T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Yüksek Lisans Tezi

MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TÜRKÇE HABERLERİN ÖZETLENMESİ

Burak ÖZDEMİR

Enformatik Anabilim Dalı

Enformatik Programı

DANIŞMAN
Doç. Dr. Çiğdem EROL

Temmuz, 2021

İSTANBUL

Bu çalışma, 8.07.2021 tarihinde aşağıdaki jüri tarafından Enformatik Anabilim Dalı, Enformatik Programında Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Jürisi

Doç. Dr. Çiğdem EROL(Danışman)
İstanbul Üniversitesi
Enformatik Bölümü

Prof. Dr. Sevinç GÜLSEÇEN
İstanbul Üniversitesi
Enformatik Bölümü

Dr. Öğr. Üyesi Yalçın ÖZKAN
İstinye Üniversitesi
Yönetim Bilişim Sistemleri Bölümü

İntihal Programı Beyanı

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

Proje Destekleri

Tezden Üretilmiş Yayınların Künye Bilgileri

ÖNSÖZ

Desteklerinden ötürü tez danışmanım Doç. Dr. Çiğdem Erol'a, aileme ve Dilruba'ya teşekkür ederim.

Temmuz 2021 Burak ÖZDEMİR

İÇİNDEKİLER

Sayfa No

ÖNSÖZ ..... **iv**
İÇİNDEKİLER ..... **v**
ŞEKİL LİSTESİ ..... **viii**
TABLO LİSTESİ ..... **ix**
SİMGE VE KISALTMA LİSTESİ ..... **x**
ÖZET ..... **xi**
SUMMARY ..... **xiii**
1. GİRİŞ **1**
2. GENEL KISIMLAR **2**
2.1. LİTERATÜR TARAMASI 2
2.1.1. Tez Çalışmaları 2
2.1.2. Uluslararası Yayınlar 4
2.1.3. Ulusal Yayınlar 5
2.2. VERİ MADENCİLİĞİ 6
2.2.1. Metin Madenciliği 7
2.3. DOĞAL DİL İŞLEME 8
2.3.1. Doğal Dil İşleme Tarihi 8
2.3.2. Doğal Dil İşleme Kullanım Alanları 8
2.3.3. Doğal Dil İşleme Adımları 9
2.3.4. Doğal Dil İşleme Araçları 10
2.3.4.1. Zemberek 10
2.3.5. Doğal Dil İşlemede Metin Ön İşleme 11
2.3.6. Doğal Dil İşleme ve Etik 12
2.4. MAKİNE ÖĞRENMESİ 13
2.4.1. Gözetimli Öğrenme 13
2.4.2. Gözetimsiz Öğrenme 14
2.4.3. Pekiştirmeli Öğrenme 14
2.4.4. Makine Öğrenmesi Araçları 14
2.4.4.1. TensorFlow 14
2.4.4.2. Keras 15

2.4.4.3. scikit-learn 15
2.5. OTOMATİK METİN ÖZETLEME 15
2.5.1. Otomatik Özetlemede Bağlam Faktörleri 16
2.5.1.1. Girdi Faktörleri 16
2.5.1.2. Amaç Faktörleri 17
2.5.1.3. Çıktı Faktörleri 17
2.5.2. Çıkarıma Dayalı Özetleme 18
2.5.2.1. tf-idf 18
2.5.2.2. Gizli Anlam Analizi 19
2.5.2.3. TextRank 20
2.5.2.4. LexRank 22
2.5.2.5. BERT 22
2.5.3. Yorumlamaya Dayalı Özetleme 23
2.5.3.1. Seq2Seq 25
2.5.4. Otomatik Özetlemede Değerlendirme Yöntemleri 26
2.5.4.1. ROUGE 26
2.5.4.2. BLEU 27
2.5.4.3. Pyramid 27
3. MALZEME VE YÖNTEM 28
3.1. OTOMATİK METİN ÖZETLEYİCİ 29
3.1.1. Veri Toplama 29
3.1.2. Veri Temizleme 35
3.1.3. Veri Analizi 35
3.2. WEB UYGULAMASI 37
3.2.1. Tasarım 37
3.2.1.1. Sunucu Yapısı 38
3.2.1.2. Arka Uç Yapısı 39
3.2.2. Uygulama 39
4. BULGULAR 42
5. TARTIŞMA VE SONUÇ 45
5.1. SONUÇLARIN DEĞERLENDİRİLMESİ 45
5.2. GELECEK ÇALIŞMALAR İÇİN ÇEŞİTLİ ÖNERİLER 47
KAYNAKLAR 49
EKLER 55

EK 1. Etkisiz Kelimeler Listesi 55
ÖZGEÇMİŞ 56

ŞEKİL LİSTESİ

Sayfa No

Şekil 2.1: Doğal dil işleme adımları [47]. 9

Şekil 2.2: PageRank için beş adet sayfadan oluşan örnek bir graf. 20

Şekil 2.3: CNN evrişim katmanı [73]. 24

Şekil 2.4: Örnek RNN yapısı [73]. 24

Şekil 2.5: Örnek Seq2Seq modeli [75]. 25

Şekil 3.1: Wikihow Türkiye ana sayfası, Şubat 2021. 29

Şekil 3.2: Wikihow Türkiye Alexa ziyaretçi istatistikleri, Şubat 2021. 30

Şekil 3.3: Wikihow Türkiye'de yer alan örnek bir paragraf. 30

Şekil 3.4: Wikihow veri seti örneği. 32

Şekil 3.5: TRT Haber gündem haberleri. 32

Şekil 3.6: TRT Haber Alexa ziyaretçi istatistikleri, Şubat 2021. 33

Şekil 3.7: TRT Haber örnek haber. 33

Şekil 3.8: TRT Haber veri seti. 35

Şekil 3.9: Veri analizi. 36

Şekil 3.10: Web uygulaması akışı. 40

Şekil 3.11: Web uygulaması masaüstü arayüzü. 41

Şekil 3.12: Web uygulama mobil arayüzü. 41

Şekil 4.1: Wikihow veri seti için Seq2Seq özetleri. 42

Şekil 4.2: TRT Haber veri seti için Seq2Seq özetleri. 42

Şekil 4.3: Farklı yöntemlerle haber özetleri. 43

TABLO LİSTESİ

Sayfa No

Tablo 2.1: Örnek graf için PageRank matrisi. 21

Tablo 3.1: Geliştirme ortamı ve kullanılan yazılıma ait bilgiler. 28

Tablo 3.2: Wikihow veri setine ait istatistikler.....31

Tablo 3.3: TRT Haber veri setine ait istatistikler. 34

Tablo 4.1: Wikihow veri seti için ortalama performans skorları.....44

Tablo 4.2: TRT Haber veri seti için ortalama performans skorları. 44

KISALTMA LİSTESİ

Kısaltmalar	Açıklama
AJAX	: Eşzamansız JavaScript ve XML (*Asynchronous JavaScript and XML*)
API	: Uygulama Programlama Arayüzü (*Application Programming Interface*)
BERT	: Transformatörlerden Çift Yönlü Kodlayıcı Gösterimleri (*Bidirectional Encoder Representations from Transformers*)
CNN	: Evrişimli Sinir Ağları (*Convolutional Neural Network*)
CSS	: Basamaklı Stil Şablonları (*Cascading Style Sheets*)
HTML	: Hiper-Metin İşaretleme Dili (*Hypertext Markup Language*)
HTTP	: Hiper-Metin Transfer Protokolü (*Hyper-Text Transfer Protocol*)
IDF	: Ters Doküman Frekansı (*Inverse Document Frequency*)
JSON	: JavaScript Nesnesi Gösterimi (*JavaScript Object Notation*)
LSA	: Gizli Anlam Analizi (*Latent Semantic Analysis*)
LSTM	: Uzun Kısa Süreli Bellek (*Long Short-Term Memory*)
NLP	: Doğal Dil İşleme (*Natural Language Processing*)
nltk	: Doğal Dil Araç Kiti (*Natural Language Toolkit*)
RAM	: Rastgele Erişimli Hafıza (*Random Access Memory*)
RNN	: Tekrarlayan Sinir Ağları (*Recurrent Neural Network*)
ROGUE	: Recall-Oriented Understudy for Gisting Evaluation
TF	: Terim Frekansı (*Term Frequency)*
ufw	: Karmaşık Olmayan Güvenlik Duvarı (*Uncomplicated Firewall)*
XML	: Genişletilebilir İşaretleme Dili (*Extensible Markup Language*))

ÖZET

YÜKSEK LİSANS TEZİ

MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TÜRKÇE HABERLERİN ÖZETLENMESİ

Burak ÖZDEMİR

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Enformatik Anabilim Dalı

Danışman : Doç. Dr. Çiğdem EROL

İnternet'in yazılı basın sektörünü derinden etkileyip, dijital yayıncılığın önem kazanmasıyla birlikte elektronik ortamda gazete okuyan kişi sayısında büyük artış gözlemlenmiştir. Dijital ortamda artan rekabet, haber sitelerini insandan çok arama motoruna yönelik habercilik yapmaya yöneltmiş; zamanla yayınlanan bu haberler olması gerektiğinden daha uzun bir hâl almıştır. Öte yandan, bireylerin daha kısa sürede daha çok bilgiye ulaşabilmelerini sağlayabilecek sistemlere her zaman ihtiyaç duyulmuştur. Yoruma veya çıkarıma dayalı yöntemlerle gerçekleşen otomatik metin özetleme, metinlerdeki ana düşünce ve önemli bilgilerin korunması şartı ile metnin boyutunun küçültülmesidir. Bu tezde, Hint-Avrupa dil ailesindeki dillere göre daha farklı bir yapısı olan Türkçe dilinde üretilmiş haberler, doğal dil işleme yöntemleri ile özetlenmiş; yakın zamanda geliştirilmiş olan BERT ve Seq2Seq modellerinin performansları, metin özetleme için kullanılan geleneksel yöntemler olan tf-idf, gizli anlam analizi, TextRank ve LexRank ile karşılaştırılmıştır. Bu karşılaştırmanın sonucunda Türkçe haberlerden oluşan bir veri setinde ortalama 0,25 ROUGE-L skoruna sahip olan, performansı ve veri setinden bağımsız girdilerde de iyi sonuçlar üretebilmesi sebebiyle tercih

edilen BERT modeli ile herkesin erişimine açık, Özetleyici adına sahip web tabanlı bir otomatik özetleyici geliştirilmiştir.

Temmuz 2021, 70. sayfa.

Anahtar kelimeler: metin özetleme, otomatik metin özetleme, türkçe metin özetleme, doğal dil işleme, bert

SUMMARY

M.Sc. THESIS

SUMMARIZATION OF TURKISH NEWS WITH MACHINE LEARNING

Burak ÖZDEMİR

İstanbul University

Institute of Graduate Studies in Sciences

Department of Informatics

Supervisor : Assoc. Prof. Dr. Çiğdem EROL

Technological advancements in media led to digital platforms offering quicker access to real-time information. As a result, people prefer to read news from credible internet resources compared with print media. Innovative media organizations develop search engine-friendly content to stay competitive and prominent in digital journalism, thus having greater accessibility to consumers. This led to an increase in the search engine optimized contents of online news lengthier than traditional news. On the contrary, people in the digitalized world desire to interact with systems that provide more information within the shortest possible time. To cope with such challenges, automatic text summarization is a technique mainly utilized to shorten the text length without compromising the integrity of the information. Through this method, this thesis summarized news produced in the Turkish language with BERT and Seq2Seq models. These models' validity and overall performances compared with traditional summarization methods, tf-idf, LSA, TextRank, and LexRank. Ultimately, this research project

led to a web-based automatic text summarizer named Özetleyici. The BERT model used in this summarizer scored an average ROUGE-L score of 0.25 in a data set consisting of Turkish news.

July 2021, 70. pages.

Keywords: text summarization, automatic text summarization, turkish text summarization, natural language processing, bert

1. GİRİŞ

İnternet'in daha fazla kişinin hayatına girmesi ile birçok medya kuruluşu da haberlerini İnternet üzerinden yayınlamaya başlamış, online haber yayınlayan içerik yayınlayıcıları arasındaki rekabet artmıştır. Rekabetin artması, aynı haberlerin farklı içerik yayınlayıcıları tarafından farklı şekilde aktarılması sonucunu doğurmuş, bu durum, okuyucuların kısıtlı zamanda erişebileceği farklı nitelikteki haber miktarını etkilemiştir. Bu sebepten ötürü, büyük veriyi yorumlamak için kısıtlı zamanın ve kısıtlı insan gücünün olduğu günümüzde, haberin anlamını bozmayacak şekilde, haberin içerisinde yer alan önemli kısımları özet halinde sunabilecek otomatik sistemler önemli hale gelmiştir.

Bu çalışmada, Türkçe dilindeki haberler doğal dil işleme yöntemleri ile otomatik olarak özetlenmiş, test edilen yöntemlerin performansı ROUGE skorlarına göre karşılaştırılmıştır. Ardından, en iyi sonuçları veren modelin üzerine kurulan, kullanıcıların istedikleri metni özetleyebilmesini sağlayan bir İnternet sitesi kurulmuştur. Bu İnternet sitesi aracılığıyla, kullanıcıların ellerindeki metinleri, belirledikleri cümle sayısı ile özetleyebilmelerine olanak tanınmıştır.

Bu tez çalışmasının genel kısımlar bölümünde yapılan literatür taraması anlatılmış, ardından bu tezin temel dinamiklerini oluşturan veri madenciliği, doğal dil işleme, makine öğrenmesi ve otomatik özetleme kavramlarına değinilmiştir.

Malzeme ve yöntem bölümünde, veri setinin nasıl oluşturulduğu ve bu verinin hangi yöntemlerle nasıl işlendiği anlatılmış, oluşturulan İnternet sitenin nasıl çalıştığına değinilmiştir.

Bulgular bölümünde özetlemede kullanılan modellerin ve algoritmaların performansları hakkında elde edilen sonuçlar ortaya konulmuştur.

Tartışma ve sonuç bölümünde elde edilen sonuçlar değerlendirilmiş, sistemin nasıl geliştirilebileceği konusundaki fikirler belirtilmiştir.

2. GENEL KISIMLAR

Bu kısımda, literatür taraması veri madenciliği, doğal dil işleme, makine öğrenmesi ve otomatik özetleme hakkında bilgi verilmektedir.

2.1. LİTERATÜR TARAMASI

2.1.1. Tez Çalışmaları

Ulusal Tez Merkezi üzerinden "özetleme" anahtar kelimesi ile sonuncusu 1 Haziran 2021'de olmak üzere yapılan literatür taramasında, Türkçe metin özetleme ile ilgili 14 adet teze ulaşılmıştır. Bu tezlerle ilgili özet bilgisi aşağıda sunulmuştur.

Akülker [1] istatistiksel tabanlı terim frekansı ile ters doküman frekansı (tf-idf) değerlerini grafik tabanlı PageRank algoritması ile birleştirerek doküman özetlemiş; oluşturduğu bu hibrit modelin, modellerin tek başlarına uygulanmaları durumuna göre daha başarılı olduğunu F-Skor değeriyle göstermiştir.

Aysu [2] Hadoop aracılığı ile depoladıkları Türkçe haber metinlerini morfolojik analiz yaparak özetlemiş, çeşitli kişilerden aldıkları geri dönüşler ile sonuçları karşılaştırmış; özetlemenin kısa metinlerde, uzun metinlere göre daha başarılı olunduğunu göstermiştir.

Baydar [3] çıkarımsal özetleme yöntemi ile Türkçe metinleri özetleyerek ölçümler yapmış; sistem için insanlar tarafından çıkarılan referans özetlerin, insanların bilgi seviyesine göre değişiklik gösterdiğini ve bunun da sistemin başarısını etkileyebileceğini belirtmiştir.

Işık [4] Türkçe, İngilizce, İspanyolca, Fransızca, İtalyanca ve Almanca içerikler için çıkarıma dayalı çeşitli özetleme yöntemlerini karşılaştırarak değerlendirmede bulunmuş; Türkçe, Almanca ve İtalyanca dillerinde çıkarıma dayalı özetlemenin diğer karşılaştırdığı dillere göre daha az başarılı olduğunu göstermiştir.

Nuzumlalı [5] farklı seviyede uygulanan kelime kökü bulma yöntemleri ile cümle sadeleştirme tekniklerinin Türkçe için yapılan özetleme işlemindeki başarısını araştırmış; LexRank ile yaptığı, insanlar tarafından oluşturulmuş veri setinde, özetlemelerde kelime sonundan harf atma tekniğinin ROUGE skorunu artırdığını belirtmiştir.

Birant [6] ise Türkçe için otomatik metin özetleyici bir yazılım geliştirmiş; geliştirdiği bu yazılımda anlambilimsel ilişki sözcüklerini doğal dil işleme araçları ile kullanarak insanlar tarafından yapılan özetlere yakın sonuçlar elde etmiştir.

Yalkın [7] çizge tabanlı yöntemlerden TextRank ile metin özetleme yapmış; birleştirici hiyerarşik kümeleme ve TextRank ile geliştirdiği hibrit modelle en iyi ROUGE sonucunu almıştır.

Güran [8] çıkarıma dayalı metin özetleme sistemleri üzerine yoğunlaşarak, Türkçe ve İngilizce veri setleri için metin özetleme yöntemlerinde kullanılabilecek yeni bir değerlendirme değeri sunmuş; yaklaşık 100'den fazla haberden oluşan ve insanlar tarafından oluşturulmuş veri setlerine dayanarak, süreyi hesaba katınca eğiticisiz sistemlerin eğiticili sistemlerden performans açısından daha pratik olduğunu belirtmiştir.

Berker [9] metinde bulunan sözcük zincirleri gibi çeşitli değerleri hesaplayarak genetik algoritmalar ile İngilizce metinlerde özetleme yapmış; cümle konumu ve cümle merkeziliği gibi bazı özelliklerin, diğer özelliklere göre özet kalitesini daha çok etkilediğini göstermiştir.

Pembe [10] Türkçe ve İngilizce dokümanlar için, bilgi isteği ve doküman yapısına dayalı özetleme sistemi sunmuş; daha uzun özetlerin, arama motorlarının arama sonuçlarında kullanıcılara gösterdiği sadece arama sorgusuna bağlı olan içeriğe göre daha faydalı olabileceğini belirtmiştir.

Tülek [11] gövdeleme algoritmalarında sözcüklerin kök ve ek birleşimlerini üreten çözümleyici ile Türkçe için özetleyici yapmış; on adet belgeden oluşan insanlar tarafından oluşturulmuş veri setinde gövdeleme yöntemi kullanımında daha iyi sonuçlar elde edildiğini göstermiştir.

Ercan [12] kelime zincirleri ile özetleme yapmış; kelime bütünlüğüne bağlı özelliklerin, anahtar sözcük çıkarma işlemini iyileştirdiğini göstermiştir.

Erhandı [13] derin öğrenme yöntemleri ile metin özetleme yapmış; derin otokodlayıcılar ile oluşturduğu, ara katmanda uzun kısa süreli bellek (LSTM) kullandığı modelinde epoch sayısını arttırarak Türkçe haberler için anlamlı başlıklar üretmiştir.

Özkan [14] Türkçe haberlerde çıkarıma dayalı özetleme yapmış; herhangi bir NLP kütüphanesine ihtiyaç duymadan yaptığı özetlemede, Zemberek NLP tarafından elde edilen

sonuçlara yakın bir başarı elde etmiş; tırnak içerisine alınmış kısımların Türkçe haberlerin özetlenmesinde sorun yaratabileceğinden bahsetmiştir.

2.1.2. Uluslararası Yayınlar

Google Scholar üzerinden "automatic text summarization" ve "Turkish text summarization" anahtar kelimeleri ile sonuncusu 1 Haziran 2021'de olmak üzere yapılan literatür taramasında aşağıdaki uluslararası yayınlara ulaşılmıştır. Bu yayınlarda Türkçe metin özetlemeye ilişkin üç yayın bulunmuştur.

Luhn [15] metin özetleme üzerine yapılan ilk çalışmalardan birisini yapmış, kelimelerin önemi ve metin içerisindeki sıklığına göre özetleme yapan bir algoritma geliştirmiş, bu algoritmanın karmaşık çıktılar üretemese de etkili olabileceğini belirtmiştir.

Gong ve Liu [16] cümle ve kelimelerin anlamsal analizleri ile özetleme yapılmış, yapılan özetler insanlar tarafından yapılan özetlerle karşılaştırarak, uzun dokümanlarda bu yöntemin daha az başarılı olduğunu belirtmiştir.

Erkan ve Radev [17] LexRank adını verdikleri graf (*graph*) tabanlı, cümlelerin benzerliklerini dikkate alan bir yöntem geliştirmiş, bu yöntemin merkezi (*centroid*) yöntemlerden daha başarılı olduğunu belirtmiştir.

Mihalcea ve Tarau [18] TextRank adını verdikleri yöntemleri ile metin içerisindeki en önemli kısımları bulan, herhangi bir veri setine ihtiyaç duymadan, metnin diline ve konusuna çok bağlı olmadan, çıkarıma dayalı iyi bir özetleme yapılabileceğini göstermiştir.

Wong ve diğ. [19] destek vektör makineleri ve Naive Bayes sınıflandırıcı kullanarak geliştirdikleri özetleyici modellerin, etiketli veri sunulması durumunda oldukça iyi sonuçlar üretebileceğini; diğer yandan üretilen özetin karakter sayısının artırılmasının ROUGE skorunu da artırdığını göstermiştir.

Nallapati ve diğ. [20] tekrarlayan yapay sinir ağları ile geliştirdikleri Summarunner adını verdikleri çıkarıma dayalı metin özetleyici ile CNN ve Daily Mail veri setinde, daha önceden yayınlanmış ve derin öğrenme ile çalışan başka modellerin birçoğundan daha iyi performans göstererek daha yüksek ROUGE skoru elde etmiştir.

Nallapati ve diğ. [21] geliştirmiş oldukları yorumlamaya dayalı Seq2Seq modelleri ile CNN ve Daily Mail veri setini kullanarak, az cümleden oluşsa da yüksek ROUGE skoruna sahip bir özetleyici geliştirmiştir.

See ve diğ. [22] yorumlamaya dayalı olan ve tekrarlayan sinir ağları ile geliştirmiş oldukları Seq2Seq modellerinde kapsama (*coverage*) kullanarak özetlenenin takibini yapmış, tekrarın önüne geçerek daha önce yayınlanmış olan yorumlamaya dayalı modellerden daha yüksek ROUGE skoruna sahip bir model geliştirmiştir.

Liu [23] Transformatörlerden Çift Yönlü Kodlayıcı Gösterimleri (BERT) ile geliştirmiş olduğu ve BertSum adını verdiği çıkarıma dayalı model ile CNN ve Daily Mail veri setinde en iyi ROUGE skoruna sahip özetleyiciyi geliştirmiştir.

Altan [24] PHP (*Hypertext Preprocessor*) ve MySQL ile web tabanlı bir özetleyici geliştirmiş; yapısal, istatiksel ve dilsel analiz ile Türkçe için özet oluşturan modelinin daha da gelişebilmesi için WordNet gibi kelimelerin ilişkisini kuran bir veritabanına ihtiyaç duyduğundan bahsetmiştir.

Ozsoy ve diğ. [25] gizli anlam analizi ile Türkçe bilimsel makaleleri özetlemiş, kendi geliştirdikleri yaklaşımlar ile en iyi ROUGE-L sonuçlarını elde etmişlerdir.

Kutlu ve diğ. [26] cümlenin öneminin, terim frekansı ve cümle pozisyonu gibi çeşitli özelliklerin dahil edildiği bir hesaplama ile belirlendiği bir özetleyici geliştirmiş; bu özetleyici ile Türkçe haberlerden oluşan veri setinde 0,561, dergi makalelerinden oluşan veri setinde 0,368 ROUGE-L skoru elde etmişlerdir.

2.1.3. Ulusal Yayınlar

DergiPark ve Google Scholar üzerinden "Türkçe metin özetleme" anahtar kelimesi ile sonuncusu 1 Haziran 2021'de olmak üzere yapılan literatür taramasında beş yayın bulunmuştur.

Uzundere ve diğ. [27] Türkçe haber metinlerinin özetinin çıkarılması için, her birine sezgisel olarak ağırlık verdiği, aralarında başlık ve ortalama uzunluk gibi 13 özellik belirlemiş, kullanıcı özetleri ile yapılan karşılaştırmada %55 benzerlik elde eden, uzun metinlerde performansı düşen bir özetleyici geliştirmiştir.

Hatipoğlu ve Omurca [28] puanlandırma ve gizli anlam analizini birleştirerek bir çıkarıma dayalı bir model geliştirmiş, bu model tarafından yapılan özetlerin, kullanıcılar tarafından yapılan özetlere %82'ye kadar benzeyebileceğini belirtmiştir.

Acı ve Çırak [29] konvolüsyonel sinir ağları ve Word2Vec kullanarak Türkçe haberleri sınıflandırmış, derin öğrenme bazlı bu sistemin geleneksel makine öğrenmesi yöntemlerinden daha başarılı olduğunu belirtmişlerdir.

Çelik ve Koç [30] çeşitli makine öğrenmesi yöntemlerini ve vektör modeli oluşturucularını karşılaştırmış; FastText ve destek vektör makinesi ile Türkçe haberlerin en iyi şekilde sınıflandırılabileceğini belirtmiştir.

Kartal ve Kutlu [31] makine öğrenmesi ile yaptıkları Türkçe haber özetlemede, 130 Türkçe haber metni ve özeti üzerinde, kelime sıklığı, cümle konumu ve başlık benzerliği gibi bazı özellikleri kullanarak lojistik regresyon, destek vektör makinesi ve rastgele orman makine öğrenmesi yöntemleri ile özetleme yapmış, cümle konumunun en önemli özellik olduğunu belirtmiş, en iyi ROUGE-L skorunu 0,32 ile rastgele orman ile elde etmiş; bu skorun gizli anlam analizinden daha yüksek olduğunu belirtmiştir.

2.2. VERİ MADENCİLİĞİ

Büyük veri kavramı ile hemen hemen her sektörde karşılaştığımız günümüzde, veri madenciliği, büyük veri kümelerindeki kalıpları çıkarmak ve keşfetmek için oldukça önemlidir.

İlk olarak 1998 yılında eski bir Silicon Graphics çalışanı olan John Mashey tarafından ortaya atılan bir kavram olarak büyük veri, yüksek miktardaki veri kümelerinin işlenmesini ve analizini ifade etmek için kullanılan bir terimdir [32]. Büyük veri kavramının ortaya çıkmasının ardından, mevcut yazılım araçlarının büyük veriyi efektif biçimde işleyememesi, beraberinde veri madenciliği kavramını önemli bir araştırma alanı haline getirmiştir [33]. Veri madenciliği, potansiyel olarak faydalı olan, ancak varlığı henüz keşfedilmemiş bilgileri büyük verinin arasından ayıklama işlemidir [34].

Veriyi efektif biçimde işleyebilmek için veri madenciliğinin çözmesi gereken bazı problemler vardır. Verinin hep aynı biçimde olmaması, verinin genellikle tek bir kaynaktan toplanamaması

ve veri madenciliği algoritmalarının değişken performansı gibi çeşitli sorunlar bunlara örnek olarak verilebilir [35].

Çeşitli veri tabanlarında veya İnternet'te yer alan verinin arasından değerli bilgilerin ortaya çıkarılması, birçok iş fırsatı ortaya çıkarmış ve çıkarmaya da devam etmektedir [35]. Özellikle İnternet'te, çok sayıda işlenmemiş, düşük kaliteli olarak tabir edilen kullanıcılar tarafından oluşturulmuş bilgi mevcuttur. Bu yüzden, birçok şirket bu veriyi kullanıp, anlamlandırarak, yüksek kâr elde edebilecekleri modeller oluşturmaya çalışmaktadır [36].

2.2.1. Metin Madenciliği

Veri madenciliğinin bir varyasyonu olan metin madenciliği, çeşitli yazılı kaynaklardan bilginin otomatik olarak çıkarılması ile önceden bilinmeyen bilgilerin bilgisayarlar tarafından keşfedilmesidir [37].

Metin madenciliği ile kullanıcılar tarafından yapılan arama motoru sorguları aynı değildir. Web aramalarında kullanıcılar zaten bilinen ve başkası tarafından yazılmış olanları aramaktadır. Diğer yandan metin madenciliğinde, bilinmeyen, henüz keşfedilmemiş çeşitli bilgiler algoritmalar aracılığıyla ortaya çıkarılmaktadır. Benzer biçimde, veri madenciliği ile metin madenciliği arasında da bazı farklar bulunmaktadır. Metin madenciliğinde, tasarım örüntülerinin belirli yapıda düzenlenmiş veri tabanları yerine, doğal dil işleme (NLP) aracılığı ile ortaya çıkarılmaktadır [37].

Metin madenciliği teknikleri en çok aşağıdaki amaçlar için kullanılmaktadır [38]:

- Özellik çıkarma
- Metinde arama
- Metin kategorizasyonu
- Metnin konusunun saptanması
- Metin özetleme

Metin madenciliği teknikleri ile çeşitli amaçlar doğrultusunda, yazılı içeriklerdeki önemli veya önemsiz bilgiler bulunabilmektedir. Metin madenciliğinin en çok kullanıldığı uygulama alanlarına örnek olarak, medya, telekomünikasyon, araştırma ve reklamcılık verilebilir [37].

2.3. DOĞAL DİL İŞLEME

Doğal dil işleme (NLP), bilgisayarların, çeşitli amaçlarla doğal dildeki metni veya konuşmayı anlamak ve işlemek için nasıl kullanılabileceğini araştıran bir araştırma alanıdır. NLP araştırmacıları, bilgisayar sistemlerinin istenen görevleri gerçekleştirmek için doğal dilleri anlamasını ve manipüle etmesini sağlamak için bilgisayar bilimleri, yapay zekâ ve dilbilim aracılığıyla çeşitli araçlar ve teknikler geliştirmektedir [39].

2.3.1. Doğal Dil İşleme Tarihi

Doğal dil işleme ile ilgili olduğu düşünülen ilk bilgisayar tabanlı uygulama, 1940'lı yılların sonlarına doğru geliştirilmiş bir makine çevirisi uygulamasıdır [40]. Chomsky, 1953'te tüm insanların doğuştan bir dil kapasitesine sahip olduğunu öne süren üretken dilbilgisi teorisini ortaya atmıştır [41]. Makine çevirisi için büyük önemi olan bu teori ile daha fazla NLP çalışması yapılmaya başlanmıştır. 1960'lı yıllarda, Chomsky'nin çalışmaları haricinde Schank'ın kavramsal bağlılık teorisi [42] gibi çalışmalar gelecekteki birçok NLP çalışmasına rehberlik etmiştir.

NLP'nin günümüzdeki popülerliğine ulaşmasındaki en büyük farkındalık ise 1980'li yıllarda olmuştur [40]. Var olan çözümlerin yetersizliğinin anlaşılmasının yanında, İnternet'in icadı ile NLP çalışmaları hız kazanmıştır. 1980'li yıllarda ortaya çıkan Bayesian ağları ve anlamsal ağlar, NLP araştırmalarındaki öne çıkan diğer gelişmeler olmuştur [43].

Günümüzde NLP en popüler araştırma konularından birisi olmaya devam etmektedir. Yakın zamanda sinir ağlarının güçlü makine öğrenmesi yöntemleri olarak tekrar araştırma konusu olması ile görüntü tanıma ve konuşma işlemi gibi alanlarda önemli çalışmalar yapılmıştır. Bu gelişmeleri, son yıllarda sinir ağlarının NLP uygulamalarında sıklıkla kullanılması takip etmektedir [44].

2.3.2. Doğal Dil İşleme Kullanım Alanları

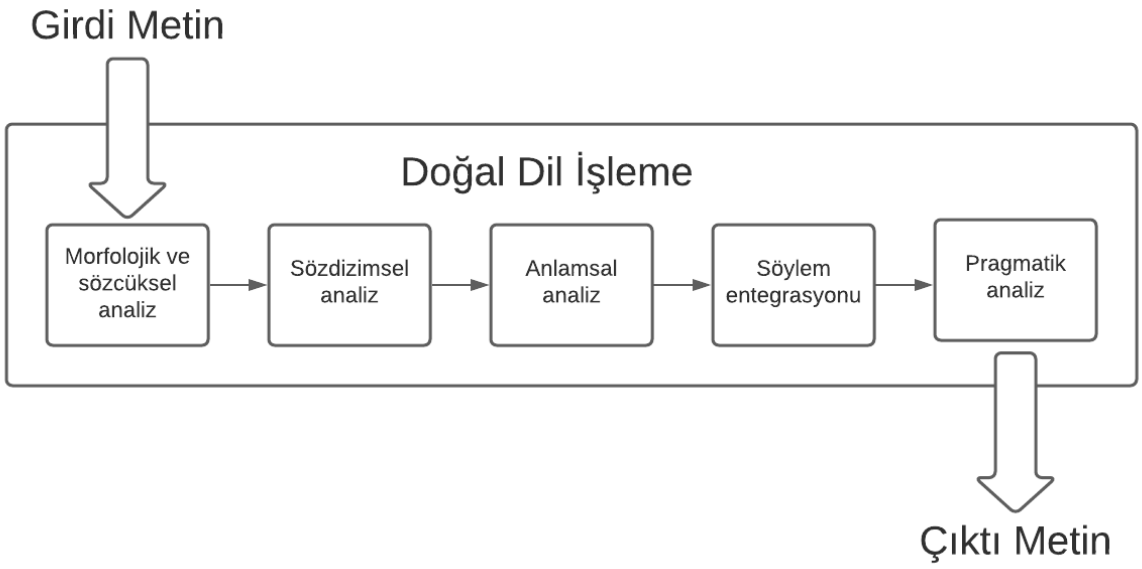
Doğal dil işlemenin kullanım alanlarının bazıları aşağıdaki gibidir [45]:

- **Makine çevirisi**: Metinlerin bir dilden bir başka dile çevrilmesidir. Bu alandaki en büyük zorluk, kelimeleri direkt olarak çevirmek yerine, anlamsal bütünlüğü koruyacak şekilde çevirebilmektir.

- **Metin kategorizasyonu**: Metnin, metin sınıflandırıcılar ile analiz edilmesi ve ardından içeriğine göre önceden tanımlanmış bir dizi kategori ile sınıflandırılmasıdır. Spam filtreleme çalışmaları metin kategorizasyonu için örnek verilebilir.
- **Bilgi çıkarma**: Bilgi çıkarma, yapılandırılmış bilgilerin yapılandırılmamış veya yarı yapılandırılmış kaynaklardan otomatik olarak çıkarılmasıdır. Bu işlem ile adlar, yerler, olaylar ve tarihler gibi başka işlemlerde işlenebilecek veri metinden çıkarılarak, yapılandırılmamış veriyle hesaplama yapılabilir. Örneğin, bir web sayfasındaki önemli bilgilerin otomatik olarak tanınması, bir arama motorunun performansını oldukça artırabilir.
- **Özetleme**: Metin gibi bir veri setinin anlamını kaybetmeden boyutunun küçültülmesidir.

2.3.3. Doğal Dil İşleme Adımları

Doğal dil işlemenin ile ilgili beş aşama vardır ve bu adımlar Şekil 2.1'de görülebilir [46]:



Şekil 2.1: Doğal dil işleme adımları [47].

1. **Morfolojik ve sözcüksel analiz**: Bir dilin kelimelerini ve ifadelerini içeren kelime hazinesidir. Morfoloji, kelimelerin yapısını analiz edip, tanımlamasını sağlar. Sözcüksel (*lexical*) analizde içerik, cümlelere, paragraflara ve kelimelere ayrıştırılır. Sözcüksel

analiz Tayca gibi kelime segmentasyonunun kolay olmadığı dillerde oldukça önemlidir [47].

2. **Sözdizimsel analiz**: Sözdisimsel (*syntactic*) analiz, cümledeki kelimelerin analizidir. Bu kelimelerin birbirleri ile olan ilişkileri ele alınır.
3. **Anlamsal analiz**: İçeriğin anlamının incelenmesidir.
4. **Söylem entegrasyonu**: Bir cümlenin anlamının ortaya çıkarılmasında, kendisinden önce gelen cümlelerin analiz edilmesidir.
5. **Pragmatik analiz**: Dilin amacına uygun kullanımının irdelenmesidir.

2.3.4. Doğal Dil İşleme Araçları

Doğal dil işleme ve makine öğrenme yöntemleri için geliştirilmiş birçok araç mevcut olup; geliştirilen bu araçların birçoğu Python ve C++ programlama dillerinde geliştirilmiştir [48].

Python ile kullanılan en popüler doğal dil işleme araçlarından ikisi doğal dil araç kiti (NLTK) ve spaCy'dir. Bu araçların sağladığı birçok işlemde Türkçe desteklenmemektedir. NLTK, sınıflandırma, metnin cümlelere ve kelimelere bölünmesi (*tokenization*), gövdeleme (*stemming*) gibi temel NLP işlemlerine olanak tanıyan, aynı zamanda Türkçe dahil birçok dilde etkisiz kelimelerin kaldırılmasını sağlamaktadır[1]. spaCy ise Python ve Cython için geliştirlen, NLP için geliştirilmiş, etiketleme, varlık tanıma gibi birçok çeşitli özellik sunan, en kapsamlı açık kaynaklı Python kütüphanelerinden birisidir[2].

Türkçe için en çok başvurulan doğal dil işleme araçlarından birisi ise Zemberek'tir.

2.3.4.1. Zemberek

Zemberek, çoğu NLP aracının sadece Hint-Avrupa dillerine yönelik olması ve Türk dili için geliştirilmiş özel bir aracın olmaması sebebiyle geliştirilmiş bir doğal dil işleme aracıdır [49].

Java programlama dili ile geliştirilen Zemberek, açık kaynaklı bir yazılım olup, Github[3] üzerinden herkesin erişimine açıktır. Her ne kadar Zemberek bir Java uygulaması olsa da Zemberek'i Python 3 ile çalıştırmak mümkündür. Bunun için Zemberek Parser[4] kullanılabilir.

---

[1] https://www.nltk.org
[2] https://spacy.io
[3] https://github.com/ahmetaa/zemberek-nlp
[4] https://github.com/kemalcanbora/zemberek_parser

Github sayfasında nasıl kullanılacağı detaylı olarak açıklanan Zemberek'in sağladığı birçok modül bulunmaktadır. Bu modüller aracılığıyla Zemberek ile Türkçe metinlerde aşağıdakileri yapmak mümkündür:

- Morfolojik analiz
- Metnin cümlelere ve kelimelere bölünmesi (*tokenization*)
- Yazım denetleme
- Kelime önerisi
- Varlık ismi tanıma (*Named entity recognition*)
- Metin sınıflandırma
- Metin dilini tanıma

2.3.5. Doğal Dil İşlemede Metin Ön İşleme

Metni veri olarak kullanan çalışmalarda, ön işleme olarak tanımlanan işlemler ile girdiler, sonuçları olumsuz etkilemeyecek şekilde, daha az karmaşık hâle getirilir. Bu işlemlerden bazıları aşağıdaki gibidir [50]:

- **Özel karakterler, noktalama işaretleri ve sayıların metinden silinmesi**: Sayılar, noktalama işaretleri, özel karakterler (#, $, vb.) gibi harf olmayan karakterler her ne kadar bazı çalışmalarda (örneğin Twitter'daki hashtag kullanımı) önemli olabilse de çoğu uygulamada bunların kaldırılması standart bir uygulamadır.
- **Küçük harfe çevirme**: Birçok uygulamada gerçekleştirilen bir başka ön işleme adımı, metindeki tüm kelimelerin küçük harfle yazılmasıdır. Bunun sebebi, ilk harfi büyük olan bir kelime ile tüm harfleri küçük olan bir kelimenin analiz esnasında iki farklı kelime gibi yorumlanmasıdır. Özel durum olarak, bir kelimenin ilk harfinin büyük olması, bu kelimenin özel bir isim olduğunu belirtebileceği için, bu gibi durumlarda küçük harfe çevirmek sonucu değiştirebilir.
- **Gövdeleme**: Bir kelimenin farklı biçimlerinin, morfolojik ekleri atılarak kelimenin kökü ile değiştirilmesi işlemine gövdeleme (*stemming*) denir. Bu sayede metin içerisindeki tekil kelime sayısı azaltılır. Ancak, bu durum bazen farklı anlama gelen, ancak aynı kelime köküne sahip farklı kelimelerin aynı kelimeymiş gibi yorumlanması ile metin içerisinde anlam sebep olabilir. Bu durum özellikle eş sesli kelimelerde

gözlenmektedir. Örneğin yüzmek kelimesinin kökü, sayı olan yüz ile aynı olduğu gibi, bir organ olan yüz ile de aynıdır.

- **Etkisiz kelimelerin çıkarılması**: Metnin kelimelerine ayrılmasından sonra, genellikle etkisiz kelime (*stop words*) olarak sınıflandırılan, fazla anlam ifade etmeyen kelimeler metinden çıkarılır. İngilizcede "the", "a", "is" gibi olan bu kelimelerin, Türkçedeki örnekleri "bir", "birçok" gibi kelimeler ve "de", "da" gibi bağlaçlardır. Bu kelimelerin belirlendiği, herkes tarafından kabul edilen standart bir liste yoktur. İngilizce için belirlenen listelerin çoğu yüz ila bin karakterden oluşmaktadır. Standart bir liste olmadığı için, geliştiriciler genellikle yazılım paketleri tarafından sağlanan listeleri kullanmaktadır.
- **n-gram özelliklerini ayıklama**: Her ne kadar birçok analizde kelimeler tek tek analiz edilse de, bazı kelimeler çıkarıldıktan sonra, bu kelimenin ardından gelen kelime anlamını yitirebilir. n-gram, metnin n uzunluğunda arka arkaya gelen parçalar halinde analiz edilmesi ile elde edilen bir dizidir. n adet öğeden oluşan bu bitişik dizilerde, n değerinin seçimi çalışmaya göre değişmekte, analist tarafından belirlenmektedir.

2.3.6. Doğal Dil İşleme ve Etik

NLP çalışmalarının ne kadar etik olduğu konusunda çeşitli soru işaretleri vardır [51]. Her ne kadar başlangıçta NLP çalışmalarında insanlar konu olmasa da sosyal medya verisinin NLP çalışmalarında sıklıkla kullanılmasıyla etik tartışmaları daha çok tartışılır olmuştur [52]. Örneğin cinsiyetin NLP çalışmalarında bir değişken olarak kullanılması bir araştırmaya konu olmuş; araştırmacıların cinsiyetin yanında din ve ırk gibi değişkenlere hassasiyetle yaklaşması önerilmiştir [53].

NLP'nin çeşitli sosyal etkileri mevcuttur [52]:

- **Dışlama**: Her dilin farklı bir demografik yapıyı temsil etmesi sebebiyle, veri setinin olası bir yan etkisi olarak ortaya çıkan bu durumda, NLP çalışmaları bilimsel çalışmaların evrenselliğini ve nesnelliğini tehdit edebilmektedir.
- **Aşırı genelleme**: Yanlış pozitiflerin etkisiyle, yapılmaması gereken genellemeler yapılabilmekte ve cevaplar, cevapsızlıktan daha kötü bir etki yaratabilmektedir.
- **Maruz kalma sorunu**: Daha fazla ilgi çeken çalışmalar ayrımcılığa sebep olabilmektedir. Örneğin, NLP araçlarının İngilizce için geliştirilmesinin en önemli

sebeplerinden birisi, İngilizcenin lingua franca kabul edilmesi ve fazla karmaşık bir dil olmamasıdır. Ancak, bu durum diğer diller için daha az NLP aracı geliştirilmesine sebebiyet vermektedir.

- **Çift kullanım sorunları**: Her ne kadar çalışmalar hiçbir zarar amacı gütmeden yapılsa da tüm toplum ile paylaşılan bilimsel çalışmaların kimlerce ne amaçla kullanıldığını bilmek mümkün değildir.

Etik ile ilgili sorunları aşmak ve bu konuda farkındalık yaratmak için, şirketlerce, araştırma aşamasından ürünü veya servisi hayata geçirme aşamasına kadar her aşamayı denetleyen etik inceleme kurullarının hayata kurulması önerilmektedir [51].

2.4. MAKİNE ÖĞRENMESİ

Makine öğrenimi, veriye dayalı olarak zaman içerisinde doğruluğu artan uygulamalar geliştirmeye odaklanan bir bilim dalıdır. Günümüzün en popüler çalışma alanlarından birisi olan ve bilgisayar bilimi ile istatistiğin kesişimi olarak nitelendirilebilecek bir alan olan makine öğrenmesi, bilim, teknoloji, sağlık hizmetleri, finansal modelleme gibi birçok sektördeki çeşitli problemleri çözmeye odaklanmaktadır. Makine öğrenmesi yaklaşımları, temel olarak gözetimli (*supervised*), gözetimsiz (*unsupervised*) ve pekiştirmeli (*reinforcement*) olarak üçe ayrılmaktadır [54].

2.4.1. Gözetimli Öğrenme

En yaygın kullanılan makine öğrenimi yöntemleri, gözetimli öğrenme yöntemleridir. İstenmeyen (*spam*) e-maillerin ayıklanması, yüz tanımlaması, tıbbi tanı sistemleri gibi çeşitli uygulamaların geliştirildiği bu yaklaşımda, örnek bir girdi-çıktı veri setine dayanarak, bir girdi için, bir çıktı tahmini yapılmaktadır [54]. Sık kullanılan gözetimli öğrenme yöntemleri arasında karar ağaçları, Bayes sınıflandırıcıları ve destek vektör makineleri yer almaktadır [55]. Ancak, son yıllarda, gözetimli öğrenme yöntemleri arasında en çok ilerleme katedilen alanlardan birisi derin öğrenme sistemleri olmuştur. Derin öğrenme sistemleri, gradyan bazlı optimizasyon algoritmalarını kullanarak, milyarlarca parametre içeren modeller ile bir çıktı üretirler. Bu sistemler, özellikle görüntü ve konuşma tanıma alanlarında büyük gelişmelere yol açmıştır [54].

2.4.2. Gözetimsiz Öğrenme

Gözetimsiz öğrenme algoritmaları, etiketlenmemiş verinin, verinin cebirsel veya olasılıksal olmaları gibi çeşitli yapısal özellikleriyle ilgili varsayımlar altında analizini içerir [54]. Bu algoritmalar, genellikle kümeleme ve özellik azaltma işlemleri için kullanılırlar. Yeni bir veri ile karşılaştıklarında, daha önceden öğrenmiş oldukları özellikleri kullanarak yeni veri sınıflarını tanımlayabilirler. K-means kümeleme ve temel bileşen analizi gibi algoritmalar, bu öğrenme yönteminde en sık kullanılan algoritmalar arasında yer almaktadır [55].

2.4.3. Pekiştirmeli Öğrenme

Pekiştirmeli öğrenme, öznelerin bir ortamda en yüksek ödüle ulaşabilmesi için nasıl hareket etmesi gerektiği ile ilgilenen temel makine öğrenmesi yaklaşımlarından birisidir. Bu öğrenme yaklaşımının gözetimli öğrenmeye göre en büyük farkı, etiketli girdi ve çıktılara ihtiyaç duymaması olup, keşfedilmeyenleri keşfetmeye odaklanmasıdır [56].

Bu öğrenme yönteminin algoritmaları genellikle kontrol teorisi literatüründeki fikirleri kullanır ve sinirbilim ile arasında çeşitli bağlantılar bulunmaktadır [54].

2.4.4. Makine Öğrenmesi Araçları

Doğal dil işleme ve makine öğrenme yöntemleri için geliştirilmiş birçok araç mevcut olup; geliştirilen bu araçların birçoğu Python ve C++ programlama dillerinde geliştirilmiştir [48].

Python programlama dili kullanılan bu tezde, Python için geliştirilmiş makine öğrenmesi araçlarına değinilmiştir. Tez kapsamında kullanılan Python kütüphanelerinden TensorFlow, Keras ve scikit-learn, geliştiricilerce de sıklıkla tercih edilen makine öğrenmesi araçlarındandır.

2.4.4.1. TensorFlow

Google Brain ekibi tarafından geliştirilen büyük ölçekli bir makine öğrenmesi aracı olan TensorFlow, CPU ve GPU gücünden faydalanarak geliştiricilerin kolay biçimde derin öğrenme yöntemlerini uygulayabilmesine olanak tanımaktadır [57].

Windows, macOS ve Linux işletim sistemlerinde çalışabilen TensorFlow, aynı zamanda Raspberry Pi için geliştirilmiş bir işletim sistemi olan Raspbian işletim sistemini de

desteklemektedir[5]. Ayrıca TensorFlow'un tarayıcıda çalışabilmesine olanak tanıyan, JavaScript bazlı farklı uyarlamaları da mevcuttur [48].

TensorFlow ile doğrusal regresyon gibi basit modeller oluşturulabileceği gibi, sinir ağlarından oluşan karmaşık modeller gerçeklenebilir [6].

2.4.4.2. Keras

TensorFlow üzerinde çalışan, Python için geliştirilmiş bir derin öğrenme kütüphanesi olan Keras, Kaggle'da en çok kullanılan derin öğrenme kütüphanelerinden birisidir [7]. Kaggle, veri bilimcilerin ihtiyaç duydukları kod ve veriye ulaşabilecekleri bir İnternet sitesi olup, veri bilimi ile ilgilenen kişilerin yarışmalara katılabileceği bir platformdur[8].

2.4.4.3. scikit-learn

scikit-learn, gözetimli ve gözetimsiz makine öğrenme yöntemlerini barındıran, Python yazılım dili için geliştirilmiş, uzman olmayanların dahi makine öğrenmesi yöntemlerini kullanabilmesine olanak tanıyan bir makine öğrenmesi kütüphanesi olup; SciPy ekosistemi üzerinde çalışmaktadır [58].

SciPy, Python programlama dili için geliştirilmiş, bilimsel hesaplamalar için kullanılan açık kaynaklı bir Python ekosistemi olup; NumPy, Matplotlib, SymPy, IPython ve pandas gibi kütüphanelerden oluşmaktadır[9].

2.5. OTOMATİK METİN ÖZETLEME

Metnin içerisindeki önemli kısımların ortaya çıkarılabilmesi için birçok yaklaşım geliştirilmiştir [59]. Otomatik özetleme, içerikte önemli olan kısımların belirlenmesi ile içeriğin boyutunun küçültülmesi işlemidir [60]. Özetleyici ise, kullanıcıların ihtiyaçları doğrultusunda, özetlenmiş bir içerik sunan sistemdir [61]. Özetlemenin amacı bir bilgi kaynağının içerisinden

[5] https://www.tensorflow.org
[6] https://github.com/aymericdamien/TensorFlow-Examples
[7] https://keras.io
[8] https://www.kaggle.com
[9] https://www.scipy.org

en önemli kısmı ayrıştırarak, bu önemli içeriği kullanıcıya yoğunlaştırılmış bir şekilde sunmaktır [62].

Otomatik metin özetleme, kaynak metnin, içeriğinin ve genel anlamının korunarak küçültülmesi işlemidir [63]. Bu alandaki çalışmalar ilk olarak 1950'li yıllarda IBM araştırma laboratuvarlarında, kelime ve sözcük öbeklerinin analizi ile yapılmıştır [15]. Günümüzde, özellikle İnternet'in yaygınlaşması ile bu alanda yapılan araştırma sayısında eksponansiyel artış gözlemlenmiştir [63].

Çıkarıma dayalı ve yorumlamaya dayalı olmak üzere, otomatik özetlemede iki temel yaklaşım vardır [64]. Çıkarıma dayalı özetlemede metin içerisindeki alt gruplar kullanılırken, yorumlamaya dayalı özetlemede metin içerisindeki ifadeler farklı şekilde belirtilmektedir. Her ne kadar insanlar tarafından yapılan özetlemeler çıkarıma dayalı olmasa da yapılan araştırmaların çoğu çıkarıma dayalı özetleme üzerine olmuştur. Bunun sebebi, doğal dil işlemeye dayalı olan yorumlamaya dayalı özetlemenin, sadece veriye dayalı olan çıkarıma dayalı özetlemeye göre daha zor olmasıdır [17].

Yakın zamana kadar otomatik metin özetlemede kullanılan geleneksel yöntemler, araştırmalardaki yerlerini yapay sinir ağları ile oluşturulan modellere bırakmaktadır [65].

2.5.1. Otomatik Özetlemede Bağlam Faktörleri

Otomatik özetleme sistemlerini etkileyen temel olarak üç bağlam faktörü vardır. Bunlar girdi, amaç ve çıktı faktörleridir [60].

2.5.1.1. Girdi Faktörleri

Girdi faktörleri kaynak yapısı, konu tipi ve birim olarak üçe ayrılmaktadır [60].

Kaynak yapısı faktöründe, girdinin yapısı, ölçeği ve türü dikkate alınır. Girdinin yapısı, girdideki alt başlıklar ve girdi içerisindeki örüntülere bağlıdır; ölçek faktöründe girdinin bir kitap veya sadece bir paragraftan oluşması dikkate alınır; tür faktöründe girdinin anlatı şeklinin farklı olması özetlemeyi etkiler [60]. Girdi yapısı ile ilgili bir diğer önemli faktör ise, girdinin dilidir. Girdi, tek dilli veya çok dilli olabilir. Tek dilli özetleyiciler, girdi ile aynı dilde özetler üretirken, çok dilli özetleyiciler aynı dildeki girdi-çıktı çiftini birden fazla dilde işleyebilir [65].

Girdinin konusu da girdi faktörlerinden birisidir. Bu husus, daha çok okuyucunun neyi bilip, neyi bilmediği için önemlidir. Örneğin içerisinde fazla teknik bilgi içeren bir içerik, bu konuda yeterli bilgi sahibi olmayan okuyucular için sadeleştirip, daha basit hâle getirilebilir [60].

Birim faktöründe, özetleme sisteminin girdi olarak kabul ettiği doküman sayısı ifade edilir. Girdinin tek tek bir dokümandan veya birden fazla dokümandan oluşması olarak, özetlemeyi etkileyen bir faktördür [65].

2.5.1.2. Amaç Faktörleri

Özetleme stratejisi için en önemli faktör amaç faktörüdür. Üretilen özetlerin değerlendirmesinin temelini oluşturan bu faktör, durum, hedef kitle ve kullanım olarak üç alt başlığa ayrılır [60].

Durum alt başlığında, özetin kullanımı ana hatlarıyla irdelenir; özetin kim tarafından, hangi amaçla, ne zaman kullanılacağı dikkate alınır. Hedef kitle alt başlığında, özetleyici tarafından üretilen özetleri okuyacak olan okuyucu kitlesi dikkate alınır. Bunun sebebi, kadınlar için içerik üreten bir dergi için üretilecek özetlerin, akademik amaçlar doğrultusunda özetler üretmesi için tasarlanmış bir sistem tarafından verimli biçimde oluşturulamayacak olmasıdır. Kullanım alt başlığında ise özetin kullanım amacı dikkate alınır [60]. Örneğin, üretilen özetlerin genel veya kullanıcı odaklı olması dikkate alınabilir. Genel özetlerde, belgedeki tüm bilgilere dayanan özetler oluşturulurken, kullanıcı odaklı bir özette, okuyucu kitlesi göz önünde bulundurulur. Benzer biçimde, özetler genel amaç veya alana özel de oluşturulabilir. Alana özel özetler, belirli bir alandaki belgeleri işlemek için tasarlanan özetleyiciler tarafından oluşturulur [65].

2.5.1.3. Çıktı Faktörleri

Materyal, format ve stil alt başlıklarında toplanan çıktı faktörleri, üretilen çıktının niteliği ile ilgilidir. Üretilecek çıktının kaynaktaki tüm önemli bilgileri kapsaması gerektiği ya da sadece bir miktarını kapsamasının yeterli olması materyal faktörü olarak tanımlanmaktadır. Format faktöründe üretilecek özetin başlıklarla birlikte oluşturulması veya oluşturulmaması gibi faktörler dikkate alınmaktadır. Stil faktöründe, özetin bilgilendirici veya gösterge niteliğinde olmasına dikkat edilebilir. Gösterge niteliğindeki bir özet, belgelerin içeriği konusunda yol göstericiyken, bilgilendirici bir özet, bir içerikten önemli kısımların çıkarılmasını sağlar [60].

2.5.2. Çıkarıma Dayalı Özetleme

Orijinal metindeki cümlelerin arasından, belirli bir alt kümenin değiştirilmeden çıkarılması ile özetlerin üretilmesi işlemine çıkarıma dayalı özetleme denir. Bu şekilde oluşturulan özetler, bir veya birden fazla belgeden oluşabilen girdinin en önemli cümlelerini içerir [64]. Çıkarıma dayalı metin özetlemede, genellikle kullanıcının çıkarılacak cümle sayısını veya yüzdesini tanımlamasına da izin verilir [66].

2.5.2.1. tf-idf

Çıkarımsal özetleme üzerine yapılan ilk araştırmalar, cümlelerin metin içerisindeki konumları, içerdikleri kelimelerin sıklıkları ve cümlelerin önemi gibi bazı özelliklere dayanmıştır. Bir cümledeki kelimenin önemini değerlendirmek için yaygın olarak kullanılan bir ölçü olan ters doküman frekansı (idf) aşağıdaki gibi tanımlanmıştır: [17]

$$idf_i = \log\left(\frac{N}{n_i}\right) \tag{2.1}$$

Denklem 2.1'de N, toplam belge sayısını ifade ederken $\boldsymbol{n_i}$, i kelimesinin yer aldığı belge sayısını ifade etmektedir. Bu denkleme göre, her belgede yer alan kelimelerin idf değeri sıfıra yakın olacakken, dokümanda sık yer almayan, dokümanlarda kendisine pek yer bulamamış, daha az kullanılan kelimeler yüksek idf değerine sahip olacaktır [17].

Bir belgede bir terimin ne sıklıkta geçtiğini ölçen terim frekansı, bir terimin bir dokümanda görülme sayısının, dokümandaki toplam terim sayısına olan oranıdır [63].

Bir kelimenin belge içerisindeki önemini gösteren bir skor parametresi olarak w_i, terim frekansı ile ters doküman frekansının çarpımı olup, denklem 2.2'deki gibi gösterilebilir [67]:

$$w_i = tf_i idf_i = tf_i \log\left(\frac{N}{n_i}\right) \tag{2.2}$$

Herhangi bir makine öğrenmesi yöntemine ihtiyaç duyulmayan, istatistiğe dayalı bir yöntem olan tf-idf ile tek bir dokümanda özetleme yapmak için aşağıdaki adımlar izlenebilir [68]:

1. Metindeki büyük harfler, küçük harfler ile değiştirilir
2. Metin içerisindeki etkisiz kelimeler (*stop words*) metinden çıkartılır

3. Metin önce cümlelere, ardından kelimelere bölünür (*tokenize*)
4. Kelimeler isim, fiil, sıfat ve zarf olarak sınıflandırıldıktan sonra, özet oluşturmak için daha ideal olan fiil ve isimler bırakılır
5. Kelimeler, kökleri ile değiştirilir (*stemming*)
6. Kalan kelimelerin, ondalık kısmında on basamak olacak şekilde tf-idf değeri hesaplanır
7. Kelimeler tf-idf değerine göre azalan biçimde sıralanarak bir sözlük oluşturulur
8. Cümleler için, cümle içerisindeki fiil ve isimlerin tf-idf değerlerinin toplamı hesaplanır
9. tf-idf değeri en yüksek olan cümleler seçilir
10. Seçilen cümlelerin doküman içinde görüldükleri konuma göre artan biçimde sıralama yapılarak, tf-idf değeri yüksek olan bu cümlelerden özet oluşturulur

2.5.2.2. Gizli Anlam Analizi

Gizli anlam analizi (LSA), bir veya birden fazla dokümandaki cümle ve kelimelerin anlamsal olarak analiz edildiği gözetimsiz bir yöntemdir [16].

Bu yöntem ile gerçeklenen özetleme işlemi ise aşağıdaki adımlarla gerçekleşmektedir [69]:

1. **Matris oluşturma**: İlk olarak her bir satırın girdideki bir kelimeyi, her bir sütunun ise girdideki bir cümleyi temsil ettiği bir kelime-cümle matrisi oluşturulur. Hücreler, kelimeler için hesaplanan tf-idf gibi önem değerleri ile doldurulur.
2. **SVD**: Tekil değer ayrışımı ile matris, denklem 2.3'teki hâle getirilir. Burada A, $m \times n$ boyutunda girdiyi tanımlayan matrisi ifade ederken; U, $m \times n$ boyutunda girdi matrisinin satırları temsil eden bir matrisi; Σ ise $n \times n$ boyutunda negatif olmayan önem değerlerinden oluşan diyagonal bir matrisi; V ile $m \times n$ boyutunda girdi matrisinin sütunlarını temsil eden bir matrisi içerir.

$$A = U\Sigma V^T \tag{2.3}$$

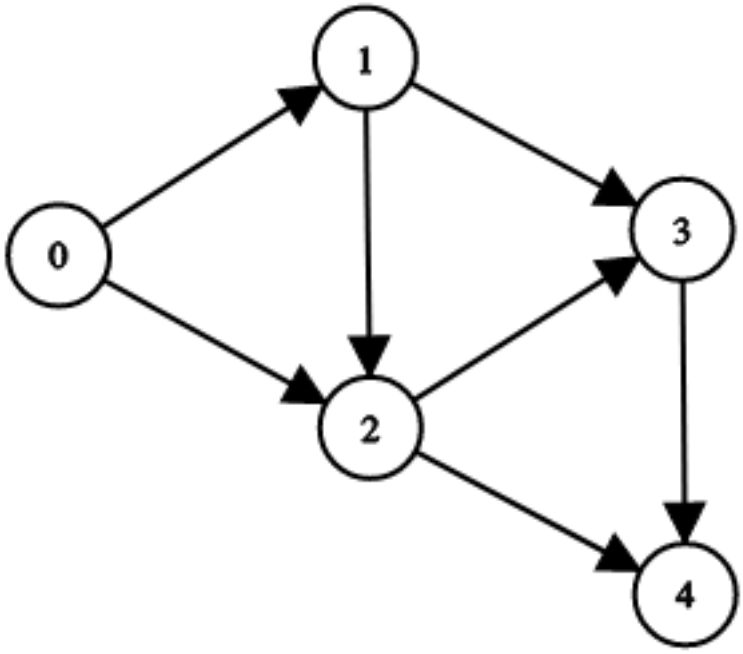
3. **Cümle seçimi**: Son olarak, genellikle girdiye göre belirlenen bir algoritma ile SVD ile elde edilen sonucun değerlendirilmesiyle, cümleler seçilir ve özetleme işlemi gerçekleştirilir.

2.5.2.3. TextRank

TextRank, metin özetleme veya anahtar kelimeleri çıkarmak için kullanılan, metin işleme için graf tabanlı bir başka sıralama modelidir [18]. Dil hakkında detaylı bir bilgiye ihtiyaç duymamasından ötürü dilden ve dolayısıyla alandan (*domain*) bağımsızdır [70]. Bu algoritma, bir zamanlar Google'ın sıralama algoritmasının önemli bir bölümünü oluşturan, aşağıdaki gibi ifade edilen PageRank algoritmasından esinlenilerek geliştirilmiştir [18].

$$S(V_i) = (1-d) + d \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \tag{2.4}$$

Denklem 2.4'te S ile web sayfasının ağırlığı ifade edilmektedir. In ile o sayfaya başka sayfalardan gelen linklerin sayısı; Out ile hesaplaması yapılan sayfadan başka sayfalara çıkan link sayısı; d ile sönümleme (*damping*) faktörü ifade edilmektedir. Sönümleme faktörü, hesaplaması yapılan sayfadan, başka sayfalara link olmaması durumunda işe yaramakta olan 0 ile 1 arasındaki bir değer olup, genellikle 0,85 değerini almaktadır.



Şekil 2.2: PageRank için beş adet sayfadan oluşan örnek bir graf.

Şekil 2.2'de gösterilen sayfalar ve sayfalara ait bağlantılar veya linkler, Tablo 2.1'de gösterildiği gibi, bir matris olarak ifade edilebilir. Burada her satırda 1 ile belirtilen, satırdaki

sayfadan, sütundaki karşılığına gelen sayfaya bir bağlantı olduğudur. Örnek olarak, 0 satırında 1 sütuna karşılık gelen değerin 1 olması, 0 sayfasından 1 sayfasına bir bağlantı olduğudur.

Tablo 2.1: Örnek graf için PageRank matrisi.

	0	1	2	3	4
0	0	1	1	0	0
1	0	0	1	1	0
2	0	0	0	1	1
3	0	0	0	0	1
4	0	0	0	0	0

PageRank'te her bir düğüm, bir sayfaya karşılık gelirken; TextRank'te her bir düğüm bir cümleyi temsil etmektedir [70]. TextRank'te hesaplanan benzerlik değeri, PageRank'teki link sayısına bağlı değerin yerine cümlelerin öneminin bulunmasında kullanılır [18].

S_i ve S_j şeklinde verilen, N_i dizisi halindeki kelimelerden oluşan ve $S_i = w_{1,}^i w_2^i, \dots, w_N^i$ şeklinde ifade edilen iki cümle arasındaki benzerlik denklem 2.5'deki gibi ifade edilebilir [18]:

$$Sim(S_i, S_j) = \frac{\left|\{W_k | W_k \in S_i \,\&\, W_k \in S_j\}\right|}{\log(|S_i|) + \log\left(\left|S_j\right|\right)} \tag{2.5}$$

Bu işlemin ardından, cümlelerin önem değerleri sıralandıktan sonra, en yüksek değere sahip olan belirli sayıdaki cümle seçilere özetleme işlemi tamamlanır [18].

Özetle, TextRank ile metin içerisindeki en önemli kısımların bulunması ile herhangi bir eğitim veri setine ihtiyaç duymadan, dilden bağımsız olarak özellikle haber gibi metinlerde özetleme işlemi gerçekleştirilebilir [70].

2.5.2.4. LexRank

LexRank, doğal dil işlemede cümlelerin ve diğer metin birimlerinin öneminin hesaplanmasında kullanılan graf tabanlı bir yöntem olup, çıkarıma dayalı özetlemede kullanılabilmektedir. TextRank'e benzeyen bu yöntemle özet oluşturmak için aşağıdaki adımlar izlenmektedir [17]:

1. Girdinin alınmasının ardından sözcükler, sözcük gömme (*word embedding*) adı verilen ve sözcüklerin vektör formatında gösterildikleri forma dönüştürülür. Bu gösterimde, birbirleri ile ilişkili olan kelimelerin vektörleri de birbirine benzemektedir.
2. Kosinüs benzerliği kullanılarak cümlelerin benzerliği hesaplanır.
3. Tüm cümlelerdeki benzerlikleri göstermek için bir bağlantı matrisi (*connectivity matrix*) oluşturulur.
4. En önemli cümleleri bulmak için özvektör merkeziliği (*eigenvector centrality*) kullanılır ve güç yineleme (*power iteration*) yöntemi ile skor hesaplaması yapılır.
5. En yüksek skora sahip cümlelerden özet oluşturulur.

2.5.2.5. BERT

Google mühendisleri tarafından, kullanıcı aramalarını daha iyi anlayabilmek için geliştirilen BERT, çeşitli NLP görevlerini gerçekleştirebilen, yayınlandığı ilk dönemde Wikipedia'daki 2,5 milyar kelime ve BookCorpus'taki 800 milyon kelime ile önceden eğitilerek, biri 12 katman ve 110 milyon parametreden oluşan, diğeriyse 24 katman ve 340 milyon parametreden oluşan iki farklı modeli sunulmuş bir dil modelidir [71].

Metin içerisindeki herhangi bir kelimenin hem sağındaki hem de solundaki kelimeyle ilişkisini algılayabilen BERT, farklı anlama sahip aynı kelimeler için farklı sözcük gömme gösterimini döndürebilme yeteneğine sahip olması sebebiyle, kelimeleri vektör uzayında ifade eden başka modellerden ayrışmaktadır.

BERT'in çalışma şekli aşağıdaki gibidir [71]:

1. Cümle modele girdiğinde, cümleyi oluşturan kelimelerin %15'i kadar kelime masked language modeling tekniği ile maskelenir ve ardından maskelenen kelimeler tahmin edilir. Maskeleme işlemi, %80 [MASK] simgesiyle, %10 rastgele bir kelimeleyle, %10 kelimede hiçbir değişiklik yapmayacak şekilde yapılır.

2. Cümlelerin arasındaki ilişkiyi anlayabilmek için, sonraki cümle tahmini (*next sentence prediction*) tekniği uygulanır. Bu teknikte, %50 oranında ikinci cümleler rastgele değiştirilir, %50 oranında değiştirilmeden bırakılır.
3. Bu iki teknik, eğitim kaybı minimum olacak şekilde uygulanır ve BERT kelimelerin ve cümlelerin ilişkisini en iyi şekilde öğrenir.

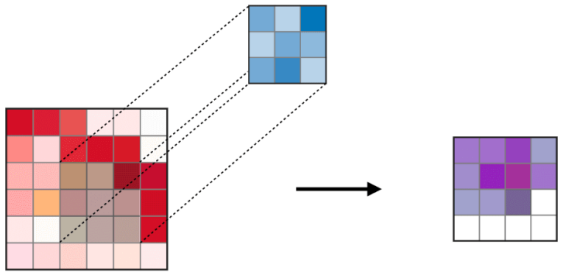
2.5.3. Yorumlamaya Dayalı Özetleme

Uzun bir süre boyunca, otomatik metin özetleme çıkarıma dayalı yöntemler aracılığıyla gerçekleştirilse de 2014 yılında bu durum değişmiş, yapay sinir ağları ile otomatik özetleme sistemlerinin geliştirilebileceği gösterilmiştir [65]. Yorumlamaya dayalı özetlemede, çıkarıma dayalı metin özetlemede olduğu gibi, metnin ana fikrini ifade eden, metnin önemli kısımlarını öne çıkaran bir özet sunulmaya çalışılmaktadır [72].

Sinir ağları ile çalışan birçok özetleyici algoritması geliştirilmiş olsa da bu sistemler genellikle aşağıdaki adımları takip ederek özetleme işlemini yapmaktadır [65]:

1. Sözcükler bir arama tablosu ile sözcük gömme vektörlerine dönüştürülür.
2. Cümleler, sözcük gömme vektörleri aracılığıyla sürekli vektörler olarak kodlanır (*encoding*). Bu aşamada genellikle evrişimli sinir ağları (CNN) veya tekrarlayan sinir ağları (RNN) kullanılır.
3. Cümle temsilleri daha sonra çıkarıma dayalı veya yorumlamaya dayalı çalışan bir model ile işleme alınır ve seçim ya da yorumlama için çözümleme (*decoding*) gerçekleştirilir.

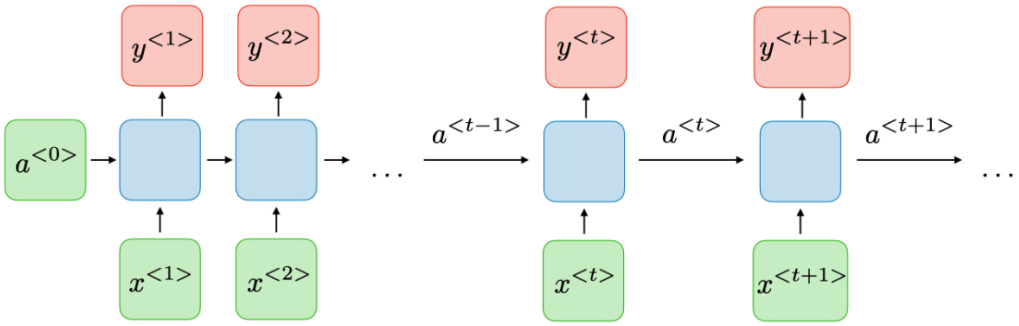
CNN'ler genellikle evrişim ve ortaklama katmanından oluşan sinir ağlarıdır [73]. Örnek bir Şekil 2.3'te örnek bir evrişim katmanı gösterilen CNN'ler, girdi dizilerine yerel bir evrişim filtresi uygulayarak otomatik özetleme işleminde kullanılırlar [65].



Şekil 2.3: CNN evrişim katmanı [73].

RNN'ler, önceki çıktıların girdi olarak kullanıldığı, herhangi uzunluktaki girdilerin hesaplanabileceği, genellikle doğal dil işlemede kullanılan yapay sinir ağlarıdır [73]. RNN'ler zamana bağlı sinir ağları sunabilmeleri sebebiyle, otomatik metin özetlemede sıklıkla kullanılırlar [65]. Benzer biçimde birçok makine çevirisi uygulamalarında da RNN kullanımına başvurulmaktadır [73].

Şekil 2.4'te örnek bir RNN yapısı görülebilir. Burada t ile zaman, a^t ile t zamanındaki aktivasyon, y^t ile t zamanındaki çıktı ifade edilmektedir.



Şekil 2.4: Örnek RNN yapısı [73].

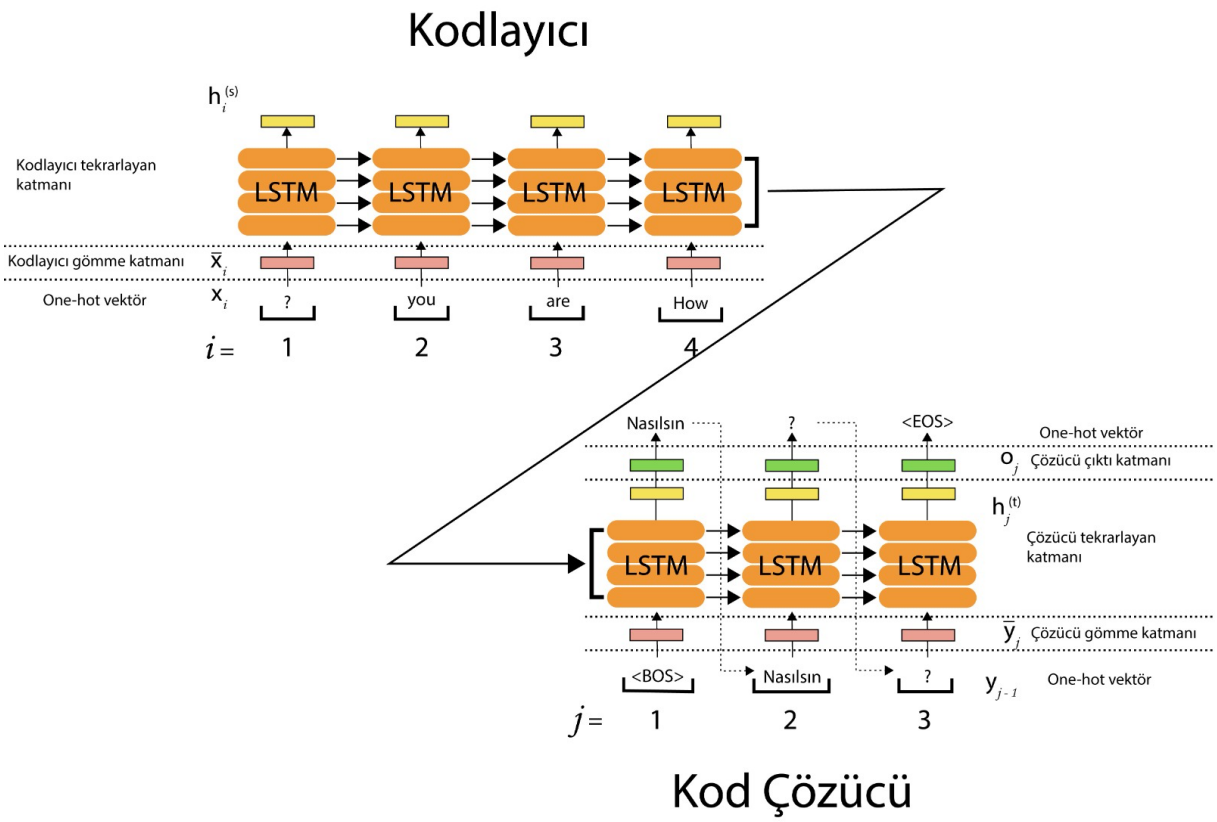
En basit RNN örneği Elman RNN'si olup Denklem 2.6'da gösterilmiştir [65]. Bu denklemde x_t giriş vektörüyken, h_t gizli katman vektörüdür. σ aktivasyon fonksiyonu olup, W parametre matrisidir.

$$h_t = \sigma(W_1 h_{t-1} + W_2 x_t) \tag{2.6}$$

2.5.3.1. Seq2Seq

Seq2Seq, doğal dil işleme işlemleri için kullanılan, Google tarafından ilk olarak İngilizceden Fransızca'ya makine çevirisi yapmak için geliştirilmiş bir modeldir [72]. Bu model için, kısaca RNN'ye dayalı kodlayıcı-kod çözücü bir yapı denilmektedir. [74]

Google mühendisleri, geri bildirim bağlarına sahip yapay bir RNN mimarisi olan LSTM kullanarak girdi dizilerini sabit boyuttaki bir vektöre çevirmiş, ardından vektördeki diziyi çözmek için de bir başka LSTM kullanarak Seq2Seq adını verdikleri bu modeli geliştirmişlerdir.



Şekil 2.5: Örnek Seq2Seq modeli [75].

Şekil 2.5'teki örnek bir Seq2Seq modelinde de olduğu gibi, Seq2Seq modelinde RNN'lerden oluşan kodlayıcı, girdi dizisinin içeriğini gizli durum vektörü ($h_i^{(s)}$) şeklinde yakalar ve bunu kod çözücüye gönderir. Ardından son karaktere gelene kadar kod çözücü her başarılı tahminin ardından bir önceki bilinmeyen durumu ($h_j^{(s)}$) kullanarak bir sonraki durumu tahmin etmeye çalışır ve son olarak çıktıyı (O_j) üretilir. Bu görevler sırayla olduğu için de modelin adı sequence

to sequence olarak belirtilmiştir. BOS (*Begin of Sequence*) vektörü kod çözücüde sıranın başlangıcını belirtirken, EOS (*End of Sequence*) vektörü kod çözücüde sıranın sonunu belirtir.

2.5.4. Otomatik Özetlemede Değerlendirme Yöntemleri

Özetlerin değerlendirilmesinde karşılaşılan bazı sorunlar vardır [62]:

- Özetleme işlemi, doğal dil işleme yöntemleri aracılığıyla oluşturulmuş bir çıktı içerir. Bu çıktı, bir soruya yanıt içeriyorsa doğruluğu ölçülebilir. Ancak, diğer durumlarda doğru çıktının ne olduğunu kestirmek zordur.
- Sistemin çıktısını değerlendirmesi için insanlara ihtiyaç duyulması, bu sistemlerin maliyetini artırabilir. Bu yüzden, insan değerlendirmesi yerine bir skor ile özetleme işleminin değerlendirilmesi daha tercih edilebilirdir.
- Özetleme işlemi sıkıştırma içerir. Bu nedenle farklı sıkıştırma oranlarında özetlemeyi değerlendirebilmek oldukça önemlidir. Bu durum ise ölçümü daha karmaşık hâle getirebilir.
- Bilgilerin özetleme sonunda kullanıcı ve uygulama ihtiyaçlarına duyarlı olacak biçimde sunulması gerektiğinden, ölçüm standartlaşmaktan çıkarak, daha karmaşık hâle gelebilir.

Günümüzde, özetleme sistemlerini değerlendirmek için hangi değerlendirme kriterinin kullanılması gerektiği konusundaki belirsizlikler devam etmektedir. Mevcut değerlendirme teknikleri içsel veya dışsal olarak ikiye ayrılmaktadır. İçsel yöntemlerde bir özetleme sisteminin sonucu doğrudan değerlendirilirken; dışsal yöntemlerde sistemin akışındaki her bir akışın performansı değerlendirilir. En yaygın içsel değerlendirme yöntemi özetleyici sistem tarafından oluşturulan özetin, insanlar tarafından oluşturulan referans özetler ile karşılaştırılarak değerlendirilmesiyken [65]; dışsal değerlendirme yöntemi olarak ROUGE ve BLEU yöntemleri öne çıkmaktadır [76].

2.5.4.1. ROUGE

Bir özetin kalitesini, genellikle bir insan tarafından üretilen ideal bir özetle karşılaştıran bir dizi yöntemden oluşan bir yazılım paketi olan ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*), özetlerin otomatik olarak değerlendirilmesine olanak tanımaktadır. ROUGE-N,

ROUGE-L, ROUGE-W ve ROUGE-S, ROUGE tarafından sağlanan çeşitli değerlendirme ölçütleridir [77].

İçsel değerlendirme yöntemi olan ROUGE ölçütlerinden ROUGE-N, ROUGE-L ve ROUGE-S'nin genişletilmişi olan ROUGE-SU, özetleme performansının ölçümünde en çok kullanılanlardır [65]:

- **ROUGE-N**: Sistem ve referans özetlerinin karşılaştırılması sonucunda, n-gram örtüşme yüzdesinin ölçümüdür. Yani, bu ölçümde, kelimelerin n-gram cinsinden ardışık eşleşmeleri gerekir. Örneğin ROUGE-1, n'nin bir olması sebebiyle sistem ve referans özetleri arasındaki her kelimenin örtüşmesini ölçerken; ROUGE-2, n'nin iki olması sebebiyle sistem ve referans özetleri arasındaki iki bitişik kelimeden oluşan gruplar anlamına gelen bigramların örtüşmesini ölçer.
- **ROUGE-L**: Karşılaştırma sonucunda, n-gram dizisinde en uzun ortak sonucun dikkate alındığı bu ölçümde, önceden belirlenmiş herhangi bir n değeri olmadan, cümle düzeyinde kelime sıraları dikkate alınarak referans ve sistem özeti karşılaştırılır.
- **ROUGE-SU**: Sistem ve referans özetlerinde, boşluğa olanak tanıyan ve skip bigram olarak tanımlanan, kelimelerin arasındaki boşluk miktarının değiştirilmesi ile oluşturulan kelime çiftleri ile tek kelimeden oluşan unigramların örtüşme yüzdesinin ölçümüdür. Bu yöntemde, boşluk miktarı genellikle dört ile sınırlandırılmaktadır.

2.5.4.2. BLEU

Çevirilerde, insanlar tarafından yapılan değerlendirmenin çok vakit almasından ötürü geliştirilen BLEU, dilden bağımsız olarak iyi bir değerlendirme sunan, otomatik bir değerlendirme yöntemidir. BLEU, belirli kelimelerin varlığına veya yokluğuna ve sırasına bakmakta, üretilen çıktının referans ile ne kadar örtüştüğünü dikkate almaktadır [78].

2.5.4.3. Pyramid

Bir dokümanın özetlenmesi sırasında referans olacak tek bir ideal kaynağın olamayacağını iddia eden Pyramid değerlendirme sistemi, içerikteki birimlerin anlamsal eşleşmelerine göre bir puanlama yaparken, birden fazla referans kaynağa ihtiyaç duymaktadır [79].

3. MALZEME VE YÖNTEM

Bu kısımda, otomatik metin özetleyici sistemi ile web uygulamasının nasıl geliştirildiği, otomatik metin özetleyicinin web uygulamasına nasıl entegre edildiği anlatılmaktadır. İşlemlerin gerçekleştirildiği geliştirme ortamına ait detaylar Tablo 3.1'de gösterilmiştir.

Tablo 3.1: Geliştirme ortamı ve kullanılan yazılıma ait bilgiler.

Ortam, Donanım veya Yazılım	**Değer**
Kullanılan Bilgisayar	MacBook Pro (2019)
İşlemci	1,4 GHz Quad-Core Intel Core i5
RAM	8 GB 2133 MHz LPDDR3
Grafik Kartı	Intel Iris Plus Graphics 645 1536 MB
İşletim Sistemi	macOS Big Sur 11.4
Makine Öğrenmesi Ortamı İşletim Sistemi	Ubuntu 20.04.2 LTS
Makine Öğrenmesi Ortamı RAM	64 GB
Makine Öğrenmesi Ortamı Grafik Kartı	NVIDIA QUADRO RTX 6000
Web Sitesi Sunucu İşletim Sistemi	Ubuntu 20.04.2 LTS
Web Sitesi Sunucu RAM	2 GB
Python	3.6.9
Numpy	1.19.5
Tensorflow	2.4.1
Keras	2.4.3
Nltk	3.5

Zemberek	0.17.1
nginx	1.18.0
PHP	7.4

3.1. OTOMATİK METİN ÖZETLEYİCİ

3.1.1. Veri Toplama

Bu çalışmada, veri setini oluşturması ve aralarında bir karşılaştırma yapılabilmesi için, veri kaynağı olarak, Wikihow'ın Türkçe versiyonu[10] ile popüler bir Türkçe haber sitesi olan TRT Haber[11] kullanılmıştır.

Şekil 3.1'de ana sayfası görülebilen Wikihow'ın seçilmesinin sebebi, Koupaee ve Wang'ın Wikihow'ın İngilizce versiyonunu kullanarak, çıkarıma dayalı özetleme işleminde oldukça başarılı sonuçlar elde etmeleridir [80]. Birden çok dilde hizmet veren bir site olarak Wikihow, herhangi bir şeyin nasıl yapıldığını, bir editöryel süreçten geçirerek kullanıcılarına sunan popüler bir İnternet sitesidir.

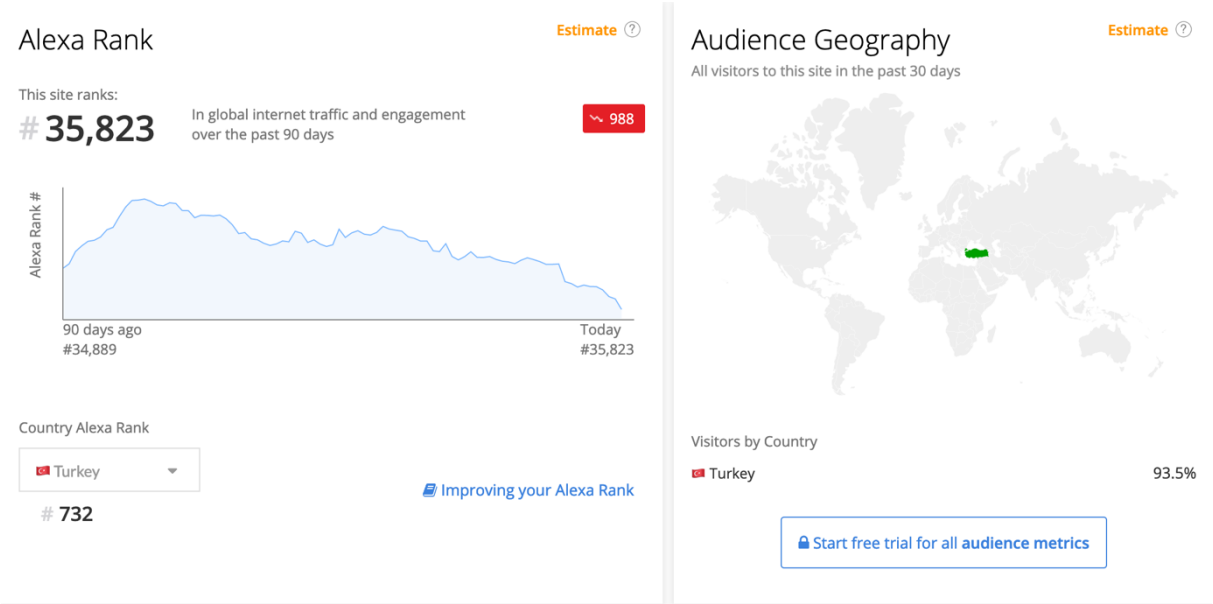


Şekil 3.1: Wikihow Türkiye ana sayfası, Şubat 2021.

Şekil 3.2'de görülebileceği gibi, Wikihow Türkiye, İnternet sitelerinin trafiklerini tahmini olarak ölçen bir şirket olan Alexa'nın sağladığı veriye göre Türkiye'nin en çok ziyaret edilen

[10] https://www.wikihow.com.tr
[11] https://www.trthaber.com/

1.000 sitesi arasında yer alırken; dünyada en çok ziyaret edilen ilk 40.000 site arasında bulunmakta; ziyaretçilerinin yüzde doksandan fazlasını Türkiye'den edinmektedir[12].



Şekil 3.2: Wikihow Türkiye Alexa ziyaretçi istatistikleri, Şubat 2021.

1 **Öksürük gibi solunum yolu belirtilerini kontrol et.** COVID-19 bir solunum yolu enfeksiyonu olduğu için, balgamlı ya da balgamsız öksürük sık görülen bir belirtidir. Ama, öksürük alerjilerin ya da başka bir solunum yolu enfeksiyonunun belirtisi de olabilir, o yüzden endişelenmemeye çalış. Öksürüğünün COVID-19'dan kaynaklanabileceğini düşünüyorsan doktorunu ara.[1]

- Hasta birinin yanında bulunup bulunmadığını düşün. Bulunduysan büyük ihtimalle bir hastalık kapmışsındır. Bu insanlar net bir şekilde hastaysa her şeyden önce onlardan uzak durmaya çalış.
- Öksürüyorsan bağışıklık sistemi zayıflamış ya da komplikasyonlar açısından yüksek risk altında olan kişilerden örneğin 65 yaş üstündekilerden, yeni doğanlardan ve çocuklardan, hamile kadınlardan ve bağışıklık sistemini baskılayan ilaçlar kullananlardan uzak dur.

Şekil 3.3: Wikihow Türkiye'de yer alan örnek bir paragraf.

Şekil 3.3'te görüldüğü gibi, Wikihow içeriklerinin paragraf başlangıcında, paragrafı özetleyen kısım bulunmaktadır. Bu durum neredeyse her Wikihow içeriğinde görülmektedir.

[12] https://alexa.com/siteinfo/wikihow.com.tr

Wikihow Türkiye'nin içerik listesi, Site Haritası[13] sayfası üzerinden tüm ziyaretçilere açıktır. Ayrıca, Wikihow'ın arama motorlarının örümceklerine (*crawler*) sunduğu genişletilebilir işaretleme dili (XML) formatındaki site haritasından da bu adresler görülebilmektedir.

Site haritası aracılığıyla Wikihow'da yer alan tüm Türkçe makalelerin linkleri bulunduktan sonra, web sayfalarından veri çekmeyi kolaylaştıran bir Python kütüphanesi olan beautifulsoup4[14] ile her bir sayfanın adresi, başlığı, genellikle yöntemleri belirten alt başlıkları, paragrafın özet kısmı ve paragrafın özet kısmı hariç olan kısmı bir Excel dosyasına, makine öğrenmesi yöntemleri ile analiz edilmesi için kaydedilmiştir.

Son aşamada, her sayfa için kaydedilen, robotlar ve insanlar dahil herkesin erişimine açık olan bu sayfalara ait veri bir dosyada birleştirilerek, makine öğrenmesi yöntemlerinde kullanılabilmesi için uygun hâle getirilmiştir.

Wikihow veri setine ilişkin istatistikler, Tablo 3.2'de görülebilir.

Tablo 3.2: Wikihow veri setine ait istatistikler.

Veri Seti İstatistiği	**Değer**
İçerik Sayısı (Tekil Başlık Sayısı)	3.200
İçeriklere Ait Toplam Metin Sayısı	32.841
Metin Cümle Sayısı	147.562
Özet Cümle Sayısı	32.849
Metin Kelime Sayısı	1.768.039
Özet Kelime Sayısı	204.526
Metin Tekil Kelime Sayısı	161.273
Özet Tekil Kelime Sayısı	32.550

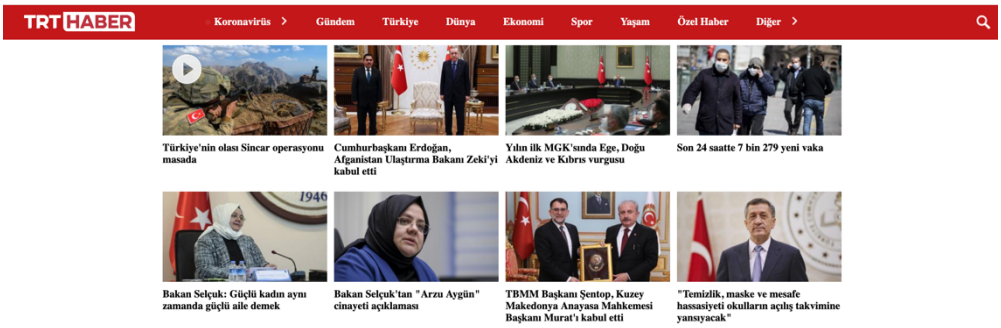
[13] https://www.wikihow.com.tr/%C3%96zel:Sitemap
[14] https://pypi.org/project/beautifulsoup4/

Wikihow veri setine ait örnek veri satırları ise Şekil 3.4'te gösterilmiştir.

Metin	Özet
Başarılı bir şirkete yatırım yapmak zengin olmak için kesin bir yoldur; ancak böyle bir şirket bulmak daha zordur. Eğer gerçekten inandığın bir şirket bulabilecek kadar şanslıysan, yatırım yapmadan önce araştırmanı yaptığından emin ol. Şirketin yönetimine inancın olması çok önemlidir. Konsept harika olsa bile kötü bir CEO şirketi batırabilir. Yatırım yapmadan önce şirketin giderleri, potansiyel gelirleri, markaları ve imajlarını çok iyi kavraman gerekir. Haklarını belirten açık bir anlaşmanın bulunduğundan emin ol. Ayrıca anlaşmayı feshetmeyle ilgili seçeneklerini de kavraman gerekir. Tüm paranı tek bir şirkete yatırma. Şirket değer kaybederse elinde hiçbir şey kalmaz.	Bir şirkete yatırım yap.
Uykusuzluk odaklanma yeteneğini olumsuz yönde etkileyebilir ve öğrenme isteğini düşürebilir. Her gece 7-8 saat uyumayı hedefle. Uykusuzsan dersler arasında 20 dakikalık güç şekerlemesi yapmaya çalış (programın müsaitse). Yatmadan önce bir veya iki saat boyunca telefonuna bakmaktan veya televizyon seyretmekten kaçın. Kitap okumak, ılık bir duş almak veya resim çizmek gibi rahatlatıcı bir şey yap. Uykusuz bir geceden sonra canlanmak için kafeine güvenmekten kaçın. Fazla miktarda kafein veya enerji içeceği içtikten sonra kendini daha enerjik hissedebilirsin, ancak altta yatan uyku yoksunluğu sebebiyle normalde olduğu kadar odaklanamazsın.	7-8 saat uyu.

Şekil 3.4: Wikihow veri seti örneği.

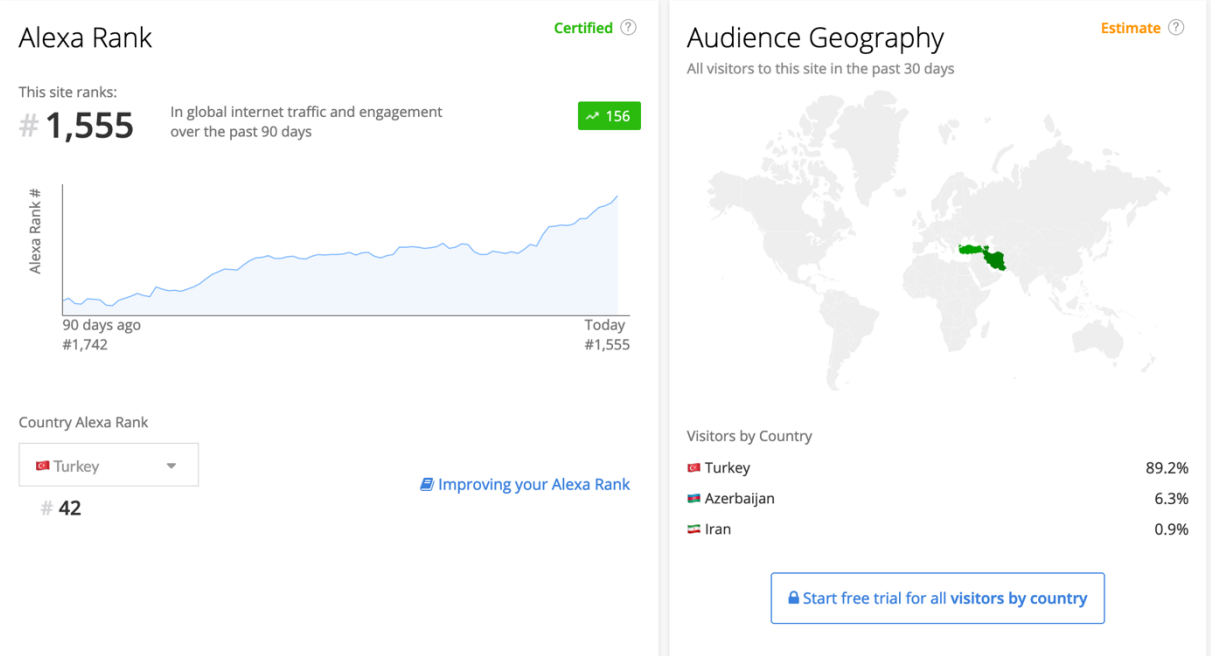
Ayrıca bu çalışmada TRT Haber'in İnternet sitesinde yer alan haberler de veri seti oluşturmak amacıyla kullanılmıştır. TRT Haber, Türkiye Radyo ve Televizyon Kurumu Haber Kanal Koordinatörlüğü tarafından yönetilen, "Türkiye'nin Haber Kaynağı" sloganı ile yayın yapan bir haber kanalı ve bir İnternet sitesidir [15]. Şekil 3.4'te örnek besleme sayfası[16] görülebilen TRT Haber tarafından yayınlanmış gündem haberleri, bu tezde referans içerik olarak kullanılmıştır.



Şekil 3.5: TRT Haber gündem haberleri.

TRT Haber, Şekil 3.5'ten de görülebileceği gibi, Alexa verisine göre Türkiye'nin en çok ziyaret edilen 50 İnternet sitesinden birisi olup; dünya genelinde en çok ziyaret edilen 1.600 İnternet sitesinden birisiyken; ziyaretçilerinin yüzde doksana yakının Türkiye'dendir[17].

[15] https://www.trthaber.com/iletisim.html
[16] https://www.trthaber.com/haber/gundem/
[17] https://www.alexa.com/siteinfo/haberler.com



Şekil 3.6: TRT Haber Alexa ziyaretçi istatistikleri, Şubat 2021.

Özetin ilk paragrafta, koyu harfler ile yazıldığı örnek bir TRT Haber haberi Şekil 3.6'da görülebilir.

Türkiye, Meis, Semadirek, Limni ve İpsara adalarının gayriaskeri statüsünün ihlal edildiğini bildiren yeni Navtex'ler yayımladı.

Yunanistan, gerilimin azaltılması için yapılan diyalog çabalarına rağmen uluslararası hukuk ve anlaşmalara aykırı adımlarına devam ediyor.

Atina yönetimi, Ege ve Akdeniz'deki adaların gayriaskeri durumunu hiçe saymayı sürdürüyor.

Türkiye, Yunanistan'ın tahriklerine yeni Navtex'ler ile cevap verdi. Semadirek, Limni ve İpsara adalarının 1923 Lozan Barış Antlaşması, Meis Adası'nın da 1947 Paris Antlaşması ile belirlenen gayriaskeri statüsünün ihlal edildiği vurgulandı.

Etiketler: Doğu Akdeniz | Navtex | Yunanistan

Şekil 3.7: TRT Haber örnek haber.

Veri kaynağı olarak TRT Haber'in seçilmesinin sebebi, haberlerinde özet kısmının yer alması ve web kazıma işlemlerini engelleyecek bir güvenlik mekanizmasının 2021 itibarıyla yer almamasıdır. Ayrıca, TRT Haber haberlerinde Türkçe dil bilgisi kurallarının titizlikle

uygulanması ve haberlerde kullanılan Türkçenin anlaşılır olması, TRT Haber'i diğer haber sitelerine kıyasla kendisini daha çok öne çıkarmaktadır.

TRT Haber, gündem, Türkiye, spor gibi haber kategori linklerini XML site haritası dosyası ile arama motoru örümceklerine sunmaktadır. Sayfalama yapılan bu kategori sayfaları aracılığıyla TRT Haber'de yayınlanan tüm haberlere erişebilmek mümkündür. Gündem kategorisi, bir Python kütüphanesi olan beautifulsoup4 ile taranmış; bu haberlerin ait olduğu sayfanın adresi, haber başlığı, haberin özet kısmı ve haberin özet kısmı hariç olan içerik kısmı bir Excel dosyasına, makine yöntemleri ile analiz edilmesi için kaydedilmiştir.

TRT Haber veri setine ilişkin istatistikler, Tablo 3.4'te görülebilir.

Tablo 3.3: TRT Haber veri setine ait istatistikler.

Veri Seti İstatistiği	**Değer**
İçerik Sayısı (Tekil Başlık Sayısı)	31.733
Metin Cümle Sayısı	702.344
Özet Cümle Sayısı	45.589
Metin Kelime Sayısı	10.935.507
Özet Kelime Sayısı	608.480
Metin Tekil Kelime Sayısı	510.688
Özet Tekil Kelime Sayısı	73.398

TRT Haber veri setine ait örnek satırlar Şekil 3.8'de görülebilir.

Metin	Özet
Yunanistan, gerilimin azaltılması için yapılan diyalog çabalarına rağmen uluslararası hukuk ve anlaşmalara aykırı adımlarına devam ediyor. Atina yönetimi, Ege ve Akdeniz'deki adaların gayriaskeri durumunu hiçe saymayı sürdürüyor. Türkiye, Yunanistan'ın tahriklerine yeni Navtex'ler ile cevap verdi. Semadirek, Limni ve İpsara adalarının 1923 Lozan Barış Antlaşması, Meis Adası'nın da 1947 Paris Antlaşması ile belirlenen gayriaskeri statüsünün ihlal edildiği vurgulandı.	Türkiye, Meis, Semadirek, Limni ve İpsara adalarının gayriaskeri statüsünün ihlal edildiğini bildiren yeni Navtex'ler yayımladı.
Cumhurbaşkanı Recep Tayyip Erdoğan, Suudi Arabistan Kralı Selman bin Abdülaziz ile bir araya geldi. Cumhurbaşkanı Erdoğan ile Kral Bin Abdülaziz, Cidde kentindeki Selam Sarayı'nda öğle yemeğinde bir araya geldi. Erdoğan ile bin Abdülaziz, daha sonra basına kapalı gerçekleşen ve Türk yetkililerin de hazır bulunduğu heyetler arası görüşmeye geçti.	Cumhurbaşkanı Recep Tayyip Erdoğan, Suudi Arabistan Kralı Selman bin Abdülaziz ile bir araya geldi.

Şekil 3.8: TRT Haber veri seti.

3.1.2. Veri Temizleme

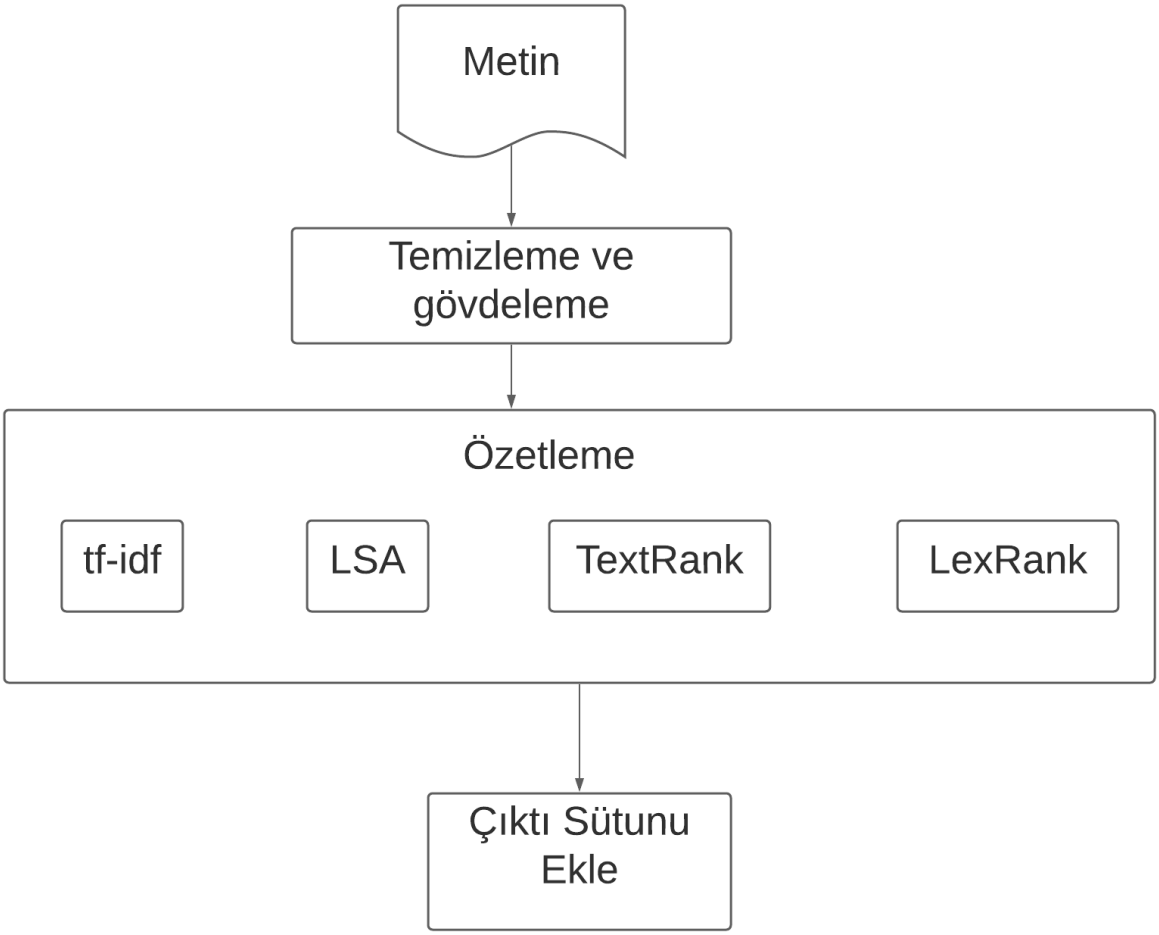
Genel olarak, veri temizleme ve hazırlamanın bir veri madenciliği çalışmasının yaklaşık %80'ini oluşturduğu düşünülmektedir [35]. Bu tezde, Wikihow ve TRT Haber veri setlerinin her ikisinde tab ile metne eklenen uzun boşluk karakteri ile yeni satıra geçilmesini sağlayan boşluk karakterleri kaldırılmıştır. Ayrıca *EK. 1*'de görülebilen etkisiz kelimeler, metinden temizlenerek kaldırılmıştır.

Yapılan incelemede TRT Haber veri setinde bazı haberlerin özet kısmının, haberin içeriğine yakın uzunlukta olduğu görülmüştür. Bunun gibi haber özetleme modelini geliştirmesi beklenmeyen habere ait satırların veri setinden çıkarılmasına karar verilmiştir. Ayrıca, yine bu veri setinde çok fazla gömülü tweet olduğu görülmüş; bu gömülü tweetler metinden kaldırılmıştır.

Wikihow veri setinde, "x kaynağı araştır" içeren kısımların yanında, köşeli parantez ve içerisindekiler içerikten ayıklanmıştır.

3.1.3. Veri Analizi

Temizlenerek veri analizi için ideal hâle getirilen Wikihow ve TRT Haber veri seti için, öncelikle tf-idf, LSA, TextRank ve LexRank ile özetleme işlemi yapılmıştır. Bu algoritmaların uygulanması sırasında gövdeleme (*stemming*) için Zemberek kullanılmıştır. Bu algoritmaların uygulanması sonucunda elde edilen özetler veri setine yeni bir sütun olarak eklenmiştir. Bu süreç, Şekil 3.9'da gösterilmiştir.



Şekil 3.9: Veri analizi.

Makine öğrenmesi ile yapılan özetlemeler, Tablo 3.1'de belirtilen makine öğrenmesi ortamında yapılmıştır. Bunun sebebi, model eğitimi esnasında GPU kullanımının bu süreci oldukça kısaltmasıdır. Örneğin GPU'suz bir ortamda yaklaşık 10 saatten fazla süren bir model iyileştirme süreci, GPU ile bir saatten daha kısa bir sürede tamamlanabilmektedir.

Yorumlamaya dayalı makine öğrenmesi yöntemi olarak seçilen Seq2Seq işlemi için, özetlemeden önce tüm karakterler küçük hâle getirilmiş, ardından her bir veri seti için aşağıdaki adımlar izlenmiştir:

1. Kelime dağılımları incelenerek, özet ve metin için maksimum kelime uzunlukları belirlenmiştir.
2. Veri seti test ve referans için ayrılmıştır. Referans veya eğitim için edilen oran %90 olurken, test için tercih edilen oran %10 olmuştur.

3. İçerik cümlelere ayrılmıştır (*tokenize*). Bu şekilde metin vektör hâline getirilebilmiştir.
4. Seq2Seq için listedeki her sıranın aynı uzunluğa sahip olabilmesi için, en uzun olan referans alınarak, her bir sıra başına yeteri kadar 0 eklenerek eşit hâle getirilmiştir.
5. Seq2Seq modeli, 100 boyuttan oluşan bir gömme katmanı, üçer adet LSTM katmanından oluşan kodlayıcı ve kod çözücü ile tanımlanmıştır. Kod çözücüde Attention mekanizması kullanılmıştır. Attention, performansı artıran bir derin öğrenme mekanizmasıdır [81].
6. GPU'suz, Wikihow veri seti 3 saat, TRT Haber veri seti 18 saat sonunda model oluşturmuştur.
7. Oluşturulan modelle özetleme işlemi yapılmıştır.

3.2. WEB UYGULAMASI

Web uygulaması, İnternet üzerinden bir Web tarayıcısı ile çalıştırılan bir uygulamadır [82]. Geçmişte, bu kavramın ne olduğu konusunda birçok tartışma olmuş; kimileri bir web sunucusu ile çalışan bir uygulama olarak tanımlarken, kimileri de bir programlama dili ile çalışan bir uygulama olarak tanımlamıştır [83]. Daha önceden sadece bilgi vermek için oluşturulmuş statik sayfalardan oluşan web sayfaları, günümüzde kullanıcıların etkileşimleri ile dinamik hâle gelmiş; web uygulamalarının sınırlarını gelişen tarayıcı teknolojileri belirler hâle gelmiştir [82].

Bu çalışmada modern web geliştirme yöntemleri ile ölçeklenebilir, birden çok isteği aynı anda işleyebilen, hızlı çalışması hedeflenen bir web uygulaması geliştirilmiş; bu web uygulaması ozetleyici.com alan adı üzerinden herkese açık bir şekilde yayına alınmıştır.

3.2.1. Tasarım

Bu web uygulamasının, arka uç (*backend*) ve ön uç (*frontend*) olarak iki farklı katmanı bulunmaktadır.

Arka uç kısmı, uygulamanın sunucu tarafı olup, kullanıcının görmediği, ancak uygulamanın düzgün çalışabilmesini sağlayan, hiper-metin transfer protokolü (HTTP) istekleri ile verinin işlendiği; ayrıca özetlemenin yapıldığı temel katmandır.

Ön uç kısmı, web uygulamasını ziyaret eden kullanıcıların gördüğü kısım olup, verinin hiper metin işaretleme dili (HTML), basamaklı stil şablonları (CSS) ve JavaScript ile kullanıcılara

yansıtıldığı kısımdır. Bu kısımda kullanıcılar, veriyi görüntüleyerek, veri ile etkileşime girebilirler. Bu uygulamadaki ön uç kısmında, kullanıcılar özetlemek istedikleri metni bir girdi (*input*) aracılığıyla sisteme yollayarak, sistem tarafından yapılan özeti görebilmektedir.

3.2.1.1. Sunucu Yapısı

Bu web uygulaması, aynı anda gelen isteklere optimum şekilde cevap verebilmesi için modern web sunucu teknolojileri ile geliştirilmiştir.

Türkçe bir uygulama olduğu için, ziyaretçilerinin çoğunun Türkiye'den gelmesi beklenen bu web uygulamasında, isteklere verilecek cevapların gecikmesini minimumda tutabilmek için Türkiye'deki sunucular tercih edilmiştir.

Web uygulamasında sunucu için seçilen işletim sistemi, Bionic Beaver kod adına sahip Ubuntu 20.04.2 olmuştur. Kelime olarak, Afrika'da başkalarına insanlık anlamına gelen Ubuntu, Linux bazlı açık kaynaklı bir işletim sistemi olup, 2004'ten beri aktif olarak geliştirilmektedir[18]. Bu uygulamada Ubuntu kullanılmasının sebebi, Ubuntu'nun web sunucuları için oldukça ideal olması ve sürekli olarak geliştirilerek, kararlı sürümlerindeki her türlü güvenlik sorununun çeşitli yamalarla düzeltilmesidir. Bionic Beaver'ın seçilmesinin sebebiyse, Ubuntu'nun her iki yılda bir en az beş yıl boyunca desteklediği LTS olarak adlandırılan kararlı sürümleri yayınlamasıdır. Stabil bir sürüm olan Ubuntu 20.04.2'in, 2023'e kadar bakım güncellemeleri, 2030'a kadar güvenlik güncellemeleri alması beklenmektedir[19].

Ubuntu, yüklendiği anda birçok yazılımla gelmiştir. Birçok ek yazılım paketi de Ubuntu Yazılım Merkezi veya Ubuntu'nun paket yönetim aracı olan APT bazlı paket yükleme araçlarıyla yüklenebilmektedir. Bu web uygulamasında, sunucunun daha iyi ve daha güvenli çalışabilmesi için, fail2ban gibi bazı ek yazılım paketleri yüklenmiştir. Ayrıca, Ubuntu'nun karmaşık olmayan güvenlik duvarı (ufw) uygulaması aktifleştirilmiş, 22, 80 ve 443 numaralı kapılar (*port*) haricinde erişim engellenmiştir.

Gelen HTTP isteklerini yönetmesi için seçilen web sunucusu nginx olmuştur. nginx, uygulama ve içerik dağıtımını hızlandıran, açık kaynaklı, güvenli bir web sunucusu olup, İnternet'teki

[18] https://ubuntu.com/about
[19] https://ubuntu.com/about/release-cycle

başka birçok İnternet sitesi tarafından da kullanılmaktadır[20]. Ubuntu sunucuya yüklenen nginx, Ubuntu paketleri ile değil, ancak kaynağından derlenerek yüklenmiştir. Kaynağından yükleme işlemi, nginx'in kaynak kodlarının derlenmesi ile olmaktadır. Derleme işlemi esnasında, nginx için geliştirilmiş bir eklenti olan PageSpeed eklentisi kullanılmıştır. PageSpeed[21], yüksek performanslı web uygulamaları geliştirmek isteyenler için Google mühendisleri tarafından geliştirilmiş, web sayfasını kullanıcıya göstermeden önce, sayfanın kaynak kodlarında çeşitli iyileştirmeler yaparak, sayfaların daha hızlı yüklenmesini sağlayan bir web teknolojisi olup; nginx haricinde, bir başka web sunucusu olan Apache HTTP sunucusu için de kullanılabilir modern bir teknolojidir.

3.2.1.2. Arka Uç Yapısı

Arka uçta PHP ve Python programlama dilleri kullanılmıştır.

Web geliştirme için oldukça uygun bir betik dili olan PHP, sunucu tarafında çalıştırılmakta olup HTML üretmek için kullanılabilmektedir[22]. Bu uygulamada PHP kullanılmasının sebebi, PHP tabanlı, açık kaynaklı bir web uygulama çatısı (*framework*) olan Laravel'in kullanılmasıdır. Laravel, kimlik doğrulama (*authentication*), önbelleğe alma (*caching*), kuyruğa görev ekleme (*task queues*) ve siteler arası istek sahteciliği (*cross site request forgery*) gibi çeşitli güvenlik sorunları için birçok çözüm sunmakta, uygulama işlevselliğinden ödün vermeden geliştirme sürecini hoş hâle getirmeyi amaçlamaktadır[23].

3.2.2. Uygulama

Kullanıcılar, daha önceden derlenmiş BERT modeli aracılığıyla veya TextRank yöntemiyle özetleme yapabilmekte, özetin kaç cümleden oluşabileceğini seçebilmektedir. Uygulamada arayüz üzerinden yapılan geçerli özetleme istekleri, eşzamansız JavaScript ve XML (AJAX) teknolojisi ile Laravel uç yapısına gönderilmekte, burada çalışan PHP kodu, özetlenecek içeriği bir metin dosyası olarak sunucuya kaydetmektedir.

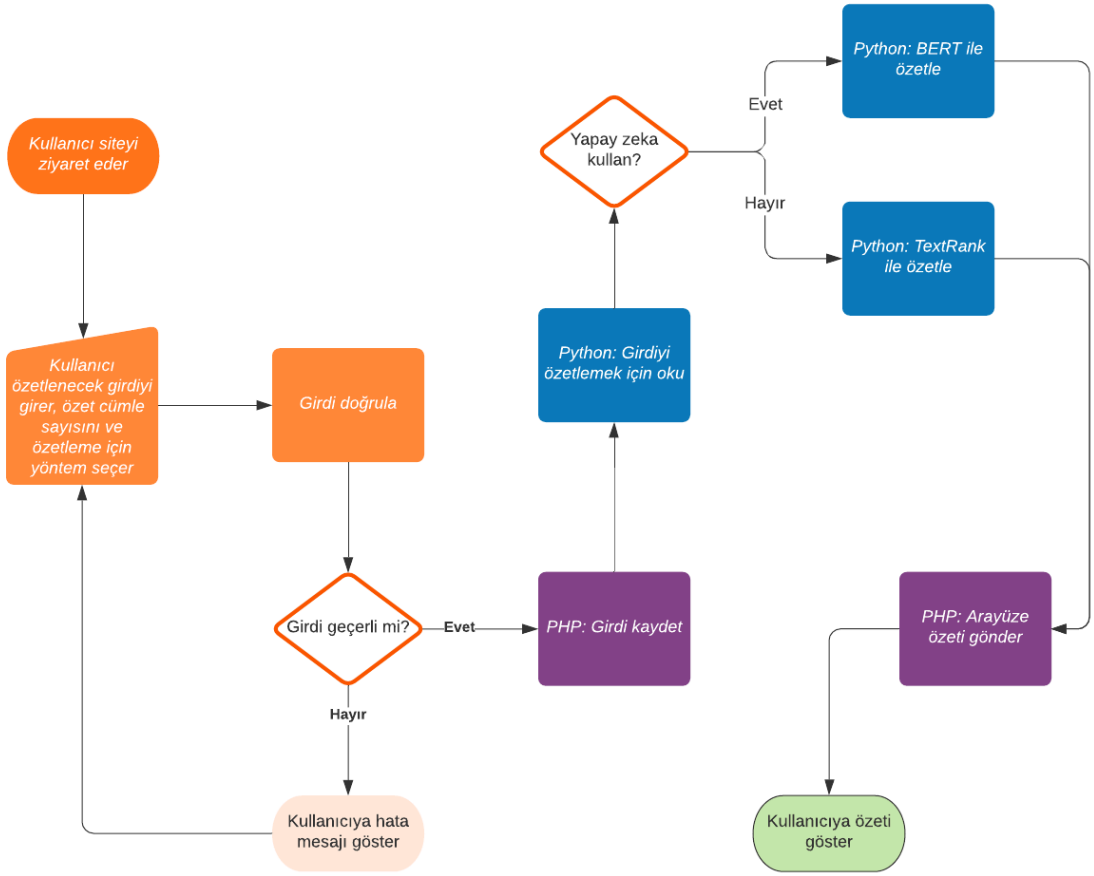
---

[20] https://www.nginx.com
[21] https://developers.google.com/speed/pagespeed/module
[22] https://www.php.net/manual/en/intro-whatis.php
[23] https://laravel.com/docs/master

Sunucuya bu dosyanın kaydedilmesinin ardından, bir tetikleyici aracılığıyla özetleme işlemini yapan Python dosyası, kaydedilen dosyadan girdiyi okuyup, kullanıcı tarafından tercih edilen özetleme yöntemiyle özetleme işlemini gerçekleştirmektedir.

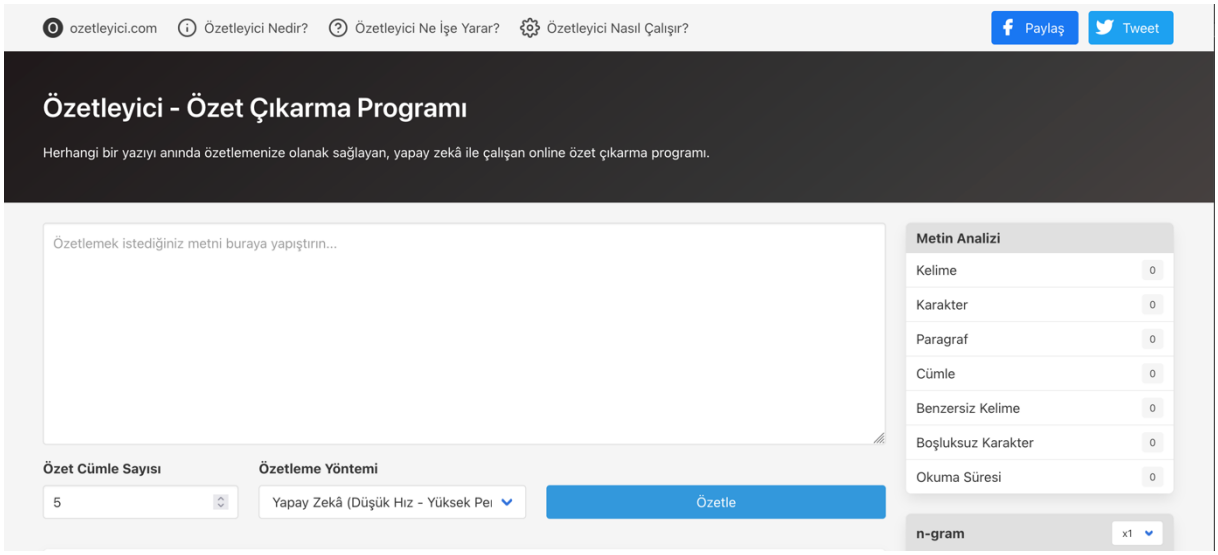
Son olarak, özetleme işleminin tamamlanmasının ardından, yeniden çalışan tetikçinin Laravel'e özetlenen içeriğin konumunu belirtmesinin ardından Laravel özet içeriği alarak, sunucu tarafından ara yüze JavaScript nesnesi gösterimi (JSON) dosya türünde cevap dönmekte ve bu cevap tarayıcıda özet olarak kullanıcıya sunulmaktadır.

Uygulamaya ait tüm bu akış Şekil 3.10'da görülebilir.



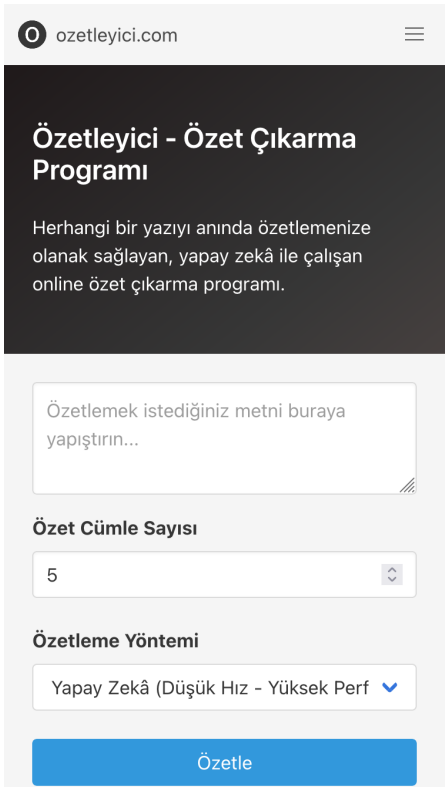
Şekil 3.10: Web uygulaması akışı.

Geliştirilen uygulamanın masaüstü bilgisayarlarda görülen arayüzü Şekil 3.11'de görülebilir.



Şekil 3.11: Web uygulaması masaüstü arayüzü.

Geliştirilen uygulamanın mobil cihazlar için gerçeklenen arayüzü Şekil 3.12'de görülebilir.



Şekil 3.12: Web uygulama mobil arayüzü.

4. BULGULAR

Yapılan çalışmada Wikihow veri seti ile elde edilen Seq2Seq özetlerine örnekler Şekil 4.1'deki gibidir. Wikihow veri setinde Seq2Seq modelinin oldukça başarılı olduğu, küçük anlamsal sorunlar haricinde mantıklı özetler ürettiği görülebilmektedir.

Metin	Beklenen Özet	Elde Edilen Özet
dişinde ağrı hassasiyet varsa tedavi edilmesi gereken bir sorunun işareti olabilir diş hekiminle bağlantıya geç diş muayenesi tedavisi randevu al	bir sorun olduğundan diş hekiminle bağlantıya geç	bir doktora görün
lcd ekranı silmek sıvı ortadan kaldırabilir ekranı alkole kaçın ekrana kalıcı hasar verebilir	izopropil alkolle sil	bir mikrofiber bezle kurula
küf nemli ılık tercih eder nem gidericiler havadaki nemi gidererek küfün yayılma azaltır	bir nem giderici kullan	bir nemlendirici sür
günümüzde büyük övgü alan tarifi denemek isteyen herkes faydalıdır îstersen sen dene midenin ihtiyacı tam olabilir bir çorba kaşığı balı bir bardak su içerisinde bir çorba kaşığı elma sirkesi karıştırmayı dene	elma sirkesi bal hazırla	bir yüz maskesi kullan
kurulum dosyası üzerinde bir exe exe uzantısı olacaktır dosya muhtemelen setup olarak adlandırılmıştır bir kere tıkladığında dosya seçilir	kurulum dosyasına bir tıkla	kurulum dosyasına çift tıkla

Şekil 4.1: Wikihow veri seti için Seq2Seq özetleri.

Yapılan çalışmada TRT Haber veri seti ile elde edilen Seq2Seq özetlerine örnekler Şekil 4.2'deki gibidir. Bu özetlerin kısmen başarılı olduğu görülse de beklenen özetteki kelime sayısı arttıkça, elde edilen özetin başarısının düştüğü görülebilmektedir.

Metin	Beklenen Özet	Elde Edilen Özet
orgeneral akar beraberindeki askeri heyetle sınır birliklerinde incelemelerde bulunarak bilgi aldı genelkurmay başkanı akar hakkari'deki birlikleri denetledi îlk olarak öncüpınar gümrük kapısı yakındaki birlikleri denetleyen akar bölgedeki gelişmelere ilişkin bilgi aldı akar buradaki incelemenin ardından beraberindekilerle helikopterle gaziantep'in îslahiye ilçesine hareket etti	genelkurmay başkanı orgeneral hulusi akar kilis'te sınır birliklerini bilgi aldı	genelkurmay başkanı orgeneral hulusi akar beraberindeki kuvvet komutanlarıyla birliklerde incelemelerde bulundu
îçişleri bakanı süleyman soylu bitlis 6 teröristin jandarma sîha ile etkisiz hale getirildiğini açıkladı soylu sosyal medya hesabından yaptığı paylaşımda şunları söyledi çoğu gitti azı kaldı jandarma silahlı însansız hava aracı j sîha ile bitlis 6 terörist etkisiz îstihbarat operasyon tam başarı	bitlis jandarma silahlı însansız hava aracı ile 6 terörist etkisiz hale getirildi	îçişleri bakanı soylu hakkari'nin çukurca ilçesinde pkk'lı teröristin etkisiz hale getirildiğini açıkladı
dışişleri bakanı mevlüt çavuşoğlu avrupa birliği ab dış îlişkiler ve güvenlik politikası yüksek temsilcisi josep borrell ile telefonda görüştü diplomatik kaynaklardan edinilen bilgiye göre bakan çavuşoğlu ile görüşmesinde doğu akdeniz konusu ele alındı	dışişleri bakanı çavuşoğlu ab komisyon başkan yardımcısı ile görüştü	dışişleri bakanı mevlüt çavuşoğlu avrupa birliği ab dış îlişkiler ve güvenlik politikası yüksek temsilcisi ile görüştü
aile ve sosyal politikalar bakan yardımcılığına mehmet atandı resmi gazete'de yayımlanan atama kararına göre açık bulunan birinci derece kadrolu aile ve sosyal politikalar bakan yardımcılığına ürün güvenliği ve denetimi genel müdürü mehmet atanması uygun görüldü söz konusu atama 657 sayılı devlet memurları kanunu'nun 59 ve 74'üncü maddeleri ile 2451 sayılı bakanlıklar ve bağlı kuruluşlarda atama usulüne îlişkin kanun'un 2'nci maddesi gereğince yapıldı	aile ve sosyal politikalar bakan yardımcılığına mehmet atandı	aile çalışma ve sosyal hizmetler bakanlığı ile bazı atama kararları resmi gazete'de yayımlandı
bakanlar kurulu başbakan binali yıldırım başkanlığında toplandı çankaya köşkü'ndeki toplantı saat 14 05'te başladı	bakanlar kurulu başbakan binali yıldırım başkanlığında toplandı	başbakan binali yıldırım başkanlığındaki bakanlar kurulu toplantısı sona erdi

Şekil 4.2: TRT Haber veri seti için Seq2Seq özetleri.

Yapılan bu çalışma sonucunda, her bir yöntemin aynı haber için ürettiği haber özetinin bir örneği Şekil 4.3'teki gibidir. BERT ve Seq2Seq her özette daha başarılı olarak öne çıkmaktadır.

Beklenen	seq2seq	BERT	LexRank	TextRank	LSA	tf-idf
Şırnak'ın Cudi Dağı bölgesinde icra edilen operasyonda 21 bölücü terör örgütü PKK işbirlikçisi gözaltına alındı	şırnak'ın ilçesinde terör örgütü pkk'ya yönelik operasyonda çok sayıda silah ve mühimmat ele geçirildi	Şırnak'ın Cudi Dağı bölgesinde icra edilen operasyonda 21 bölücü terör örgütü PKK işbirlikçisi gözaltına alındı	Şırnak'ın Cudi Dağı bölgesinde icra edilen operasyonda da 21 PKK işbirlikçisi yakalandı	Şırnak'ın Uludere ilçesindeki temas aramasında bulunan bir el yapımı patlayıcı ile Mardin'in Dargeçit ilçesinde tespit edilen büyük tüpe, yaklaşık 50 kilogram amonyum nitratla hazırlanmış patlayıcı imha edildi	Türk Silahlı Kuvvetlerinden (TSK) yapılan bilgilendirmeye göre, Şırnak'ın Beytüşşebap ilçesi İncebel (Kato) Dağı bölgesindeki operasyonda iki mağara tespit edildi	Kilis'te, Suriye'den Türkiye'ye yasa dışı yollardan giren yabancı uyruklu 5 DEAŞ mensubu yakalandı
Cumhurbaşkanı Erdoğan, Bahçeli'yi evinde ziyaret etti	cumhurbaşkanı recep tayyip erdoğan mhp genel başkanı devlet bahçeli ile görüştü	Cumhurbaşkanı Recep Tayyip Erdoğan, MHP Genel Başkanı Devlet Bahçeli ile bir araya geldi	Cumhurbaşkanı Recep Tayyip Erdoğan, MHP Genel Başkanı Devlet Bahçeli ile bir araya geldi	Erdoğan, Bahçeli ile selamlaştıktan sonra yağmur nedeniyle şemsiyesini alarak "Elhamdülillah rahmetle geldik" dedi	Cumhurbaşkanı Recep Tayyip Erdoğan, MHP Genel Başkanı Devlet Bahçeli ile bir araya geldi	Basına kapalı gerçekleşen ziyaret yaklaşık 1 saat 10 dakika sürdü
İçmek için sade bir votka seç	bir böcek ilacı seç	Votka meraklılarının çoğu saf hâlde votka içmenin bu içeceğin tadını çıkarmanın en uygun yolu olduğuna inanmaktadır.	Tahıl bazlı votka yumuşak ve meyvemsi bir tada sahipken sebze bazlı votkanın tadı sert veya ilaç gibi olabilir.	Tahıl bazlı votka yumuşak ve meyvemsi bir tada sahipken sebze bazlı votkanın tadı sert veya ilaç gibi olabilir.	Çoğu insan İstanblue, Binboa, Grey Goose, Absolut, Smirnoff, Titos, Ketel One veya Stolichnaya'nın tadını tercih eder.	Eğer votka damağını yakarsa büyük olasılıkla kalitesi düşüktür.
Düzenli olarak egzersiz yap	her gün egzersiz yap	Egzersiz rutinine baş aşağı durmayı da kat.	Baş aşağı durmak, yüz kaslarını gevşemeye de zorlar ve böylece karışıklığa yol açan baskıyı azaltır.	Düzenli, orta tempolu egzersiz, genel olarak sağlıklı bir hatay tarzının önemli bir parçasıdır ve genel sağlığını iyileştirmek, tüm vücudundaki dolaşımı ve alnınla yüzünün geri kalanı dâhil, cildinin esnekliğini artırabilir.	Düzenli, orta tempolu egzersiz, genel olarak sağlıklı bir hatay tarzının önemli bir parçasıdır ve genel sağlığını iyileştirmek, tüm vücudundaki dolaşımı ve alnınla yüzünün geri kalanı dâhil, cildinin esnekliğini artırabilir.	Düzenli, orta tempolu egzersiz, genel olarak sağlıklı bir hatay tarzının önemli bir parçasıdır ve genel sağlığını iyileştirmek, tüm vücudundaki dolaşımı ve alnınla yüzünün geri kalanı dâhil, cildinin esnekliğini artırabilir.
Bütün işi kendin yapmak yerine hoşlandığın kişinin senin peşinden koşmasını sağla	ona bir dinleyici ol	Flört etmenin amacı karşındaki kişiye ondan hoşlandığını belli etmek de olsa her şeyi sen yapıyormuşsun gibi gözükmemelisin	Aşağıda nesnel ve öznel iltifatlara örnekler bulabilirsin	İlgini belli edecek kadar onun hoşuna gidecek şeyler söylemeli ama onu ne kadar umursadığın konusunda merakta bırakmalısın	Fakat romantik iltifatlar kullanırken en çok yapılan hata, sürekli "sevdiğin şeyi buraya yaz hoşuma gidiyor/seviyorum" tarzı kelimeleri kullanmaktır	Harika gözlerin var. Çok güzeller

Şekil 4.3: Farklı yöntemlerle haber özetleri.

Yapılan tüm bu özetleme işlemlerinde, Wikihow veri seti için yöntemlerin elde ettiği ROUGE skorları ise Tablo 4.1'deki gibi olmuş; her bir değerlendirme yöntemi için Seq2Seq en iyi skoru elde etmiştir.

Tablo 4.1: Wikihow veri seti için ortalama performans skorları.

	Seq2Seq	**BERT**	**LexRank**	**TextRank**	**LSA**	**tf-idf**
Rouge-1	**0,16**	0,14	0,11	0,1	0,09	0,05
Rouge-2	**0,05**	0,05	0,03	0,03	0,03	0,01
Rouge-L	**0,16**	0,13	0,1	0,09	0,08	0,04
BLEU	**3,38**	1,04	0,85	0,68	0,65	0,43

Yapılan tüm bu özetleme işlemlerinde, TRT Haber veri seti için yöntemlerin elde ettiği ROUGE skorları ise Tablo 4.2'deki gibi olmuş; Seq2Seq ROUGE değerlendirmesinde en iyi skorları alsa da BERT en iyi BLEU skorunu elde etmiştir.

Tablo 4.2: TRT Haber veri seti için ortalama performans skorları.

	Seq2Seq	**BERT**	**LexRank**	**TextRank**	**LSA**	**tf-idf**
Rouge-1	**0,32**	0,29	0,26	0,21	0,22	0,07
Rouge-2	**0,19**	0,19	0,17	0,13	0,13	0,03
Rouge-L	**0,3**	0,25	0,23	0,18	0,19	0,06
BLEU	8,59	**9,74**	9,23	6,34	6,38	1,48

5. TARTIŞMA VE SONUÇ

Bu bölümde elde edilen sonuçlar değerlendirilmiş, elde edilen sonuçlar literatür ile karşılaştırılmış, gelecek çalışmalar için çeşitli önerilerde bulunulmuştur.

5.1. SONUÇLARIN DEĞERLENDİRİLMESİ

Bu tez çalışmasında, insanlardan çok, arama motorları için içerik üreten haber sitelerinde yer alan Türkçe metinleri özetleyebilen, nginx sunucuda çalışan, PHP tabanlı bir web sitesi geliştirilmiş, ozetleyici.com adresinden erişilebilen, kullanıcıların özet cümle sayısını ve iki ayrı özetleme yönteminden birisini seçebilecekleri bir İnternet sitesi kullanıcıların hizmetine açılmıştır.

Python ile geliştirilmiş özetleyici sistemin oluşturulması aşamasında, yakın zamanda doğal dil işleme literatürüne katılmış olan, daha önceden yapılmış Türkçe metin özetleme tezlerinde yer almayan, BERT ve Seq2Seq modellerinin performansları, metin özetlemede kullanılan eski yöntemlerden tf-idf, gizli anlam analizi, TextRank ve LexRank ile karşılaştırılmıştır.

Literatürde görüldüğü üzere, dünyanın en çok ziyaret edilen haber siteleri olan CNN ve Daily Mail veri seti kullanılırken [20, 21, 23], Koupaee ve Wang [80] Wikihow veri setini literatüre kazandırmıştır. Ulusal çalışmalarda farklı haber sitelerinden derlenen ve birçoğu 200'den az farklı haberden oluşan Türkçe haber metinleri [27, 29, 30, 31] ile Vikipedi'de yer alan içeriklerden [28] özetleme yapılmıştır. Türkiye'de yapılan tez çalışmalarında ise, benzer biçimde, çoğu 200'den az satırdan oluşan ve çeşitli haber sitelerinden derlenen haberlerin [1, 14] yanında, elle hazırlanmış veri seti [2, 3], Türkçe ve İngilizce haberlerden oluşan veri seti [4, 8, 13], DUC 2004 veri seti [5, 7, 12] ile haber ve bilimsel makalelerden oluşan veri setleriyle [6] çalışmalar yapılmıştır. Bu çalışmada, daha önceden yapılan birçok çalışmanın aksine, herhangi bir hazır veri seti kullanılmamış; web kazıma ile Türkiye'nin en çok ziyaret edilen haber sitelerinden olan TRT Haber'deki haberlerden ve Wikihow'daki Türkçe içeriklerden oluşan, her biri 31.000'den fazla satırdan oluşan yeni veri setleri oluşturulmuştur. Türkçe literatürde daha önceden hakkında metin özetleme için çalışma yapılmamış olan, bu tezde karşılaştırma için kullanılmış Seq2Seq modelinin eğitimi için oldukça fazla veriye ihtiyaç duyulduğu için ve mevcut veri setlerinin bu modelin ihtiyaç duyduğu veri miktarını karşılamaması sebebiyle, veri seti bu şekilde oluşturulmuştur.

Uluslararası yayınlarda, geçmiş yıllarda yapılan özetleme çalışmalarının birçoğunun çıkarıma dayalı [17, 18, 19] olduğu, derin öğrenmenin popülerleşmesiyle yorumlamaya dayalı özetleme yöntemlerinin öne çıktığı gözlemlenmiştir [21, 22]. Ancak, Google mühendisleri tarafından geliştirilen BERT modeli ile çıkarıma dayalı yöntemlerin yeni bir hâl aldığı görülmüştür [23]. Ulusal yayınlarda ve tez çalışmalarında yapılan metin özetleme çalışmalarının neredeyse tamamının çıkarıma dayalı olduğu, yorumlamaya dayalı özetlemenin melez bir yöntemde [28] kullanılması haricinde pek tercih edilmediği, makine öğrenmesi yöntemlerinin metin sınıflandırmada [29, 30] ve metin özetlemede kullanıldığı gözlemlenmiştir [31]. Diğer yandan, yorumlamaya dayalı Türkçe özetlemenin pek tercih edilmemesinin yanında, Türkçe özetleme çalışmalarında çıkarıma ve yorumlamaya dayalı yöntemlerin arasında bir karşılaştırma yapılmadığı da gözlemlenmiş, bu sebeple bu tezde hem yorumlaya dayalı özetleme yapılmış, hem de çıkarıma dayalı yöntemlerle performansı karşılaştırılmıştır.

Geçmişte yapılan uluslararası yayınlarda LexRank [17], TextRank [18] ve makine öğrenmesi yöntemleri [19] öne çıkarken, son zamanlarda yapay sinir ağlarının tercih edildiği [20], Seq2Seq [21] ve BERT [23] modelleri ile bu çalışmalara farklı bir boyut eklendiği görülmüştür. Türkçe metinler için yapılan ulusal ve uluslararası yayınlarda gizli anlam analizi [25, 28] ve makine öğrenmesi yöntemlerinin [31] tercih edildiği; tez çalışmalarında ise tf-idf [1], LexRank [5], TextRank [7] gibi çeşitli yöntemlerin haricinde, yapay sinir ağlarının da tercih edildiği görülmüştür [13]. Ancak, Türkçe metin özetlemede BERT ve Seq2Seq modelleri ile bir çalışma yapılmadığı gözlemlendiği için, bu modellerin performansları tez dahilinde değerlendirilmiştir.

Bu tez çalışmasında 31.000'nin üzerinde paragraf/haber metni içeren veri setleri kullanılmıştır. Literatürde belirlenen en yüksek veri setinin boyutunun ise 120 haber ve 100 dergi makalesinden oluştuğu görülmektedir [26]. Her iki veri setinde yapılan karşılaştırma kapsamında, BERT ve Seq2Seq'in, eski yöntemler olan tf-idf, LexRank, TextRank gibi yöntemlere göre çok daha başarılı olduğu hem ROUGE hem de BLEU skoru ile görülmüştür. Wikihow ve TRT Haber veri setleri için en iyi ROUGE skoruna sahip olan modelin Seq2Seq'in olduğu görülmüştür. Seq2Seq, her ne kadar özet kelime sayısı arttıkça daha kötü performans gösterse de Türkçe haberler için ortalama 0,3 ROUGE-L skoru elde etmiştir. Bu değer, Türkçe metin özetleme çalışmaları tarafından elde edilen ve en yükseği Türkçe haberlerden oluşan veri setinde 0,561; dergi makalelerinden oluşan veri setinde 0,368 olan [26] ROUGE-L skorlarından

daha azdır. Bunun sebebinin, daha önceden yapılan çalışmalarda kullanılan veri setlerinin çok daha az sayıda metinden oluşması ve genellikle insanlar tarafından oluşturulması ya da düzenlemiş olması olduğu düşünülmektedir.

Seq2Seq modeli her iki veri seti için en iyi sonucu vermiş, web sitesi ilk aşamada Seq2Seq modeli ile özetleme yapacak şekilde tasarlanmıştır. Ancak, bu modelin sonuç üretirken oldukça yavaş çalıştığı ve sadece eğitim veri seti ile ilişkili konularda iyi performans gösterdiği gözlemlenmiştir. Bu sebeple, çıkarıma dayalı modellerin farklı konularda metin özetlemelerinde daha başarılı olabileceği göz önünde bulundurularak, web sitesi BERT ile özetleme yapacak şekilde yeniden tasarlanmıştır. BERT'in de Seq2Seq gibi GPU ile çalışmayan bir sunucuda yavaş özetleme yapması sebebiyle, çıkarıma dayalı bir başka yöntem olan TextRank algoritması alternatif özetleyici olarak kullanıcıların hizmetine sunulmuştur.

Her ne kadar bu tezin başlığı haberler ile ilgili olsa da; geliştirilen sistem, haberler haricinde de başarılı sonuçlar elde edebilmektedir. Bu tez kapsamında oluşturulan İnternet sitesi Nisan 2021'de hizmete açılmış, Temmuz 2021'e kadar 30.000'den fazla Türkçe metin özetleme işlemini başarıyla tamamlamış ve tamamlamaya devam etmektedir.

5.2. GELECEK ÇALIŞMALAR İÇİN ÇEŞİTLİ ÖNERİLER

Sistemin performansının arttırılabilmesi için, insanlar tarafından kontrol edilmiş, daha iyi referans özetler kullanılabilir. Öte yandan, akademik veya hukuki konular gibi niş bir alanda yorumlamaya dayalı özet üreteceklerin, referans özetlerini bu alanlardan sağlamaları oldukça önemlidir.

Bu tez sonucunda geliştirilen web uygulamasında kullanıcı tarafından sisteme sunulan girdi, sadece yazı girdisi veya bir karakter dizisi olarak İnternet sitesi üzerinden sisteme sunulmaktadır. Bu işlem, özetlenecek metnin yer aldığı web adresinin girilmesi ile kısaltılarak, geliştirilebilir. Buradaki problem, her İnternet sitesinin farklı yapıda olması ve farklı yapılara sahip bu İnternet sitelerinden hangi kısmın örümceklerce taranması gerektiğidir. Python'daki trafilatura gibi çeşitli yazılım kütüphaneleri kullanılarak bu sorunun çözülmesine katkı sağlanabilir.

Diğer yandan, web tabanlı sistemlerde hiçbir zaman kullanıcı girdisine güvenilmemesi gerektiği, bazı kullanıcıların sistemi kötüye kullanabileceği asla unutulmamalıdır.

Son olarak, geliştirilen özetleyici sistemler, istek sayısına kota koymak şartıyla, bir uygulama programlama arayüzü (API) aracılığıyla başka geliştiricilere açılabilir.

KAYNAKLAR

[1]. Akülker, E., 2019, *TF-IDF ve Pagerank algoritmaları kullanılarak Türkçe için text özetleme*, Yüksek Lisans Tezi, Atılım Üniversitesi Fen Bilimleri Enstitüsü.

[2]. Aysu, E., 2018, *Dev veri depolama ve Türkçe metin için otomatik özetleme*, Yüksek Lisans Tezi, Işık Üniversitesi Fen Bilimleri Enstitüsü.

[3]. Baydar, E., 2018, *Genetik algoritma kullanarak cümle seçme yaklaşımı ile otomatik metin özetleme*, Yüksek Lisans Tezi, Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü.

[4]. Işık, Y. E., 2018, *Otomatik doküman özetleme yöntemlerinin karşılaştırılması*, Yüksek Lisans Tezi, Cumhuriyet Üniversitesi Sosyal Bilimler Enstitüsü.

[5]. Nuzumlalı, M. Y., 2015, *Kök bulma ve cümle sadeleştirme yöntemlerinin Türkçe çoklu belge özetleme üzerine etkileri*, Yüksek Lisans Tezi, Boğaziçi Üniversitesi Fen Bilimleri Enstitüsü.

[6]. Birant, Ç. C., 2015, *Türkçe için kural tabanlı metin özetleme*, Doktora Tezi, Dokuz Eylül Üniversitesi Fen Bilimleri Enstitüsü.

[7]. Yalkın, C., 2014, *Çizge tabanlı metin özetleme*, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü.

[8]. Güran, A., 2013, *Otomatik metin özetleme sistemi*, Doktora Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü.

[9]. Berker, M., 2011, *Otomatik metin özetleme için genetik algoritmaların sözcük zincirleri ile kullanımı*, Doktora Tezi, Boğaziçi Üniversitesi Fen Bilimleri Enstitüsü.

[10]. Pembe, F. C., 2010, *Arama motorları için bilgi isteğine ve metin yapısına dayalı olarak otomatik doküman özetlenmesi*, Doktora Tezi, Boğaziçi Üniversitesi Fen Bilimleri Enstitüsü.

[11]. Tülek, M., 2007, *Türkçe için metin özetleme*, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü.

[12]. Ercan, G., 2006, *Otomatik özet ve anahtar kelime çıkarma*, Yüksek Lisans Tezi, İhsan Doğramacı Bilkent Üniversitesi Fen Bilimleri Enstitüsü.

[13]. Erhandı, B., 2020, *Derin öğrenme ile metin özetleme*, Yüksek Lisans Tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü.

[14]. Özkan, C., 2019, *İnternet tabanlı Türkçe metinler için otomatik özetleme tekniği*, Yüksek Lisans Tezi, Maltepe Üniversitesi Fen Bilimleri Enstitüsü.

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

[16]. Gong, Y., & Liu, X., 2001, Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 19-25).

[17]. Erkan, G., & Radev, D. R., 2004, Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, *22*, 457-479.

[18]. Mihalcea, R., & Tarau, P., 2004, Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404-411).

[19]. Wong, K. F., Wu, M., & Li, W., 2008, Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd international conference on computational linguistics (Coling 2008)* (pp. 985-992).

[20]. Nallapati, R., Zhai, F., & Zhou, B., 2017, Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1).

[21]. Nallapati, R., Zhou, B., Gulcehre, C., & Xiang, B., 2016, Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

[22]. See, A., Liu, P. J., & Manning, C. D., 2017, Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

[23]. Liu, Y., 2019, Fine-tune BERT for extractive summarization. *arXiv preprint arXiv:1903.10318*.

[24]. Altan, Z., 2004, A Turkish automatic text summarization system. In *IASTED International Conference on*.

[25]. Ozsoy, M., Cicekli, I., & Alpaslan, F., 2010, Text summarization of Turkish texts using latent semantic analysis. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)* (pp. 869-876).

[26]. Kutlu, M., Cığır, C., & Cicekli, I., 2010, Generic text summarization for Turkish. *The Computer Journal*, *53*(8), 1315-1323.

[27]. Uzundere, E., Dedja, E., Diri, B., & Amasyalı, M. F., 2008, Türkçe haber metinleri için otomatik özetleme. *Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu, Isparta, Türkiye*, 1-3.

[28]. Hatipoğlu, A., & Omurca, S. İ., 2015, Türkçe metin özetlemede melez modelleme. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, *17*(50), 95-108.

[29]. Çiğdem, A. C. I., & Çırak, A., 2019, Türkçe haber metinlerinin konvolüsyonel sinir ağları ve Word2vec kullanılarak sınıflandırılması. *Bilişim Teknolojileri Dergisi*, *12*(3), 219-228.

[30]. Çelik, Ö., & Koç, B. C. TF-IDF, Word2vec ve Fasttext vektör model yöntemleri ile Türkçe haber metinlerinin sınıflandırılması, 2020, *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, *23*(67), 121-127.

[31]. Kartal, Y. S., & Kutlu, M., 2020, Türkçe haber metinleri için makine öğrenmesi temelli özetleme. In *2020 28th Signal Processing and Communications Applications Conference, SIU 2020-Proceedings*. Institute of Electrical and Electronics Engineers Inc.

[32]. Diebold, F. X., 2012, On the origin(s) and development of the term 'big data'.

[33]. Wu, X., Zhu, X., Wu, G. Q., & Ding, W., 2013, Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), 97-107.

[34]. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P., 1996, From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37-37.

[35]. Chen, M. S., Han, J., & Yu, P. S., 1996, Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*, *8*(6), 866-883.

[36]. Zhang, S., Zhang, C., & Yang, Q., 2003, Data preparation for data mining. *Applied artificial intelligence*, *17*(5-6), 375-381.

[37]. Hearst, M., 2003, What is text mining. *SIMS, UC Berkeley*, *5*.

[38]. Hashimi, H., Hafez, A., & Mathkour, H., 2015, Selection criteria for text mining approaches. *Computers in Human Behavior*, *51*, 729-733.

[39]. Chowdhury, G. G., 2003, Natural language processing. *Annual review of information science and technology*, *37*(1), 51-89.

[40]. Joseph, S. R., Hlomani, H., Letsholo, K., Kaniwa, F., & Sedimo, K. (2016). Natural language processing: A review. *Natural Language Processing: A Review*, *6*, 207-210.

[41]. Chomsky, N., 1957, *Syntactic structures*. The Hague: Mouton.

[42]. Schank, R. C., 1972, Conceptual dependency: A theory of natural language understanding. *Cognitive psychology*, *3*(4), 552-631.

[43]. Cambria, E., & White, B., 2014, Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, *9*(2), 48-57.

[44]. Goldberg, Y., 2016, A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*, 345-420.

[45]. Khurana, D., Koli, A., Khatter, K., & Singh, S., 2017, Natural language processing: State of the art, current trends and challenges. *arXiv preprint arXiv:1708.05148*.

[46]. Chopra, A., Prashar, A., & Sain, C., 2013, Natural language processing. *International journal of technology enhancements and emerging engineering research*, *1*(4), 131-134.

[47]. Tapsai, C., Meesad, P., & Haruechaiyasak, C., 2016, TLS-ART: Thai language segmentation by automatic ranking trie. In *9th International Conference Autonomous Systems*.

[48]. Smilkov, D., Thorat, N., Assogba, Y., Yuan, A., Kreeger, N., Yu, P., ... & Wattenberg, M., 2019, Tensorflow. js: Machine learning for the web and beyond. *arXiv preprint arXiv:1901.05350*.

[49]. Akın, A. A., & Akın, M. D., 2007, Zemberek, an open source NLP framework for Turkic languages. *Structure*, *10*, 1-5.

[50]. Denny, M., & Spirling, A., 2017, Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *When It Misleads, and What to Do about It (September 27, 2017)*.

[51]. Leidner, J. L., & Plachouras, V., 2017, Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 30-40).

[52]. Hovy, D., & Spruit, S. L., 2016, The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 591-598).

[53]. Larson, B. N., 2017, Gender as a variable in natural-language processing: Ethical considerations.

[54]. Jordan, M. I., & Mitchell, T. M., 2015, Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

[55]. Dey, A., 2016, Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174-1179.

[56]. Kaelbling, L. P., Littman, M. L., & Moore, A. W., 1996, Reinforcement learning: A survey. *Journal of artificial intelligence research*, *4*, 237-285.

[57]. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X, 2016, Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)* (pp. 265-283).

[58]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E., 2011, Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

[59]. Nenkova, A., & McKeown, K., 2012, A survey of text summarization techniques. In *Mining text data* (pp. 43-76). Springer, Boston, MA.

[60]. Jones, K. S., 1999, Automatic summarizing: factors and directions. *Advances in automatic text summarization*, 1-12.

[61]. Neto, J. L., Freitas, A. A., & Kaestner, C. A., 2002, Automatic text summarization using a machine learning approach. In *Brazilian symposium on artificial intelligence* (pp. 205-215). Springer, Berlin, Heidelberg.

[62]. Mani, I., 2001, Summarization evaluation: An overview.

[63]. Tas, O., & Kiyani, F., 2007, A survey automatic text summarization. *PressAcademia Procedia*, *5*(1), 205-213.

[64]. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K., 2017, Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

[65]. Dong, Y., 2018, A survey on neural network-based summarization methods. *arXiv preprint arXiv:1804.04589*.

[66]. Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1*(1), 60-76.

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

[68]. Christian, H., Agus, M. P., & Suhartono, D., 2016, Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications*, *7*(4), 285-294.

[69]. Steinberger, J., & Jezek, K., 2004, Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, *4*, 93-100.

[70]. Barrios, F., López, F., Argerich, L., & Wachenchauzer, R., 2016, Variations of the similarity function of textrank for automated summarization. *arXiv preprint arXiv:1602.03606*.

[71]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., 2018, BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[72]. Zhang, Y., Li, D., Wang, Y., Fang, Y., & Xiao, W., 2019, Abstract text summarization with a convolutional Seq2Seq model. *Applied Sciences*, *9*(8), 1665.

[73]. Amidi, M. & Amidi, S., 2018, Deep Learning cheatsheets for Stanford's CS 230, https://stanford.edu/~shervine/teaching/cs-230/, [Ziyaret tarihi: 25 Ocak 2021].

[74]. Sutskever, I., Vinyals, O., & Le, Q. V., 2014, Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

[75]. Umezawa, K. & Saito, S., 2018, Write a sequence to sequence (Seq2Seq) model, https://docs.chainer.org/en/stable/examples/Seq2Seq.html, [Ziyaret tarihi: 3 Şubat 2021].

[76]. Graham, Y., 2015, Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 128-137).

[77]. Lin, C. Y., 2004, Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81).

[78]. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J., 2002, BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

[79]. Nenkova, A., & Passonneau, R. J., 2004, Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004* (pp. 145-152).

[80]. Koupaee, M., & Wang, W. Y., 2018, Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

[81]. Bahdanau, D., Cho, K., & Bengio, Y., 2014, Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[82]. Jazayeri, M., 2007, Some trends in web application development. In *Future of Software Engineering (FOSE'07)* (pp. 199-213). IEEE.

[83]. Conallen, J., 1999, Modeling web application architectures with UML. *Communications of the ACM*, *42*(10), 63-70.

EKLER

EK 1. Etkisiz Kelimeler Listesi

Acaba, ama, aslında, az, bazı, belki, biri, birkaç, birşey, biz, bu, çok, çünkü, da, daha, de, defa, diye, eğer, en, gibi, hem, hep, hepsi, her, hiç, için, ile, ise, kez, ki, kim, mı, mu, mü, nasıl, ne, neden, nerde, nerede, nereye, niçin, niye, o, sanki, şey, siz, şu, tüm, ve, veya, ya, yani[24]

[24] https://github.com/nltk/nltk

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Burak Özdemir
Doğum Yeri	
Doğum Tarihi	
Uyruğu	☑ T.C. ❒ Diğer:
Telefon	
E-Posta Adresi	
Web Adresi	

Eğitim Bilgileri	
Lisans	
Üniversite	İstanbul Teknik Üniversitesi
Fakülte	Bilgisayar ve Bilişim Fakültesi
Bölümü	Bilgisayar Mühendisliği
Mezuniyet Yılı	2015

Yüksek Lisans	
Üniversite	İstanbul Üniversitesi
Enstitü Adı	Fen Bilimleri
Anabilim Dalı	Enformatik Anabilim Dalı
Programı	Enformatik Programı