TÜRKÇE METİNLERDE OTOMATİK ÖZETLEME

Dil İşleme Yaklaşımları ve Çıkarımsal Yöntemler

DR. ERTÜRK ERDAĞI
DOÇ. DR. VOLKAN TUNALI

EĞİTİM
yayınevi

TÜRKÇE METİNLERDE OTOMATİK ÖZETLEME
Doğal Dil İşleme Yaklaşımları ve Çıkarımsal Yöntemler

Dr. Ertürk Erdağı (İstanbul Medeniyet Üniversitesi)

Doç. Dr. Volkan Tunalı (University of the West of Scotland)

Yayınevi Grubu Genel Başkanı: Yusuf Ziya Aydoğan (yza@egitimyayinevi.com)
Genel Yayın Yönetmeni: Yusuf Yavuz (yusufyavuz@egitimyayinevi.com)
Sayfa Tasarımı: Kübra Konca Nam
Kapak Tasarımı: Eğitim Yayınevi Tasarım Bilimi

T.C. Kültür ve Turizm Bakanlığı
Yayıncı Sertifika No: 76780

E-ISBN: 978-625-6382-34-3
1. Baskı, Ağustos 2025

Kütüphane Kimlik Kartı

TÜRKÇE METİNLERDE OTOMATİK ÖZETLEME
Doğal Dil İşleme Yaklaşımları ve Çıkarımsal Yöntemler

Dr. Ertürk Erdağı, Doç. Dr. Volkan Tunalı
V+102 s., 135x215 mm
Kaynakça var, dizin yok.
E-ISBN: 978-625-6382-34-3

Bu kitap "Türkçe metinlerde çıkarım tabanlı otomatik metin özetleme (2023)" adlı doktora tezinden türetilmiştir.

Copyright © Bu kitabın Türkiye'deki her türlü yayın hakkı Eğitim Yayınevi'ne aittir. Bütün hakları saklıdır. Kitabın tamamı veya bir kısmı 5846 sayılı yasanın hükümlerine göre kitabı yayımlayan firmanın ve yazarlarının önceden izni olmadan elektronik/mekanik yolla, fotokopi yoluyla ya da herhangi bir kayıt sistemi ile çoğaltılamaz, yayımlanamaz.

EĞİTİM
yayınevi

Yayınevi Türkiye Ofis: İstanbul: Eğitim Yayınevi Tic. Ltd. Şti., Atakent mah. Yasemen sok. No: 4/B, Ümraniye, İstanbul, Türkiye

Konya: Eğitim Yayınevi Tic. Ltd. Şti., Fevzi Çakmak Mah. 10721 Sok. B Blok, No: 16/B, Safakent, Karatay, Konya, Türkiye
+90 332 351 92 85, +90 533 151 50 42
bilgi@egitimyayinevi.com

Yayınevi Amerika Ofis: New York: Egitim Publishing Group, Inc.
P.O. Box 768/Armonk, New York, 10504-0768, United States of America
americaoffice@egitimyayinevi.com

Lojistik ve Sevkiyat Merkezi: Kitapmatik Lojistik ve Sevkiyat Merkezi, Fevzi Çakmak Mah. 10721 Sok. B Blok, No: 16/B, Safakent, Karatay, Konya, Türkiye
sevkiyat@egitimyayinevi.com

Kitabevi Şubesi: Eğitim Kitabevi, Şükran mah. Rampalı 121, Meram, Konya, Türkiye
+90 332 499 90 00
bilgi@egitimkitabevi.com

İnternet Satış: www.kitapmatik.com.tr
bilgi@kitapmatik.com.tr

EĞİTİM YAYINEVİ GRUBU

İÇİNDEKİLER

KISALTMALAR

DUC	**:** Document Understanding Conference
GloVe	**:** Global Vectors for Word Representation
HTML	**:** Hyper Text Markup Language
IDF	**:** Inverse Document Frequency
NLTK	**:** Natural Language Toolkit
ROUGE	**:** Recall-Oriented Understudy for Gisting Evaluation
ROUGE 1-F	**:** ROUGE-1 F Skoru
ROUGE 1-K	**:** ROUGE-1 Keskinlik Skoru
ROUGE 1-G	**:** ROUGE-1 Geri Çağırma Skoru
ROUGE 2-F	**:** ROUGE-2 F Skoru
ROUGE 2-K	**:** ROUGE-2 Keskinlik Skoru
ROUGE 2-G	**:** ROUGE-2 Geri Çağırma Skoru
SIGIR	**:** Special Interest Group on Information Retrieval
TBF	**:** Ters Belge Frekansı
TF	**:** Term Frequency
TIDSumm	**:** Tor Illegal Documents Summarization
TS	**:** Terim Sıklığı

BÖLÜM 1.

GİRİŞ

Çevrimiçi ortamda gün geçtikçe bilginin miktarı ve boyutu artmaktadır. Artan bu verideki boyut parametresi verinin yalnızca kapladığı alan olarak değil, birçok açıdan farklılaşan özelliği ile ilgilidir. Çevrimiçi ortamda bulunan yüksek miktar ve hacimli veri içerisinden önemli bilginin elde edilmesi son yıllarda çözülmesi gereken bir problem olarak karşımıza çıkmaktadır. Önemli bilginin elde edilmesi süreci metin tabanlı bir verinin özetlenmesi olarak örnek verilebilir. Özetleme işlemi bir veri kaynağı içerisinden önemli bilgilerin alınarak kısaltılma işlemi olarak tanımlanabilir (Çetiner, 2002).

Özetleme işlemi haber metinleri ve makale gibi yüksek hacimde metin bulunduran yapılar içerisinden önemli bilginin elde edilmesinde sıklıkla kullanılmaktadır. İnsanlar tarafından yapılan özetleme süreci, kişinin metnin tamamının okuması, sonrasında metnin bahsettiği konuya ilişkin çıkarımda bulunması ve metindeki cümleyi değiştirmeden alması ya da kendi düşünceleri ile birlikte yorumlayarak yeni sözcüklerle ifade etmesi şeklinde özetlemesi olarak sıralanmaktadır (Hark, Seyyarer, Uçkan, Myo, ve Karci, 2017). Bu durum zaman ve maliyet unsurları açısından olumsuz bir etki oluşturmaktadır. Bunun önüne geçmek adına otomatik metin özetleme sistemleri son yıllarda yaygın olarak kullanılmaktadır. Doğal Dil İşleme alanında kullanılan bu sistem verinin boyutunun giderek artması, farklı yazım tekniklerinin olması sebebiyle geliştirilmeye ihtiyaç duymaktadır.

Otomatik metin özetleme sistemlerinde temel olarak iki yöntem bulunmaktadır (Joshi, Fidalgo, Alegre, ve Fernández-Robles, 2019). Bu yöntemlerden ilki metin içerisinde çeşitli yöntemler ile önemli olarak tespit edilen cümle ya da cümlelerin herhangi bir değişiklik yapılmadan özet kısmına alınmasıdır. Çıkarım tabanlı özetleme adı verilen bu yöntem literatürde geniş kullanım alanına sahiptir. İkinci yöntem ise metnin çeşitli yöntemlerle anlamlandırılması ve özet kısmının yeni sözcükler kullanılarak yazılması sürecini kapsamaktadır. Bu yöntem ise soyutlama tabanlı özetleme olarak adlandırılmaktadır.

Soyutlama tabanlı özetleme, çıkarım tabanlı özetlemeye göre daha az kullanım ve çalışma alanına sahiptir. Bunun sebebi soyutlama tabanlı özetlemede yeni sözcüklerden cümle oluşturma sürecinde anlamsal problemlerin olması ve dilbilgisi kurallarının tam anlamıyla uygulanamama durumunun olmasıdır. Soyutlama tabanlı özetlemedeki bu durum ana metinden kısmen bağımsız olma özelliğini taşımaktadır. Bu yöntemde metnin anlamsal ve dilbilgisi özelliklerinin yanında metne dair kapsamın doğru olarak yansıtılması ve kendi içerisinde tutarlılığının yüksek olması hedeflenmektedir. Zor olarak görülen bu hedef için çalışma yapılan dil içerisinde zengin bir sembolik kelime yapısı bulunmalıdır.

Çıkarım tabanlı özetleme temelde metnin çeşitli yöntemlerle kısaltılmış ve anlamlı bir bütününü yakalayabilmeyi hedefler. Bu hedefi soyutlama tabanlı özetlemeye göre daha hızlı ve düşük maliyette gerçekleştirebilmektedir. Kullanım alanının fazla olmasının temel sebebi sağladığı bu avantajlar olarak sıralanabilmektedir. Bu yöntemde metin içerisindeki unsurlar çeşitli yöntemler ile derecelendirilmekte, istatistiki ve sezgisel yaklaşımlar ile metin parçalarının özet kısmında yer alması için seçilmektedir.

Her iki yöntemde de metin özetlemeleri tekli ve çoklu doküman olmak üzere ikiye ayrılmaktadır. Tek bir doküman üzerinden bir metnin özetlenmesi yapılabileceği gibi, kapsamının aynı olduğu birden fazla doküman içerisinden

de özetlemenin yapılarak tek bir ürün elde edilmesi de mümkündür. Çoklu doküman özetlemede verilerin saklanma durumu ve farklı yazım tekniklerinin olması özetleme işlemi sonucunda tek bir dokümana indirgenmesi kısmi olarak dezavantaj oluşturmaktadır.

Yöntem ve modeller ışığında otomatik metin özetleme işlemi sondan eklemeli dil yapısında olan Türkçe için sınırlamalar barındırmaktadır. Bu sınırlamalar çalışmaların son yıllarda biraz daha artmasıyla kısmen de olsa kalkmasına karşın, İngilizce üzerine yapılan çalışmalar henüz yakalanamamıştır. Bu kısımda İngilizce dilindeki yapısal kolaylık ve kullanım alanının genişliğinin sağladığı avantaj büyük rol oynamaktadır.

Otomatik metin özetlemede amaç, insanlar tarafından oluşturulan özet ile karşılaştırılabilir nitelikte tutarlı ve anlamlı bir özet oluşturmaktır. Çalışma kapsamında bu amaç doğrultusunda metin özetleme yöntemlerinden çıkarım tabanlı metin özetleme için mevcut durumda kullanılan algoritmaların deneysel olarak incelenmesi ve bu incelemeler ışığında yeni bir algoritma oluşturulması veya mevcut bir algoritmada zaman veya maliyet konularında iyileştirme yapılması amaçlanmaktadır.

Literatüre kazandırılması hedeflenen iyileştirme veya yenilik çalışmasına ilişkin çalışmayı sınırlandırabilecek ya da engel olabilecek bazı hususlar bulunmaktadır.

Eş Dizim Problemi: Metin içerisinde aynı anlam ve kapsam içerisindeki cümlelerin birçok kez tekrarlanmasından kaynaklanan, cümlenin özet kapsamında yer almaması gerekirken yüksek frekans değerinden ötürü özette yer alması (Önal, 2019).

Türkçe'nin Sondan Eklemeli Dil Olması: Türkçe'de yapım ve çekim ekleri kelimenin sonuna eklenir. Pekiştirme harici kullanılan eklerde sondan eklenen bu eklerin ayrıştırılması zorluğu bulunmaktadır (Tülek, 2007).

Anahtar Sözcük: Metnin ana unsurları denilebilecek anahtar kelimeler özet oluşturma için bir avantaj sağlarken tüm metin yapılarında anahtar sözcük bulunmamaktadır. Öncelerde yalnızca akademik çalışmalar içerisinde kullanılan anahtar sözcükleri şu anda haber sitelerinde arama motoru optimizasyonu için kullanılmaktadır. Fakat bir kitaptaki metin parçasında bu yapının olmaması temel anlamda bu kıstası kullanan yöntemler için dezavantaj oluşturmaktadır (Uzundere ve Dedja, 2008)

Düzensiz Metin: Özellikle çevrimiçi paylaşım sitelerinde yazıların belli bir düzen olmaksızın; noktalama işaretleri ve imla kurallarına dikkat edilmeden oluşturulan metinler bu ortam üzerinde yapılan özetleme çalışmalarında dezavantaj oluşturmaktadır (Doğan, 2019).

Paragraf Yapısının Olmaması: Metnin ayrıştırılarak puanlanması ve paragraf içerisinde en önemli cümlenin tespiti gibi istatistiksel yaklaşımlar paragraf yapısı olmadan (tek bir parça halinde) yazılan metinler için ayrıştırma işleminin yapılmaması çalışmalarda bir engel olarak görülmektedir (Turan, 2015).

Özet Miktarı: Çıkarım tabanlı metin özetleme yaklaşımlarımla ana metnin belli bir oranda sıkıştırılması hedeflenmektedir. Bu hedef doğrultusunda istatistiki ve sezgisel yaklaşımlarla yapılan puanlamalar verilen oran doğrultusunda en yüksek değerli metin parçalarının seçilmesi işlemi yürütülmektedir. Bu durum sayısal bir ifade ile metnin kapsamı konusunda net bir fikir vermeyebilir ve istenilen asıl ifadeler özet kısmında yer almayabilir (Gündoğdu ve Duru, 2016).

Veri Seti Problemi: Türkçe dilinde özetleme çalışmalarında kullanılmak üzere veri setleri sınırlı bir sayıda bulunmaktadır. Yapılan çalışmaların son yıllarda artması bir nebze de olsa bu sayının artmasını sağlarken, özellikle Türkçe'deki çalışma zorluğu veri setlerinde de sınırlamalara sebep olmuştur (Gündoğdu ve Duru, 2016).

Bu tez çalışmasında Türkçe metinlerde çıkarım tabanlı özetleme üzerinde deneysel çalışmalar yürütülmüştür. Yürütülen çalışmada Türkçe diline özgün bir durumun olup olmadığı incelenmiş, literatürde bulunan yöntemlere ek olarak sunulabilecek katkı ya da geliştirilebilecek hususlar üzerinde inceleme yapılmıştır. Bu çalışma kapsamında;

- Veri seti üzerinde çeşitli önişleme aşamaları gerçekleştirilmiş, veri düzenleme, aykırı verilerin kaldırılması gibi teknikler uygulanmıştır.
- Önerilecek yöntem ve tekniklerin karşılaştırmalarını yapabilmek adına bir çalışma grubu kurulmuş, üç kişilik bu grup ile insan özetleri oluşturulmuştur.
- Metin özetleme çalışması için literatürde bulunan yöntemler incelenmiş, veri seti üzerinde bu yöntemler uygulanmıştır.
- Türkçe diline özgü bir farklılık ortaya konulması adına büyük ünlü ve küçük ünlü uyumlarının dilbilgisi yönünden kuralları incelenmiş ve veri seti üzerinde uygulanabilmesi için programlanmıştır.
- Literatürde bulunan yöntemler ve tez çalışması kapsamında önerilen yöntemlerin başarım değerlendirmesi için iki farklı metrik üzerinde çalışma yapılmıştır.
- Elde edilen özetler bir sınıflandırma yaklaşımıyla değerlendirilmiş ve dört farklı sınıflandırma algoritması ile elde edilen sonuçlar değerlendirilmiştir.
- Sonuçlar Türkçe dili için önerilen üç yöntemin literatürde bulunan diğer yöntemlerden daha başarılı bir durum sergilemiş, literatüre bu noktada katkı sağlamıştır.

Tez beş bölümden oluşmaktadır. Bölümlere ilişkin tanımlayıcı bilgiler aşağıda sunulmuştur.

Birinci bölümde konuya ilişkin genel kavramlar, tezin literatüre katkısı ve amacı açıklanmıştır.

İkinci bölümde tez çalışması için literatür taramasına yer verilmiş, farklı dil ve farklı yöntemlerin sonuçları sunulmuştur.

Tezin üçüncü bölümünde otomatik metin özetleme kavramı, çeşitleri ve yapısal farklılıkları ortaya konulmuştur.

Dördüncü bölümde tezin çalışma kapsamında uygulanan deneysel yöntem ve teknikler açıklanmış, önerilen yöntem ve teknikler örneklerle sunulmuş, elde edilen sonuçlar metriklerle değerlendirilmiş ve karşılaştırmalar yapılmıştır.

Beşinci bölümde yapılan tüm çalışmanın katkısı özetlenmiş ve yapılacak yeni çalışmalar için öneriler sunulmuştur.

BÖLÜM 2.

LİTERATÜR TARAMASI

Otomatik metin özetleme çalışmalarında çıkarım tabanlı ve soyutlama tabanlı olmak üzere iki yaklaşım bulunmaktadır. Çıkarım tabanlı yaklaşım mevcut metin parçası içerisinden çeşitli unsurlar göz önüne alınarak yapılan çalışma neticesinde önemli olduğu düşünülen cümle ya da paragrafın alınmasına dayanmaktadır. Soyutlama tabanlı yaklaşımda ise metnin çeşitli yöntemler ile anlamlandırılması ve sonrasında özet için ana metindeki cümle ya da paragrafın aynısını alarak değil, kişilerin günlük hayattaki kendi cümleleri ile ifade etme yolu bulunmaktadır. İki farklı yaklaşımın bu durumu soyutlama tabanlı özetleme çalışmalarındaki birçok ek çalışmayı da beraberinde getirmektedir.

Otomatik metin özetleme çalışmaları 1958 yılında Luhn'un çalışmasıyla başlamıştır. Literatürdeki ilk çalışma özelliğini taşıyan bu çalışmada özetleme işlemi iki adımda yapılmıştır. İlk adımda doküman içerisindeki metin Ön işlemeye alınmıştır. Ön işleme kapsamında durak kelimeler olarak adlandırılan metin içerisinde etkileyici anlama sahip olmayan bağlaç, zamir vb. kısımlar temizlenmiştir. Temizleme sonrasında önek ve harf sayısı gibi ölçütler üzerinde gruplama işlemi yapılmıştır. İkinci adımda gruplama yapılan kelimelerden yüksek frekanslı olanlar seçilmiştir. Burada yüksek frekansın alınmasının sebebi bir terimin metin içerisinde çok sık geçmesi bu terimin önemli olduğunun düşünülmesidir. Yüksek frekanstaki kelimelerin

bulundukları cümlelerin özet içerisinde geçmesi sağlanmıştır. Literatüre Term Frequency – TF (Terim Sıklığı – TS) kavramının kazandırılmasına ve metin özetleme çalışmalarında günümüzde de kullanımına olanak sağlamıştır (Luhn, 1958).

Baxendale metin içerisinde bulunan cümlenin bulunduğu pozisyonun önem derecesinde büyük rol oynadığı fikrini öne sürmüştür. Metin üzerinde vurgulayıcı ve dikkat çekici yönün genellikle ilk cümle ya da son cümlede verilmesi fikrinden yola çıkılarak yapılan çalışmada özellikle isimlerin ve basit tamlamaların ön derecesini artıran etmen olduğunu gözlemlemiştir (Baxendale, 1958).

Edmundson tarafından yapılan çalışmada Luhn'un çalışması (Luhn, 1958) doğrultusunda kelimelerin bulunma sıklığı üzerinde çalışmıştır. Terim sıklığına ek olarak metnin başlık bölümündeki kelimelerin metin içerisinde bulunma durumunun önem derecesini artırdığı ve anahtar kelime yaklaşımı gibi parametreler dâhil edilmiştir (Edmundson, 1969).

Jones literatüre IDF (Inverse Document Frequency) – TBF (Ters Belge Frekansı) kavramını kazandırmıştır. Luhn'un çalışmasındaki (Luhn, 1958) terim sıklığı yönteminin normalizasyon işlemi uygulandıktan sonra kullanılmasının daha yararlı olduğu fikrini ortaya atmıştır. Bu noktada ana metin içerisinde çok sık geçen kelimelerdeki ağırlık değerlerinin düşürülmesi, daha az frekansta bulunan kelimelerin ağırlık değerini yükseltilmesi yönünde bir çalışma yürütmüştür. Luhn'un çalışması ile birlikte kendilerinden sonraki çalışmalar için TFxIDF kavramının birlikte kullanımına olanak sağlamıştır (Jones, 1972).

Salton ve diğ. tarafından yapılan çalışmada kelimelerin ve cümlelerin özet içerisinde yer alma durumları tahmini istatistiki yöntemler kullanılarak gerçekleştirilmiştir. Çalışma kapsamında kullanılan graf yapısında kelime ve metin parçalarının ağırlıklarının tahmini yöntemler ortaya konularak gerçekleştirilmesi ve denetimli öğrenme yöntemi ile desteklenerek insanların yaptıkları özet ile karşılaştırılmasıyla

elde edilmiştir (Salton, Singhal, Buckley, Mitra, ve Mitra, 1996).

Fukumoto ve Suzuki özetlemede yapısal olarak paragrafları kullanmışlardır. Paragraf öbekleri içerisinde özette yer alacak kısmın paragrafta en fazla frekansa sahip terim üzerinden konu kapsamı dâhilinde gerçekleştirmiştir (Fukumoto ve Suzuki, 2000).

Carbonell ve Goldstein metin içerisinden, belirlenen bazı ölçütler doğrultusunda sorgu temelli yapılabilen özetleme çalışmalarına alternatif olarak herhangi bir sorgulama olmaksızın metin içerisindeki kelime, cümle, paragraf ya da metin parçalarının konuyu temsiline ilişkin derecelendirme üzerinden otomatik olarak yapılmasını sağlayan yöntem önermişlerdir. Çalışmada The Maximal Marginal Relevance (Maksimum Marjinal Alaka Düzeyi) isminde en yüksek düzeyde ilişkili öğelerin tespitine yönelik bir ölçüt önermişlerdir (Carbonell ve Goldstein, 1998).

Hovy ve Lin, SUMMARIST ismini verdikleri yöntem kapsamında özetleme çalışması için üç aşama belirlemişlerdir. Bu aşamalardan ilki metne dair konu kapsamının belirlenmesidir. İkinci aşama metin içerisindeki kavramlara ilişkin çıkarım tabanlı soyutlama çalışmasıdır. Üçüncü ve son aşama ise iki aşamada elde edilen değerler doğrultusunda özetleme işleminin gerçekleştirilmesidir (Hovy ve Lin, 1996).

Jaruskulchai ve Kruengkrai, Tay dilinde ana metinden konuyu en iyi yansıtabilecek paragrafın belirlenmesine yönelik çıkarım tekniği uygulamışlardır. Çıkarım tabanlı metin özetleme yaklaşımı doğrultusunda önerilen yöntemde bir paragrafta önemli kelimelerin tespitine yönelik hem yerel hem de genel özelliklerin keşfedilmesi gerektiği düşünülmüştür. Yerel özellik olarak paragrafın kendi bütünlüğü içinde tespit edilebildiği konuyu yansıtacak en önemli sözcük kümesinin belirlenmesi iken, genel özellikte paragraflar arasındaki ilişkiden faydalanılarak oluşturulmuştur. Terim sıklığı yerine konum gibi farklı derecelendirme yöntemi kullanılmıştır. Daha

sık kullanılan kelimelerin daha kısa olma eğiliminde olduğu varsayımı doğrultusunda çalışma yürütmüşlerdir. Global özellik derecelendirme kapsamında her bir paragrafın graf üzerinde temsil edilmesi ve kosinüs benzerliği ile paragrafların birbirleri ile ilişkileri derecelendirilmiştir (Jaruskulchai ve Kruengkrai, 2003).

Mihalcea ve Tarau, TextRank isminde bir yöntem önermişlerdir. Bu öneri kapsamında metin içerisindeki cümleler puanlanmakta ve bu puan ağırlığına göre özetleme oluşturulmaktadır. Çalışmada PageRank yönteminden (Page ve Nicholson, 1998) esinlenilmiştir. Bu yöntemde ağ üzerinde bulunan sayfaların birbirlerine verdikleri bağlantının izlenmesi üzerinde derecelendirme yapılmaktadır. Önerilen TextRank yöntemi de metin içerisindeki cümlelerin birbirleri ile olan bağlantılarını ve ilişkilerini modelleyerek değerlendirilmiştir (Mihalcea ve Tarau, 2004).

Orrú tarafından Portekiz dilindeki haber metinlerinin özetlenmesi üzerinde çalışma yapılmıştır. Yapay Sinir Ağları temelinde yapılan çalışmada metin özetlemede en sık kullanılan yöntemlerinden biri olan metin parçası pozisyonu kullanılmıştır (Orrú, Rosa ve Netto, 2006).

Medelyan tarafından yapılan çalışmada graf temelli anlamsal yaklaşım ile birlikte metin içerisinde geçen terimlerin düğümler üzerine yerleştirilmesi ve düğümler arasındaki kenar ağırlıklarının anlamsal ilişkiler üzerine temellendirilmesi yer almıştır. Bu durumda en kısa yol - en uzun yol ile güç ve zayıf bağlar arasındaki ilişkiler kullanılarak özetlemede yer alacak metin parçalarının tespiti gerçekleştirilmiştir (Medelyan, 2007).

Uzundere ve Dedja tarafından yapılan çalışmada denetimli bir yöntem yaklaşımı benimsenmiştir. Metin içerisindeki kelime, cümle ve paragraf yapıları puanlanmış ve puanlama doğrultusunda en yüksek değerlerin özet içerisinden yer alması sağlanmıştır. Elde edilen sonuçlar insanların yapmış oldukları

özetler ile karşılaştırılmış ve %55'lik bir başarı değerine ulaşılmıştır (Uzundere ve Dedja, 2008).

Adalı Türkçe metin özetleme için alan ontolojisi ve varlık ontolojisinin birlikte kullanıldığı tümleşik bir mimari önermiştir. Yöntem kapsamında doküman içerisindeki yapısal birtakım özelliklerden yola çıkılarak özetleme çalışması yapılmasının üzerinde durmuştur. Metin parçaları üzerinde varlık yönetimi işlemi sağlandığı yöntem ile elde edilen varlıklar orijinal metin üzerinde karşılaştırmıştır. 2000 doküman üzerinde 23426 varlık çıkarımı yapmıştır. Çıkarımı yapılan varlıkların ayrıca ilişki durumları da kontrol edildiği tümleşik yöntemin değerlendirmesinde başarılı bir sonuç elde etmiştir (Adalı, 2009).

Aliguliyev metinlerin kümelenmesi işlemini Parçacık Sürü Optimizasyon Algoritması kullanılarak gerçekleştirmiştir. Çoklu doküman özetleme çalışması üzerinde gerçekleştirilen bu yöntemde kümeleme neticesinde kümelerin birbirleri ile olan benzerliklerinin optimize edilmesine dayanan bir yöntem önermiştir (Aliguliyev, 2010).

Pembe ayrı ayrı kullanılan doküman yapısı ve bilgi çıkarım yönteminin beraber kullanılmasını önermiştir. İnternet üzerinden elde edilen dokümanların başlık ve alt başlık yapısıyla birlikte Ön işleme uygulanması aşaması ile başlayan yöntemde sıralı ağaç yapısını kullanmıştır. Kural tabanlı yaklaşım ile destek vektör makinesi yardımıyla oluşturulan yapı sonrasında cümle bazında ve paragraf bazında olmak üzere iki farklı puanlama gerçekleştirmiştir. Yapılan çalışma hem Türkçe hem de İngilizce dilindeki veri seti üzerinde denenmiştir. Önerilen özetleri, Google özeti ve doküman yapısının kullanılmadığı benzer çalışmalar ile karşılaştırdığında yüksek performans sonucu elde etmiştir (Pembe, 2010).

Huang ve diğ. tarafından yapılan çalışmada metin içerisindeki bilgi kapsamı, terim ve cümlelerin önemi, uyum gibi parametreler üzerinden optimizasyon problemi olarak ele

alınan bir özetleme çalışması yapılmıştır (Huang, He, Wei, ve Li, 2010).

Kogilavani ve Balasubramani genetik algoritma temelli çoklu doküman özetleme sağlayan bir yöntem önermişlerdir. Bu yöntem dâhilinde özellikle veri madenciliği alanında kullanılan kümeleme algoritması ile çoklu dokümanlar içerisinde benzer yapıların gruplaması yapılmıştır. Her bir küme içerisinde önem derecesine ve metni temsil kabiliyetine göre puanlama yapılmış ve özeti oluşturacak cümleler belirlenmiştir (Kogilavani ve Balasubramani, 2010).

Lloret ve Palomar özetleme için metin yapısı olarak cümleyi kullanmışlardır. Cümlenin seçimi ile ilgili Code Quality Princible (Kod Kalitesi İlkesi) prensibi doğrultusunda bir cümle ana metne dönük ne kadar fazla bilgi içeriyorsa o derece özet içerisinde yer almalıdır düşüncesini benimsemişlerdir (Lloret ve Palomar, 2012).

Sami ve Diri, Türkçe dilindeki web sayfalarının özetleme işlemi gerçekleştirmiştir. HTML formatında elde edilen verilerin etiketler arasındaki verilere doğru bir şekilde ulaşabilmesi ve okunurluğu artırabilmek amacıyla etkilet bilgilerini içeren bir sözlük üzerinden çalışma yapmış ve öncelikle metin içeriği bu etiket listesi ile karşılaştırmıştır. Elde edilen veri üzerinde konu belirleme, soyutlama ve özet üretme olmak üzere üç adımda çalışma gerçekleştirmiştir. Konu belirleme aşamasında ana metin kısmında konuyu vurgulayan anahtar kelimelerin tespiti ile konu kapsamının belirlenmesi, soyutlama kısmında metni oluşturan cümleler arasındaki ilişki doğrultusunda çıkarım ve kısmi yorumlayıcı tabanlı yöntem ile cümlelerin oluşturulması, özet üretme kısmında ise belirlenen yöntemler doğrultusunda ve cümle ağırlıkları üzerinde özet metninin oluşturulmasını gerçekleştirmiştir. Cümlelerin puanlandırılması aşamasında 12 farklı özellik üzerinden oluşturulan yöntem doğrultusunda sezgisel yöntemler ile bu özellikleri derecelendirmiştir. Sonuç olarak sistemin genel başarısını %59 olarak ölçümlemiştir (Sami ve Diri, 2010).

Shardan ve Kulkarni özetlenecek metin içerisindeki parçanın ana metin içerisindeki yeri, geçme sayısı gibi özellikleri kullanılarak istatistiksel yaklaşımlı çözüm önerileri sunmuştur (Shardan ve Kulkarni, 2010).

Zhao ve Tang genetik algoritma yapısı temelinde sorgu yapısını kullanılmışlardır. Çoklu doküman özetlemesine imkân sağlayan bu çalışmada maksimum düzeyde kapsam sağlamak için metin parçalarının uzunluk ve kapsam değerini kullanılmışlardır (Zhao ve Tang, 2010).

Alguliev ve diğ. tarafından yapılan çalışmada çoklu doküman üzerinde maksimum kapsam genişliği üzerinde literatürde bulunan genetik algoritma ve doğadan ilham alınan Parçacık Sürü Optimizasyonu gibi optimizasyon algoritmaları kullanmışlardır (Alguliev, Aliguliyev ve Isazade, 2013).

Arslan, özetleme işlemini üç adımda gerçekleştirmiştir. İlk adımda varlık adı verilen isim niteliğindeki öğeleri çıkarmış, ikinci adımda bulunan varlıklar arasındaki ilişkinin çıkarımını gerçekleştirmiş, üçüncü adımda ise varlıkların içinde bulunduğu olayları belirlemiştir. Her cümle içerisinde iki sözcük seçmiş, bu sözcüklerin ilişki modelini oluşturmuş, terim sıklığı gibi metrikler ile önemli varlıkları çıkarmıştır. Varlık oluşturma aşamasında insan, kurum, yer, araç, coğrafi konum gibi beş temel sınıf üzerinden çalışma yürütmüştür. Bu varlıklar belirlenen eşik değeri doğrultusunda destek vektör makinesi algoritması ile doğrulanmıştır. Varlık çıkarımında kişi deneyimleri ile oluşturulan özetler ve gazete haber metinlerinden derleme yoluyla oluşturulan veri seti üzerinde çalışma yürütmüştür. Çalışma sonucunu etkileyen eşik değerini deneysel yollarla belirlemiştir (Arslan, 2011).

Boudin ve diğ. graf tabanlı yaklaşımla çoklu doküman üzerinde İngilizce dilinde hazırlanmış içeriklerin Fransızca özetlerinin sunulmasını sağlamıştır. Graf tabanlı ve denetimli bir öğrenme yaklaşımı kullanmıştır. Önerilen model iki aşamalı olup; ilk aşamada cümlelerin derecelendirilmesini gerçekleştirmiş, ardından en yüksek dereceli cümlelerin

seçilmesini sağlamıştır. Her cümlenin diğer dile çevrilmesi ve çevrilen kısmın kalitesinin değerlendirilmesi için ek bir adım oluşturmuştur. Ön işleme aşamasında NLTK kütüphanesindeki (Bird ve Loper, 2004) punkt cümle sınır belirleme yöntemini kullanarak metni cümlelere ayırmıştır. İngilizce içeriği Google Çeviri kullanarak Fransızca diline çevirmiştir. Fransızca cümlelerin hem çeviri doğruluğunu hem de akıcılığını tahmin etmek için her cümle için bir puan hesaplamıştır. Bu puan ile özetleme sürecinde kolayca okunabilen ve anlaşılabilen cümleleri teşvik etmeyi amaçlamıştır. Bunu elde etmek için, her cümle için kaynak cümlenin ne kadar zor olduğu ve üretilen çevirinin ne kadar akıcı olduğu hakkında bilgi sağlayan özellikler hesaplamıştır. Bu özelliklerden bazıları kelimelere göre kaynak dildeki cümle uzunluğu, kaynak ve hedef metin içeriğinin uzunluk oranı, kaynak dildeki cümlelerde bulunan noktalama işaretlerinin sayısı, hedef cümlede bulunan kaynak numaralarının ve noktalama işaretlerinin oranı olarak sıralamıştır. DUC-2004 veri setinden çevrilmiş 16 kümeden oluşan bir alt küme üzerinde yapılan değerlendirmede içerik yapısı bozulmadan oluşturulan özetlerin okunabilirlik özelliğini artırdığını gözlemlemiştir (Boudin, Huet ve Torres-Moreno, 2011).

Çelikyılmaz ve Tür çoklu doküman üzerinde çıkarım tabanlı metin özetleme işlemi için belgeler arasında gizli soyut kavramları ve bu kavramlar arasındaki ilişkiyi modellemek için denetimsiz bir olasılık yaklaşımını sunmuşlardır. İki Katmanlı Konu Modeli adı verilen yöntem için kişi deneyimlerine bağlı olarak oluşturulan özetler üzerindeki değerlendirmelerde tutarlılık, okunabilirlik ve fazlalık giderme açısından yüksek dil kalitesine sahip özetler üretmişlerdir. Model, dönemin son teknoloji denetimli modelleri ile karşılaştırmışlar, DUC-2007 veri seti üzerinde ROUGE metriği (Lin C. , 2004) ile %44,1'lik sonuç elde etmişlerdir (Çelikyılmaz ve Hakkani-Tur, 2010).

Çakır ve Çelebi tüm dillerde kullanılabilecek nitelikte, doküman kümeleme için kullanılmakta olan Cover Coefficient

Based Clustering Methodology (Kapak Katsayısına Dayalı Kümeleme Metodolojisi) kümeleme algoritmasını kullanmışlardır. Kümeleme gerçekleştikten sonra kümeyi temsilen seçilen cümleyi özet kısmına dâhil etmişlerdir. Önerilen yönteme ilişkin performans değerlendirmeleri 10 farklı kişi ile gerçekleştirmişlerdir. Türkçe veri setinin kullanıldığı yöntemde ROUGE-1 ve ROUGE-2 metrikleri üzerinde yürütülen değerlendirmelerde ROUGE-2 için daha iyi sonuçlar elde etmişlerdir (Çakır ve Çelebi, 2011).

Durmaz, terim frekansı ve ters doküman frekansı vektörleri ile çalışma yapmıştır. Bu vektörlerin boyutunu azaltarak sınıflandırma aşamasında maliyeti düşürmek için özellik azaltma yoluna gitmiştir. Bu kapsamda Proportion of Variance (Varyans Oranı) ve Discrete Cosine Transform (Ayrık Kosinüs Dönüşümü) yöntemlerini kullanmıştır. Bu yöntemler boyut azaltmanın yanında sonucun elde edilme süresinde de avantaj sağlamıştır (Durmaz, 2011).

Güran ve diğ. çıkarım tabanlı metin özetleme yöntemi çalışması yürütmüşlerdir. Ön işleme aşamasında durak kelimelerin temizlenmesi, kök ayırma işlemi ve Ardışık Sözcük Algılama adımları gerçekleştirmişlerdir. Ardışık Sözcük Algılama yöntemi yeni bir model olarak sunulmuş olup; belgede yaygın olarak bulunan ardışık kelimelerin tek bir terim olarak temsil edilmesine olanak tanımışlardır. Negatif Olmayan Matris Çarpımı algoritması doğrultusunda her satırda benzersiz bir kelime, her sütunda ise benzersiz bir cümle temsil edilerek terim-cümle matrisini elde etmişlerdir. Bu algoritmada girdi değeri olarak hem yerel hem genel dereceler kullanmışlardır. Logaritma ağırlığı, ağırlıksız durum, Ters Belge Frekansı yöntemini derecelendirme için kullanmışlardır. Çevrimiçi haber sitelerinden ve bazı haber sitelerinden elde edilen 100 dokümanı veri seti olarak kullanılmışlardır. Özet için %33'lük bir indirgeme oranı belirlemişlerdir. Ağırlıkların farklı kombinasyonları ile kullanımında %50-%51 aralığında değer elde etmişlerdir (Güran, Arslan, Kılıç, ve Diri, 2014).

Tatar, Türkçe metin içerisinde iki temel bilginin edinilmesi ile özetleme çalışması geliştirmiştir. Metin içerisinde yer alan varlıkların tespitine yönelik ad tanıma ve metin içerisindeki varlıklar arasındaki ilişkilerin tespitine yönelik ilişki bulma yöntemi olmak üzere iki yöntem üzerinde durmuştur. Bu iki yöntem metinlerde bulunan kalıpları tanımak için kuralları otomatik olarak öğrenmeyi ve kalıplar arasındaki benzerlikleri ve farklılıkları işleyerek genelleştirmeyi amaçlamıştır. Terörizm ile ilgili çevrimiçi ve basılı medya araçları üzerindeki 355 adet haber içeriğinden oluşturulan TurkIE adında bir veri seti üzerinde deneysel çalışmalar yürütmüştür (Tatar, 2011).

Wang ve diğ. çoklu doküman özetleme işleminde öncelikle dokümanların kümelenmesi ve ardından her bir küme için özetleme çalışmasının yürütülmesi üzerinde durmuştur. Mevcut özet yöntemlerinin ağırlıklı olarak cümle terim matrisi üzerinde durmasının tek başına yeterli olmadığını, ek olarak doküman terim matrisinin de kullanımının başarıyı artırdığının gözlemlendiği bir model önermiştir (Wang, Zhu, Li, Chi, ve Gong, 2011).

Galanis ve diğ. çoklu doküman üzerinde çıkarım tabanlı metin özetleme için belirlenen maksimum özet miktarını aşmadan Tamsayı Doğrusal Programlama yöntemini kullanmışlardır. Her bir cümlenin derecelendirmesinde kişi deneyimleri doğrultusunda ortaya konmuş özetler üzerinde Destek Vektör modeli kullanılarak eğitim işlemini yürütmüşlerdir. Destek Vektör modelinde insanlar tarafından oluşturulan özetlere karşı makina tabanlı özetlerin oluşturulmasında çok sık kullanılan ROUGE-2 ve ROUGE-SU4 metriklerinin ortalamasını kullanılmışlardır. Destek Vektör modeline verilecek vektör için cümle konumu, varlık tanımlamaları, Levenshtein uzaklığı (Yujian ve Bo, 2007), kelime çakışması, kelime sıklığı özelliklerini kullanmışlardır. Deneysel çalışmalarda ILP1 adını verdikleri yöntemde özellikle kısa cümlelerin özet için seçildiği, bu cümlelerin ROUGE metriklerinde düşük derece aldığı; bu doğrultuda yeni bir yaklaşımla cümle uzunluğuna

bağlı bir olarak ILP2 yöntemini önermişlerdir. Önerilen modelin dönemin son teknoloji yöntemlerinin başarısına ulaştığını, hesaplama kısmında ise büyük verimlilik sağladığını gözlemlemişlerdir (Galanis, Lampouras, ve Androutsopoulos, 2012).

Kiabod ve diğ. çıkarım tabanlı metin özetleme işlemi kapsamında kelimenin hem yerel hem de genel özelliklerinden faydalanılarak anlamlı kelimelerin belirlenmesi ve bu kelimenin özette yer alması için bir yöntem önermişlerdir. Kelime için yerel özellikten kasıt normalizasyon sonrasında terim sıklık ağırlığı ile sözcüğün bulunduğu cümlelerin normalizasyon sonrasındaki sayılarının ağırlığının çarpımıyla elde edilen tüm değerlerin toplamına dayanan sayısal bir veridir. Yerel özellik kapsamında kelimenin puanı, belirlenen eşik değerinin altında kalırsa özette yer almaması için kaldırmışlardır. Global özellik kapsamında ise metnin başlığındaki kelimeler ile o kelime arasındaki anlamsal benzerlik incelemişlerdir. Algoritma dâhilinde önemli kelimelerin belirlenmesi için yinelemeli bir yöntem sunmuşlardır. Ağırlıkların sayısal hesaplamaları için yapay sinir ağı modelinden yararlanmışlardır. Geri yayılımlı olarak tasarlanan bu modelde gizli katman yapısında üç adet nöron bulunmaktadır. Önerilen yöntem karşılaştırıldığı algoritmalardan daha başarılı sonuçlar üretmiştir (Kiabod, Dehkordi, ve Sharafi, 2012).

Plaza ve diğ. biyomedikal alan için özelleştirilmiş, sözlük tabanlı bir uygulama sonucu elde edilen graf yapısı üzerinde PageRank algoritması kullanılarak oluşturulan bir özetleme mekanizması önerilmişlerdir (Plaza, Stevenson, ve Díaz, 2012).

Alguliev ve diğ. diferansiyel evrim adı verilen yeni bir algoritma yaklaşımıyla optimizasyon odaklı özetleme çalışması gerçekleştirmişlerdir (Alguliev, Aliguliyev, ve Isazade, 2013).

Hariharan ve diğ. LexRank (Erkan ve Radev, 2004) yöntemine iki ek özellik eklemişlerdir. Çoklu doküman özetleme yaklaşımına dayanan bu ek özelliklerden ilki fazlalık

kısımlarının azaltılmasına yönelik bir indirgeme yaklaşımıdır. Bu kısımda özet kısmında yer alması için belirlenen bir cümleden sonra özete eklenecek ikincisi cümle için aynı bilginin olmaması ve özetin daha sade olmasına yönelik bir yaklaşım önermişlerdir. İkinci yenilik ise metin parçalarının bulundukları pozisyona göre ağırlık kazandırılmasına yönelik değişikliği kapsamaktadır. Deneysel çalışmaları iki veri seti üzerinden yürütmüşlerdir. Altı algoritmanın karşılaştırıldığı çalışmada indirgeme özelliğinin bulunduğu yaklaşım eşik değerine sahip LexRank algoritmasından daha iyi sonuç vermiştir (Hariharan, Ramkumar, ve Srinivasan, 2013).

Mikolov ve diğ. metin özetleme konusunda iki yöntem önermişlerdir. Literatürde kullanılan ve word2vec olarak adlandırılan kütüphane desteği bulunan, kelimelerin vektör olarak ifade edilmesi ve sayısal olarak bu vektörler üzerinde işlemler yapılmasını sağlayan bu yöntem kelimelerin temsil yeteneği ve metin içerisinde birbirlerine olan yakınlıklarının tespitinde önemli yer tutmaktadır. Hesaplama maliyetinin düşürüldüğü bu çalışmada 1,6 milyar kelime içeren bir veri setinin vektörlerle temsili ve doğruluk işlemleri bir günden az sürmektedir. Çalışmayla literatürde o zamana kadar sıkça rastlanan terim benzerliği kavramını genişletmişlerdir. Özellikle İngilizce'de big-bigger gibi ek alan kelimeler üzerinde yapılan çalışmada kelimelerin temsil ettiği vektörlerin yakınlıklarını kosinüs uzaklığı ile ölçümlemişlerdir (Mikolov, Chen, Corrado, ve Dean, 2013).

Nallapati ve diğ. SummaRuNNer isminde tekrarlayan sinir ağı üzerinde çalışan çıkarım tabanlı metin özetleme yaklaşımı sunmuşlardır. Modelde soyutlama tabanlı özetler kullanarak uçtan uca eğitime olanak tanıyan yeni bir eğitim mekanizması sunmuşlardır. Tekrarlayan sinir ağının birinci katmanı kelime seviyesinde çalışan ve gizli durum temsillerini, mevcut kelime gömmeleri ve önceki gizli duruma bağlı olarak, her kelime konumunda sıralı olarak hesaplama işlemini gerçekleştirir. Bu yapının yanında son kelimeden ilk kelimeye

doğru giden, kelime düzeyinde başka bir tekrarlayan sinir ağı kullanmışlardır. Model aynı zamanda, cümle düzeyinde çalışan ve çift yönlü kelime düzeyindeki tekrarlayan sinir ağlarının sıralı gizli durumlarını bir parametre olarak kabul eden ikinci bir çift yönlü tekrarlayan sinir ağı katmanından oluşur. İkinci tekrarlayan sinir ağı katmanının gizli durumları, belgedeki cümlelerin temsillerini kodlar. Modelin deneysel çalışmaları için DUC-2002 ve CNN/DailyMail veri setleri ve ROUGE metriğini kullanmışlardır. CNN/DailyMail veri seti üzerinde ROUGE-1 metriğinde %39,6, DUC-2002 veri seti üzerinde ROUGE-1 metriğinde %48,5'lik sonuç elde etmişlerdir (Nallapati, Zhou, Santos, Gulcehre, ve Xiang, 2016).

Güran ve diğ. ana metin içerisinden özeti oluşturan cümlelerin seçimi için kullanılan yöntemleri karşılaştırmışlardır. Cümle konumu, cümle uzunluğu, ilk ve son cümleye benzerlik gibi toplam 15 yöntem üzerinde analiz yapmışlardır. Çalışmayı çeşitli kaynaklardan elde edilen 20 haber metni üzerinde uygulamışlardır. Çalışma kapsamında durak kelimelerin silinmesi ve kök ayırma işlemini gerçekleştirmişlerdir. Ana metin üzerinden %35'lik bir özet miktarı 15'i erkek, 15'i kadın olmak üzere 30 kişilik bir ekip tarafından gerçekleştirilen özetleri karşılaştırmışlardır. Çalışma kapsamında en iyi sonucu kelime frekans yöntemi ile elde etmişlerdir. Ayrıca çalışmayı kadın ve erkek değerlendiriciler üzerinde ayrı ölçümlemişler ve sonucun cinsiyet faktörünü etkilemediği gözlemlemişlerdir. Farklı 15 yöntemi çeşitli denemeler ile gruplara ayırmış ve yapılan denemelerde hibrit şekilde uygulanan yöntemlerin daha başarılı sonuç sağladığını gözlemlemişlerdir (Güran, Arslan, Kılıç, ve Diri, 2014).

Nuzumlalı ve Özgür çoklu doküman üzerinde çıkarım tabanlı özetleme çalışması yürütmüşler, morfolojik analiz ve sabit uzunluktaki sözcük kesiminin etkilerini araştırmışlardır. Farklı konularda 21 küme ve her bir kümede yaklaşık 10 dokümanın bulunduğu bir veri seti oluşturmuşlardır. İki katmanlı morfolojik analiz yöntemi kullanılmışlardır. Terim sıklığı

üzerinden kosinüs benzerliği ile ölçüm yapmışlardır. ROUGE metriği ile yapılan değerlendirmede ROUGE-1 ölçütünün daha iyi sonuç verdiğini gözlemlemişlerdir. Yapılan araştırma sonucunda sabit uzunluktaki sözcük kesimi yaklaşımının köksüz yaklaşımdan daha iyi performans gösterdiği sonucuna varmışlardır (Nuzumlalı ve Özgür, 2014).

Pennington ve diğ. Mikolov ve diğ. tarafından önerilen Skip-Gram ve CBOW modellerindeki (Mikolov vd. 2013) anlamsal bilgileri yakalarken birlikte kullanım istatistiklerini elde edememe zafiyeti konusunda çalışmışlardır. Önerdikleri yöntemde olasılık yöntemlerinden yararlanılarak bu zafiyetin gidermesini amaçlamışlar ve literatüre GloVe modelini kazandırmışlardır (Pennington, Socher, ve Manning, 2014).

Tutkan ve diğ. anlamsal özellik seçimi yöntemi doğrultusunda Gestalt teorisini kullanılmışlardır. Çalışma kapsamında eğitimsiz bir model önermişlerdir. Yöntem doğrultusunda özellik seçimi için fazla sayıdaki nitelik içerisinden hangisinin önem derecesinin yüksek olduğunu belirlemişler ve bu kapsamda özellik sayısını azaltarak sınıflandırma için harcanan zamanı azaltmışlardır. İki farklı dilde veri seti kullanıldıkları çalışmada İngilizce için beş farklı sınıf üzerinde 1150 haber, Türkçe için yedi sınıflı 927 metin kullanmışlardır. Türkçe veri setinde kelime sayısı arttıkça sınıflandırma doğruluğunun %95'e kadar ulaştığını gözlemlemişlerdir (Tutkan, Ganiz, ve Akyokuş, 2014).

Attokurov çoklu doküman üzerinde özetleme çalışması yürütmüştür. Soyutlama tabanlı özetleme çalışmalarındaki bilgi tekrarı probleminin giderilmesine yönelik olarak Optimal Tree Pruning (Optimal Ağaç Budama) Algoritması ve Hierarchical Agglomerative Clustering (Hiyerarşik Yığılma Kümeleme) Algoritmasının birlikte kullanımını incelemiştir. Hiyerarşik Yığılma Kümeleme Algoritması'nı özet içerisinde tekrarlanan metin parçalarının ayıklama işleminde kullanırken, Optimal Ağaç Budaması Algoritması'nı ise bu metin parçalarının özet bölümünde azaltılması için kullanmıştır. Vektörler arasındaki

benzerlik ölçütü olarak kosinüs benzerlik katsayısını kullanmıştır. Optimal Ağaç Budama algoritmasında benzer cümlelerin aynı ağaç içerisinde olması sebebiyle tekrarlama olasılığı bu kısımda azaltılmıştır. Benzerlik katsayısına göre bozulum değerine göre budama işlemini tekrar yapılandırmış ve son aşamada özeti oluşturmuştur. Çalışmayı ROUGE metriği ve DUC-2002 veri seti ile denemiştir. Soyutlama tabanlı 200 ve 400 kelimelik özet çalışmalarının bulunduğu sınıfı kullanarak yaptığı değerlendirmede iki aşamalı bir işlem gerçekleştirmiştir (Attokurov, 2014).

Babar ve Patil çıkarım tabanlı metin özetleme işleminde bulanık mantığın kullanımının özet kalitesinin artırdığı düşünmüşlerdir. Önerilen algoritma dâhilinde gizli anlamsal analiz ile ana metin içerisindeki kavramlar arasındaki anlamsal ilişkileri ortaya çıkarmak için kullanılan bulanık mantık sistemine dâhil etmişlerdir. Algoritmayı çevrimiçi ortamda özetleme yapan iki sistem ile karşılaştırdıklarında daha iyi sonuçlar almışlardır. Ana metin içerisindeki her bir cümleyi sekiz özellik üzerinden derecelendirmişlerdir. Bu özellikleri başlık benzerliği, cümle uzunluğu, terim ağırlığı, cümle konumu, cümleler arası benzerlik, özel isim, tematik kelime ve sayısal veri olarak sıralamışlardır. Belirtilen özellikler doğrultusunda sonucu bulanık mantık sistemine, sonrasında çıkarım alanına, son olarak da bulanık çözücü alana aktarmışlardır. Bu kısım ile birlikte her bir cümle bir puan elde etmiş ve cümleleri elde edilen puan doğrultusunda özet kısmına almışlardır. On farklı veri seti üzerinde yapılan deneysel çalışmada ortalama kesinlik değeri %90'lara ulaşmıştır (Babar ve Patil 2015).

Cao ve diğ. çoklu doküman üzerinde çıkarım tabanlı özetleme işleminde ana metin içerisinden seçilecek cümlenin genel metin bağlamından bağımsız değerlendirilmesine yönelik evrişimli sinir ağı modeline dayanan PriorSum isminde bir model önermişlerdir. Önceki özet deneyimlerine bağlı olarak ve ana metin içerisindeki özelliklerin birleşimiyle meydana gelen yöntemde DUC-2001, DUC-2002 ve DUC-2004

veri setleri ile deneysel çalışmalar yürütmüşlerdir. Yapılan değerlendirmelerde ROUGE-1 ve ROUGE-2 metriklerini kullanmışlardır. ROUGE-1 metriği üzerinde %35-%40 aralığında, ROUGE-2 metriği üzerinde %7-%10 aralığında değer elde edilerek karşılaştırıldığı dönemin son teknoloji özetleme tekniklerine göre üstünlük sağlamıştır (Cao, ve diğerleri, 2015).

Birant, Türkçe dili üzerinde kural tabanlı bir özetleme yazılımı gerçekleştirmiştir. Çalışmada yapılan yazılım içerisinde literatürde bulunan doğal dil işleme araçları performanslarını iyileştirerek kullanmıştır. Çalışmada Aktaş ve Çebi'nin Rule-Based Sentence Detection Method for Turkish (Türkçe için Kural Tabanlı Cümle Tespit Yöntemi) (Aktaş ve Çebi, 2013) çalışmasını kullanmıştır. Yapılan yazılım içerisinde kullanıcının anahtar kelime girmesine olanak tanınarak önemli terimlerin ve özet kısmında yer alması kaçınılmaz noktaların belirlenmesi sağlamıştır. Anahtar kelime ve ana metin girişi sonrasında ilk aşamada metni cümlelere ayırmış ve XML yapısında göstermiştir. Ardından cümle bazlı puanlama işlemini paragraf bazında ve tüm metin bazında iki farklı şekilde yapmıştır. Elde edilen puanlarda cümleler yüksek puandan düşük puana doğru sıralamış ve özette yer alacak cümleler belirlemiştir. Değerlendirme aşamasında ROUGE metriği ve kişi deneyimleri kullanılmıştır. Üç bilimsel makale, iki haber metni olmak üzere toplam beş metin kullanmıştır. Kişi deneyimi kullanımının maliyet ve zaman problemleri oluşturması açısından kısıtlı düzeyde bir veri seti kullanmıştır. Haber metinlerinin kısa olması sebebiyle daha iyi sonuçlar verdiğini gözlemlemiştir. Bilimsel makalelerde matematiksel formül ve karmaşık düzeyde cümlelerin yer alması performans olarak dezavantajlı bir durum oluşturmuştur. Kişi deneyimleri üzerinden 10 öğrencinin ilgili metinler üzerinden yaptıkları özet çalışması ile yazılımın ürettiği özet metinlerini karşılaştırmıştır. Sonuç olarak otomatik üretilen özetler ile kişi deneyimlerinin çok yakın sonuçlar ürettiğini gözlemlemiştir (Birant, 2015).

Hatipoğlu ve Omurca metin özetleme için cümlelerin istatistiksel yöntemler ile derecelendirilmesine dayanan model ve sezgisel yöntemlerle anlam üzerinde çalışan modelin birleştirilmesi sonucu melez bir yöntem önermişlerdir. Cümlenin uzunluğu, konumu gibi sayısal değerler üzerinden ve anlamsal niteliği ortaya koyacak LSA temelli çalışmayı birlikte yürütmüşlerdir. Uygulama sonucunda en yüksek puan alan cümleden başlayarak istenilen oranda özete yansıyacak metin parçalarını belirlemişlerdir. Çalışmada Türkçe Vikipedi üzerindeki metinlerden oluşan veri setini kullanmışlardır. Model performansını ana metnin 10 farklı kişi üzerinden özetlenmesi ile karşılaştırılmışlardır. İki farklı metin üzerinde yapılan karşılaştırma sonucu güneş sistemine ilişkin bir metin üzerinde yapılan çalışmada cümle derecelendirme yöntemi ile kişilerin özetleri arasında %77,5 oranında, Charles Bukowski metni üzerinde ise %82 oranında örtüşme tespit etmişlerdir (Hatipoğlu ve Omurca, 2015).

Meena ve Gopalani çıkarım tabanlı metin özetleme işlemi için evrimsel algoritmalarda bazı özelliklerin kullanılmasını önermişlerdir. Bu özelliklerin bazıları geçmişten beri süregelen kullanım alanlarına sahipken, çalışmada hibrit bir model önermişlerdir. Özellik kümesinde terim sıklığı, cümle konumu, işaret ifadesi (sonuç olarak, özetle vb.), başlık benzerliği, özel isim, eş anlamlı kelimeler, cümle benzerliği, cümle içi sayısal ifadeler, yazı tipi, sözcük benzerliği, TextRank (Mihalcea ve Tarau, 2004) ile graf yapısında temsil edilen cümlelerin derecelendirilmesi, cümle uzunluğu, olumlu anahtar kelime, olumsuz anahtar kelime, genel benzerlik, cümleler arası kelime benzerliği, paragraflar arası kelime benzerliği, yinelemeli sorgu tabanlı derecelendirme, tematik özellikler, varlık bilgisi gibi özelliklerin dâhil edilmesini önermişlerdir. DUC-2002 veri seti içerisindeki on doküman üzerinde deneysel çalışmalar yürütülmüşlerdir. Öncelikle tüm özellikleri eşit önem derecesiyle uygulamışlar, ardından genetik algoritma

yapısını dâhil etmişlerdir. 100 yineleme sonrasında süreci durdurmuş, sonuçları gözden geçirilmiş ve ROUGE metriği ile değerlendirmişlerdir. Veri seti içerisindeki ikinci belge dışında tüm belgelerde yüksek başarı elde etmişlerdir. Bazı özellikler (özel isim, cümle konumu gibi) çok daha fazla bilgilendirici unsur içerdiğinden ağırlıkları diğerlerine oranla artmıştır (Meena ve Gopalani, 2015).

Turan, İngilizce dili için çıkarım tabanlı bir özetleme yöntemi üzerinde durmuştur. Dokümana dair bir vektör kümesi temsil edildikten sonra kelimelerin temsil sıklığı üzerinde fazla oran elde edilenleri vektör kümesine dâhil etmiştir. Aykırı olabilecek vektörleri uzaklık hesaplaması sonucu devre dışı bırakmıştır. Son aşama olarak paragraf terim vektörlerinin ilk aşamada temsil edilen doküman üzerindeki vektör kümesi ile benzerliğini hesaplamıştır. Bu benzerlik kısmını Eşleşme Yüzdesi adını verdiği özgün bir yöntem ile oluşturmuştur. Hedeflenen özet oranına ulaşana dek en yüksek değerdeki cümleleri özet bölümüne eklemiştir. Veri seti olarak DUC-2006 kullanılmış ve performans ölçütü için ROUGE metriğini kullanmış ve %75 sonucunu elde etmiştir (Turan, 2015).

Mirshojaei ve Masoomi, Guguk Kuşu Algoritması'nı kullanılmışlar, optimizasyon temelli bu yaklaşımda özetleme çalışması için performans artırıcı bir yöntem önermişlerdir (Mirshojaei ve Masoomi, 2015).

Rautray ve diğ. özetlenecek metin içerisindeki parçanın ana metin içerisindeki yeri, geçme sayısı gibi özellikleri kullanılarak istatistiksel yaklaşımlı çözüm önerileri önermişlerdir (Rautray, Balabantaray, ve Bhardwaj, 2015).

Rush ve diğ. soyutlama tabanlı yeni bir model önerilmişlerdir. Bu model tamamıyla veri odaklı, girdi cümlesine bağlı olarak özette yer alacak her bir kelimenin üretimini sağlayan dikkat temelli bir modeldir. Modelde uçtan uca kolayca eğitim sağlanabilir iken, aynı zamanda büyük miktardaki veri ile işlem yapılabilmektedir. Önerilen model oluşturulan özetin kelime dağarcığı hakkında herhangi bir varsayımda bulunmadığından,

doğrudan herhangi bir belge-özet çifti üzerinde eğitilebilir. Oluşturulan yapay sinir ağı hem olasılık tabanlı bir dil modeli hem de koşullu özetleme modeli olarak hareket eden bir kodlayıcı içermektedir. Bag-of-words (kelime çantası) tekniğinde yer alan ana metindeki sözcüklerin sıralama ve birbirleri ile ilişkinin göz ardı edilmesinin oluşturduğu sorunun önüne geçerek ilgili teknik üzerinde değişiklikler yapmışlardır. DUC-2003 veri setinin kullanıldığı çalışma ROUGE metriği ile değerlendirdiklerinde önemli performans kazanımları elde edildiği gözlemlemişlerdir (Rush, Chopra, ve Weston, 2015).

Saleh ve diğ. özgün bir optimizasyon modeli sunmuşlardır. Baskın Olmayan Sıralı Genetik Algoritma adını verdikleri bu modelde çıkarım tabanlı metin özetleme yöntemi kullanmışlardır (Saleh, Kadhim, ve Attea, 2015).

Altıntop sağlık birimleri tarafından oluşturulan yönetimsel ve ekonomik içerikli verilerin özetlenmesi amaçlamıştır. Çalışma genetik algoritma tabanlı bir optimizasyon yöntemine dayanmaktadır. Metne dair öznitelikleri bulanık mantık üzerinden dinamik olarak işlemiştir. Yapılan deneysel çalışmalar sonucunda genetik algoritma için en uygun parametre değerleri tespit etmiştir. Elde edilen sonuç değerlendirildiğinde %75 oranında bir başarı elde etmiştir (Altıntop, 2015).

Umam ve diğ. metin özetleme için yöntem olarak metin içerisindeki bilgiye dair kapsam, varyans gibi parametrelerin kullanılarak çoklu doküman özetlemeye ilişkin bir optimizasyon algoritması önermişlerdir (Umam, Putro, Pratamasunu, Arifin, ve Purwitasari, 2015).

Barrios ve diğ. graf tabanlı TextRank yöntemi üzerinde yapılan değişiklikleri açıklamışlardır. Önerilen değişiklik graf ile temsil edilen cümleler arasındaki benzerlik oranının ölçümüne yöneliktir. Bu sayede graf ile temsil edilen yapının kenar ağırlıkları da farklı bir şekilde hesaplanacaktır. Bu değişiklikler mevcut TextRank algoritmasının performansında iyileşme sağlamıştır. Terim frekansı sonrasında sözcüklerin temsil edildiği vektörler arasındaki benzerliğin kosinüs

uzaklığı yöntemi ile bulunmasını önermişlerdir. Puanlama niteliğine dayalı BM25 fonksiyonunda bir terimin belgenin yarısından fazlasında görülme durumunda negatif bir değere sahip olmasından kaynaklanan sorunun önüne geçmek için formül üzerinde düzenleme yapmışlardır. DUC-2002 veri seti üzerinde yapılan çalışma ROUGE metriği üzerinden değerlendirmişlerdir. Farklı yöntemlerin denediği çalışmada BM25 ve BM25+ teknikleri mevcut TextRank algoritmasından %2,92 oranında, Kosinüs uzaklığı %2,54 oranında daha başarılı sonuç vermiştir (Barrios, Lopez, Argerich, ve Wachenchauzer, 2015).

Cheng ve Lapata yapay sinir ağı tabanlı ve veri odaklı çıkarım tabanlı bir özetleme modeli önermişlerdir. Çalışmada iki veri seti kullanılmış, birinde kelime bazlı çıkarım, diğerinde cümle bazlı çıkarım işlemini yürütmüşlerdir. Modelin kod çözücü kısmındaki önemli yenilik tüm kelime dağarcığı yerine belge içerisindeki semboller üzerinden hareket etmesidir. Cümle uzunluğu, cümlede geçen varlık tanımlaması, cümlenin doküman ile ilişkisi, cümle-cümle benzerliği gibi çeşitli metrikler kullanılmışlardır. Kullanılan veri setlerinde %90 oranında eğitim, %5 oranında doğrulama, %5 oranında test verisi ayırmışlardır. ROUGE metriğinin yanı sıra 20 katılımcı ile kişi deneyimleri üzerinden de değerlendirme yapmışlardır. DUC-2002 ve CNN/DailyMail olmak üzere iki veri seti üzerinde yapılan çalışmalar rekabetçi bir sonuç ortaya çıkarmıştır. Cümle çıkarım yönteminin kelime çıkarım yöntemine göre daha başarılı olduğunu gözlemlemişlerdir (Cheng ve Lapata, 2016).

Nallapati, Tekrarlayan Sinir Ağı yapısı üzerinde soyutlama tabanlı özetleme çalışması gerçekleştirmiştir. Anahtar kelime modelleme, cümle-kelime hiyerarşisinin oluşturulması, eğitim sırasında nadir kelimelerin yayılmasının engellenmesi gibi kritik sorunları ele alan birkaç yeni model önermiştir. Veri seti olarak tek cümle hedefi olarak Gigaword (Pennington, Socher, ve Manning, 2014) ve DUC-2003 ile DUC-2004 kullanmıştır.

Çoklu cümle özeti için CNN/DailyMail veri setini kullanmıştır. İki farklı metin yapısı üzerinde dönemin son teknoloji teknikleri ile yarışabilecek nitelikte model olduğunu gözlemlemiştir (Nallapati, Zhou, Santos, Gulcehre, ve Xiang, 2016).

Oliveira ve diğ. çıkarım tabanlı özetleme modellerinde ana metindeki cümlenin önem derecesini belirlemek için kullanılan 18 tekniğin karşılaştırmasını yapmışlardır. Yapılan karşılaştırmada hem tekli hem de çoklu dokümanlarda özetleme yaklaşımı için kullanılan teknikler yer almıştır. Teknikler için hem ayrı, hem de çeşitli kombinasyonları içeren deneyler gerçekleştirilmişlerdir. Tek doküman özetleme modelleri için DUC-2001, DUC-2002 ve CNN/DailyMail veri setleri, çoklu doküman özetleme modelleri için DUC-2001 ve DUC-2004 veri setlerini kullanmışlardır. Yapılan değerlendirme bireysel olarak terim sıklık modelinin diğer modellere göre daha başarılı sonuç verdiği görülmüş; hibrit model denebilecek farklı modellerin bir arada kullanılmasının, geleneksel modellerin tek başlarına kullanılmasından daha iyi performans gösterdiğini kanıtlamışlardır (Oliveira, ve diğerleri, 2016).

Özateş ve diğ. çoklu doküman üzerinde çıkarım tabanlı özetleme işlemi için cümle benzerliği üzerinden bir yöntem önermişlerdir. Dilbilgisi bağımlılık temsilinden yararlanılmışlardır. Önerilen yeni yöntemi LexRank içerisindeki kelime tabanlı cümle benzerliği ile karşılaştırılmışlardır. Dependency Tree Kernel (Culotta ve Sorensen, 2004) ve Dependency Trigram Kernel (Choi, Kim, ve Croft, 2012) isimli iki bağımlılık ağacın ilişki çıkarımı için yüksek performans sergilemesine karşın ROUGE-1 ve ROUGE-2 metrikleri üzerinde kelime tabanlı kosinüs benzerliğinden daha yüksek bir başarı elde edememiştir. Önerilen cümle benzerlik yöntemi iki bağımlılık ağacından, LexRank orijinal benzerlik yönteminden daha iyi performans göstermiştir (Özateş, Radev, ve Özgür, 2016).

Fang ve diğ. cümle puanlama tekniğini ele almışlardır. Kelime-cümle ilişkisini graf tabanlı denetimsiz sıralama

modeli ile birleştiren CoRank adında yeni bir sıralama modeli önermişlerdir. Modelde cümle puanlama modelini geliştirmek için kelime ile cümle arasındaki ilişki cümlenin temsil edildiği graf modeline eklemişlerdir. Karşılıklı geri bildirim esasına dayanan modelde yüksek önem derecesine sahip bir kelimenin bulunduğu cümle de aynı derecede öneme sahip olabilmekte, yüksek önem derecesine sahip bir cümledeki kelimeler de aynı şekilde önemli kelimeler bölümüne alınmaktadır. TextRank modeline göre her kelimenin aynı ağırlığa sahip olmadığı, farklı kelimelerin ağırlıklarının değişken olduğunu vurgulamışlardır. Özetleme işleminde gereksiz metin parçalarının tespiti için fazlalık giderme tekniği kullanılmışlardır. Bu teknikte eşik değeri %80 olarak belirlenmiş ve özette yer alacak cümlelerin metrik değerlendirmesi sonucu bu eşik değerini aşmasını beklemişlerdir. Eklenen bu kısımla model CoRank+ ismini almıştır. İki farklı isimle anılması modelin yapısının modüler olması ve istenildiğinde birbirinden bağımsız şekilde çalışabilmesine olanak sağlamasından gelmektedir. 10 Çince doküman içeren bir veri seti ile 567 İngilizce doküman içeren bir veri seti üzerinde çalışmayı gerçekleştirmişlerdir (Fang, Mu, Deng, ve Wu, 2017).

Hark ve diğ. metin özetleme işlemlerinin iki yönteminden biri olan çıkarım tabanlı yaklaşım için temel teşkil edebilecek farklı dokümanlar arasındaki yapısal benzerliğin tespitine yönelik R dili üzerinde bir yöntem önermişlerdir. Jaccard benzerliği kullanılan yaklaşım benzer çalışmalara oranla iyi sonuç vermiştir (Hark, Seyyarer, Uçkan, Myo, ve Karci, 2017).

Hu ve diğ. tatil ve gezi için çevrimiçi ortamlarda yapılan eleştiri ve görüşlerin özet halinde sunulması kişilerin seçim yapmalarını kolaylaştıracak bir çalışma yürütmüşlerdir. Çalışma kapsamında metin içerisinde en önemli cümlelerin tespit edilmesi ve kullanıcıya sunulmasını sağlayan yeni bir model önerilmişlerdir. Önceki çalışmalarda yazarın güvenirliği, çelişkili ifadeler gibi bilginin önem derecesini etkileyecek faktörlerin göz ardı edildiği belirtilmiş ve bu doğrultuda yeni bir önem metriği oluşturmuşlardır. Çalışmada tripadvisor.

com web sitesinde bulunan iki otel için yapılan yorumları kullanılmışlardır. Üç farklı yaklaşım yürütülmüşlerdir. İlk yaklaşımda tüm yorumlar tek bir doküman içerisinde gibi düşünerek en önemli cümleleri puanlama ile elde etmişlerdir. İkinci yaklaşım doğrultusunda elde edilen veriler için k-medoids kümeleme algoritmasını kullanmışlardır. Kümeler üzerinde ağırlık merkezine minimum mesafede bulunan cümleyi temsili cümle olarak seçmişlerdir. Üçüncü ve farklı olan yaklaşımda ise cümleler için önem derecesini belirleyen terim frekansı, cümle uzunluğu vb. bilgilere ek olarak yazar güvenirliği, tutarlı ifadeleri eklemişlerdir. Önerilen özetin değerlendirmesini 20 farklı kişiyle iki geleneksel yöntemin karşılaştırılması şeklinde gerçekleştirmişlerdir. Sonucun diğer iki yönteme göre daha iyi performans gösterdiği gözlemlenmişlerdir (Hu, Chen, ve Chou, 2017).

Kaynar ve diğ. graf tabanlı metin özetleme yöntemlerini karşılaştırılmışlardır. LexRank yöntemi ve farklılaştırılmış beş TextRank algoritmasını kullanmışlardır. TextRank algoritması içerisinde Longest Common Subsequence (En Uzun Ortak Sonuç) benzerlik ölçümü graf üzerindeki düğümleri arasında benzerlik için kullanmışlardır. Yapılan çalışmayı iki farklı dil üzerindeki veri seti ile ölçümlemişlerdir. İngilizce için DUC-2002 veri seti, Türkçe için farklı kaynaklardan elde edilmiş 120 haber metnini kullanmışlardır. En iyi başarıyı Longest Common Subsequence (En Uzun Ortak Sonuç) ile yapılan benzerlik ölçütü ile elde etmişlerdir. İngilizce veri seti için ROUGE metrikleri ile değerlendirmişler, İngilizce için %51 ROUGE-1, %26 ROUGE-2 değerleri, Türkçe veri seti için ise %74 ROUGE-1, %67 ROUGE-2 değerlerini elde etmişlerdir (Kaynar, Görmez, Işık, ve Demirkoparan, 2017).

Rautray ve Balabantaray metin özetleme çalışmasını çıkarım tabanlı yaklaşım ile bir optimizasyon problemi olarak ele almışlardır. Buradaki çalışma çoklu doküman özetlemeye dayanan bir algoritma optimizasyon önerisini kapsamaktadır (Rautray ve Balabantaray, 2017).

Kısayol metin özetleme işlemi için ana metinde geçen yapıların paragraf bazında önem derecesine göre seçilmesi ve özet metnine aktarılmasını gerçekleştirmiştir. İngilizce dilinde yazılan makale ve haberlerde ana metin içerisindeki kelimelerin kök yapısına göre metin içerisinde geçme frekanslarını değerlendirmiş, yüksek frekanslı kelimelerin bulunduğu paragrafların temsil kısmında yer almasını sağlamıştır. Metin üzerinde Ön işleme çalışmaları kapsamında HTML ayrıştırıcı kütüphane kullanmıştır. Çalışma içerisindeki iki farklı yöntem ile karşılaştırma yapmıştır. İlk yöntemde frekansı en yüksek 10 kelimenin bulunduğu paragrafı seçmiş, ikinci yöntemde ise oransal olarak talep edilen özet paragraf bazında kümeleme algoritması ile ayrıştırılmış ve seçilecek kümeyi Jaccard benzerliği ile tespit etmiştir. Çalışmada ikinci yöntemin daha iyi sonuç verdiğini gözlemlemiştir (Kısayol, 2018).

Baydar çıkarım tabanlı metin özetleme yaklaşımı için üç farklı yöntem üzerinde değerlendirme yapmıştır. Sezgisel yöntemler ile elde edilen özetleme başarısını genetik algoritma ile destekleyerek başarımı artırmayı hedeflemiştir. Cümlelerin puanlama işlemini genetik algoritma ile gerçekleştirmiştir. Sabit puanlı değerlendirme yönteminde ana metin içerisindeki cümlelere sezgisel yöntemler ile puanlama yaparken, sezgisel rastgele puan aralığı yönteminde ilk yönteme ek olarak puan aralığı eklenerek çalışmayı sürdürmüştür. Genel aralıklı rastgele puanlama yönteminde ise puan aralığını 1 ile 100 arasında belirlemiştir. Özet için ana metin üzerinden %35 oranını belirlemiştir. Çalışmayı f-skoru, ROUGE metriği ve kosinüs benzerliği gibi yöntemler ile değerlendirmiştir. İlk yöntem ile %42,2, ikinci yöntem ile %51,1, üçüncü yöntem ile %59,2 sonucunu elde etmiştir (Baydar, 2018).

Aysu, Türkçe haber metinleri üzerinde özetleme çalışması yürütmüştür. Cümle bazlı puanlama işlemi gerçekleştirmiş, 20 farklı kişiden oluşan grup tarafından haber metinleri içerisinde en önemli üç cümlenin belirlenmesi istemiş ve özetleme çalışması sonrasında bu grup ile karşılaştırma

yaparak değerlendirmiştir. Değerlendirme sonucunda özetlerin benzerlik oranını %36,5 olarak tespit etmiştir (Aysu, 2018).

Kaynar ve diğ. mobil platform üzerinde bir özetleme çalışması gerçekleştirmişlerdir. Android tabanlı yapılan çalışma kapsamında çalışacak algoritmayı kullanıcıya seçenek olarak sunmuşlardır. TextRank , LextRank , öznitelik tabanlı ve kümeleme olmak üzere dört farklı algoritma sunmuşlardır. ROUGE metriği üzerinden yapılan değerlendirmede sunulan dört farklı algoritmadan en iyi sonucu %66 ile öznitelik tabanlı algoritma, en kötü sonucu ise %56 ile kümeleme algoritması ile elde etmişlerdir. Dört farklı algoritma için hesaplama sürelerini karşılaştırdıklarında 25 saniye ile kümeleme algoritmasının, diğer üç algoritmanın 15 saniyede tamamlandığını gözlemlemişlerdir (Kaynar, Işık, Görmez, ve Kuş, 2018).

Koupaee ve Wang özetleme çalışmalarında veri seti olarak sıklıkla haber metinlerinin kullanılmasının aynı yazım şeklinin getirdiği dezavantajın kısıtlı bir geliştirme ortamına neden olması durumundan yola çıkmışlardır. Bu kapsamda 230.000'den fazla makalenin kişiler tarafından çıkarılan özetlerinin sunulduğu bir veri seti geliştirmişlerdir. Mevcut veri setlerinden farkı birçok kategorideki konuları kapsaması, farklı yazım ve özet şekillerini barındırmasıdır. Soyutlama tabanlı özetleme çalışmaları için kaynak oluşturabilecek veri setinin sunmalarının yanında, hem çıkarım tabanlı hem de soyutlama tabanlı özetleme modelleri ile performans değerlendirmesi yapmışlardır. Performans değerlendirmesi sırasında CNN/DailyMail veri setini kullanmışlardır. ROUGE metriği ile yapılan değerlendirmede sık kullanılan yöntemler ile yaklaşık 10 puanlık bir fark bulunduğunu, bu durumun farklı kişilerin farklı özet stillerinin bulunmasından kaynaklandığı gözlemlemişlerdir (Koupaee ve Wang, 2018).

Kumbhar ve diğ. metin özetleme çalışmalarında cümlelere puanlama verilmesiyle önem derecesinin belirlendiği bir metrik kullanan model geliştirmiş ve bu modeli makina öğrenimi ve derin öğrenme modellerine uygulamışlardır. Bu kapsamda en

uygun model seçiminin sağlanması ve bu modelin niçin daha yüksek performans sağladığını da analiz etmişlerdir. Ön işleme aşamasında kelime bazlı ayrım yapmış ve durak kelimeleri temizlemişlerdir. Sonrasında özellik matrisini oluşturmuş ve belirledikleri altı özellik üzerinde değerlendirmişlerdir. Bu özellikleri kelime frekansı, önemli kelimeler, cümle uzunluğu, cümle yerleşimi, gereksiz kelime, özel isimler olarak sıralamışlardır. Özellik matrisi üzerinde yapılan ölçümlerde en yüksek değeri Yapay Sinir Ağı, ikinci en yüksek değeri Rastgele Orman Algoritması elde etmiştir. Metrik değerlendirmesinde ise kelime frekansı özelliğinin en yüksek derecede öneme sahip olduğu gözlemlemişlerdir (Kumbhar, Bordia, Kapse, ve Paatil, 2018).

Lin ve diğ. soyutlama tabanlı özetleme çalışmalarındaki tekrarlama ve anlamsal ilgisizlik sorunu üzerinde durmuşlardır. Oluşturdukları model seq2seq yapısına dayanmaktadır. Yönsüz bir Long Short-Term Memory (Uzun Kısa Süreli Bellek) çözücü yapısını kelime okuma ve özet için sözcükleri oluşturma işleminde kullanmışlardır. Çalışma Gigaword (Graff ve Cieri, 2003) veri seti üzerinde gerçekleştirmiş ve ROUGE metriği ile yaptıkları değerlendirmede ROUGE-2 için %26,8 ve %17,8 ile dönemin sık kullanılan yöntemlerinden daha iyi performans gösterdiğini gözlemlemişlerdir (Lin, Sun, Ma, ve Su, 2018).

Narayan ve diğ. çıkarım tabanlı özetleme için cümle puan ile derecelendirme işlemi gerçekleştirmişler ve pekiştirmeli öğrenme tabanlı yeni bir eğitim algoritması önermişlerdir. CNN/DailyMail veri setleri üzerinde çalışma gerçekleştirmişlerdir. Uzun parametre miktarlarında eğitim sırasında ortaya çıkan gradyan sorununu gidermek için Long Short-Term Memory (Uzun Kısa Süreli Bellek) hücrelerine sahip tekrarlayan bir sinir ağı modeli kullanmışlardır. Pekiştirmeli öğrenmenin katkısı doğrudan değerlendirme ölçütünün optimize edilebilmesine olanak sağlaması ve cümleler arasında ayrım yapma kısmında daha iyi bir ortam oluşturmasıdır. REFRESH ismini verdikleri modelde %35-%40 arasında sonuç elde etmişlerdir (Narayan, Cohen, ve Lapata, 2018).

Sanchez-Gomez ve diğ. çoklu doküman üzerinde özetleme çalışması kapsamında çıkarım tabanlı yöntem kullanmışlardır. Yöntem dâhilinde doğadan esinlenilen yapay arı kolonisi algoritmasını kullanmışlardır (Sanchez-Gomez, Vega-Rodríguez, ve Pérez, 2018).

Sinha ve diğ. geleneksel yöntemlerdeki özellik bağlamlı özetleme yerine tamamen veriye dayalı ve ileri beslemeli yapay sinir ağı üzerinde çalışma gerçekleştirmişlerdir. Çalışmayı tek doküman üzerinde ve DUC-2002 veri seti kullanılarak yürütmüşlerdir. Yaptıkları değerlendirmelerde dönemin son teknoloji yöntemleriyle karşılaştırılabilecek nitelikte sonuçlar ortaya çıkmıştır. Farklı dokümanlarda farklı boyutlarda cümlelerin olma durumuna karşın ölçeklenebilir bir model üzerinden ana metni sabit boyutlu parçalara bölmüşlerdir. Bu kapsamda her bir bölüm üzerinde aynı miktarda cümle sayısı bulunmaktadır. ROUGE metriği üzerinde yaptıkları değerlendirmede %55'lik bir sonuç elde etmişler ve dört model ile karşılaştırdıklarında en yüksek değeri elde etmişlerdir (Sinha, Yadav, ve Gahlot, 2018).

Torun ve Inner, farklı kaynaklar üzerinden elde ettikleri 12.000 haber üzerinden özetleme çalışması yapmış ve elde ettikleri özetler üzerinden haber benzerliklerini karşılaştırmışlardır. Değerlendirmede 12.000 haber içerisinde 1.500'ünün benzer olduğunu gözlemlemişlerdir. Özetleme çalışmasında ana metin içerisindeki cümleler içerdikleri kelimelerin frekansları doğrultusunda derecelendirilmiş, cümlelerin aldıkları bu değer doğrultusunda özet kısmında yer alıp almayacaklarını belirlemişlerdir. Özetleme çalışması için tüm metin içerisinden beş cümle seçmişlerdir (Torun ve İnner, 2018).

Tozyılmaz, metin özetleme için soyutlama tabanlı bir yöntem önermiştir. Yapay sinir ağı temelli çalışma kapsamında cümle gömme uygulamasını kullanmış ve WikiHow (Koupaee ve Wang, 2018) veri seti üzerinde literatürde özetleme yöntemlerinin sonuçlarına yakın bir değer elde

etmiştir. Çalışmayı ROUGE metriği ile önce kelime gömmesi üzerinde denemiş ve seq2seq modeline göre daha düşük bir değer elde etmiştir. Başarıyı artıran faktör olarak yöntemin cümle gömmesi olarak değiştirmesiyle büyük bir parametre verilmesi durumunda sistemin daha performanslı çalışacağını düşünmüştür. Greedy search (açgözlü arama) algoritmasına alternatif olarak beam search (ışın arama) algoritmasını kullanmıştır. Cümlelerin vektörlere dönüştürülmesi sonrasında en yaygın 25000 kelimenin seçilmesi ve ID atanma işlemini gerçekleştirmiştir. Yapı olarak başlangıç, bitiş ve belirsiz olmak üzere üç yer tutucu kullanmıştır. Belirsiz kısımda en yaygın kelimelerin bulunmadığı cümleleri kategorize etmiştir (Tozyılmaz, 2019).

Özkan, geleneksel yöntem olarak kelimelerin kök ve ek ayıklaması gibi Ön işleme süreçleri yerine kelimelerin harf bazında karşılaştırılmasının yapılması sağlamış, karşılaştırma sonucunda ortak noktadaki harflerin kelime içerisindeki konumlarına göre derecelendirme yapmıştır. Çalışma sonucunda elde edilen özetler karşılaştırmıştır. Önerilen yöntem daha başarılı bir sonuca ulaşamamış olmasına karşın, yöntemin kabul edilebilir düzeyde ve kullanılmaya değer olduğunu gözlemlemiştir (Özkan, 2019).

Doğan, hem duygu analizi hem de metin özetlemenin bir arada yapıldığı bir çalışma yürütmüştür. Çalışmada metin özetleme kısmı için Latent Semantic Analysis (Gizli Anlamsal Analiz) kullanmıştır. Twitter üzerinde hashtag (etiket) ile alınan veriler üzerinde gerçekleştirdiği yöntemde ikili sınıflandırma yöntemiyle çeşitli kaynaklardan elde edilmiş 60.000 yorum üzerinde çalışma gerçekleştirmiştir. Ön işleme aşaması sonrasında sayısal temsil için Word2Vec ve FastText kullanmış ve kelime bazlı gömme durumu üzerinden ilerleme sağlamıştır. Sosyal ağ üzerinde kişilerin dilbilgisi ve yazım kurallarına çok fazla dikkat etmemesi sebebiyle düzenli bir şekilde yazılan makale vb. yazılı metne göre daha düşük başarı gözlemlemiştir (Doğan, 2019).

Goularte ve diğ. bulanık tabanlı bir yöntem önermişler ve bu noktada çok boyutluluk sorununun önüne geçmeye çalışmışlardır. Bir sanal öğrenme ortamında öğrencilere atanan görevler doğrultusunda verdikleri yanıt üzerinde Brezilya Portekizcesi dilindeki metinlerden bir veri kümesi oluşturmuşlardır. Ön işleme, kelime ve cümle puanlama, bulanık analiz ve metin özetleme olmak üzere dört adımda çalışmayı tamamlamışlardır. Kelime ve cümle puanlama kısmında istatistiksel ve deneysel çalışma yürütülmüş; cümle ya da kelime konumu, cümle uzunluğu, kelime frekansı ve cümleler arası benzerlik ölçütlerinden yararlanmışlardır. ROUGE metriği ile yaptıkları ölçümlerde bulanık tabanlı yaklaşımın %55'e yakın sonuç elde etmişlerdir (Goularte, Nassar, Fileto, ve Saggion, 2019).

Hark ve diğ. denetimli özetleme olarak adlandırılabilecek yaklaşımda metin içerisinde özet kısmında dahil edilmesi gereken metin parçalarının seçilme işlemini gerçekleştirmişlerdir. Denetimsiz özetlemede çeşitli ağırlık kazandırma yöntemi ile metin içerisindeki öğelerin önem dereceleri belirleyerek özet çalışmasını yürütmüşlerdir. Metin içerisinde özet niteliğini oluşturacak cümle ve metin parçalarının tespitine yönelik graf yapısındaki her bir düğüm üzerinde graf yapısından ayırma işlemi yaparak genel anlamda etkinin görülmesi ve sayısal olarak ölçümlemeyi sağlamışlardır. Bu çalışma ile hangi metin parçasının özetleme için ana metne dair temsil yeteneğinin yüksek olduğu da gözlemlenebilmiş ve özet kısmında dâhil edilebilmiştir (Hark, Uckan, Seyyarer, ve Karci, 2019).

Joshi ve diğ. SummCoder isminde tek doküman üzerinden çıkarım tabanlı metin özetleme için yeni bir yaklaşım sunmuşlardır. Natural Language Toolkit (NLTK) kütüphanesi ile yaptıkları Ön işleme sürecinde web ortamından elde ettikleri dokümanlardan HTML ve XML formatında olanlar için etiket temizleme ve gereksiz karakterlerin kaldırılma sürecini gerçekleştirmişlerdir. Cümle içerik bağımlılığı, cümle yeniliği ve cümle konum ilişkisi olmak üzere üç düzeyde cümle seçim

yöntemi kullanmışlardır. Cümle içerik bağımlılığını derin otomatik kodlayıcı ağı üzerinden, cümle yeniliğini cümleler arasındaki benzerlik oranı üzerinden, cümle konum ilişkisini ise belge uzunluğuna bağlı olarak dinamik bir yöntemle elde etmişlerdir. Bu üç metrik üzerinden cümleleri puanlamış ve özette yer alma durumlarını belirlemişlerdir. CNN/DailyMail , DUC-2002 veri setleri üzerinden karşılaştırma yapmış; ayrıca Tor Illegal Documents Summarization (TIDSumm) ismiyle yeni bir veri seti oluşturmuşlardır. ROUGE metriği ile yaptıkları değerlendirmede DUC-2002 veri setinde başarılı sonuçlar çıkmasa da, diğer veri setleri üzerinde yapılan çalışmalar literatürde bulunan yedi farklı yöntemle karşılaştırılmış ve ortaya rekabetçi bir sonuç çıkmıştır (Joshi, Fidalgo, Alegre, ve Fernández-Robles, 2019).

Karakoç ve Yılmaz, kodlayıcı-kod çözücü yöntem üzerinden soyutlama tabanlı başlık tahmin çalışması yapmışlardır. Veri seti üzerindeki metinleri Ön işlemeden geçirdikten sonra FastText kütüphanesi kullanarak kelimeleri vektörlere dönüştürmüşlerdir. Long Short-Term Memory (Uzun Kısa Süreli Bellek) (Lin vd. 2018) ve ışın aramasının kullanıldığı çalışma sonucunda elde edilen başlıklar ana metnin ilk cümlesi, ilk iki cümlesi ve metnin tamamı olmak üzere üç farklı parça üzerinden test etmişlerdir. Yapılan bu test işlemi ROUGE metriği üzerinden ölçülmüş ve metnin tamamının kullanılarak hazırlanan yöntemin daha başarılı olduğu gözlemlemişlerdir (Karakoç ve Yılmaz, 2019).

Mutlu, çoklu doküman üzerinde çıkarım tabanlı özetleme için cümleleri çeşitli özelliklerine göre vektör olarak ifade etmiştir. Birçok çalışmada kullanılan özellikleri bir araya getirmiştir. Bunlar terim frekansı, cümle pozisyonu, başlık uyumu, cümle konumu, cümle uzunluğu, cümle benzerliği gibi özellikleri kapsamaktadır. Bu özelliklerden yola çıkarak çeşitli kombinasyonlar oluşturmuş ve DUC-2002 veri seti üzerinde denemiştir. Yapay sinir ağı deneyimlerinin kişi deneyimlerine bağlı olarak özette yer alması için seçilen cümleler dışında

metin parçalarını çıkardığı gözlemlemiştir. Bulanık çıkarım sisteminin yapay sinir ağından daha iyi sonuç verdiğini, bulanık yapının ise literatürdeki mevcut bulanık sistemlerden daha iyi sonuca ulaştığını tespit etmiştir (Mutlu, 2020).

Song ve diğ. soyutlama tabanlı metin özetleme çalışması yürütmüşlerdir. Özetleme işlemi Long Short-Term Memory (Uzun Kısa Süreli Bellek) ve Convolutional Neural Network (Yapay Sinir Ağı) yapısının birlikte kullanarak mevcut cümlelerden daha küçük yapıda anlamsal bütünlükte yapılar oluşturmuşlardır. Çalışmanın dönemin mevcut soyutlama çalışmalardan ayrıldığı nokta ilk aşamada Ön işleme adımları gerçekleştirilmesi, ikinci aşamada kaynak metin içerisinden ifadeler çıkarılması, son aşamada da derin öğrenme yaklaşımı kullanarak özet oluşturulmasıdır. Ön işleme aşamasında CoreNLP kütüphanesini kullanmışlardır. İkinci aşamada ifade çıkarımı kısmında yeni bir yöntem olarak Multiple Order Semantic Parsing (Çoklu Sıralı Anlamsal Ayrıştırma) önermişlerdir. Bu yöntem ile elde edilen ifadelerden sözdizimsel ve anlamsal hatalardan dolayı bazı bölümlerini silmişlerdir. Son aşama olan özet üretme sürecinde veriler üzerinde eğitim işlemi sekiz dakika civarında sürmüştür. Bunun üzerine GPU üzerinde eğitim işlemine devam etmişlerdir. İki ve üç katmanlı bir derin öğrenme yapısını aşırı öğrenmeyi engellemek için kullanmışlardır. Gigaword (Graff ve Cieri, 2003) ve DUC veri setlerinde soyutlama tabanlı ve tek cümlelik özetler üzerinde yapılan çalışmalarda başarılı sonuçlar elde etmişlerdir. CNN/DailyMail haber metinleri kullanarak kişilerin oluşturduğu özet çalışmaları üzerinde yapılan çalışma anlamsal ve sözdizimsel yapılarda rekabetçi bir sonuç ortaya koymuştur. Çalışmayı ROUGE metriği üzerinde değerlendirmişler ve ROUGE-1 için %34,9 değeri, ROUGE-2 için %17,8 değerini elde etmişlerdir. Bu değerler o zamanki mevcut yöntemler üzerinde az bir farkla da olsa üstün duruma gelmiştir (Song, Huang, ve Ruan, 2019).

Mutlu ve diğ. metin özetlemede kullanılan çıkarım tabanlı yaklaşıma ek olarak anlamsal ilişki gibi parametreleri birlikte

kullanmıştır. Çıkarım tabanlı yaklaşım ile cümle içerisindeki kelimelerin sözdizimsel özelliklerinin yanında anlamsal özelliklerini de kullanmıştır. Literatürde yapılan çalışmalar bu iki yöntemden birine eğilmekteyken, bu çalışmada melez bir yöntem kullanılmıştır. Çalışmada Long Short-Term Memory (Uzun Kısa Süreli Bellek) ve yapay sinir ağı yapısının birlikte kullanımına dair bir yöntem önermiştir. Bulanık kural tabanlı yapay sinir ağı üzerinde yapılan çalışma hem literatürde hazır bulunan veri seti olan SIGIR (Wan ve Yang, 2008) ile hem de deneysel nitelikte üç kişinin özetlemeleri ile karşılaştırmaya olanak veren yeni bir veri seti üzerinde değerlendirmiştir. Değerlendirme sonucunda hem söz dizilim hem de anlamsal yaklaşımda başarılı sonuçlar elde edildiğini gözlemlemiştir (Mutlu, Sezer, ve Akcayol, 2019).

Erhandı derin öğrenme yapısını kullanarak metin içerisinden çıkarım tabanlı olarak başlık oluşturma çalışması yapmıştır. Epoch (periyot) sayısının parametrik olarak artırılmasına bağlı olarak sonucu gözlemlediği çalışmada epoch (periyot) sayısının artırılmasına yönelik çalışmayla daha iyi sonuçlar alınacağını belirtmiştir (Erhandı, 2020).

Hark genel özetleme yöntemi üzerinde maksimum düzeyde kapsamın belirlenmesi ile boyut azaltma işlemi uygulanan ve Fidan Gelişim Algoritması adı verilen optimizasyon temelli çalışma gerçekleştirmiştir (Hark, 2020).

Dudak, sezgisel Gri Kurt Optimizasyonu'nu kullanmıştır. Çıkarım tabanlı ve tek doküman üzerinde gerçekleştirdiği çalışmada sezgisel algoritmanın kümeleme yeteneği ile istatistiki birkaç parametreyi birleştirmiştir. K-means kümeleme algoritması ile de karşılaştırdığı çalışmada Gri Kurt Optimizasyon algoritmasının %6 oranında daha verimli olduğunu gözlemlemiştir. Önerilen yöntemi BBC News veri seti (ML Resources, 2010) üzerinde ROUGE metriği ile değerlendirmiştir. Elde edilen özetler Long Short-Term Memory (Uzun Kısa Süreli Bellek) ile sınıflandırmış ve

sınıflandırma sonucuna göre önerdiği yöntemin verimli olduğu ortaya koymuştur (Dudak, 2020).

Lamsiyah ve diğ. yapay sinir ağı temelli çıkarım tabanlı tek doküman özetleme ile ilgili bir yöntem önermişlerdir. DUC-2002 veri setinin kullanıldığı çalışmada metnin ilk sayısallaştırma aşamasında kelimelerin sırasının ve anlamının göz ardı edilmesi konusundaki eksikliği gidermeye çalışmışlardır. İleri beslemeli yapay sinir ağı kullandıkları modelde Ön işleme yapmış ve doküman içerisindeki metni cümlelere ayırmışlardır. Ardından NLTK (Natural Language Toolkit) kütüphanesini kullanarak cümleler içerisindeki tüm kelimeleri küçük harflere çevirmiş ve özel karakterler ile gereksiz metin parçalarını temizlemişlerdir. Cümle bazlı vektör ile doküman bazlı vektörlerin oluşturulması sonrasında cümle bazlı vektör ile doküman vektörünün çarpımı ve farklarının mutlak değerlerinin alındığı iki farklı eşleştirme operasyonu kullanılmışlardır. Cümlelerin puanlanmasında kullandıkları ileri beslemeli yapay sinir ağında üç gizli katman için ReLu aktivasyon fonksiyonu, çıkış katmanı için Sigmoid aktivasyon fonksiyonu, kayıp fonksiyonu için binary cross-entropy kullanmışlardır. Ortaya çıkan özet çalışmasını değerlendirmek için ROUGE metriği kullanmışlardır. Değerlendirme sonucunda önerdikleri yöntemin literatürde sıklıkla kullanılan sekiz yönteme göre daha başarılı sonuçlar verdiği görmüşlerdir (Lamsiyah, El Mahdaouy, El Alaoui, Espinasse, ve Ezziyyani, 2020).

Uçkan çıkarım tabanlı ve denetimsiz iki farklı yöntem önermiştir. Önerdiği yöntemleri graf tabanlı hazırlamıştır. Metni temsil edilen parçalar için normalizasyon işlemi, kümeleme algoritması ve ağırlıklandırma yöntemini kullanmıştır. Merkezilik değerlerine bağlı olarak cümlelerin özet bölümünde yer almasına karar vermiştir. Ayrık olarak değerlendirilen kümeleme dışındaki parçaların özet içerisinde yer almaması sağlamıştır. Yapılan çalışmayı DUC-2002 ve DUC 2004 veri setleri üzerinde ROUGE metriği ile test etmiştir.

Değerlendirme sonucunda özellikle 200 ve 400 kelimelik özetlerde iyi sonuçlar almıştır (Uçkan, 2020).

Xie ve diğ. metin özetleme için yapay sinir ağı modelini kullanmışlardır. Doğal Dil İşleme alanında sıkça kullanılan self-attention (öz-dikkat) modelini temele alan bir yöntem önermişlerdir. Yöntem kapsamında mevcut self-attention (öz-dikkat) yapısına şartlı durum mekanizmasını eklemişlerdir. Yöntem kapsamında parametre durumundaki metinlerin parametre ile sonuç kısımlarında eşleşme durumları için puanlama yapmışlardır. Mevcut self-attention (öz-dikkat) yapılarından farklı olarak çapraz bir bileşen uygulamışlardır. Verilen bir parametre üzerinden, önce çapraz bileşen doğrultusunda bağımlılık puanı hesaplanmış ve ölçeklendirme yapmışlardır. Sonrasında bu ölçek üzerinden self-attention (öz-dikkat) yapısını uygulamışlardır. Bu şekilde koşul bağlamında ilişkili parametrelerin seçimi mümkün olmaktadır. Çalışmanın özgün olan diğer yapısı ise yöntemde bağımlılık kısmının hem yerel hem de genel düzeyde erişilebilir olması sebebiyle avantajlı durumda olan kısmın hemen seçilebilmesidir. Geleneksel yöntemde bu iki bilgi için iki farklı modülün tasarlanması gerekmektedir. Değerlendirmelerde sorgu tabanlı yöntemin diğer yöntemlere göre başarılı sonuçlar ürettiğini gözlemlemişlerdir (Xie, Zhou, Mao, ve Chen, 2020).

Xu ve Lapata sorgu temelli çoklu doküman özetleme yaklaşımı kullanmışlardır. Sorgu ile bilgi elde etme yöntemi içerisinde uzak denetimden yararlanmışlardır. Sorgu ile elde edilen sonuçların ilgi düzeylerini tespit etmek için genelden özele yaklaşım kullanan bir model sunmuşlardır. Tahmin mekanizması üzerinden özette hangi bölümlerin yer alması gerektiğini çıkarmışlardır. Tahmin mekanizmasına ek olarak çalışmanın özgünlüğünü ortaya çıkaran eğitilebilir bir kanıt bulma katmanı eklemişlerdir. Bu katman anlamsal ilişkileri de kullanarak nihai olarak özete eklenecek metin parçalarını belirlemiştir. Bu kısım cümle tabanlı ya da paragraf tabanlı çalışmaktadır. Çıkarım tabanlı yöntem kullanmışlardır.

Geleneksel yöntemler üzerindeki sorgu tabanlı çalışmalar Terim Frekansı / Ters Doküman Frekansı kullanırken bu yöntemde merkezilik doğrultusunda soruya en iyi cevabı verme ve bunun tahmin-kanıtlayıcı modelle sunulması bu çalışmayı özgün kılmıştır. Geliştirme sürecinde DUC-2005 veri seti, eğitim ve test aşamasında DUC-2006 ve DUC-2007 veri setleri üzerinde çalışma yapmışlardır (Xu ve Lapata, 2020).

Yang ve diğ. PGAN-ATSMT isminde bir yapı önermişlerdir. Bu yapı kapsamında dilbilgisi kurallarına uygun ana doküman içerisindeki önemli bilgilerin elde edilirken yüksek performans kazancı sağlanacağını vurgulamışlardır. Soyutlama tabanlı özetleme tekniğini kullandıkları çalışmada ilk olarak olasılık tabanlı üretken çekişmeli bir ağ modeli sunmuşlardır. Ayırt edici modeli temsilen D, üretken modeli temsilen G modelini oluşturmuşlardır. Çekişmeli öğrenme yolu ile birlikte iki modelin birlikte eğitilmesi gerçekleşirken G modeli girdi üzerinden bir metnin özetini çıkarmıştır. G modeli üretilen özet ve ana metin üzerinde çalışarak rekabetçi bir yöntem ile G modelinin en iyi özeti oluşturmasını sağlamıştır. Bu aşama sonrasında Long Short-Term Memory (Uzun Kısa Süreli Bellek) kod çözücüyü kullanarak kategorilendirme işlemini gerçekleştirmiş, ardından sözdizimi oluşturmayı tamamlamışlardır. Eksik cümleler, yinelenen kelimeler ve bilgilerin atılmasını bu aşamada gerçekleştirmişlerdir. Önerilen yöntem CNN/Daily Mail ve Gigaword veri setleri üzerinde çalıştırılmış ve ROUGE-1, ROUGE-2 ve ROUGE-L metriğini kullanarak yaptıkları değerlendirme sonucunda alanında benzer çalışmalara göre olumlu yönde bir başarı elde etmişlerdir (Yang, ve diğerleri, 2020).

Bu tez çalışması kapsamında Türkçe dilinde yazılmış metinler üzerinde özetleme çalışması yürütülmüştür. Yürütülen çalışmada Türkçe dilinin sondan eklemeli bir dil olması ve kök bulmanın zor olması gibi durumlar değerlendirilmiştir. Literatürde yapılan çalışmalarda hemen hemen tüm diller için geçerli olabilecek yöntem ve algoritma kullanılırken, tez

kapsamında önerilen büyük ünlü uyumu, küçük ünlü uyumu ve her iki yöntemin birlikte kullanıldığı hibrit yöntem deneysel çalışma kapsamında yürütülmüştür. Süreçte iki kuralın Türkçe diline özgü kurallar barındırması ve bir kelimenin öz Türkçe nitelikte olup olmadığına karar verilirken ilk yol olması bakımından kullanımı değerlendirilmiştir. Önerilen yöntemlerin literatürde bulunan sekiz farklı yöntemle karşılaştırması yapılmış ve bu yöntemlere göre daha fazla başarı elde edilmiştir. Cümle uzunluğu, ilk cümle, son cümle gibi literatürde bulunan yöntemler Türkçe diline özgü nitelikte bir unsur barındırmadığından, bu yöntemlerin kullanımı dile özgü olmayan bir çalışma niteliğinde olacaktır. Tez kapsamında önerilen yöntemler Türkçe diline özgü olup literatüre katkı sağlayacaktır.

BÖLÜM 3.

OTOMATİK METİN ÖZETLEME

Büyük çapta sayılabilecek belgelerin insanlar tarafından okunması ve özetlenmesinin uzun zaman alması ve yorucu olması sebebiyle metin özetleme çalışmaları yürütülmektedir. Metin özetlemenin genel amacı ana kaynak olarak nitelendirilen dokümanın önemli noktalarının kaybedilmeden kısa bir türünün oluşturulmasıdır. Joshi otomatik metin özetleme çalışmasını, büyük hacimdeki dokümanların daha küçük hacimli bir türünü elde etmek için gerçekleştirilen Doğal Dil İşleme dalı olarak tanımlanmıştır (Joshi, Fidalgo, Alegre, ve Fernández-Robles, 2019).

Otomatik metin özetleme çalışması özetlenecek doküman sayısı, özet türü ve özet içeriği noktalarında alanlara ayrılmaktadır. Bu ayrım Şekil 1'de sunulmuştur.

Şekil 1. Otomatik Metin Özetleme Türleri

3.1. Kaynak Çeşitliliğine Bağlı Özetleme

Otomatik metin özetleme işlemi doküman sayısına bağlı olarak tek belgeli ve çok belgeli olmak üzere ikiye ayrılmaktadır.

3.1.1. Tek belge özetleme

Tek belge özetlemede yalnızca tek bir dokümanın özetleme çalışmasına odaklanılmaktadır. Tez çalışması kapsamında elde edilen haber metinlerinin oluşturulduğu veri seti bu yapıdadır.

3.1.2. Çok belgeli özetleme

Çok belgeli özetlemede en az iki doküman içeren bir doküman kümesi üzerinde özetleme çalışması yürütülmektedir. Bu dokümanlar genellikle benzer konu içeriklerinin farklı yazar ya da kurumlar tarafından kaleme alındığı, farklı kaynaklardan beslenerek konuya dair özetin çıkarılmasını sağlar.

3.2. İçeriğe Bağlı Özetleme

Otomatik metin özetleme işleminde oluşturulan özetlerin içeriğine bağlı olarak alana özgü özet, sorgu bazlı özet ve genel özet olmak üzere üç farklı ayrım bulunmaktadır.

3.2.1. Alana özgü özetleme

Alana özgü özet için belirli bir alan üzerinde yazılmış bir dokümanın özetlenmesi yürütülmektedir. Diğer türlerden farklı olarak bu alan için önemli kavramların belirlenmesi ve alana özgü bazı özelleştirme çalışmalarını da içermektedir. Örneğin nükleer enerji konusunda bir dokümanın özetlenme sürecinde alana özgü kavramların belirlenmesi, bu alana özgü bazı özelleştirme çalışmaları da yapılabilmektedir.

3.2.2. Sorgu bazlı özetleme

Sorgu bazlı özetlemede son kullanıcı tarafından üretilen ya da iletilen sorgular doğrultusunda özet oluşturulmaktadır. Bu sorgular önceden yapılandırılmış sabit nitelikteki sorgular olabilirken, kullanıcıların özelleştirebildiği yapıları da kapsayabilmektedir.

3.2.3. Genel özet

Genel özet ise herhangi bir ek bilgi olmaksızın (alana özgü durum, sorgu vb.) dokümanın özetlenmesi işlemidir. Bu yöntemde metin içerisindeki bilginin genel hatlarıyla özetlenmesi amaçlanmaktadır.

3.3. Çözüm Stratejisine Göre Özetleme

Otomatik metin özetlemede üretilen özetin türüne göre soyutlama tabanlı ve çıkarım tabanlı olmak üzere iki farklı özet türü bulunmaktadır.

3.3.1. Soyutlama tabanlı metin özetleme

Bu özet türünde dokümandaki genel kapsama bağlı olarak yeni kelime ve cümle temsilleriyle özet oluşturulur. Yeni bir temsil işlemi olduğundan dilbilimsel açıdan zorluklar bulunmakta ve zor bir görev olarak nitelendirilebilir. Nitekim insanlar bir metni okurken içerikte bulunan ifadelerin neden-sonuç ilişkilerini özümser, bunu başka bir kişiye sözlü olarak anlatmak ya da yeniden kaleme almak istediklerinde bu neden sonuç bağlamında yeni kelime ve cümleler ile ifade etmeye çalışırlar. Genel olay ve konu örgüsünü bozmadan yapılan bu özetleme soyutlama tabanlı özetleme ile paralellik göstermektedir. Soyutlama tabanlı özetleme insanlar tarafından oluşturulmuş doğal görünümlü özet oluşturmayı amaçlamaktadır (Song, Huang, ve Ruan, 2019). Soyutlama tabanlı özetlemede yapısal kısımdan çok anlamsal yapı kullanılmaktadır. Özetlemede ilk önce doküman içerisinde adlandırılmış varlık niteliğindeki kısımlar belirlenir ve bunlar üzerinde anlamsal ilişkiler kurulur. Kurulan bu ilişki üzerinden özetleme yapılacak dilin kuralları üzerinden yeni kelime ve cümleler ile ifade etme çalışmaları yürütülür (Yang, ve diğerleri, 2020). Dilin kurallarına tam anlamıyla uyulması gerekliliğinden, ilgili konunun tam olarak yansıtılmasında eksikler olabileceği bu yöntemin zorluklarını oluşturmaktadır.

3.3.2. Çıkarım tabanlı özetleme

Çıkarım tabanlı özetlemede dokümanın kapsamına en uygun bir alt kümesi seçilir. Bu alt küme seçim işlemi çeşitli yöntemler ile elde edilmekte ve konunun yeni cümle ve kelimeler ile ifadesi olmaksızın gerçekleştirildiğinden soyutlama tabanlı özetlemeye göre daha kolay bir işlemdir. Hali hazırda doğal dile uygun yazılan bir metnin içerisinden cümle ya da cümleler alındığından doğal dile uygunluk gibi bir durum için ek çalışmaya ihtiyaç yoktur. Bu kolaylığından ötürü daha fazla tercih edilmektedir.

Çıkarım tabanlı özetlemede avantajlar olduğu gibi bazı noktalarda da çeşitli zorluklar bulunmaktadır. Genellikle uzun cümleler çıkarım tabanlı cümlelerde en fazla tercih edilen yapılar olmaktadır. Bunun nedeni uzun cümle içerisinde çok fazla bilginin yer alması ve çoğunlukla bir cümle ile birkaç cümlede ifade edilen yapının tek cümlede yer almasıdır. Bu durum içerisinde gereksiz kelimelerin ya da kelime öbeklerinin tekrarlanmasıyla bir dezavantaja da dönüşebilmektedir.

Tez kapsamında yapılan deneysel çalışmaların tamamında tek belgeli, çıkarım tabanlı ve genel özet türünde otomatik metin özetleme çalışması yürütülmüştür.

BÖLÜM 4.

YÖNTEM VE UYGULAMALAR

4.1. Çalışma Grubunun Oluşturulması

Çalışma kapsamında önerilen otomatik metin özetleme yönteminin başarımının değerlendirilmesi için gerekli referans setlerini oluşturmak için bir çalışma grubu oluşturulmuştur. Çalışma grubu üyelerin veri setinde bulunan haberlerin özetleme çalışmasını yürütmesi için alanında uzman niteliğinde kişiler seçilmiştir. Bunun için Türk Dili ve Edebiyatı Öğretmeni olan iki üye, ardından bir Psikolojik Danışman çalışma grubuna eklenmiştir. Üyelere ait bazı özellikler Tablo 1'de sunulmuştur.

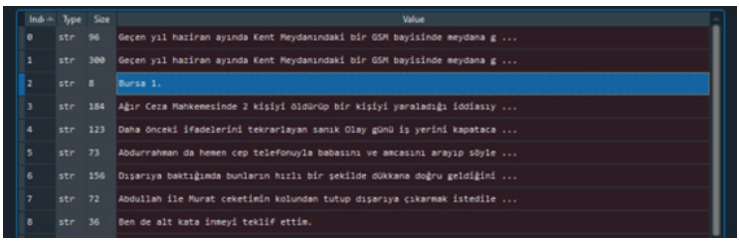
Tablo 1. Çalışma grubu üye özellikleri

Üye	Mesleği	Yaşı	Uzmanlığı
Üye 1	Öğretmen	32	Türk Dili ve Edebiyatı
Üye 2	Öğretmen	29	Psikolojik Danışman
Üye 3	Öğretmen	34	Türk Dili ve Edebiyatı

4.2. Veri Setinin Oluşturulması

Fırat Üniversitesi Büyük Veri ve Yapay Zeka Laboratuvarı tarafından hazırlanan ve içeriğinde ulusal haber sitelerinden derlenen, her bir habere ait başlık, haber içeriği ve anahtar kelimelerin bulunduğu bir veri setinin bir ölçeği kullanılmıştır

(Fırat Üniversitesi - Büyük Veri ve Yapay Zeka Labo, 2015). Veri seti içerisinden rastgele seçilen 215 adet haber üzerinde çalışma yürütülmüştür. Veri setinde bulunan haberler herhangi bir önişleme işlemi olmaksızın cümlelere ayrılmış ve çalışma grubuna cümle bazında sunulmuştur. Cümlelere ayırma işlemi için NLTK kütüphanesi kullanılmış, ancak kütüphanenin sonuçları değerlendirildiğinde bazı haberlerde hatalar olduğu tespit edilmiştir. Şekil 2'de sunulan örnek bir hata için ek bir modül oluşturulmuş ve bu sorunun önüne geçilmiş ve cümleler doğru şekilde ayrılmıştır. Bu örnekte ilgili fonksiyonun cümlelere ayırması beklenirken, bir cümlenin "Bursa 1." olarak ayrılması ve bu metin parçasının bir cümle olmaması yapılan işlemde hata olmasına neden olmuştur. Bu kapsamda cümle sonlarının ve bir sonraki cümlenin başlangıcının tespit edilmesi için ek bir fonksiyon yazılmış ve bu sorun giderilerek cümleler sağlıklı bir şekilde ayrılmıştır.



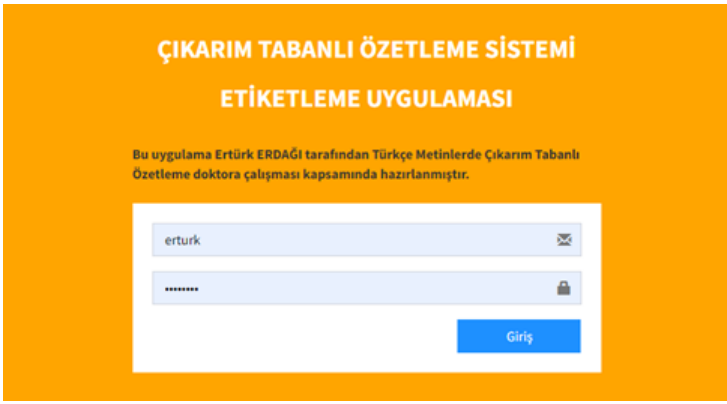
Ind	Type	Size	Value
0	str	96	Geçen yıl haziran ayında Kent Meydanındaki bir GSM bayisinde meydana g ...
1	str	300	Geçen yıl haziran ayında Kent Meydanındaki bir GSM bayisinde meydana g ...
2	str	8	Bursa 1.
3	str	184	Ağır Ceza Mahkemesinde 2 kişiyi öldürüp bir kişiyi yaraladığı iddiasıy ...
4	str	123	Daha önceki ifadelerini tekrarlayan sanık Olay günü iş yerini kapataca ...
5	str	73	Abdurrahman da hemen cep telefonuyla babasını ve amcasını arayıp söyle ...
6	str	156	Dışarıya baktığımda bunların hızlı bir şekilde dükkana doğru geldiğini ...
7	str	72	Abdullah ile Murat ceketimin kolundan tutup dışarıya çıkarmak istedile ...
8	str	36	Ben de alt kata inmeyi teklif ettim.

Şekil 2. Haberlerin Cümlelere Ayrılmasındaki Hata

4.3. Web Uygulamasının Oluşturulması

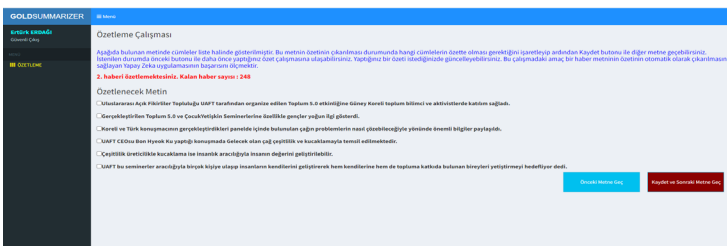
Veri seti içerisinde bulunan 215 adet haberin otomatik metin özetleme yöntemleri ile özetlenmesinin ardından özetin başarımını ölçümlemek için çalışma grubu tarafından oluşturulan özetlere ihtiyaç bulunmaktadır. Çalışma grubunun veri seti üzerinde özetleme çalışmalarını yürütmesi ve elde edilen verilerin üye bazında saklanması için bir web uygulaması hazırlanmıştır. Çalışma grubu üyeleri her bir haberin özetleme çalışması için etiketleme çalışması yürütmüştür. Üyeler Şekil

3'te sunulduğu üzere giriş işlemlerini kendilerine verilen kullanıcı adı ve şifre ile gerçekleştirmişlerdir.



Şekil 3. Web uygulaması kullanıcı giriş modülü

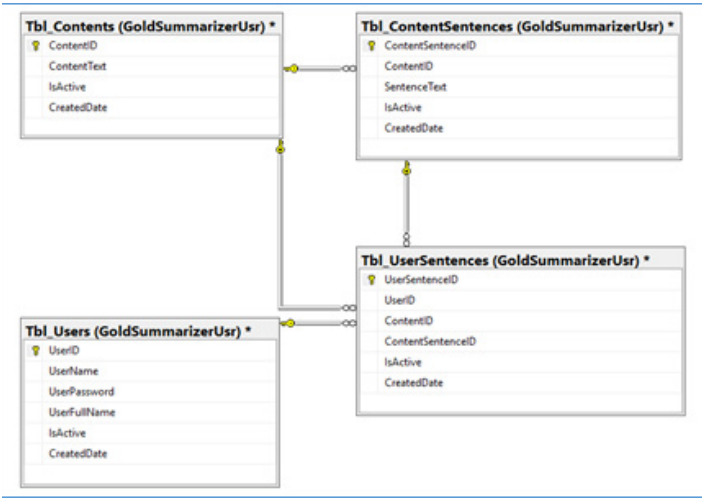
Uygulamaya giriş yapıldıktan sonra grup üyesinin hangi haberleri özetlediği veri tabanı üzerinden okunmakta ve etiketleme işleminin yapılmadığı kontrol etmekte ve sıradaki haber kullanıcıya sunulmaktadır. Şekil 4'te sunulan modülde kullanıcı özette yer alması gerektiğini düşündüğü cümleyi seçerek işlemi tamamlayabilmektedir. Bu kısımda kullanıcılara bir özet oranı kısıtlaması getirilmemiş olup, özet oranının haber özelinde değerlendirmesi gerçekleştirilmiştir.



Şekil 4. Web uygulaması özetleme (etiketleme) modülü

Kullanıcıların tüm seçimleri işlem bazında veri tabanında saklanmış ve sonrasında veri seti üzerine eklenmiştir. Şekil 5'te sunulan veri tabanı yapısında veri setinde bulunan haberler,

haberlere ait cümleler ve kullanıcıların özet kısmı için seçtikleri cümleler saklanmıştır.



Şekil 5. Web uygulaması veri tabanı yapısı

Veri tabanındaki Tbl_Contents tablosunda veri setinde bulunan haberlerin herhangi bir işlem uygulanmamış hali yer almaktadır. Tbl_ContentSentences tablosunda haberlerin cümlelere ayrılmış, ancak herhangi bir ön işleme aşamasına tabi tutulmadığı hali saklanmıştır. Burada ContentSentenceID ile her bir cümleye nümerik ve benzersiz bir değerin atanması sağlanmıştır. Tbl_Users tablosunda kullanıcıların web uygulamasına girişte kullanacakları kullanıcı adı ve şifre bilgileri saklanırken, yapılan özetleme çalışmalarında her bir kullanıcıya ait verinin saklanması için UserID isminde benzersiz bir nümerik değer atanmıştır. Tbl_UserSentences tablosunda kullanıcının bir içerik için hangi cümleleri seçtiği bilgisi saklanmıştır.

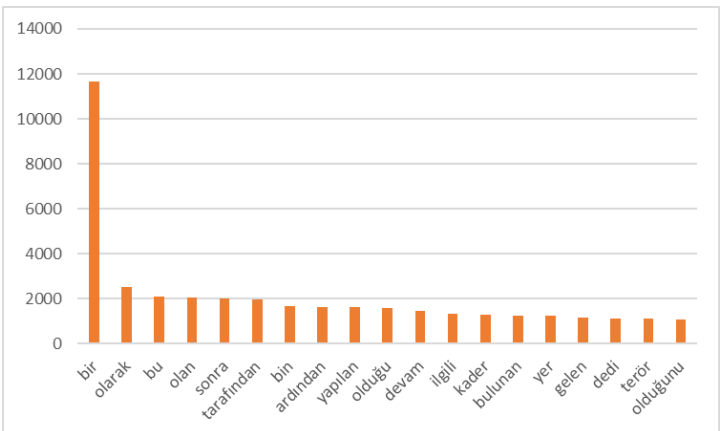
Tüm kullanıcıların etiketleme çalışmasının tamamlanması ardından elde edilen sonuçlar raporlanmış ve bu veriler analiz edilmiştir. Elde edilen sonuçların tek bir özet kısmına indirgenmesi için ek bir çalışma yürütülmüştür. Bu çalışmada özetlenen haberlerde üç kullanıcının da seçtiği cümleler analiz

edilmiş, eğer bir haber için aynı özet yapılmışsa tek özete indirgenmiş, farklı özetler için yoğunlaşılan noktalarda ortak değer bulunması için minimum özet cümle sayısı baz alınmış ve bu cümleye ek olarak gelen cümlelerin metrik üzerinde değerlendirilmesi sağlanmıştır. ROUGE metriği kullanılarak yapılan bu çalışmada ilk elde edilen ortalama ROUGE-1 değeri üzerinden yapılan değerlendirme sonrasında her bir cümlenin eklenmesi ile ilk oranın ortalama bazda daha iyi sonuç verene kadar cümle eklenmiştir. Genel bir analiz yapıldığında kullanıcıların birbirlerinden çok farklı oranda özetleme yapmadıkları gözlemlenmiş ve belirtilen bu ek çalışma veri setinin küçük bir kısmına uygulanmıştır.

Veri setine son referans özet ve bu özetin içerdiği cümle sayısı eklenerek veri seti ön işleme aşamasına hazırlanmıştır.

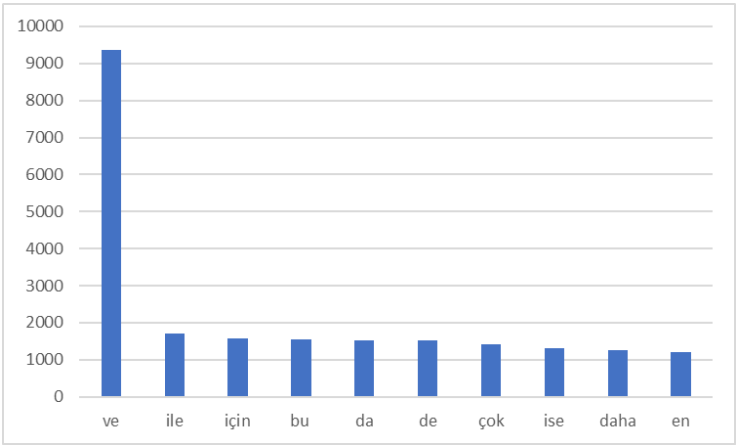
4.4. Ön İşleme Öncesi Analiz

Veri setinde bulunan haberler ön işleme öncesinde kelime ve cümle bazında analiz edilmiştir. Şekil 6'da sunulan kelime frekans grafiği incelendiğinde bazı durak kelimelerin yoğun olduğu gözlemlenirken, bunun dışında var olan kelimelerin veri setinin haber metinleri içermesi sebebiyle yoğun olduğu çıkarımı yapılmıştır.



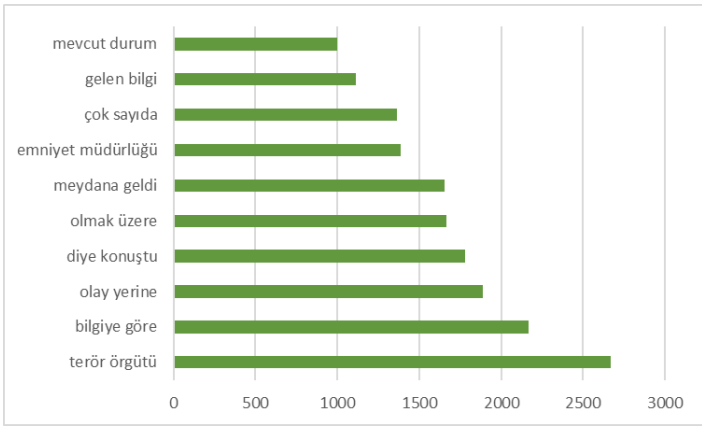
Şekil 6. Ön İşleme Aşaması Öncesinde Kelime Frekans Grafiği

Şekil 7'de sunulan durak kelime frekans grafiğinde cümleye anlam katmayan ya da sık tekrarlanmasının terim frekansı özetleme için yanlış yönlendirme içerecek kelimelerin analizi sunulmuştur. Bu kelimelerin bağlaç ve zamir türünde ağırlıklı olduğu gözlemlenmiştir. Durak kelimelerin tespitinde NLTK kütüphanesi kullanılmıştır. Bu kelimeler ön işleme aşamasında saf dışı bırakılmıştır.

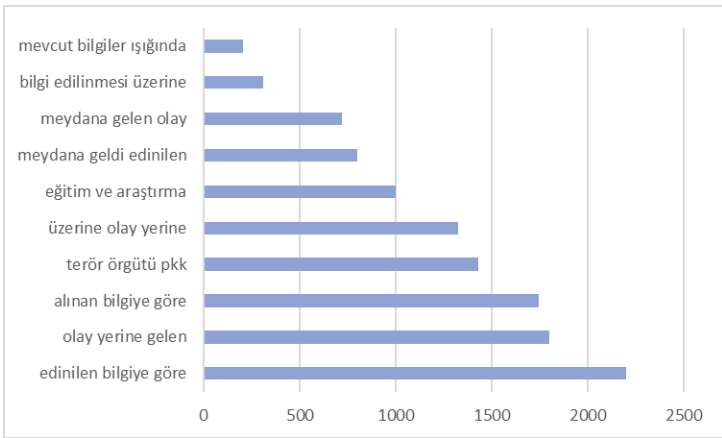


Şekil 7. Durak Kelime Frekans Grafiği

Veri setinde kalıplaşmış ifade ve ardışık sıklıkla kullanılan yapıların tespiti için n-Gram modeli kullanılmıştır. Bu kapsamda veri seti üzerinde aşağıda sunulan Şekil 8 ve Şekil 9 grafiklerinde bulunan bigrams ve trigrams grafikleri üzerinde yapılan incelemelerde haberin işaret edilme şekline bağlı olarak kalıp ifadelerin sıklıkla kullanıldığı görülmektedir. Bu ifadeler durak kelimeler ile karşılaştırılmış ve çok fazla ortak nokta bulunmadığından çalışmayı olumsuz yönde etkileyecek bir unsur olarak değerlendirilmemiştir.

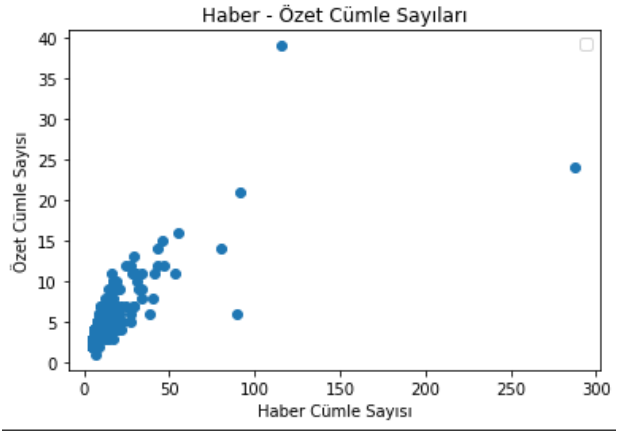


Şekil 8. Bigrams grafiği



Şekil 9. Trigrams grafiği

Şekil 11'de bulunan haber cümle sayısı ve özet cümle sayısı grafiği incelendiğinde haberlerde bulunan cümle sayıları ile özet cümle sayıları paralellik göstermektedir. Bu durum haber metninde yer verilecek çok fazla unsurun olduğu durumlarda özetin de buna bağlı olarak genişlediğini göstermektedir. Özellikle adlandırılmış varlık olarak nitelendirilen kurum, kuruluş, yer, zaman vb. durumlar ilgili haber metninde kritik öneme sahip olduğundan bu bilgilerin özette de yer alması kaçınılmaz olacak ve özetteki cümle sayısı artacaktır.



Şekil 10. Haber ve İnsan Özetlerindeki Cümlelerin Sayı Yönünden Karşılaştırılması

4.5. Ön İşleme Çalışması

Ön işleme aşamasında ilk olarak metin içerisindeki durak kelimeler çıkarılmıştır. Bu kelimeler çeşitli kütüphane ve üçüncü parti yazılımlarda her dil için hazırlanmış bir veri kümesi üzerinden gerçekleştirilmektedir. Bu kelimelere ek olarak noktalama işaretleri, seslenme ve ünlem içeren kelimeler de atılmıştır.

Kelimeler aldıkları ek yapılarına göre çeşitli anlam ve yapısal farklılıklara büründüklerinden kelimelerinin asıl köklerinin anlam kaybı olmaksızın tespit edilmesi hem yapısal hem de anlamsal olarak avantaj sağlamaktadır. Kelime köklerinin tespitine yönelik TRNLP (Github - TRNLP, 2008) kütüphanesi kullanılmıştır.

4.6. Değerlendirme Metrikleri

Çalışma kapsamında üretilen özetlerin değerlendirmesi için ROUGE ve BLEU metrikleri kullanılmıştır. ROUGE, özetlenen verilerin orijinal metne ne kadar uygun olduğunu ölçmek için kullanılan, özetler arasında otomatik değerlendirme sağlamak için geliştirilmiş bir kütüphanedir (Lin C. , 2004).

ROUGE-N geri çağırmaya dayalı ve n-gram karşılaştırmasını temel alan bir ölçüdür. Bu yöntemde verilen bir özetin istenilen özet değerine uygunluğu n-gram kelimelerinin örtüşme oranına bakılarak belirlenir. Elde edilen f-skoru 0-1 arasındadır, amacına ulaşmış bir özetin f-skoru 1'e yaklaşmaktadır. ROUGE-N için unigram (tekli), bigram (ikili) ve trigram(üçlü) olmak üzere üç farklı varyasyon bulunmaktadır.

Bu çalışma kapsamında ROUGE-1 ve ROUGE-2 kullanılmıştır. ROUGE-1 ve ROUGE-2 ile tekli ve ikili söz öbeklerinin insan özetleri ile daha fazla örtüşmesi, üçlü kelimelerin örtüşme oranının diğer iki yapıya göre daha az olması, üçlü yapının ağırlıklı olarak zincirleme isim tamlaması dışında çok fazla örtüşme sağlayamayacağı düşünülmüştür.

IBM tarafından önerilen BLEU, sistem çıktısının benzerliğini, çevirmenler tarafından daha önce oluşturulan "n" adet referans çeviriyle ölçer (Papineni, Roukos, Ward, ve Zhu, 2002). Benzerlik, kelime ve deyimleri eşleştirerek ölçülür. Temelde hassas hesaplamaya dayanır. Kesinlik hesaplaması, Denklem 4.6.1'de sunulduğu gibi referans cümle içerisinde özete aday cümledeki kelimelerin kesişim sayısının, özete aday cümledeki toplam kelime sayısına bölünmesiyle elde edilir.

$$\text{BLEU} = \frac{\textit{Özete aday cümlede kelimeler} \cap \textit{İnsan özetindeki kelimeler}}{\textit{Özete aday cümledeki toplam kelime sayısı}} \quad (4.6.1)$$

4.7. Cümle Derecelendirme Yöntemleri

Çalışma kapsamında cümlelerin özette yer alıp almamasını belirlemek için derecelendirme işlemleri uygulanmıştır. Bu yöntemlerde geleneksel yöntemlerin yanı sıra literatüre katkı niteliğinde olabilecek üç farklı derecelendirme yöntemi de bulunmaktadır.

4.7.1. Terim frekansı yöntemi

Metin verileri içerisinde özellik çıkarımı için genellikle bir kelimenin terim sıklığı veya belge sıklığı kullanılır (Azam

ve Yao, 2012). Terimin ilgili metin içerisinde sık geçmesinin terimin önemli bir unsur olduğu varsayımına dayanan bu yöntemde ilgili kelimelerin geçtiği cümlelerin de önem derecesi artmaktadır. Bu kapsamda metin içerisinde sık geçen ve herhangi bir anlama veya katkıya sahip olmayan durak kelimelerin önişleme aşamasında cümleden çıkarılması gerekmektedir. Aksi durumda bu kelimelerin terim frekansı yüksek olacak ve bu kelimeler cümle ağırlıklandırma aşamasında yanlış sonuç elde edilmesine neden olacaktır.

Veri setinde bulunan haberlerin tamamı için ayrılan cümleler bazında terim sıklığı işlemi uygulanmış ve bu terim sıklığının tüm haber içerisindeki oranı o cümle için terim sıklık puanı olarak belirlenmiştir. Tablo 2'de veri setinde bulunan bir haberin puanlaması örnek olarak sunulmuştur.

Tablo 2. Terim Frekansı Yöntemi ile Cümle Derecelendirme Örnek Verisi

Haber No	Cümle Sırası	Cümle	Terim Frekansı Puanı
215	0	Ali Haydar Paksoy 41…	0,2307
215	1	Pist alanından havalana…	0,2153
215	2	Paksoy ihbar üzerine…	0,2
215	3	Evli ve 2 çocuk babası…	0,2769
215	4	Paksoyun paraşütünün…	0,2153
215	5	Öte yandan Paksoyun…	0,2

4.7.2. Başlık yöntemi

Her bir habere ait başlık kısmındaki kelimeler konu kapsamını geniş ölçüde yansıtmaktadır. Bu yöntem kapsamında başlığın ilgili haber cümlesine vektörel yakınlığı kullanılmıştır.

Türkçe için oluşturulmuş bir BERT modeli üzerinde ilgili başlığın ve referans noktasındaki cümlenin kosinüs benzerliği hesaplanmış ve ilgili cümlenin puanı olarak belirlenmiştir.

Tablo 3'te veri setinden bir örnek üzerinde başlık yönteminin uygulanması ile cümle bazında elde edilen sonuç sunulmuştur.

Tablo 3. Başlık Yöntemi ile Cümle Derecelendirme Örnek Verisi

Haber No	Cümle Sırası	Cümle	Başlık Puanı
215	0	Ali Haydar Paksoy 41 Pamukkale…	0,5125
215	1	Pist alanından havalanan Paksoy…	0,5820
215	2	Paksoy ihbar üzerine olay…	0,5310
215	3	Evli ve 2 çocuk babası olduğu…	0,4676
215	4	Paksoyun paraşütünün ters…	0,5750
215	5	Öte yandan Paksoyun düşme…	0,5412

4.7.3. Anahtar kelime yöntemi

Veri seti içerisinde her bir haber için anahtar kelime bilgisi bulunmaktadır. Kimi haber için tek bir anahtar kelime varken, konu kapsamı geniş olan haberler için birden fazla anahtar kelime bulunmaktadır.

Anahtar kelimeler çoğunlukla haberin arama motorlarında kolaylıkla bulunması, bir haberdeki anahtar kelimeden benzer nitelikteki habere ulaşım, kategorik olabilecek haberlerin etiketlenmesini sağlayan yapıdır.

Veri seti içerisindeki anahtar kelimeler incelendiğinde özellikle haber içerikleri ve yer adlarına yönelik yoğun bir kullanım olduğu gözlemlenmiştir. Bu kısım için özellikle derecelendirme yöntemlerinden biri olan varlık tespitinin önemli olacağı çıkarımı yapılabilmektedir.

Anahtar kelimeler üzerinde BERT modeli kullanılmış, cümlelerin puanlanmasında cümle ve anahtar kelimeleri arasında kosinüs benzerliği hesaplanmıştır. Tablo 4'te veri

setinden bir örnek üzerinde anahtar kelime yönteminin uygulanması ile cümle bazında elde edilen sonuç sunulmuştur.

Tablo 4. Anahtar Kelime Yöntemi ile Cümle Derecelendirme Örnek Verisi

Haber No	Cümle Sırası	Cümle	Anahtar Kelime Puanı
215	0	Ali Haydar Paksoy 41…	0,4961
215	1	Pist alanından havalanan…	0,5575
215	2	Paksoy ihbar üzerine…	0,4835
215	3	Evli ve 2 çocuk babası…	0,4065
215	4	Paksoyun paraşütünün…	0,6024
215	5	Öte yandan Paksoyun…	0,4933

4.7.4. Cümle konumu (ilk cümle) yöntemi

Metinde çoğunlukla giriş cümlesi konuya dair önemli ipuçları vermektedir. Bu kısımda yer alan kavram, yer, tarih, kişi ve zaman bilgisi ilerleyen cümlelerde açıklanabilecek durumdadır.

Metinlerde gelişme ve sonuç bölümlerine ışık tutacak ilk cümledeki kelimeler üzerinde BERT modeli kullanılmış, cümlelerin puanlanmasında ilk cümle ile kosinüs benzerliği hesaplanmıştır. Tablo 5'te veri setinden bir örnekte cümle konumu (ilk cümle) yönteminin uygulanması ile cümle bazında elde edilen sonuç sunulmuştur.

Tablo 5. Cümle Konumu (İlk) Yöntemi ile Cümle Derecelendirme Örnek Verisi

Haber No	Cümle Sırası	Cümle	Cümle Konumu (İlk) Puanı
215	0	Ali Haydar Paksoy …	1
215	1	Pist alanından…	0,7897
215	2	Paksoy ihbar…	0,6550
215	3	Evli ve 2 çocuk…	0,7837
215	4	Paksoyun…	0,8084
215	5	Öte yandan…	0,8045

4.7.5. Cümle konumu (son cümle) yöntemi

Son cümlenin önemli bilgi içerdiği varsayımı doğrultusunda bu cümlede geçen kelimelerin tüm metnin içerisindeki kelimelerin önemli olduğu varsayımı üzerinden çalışma yürütülmüştür. Bu cümledeki kelimeler üzerinde BERT modeli kullanılmış, cümlelerin puanlanmasında ilk cümle ile kosinüs benzerliği hesaplanmıştır. Tablo 6'da veri setinden bir örnekte cümle konumu (son cümle) yönteminin uygulanması ile cümle bazında elde edilen sonuç sunulmuştur.

Tablo 6. Cümle Konumu (Son) Yöntemi ile Cümle Derecelendirme Örnek Verisi

Haber No	Cümle Sırası	Cümle	Cümle Konumu (Son) Puanı
215	0	Ali Haydar…	1
215	1	Pist alanından…	0,7897
215	2	Paksoy ihbar…	0,6550
215	3	Evli ve 2 çocuk…	0,7837
215	4	Paksoyun…	0,8084
215	5	Öte yandan…	0,8045

4.7.6. Cümle uzunluğu yöntemi

Cümlelerdeki kelime sayısına bağlı olarak yapılan çalışmanın kelimenin uzunluğunun da bir etki olabileceği varsayımına dayanılarak harf sayısı yöntemi oluşturulmuştur. Yöntem kapsamında harfin fazla olmasının gereksiz kelime ve tekrarlayıcı ifadenin olmasının dezavantaj oluşturabileceği düşünülerek bir eşik değeri tespit edilmiştir.

Eşik değeri cümlelerdeki harf sayısının ortalama değeri üzerinden tespit edilmiştir. Bu eşik değerinin çok fazla üzerinde olan cümlelerin tekrar olması, eşik değerinin çok altında olan cümlelerin kısıtlı bilgi içeriyor olabileceği varsayımı üzerinden ilgili haber metni için istenilen özet oranında ortalama değere yakın doğrultudaki cümlelerin özet kısmında yer alması sağlanmıştır.

Tablo 7'de veri setinden bir örnekte cümle uzunluğu yönteminin uygulanması ile cümle bazında elde edilen sonuç sunulmuştur.

Tablo 7. Cümle Uzunluğu Yöntemi ile Cümle Derecelendirme Örnek Verisi

Haber No	Cümle Sırası	Cümle	Cümle Uzunluğu Puanı
215	0	Ali Haydar Paksoy 41…	0,0833
215	1	Pist alanından…	0,1666
215	2	Paksoy ihbar üzerine…	0,25
215	3	Evli ve 2 çocuk babası…	0,3333
215	4	Paksoyun paraşütünün…	0,4166
215	5	Öte yandan Paksoyun…	0,5

4.7.7. Adlandırılmış varlık yöntemi

Yöntem kapsamında haber metni içerisinde özel isim, kurum, kuruluş, tarih gibi varlıkların yoğun kullanımının haberi yansıtacak önemli cümle olduğu varsayımı kullanılmıştır.

Cümleler içerisinde belirtilen kategorik varlıklar üzerinde tespitler yapılmış ve durak kelime kapsamı dışında yoğunlukta olan cümlelerin ağırlığı artırılmış ve özet kısmında yer alması sağlanmıştır. Uygulanan yöntemde her bir cümle için belirtilen nitelikte adlandırılmış varlık sayısı bulunmuş ve o cümle için adlandırılmış varlık puanı olarak kayıt altına alınmıştır.

Tablo 8'de veri setinden bir örnekte adlandırılmış varlık yönteminin uygulanması ile cümle bazında elde edilen sonuç sunulmuştur.

Tablo 8. Adlandırılmış Varlık Yöntemi ile Cümle Derecelendirme Örnek Verisi

Haber No	Cümle Sırası	Cümle	Adlandırılmış Varlık Puanı
215	0	Ali Haydar Paksoy…	3
215	1	Pist alanından…	1
215	2	Paksoy ihbar…	1
215	3	Evli ve 2 çocuk…	2
215	4	Paksoyun…	1
215	5	Öte yandan…	1

4.7.8. Tekil-çoğul yöntemi

Haber metni içerisinde yer alan kelimelerin tekil çoğul ayrımı yapılarak anlam noktasında ne kadar geniş ölçüye ulaştığı kontrol edilmiştir. Kelime yapısal olarak tekil iken anlam olarak birden fazla unsuru temsil edebileceğinden daha güçlü bir kavram ortaya koymaktadır. Sürü, orman vb. gibi topluluk isimleri bu yapıya örnek gösterilebilir. Kelimelerin bu yapısı TRNLP kütüphanesi yardımıyla sorgulanmış ve her kelime için 0 ya da 1 olmak üzere bir değer üretilmiştir. Çoğul ve topluluk isimleri için 1 değeri üretilmiştir. Üretilen bu değerler habere ait her bir cümlenin puanı olmuştur.

Tablo 9'da veri setinden bir örnekte tekil-çoğul yönteminin uygulanması ile cümle bazında elde edilen sonuç sunulmuştur.

Tablo 9. Tekil-Çoğul Yöntemi ile Cümle Derecelendirme Örnek Verisi

Haber No	Cümle Sırası	Cümle	Tekil-Çoğul Puanı
215	0	Ali Haydar Paksoy 41…	2
215	1	Pist alanından havalanan…	0
215	2	Paksoy ihbar üzerine olay…	1
215	3	Evli ve 2 çocuk babası…	0
215	4	Paksoyun paraşütünün ters…	0
215	5	Öte yandan Paksoyun…	0

4.7.9. Büyük ünlü uyumu yöntemi

Türkçe dilinin önemli bir özelliği, pek çok dilden ayrıldığı ünlü uyumudur. Alfabede bulunan sekiz adet ünlü harfin kelimenin ilerleyen hecelerindeki yerleşiminde büyük ünlü uyumu kuralı aranmaktadır. Ünlü harflerin kalınlık-incelik bakımından benzeşmesi anlamına gelen bu kural Türkçe'de kelimeye gelecek ek yapısının şekillenmesi noktasında önem arz etmektedir. Yalın halde ya da ekler ile zenginleştirilmiş kelimelerin ilk hecede bulunan ünlü harfin taşıdığı kalınlık-incelik durumu sonraki hecelerde de aynı şekilde devam etmesi kuralıdır (Nakiboğlu, 2007). Bu kural için bazı istisnai durumlar da bulunmaktadır (Türk Dil Kurumu, 1932).

a) Türkçe olmasına karşın büyük ünlü uyumuna uymayan istisnai kelimeler vardır. (şişman, anne, elma, dahi, hangi, hani, inanmak …)
b) Alıntı kelimelerde bu kural aranmaz (pehlivan, selam, tiyatro …)

c) Bitişik halde yazılan birleşik kelimelerde bu kural aranmaz (bilgisayar, çekyat …)

d) Bazı ekler bu kurala uymaz (-gil, -leyin, -yor, -ki …)

Bunun gibi istisnai durumlar dışında bir kelimenin Türkçe olup olmadığının kontrolünde veya sonrasında gelecek ekin ses uyumu için bu kural uygulanır.

Tez çalışması kapsamında geleneksel yöntemlere ek olarak önerilen yöntemlerden ilki olan büyük ünlü uyumu ile her bir cümledeki kelimelerin bu kurala uyup uymadıkları kontrol edilmiştir. Kurala uyan kelimelerin sayısı o cümlenin puanını belirlenmiştir. Tablo 10'da veri setinden bir örnekte büyük ünlü uyumu yönteminin uygulanması ile cümle bazında elde edilen sonuç sunulmuştur.

Tablo 10. Büyük Ünlü Uyumu Yöntemi ile Cümle Derecelendirme Örnek Verisi

Haber No	Cümle Sırası	Cümle	Büyük Ünlü Uyumu Puanı
215	0	Ali Haydar Paksoy…	10
215	1	Pist alanından…	9
215	2	Paksoy ihbar üzerine…	8
215	3	Evli ve 2 çocuk…	12
215	4	Paksoyun…	8
215	5	Öte yandan…	7

4.7.10. Küçük ünlü uyumu yöntemi

Bir kelimenin öz Türkçe olup olmamasına dair istisnai durumlar dışında büyük ünlü uyumu sonrasında küçük ünlü uyumu kontrol edilmektedir. Düzlük-yuvarlaklık kuralı olarak da adlandırılan bu uyumda aranan kurallar şunlardır (Nakiboğlu, 2007) :

a) Sözcüğün herhangi bir hecesinde düz ünlü (a, e, ı, i) varsa sonraki hecelerinde de düz ünlü bulunmalıdır.

b) Sözcüğün herhangi bir hecesinde yuvarlak ünlü (o, ö, u, ü) varsa sonraki ilk hecede geniş düz (a, e) ya da dar yuvarlak (u, ü) bulunmalıdır.

Düz ünlü kuralı doğrultusunda bir düz ünlü sonrasında tüm hecelerde düz ünlü bulunması gerekmektedir. Ancak yuvarlak ünlü kuralında her hece kendinden bir sonraki ünlü ile karşılaştırılır. Bu kural için büyük ünlü uyumundaki gibi bazı istisnai durumlar bulunmaktadır. Bazı ekler küçük ünlü uyumunu bozmaktadır (-gil, -mtırak, -ki, -yor). Bazı kelimeler (kavuk, kavun, çamur …) bu kurala uymamaktadır (Türk Dil Kurumu, 1932)

Geleneksel yöntemlere ek olarak önerilen yöntemlerden ikincisi olan küçük ünlü uyumu ile her bir cümledeki kelimelerin bu kurala uyup uymadıkları kontrol edilmiştir. Kurala uyan kelimelerin sayısı o cümlenin puanını belirlenmiştir. Tablo 11'de veri setinden bir örnekte küçük ünlü uyumu yönteminin uygulanması ile cümle bazında elde edilen sonuç sunulmuştur.

Tablo 11. Küçük Ünlü Uyumu Yöntemi ile Cümle Derecelendirme Örnek Verisi

Haber No	Cümle Sırası	Cümle	Küçük Ünlü Uyumu Puanı
215	0	Ali Haydar Paksoy…	7
215	1	Pist alanından…	4
215	2	Paksoy ihbar üzerine…	9
215	3	Evli ve 2 çocuk…	8
215	4	Paksoyun…	4
215	5	Öte yandan…	3

4.7.11. Büyük ve küçük ünlü uyumu (hibrit) yöntemi

Büyük ünlü uyumu ve küçük ünlü uyumu kuralı birlikte uygulanarak cümledeki kelimelerin hem büyük hem de küçük ünlü uyumuna uyup uymadıkları kontrol edilmiştir. Her iki kurala uyan kelimelerin sayısı bu yöntem dâhilinde ilgili cümlenin puanı olmuştur.

Tablo 12'de veri setinden bir örnekte büyük ve küçük ünlü uyumu yönteminin uygulanması ile cümle bazında elde edilen sonuç sunulmuştur.

Tablo 12. Büyük ve Küçük Ünlü Uyumu (Hibrit) Yöntemi ile Cümle Derecelendirme Örnek Verisi

Haber No	Cümle Sırası	Cümle	Büyük ve Küçük Ünlü Uyumu Puanı
215	0	Ali…	6
215	1	Pist…	4
215	2	Paksoy…	6
215	3	Evli ve 2…	7
215	4	Paksoyun…	4
215	5	Öte…	2

4.8. Cümle Derecelendirmelerinin Birleştirilmesi

Literatürde mevcut bulunan sekiz yöntem ve tez kapsamında önerilen üç yöntem olmak üzere toplamda 11 yöntemin uygulanması ile elde edilen cümle puanlarına her bir yönteme dair veriler kendi içerisinde olmak üzere normalizasyon işlemi uygulanmıştır. MinMax Scaling yöntemi ile veriler 0 ile 1 arasında olmak üzere ölçeklendirilmiştir (Yavuz ve Deveci, 2012). Ölçeklendirme sonrasında haberlere ilişkin her bir cümleye ait üç farklı cümle puanı oluşturulmuştur. Bunlardan ilki tamamen literatürde bulunan yöntemler ile elde edilen

sekiz puanın aritmetik ortalama ile elde edilmesiyle, ikincisi yalnızca tez kapsamında önerilen üç yöntem ile elde edilen puanların aritmetik ortalama ile elde edilmesiyle ve üçüncüsü ise hem literatürdeki yöntemlerin hem de önerilen yöntemlerin birlikte değerlendirildiği ve aritmetik ortalama ile elde edildiği puanlama sistemidir.

4.9. Özet İçin Cümle Seçimleri

Üç farklı cümle derecelendirme puanı üzerinden haber metinlerinde özete dâhil edilecek cümleler belirlenmiştir. Burada öncelikle habere ait her cümlenin puanı üç farklı yöntem için en yüksekten en düşüğe doğru sıralanmıştır. Bu sıralama sonrasında haber içerisinde ilgili cümlenin yerinin önemli olduğu ve özet kısmında da bu sıraya göre tekrar düzenlenmesi gerektiğinden bunun için bir sıralama index bilgisi eklenmiştir. İkinci aşamada insanlar tarafından oluşturulan özetlerden elde edilen özet oranı ve cümle sayısı bilgisi alınmış ve ilgili haber için özette yer alacak cümle sayısı kadar özete aday cümle alınmıştır. Alınan cümleler daha sonra index bilgisi kullanılarak ana metindeki sıralamaya uygun olarak tekrar sıralanmıştır. Buradaki sıralamanın tekrar yapılmasının amacı haberdeki giriş, gelişme ve sonuç odağında belirli bir akışın özet kısmında da yer almasını sağlamaktır. Üç kategoride elde edilen özetler için sonrasında değerlendirme işlemleri uygulanmıştır.

4.10. Özetlerin Başarı Değerlendirmesi

Tez kapsamında yürütülen çıkarım tabanlı otomatik metin özetleme çalışması için literatürde bulunan sekiz yöntem ve önerilen üç farklı yöntemin çeşitli özet oranları kullanılarak ROUGE ve BLEU metrikleri ile değerlendirmeleri gerçekleştirilmiştir.

Terim frekansı yönteminde en büyük artış ivmesinin tüm metriklerde %20 özet oranı ile %30 özet oranı arasında olduğu gözlemlenmiştir. Bu orandan sonra özette yer alacak miktar artsa da haber metninde önemli sayılabilecek terim sayısı olay kapsamı dışına çıkmadığı için yüksek ivmeli bir artış

gerçekleşmemiştir. BLEU metriği ile elde edilen sonuçlar da %20 özet oranı ile %30 özet oranı arasında en yüksek ivmeli artış gözlemlenirken, bu artış hızı diğer oranlarda daha yavaş gerçekleşmiştir. Terim frekansı yöntemi ile oransal özet değerlendirmesine ilişkin sonuçlar Tablo 13'te sunulmuştur.

Tablo 13. Terim Frekansı Yöntemi ile Oransal Özet Değerlendirmesi

Özet Oranı	ROUGE 1-F	ROUGE 1-K	ROUGE 1-G	ROUGE 2-F	ROUGE 2-K	ROUGE 2-G	BLEU
20	%39,6	%99,31	%25,39	%38,37	%97,97	%24,53	%6
30	%54,61	%97,93	%38,72	%53,03	%95,85	%37,49	%17
40	%62,51	%95,97	%47,27	%60,53	%93,51	%45,65	%26
50	%68,87	%92,63	%55,93	%66,44	%89,77	%53,82	%36
60	**%73,8**	%87,83	%64,67	%70,91	%84,62	%62,03	**%46**

Başlık özelliğine temelde terim frekansı yöntemi ile paralellik göstermektedir. Dolayısıyla elde edilen sonuçların terim frekansı ile benzer olması beklenmektedir. Metni ifade etmek adına başlıkta özelleştirilmiş kelimeler bulunmaktadır. Bu kelimeler genellikle adlandırılmış varlıkları da içermektedir. Başlık yöntemi ile oransal özet değerlendirmesine ilişkin sonuçlar Tablo 14'te sunulmuştur.

Tablo 14. Başlık Yöntemi ile Oransal Özet Değerlendirmesi

Özet Oranı	ROUGE 1-F	ROUGE 1-K	ROUGE 1-G	ROUGE 2-F	ROUGE 2-K	ROUGE 2-G	BLEU
20	%39,65	%99,7	%25,41	%38,32	%98,11	%24,49	%06
30	%54,92	%99,51	%38,72	%53,42	%97,51	%37,57	%17
40	%63,35	%99,43	%47,26	%61,86	%97,6	%46,05	%25
50	%71,47	%99,36	%56,62	%70,11	%97,81	%55,45	%35
60	**%79,78**	%99,29	%67,28	%78,63	%98,05	%66,23	**%46**

Anahtar kelime yöntemi kelime frekansı ve başlık özelliği ile yapısal benzerlik içerdiğinden yakın sonuçlar elde edilmiştir. Anahtar kelime yöntemi ile oransal özet değerlendirmesine ilişkin sonuçlar Tablo 15'te sunulmuştur.

Tablo 15. Anahtar Kelime Yöntemi ile Oransal Özet Değerlendirmesi

Özet Oranı	ROUGE 1-F	ROUGE 1-K	ROUGE 1-G	ROUGE 2-F	ROUGE 2-K	ROUGE 2-G	BLEU
20	%38,86	%99,69	%24,79	%37,65	%98,38	%23,94	%6
30	%53,78	%99,46	%37,66	%52,52	%97,92	%36,69	%16
40	%62,07	%99,38	%45,91	%60,85	%97,99	%44,91	%23
50	%70,22	%99,32	%55,13	%69,13	%98,17	%54,18	%33
60	**%78,68**	%99,26	%65,81	%77,79	%98,35	%64,98	**%44**

Cümle konumu (ilk) özelliğinde her özet oranı için dengeli bir artış gözlemlenmiştir. İlk cümlede belgeyi yansıtacak bilgilerin kısıtlı olması sonuçlara yansımıştır. Cümle konumu (ilk) yöntemi ile oransal özet değerlendirmesine ilişkin sonuçlar Tablo 16'da sunulmuştur.

Tablo 16. Cümle Konumu (İlk) Yöntemi ile Oransal Özet Değerlendirmesi

Özet Oranı	ROUGE 1-F	ROUGE 1-K	ROUGE 1-G	ROUGE 2-F	ROUGE 2-K	ROUGE 2-G	BLEU
20	%37,14	%94,69	%23,81	%35,09	%91,36	%22,42	%05
30	%49,17	%88,45	%34,94	%46,05	%83,45	%32,64	%15
40	%56,95	%88,04	%43,01	%53,72	%83,43	%40,5	%23
50	%64,29	%87,66	%51,77	%61,16	%83,62	%49,18	%32
60	**%71,42**	%86,97	%61,4	%68,51	%83,56	%58,85	**%42**

Cümle konumu (son) yönteminde özellikle %20'lik oran için düşük bir oran gözlemlenmiştir. %30'luk orandan itibaren dengeli bir artış gözlemlenmiştir. Düşük orandaki özette son cümledeki kelimelerin bağlamsal olarak içeriği yansıtmadaki eksikliği bu dezavantajı oluşturmuştur. Cümle konumu (son) yöntemi ile oransal özet değerlendirmesine ilişkin sonuçlar Tablo 17'de sunulmuştur.

Tablo 17. Cümle Konumu (Son) Yöntemi ile Oransal Özet Değerlendirmesi

Özet Oranı	ROUGE 1-F	ROUGE 1-K	ROUGE 1-G	ROUGE 2-F	ROUGE 2-K	ROUGE 2-G	BLEU
20	%26,37	%83,54	%16,44	%23,78	%77,5	%14,80	%3
30	%52,27	%99,61	%36,35	%50,35	%96,7	%34,95	%14
40	%56,92	%99,57	%40,95	%54,99	%96,8	%39,49	%23
50	%61,45	%99,52	%45,84	%59,55	%96,93	%44,36	%34
60	**%66**	%99,48	%51,16	%64,16	%97,07	%49,69	**%46**

Cümle uzunluğu yöntemine ilişkin sonuçlar incelendiğinde özetleme oranı arttıkça dengeli bir artışın olduğu gözlemlenmiştir. Cümle uzunluğu yöntemi ile oransal özet değerlendirmesine ilişkin sonuçlar Tablo 18'de sunulmuştur.

Tablo 18. Cümle Uzunluğu Yöntemi ile Oransal Özet Değerlendirmesi

Özet Oranı	ROUGE 1-F	ROUGE 1-K	ROUGE 1-G	ROUGE 2-F	ROUGE 2-K	ROUGE 2-G	BLEU
20	%34,78	%99,9	%21,24	%29,13	%96,67	%17,29	%1
30	%48,17	%99,87	%32,03	%41,56	%95,25	%26,82	%5
40	%52,49	%99,87	%36,03	%45,73	%95,08	%30,46	%11
50	%60,37	%99,65	%43,78	%56,92	%94,52	%41,19	%31
60	**%74,87**	%99,85	%60,3	%68,09	%94,5	%53,6	**%41**

Adlandırılmış varlık yönteminde özet oranı arttıkça sonuçlarda dengeli bir artış gözlenmiştir. Haberlerde bu kelimeler fazla sayıda bulunduğundan, özet oranı arttıkça metrikler ile elde edilen değerler de artmaktadır. Adlandırılmış varlık yöntemi ile oransal özet değerlendirmesine ilişkin sonuçlar Tablo 19'da sunulmuştur.

Tablo 19. Adlandırılmış Varlık Yöntemi ile Oransal Özet Değerlendirmesi

Özet Oranı	ROUGE 1-F	ROUGE 1-K	ROUGE 1-G	ROUGE 2-F	ROUGE 2-K	ROUGE 2-G	BLEU
20	%44,60	%99,64	%29,44	%43,28	%98,10	%28,48	%1
30	%59,25	%99,43	%42,95	%57,65	%97,33	%41,70	%5
40	%67,09	%99,37	%51,34	%65,44	%97,35	%49,98	%11
50	%74,48	%99,31	%60,29	%72,90	%97,50	%58,93	%31
60	**%81,89**	%99,24	%70,19	%80,52	%97,75	%68,94	**%41**

Haberdeki kelimeler yapısal olarak tekil, anlamca birden fazla unsuru temsil edebilir. Özet oranlarına göre yapılan bu değerlendirmede sonuçların dengeli bir şekilde özet oranına bağlı olarak arttığı gözlemlenmiştir. Tekil-çoğul yöntemi ile oransal özet değerlendirmesine ilişkin sonuçlar Tablo 20'de sunulmuştur.

Tablo 20. Tekil-Çoğul Yöntemi ile Oransal Özet Değerlendirmesi

Özet Oranı	ROUGE 1-F	ROUGE 1-K	ROUGE 1-G	ROUGE 2-F	ROUGE 2-K	ROUGE 2-G	BLEU
20	%36,73	%99,63	%23,11	%35,54	%98,28	%22,3	%4
30	%51,32	%99,39	%35,38	%50,28	%98,22	%34,57	%13
40	%59,64	%99,33	%43,35	%58,71	%98,4	%42,58	%20
50	%67,92	%99,28	%52,43	%67,17	%98,6	%51,75	%30
60	**%76,82**	%99,24	%63,35	%76,23	%98,7	%62,77	**%41**

Büyük ünlü uyumu yöntemine ilişkin oransal özet çalışmasında sonuçların dengeli bir şekilde arttığı gözlemlenmiştir. Her iki metrik için de gözlemlenen bu durum ile elde edilen sonuçların diğer yöntemlerden daha başarılı olmuştur. Büyük ünlü uyumu yöntemi ile oransal özet değerlendirmesine ilişkin sonuçlar Tablo 21'de sunulmuştur.

Tablo 21. Büyük Ünlü Uyumu Yöntemi ile Oransal Özet Değerlendirmesi

Özet Oranı	ROUGE 1-F	ROUGE 1-K	ROUGE 1-G	ROUGE 2-F	ROUGE 2-K	ROUGE 2-G	BLEU
20	%51,62	%99,71	%35,39	%50,41	%98,45	%34,46	%14
30	%66,36	%99,55	%50,28	%64,85	%97,73	%49,04	%30
40	%73,92	%99,49	%59,28	%72,33	%97,65	%57,9	%40
50	%80,72	%99,42	%68,41	%79,14	%97,68	%66,99	%50
60	**%87,2**	%99,35	%77,96	%85,74	%97,8	%76,59	**%59**

Küçük ünlü uyumu da büyük ünlü uyumu özelliği ile yakın sonuçlara ulaşmıştır. Küçük ünlü uyumu yöntemi ile oransal özet değerlendirmesine ilişkin sonuçlar Tablo 22'de sunulmuştur.

Tablo 22. Küçük Ünlü Uyumu Yöntemi ile Oransal Özet Değerlendirmesi

Özet Oranı	ROUGE 1-F	ROUGE 1-K	ROUGE 1-G	ROUGE 2-F	ROUGE 2-K	ROUGE 2-G	BLEU
20	%51,64	%99,71	%35,41	%50,43	%98,46	%34,48	%14
30	%66,37	%99,54	%50,29	%64,87	%97,73	%49,06	%30
40	%73,92	%99,48	%59,28	%72,34	%97,66	%57,91	%40
50	%80,71	%99,42	%68,39	%79,13	%97,68	%66,97	%50
60	**%87,18**	%99,35	%77,93	%85,72	%97,79	%76,56	**%59**

Büyük ve küçük ünlü uyumu yöntemi ile oransal özet değerlendirmesine ilişkin sonuçlar Tablo 23'te sunulmuştur.

Tablo 23. Büyük ve Küçük Ünlü Uyumu Yöntemi ile Oransal Özet Değerlendirmesi

Özet Oranı	ROUGE 1-F	ROUGE 1-K	ROUGE 1-G	ROUGE 2-F	ROUGE 2-K	ROUGE 2-G	BLEU
20	%51,64	%99,71	%35,41	%50,43	%98,46	%34,48	%14
30	%66,37	%99,54	%50,29	%64,87	%97,73	%49,06	%30
40	%73,92	%99,48	%59,28	%72,34	%97,66	%57,91	%40
50	%80,71	%99,42	%68,39	%79,13	%97,68	%66,97	%50
60	**%87,18**	%99,35	%77,93		%97,79	%76,56	**%59**

Tüm yöntemlere ilişkin elde edilen sonuçların ortalama değerleri alınmış ve bu ortalama değerler üzerinden yöntemler arasında karşılaştırma yapılmıştır. Elde edilen ortalama veriler Tablo 24'te sunulmuştur.

Tablo 24. Oransal Özet Değerlendirmesine İlişkin Tüm Yöntemlerin Karşılaştırması

Yöntem	ROUGE 1-F	ROUGE 1-K	ROUGE 1-G	ROUGE 2-F	ROUGE 2-K	ROUGE 2-G	BLEU
Terim Frekansı	%59,88	%94,73	%46,39	%57,85	%92,34	%44,7	%26
Anahtar Kelime	%60,72	%99,42	%45,86	%59,59	%98,16	%44,94	%24
Başlık	%61,84	%99,46	%47,06	%60,47	%97,82	%45,96	%26

Cümle Konumu (İlk)	%55,79	%89,16	%42,99	%52,91	%85,08	%40,72	%23
Cümle Konumu (Son)	%52,60	%96,34	%38,15	%50,57	%93,00	%36,66	%24
Cümle Uzun.	%54,13	%99,82	%38,67	%48,28	%95,2	%33,87	%17
Adlan. Varlık	%65,46	%99,40	%50,84	%63,96	%97,61	%49,61	%30
Tekil-Çoğul	%58,49	%99,38	%43,52	%57,59	%98,44	%42,79	%22
Büyük Ünlü Uyumu	**%71,97**	**%99,50**	**%58,27**	**%70,49**	**%97,86**	**%57,00**	**%39**
Küçük Ünlü Uyumu	**%71,97**	**%99,50**	**%58,26**	**%70,50**	**%97,87**	**%57,00**	**%37**
Büyük-Küçük Ünlü Uyumu	**%71,53**	**%99,51**	**%57,74**	**%70,05**	**%97,86**	**%56,48**	**%38**

Büyük ünlü uyumu özelliği, küçük ünlü uyumu özelliği ve bu iki özelliğin hibrit bir modeli, diğer sekiz özellikten daha iyi performans göstermiştir. Bu sonuçlar hem ROUGE hem de BLEU metriklerinde gözlemlenmiştir. Adlandırılmış varlık özelliği, önerilen üç özelliğe en yakın sonuca ulaşmıştır.

Varlık olarak adlandırılan yer, zaman, kurum adı gibi unsurlar haber metinlerinde sıklıkla kullanıldığı için özet performansının yüksek olması doğaldır. Önerilen üç yöntemde ROUGE-1 metriğinde yaklaşık %71 f-skoru elde edilirken adlandırılmış varlık özelliğinde sonuç %65 olmuştur. Üç yöntem için de ROUGE-2 metriğinde yaklaşık %70 f-skoru elde edilirken, adlandırılmış varlıkta %68 f-skoru elde edilmiştir.

ROUGE için elde edilen bu sonuçlar BLEU metriği için de paralellik göstermektedir. Üç yöntem için %37 ile %39 arasında değerler gözlenirken en yakın sonuç adlandırılmış varlık yöntemi ile %30 sonucu elde edilmiştir.

Elde edilen üç farklı kategorideki özetler ROUGE ve BLEU metrikleri ile değerlendirilmiştir. ROUGE metriği ile yapılan f-skoru değerlendirme sonuçları Tablo 25'te sunulmuştur.

Tablo 25. ROUGE Metriği ile Değerlendirme Sonuçları

Yöntem	ROUGE 1-F	ROUGE 2-F
Literatürdeki 8 Yöntem	%58,61	%56,40
Önerilen 3 Yöntem	**%71,82**	**%70,34**
Tüm Yöntemlerin Birleşimi	%68,21	%79,07

Tablo 13'te sunulan veriler de önerilen üç yöntemin içinde en iyi sonucun yalnızca önerilen üç yöntemin kullandığı kısım olduğunu göstermektedir. Literatürde bulunan yöntemlerin başarım oranında önerilen yöntemlerin içerisine dâhil edildiği hibrit yöntem ile artış sağlanmıştır.

BLUE metriği ile yapılan değerlendirme sonuçları Tablo 26'da sunulmuştur.

Tablo 26. BLEU Metriği ile Değerlendirme Sonuçları

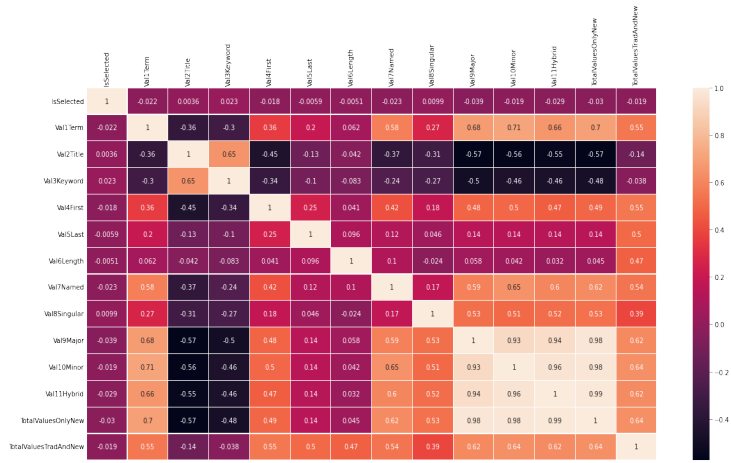
Yöntem	BLEU Metriği
Literatürdeki 8 Yöntem	%24
Önerilen 3 Yöntem	**%38**
Tüm Yöntemlerin Birleşimi	%32

BLUE metriği ile yapılan değerlendirme incelendiğinde ROUGE metriği ile elde edilen sonuçlara paralel olarak en iyi sonucun önerilen üç yöntem ile elde edildiği gözlemlenmektedir. Literatürdeki sekiz yöntemin ile elde edilen başarım oranı önerilen üç yöntemin de dâhil edildiği hibrit yöntem ile artırılmıştır.

4.11. Özetleme için Sınıflandırma Yöntemi

Özetleme çalışması önceki bölümde cümle bazında çeşitli yöntemlerle derecelendirme üzerinden yürütülmüştür. Bu çalışma bir sınıflandırma problemi olarak da ele alınabilmektedir. Bu kısımda sınıflandırma için bir cümlenin özette yer alması ya da almaması şeklinde ikili bir sınıflandırma üzerinden çalışma yürütülmüştür. Burada elde edilen nihai sonuç cümlenin özette yer alıp almamasını karar verici sınıflandırma algoritma yardımıyla sonucun sonrasında değerlendirmesi şeklinde ilerlemiştir.

Çalışma kapsamında Karar Ağacı Sınıflayıcısı, K-En Yakın Komşu Sınıflayıcısı, Naive Bayes Sınıflayıcısı ve Rastgele Orman Sınıflayıcısı olmak üzere dört farklı sınıflayıcı yöntem uygulanmıştır. Sınıflandırma işlemi için %75 eğitim, %25 oranında test verisi ayrımı yapılmıştır. Özellikler arası ilişkilerin gözlemlenmesi adına bir korelasyon grafiği oluşturulmuştur. Şekil 10'da sunulan bu grafikte terim frekansı, adlandırılmış varlık ve tekil çoğul yöntemlerinin önerilen yöntemlere ve sonuçlar ile daha yakın bir ilişkide olduğu gözlemlenmiştir.



Şekil 11. Korelasyon Grafiği

4.11.1. Karar ağacı sınıflayıcısı

Karar ağacı yapı olarak akış şemalarına benzemektedir. Bu yapı içerisindeki düğüm olarak nitelendirilen yapılar ilgili veriye ait öznitelik olarak adlandırılmaktadır (Sharma ve Kumar, 2016).

Olası sonuçlar ve bu sonuçların etkilerine göre oluşan karar yapısını ifade eden karar ağaçlarında düğüm, yaprak ve dal yapısından ötürü ağaç ismi verilmiştir. Örneğin bir iş başvurusu için kabul edilme ihtimaliniz ihtimallerin değerlendirilmesi, bu ihtimalleri oluşturan veriler bütününün değerlendirilerek sonuca ulaştırabilmektedir.

Karar Ağacı yönteminde cümle puanlama yönteminde olduğu gibi literatürdeki sekiz yöntem, önerilen üç yöntem ve tüm yöntemlerin bir arada kullanıldığı hibrit yöntem olmak üzere üç farklı kategoride değerlendirme yapılmıştır. Elde edilen sonuçlar Tablo 27'de sunulmuştur.

Tablo 27. Karar Ağacı Yöntemi Üzerinde Değerlendirme

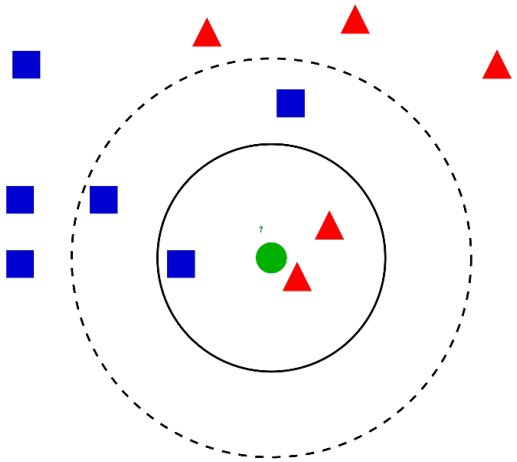
Yöntem	f1 Skoru	Kesinlik	Duyarlılık	Doğruluk
Literatürdeki 8 Yöntem	%48,85	%43,51	%35,95	%49,31
Önerilen 3 Yöntem	**%51,78**	**%46,93**	**%39,85**	**%51,21**
Tüm Yöntemlerin Birleşimi	%47,51	%41,86	%38,66	%47,65

Karar Ağacı yönteminin sonuçları değerlendirildiğinde önerilen üç yöntemin daha iyi bir başarı sağladığı gözlemlenirken, hibrit yöntemin cümle puanlama yöntemindeki gibi olumlu bir katkısı gözlemlenememiştir.

4.11.2. K-en yakın komşu sınıflayıcısı

K-En Yakın Komşu Sınıflayıcısı yönteminde verinin sınıfını belirlemek için ilgili verinin diğer verilere olan uzaklığı hesaplanır. Elde edilen bu uzaklık dâhilinde verinin en uygun sınıfı tahmin edilir (Cover ve Hart, 1967).

Yöntemde bulunan *k* komşuluk değeri optimal değerin bulunması için kullanılırken uzaklık yöntemi için genellikle Öklid Uzaklığı kullanılmaktadır. K-En Yakın Komşu Sınıflayıcısı'nda öncelikle bir k değeri belirlenir. Bu değer genellikle optimal değere ulaşana dek sürebilir. Optimal değer tespitinin ardından öznitelikler arası uzaklıklar hesaplanır. Elde edilen uzaklık bilgisi küçükten büyüğe sıralanır, en yakın uzaklığa bağlı olarak ilgili veri için komşuluklar belirlenir. Komşuluklar için uzaklık verisi en küçük olan 1. komşu, en büyük olan veri n. komşu olarak belirlenir. Ardından k-en yakın komşu sınıfları toplanırve en uygun komşu kategorisi seçilir (Martínez, Frías, Pérez-Godoy, ve Rivera, 2018). Şekil 12'de K-En Yakın Komşu Sınıflayıcısı'na ait örnek bir yapı sunulmuştur.



Şekil 12. K-En Yakın Komşu Sınıflayıcısı Yapısı (Analytics Vidhya, 2012)

K-En Yakın Komşu Sınıflayıcısı kullanılarak elde edilen sonuçlar Tablo 28'de sunulmuştur.

Tablo 28. K-En Yakın Komşu Sınıflayıcısı Yöntemi Üzerinde Değerlendirme

Yöntem	f1 Skoru	Kesinlik	Duyarlılık	Doğruluk
Literatürdeki 8 Yöntem	%48,85	%43,51	%35,95	%49,31
Önerilen 3 Yöntem	**%50,83**	**%45,79**	**%44,87**	**%50,92**
Tüm Yöntemlerin Birleşimi	%49,56	%44,51	%38,66	%47,65

K-En Yakın Komşu Sınıflayıcısı yönteminin sonuçları değerlendirildiğinde önerilen üç yöntemin daha iyi bir başarı sağladığı gözlemlenirken, Karar Ağacı yönteminin aksine hibrit yöntemin cümle puanlama yöntemindeki gibi olumlu bir katkısı gözlemlenmiştir.

4.11.3. Naive bayes sınıflayıcısı

Bu yöntem olasılık temellerini kullanan bir yaklaşımı temel almaktadır. Özniteliklerin istatistiki açıdan birbirinden bağımsız olduğu veri kümelerinde kullanılan yöntem Bayes teoremini kullanmaktadır. Bu teoremin denklemi Denklem 4.11.3.1'de sunulmuştur:

$$P(X \mid Y) = \frac{P(X)P(X)}{P(Y)} \qquad (4.11.3.1)$$

P(X | Y); Y olayı gerçekleştiğinde X olayının gerçekleşme olasılığı

P(Y | X); X olayı gerçekleştiğinde Y olayının gerçekleşme olasılığı

P(X), P(Y); X ve Y olaylarının gerçekleşme olasılıkları

Naive Bayes Sınıflayıcısı ile edilen sonuçlar Tablo 29'da sunulmuştur.

Tablo 29. Naive Bayes Sınıflayıcısı Yöntemi Üzerinde Değerlendirme

Yöntem	f1 Skoru	Kesinlik	Duyarlılık	Doğruluk
Literatürdeki Sekiz Yöntem	%45,92	%46,17	%38,76	%49,75
Önerilen Üç Yöntem	%46,40	%42,45	%51,07	%50,85
Tüm Yöntemlerin Birleşimi	**%46,60**	**%46,73**	**%59,20**	**%50,89**

Naive Bayes Sınıflayıcısı yönteminin sonuçları değerlendirildiğinde en iyi sonucun hibrit yöntem olduğu gözlemlenmiştir. Naive Bayes Sınıflayıcı temelinde önerilen üç yöntemin hibrit yöntemin ardından takip ettiği ve literatürdeki mevcut yöntemlerden daha iyi başarım sergilemesi, önerilen yöntemin salt olarak en iyi başarıyı sağlamasa da

mevcut yöntemlerin başarısını artırmada etkili olduğunu göstermektedir.

4.11.4. Rastgele orman sınıflayıcısı

Rastgele Orman Sınıflayıcısı bir veri kümesinde rastgele seçilen veriler ışığında oluşturulan karar ağaçlarında bir tahmin kümesi oluşturmayı sağlayan bir yöntemdir (Breiman, 2001). Rastgele Orman Sınıflayıcısı ile edilen sonuçlar Tablo 30'da sunulmuştur.

Tablo 30. Rastgele Orman Sınıflayıcısı Yöntemi Üzerinde Değerlendirme

Yöntem	f1 Skoru	Kesinlik	Duyarlılık	Doğruluk
Literatürdeki 8 Yöntem	%46,45	%41,11	%32,25	%44,20
Önerilen 3 Yöntem	**%49,84**	**%44,50**	**%40,90**	**%49,70**
Tüm Yöntemlerin Birleşimi	%47,51	%42,42	%31,92	%46,50

Rastgele Orman Sınıflayıcısı yönteminin sonuçları değerlendirildiğinde önerilen üç yöntemin daha iyi bir başarı sağladığı gözlemlenirken, Karar Ağacı yönteminin aksine hibrit yöntemin cümle puanlama yöntemindeki gibi olumlu katkısının olduğu gözlemlenmiştir.

Sınıflandırma için kullanılan dört farklı yöntem içerisinde Karar Ağacı ve K-En Yakın Komşu yönteminin en iyi başarımda bulunan yöntemler olduğu gözlemlenmektedir.

Tablo 31'de sunulan verilerden iki yöntemin her birinin iki kategoride en yüksek başarıma ulaştığı gözlemlenmiştir.

Tablo 31. Dört Sınıflayıcı için f1 Skoru Değerlendirmesi

Yöntem / f1 Skoru	Karar Ağacı	K En Yakın Komşu	Naive Bayes	Rastgele Orman
Literatürdeki 8 Yöntem	%48,85	%48,85	%45,92	%46,45
Önerilen 3 Yöntem	**%51,78**	**%50,83**	%46,40	**%49,84**
Tüm Yöntemlerin Birleşimi	%47,51	%49,56	**%46,60**	%47,51

BÖLÜM 5.

SONUÇ ve ÖNERİLER

Tez kapsamında Türkçe dilinde yazılmış 215 haberden oluşan bir veri seti üzerinde önce insanlar tarafından oluşturulan özetlerinin oluşması sağlanmıştır. Bunun için üç öğretmenden oluşan bir çalışma grubu oluşturulmuş, öğretmenlerin birbirlerinden bağımsız şekilde özetleme çalışmaları yapmaları sağlanmıştır. Her bir haber için cümlelere ayrılma işlemi yapılmış ve çalışma grubu üyelerinin özette yer alması gerektiği cümleleri yalnızca seçim yöntemiyle oluşturmaları istenmiştir. Elde edilecek özetlerin dijital ortamda alınması, rahatlıkla raporlanabilmesi ve işlenebilmesi için çalışma ASP. NET tabanlı bir uygulama üzerinde gerçekleştirilmiş ve yapılan işlemlerin tamamı SQL Server'da veritabanına kaydedilmiştir. Üç farklı kişiden alınan özetlerin istatistiki yöntemler ile (optimal değerin bulunması) tek bir özete düşürülmesi ardından özet oranları haber özelinde oluşturulmuştur. Özet oranları ana metin baz alınarak cümle sayısına göre oluşturulmuştur.

Haber metinleri için çıkarım tabanlı, tek belgeli ve genel özet yapısında özetleme yapılmıştır. Özetleme aşamasından önce haber metinleri içerisinde çeşitli analiz işlemleri gerçekleştirilmiş ve aykırı verilerin olup olmadığı tespit edilmiştir. Metin içerisinde ön işleme çalışmaları (durak kelimelerin kaldırılması, tüm harflerin küçük harfe dönüştürülmesi, noktalama işaretlerinin kaldırılması vb.) yapılması sonrasında Bu özetlemeler için literatürde hali hazırda

bulunan sekiz farklı yöntem ve tez çalışması kapsamında önerilen üç farklı yöntem olmak üzere toplamda 11 yöntem birbirlerinden bağımsız olarak uygulanmıştır.

Literatürde bulunan Terim Frekansı, Başlık, Anahtar Kelime, Cümle Konumu (İlk Cümle), Cümle Konumu (Son Cümle), Cümle Uzunluğu, Adlandırılmış Varlık, Tekil-Çoğul yöntemleri ve tez çalışması kapsamında önerilen Büyük Ünlü Uyumu, Küçük Ünlü Uyumu ile Büyük ve Küçük Ünlü Uyumu (Hibrit) Yöntemi uygulanmıştır. Uygulama sonrasında her bir cümle ilgili yöntem için bir puan değerine sahip olmuştur. Elde edilen cümle puanları için yöntem içlerindeki verilerde olmak üzere normalizasyon işlemi uygulanmıştır. MinMax Scaling yöntemi ile veriler 0 ile 1 arasında olmak üzere ölçeklendirilmiştir. Ölçeklendirme sonrasında haberlere ilişkin her bir cümleye ait literatürde bulunan yöntemler ile elde edilen sekiz puanın aritmetik ortalaması, yalnızca tez kapsamında önerilen üç yöntem ile elde edilen puanların aritmetik ortalaması ve hem literatürdeki yöntemlerin hem de önerilen yöntemlerin birlikte değerlendirildiği ve aritmetik ortalamasının hesaplandığı üç farklı puan elde edilmiştir.

Üç farklı cümle derecelendirme puanı üzerinden haber metinlerinde özete dâhil edilecek cümleler belirlenmiştir. Habere ait her cümlenin puanı üç farklı yöntem için en yüksekten en düşüğe doğru sıralanmıştır. Haberlerin akışının korunması için bu kısımda index bilgisi ile sıralama korunmuştur. Haber için yapılan kişi özetlerindeki cümle sayısı bilgisi alınmış ve ilgili haber için özette yer alacak cümle sayısı kadar özete aday cümle alınmıştır. Alınan cümleler daha sonra index bilgisi kullanılarak ana metindeki sıralamaya uygun olarak tekrar sıralanmıştır.

Üç kategoride elde edilen özetler için değerlendirme işlemi ROUGE ve BLEU metrikleri kullanılarak gerçekleştirilmiştir.

ROUGE metriğinde ROUGE-1, ROUGE-2 ve ROUGE-L alt metrikleri kullanılmış ve f-skoru üzerinden karşılaştırma yapılmıştır. Önerilen yöntemin ROUGE-1 için %93,24 değeri,

ROUGE-2 için %88,05 değeri, ROUGE-L için %93,24 f-skoru değeri ile en iyi başarıyı elde ettiği gözlemlenmiştir. Literatürde bulunan yöntemlerin başarım oranı önerilen yöntemlerin dâhil edildiği hibrit yöntem ile artmıştır. ROUGE-1 için %83,82'den %83,94'e, ROUGE-2 için %78,08'den %79,07'ye, ROUGE-L için %83,82'den %84,13'e artış sağlanmıştır.

BLEU metriği ile yapılan karşılaştırmada ROUGE metriğine paralel olarak %51 değeri ile en iyi başarıyı önerilen yöntemin elde ettiği gözlemlenmiştir. Literatürdeki sekiz yöntemin başarım oranı önerilen üç yöntemin de dâhil edildiği hibrit yöntem ile %48'den %50 değerine doğru artmıştır.

Özetleme çalışması bir sınıflandırma problemi olarak da ele alınabilmektedir. Bir cümlenin özette yer alması ya da almaması şeklinde ikili bir sınıflandırma üzerinden çalışma yürütülmüştür. Çalışma kapsamında Karar Ağacı Sınıflayıcısı, K-En Yakın Komşu Sınıflayıcısı, Naive Bayes Sınıflayıcısı ve Rastgele Orman Sınıflayıcısı olmak üzere dört farklı sınıflayıcı yöntem uygulanmıştır. Sınıflandırma işlemi için %75 eğitim, %25 oranında test verisi ayrımı yapılmıştır. Sınıflandırma için kullanılan dört farklı yöntem içerisinde Karar Ağacı ve K-En Yakın Komşu yöntemleri en iyi başarımı göstermiştir. İki yöntemin her birinin iki kategoride en yüksek başarıma ulaştığı gözlemlenmiştir.

Yöntem bazında değerlendirildiğinde literatürde mevcut bulunan yöntemler için en iyi başarımı %48,85 f1-skor değeri ile Karar Ağacı ve K-En Yakın Komşu Sınıflayıcısı'nın birlikte, önerilen üç yöntem için en iyi başarımı %51,78 f1-skor değeri ile Karar Ağacı Sınıflayıcısı'nın, hibrit yöntemde en iyi başarımı %49,56 f1-skor değeri ile K-En Yakın Komşu Sınıflayıcısı'nın elde ettiği gözlemlenmiştir.

Sınıflayıcı bazında değerlendirildiğinde Karar Ağacı Sınıflayıcısı için %51,78 f1-skor, K-En Yakın Komşu Sınıflayıcısı için %50,83 f1-skor, Rastgele Orman Sınıflayıcısı için %49,84 f1-skor değerleri ile önerilen üç yöntemde en iyi başarımın sağlandığı, Naive Bayes Sınıflayıcısı için %46,60

f1-skor değeri ile hibrit yöntemde en iyi sonucun elde edildiği gözlemlenmiştir.

Hem cümle derecelendirme, hem de sınıflandırma yöntemlerinin sonuçlarında önerilen yöntemin mevcut yöntemlere oranla daha fazla başarım sağladığı gözlemlenmiştir. Bu durum Türkçe dilinde kelimenin dilin öz yapısına uygunluğunun kontrolünün yapılırken kullanılması ve haber metinleri gibi her okuyucu kitlesinin kolayca anlayabileceği düzeyde basit ve öz kelimelerin kullanıldığını desteklemektedir.

Çalışmada büyük ünlü uyumu ve küçük ünlü uyumuna uymayan ancak adlandırılmış varlık olarak nitelendirilebilecek özel nitelikteki isimlerin anlam noktasındaki katkılarının ek özetleme yöntemleri ile sorgulanması ilerideki çalışmalar kapsamında değerlendirilebilir. Ayrıca veri setinin çeşitli konu kapsamında da uygulanması ve teknik tabirlerin önerilen yöntemler kapsamında bir sözlüğe dahil edilmesinin başarım oranının daha fazla artırabileceği öngörülmektedir.

KAYNAKLAR

Adalı, Ş. (2009). *An Integrated Archıtecture For Informatıon Extractıon From Documents In Turkısh* (Yayın no. 293696) [Yayınlanmış doktora tezi, İstanbul Teknik Üniversitesi]. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Aktaş, Ö., ve Çebi, Y. (2013). Rule-Based Sentence Detection Method (RBSDM) for Turkish. *International Journal of Language and Linguistics*, 1. doi:https://doi.org/10.11648/j.ijll.20130101.11

Alguliev, R. M., Aliguliyev, R. M., ve Isazade, N. R. (2013). Multiple documents summarization based on evolutionary optimization algorithm. *Expert Systems with Applications*, 1675–1689. doi:https://doi.org/10.1016/j.eswa.2012.09.014

Aliguliyev, R. M. (2010). Clustering Techniques And Discrete Particle Swarm Optimization Algorithm For Multi-Document Summarization. *Computational Intelligence*, 420-448. doi:https://doi.wiley.com/10.1111/j.1467-8640.2010.00365.x

Altıntop, T. (2015). *Genetik algoritma ile dilsel özetlerin çıkarılması* (Yayın no. 397050) [Yayınlanmış yüksek lisans tezi, Gazi Üniversitesi]. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Analytics Vidhya. (2012). *Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm (with implementation in Python & R).* Nisan 3, 2022 tarihinde Analytics Vidhya: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/ adresinden alındı

Arslan, A. A. (2011). *Türkçe Metinlerden Anlamsal Bilgi Çıkarımı İçin Bir Veri Madenciliği Uygulaması* (Yayın no. 301598) [Yayınlanmış yüksek lisans tezi, Başkent Üniversitesi]. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Attokurov, U. (2014). *Multi-Document Summarization Using Distortion-Rate Ratio* (Yayın no. 353807) [Yayınlanmış yüksek lisans tezi, İstanbul Teknik Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Aysu, E. (2018). *Big data storage and automated text summarization in turkish text* (Yayın no. 507693) [Yayınlanmış yüksek lisans tezi, Işık Üniversitesi]. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Azam, N., ve Yao, J. (2012). Comparison of term frequency and document frequency based feature selection metrics in text categorization. *Expert Systems with Applications*, 4760–4768. doi:https://doi.org/10.1016/j.eswa.2011.09.160

Babar, S. A., ve Patil, P. D. (2015). Improving Performance of Text Summarization. *International Conference on Information and Communication Technologies (ICICT 2014)* (s. 354-363). Procedia Computer Science. doi:https://doi:10.1016/j.procs.2015.02.031

Barrios, F., Lopez, F., Argerich, L., ve Wachenchauzer, R. (2015). Variations of the Similarity Function of TextRank for Automated Summarization. Simposio Argentino de Inteligencia Artificial. Nisan 3, 2022 tarihinde https://44jaiio.sadio.org.ar/sites/default/files/asai65-72.pdf adresinden alındı

Baxendale, P. B. (1958). Machine-Made Index for Technical Literature—An Experiment. *IBM Journal of Research and Development*, 354-361. doi:https://doi.org/10.1147/rd.24.0354

Baydar, E. (2018). *Genetik algoritma kullanarak cümle seçme yaklaşımı ile otomatik metin özetleme* (Yayın no. 509691) [Yayınlanmış yüksek lisans tezi, Van Yüzüncü Yıl Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Birant, Ç. C. (2015). *Rule-based text summarization in Turkish* (Yayın no. 410518) [Yayınlanmış doktora tezi, Dokuz Eylül Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Bird, S., ve Loper, E. (2004). NLTK: The Natural Language Toolkit. *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 214–217. Nisan 3, 2022 tarihinde https://aclanthology.org/P04-3031 adresinden alındı

Boudin, F., Huet, S., ve Torres-Moreno, J. M. (2011). A Graph-based Approach to Cross-language Multi-document Summarization. *Polibits*, 113-118. doi:https://doi.org/10.17562/PB-43-16

Breiman, L. (2001). Random Forests. *Machine Learning*, 5-32. doi:https://doi.org/10.1023/A:1010933404324

Cao, Z., Wei, F., Li, S., Li, W., Zhou, M., ve Wang, H. (2015). Learning Summary Prior Representation for Extractive Summarization. *ACL-IJCNLP 2015* (s. 829–833). Beijing, Çin: Association for Computational Linguistics. doi:https://doi.org/10.3115/v1/P15-2136

Carbonell, J., ve Goldstein, J. (1998). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development*, (s. 335–336). doi:https://doi.org/10.1145/290941.291025

Cheng, J., ve Lapata, M. (2016). Neural Summarization by Extracting Sentences and Words. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (s. 484-494). Berlin, Almanya: Association for Computational Linguistics. doi:http://dx.doi.org/10.18653/v1/P16-1046

Choi, M., Kim, H., ve Croft, W. (2012). Dependency trigram model for social relation extraction from news articles. *SIGIR '12: Proceedings of the 35th international ACM*

SIGIR conference on Research and development in information retrieval, (s. 1047–1048). doi:https://doi.org/10.1145/2348283.2348462

Cover, T., ve Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 21 - 27. doi:https://doi.org/10.1109/TIT.1967.1053964

Culotta, A., ve Sorensen, J. (2004). Dependency Tree Kernels for Relation Extraction. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, (s. 423–429). doi:https://doi.org/10.3115/1218955.1219009

Çakır, M., ve Çelebi, E. (2011). Kapsama katsayisi tabanli kümeleme ile belge özetleme. *2011 IEEE 19th Signal Processing and Communications Applications Conference (SIU 2011)*, (s. 186-189). doi:https://doi.org/10.1109/SIU.2011.5929618

Çelikyılmaz, A., ve Hakkani-Tur, D. (2010). A Hybrid Hierarchical Model for Multi-Document Summarization. *ACL '10: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, (s. 815–824). doi:https://doi.org/10.5555/1858681.1858765

Çetiner, B. (2002). *Teknik metinleri otomatik olarak özetleyen bir paket program geliştirilmesi* (Yayın no. 121125) [Yayınlanmış yüksek lisans tezi, İstanbul Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Doğan, E. (2019). *Derin öğrenme yöntemiyle çevrimiçi sosyal ağlarda duygu analizi ve metin özetleme* (Yayın no. 566592) [Yayınlanmış yüksek lisans tezi, Fırat Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Dudak, E. (2020). *Gri kurt optimizasyon algoritması ile çıkarımsal metin özetleme ve özetlerin derin öğrenme ile sınıflandırılması* (Yayın no. 634927) [Yayınlanmış yüksek

lisans tezi, Düzce Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Durmaz, O. (2011). *Metin sınıflandırmada boyut azaltmanın etkisi ve özellik seçimi* (Yayın no. 312793) [Yayınlanmış yüksek lisans tezi, Gazi Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Edmundson, H. P. (1969). New Methods in Automatic Extracting. *Journal of the ACM*, 264–285. doi:https://doi.org/10.1145/321510.321519

Erhandı, B. (2020). Derin öğrenme ile metin özetleme (Yayın no. 638659) [Yayınlanmış yüksek lisans tezi, Sakarya Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Erkan, G., ve Radev, D. R. (2004). LexRank: Graph-based Lexical CentralityasSalience in TextSummarization. *Journal of Artificial Intelligence Research*, 457-479. doi:https://doi.org/10.1613/jair.1523

Fang, C., Mu, D., Deng, Z., ve Wu, Z. (2017). Word-sentence co-ranking for automatic extractive text summarization. *Expert Systems with Applications*, 189-195. doi:https://doi.org/10.1016/j.eswa.2016.12.021

Fırat Üniversitesi - Büyük Veri ve Yapay Zeka Labo. (2015). *Büyük Veri ve Yapay Zeka Laboratuvarı*. Nisan 3, 2022 tarihinde http://buyukveri.firat.edu.tr/veri-setleri/ adresinden alındı

Fukumoto, F., ve Suzuki, Y. (2000). Extracting Key Paragraph based on Topic and Event Detection Towards Multi-Document Summarization. *NAACL-ANLP 2000 Workshop: Automatic Summarization.* Nisan 3, 2022 tarihinde https://aclanthology.org/W00-04 adresinden alındı

Galanis, D., Lampouras, G., ve Androutsopoulos, I. (2012). Extractive Multi-Document Summarization with Integer Linear Programming and Support Vector Regression. *COLING 2012* (s. 911–926). The COLING 2012 Organizing

Committee. Nisan 3, 2022 tarihinde https://www.aclweb.org/anthology/C12-1056 adresinden alındı

Github - TRNLP. (2008, Şubat 8). *TRNLP.* Nisan 3, 2022 tarihinde https://github.com/brolin59/trnlp adresinden alındı

Goularte, F. B., Nassar, S. M., Fileto, R., ve Saggion, H. (2019). A text summarization method based on fuzzy rules and applicable to automated assessment. *Expert Systems with Applications*, 264-275. doi:https://doi.org/10.1016/j.eswa.2018.07.047

Graff, D., ve Cieri, C. (2003). English Gigaword. *Linguistic Data Consortium*. doi:https://doi.org/10.35111/0z6y-q265

Gündoğdu, Ö. E., ve Duru, N. (2016). Türkçe Metin Özetlemede Kullanılan Yöntemler., (s. 8). Nisan 3, 2022 tarihinde https://ab.org.tr/ab16/bildiri/36.pdf adresinden alındı

Güran, A., Arslan, S. N., Kılıç, E., ve Diri, B. (2014). Sentence selection methods for text summarization. *2014 22nd Signal Processing and Communications Applications Conference (SIU).* doi:https://doi.org/10.1109/SIU.2014.6830198

Güran, A., Güler, N. B., ve Bekar, E. (2011). Automatic summarization of Turkish documents using non-negative matrix factorization. *2011 International Symposium on Innovations in Intelligent Systems and Applications.* doi:https://doi.org/10.1109/INISTA.2011.5946121

Hariharan, S., Ramkumar, T., ve Srinivasan, R. (2013). Enhanced Graph Based Approach for Multi. *The International Arab Journal of Information Technology*, 334-341. Nisan 3, 2022 tarihinde https://iajit.org/portal/PDF/vol.10,no.4/4460.pdf adresinden alındı

Hark, C. (2020). *Metin çizgelerinde entropi ve optimizasyon tabanlı çıkarımsal metin özetleme* (Yayın no. 619018) [Yayınlanmış doktora tezi, İnönü Üniversitesi]. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Hark, C., Seyyarer, A., Uçkan, T., Myo, B., ve Karci, A. (2017). Doğal Dil İşleme Yaklaşımları ile Yapısal Olmayan Dökümanların Benzerliği. 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). doi:https://doi.org/10.1109/IDAP.2017.8090306

Hark, C., Uckan, T., Seyyarer, E., ve Karci, A. (2019). Extractive Text Summarization via Graph Entropy Çizge Entropi ile Çıkarıcı Metin Özetleme. *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, (s. 1-5). doi:https://doi.org/10.1109/IDAP.2019.8875936

Hark, C., Uçkan, T., Seyyarer, E., ve Karci, A. (2019). Metin Özetlemesi için Düğüm Merkezliklerine Dayalı Denetimsiz Bir Yaklaşım. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, 1109-1118. doi:https://doi.org/10.17798/bitlisfen.568883

Hatipoğlu, A., ve Omurca, S. İ. (2015). Türkçe Metin Özetlemede Melez Modelleme. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, 95-108. Nisan 3, 2022 tarihinde https://dergipark.org.tr/tr/pub/deumffmd/issue/40786/492043 adresinden alındı

Hovy, E., ve Lin, C.-Y. (1996). Automated text summarization and the SUMMARIST system. *Association for Computational Linguistics*, (s. 197). doi:https://doi.org/10.3115/1119089.1119121

Hu, Y.-H., Chen, Y.-L., ve Chou, H.-L. (2017). Opinion mining from online hotel reviews – A text summarization approach. *Information Processing & Management*, 436-449. doi:https://doi.org/10.1016/j.ipm.2016.12.002

Huang, L., He, Y., Wei, F., ve Li, W. (2010). Modeling Document Summarization as Multi-objective Optimization. *2010 Third International Symposium on Intelligent Information Technology and Security Informatics (IITSI)*, (s. 382-386). doi:https://doi.org/10.1109/IITSI.2010.80

Jaruskulchai, C., ve Kruengkrai, C. (2003). A Practical Text Summarizer by Paragraph Extraction for Thai. *Proceedings*

of the Sixth International Workshop on Information Retrieval with Asian Languages, (s. 9–16). doi:https://doi.org/10.3115/1118935.1118937

Jones, K. S. (1972). A Statistical Interpretation Of Term Specificity And Its Application In Retrieval. *Journal of Documentation*, 11-21. doi:https://doi.org/10.1108/eb026526

Joshi, A., Fidalgo, E., Alegre, E., ve Fernández-Robles, L. (2019). SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders. *Expert Systems with Applications*, 200-215. doi:https://doi.org/10.1016/j.eswa.2019.03.045

Karakoç, E., ve Yılmaz, B. (2019). Deep Learning Based Abstractive Turkish News Summarization., (s. 1-4). doi:https://doi.org/10.1109/SIU.2019.8806510

Kaynar, O., Görmez, Y., Işık, Y. E., ve Demirkoparan, F. (2017). Comparison of graph based document summarization method. *2017 International Conference on Computer Science and Engineering (UBMK)*, (s. 598-603). doi:https://doi.org/10.1109/UBMK.2017.8093475

Kaynar, O., Işık, Y. E., Görmez, Y., ve Kuş, E. (2018). Mobil application for automatic document summarization systems. *2018 26th Signal Processing and Communications Applications Conference (SIU)*, (s. 1-4). doi:https://doi.org/10.1109/SIU.2018.8404703

Kiabod, M., Dehkordi, M., ve Sharafi, M. (2012). A Novel Method of Significant Words Identification in Text Summarization. *Journal of Emerging Technologies in Web Intelligence*. doi:https://doi.org/10.4304/jetwi.4.3.252-258

Kısayol, A. İ. (2018). *Dokümanları çıkarımsal özetlemek için paragrafları öneme göre sıralama* (Yayın no. 527253) [Yayınlanmış yüksek lisans tezi, İstanbul Ticaret Üniversitesi]. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Kogilavani, A., ve Balasubramani, P. (2010). Clustering and Feature Specific Sentence Extraction Based Summarization of Multiple Documents. *International journal of computer science & information Technology*, 99-111. doi:https://doi.org/10.5121/ijcsit.2010.2409

Koupaee, M., ve Wang, W. Y. (2018). WikiHow: A Large Scale Text Summarization Dataset. Nisan 3, 2022 tarihinde https://arxiv.org/pdf/1810.09305.pdf adresinden alındı

Kumbhar, S., Bordia, A., Kapse, S., ve Paatil, F. (2018). Comparison of Extractive Text Summarization Methods. *International Journal of Advances in Electronics and Computer Science*. Nisan 3, 2022 tarihinde http://www.iraj.in/journal/journal_file/journal_pdf/12-485-153560965960-62.pdf adresinden alındı

Lamsiyah, S., El Mahdaouy, A., El Alaoui, S., Espinasse, B., ve Ezziyyani, M. (2020). A Supervised Method for Extractive Single Document Summarization Based on Sentence Embeddings and Neural Networks. *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019)* (s. 75-88). Springer International Publishing. doi:https://doi.org/10.1007/978-3-030-36674-2_8

Lin, C. (2004). ROUGE: A Package for Automatic Evaluation of summaries. Nisan 3, 2022 tarihinde https://aclanthology.org/W04-1013.pdf adresinden alındı

Lin, J., Sun, X., Ma, S., ve Su, Q. (2018). Global Encoding for Abstractive Summarization. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (s. 163–169). Association for Computational Linguistics. doi:https://doi.org/10.18653/v1/P18-2027

Lloret, E., ve Palomar, M. (2012). Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 1-41. doi:https://doi.org/10.1007/s10462-011-9216-z

Luhn, H. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 159-165. doi:https://doi.org/10.1147/rd.22.0159

Martínez, F., Frías, M., Pérez-Godoy, M., ve Rivera, A. (2018). Dealing with seasonality by narrowing the training set in time series forecasting with kNN. *Expert Systems with Applications*, 38-48. doi:https://doi.org/10.1016/j.eswa.2018.03.005

Medelyan, O. (2007). Computing Lexical Chains with Graph Clustering. *Proceedings of the ACL 2007 Student Research Workshop* (s. 85–90). Association for Computational Linguistics. Nisan 3, 2022 tarihinde https://aclanthology.org/P07-3015.pdf adresinden alındı

Meena, Y., ve Gopalani, D. (2015). Evolutionary Algorithms for Extractive Automatic Text Summarization. *Procedia Computer Science*, 244-249. doi:https://doi.org/10.1016/j.procs.2015.04.177

Mihalcea, R., ve Tarau, P. (2004). TextRank: Bringing Order into Text. (s. 404–411). Association for Computational Linguistics. Nisan 3, 2022 tarihinde https://aclanthology.org/W04-3252.pdf adresinden alındı

Mikolov, T., Chen, K., Corrado, G., ve Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Nisan 3, 2022 tarihinde https://dblp.org/rec/journals/corr/abs-1301-3781.html adresinden alındı

Mirshojaei, S., ve Masoomi, B. (2015). Text Summarization Using Cuckoo Search Optimization Algorithm. *Journal of Computer & Robotics*, 6. Nisan 3, 2022 tarihinde https://jcr.qazvin.iau.ir/article_683_e08cfc39b8a850adf76246e0096d3d22.pdf adresinden alındı

ML Resources. (2010). *BBC Datasets*. Nisan 3, 2022 tarihinde http://mlg.ucd.ie/datasets/bbc.html adresinden alındı

Mutlu, B. (2020). *Hibrit Zeki Sistem İle Metin Özetleme* (Yayın no. 633762) [Yayınlanmış doktora tezi, Gazi Üniversitesi]. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Mutlu, B., Sezer, E., ve Akcayol, M. (2019). Multi-document extractive text summarization: A comparative assessment on features. *Knowledge-Based Systems*, 104848. doi:https://doi.org/10.1016/j.knosys.2019.07.019

Nakiboğlu, S. (2007). Türkçede Ünlü Uyumu Ve Türkiye Türkçesi Ağızlarında Ünlü Uyumunun Bozulması. *Vowel Harmony In Turkish And The Degeneration Of Vowel Harmony In The Dialects Of In Turkey Turkish* (Cilt 171, s. 95-106). içinde

Nallapati, R., Zhou, B., Santos, C., Gulcehre, C., ve Xiang, B. (2016). Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning* (s. 280–290). Berlin, Almanya: Association for Computational Linguistics. doi:http://dx.doi.org/10.18653/v1/K16-1028

Narayan, S., Cohen, S., ve Lapata, M. (2018). Ranking Sentences for Extractive Summarization with Reinforcement Learning. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1747–1759. doi:http://dx.doi.org/10.18653/v1/N18-1158

Nuzumlalı, M., ve Özgür, A. (2014). Analyzing Stemming Approaches for Turkish Multi-Document Summarization. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (s. 702-706). Association for Computational Linguistics. doi:https://doi.org/10.3115/v1/D14-1077

Oliveira, H., Ferreira, R., Lima, R., Lins, R. D., Freitas, F., Riss, M., ve Simske, S. J. (2016). Assessing shallow

sentence scoring techniques and combinations for single and multi-document summarization. *Expert Systems with Applications: An International Journal*, 68-86. doi:https://doi.org/10.1016/j.eswa.2016.08.030

Orrú, T., Rosa, J. L., ve Netto, M. L. (2006). SABio: An Automatic Portuguese Text Summarizer Through Artificial Neural Networks in a More Biologically Plausible Model. *Computational Processing of the Portuguese Language* (Cilt 3960, s. 11-20). içinde Springer Berlin Heidelberg. doi:https://doi.org/10.1007/11751984_2

Önal, I. O. (2019). Collocation in Turkish linguistics: problems of definition and main areas of research. *Liberal Arts in Russia, 8*, 191-202. doi:https://doi.org/10.15643/libartrus-2019.3.3

Özateş, Ş. B., Radev, D. R., ve Özgür, A. (2016). Sentence Similarity based on Dependency Tree Kernels for Multi-document Summarization. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (s. 2833–2838). European Language Resources Association (ELRA). Nisan 3, 2022 tarihinde https://aclanthology.org/L16-1452.pdf adresinden alındı

Özkan, C. (2019). *İnternet tabanlı Türkçe metinler için otomatik özetleme tekniği* (Yayın no. 611878) [Yayınlanmış yüksek lisans tezi, Maltepe Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Page, L., ve Nicholson, J. (1998). The PageRank Citation Ranking: Bringing Order to the Web. *Stanford Digital Library Technologies Project*. Nisan 3, 2022 tarihinde http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf adresinden alındı

Papineni, K., Roukos, S., Ward, T., ve Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. *ACL 2002* (s. 311–318). Association for Computational Linguistics. doi:https://doi.org/10.3115/1073083.1073135

Pembe, F. C. (2010). *Automated query-biased and structure-preserving document summarization for web search tasks* (Yayın no. 270506) [Yayınlanmış doktora tezi, Boğaziçi Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Pennington, J., Socher, R., ve Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (s. 1532–1543). Association for Computational Linguistics. doi:https://doi.org/10.3115/v1/D14-1162

Plaza, L., Stevenson, M., ve Díaz, A. (2012). Resolving ambiguity in biomedical text to improve summarization. *Information Processing & Management*, 755-766. doi:https://doi.org/10.1016/j.ipm.2011.09.005

Rautray, R., ve Balabantaray, R. C. (2017). Cat swarm optimization based evolutionary framework for multi document summarization. *Physica A: Statistical Mechanics and its Applications*, 174-186. doi:https://doi.org/10.1016/j.physa.2017.02.056

Rautray, R., Balabantaray, R., ve Bhardwaj, A. (2015). Document Summarization Using Sentence Features. *International Journal of Information Retrieval Research (IJIRR)*, 36-47. Nisan 3, 2022 tarihinde https://ideas.repec.org/a/igg/jirr00/v5y2015i1p36-47.html adresinden alındı

Rush, A. M., Chopra, S., ve Weston, J. (2015). A Neural Attention Model for Abstractive Sentence Summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (s. 379–389). Association for Computational Linguistics. doi:https://doi.org/10.18653/v1/D15-1044

Saleh, H., Kadhim, N., ve Attea, B. (2015). Genetic Based Optimization Models for Enhancing Multi- Document Text Summarization. *Engineering and Technology Journal, 33*(8), 1489-1498.

Salton, G., Singhal, A., Buckley, C., Mitra, M., ve Mitra, A. (1996). Automatic Text Decomposition Using Text Segments and Text Themes. *HYPERTEXT '96: Proceedings of the the seventh ACM conference on Hypertext*, (s. 53–65). doi:https://doi.org/10.1145/234828.234834

Sami, M. V., ve Diri, B. (2010). Web Tabanlı Otomatik Özet Çıkarma Sistemi. *Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu*, (s. 4). Kayseri. Nisan 3, 2022 tarihinde http://www.kemik.yildiz.edu.tr/data/File/publications/Text%20Summarization/HTML%20Dokumanlarin%20Otomatik%20Ozetlenmesii.pdf adresinden alındı

Sanchez-Gomez, J., Vega-Rodríguez, M., ve Pérez, C. (2018). Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach. *Knowledge-Based Systems*, 1-8. doi:https://doi.org/10.1016/j.knosys.2017.11.029

Shardan, R., ve Kulkarni, U. (2010). Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization. *Journal of Computer Science*, 1366-1376. Nisan 3, 2022 tarihinde https://thescipub.com/pdf/10.3844/jcssp.2010.1366.1376 adresinden alındı

Sharma, H., ve Kumar, S. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. *International Journal of Science and Research (IJSR)*, 1-4. doi:https://doi.org/10.21275/v5i4.nov162954

Sinha, A., Yadav, A., ve Gahlot, A. (2018). Extractive Text Summarization using Neural Networks. Nisan 3, 2022 tarihinde https://arxiv.org/ftp/arxiv/papers/1802/1802.10137.pdf adresinden alındı

Song, S., Huang, H., ve Ruan, T. (2019). Abstractive text summarization using LSTM-CNN based deep learning. *Multimedia Tools and Applications*, 857-875. doi:https://doi.org/10.1007/s11042-018-5749-3

Tatar, S. (2011). *Automating information extraction task for Turkish texts* (Yayın no. 277012) [Yayınlanmış doktora tezi, İhsan Doğramacı Bilkent Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Torun, H., ve İnner, A. B. (2018). Detecting similar news by summarizing Turkish news. *2018 26th Signal Processing and Communications Applications Conference (SIU)*, (s. 1-4). doi:https://doi.org/10.1109/SIU.2018.8404826

Tozyılmaz, B. (2019). *Abstractive text summarization on WikiHow dataset using sentence embeddings* (Yayın no. 583291) [Yayınlanmış yüksek lisans tezi, Orta Doğu Teknik Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Turan, M. (2015). *Özgün paragraf tabanlı çıkarım tekniği kullanarak otomatik çoklu doküman özetleme* (Yayın no. 442543) [Yayınlanmış doktora tezi, Yıldız Teknik Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Tutkan, M., Ganiz, M. C., ve Akyokuş, S. (2014). Metin Sınıflandırma için Eğitimsiz bir Anlamsal Özellik Seçimi Yöntemi An Unsupervised Semantic Attribute Selection Method for Text Classification. *Bilgisayar ve Biyomedikal Mühendisliği Sempozyumu* (s. 1-4). Eleco 2014 Elektrik. Nisan 3, 2022 tarihinde https://www.emo.org.tr/ekler/71cef18f4837c7a_ek.pdf adresinden alındı

Tülek, M. (2007). *Türkçe için Metin Özetleme* (Yayın no. 222007) [Yayınlanmış yüksek lisans tezi, İstanbul Teknik Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Türk Dil Kurumu. (1932). *Büyük Ünlü Uyumu*. Nisan 3, 2022 tarihinde Türk Dil Kurumu: https://www.tdk.gov.tr/icerik/yazim-kurallari/buyuk-unlu-uyumu/ adresinden alındı

Türk Dil Kurumu. (1932). *Küçük Ünlü Uyumu*. Nisan 3, 2022 tarihinde Türk Dil Kurumu: https://www.tdk.gov.tr/icerik/yazim-kurallari/kucuk-unlu-uyumu/ adresinden alındı

Uçkan, T. (2020). *Metin çizgelerinde bağımsız kümelere dayalı çıkarımsal metin özetleme* (Yayın no. 619020) [Yayınlanmış doktora tezi, İnönü Üniversitesi] https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Umam, K., Putro, F., Pratamasunu, G., Arifin, A., ve Purwitasari, D. (2015). Coverage, Dıversıty, And Coherence Optımızatıon For Multı-Document Summarızatıon. *Jurnal Ilmu Komputer dan Informasi, 8*. doi:https://doi.org/10.21609/jiki.v8i1.278

Uzundere, E., ve Dedja, E. (2008). Türkçe Haber Metinleri İçin Otomatik Özetleme. *Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu*, (s. 1-4). Isparta. Nisan 3, 2022 tarihinde https://avesis.yildiz.edu.tr/yayin/46943024-5b6a-4073-ae2d-ec5702d4641b/turkce-haber-metinleri-icin-otomatik-ozetleme adresinden alındı

Wan, X., ve Yang, J. (2008). Multi-document summarization using cluster-based link analysis. *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, (s. 299–306). doi:https://doi.org/10.1145/1390334.1390386

Wang, D., Zhu, S., Li, T., Chi, Y., ve Gong, Y. (2011). Integrating Document Clustering and Multidocument Summarization. *ACM Transactions on Knowledge Discovery from Data*, 1-26. doi:https://doi.org/10.1145/1993077.1993078

Xie, Y., Zhou, T., Mao, Y., ve Chen, W. (2020). Conditional Self-Attention for Query-based Summarization. Nisan 3, 2022 tarihinde https://arxiv.org/pdf/2002.07338.pdf adresinden alındı

Xu, Y., ve Lapata, M. (2020). Coarse-to-Fine Query Focused Multi-Document Summarization. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), 3632–3645. doi:http://dx.doi.org/10.18653/v1/2020.emnlp-main.296

Yang, M., Wang, X., Lu, Y., Lv, J., Shen, Y., ve Li, C. (2020). Plausibility-promoting generative adversarial network for abstractive text summarization with multi-task constraint. *Information Sciences*, 46-61. doi:https://doi.org/10.1016/j.ins.2020.02.040

Yavuz, S., ve Deveci, M. (2012). İstatiksel Normalizasyon Tekniklerinin Yapay Sinir Ağın Performansına Etkisi. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 22. Nisan 3, 2022 tarihinde http://iibf.erciyes.edu.tr/dergi/sayi40/ERUJFEAS_Jun2012_167to187.pdf adresinden alındı

Yujian, L., ve Bo, L. (2007). A Normalized Levenshtein Distance Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1091-1095. doi:https://doi.org/10.1109/TPAMI.2007.1078

Zhao, X., ve Tang, J. (2010). Query-focused Summarization Based on Genetic Algorithm. *2010 International Conference on Measuring Technology and Mechatronics Automation (ICMTMA 2010)* (s. 968-971). Changsha, Çin: IEEE. doi:https://doi.org/10.1109/ICMTMA.2010.429