenshtein mesafesi

Levenshtein mesafesi, iki kelimeyi birbirine benzetmek için minimum kaç harfte değişiklik yapılacağının (ekleme, silme veya değiştirme) sayısıdır. Şekil 1' de Levenshtein mesafesinin kelimeler için nasıl hesaplandığını gösterilmektedir. "ÇARŞAMBA" kelimesini "PERŞEMBE" kelimesine dönüştürebilmek için minimum 4 adet harfte değişiklik yapılması gerekmektedir. Bu nedenle "ÇARŞAMBA" ve "PERŞEMBE" kelimelerinin

birbirlerine olan Levenshtein mesafesi 4 olarak hesaplanmıştır. Cümleler arasındaki Levenshtein mesafesi hesaplanarak cümlelerin benzerlikleri hakkında bilgi sahibi olunabileceği varsayımı ortaya atılmıştır [17].



Şekil 1. Levenshtein mesafesi

3.2. Merkezilik ölçütleri

Merkezilik ölçütleri, bir ağ içindeki bazı düğümlerin konum ve bağlantıları nedeniyle diğer düğümlerden ne kadar önemli olduğunu belirlemek için kullanılmaktadır. Bavelas, sosyal ağdaki bir bireyin konumunun, bu kişinin diğer kişilere etkisini nasıl belirlediğini anlamak için merkeziyet kavramını ortaya atmıştır [22]. Merkezilik ölçümleri konusunda çok sayıda çalışma yapılmıştır. Her çalışma farklı bir fikri savunmaktadır. Merkezilik ölçümleri farklı alanlarda merkez düğümleri bulmak amacı ile kullanılmaktadır [23-24]. Bu çalışmada metin çizgelerinin en değerli düğümlerini tespit etmek amacıyla Arasındalık Merkeziliği, Yakınlık Merkeziliği, Derece Merkeziliği ve Özvektör Merkeziliği ölçümleri kullanılmaktadır.

3.2.1. Derece merkeziliği

Çizgedeki bir düğüm ile diğer düğümler arasındaki bağlantılarının sayısı olarak tanımlanmaktadır. Freeman, merkez düğümün derecesinin bağlı olduğu diğer düğümlerin sayısı olduğunu ortaya atmıştır. Düğüm derecesi bir ağı incelerken ilk adım olarak kullanılır [25]. N adet düğümden oluşan bir $G = (V, E)$ çizgesi olsun ve ν düğümünün derece merkeziliği $C_D(\nu)$ Denklem 4 ile hesaplanmaktadır.

$$C_D(v) = \frac{derece(v_i)}{N-1} \tag{4}$$

3.2.2. Arasındalık merkeziliği

Arasındalık merkeziliği, bir çizgedeki merkeziliği ölçmek için kullanılan yöntemlerden biridir. Bu yöntemdeki temel fikir x düğümünden y düğümüne en kısa yoldan giderken çoğunda V düğümünden de geçiliyorsa V düğümü önemlidir. Bu yöntemde önemli düğümler diğer düğümlerle bağlantılı olduğu varsayılır [26]. V düğümü için arasındalık merkeziliği Denklem 5 ile hesaplanmaktadır. N çizgedeki düğümler kümesidir.

$$C_b(v) = \sum_{x,y \in N} \frac{G_{x,y}(v)}{G_{x.y}} \tag{5}$$

3.2.3. Yakınlık merkeziliği

Yakınlık merkeziliği, düğümler arasındaki en kısa mesafelerin toplamıdır [27]. Yakınlık merkeziliği değeri, hesaplanmak istenen düğümün çizgede bulunan diğer düğümler ile olan en kısa yol mesafelerinin toplanması ile elde edilebilir [18]. V bir düğüm olmak üzere, v düğümünden diğer düğümlere gidilebilen en kısa yol olarak tanımlanır [28]. İki düğüm arasındaki en kısa mesafe $mesafe\ (Vi, Vj)$ olmak üzere yakınlık merkeziliği Denklem 6'daki gibi hesaplanmaktadır [27].

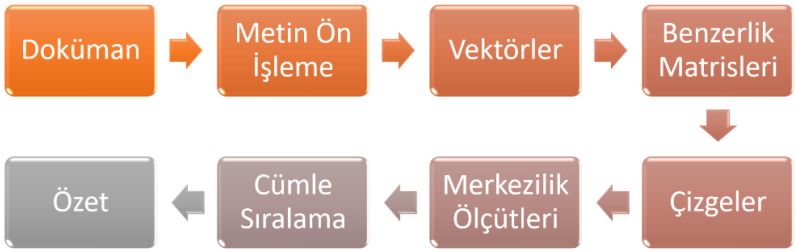
$$C_c(v_i) = \frac{N-1}{\sum_{V_j \in v} mesafe(v_i, v_j}} \tag{6}$$

3.2.4. Özvektör merkeziliği

Özvektör merkeziliği, bir çizgedeki düğümün önemini belirlemek için kullanılan yöntemlerden bir diğeridir. Bu yöntem önemli bir düğüm ile olan bağlantının düşük öneme sahip olan bir düğüme olan bağlantıdan daha değerli olduğu prensibine dayanmaktadır [28]. Google tarafından sunulan PageRank, özvektör merkeziliği kullanan bir yöntemdir [29]. Özvektör merkezliliği, Vj ve Vi düğümler arasındaki bağlantıların ağırlığı ve λ bir sabit olmak üzere, Denklem 7 ile hesaplanmaktadır [27].

$$C_e(v_i) = \frac{1}{\lambda}\sum_{v_j \in N(v_i)} v_{ij} x C_e(v_j) \quad (7)$$

Şekil 2'de, yapılan çalışmaya ait aşamalar genel hatlarıyla gösterilmektedir. Öncelikle özetlenecek dokümanlara metin ön işleme adımları uygulanır. Bu aşamada noktalama işaretleri ve anlam ifade etmeyen kelimeler cümlelerden çıkarılır. Ön işlemden geçen cümleler vektörlere dönüştürülür. Vektörler kullanılarak benzerlik yöntemleri ile birbirlerinden farklı benzerlik matrisleri oluşturulur. Bu çalışmada 4 farklı benzerlik yöntemi kullanılmıştır. Benzerlik matrisleri çizgelere dönüştürülür. Böylece her bir cümle çizgedeki bir düğüm ile temsil edilmiş olur. Benzerlik matrislerinde tutulan değerler çizgenin düğümleri arasındaki bağlantıların ağırlıklarıdır. Farklı benzerlik yöntemleri ile elde edilen çizgelerdeki merkezi düğümlerin tespit edilmesi için yukarıda bahsedilen merkezilik ölçümleri uygulanarak en değerli düğümler tespit edilir. Çalışmada farklı benzerlik yöntemi ve merkezilik ölçütleri kullanılarak 16 farklı özet oluşturulmuştur. Böylece benzerlik ve merkezilik ölçütleri birlikte kullanılarak oluşturulan özetler birbirleri ile karşılaştırılmıştır.



Şekil 2. Benzerlik yöntemleri ve merkezilik ölçütleri kullanarak metin özetleme

4. Deneysel Çalışmalar

Çalışmanın bu bölümünde farklı benzerlik ölçütleri kullanılarak oluşturulan çizgelerin merkezilik değerleri ile oluşturulan özetlerin başarısı değerlendirmektedir. Yöntemlerin başarısını test edebilmek için Document Understanding Conference (DUC-2002) veri seti kullanılmıştır. Bu veri setinde metin özetlemeye yönelik dokümanlar bulunmaktadır. Çalışma, çıkarımsal özetleme yöntemleri kullanılarak yapıldığından veri setinin ilgili dosyalarından yararlanılmıştır. DUC-2002 veri seti çıkarımsal özeti farklı değerlendiriciler tarafından oluşturulan 59 adet dosyaya sahiptir. Ayrıca bu dosyalara ait her birinden iki farklı değerlendiricinin oluşturduğu 200 ve 400 kelimelik özetler bulunmaktadır.

Çalışmada DUC-2002 veri setinden 10 adet dosyanın farklı benzerlik ölçütleri ile oluşturulan çizgelerin merkezilik değerleri ile özetleri oluşturulmuştur. Oluşturulan özetler, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) performans ölçütlerinden ROUGE-1, ROUGE-2, ROUGE L ve ROUGE-W-1.2 ile değerlendirilmiştir.

5. Sonuçlar

Bu çalışmanın amacı, farkı benzerlik ölçütleri ile oluşturulan çizgeler ve bu çizgelerin merkezilik değerleri ile çıkarımsal çizge tabanlı özet oluşturmaktır. DUC-2002 veri setinde yer alan metinler üzerinde deneysel çalışmalar gerçekleştirilmiştir. Metinler üzerinde çalışabilecek biçimlere dönüştürülmesi amacıyla bir takım metin ön işleme işlemlerine tabi tutulmuştur. TF-IDF algoritması kullanılarak metin içerisindeki cümleler vektör olarak temsil edilmiştir. Vektör olarak temsil edilen cümleler kosinüs benzerliği, Jaccard benzerliği, levenshtein benzerliği, pearson korelasyon katsayısı ölçütleri kullanılarak çizgeler oluşturulmuştur. Oluşturulan bu çizgelerdeki en değerli düğümleri tespit etmek amacı ile Derece Merkeziliği, Arasındalık Merkeziliği, Yakınlık Merkeziliği ve Özvektör Merkeziliği ölçütleri kullanılmıştır.

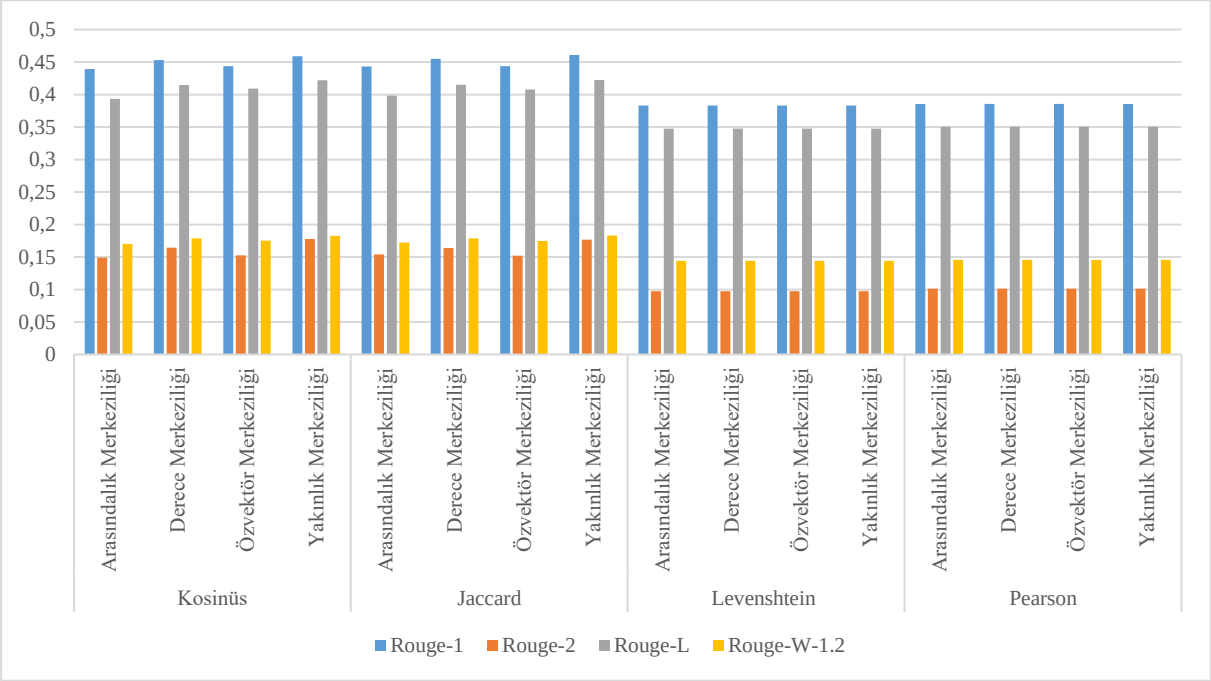
Literatürdeki en yaygın performans metrikleri kullanılarak elde edilen özetlerin başarısı hesaplanarak yaklaşımların performansı birbirleri ile karşılaştırılmıştır. Her bir yaklaşıma ait 200 ve 400 kelimelik özetler oluşturulmuştur. Gerçekleştirilen çalışma ile DUC-2002 veri setindeki metinlere ait referans özetler (200,400 kelimelik) ve bu çalışma ile oluşturulan özetler karşılaştırılarak ROUGE metrikleri kullanılarak elde edilen özetlerin performansını değerlendirilmiştir. Kosinüs Benzerliği, Jaccard Benzerliği, Levenshtein Benzerliği ve Pearson Korelasyon Katsayısı ölçütleri ile merkezilik değerlerine göre oluşturulan modellerin 200 kelimelik özetleme performanslarına Tablo 1'de, 400 kelimelik özetleme performanslarına ise Tablo 2'de yer verilmiştir.

Tablo 1. 200 kelimeden oluşan özetler için farklı benzerlik yöntemleri kullanılarak oluşturulan çizgelerin merkezilik ölçütlerine göre performansları

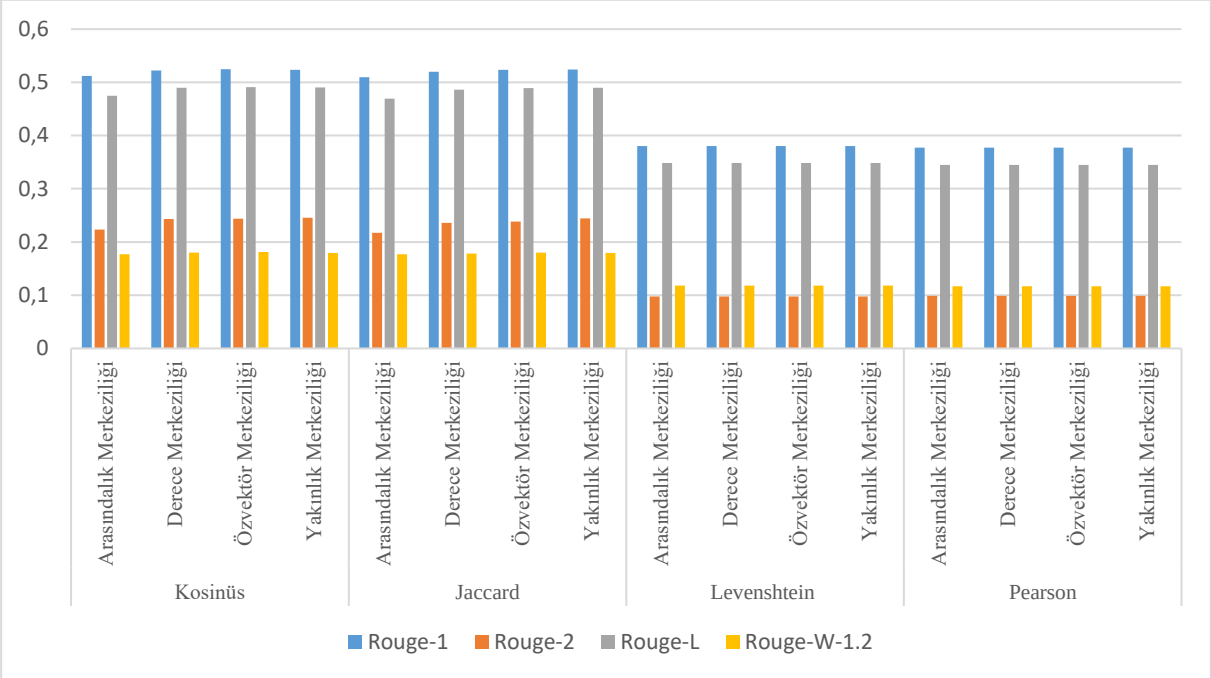
		Rouge-1	**Rouge-2**	**Rouge-L**	**Rouge-W-1.2**
Kosinüs	*Arasındalık Merkeziliği*	0,43934	0,14897	0,39362	0,17011
	Derece Merkeziliği	0,45309	0,16436	0,4145	0,17847
	Özvektör Merkeziliği	0,44369	0,15241	0,40893	0,17532
	Yakınlık Merkeziliği	0,45876	0,17765	0,42197	0,18264
Jaccard	*Arasındalık Merkeziliği*	0,44309	0,15411	0,39825	0,17218
	Derece Merkeziliği	0,45507	0,16362	0,41486	0,17874
	Özvektör Merkeziliği	0,44366	0,15213	0,40764	0,17487
	Yakınlık Merkeziliği	0,46091	0,17648	0,42249	0,18284
Levenshtein	*Arasındalık Merkeziliği*	0,38284	0,09753	0,34754	0,14433
	Derece Merkeziliği	0,38284	0,09753	0,34754	0,14433
	Özvektör Merkeziliği	0,38284	0,09753	0,34754	0,14433
	Yakınlık Merkeziliği	0,38284	0,09753	0,34754	0,14433
Pearson	*Arasındalık Merkeziliği*	0,38552	0,10106	0,3507	0,14576
	Derece Merkeziliği	0,38552	0,10106	0,3507	0,14576
	Özvektör Merkeziliği	0,38552	0,10106	0,3507	0,14576
	Yakınlık Merkeziliği	0,38552	0,10106	0,3507	0,14576

Tablo 2. 400 kelimeden oluşan özetler için farklı benzerlik yöntemleri kullanılarak oluşturulan çizgelerin merkezilik ölçütlerine göre performansları

		Rouge-1	Rouge-2	Rouge-L	Rouge-W-1.2
Kosinüs	*Arasındalık Merkeziliği*	0,51221	0,22353	0,47469	0,17705
	Derece Merkeziliği	0,52245	0,24289	0,48987	0,1798
	Özvektör Merkeziliği	0,52485	0,2439	0,49092	0,18123
	Yakınlık Merkeziliği	0,52357	0,24578	0,49026	0,17964
Jaccard	*Arasındalık Merkeziliği*	0,50963	0,21711	0,46918	0,17674
	Derece Merkeziliği	0,52013	0,23619	0,48588	0,17831
	Özvektör Merkeziliği	0,52335	0,23827	0,48894	0,18013
	Yakınlık Merkeziliği	0,52376	0,24448	0,48994	0,17958
Levenshtein	*Arasındalık Merkeziliği*	0,38037	0,09744	0,3483	0,11819
	Derece Merkeziliği	0,38037	0,09744	0,3483	0,11819
	Özvektör Merkeziliği	0,38037	0,09744	0,3483	0,11819
	Yakınlık Merkeziliği	0,38037	0,09744	0,3483	0,11819
Pearson	*Arasındalık Merkeziliği*	0,37712	0,09877	0,34471	0,11698
	Derece Merkeziliği	0,37712	0,09877	0,34471	0,11698
	Özvektör Merkeziliği	0,37712	0,09877	0,34471	0,11698
	Yakınlık Merkeziliği	0,37712	0,09877	0,34471	0,11698



Şekil 3. 200 kelimeden oluşan özetlerin ROUGE metriklerine göre F-skor değerleri



Şekil 4. 400 kelimeden oluşan özetlerin ROUGE metriklerine göre F-skor değerleri

Çizge tabanlı metin özetleme çalışmalarında cümlelerin birbirlerine olan benzerlikleri genellikle Kosinüs Benzerliği ve Jaccard Benzerliği ile hesaplanılır. Belwal ve diğ. [8] ve Yalkın [15], Kosinüs ve Jaccard benzerliklerinden yaralanarak özetler oluşturmuşlardır. Ayrıca Kaynar ve diğ. [17], farklı benzerlik yöntemlerinin metin özetlemeye etkileri üzerine çalışma yapmışlardır. Ancak bu çalışmada Kosinüs, Jaccard ve Levenshtein benzerlik yöntemlerinin yanı sıra Pearson Korelasyon katsayısı da kullanılarak çizgeler oluşturulmuştur. Elde edilen çizgelerin her biri için Derece Merkeziliği, Arasındalık Merkeziliği, Yakınlık Merkeziliği ve Özvektör Merkeziliği yaklaşımları ile en değerli düğümler tespit edilerek cümlelerin metindeki sıralarına göre özetler oluşturulmuştur.

Elde edilen sonuçlara göre Jaccard Benzerliği ile Özvektör Merkeziliği yaklaşımı ve Kosinüs Benzerliği ile Yakınlık Merkeziliği yaklaşımı birlikte kullanıldığında diğer yaklaşımlara göre daha başarılı özetler elde edildiği görülmüştür.

6. Tartışma

Bu çalışmada, çıkarımsal metin özetleme yapmak amacıyla çizge tabanlı bir süreç sunulmaktadır. Dokümanlar farklı benzerlik yöntemleri kullanılarak çizge temsilleri oluşturulmuş ve en değerli cümlelerin tespit edilmesi için düğüm merkezilik değerleri hesaplanmıştır. Merkezilik ölçümleri en yüksek değerli olan cümleler seçilerek 200 kelimelik ve 400 kelimelik özetler oluşturulmuştur. Elde edilen özetler birbirleri ile karşılaştırılmıştır.

Tablo 1' de görüldüğü üzere çizgelerin merkezilik değerleri ile oluşturulan 200 kelimelik özetlerde Jaccard Benzerliği ve Yakınlık merkeziliği yaklaşımı diğer yaklaşımlara göre başarılı performans göstermiştir. Bu yaklaşım ile elde edilen ROUGE-1 metriğinin F-skor değeri 0,46091 olarak hesaplanmıştır.

Tablo 2' de görüldüğü üzere ise Kosinüs Benzerliği kullanılarak oluşturulan çizgenin Özvektör Merkeziliği değerleri ile oluşturulan 400 kelimelik özetler en yüksek performansı göstermiştir. ROUGE-1 metriğinin F-skor değeri 0,52485 olarak hesaplanmıştır. Özvektör merkezilik değeri yüksek olan düğümler hem kendileri hem de komşuları çok fazla bağlantıya sahiptir. Bu da o düğümün önemli ve etkili olduğunu ifade eder. En yüksek performans özvektör merkeziliği yaklaşımı ile ölçülmüştür.

Tablolar ve şekillerde görüldüğü üzere kullanılan performans metrikleri açısından, farklı benzerlik ölçütleri kullanılarak oluşturulan çizgelerin düğüm merkezilik değerleri ile elde edilen özetlerin performansı metin özetlemede kullanışlı ve etkin bir yöntem olduğunu göstermiştir.

Kaynaklar

[1] Sunitha C, Jaya A, Ganesh A. A study on abstractive summarization techniques in Indian languages. Procedia Computer Science. 2016;87:25-31.
[2] Luhn HP. The automatic creation of literature abstracts. IBM Journal of research and development. 1958;2(2):159-65.
[3] Nenkova A, McKeown K. Automatic summarization: Now Publishers Inc; 2011.
[4] Çelik Ö, Koç BC. TF-IDF, Word2vec ve Fasttext Vektör Model Yöntemleri ile Türkçe Haber Metinlerinin Sınıflandırılması. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi. 2021;23(67):121-7.
[5] Dharma EM, Gaol FL, Leslie H, Warnars H, Soewito B. The accuracy comparison among Word2vec, Glove, and Fasttext towards convolution neural network (CNN) text classification. J Theor Appl Inf Technol. 2022;31(2).
[6] Gautam AK, Bansal A. Effect of Features Extraction Techniques on Cyberstalking Detection Using Machine Learning Framework. Journal of Advances in Information Technology Vol. 2022;13(5).
[7] El-Kassas WS, Salama CR, Rafea AA, Mohamed HK. Automatic text summarization: A comprehensive survey. Expert Systems with Applications. 2021;165:113679.
[8] Belwal RC, Rai S, Gupta A. A new graph-based extractive text summarization using keywords or topic modeling. Journal of Ambient Intelligence and Humanized Computing. 2021;12(10):8975-90.
[9] Joshi A, Fidalgo E, Alegre E, Alaiz-Rodriguez R. RankSum—An unsupervised extractive text summarization based on rank fusion. Expert Systems with Applications. 2022;200:116846.
[10] Azadani MN, Ghadiri N, Davoodijam E. Graph-based biomedical text summarization: An itemset mining and sentence clustering approach. Journal of biomedical informatics. 2018;84:42-58.
[11] Edmundson HP. New methods in automatic extracting. Journal of the ACM (JACM). 1969;16(2):264-85.
[12] Lin C-Y, editor Rouge: A package for automatic evaluation of summaries. Text summarization branches out; 2004.
[13] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information processing & management. 1988;24(5):513-23.
[14] Mihalcea R, Tarau P, editors. Textrank: Bringing order into text. Proceedings of the 2004 conference on empirical methods in natural language processing; 2004.
[15] Yalkın C. Çizge tabanlı metin özetleme Yüksek Lisans Tezi. Fırat Üniversitesi, 2014.
[16] Kupiec J, Pedersen J, Chen F, editors. A trainable document summarizer. Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval; 1995.
[17] Kaynar O, Işik YE, Görmez Y, editors. Graph based automatic document summarization with different similarity methods. 2017 25th Signal Processing and Communications Applications Conference (SIU); 2017: IEEE.
[18] Cengiz H, Uckan T, Seyyarer E, Karci A, editors. Graph-based suggestion for text summarization. 2018 International Conference on Artificial Intelligence and Data Processing (IDAP); 2018: Ieee.
[19] Singhal A. Modern information retrieval: A brief overview. IEEE Data Eng Bull. 2001;24(4):35-43.
[20] Bag S, Kumar SK, Tiwari MK. An efficient recommendation generation using relevant Jaccard similarity. Information Sciences. 2019;483:53-64.
[21] Zhou H, Deng Z, Xia Y, Fu M. A new sampling method in particle filter based on Pearson correlation coefficient. Neurocomputing. 2016;216:208-15.
[22] Bavelas A. A mathematical model for group structures. Human organization. 1948;7(3):16-30.
[23] Erkan G, Radev DR. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research. 2004;22:457-79.
[24] Kutlu M, Cığır C, Cicekli I. Generic text summarization for Turkish. The Computer Journal. 2010;53(8):1315-23.
[25] Freeman LC. Centrality in social networks conceptual clarification. Social networks. 1978;1(3):215-39.
[26] Feo TA, Resende MG, Smith SH. A greedy randomized adaptive search procedure for maximum independent set. Operations Research. 1994;42(5):860-78.
[27] Boudin F, editor A comparison of centrality measures for graph-based keyphrase extraction. International joint conference on natural language processing (IJCNLP); 2013.
[28] Kosorukoff A. Social network analysis: theory and applications: Passmore, D. L; 2011.
[29] See A, Liu PJ, Manning CD. Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:170404368. 2017.