TÜRK DİLİNDE DERİN ÖĞRENME İLE METİN ÖZETLEME

Neda ALIPOUR

Doktora Tezi
Yönetim Bilişim Sistemleri Ana Bilim Dalı
Dr. Öğr. Üyesi Serdar AYDIN
2023

T.C.
ATATÜK ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
YÖNETİM BİLİŞİM SİSTEMLERİ ANA BİLİM DALI

Neda ALIPOUR

TÜRK DİLİNDE DERİN ÖĞRENME İLE METİN ÖZETLEME

DOKTORA TEZİ

TEZ YÖNETİCİSİ
Dr. Öğr. Üyesi Serdar AYDIN

ERZURUM – 2023

SOSYAL BİLİMLER ENSTİTÜSÜ
Graduate School of Social Sciences

TEZ BEYAN FORMU

SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE

BİLDİRİM

Atatürk Üniversitesi Lisansüstü Eğitim ve Öğretim Uygulama Esaslarının ilgili maddelerine göre hazırlamış olduğum "Türk Dilinde Derin Öğrenme ile Metin Özetleme" adlı tezin/raporun tamamen kendi çalışmam olduğunu ve her alıntıya kaynak gösterdiğimi taahhüt eder, tezimin/raporumun kâğıt ve elektronik kopyalarının aşağıda belirttiğim koşullarda saklanmasına izin verdiğimi onaylarım.

Gereğini bilgilerinize arz ederim *.

☐ Tezimin/Raporumun tamamı her yerden erişime açılabilir.

☒ Tezimin/Raporumun makale için **altı ay,** patent için **iki yıl** süreyle erişiminin ertelenmesini istiyorum.

Neda ALIPOUR

Aslı Islak İmzalıdır

---

* **LİSANSÜSTÜ TEZLERİN ELEKTRONİK ORTAMDA TOPLANMASI, DÜZENLENMESİ VE ERİŞİME AÇILMASINA İLİŞKİN YÖNERGE**

.....

ÜÇÜNCÜ BÖLÜM

Çeşitli ve Son Hükümler

Lisansüstü tezlerin erişime açılmasının ertelenmesi MADDE 6– (1) Lisansüstü tezle ilgili **patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda,** tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu **iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.**

(2) Yeni teknik, materyal ve metotların kullanıldığı, henüz **makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış** ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile **altı ayı aşmamak üzere tezin erişime açılması engellenebilir.**

Gizlilik dereceli tezler MADDE 7– (1) Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.

(2) Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir.

SOSYAL BİLİMLER ENSTİTÜSÜ
Graduate School of Social Sciences

TEZ KABUL TUTANAĞI

SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE

Dr. Öğr. Üyesi Serdar AYDIN danışmanlığında, Neda ALIPOUR tarafından hazırlanan bu çalışma 13 / 01 / 2023 tarihinde aşağıda isimleri yazılı jüri tarafından Yönetim Bilişim Sistemleri Anabilim Dalı'nda Doktora Tezi olarak kabul edilmiştir.

Başkan	: Prof. Dr. Üstün ÖZEN	İmza: Aslı Islak İmzalıdır
Jüri Üyesi	: Dr. Öğr. Üyesi Serdar AYDIN	İmza: Aslı Islak İmzalıdır
Jüri Üyesi	: Doç. Dr. Handan ÇAM	İmza: Aslı Islak İmzalıdır
Jüri Üyesi	: Dr. Öğr. Üyesi Bilal USANMAZ	İmza: cAslı Islak İmzalıdır
Jüri Üyesi	: Dr. Öğr. Üyesi Fulya ASLAY	İmza: Aslı Islak İmzalıdır

Prof. Dr. Sait UYLAŞ
Enstitü Müdürü
Aslı Islak İmzalıdır

İÇİNDEKİLER

ÖZET

DOKTORA TEZİ

TÜRK DİLİNDE DERİN ÖĞRENME İLE METİN ÖZETLEME

Neda ALIPOUR

Danışman: Dr. Öğr. Üyesi Serdar AYDIN

2023, 97 Sayfa

Jüri: Dr. Öğr. Üyesi Serdar AYDIN (Danışman)
Prof. Dr. Üstün ÖZEN
Doç. Dr. Handan ÇAM
Dr. Öğr. Üyesi Bilal USANMAZ
Dr. Öğr. Üyesi Fulya ASLAY

Doğal dil işleme alanındaki en önemli alanlardan biri metin özetlemedir. Çeşitli alanlarda kullanılmaya başlayan metin özetleme zaman ve bütçede tasarruf sağlayarak yöneticilerin hızlı bir şekilde karar vermelerine imkân sağlamaktadır. Metin özetleme tek bir paragraf, tek sayfalı veya çok sayfalı dokümantasyonlar üzerinde uygulanabilir. Günümüzde metin özetleme çalışmaları daha çok haber veri setleri üzerinde gerçekleştirilmektedir. Ancak bu çalışmada Türkçe dili için makale veri seti yazar tarafından hazırlanmış ve literatüre kazandırılmıştır.

Bu çalışmada haber ve makale olmak üzere iki faklı veri seti kullanılmış ve başarısı incelenmiştir. Haber veri setinde 80000 eğitim, 20000 doğrulama ve 2000 test kullanılmış, ancak makale veri setinde 818 eğitim, 91 doğrulama ve 101 test veri seti olarak sistemin eğitilmesi ve denenmesi için kullanılmıştır. Makale veri setinin uzun olması nedeniyle ilk olarak BERT çıkarımsal özetleme yöntemi ile metinlerin uzunluğu azaltılarak sistemin eğitilmesi için hazırlanmıştır.

Çalışmada mT5 mimarisi kullanılmış ve ilk olarak haber veri seti 8 batch-size ve 16 batch-size ile çalıştırılmıştır. Daha sonra öğrenme oranı değiştirilerek sistemin başarısı ROUGE metrikleriyle incelenmiştir. Bu çerçevede 8 batch-size, 16 batch-size'a göre daha başarılı olmuştur. Öğrenme oranı değiştirildikten sonra sistemin başarısı önemli ölçüde artmıştır. Devamında makale veri seti 8 batch-size ile çalıştırılmış ve başarısı ROUGE metrikleri ile incelenmiştir.

Anahtar Kelimeler: Derin Öğrenme, Metin Özetleme, Transformer, Önceden Eğitilmiş, Veri Seti.

ABSTRACT

Ph. D. DISSERTATION

TEXT SUMMARIZATION USING DEEP LEARNING IN TURKISH LANGUAGE

Neda ALIPOUR

Advisor: Assist. Prof. Serdar AYDIN

2023, 97 Pages

Jury: Assist. Prof. Dr. Serdar AYDIN (Advisor)
Prof. Dr. Üstün ÖZEN
Assoc. Prof. Dr. Handan ÇAM
Assist. Prof. Dr. Bilal USANMAZ
Assist. Prof. Dr. Fulya ASLAY

One of the most important areas in natural language processing is text summarization. Text summarization, which has started to be used in various fields, allows managers to make quick decisions by saving time and budget. Text summarization can be implemented on single-paragraph, single-page, or multi-pages documentation. Today, text summarization studies are mostly carried out on news data sets. However, in this study, the article data set for the Turkish language was prepared by the author to the literature.

In this study, two different data sets, news and articles, were used and its success was examined. In the news dataset, 80000 training, 20000 validation and 2000 tests were used, but in the article dataset, 818 training, 91 validation and 101 test datasets were used to train and test the system. Due to the long article data set, it was first prepared to train the system by reducing the length of the texts with the BERT extractive summarization method.

In the study, mT5 architecture was used and the news data set was run with 8 batch-size and 16 batch-size. Afterwards, the learning rate was changed and the success of the system was examined with ROUGE metrics. In this framework, 8 batch-sizes were more successful than 16 batch-sizes. After the learning rate was changed, the success of the system increased significantly. Afterwards, the article dataset was run with 8 batch-sizes and its success was examined with ROUGE metrics.

Keywords: Deep Learning, Text Summarization, Transformer, Pre-trained, Dataset.

KISALTMALAR VE SİMGELER DİZİNİ

ARG	: America's Research Group
Bi-LSTM	: Bidirectional Long Short-Term Memory
CNN	: Cable News Network
DUC	: Belge Anlama Konferansı
FLSTM	: Fuzzy-Based LSTM
FRUMP	: Fast Reading Understanding And Memory Program
HTML	: The HyperText Markup Language or HTML
LSA	: Latent Semantic Analysis
LSTM	: Long Short-Term Memory
MEAD	: MEAD, özetleme için halka açık bir çerçevedir.
mT5	: Multilingual Text-to-Text Transfer Transformer
NISO	: National Information Standards Organization
NLP	: Natural language processing
P	: Precision
R	: Recall
RL	: Reinforcement Learning
RNN	: Recurrent Neural Network
ROUGE	: Recall-Oriented Understudy for Gisting Evaluation
Seq2Seq	: Sequence To Sequence
SCUs	: Summarization Content Units
SUMMONS	: Summarizing Online News articles
TF-IDF	: Term frequency- Inverse Document Frequency
THV	: Türkçe Haber Veri Seti
TLMs	: Transformer-based Language Models
TREC	: Text Retrieval Conference
UNESCO	: The United Nations Educational, Scientific and Cultural Organization

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

ÖNSÖZ

Yüksek lisans ve Doktora eğitim süreci boyunca bana olan güvenini ve desteğini hep hissettiğim, mütevazı kişiliği ve bilgeliği ile bana kılavuz olan, tezimin ortaya çıkmasında, geliştirilmesinde ve sonuçlandırılmasında her türlü özveriyi gösteren ve yoğun akademik yaşamında değerli zamanını her türlü problemimi çözmeye ayıran çok değerli danışman hocam **Dr. Öğr. Üyesi Serdar AYDIN**'e sonsuz teşekkürlerimi sunuyorum.

Çok değerli hocam **Prof. Dr. Üstün ÖZEN**'e, yüksek lisansa başladığım günden itibaren bana güvendiği ve desteklediği, en kritik dönemeçlerde en doğru kararları almamı sağladığı, görüşmelerimizde tezin gidişatı hakkında ne kadar karamsar olsam da olumlu bakış açısı, hem manevi desteğini hem de doğru yönlendirme ve tecrübeleriyle yardımlarını hiç bir zaman esirgemediği için sonsuz şükranlarımı sunarım.

Tezim ile ilgili tüm süreçlerde yanımda yer alan, bana sabreden, her zaman bana moral ve destek veren, beni her türlü sorunun üstesinden gelebileceğim konusunda cesaretlendiren sevgili eşim Hadi POURMOUSA'ya teşekkür ederim.

Erzurum – 2023 **Neda ALIPOUR**

GİRİŞ

Yıllarca, özetleme insanlar tarafından elle yapılırdı. Manuel özetleme, çok sayıda becerikli uzman, ciddi zaman ve bütçe gerektirir. Ayrıca günümüzde dünya genelinde bilgi miktarı, internet ve diğer kaynaklar tarafından giderek artmaktadır. Bu sorunun üstesinden gelmek için, otomatik metin özetleme gibi otomatik tekniklerin uygulanması kaçınılmaz hale gelmiştir. Otomatik metin özetleme, orijinal belgelerin yoğunlaştırılmış bir versiyonunu oluşturmak için kullanılan otomatik bir tekniktir. Doğal dil işleme alanında metin özetleme, sağlanan ekstra bilgiyi incelemeye gerek kalmadan kaynak metnin konusunu almanın bir yolu olarak tanımlanmaktadır. Metin özetlemede kullanılan yöntemler, Luhn tarafından ilk kez tanıtılmasından bu yana gelişti. Manuel metin özetleme 1958 tarihine dayanmaktadır (Luhn, 1958) ve önerilen teknikler kapsamlı bir şekilde incelendi (Lloret ve Palomar, 2012; Nenkova ve McKeown, 2011). Metin özetleme, önemli noktaları koruyarak bir girdi belgesinden otomatik olarak doğal dil özetleri oluşturma işlemidir. Özetleme çok miktarda bilgiyi kısa, bilgilendirici özetlere yoğunlaştırmak suretiyle, haber özeti oluşturma, arama ve rapor oluşturma gibi pek çok konuda yardımcı olabilir.

Bir özet, "göze çarpan ve önemli olan cümleleri barındıran" kısa bir düzeltme olarak tanımlanabilir (Kiani ve Akbarzadeh, 2006). Otomatik metin özetleme, bir bilgisayarın otomatik olarak metinden bir özet ürettiği süreçtir. Otomatik özetlemenin amacı, bir bilgi kaynağının içeriğinden kullanıcıya veya uygulamanın ihtiyaçlarına göre en önemli noktaları duyarlı bir şekilde sunmaktır (Mani, 2001a).

Literatürde farklı dillerde metin özetleme çalışması yapılmış olup bu amaçla kullanılan veri setleri bulunmaktadır. Çalışmaların çoğu haber veri setleri üzerinde gerçekleşmiş ve makale veri seti, Türkçe dâhil çoğu dilde bulunmamaktadır. Bu amaç doğrultusunda Türkçe makale veri seti ilk kez yazar tarafından oluşturulmuş ve literatüre kazandırılmıştır.

Çalışmanın ilk bölümünde metin özetlemeye neden ihtiyaç duyulduğu, özetlemenin türleri, metin özetlemenin tarihçesi anlatılmış ve literatür taraması gerçekleştirilmiştir. İkinci bölümde derin öğrenmenin kavramı ve metin özetlemenin konusunda kullanılan Uzun Kısa Hafızalı Bellek anlatılmış ve metin özetlemenin başarısını ölçen metriklerden bahsedilmiştir. Çalışmanın üçüncü bölümünde araştırmanın önemi ve amacı vurgulanmış

ve veri setinin nasıl hazırlandığı, hangi mimarinin kullanıldığı, sistemin nasıl eğitildi anlatılmış ve elde edilen sonuçlar sunulmuştur.

BİRİNCİ BÖLÜM

NEDEN ÖZETE İHTİYAÇ VARDIR

İnsanoğlu sınırlı veri işleme kapasitesine sahiptir ve bu nedenle her gün yaratılan çok büyük miktarda veriyi işleyemez. Herhangi bir metnin özetini elde etmek için metnin tamamını okumak gerekir. Bu kısa metinler ve hatta bazı uzun metinler için kolay bir görev gibi görünebilir, ancak bir insanın dünyadaki tüm metinleri okuması ve işlemesi, hatta belirli bir türdeki metinleri okuması imkânsızdır. Ayrıca farklı dillerdeki belgeler de bu zorluğu arttırmaktadır. Eğer birisi sadece boş zaman aktivitesi için değil, bir görevi gerçekleştirmek için okuyorsa, okuma zaman ve kaynak tüketen bir görev haline gelir. Bu nedenle, büyük miktarda metni daha verimli bir şekilde işlemeye yardımcı olacak bir araca ihtiyacı vardır. Bu noktada özetler devreye girer. Ulusal Bilgi Standartları Örgütü (NISO) belirttiği gibi, "iyi hazırlanmış bir özet, okuyucuların bir belgenin temel içeriğini hızlı bir şekilde tanımlamalarını, ilgi alanlarına uygunluğunu belirlemelerini ve böylece belgenin tamamını okumaları gerekip gerekmediğine karar vermelerini sağlar". Yani özet, dokümanın marjinal ilgisini çeken okuyucular için konusuna giriş veya görüşüne genel bir bakış sunabilir (Birant, 2015).

Metin özetleme 90'lı yıllardan beri popüler bir konu olmasına rağmen, çalışmaların çoğu sadece küçük bir dil grubuna uygulanmıştır (çoğunlukla İngilizce, ayrıca Çince, Arapça ve İspanyolca). Önerilen özetleme tekniklerinin diğer dillere uygulanmasıyla ilgili en önemli sorun, özetleme görevi için insan tarafından hazırlanan bir veri kümesinin oluşturulmasının çok zor olmasıdır. Diğer diller üzerinde gelecekteki çok belgeli özetleme çalışmaları için el ile açıklamalı çok dilli başlıklar toplamaya çalışan yeni çalışmalar bulumaktadır (Giannakopoulos vd., 2011).

Mevcut bilgi tabanını başka dillere taşımak için diğer bir zorluk, dillerin farklı biçim-sözdizimsel (morphosyntactic) özelliklerinden kaynaklanan uyumluluk sorunlarıdır. Doğal Dil İşleme (NLP) alanındaki önceki çalışmalar, İngilizce gibi diller için önerilen yöntemlerin genellikle Fince, Türkçe ve Çekçe gibi morfolojik açıdan zengin diller için iyi çalışmadığını göstermiştir, bu nedenle bu dillerin morfolojik yapılarını dikkate alan ek yöntemlere ihtiyaç vardır (Eryiğit, Nivre ve oflazer, 2008). Mesela; Türkçe, kök kelimelerin birçok türev ve çekim eki alabildiği eklemeli bir dildir.

Bu özellik, sonuçta veri seyrekliği sorununa yol açan çok sayıda farklı kelime yüzey formu oluşturur. Hakkani-Tür, Oflazer ve G. Tür (2000) Türkçe ve İngilizce'ye özgü terimlerin sayısını analiz etti ve 1 milyonluk bir kelime topluluğu için Türkçe terim sayısının İngilizce'den üç kat daha fazla olduğunu gösterdi. Türkçede metin özetlemeye odaklanan sadece birkaç çalışma vardır, bunların hepsi tek belgeli özetleme ile ilgilidir. Altan(2004), Cığır, Kutlu ve Ciçekli (2009) özellik tabanlı yaklaşımları önerdiler, Ozsoy, Ciçekli ve Alpaslan (2010), Güran, Bekar ve Akyoku (2010) Latent Semantic Analysis (LSA) 'ı kullandı, Güran, Bayazıt ve Bekar (2011), negatif olmayan matris faktörünü uyguladı ve ön işleme adımı olarak ardışık kelimeler saptama kullandı. Bu çalışmaların bazıları morfolojik analiz metotları uygulasa da hiçbiri etkilerini detaylı olarak analiz etmedi.

1.1. ÖZETLEMENİN ANA YÖNTEMLERİ

Metin özetleme yöntemleri, iki kategoride sınıflandırılmıştır. Kategoriler özetsel ve çıkarımsal tekniklerdir. Bu iki tür sınıflandırma yöntemleri detaylı bir şekilde aşağıda açıklanmıştır.

1.1.1. Özetsel Metin Özetleme

Orijinal metinde bulunmayan cümleleri kullanarak yeni ifadeler üretilir (Chopra, Auli ve Rush 2016; Nallapati vd., 2016). Yani, özetleyicinin çıkarılan içeriği veya metni yeniden üretmesi gerekir. Özetsel özetleme, metni anlama ve inceleme için kullanılan dilbilimsel yöntemlerin yardımıyla orijinal metni anlayarak özetleme yapar (Siddharthan, 2017). Özetsel metin özetleri, çıkarımsal özetinde ortaya çıkan problemleri ortadan kaldırmaya ve insana benzer özetlere yaklaşmaya çalışmaktadır. İlgili ana adımlar, temel özelliklerin kullanılmasıyla kullanılacak içeriğin seçilmesi, alınması ve tutarlı bir özeti çıkarmak için seçilen cümlenin azaltılması ve yeniden düzenlenmesidir. Orijinal metinde aynı kelimeler kullanılmayabilir. Özetsel metin özetlemesi ile uğraşırken, özetleri oluşturmak için yapılan bazı büyük kes ve yapıştır işleri vardır (Kasture, vd., 2014).

1.1.2. Çıkarımsal Metin Özetleme

Çıkarımsal özetler kaynaktan en önemli cümleleri tanımlar ve özeti oluşturmak için bunları bir araya getirir, birbirlerine dayanan ancak birbirlerinin bir uzantısı olarak çalışan üç adımdan oluşur. İlk adım, yalnızca önemli yönleri yakalayan girdi metninin ara ifadesinin oluşturulmasını içerir. Daha sonra, ikinci adım bir cümle puanlama mekanizması oluşturur ve aday cümleleri verir ve son adım ise, belirtilen özelliklere ve ikinci aşamadaki cümle puanlarına dayalı olarak gereksinimin en üstündeki aday cümleleri seçer (Abdı Omar, 2018).

1.1.2.1. Çıkarımsal Özetlemenın Zorlukları

Çıkarımsal özette, cümlelerin en göze çarpan bilgilere göre sıralanması gerekir (Dorr, Zajic ve Schwartz, 2003; Nallapati vd., 2017; Over ve Liggett 2002). Çıkarımsal özetleme esnektir ve özetsel özetlemeye göre daha az zaman harcanır (Patil ve Brazdil, 2007). Çıkarımsal özetlemede tüm cümleler bir matris formunda ele alınır ve bazı özellik vektörlerine dayanarak, gerekli veya önemli cümleler çıkarılır. Bir özellik vektörü, bir nesneyi temsil eden sayısal özelliklerin n boyutlu bir vektörüdür. Çıkarımsal özetlemenin asıl amacı, bir kullanıcının gereksinimine göre uygun cümleyi seçmektir (Mani, 2001a; 2001b). Birçok araştırma çalışmasında çıkarımsal özet, cümle sıralaması olarak da adlandırılır (Edmundson, 1969; Mani ve Maybury, 1999). Burada, özetlemenin analiz aşaması önemlidir. Edmundson (1969), çıkarımsal özetleme için bir çerçeve oluşturmuştur. Örneğin, özetleyicilerini eğitmek ve genetik programlamayı kullanırken özellik ağırlıklarını öğrenmek için genetik algoritmaları ve matematiksel regresyon modellerini kullanmıştır (Fattah an Ren, 2008; Dehkordi, Khosravi ve Kumarci, 2009). Çıkarımsal özet, cümlenin altında yatan anlamı kontrol etmez, ancak cümleyi önemli ya da önemsiz olarak niteleyen özelliklere dayanır; Zorluklardan biri, sık sözcük içeriğinden dolayı uzun cümleler dâhil edildiğinde ortaya çıkar. Bu, bir özetin bir bölümünü oluşturan ve bilgilerin aşırı yüklenmesine neden olan bir belgenin önemsiz kısımlarına yol açar. Çıkarımsal özette en iyi cümleler seçileceği ve uzun belgelerde bilgi yayıldığı için önemli detaylar kaçırılabilir. İsimler ve uygun isimler karıştırıldığı zaman çelişkili bilgilerin yanlışlığı veya yanlış beyanı, sarkan zamirler olarak adlandırılır, eksik parça eksik olduğunda bilgi yanlış yorumlanabilir. Çıkarımsal özet çoğunlukla tutarlı bir yapıya

sahiptir; çünkü cümleler, rastgele seçilir ve özette işlem sonrası ifadeleri olmadan birlikte sabitlenir. (Abdı Omer, 2018).

1.2. ÖZET TÜRLERİ

Torres-Moreno (2014) aşağıda özetlediğimiz özetleme türlerinin kapsamlı bir listesini sunmaktadır:

Giriş Belge Sayısı:

Özet sistemin girdisine göre tek belgeli özetleme ve çok belgeli özetleme olarak iki gruba ayrılır. Tek belgeli özetleme, sadece tek bir belge ile ilgilenen bir süreçtir. Çok belgeli özetleme, sadece tek bir belgeyi değil, aynı zamanda ilgili belgelerin bir özetini de tek bir özet halinde kısaltma yöntemidir (Ou, Khoo ve Goh, 2008, McKeown ve Radev, 1995). Konsept kolay görünüyor, ancak uygulama sırasında derlenmesi zor bir iştir. Bazen istediğimiz hedefi yerine getiremeyebilir. Tek belgeli özetlemede kullanılan benzer tekniklerin çoğu, çok belgeli özetlemede de kullanılır. Kayda değer farklılıklar vardır (Goldstein vd., 2000): (1) Bir konuyla ilgili makaleler grubunda bulunan fazlalık derecesi, bir makale içindeki fazlalık derecesinden oldukça fazladır, çünkü her makale en önemli noktayı ve ayrıca gereken paylaşılan arka planı göstermek için uygundur. Bu nedenle, fazlalık önleme yöntemleri hayati bir rol oynamaktadır. (2) Sıkıştırma oranı (belirlenen belgenin boyutuna göre özet boyuttu), konuyla ilgili geniş kapsamlı belgeler için tek belgeli özetlerinden önemli ölçüde daha düşük olacaktır.

Giriş Dili Sayısı:

Özet, bir dili kabul etme sınırlamasına bağlı olarak tek dilli veya çok dilli olarak sınıflandırılabilir. Daha önce metin özetlemesinde yapılan çalışmaların hepsi temelde ve çoğunlukla İngilizce metin olan bir dille sınırlıydı ve daha sonra özet, Bangali ve Arapça gibi diğer dillere de kabul edildi. Zaman içinde bir özet, farklı dillerden kaynak belgeyi seçip belirtilen dil özetinde özetlenebilir (Radev vd., 2004, Alami vd., 2015).

Tür / Giriş Metni Sınırlaması

Bir belgenin özetleri, bilimsel makaleler veya belirli haberler gibi belirli bir türle sınırlandırılmış olarak tanımlanabilir veya biyomedikal gibi belirli alanlar ile sınırlanabilir veya giriş metninin alanını sınırlamayan herhangi bir alanın açık şablon tabanlı bir özeti olabilir (Meena ve Gopalani, 2015).

Hedef kitleye göre:

Özet etki alanına göre ikiye ayrılır: kaynak belgelerden gelen bilgilere göre benzersiz bir özet, profil olmayan özetler olarak adlandırılır. Öte yandan kullanıcı profiline göre, özel bir alanla ilgilenen kullanıcıları hedefleyen bir özette vardır.

Özetleyicinin türüne göre:

Özetleyicinin türüne göre özetler üçe ayrılır:

1) Yazarın özeti: Belgenin yazarı tarafından yazılmış özetlerdir.
2) Uzman özeti: Belge alanında uzmanlaşmış biri tarafından yazılmış özetlerdir.
3) Profesyonel özeti: Özet üretme tekniklerini bilen profesyonel bir özetleyici tarafından yazılmış özetlerdir.

İçeriğe göre: (According to context)

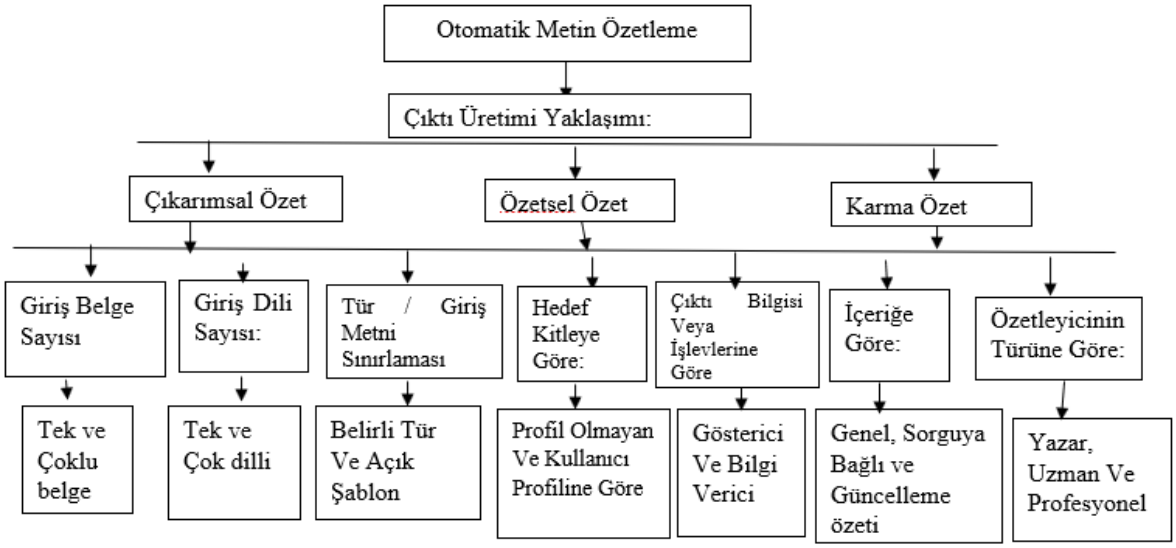
Belirli bir noktaya değinmeyen ve hedef kitlesi olmayan belgelerin özeti genel özetler olarak adlandırılır. Bu özet belgenin temel noktaları hakkında genel ayrıntıları verir. Öte yandan, sorgu odaklı sistemler, bir kullanıcı sorgusu ile ve sorguya bağlı içerikle ilgili belgelerin özetlerini üretir. Sorgu odaklı özetler, ders kitabı veya kullanma kılavuzu gibi büyük veya çeşitli konuları olan belgelerle ilgilenirken çok yararlıdırlar (Jackson ve Moulinier, 2002). Güncelleme özeti: yalnızca önemli yeni bilgileri gösteren ve tekrarlayan bilgileri engelleyen bir özettir.

Çıktı Bilgisi veya İşlevlerine göre:

Çıktının stili dikkate alındığında özet iki sınıfa ayrılabilir: çıktı, belgelerin bir taslağı veya ön izlemesi olabilir ve belge içeriğinde konu olan genel başlıklar belirlenmektedir. Böylece izleyiciye bir belgede ne bekleneceğini ve kaynak belgeyi daha derine inip okuması veya tamamen atması için bir karar vermesini sağlamaktadır. Bu bir gösterici (indicative) özeti olarak adlandırılır. İçindekiler tablosuna benzeyen ve belgedeki konular hakkında bilgi veren bir özettir. Öte yandan, özetin kaynak belgede ifade edilen tüm önemli bilgi ve fikirleri sağlayabildiği ve insanın isteğine bağlı olan konularla ilgili bilgi verici (informative) bir özet vardır yani kaynak metnin içeriğini yansıtan belgenin kısa bir versiyonudur.

Çıktı Üretimi Yaklaşımı:

Özetleme sistemi özetleri çıktı üretimi yaklaşımına göre özetsel, çıkarımsal veya karma özetler olarak sınıflandırmaktadır. Çıkarımsal özetler daha fazla araştırılmış ve uygulanması diğerlerine göre daha kolay olduğu için tutarlı bir şekilde kullanılmıştır. Ayrıca çıkarımsal özetler belgeden cümle seçmeye bağlıdır ve önemli cümleler değiştirilmeden sezgisel çıkarımlarla, istatistiksel yöntemlerle veya bunların birleşimiyle seçilmektedir. Öte yandan özetsel özet, cümle oluşturma tekniklerini kullanır ve neredeyse insan özetine yakın bir şekilde özet verilir, amaçlanan anlamı ya da çıkarımsal özetten daha tutarlı bir cümle vermek için bazen yeniden yapılandırılır ya da yeniden ifade edilir. Mesela, "Neda ışığı kapattı, yatağa geçti, güzlerini kapattı" cümlesinin yorumlanmış hali "Neda uydu" olacaktır. Hibrit özetlemede, sistem, çıkarımsal yöntemlerinin kullanımıyla özetsel özetlemede en üst düzeye çıkarılır, çıkarım aşamasında çıkarım teknikleri kullanılır ve çıktı cümleleri yeniden düzelten ve özetleri üreten özetsel bir jeneratörden beslenir (Abdı Omer, 2018).



Şekil 1.1. Özetleme Kategorilerinin Resimsel Görünümü (Abdı Omer, 2018)

1.3. METİN ÖZETLEMENİN GERÇEK DÜNYADA UYGULANMASI

Bir özetleme fikri, milyarlarca metin belgesinin incelemesi için gereken süreyi azaltır. Bu, gerçek dünyada doğrudan, belirli bir alanda çalışan araştırmacılar ve öğrencilerin yararına olduğu anlamına gelir ve çalışmaları için iyi kaynaklar bulmak için çaba harcamalarına gerek kalmaz. Aynı durumda, bir gazeteci, dünyadaki bir olay ya da haberin içeriğindeki en son haberlere zahmetsizce devam edebilir. Bir asistan rahatça

sorunları tartışabilir ve sağlık sektöründe doktor, yıllarca belgelenmiş sağlık detaylarına göz atmak yerine bir hastanın geçmişini bir özette alabilir. Daha önce yapılan çalışmalarda, otomatik özetlemenin, öğretimde örneğin İngilizce öğretmenlerinin daha hızlı bir yöntemle bilginin özetlenmesi, çıkarılması ve analizinde etkili bir şekilde yardımcı olabileceğini göstermektedir. Ayrıca Google arama motorlarında anahtar sözcük çıkarma işlemiyle yapılan metin özetlemesinden yararlanır (Moratanch, Chitrakala, 2016).

1.4. OTOMATİK METİN ÖZETLEME

Otomatik metin özetleme, uzun yıllardır aktif bir araştırma alanı olmuştur ve internetin hemen hemen tüm alanlarında kendi uygulamalarına sahiptir. Arama motorları, arama deneyiminin bir parçası olarak sorguya ve içeriğe özel özet parçacıkları sağlar, haber web siteleri makaleleri özetlemek için özetler kullanır, sosyal medya bunları içerik hedeflemesi için kullanır, e-ticaret web siteleri de öğe veya ürün vurgulamaları yoluyla daha iyi göz atma deneyimi için özetler kullanır. Otomatik özetleme, belirsizliklere katkıda bulunan metin biçimlerindeki, ifadelerdeki ve baskılardaki değişikliklerden oluşan çeşitli zorluklar getiren metin anlayışıyla yakından ilişkilidir. Metin özetlemesindeki araştırmacılar, bu soruna doğal dil işleme (Zhang, Wang, ve Li, 2011), istatistik (Darling ve Song, 2011), makine öğrenmesi ve metin analizi gibi birçok yönden yaklaşmış, metinlerin odağını belirleyen temel konudur.

Belgelerin otomatik olarak özetlenmesi için birkaç geçerli neden ortaya çıktı (Torres-Moreno, 2014):

i) Özetler okuma süresini kısaltır.
ii) Belgeleri araştırırken, özetler seçim sürecini kolaylaştırır.
iii) Otomatik özetleme endekslemenin etkinliğini arttırır.
iv) Otomatik özetleme algoritmaları, insan özetleyicilere göre daha az önyargılıdır.
v) Kişiselleştirilmiş özetler, kişiselleştirilmiş bilgiler sağladıkları için soru cevaplama sistemlerinde yararlıdır.
vi) Otomatik veya yarı otomatik özetleme sistemlerinin kullanılması, işlenen metin sayısını arttırır.

1.5. OTOMATİK METİN ÖZETLEMENİN TARİHÇESİ

Otomatik metin özetleme, başlangıcından farklı olarak, günümüzde bilgisayar bilimi, yapay zekâ, istatistik, bilişsel bilimler, doğal dil üretimi, makine öğrenimi, dil bilimi ve söylem analizi gibi Doğal Dil İşleme (NLP) dışındaki birçok alanın uzmanlığından yararlanan, disiplinler arası bir araştırma alanıdır (Birant, 2015).

Otomatik metin özetleme kavramı Luhn'a (1958) özet elde etmenin otomatik yöntemleri hakkında bazı araştırmalar yapmasına dayanır. Aynı zamanda Baxendale (1958) orijinal kaynaktan bilgi eklemeye çalışmıştır. Vasiliev 1963'te, UNESCO'ya bir rapor sunmuş ve o sırada otomatik özetleme durumuna geçmiştir. Bu çalışma, istatistiksel, tanımlayıcı ve otomatik metin özetlemesine anlamsal-mantıksal bir yaklaşımdan bahsetmiştir (Vasiliev, 1963).

1969'da Edmundson otomatik özetleme için yeni bir yaklaşım geliştirmiştir. Edmundson (1969), önceki çalışmaların aksine, sadece anahtar kelimeler gibi yüksek frekanslı içeriğin varlığına değil, aynı zamanda pragmatik kelimelerin (işaret kelimeleri); başlık kelimeleri ve yapısal göstergelere (bir cümle içindeki kelimelerin yerine) dikkat etmiştir. Burada önerilen yöntem, teknik belgelerin özetini almak için özelliklerin bir arada kullanılmasını gerektirdi, özellik kümeleri, işaret sözcüklerini, kelimenin konumunu ve sıklığını içeriyordu (Mani ve Bloedorn, 1997). Bu çalışmanın ardından, Rush, Salvador ve Zamora (1971) belgedeki cümleleri seçmek veya kaldırmak için bağlamsal çıkarım, konum yöntemi, işaret yöntemi, sıklık verileri ve tutarlılık hususlarını kullanarak bir algoritma geliştirmiştir.

Daha sonra, Pollock ve Zamora (1975), kimya konusuda, daha özel olarak farmakodinamik için Rush-Salvador-Zamora algoritmasını (Rush, Salvador ve Zamora, 1971) özelleştirmiştir. Çalışmada, önceki çalışmalarda romanlardan, ders kitaplarından, dergi veya gazete makalelerinden oluşan daha genel veri tabanları kullanılan farmakodinamik metinlerden oluşan özel bir veri tabanı kullanılmıştır. Çalışma, otomatik özetlemenin bazı metin türlerinde diğerlerinden daha başarılı olduğunu iddia etti.

1978'de Yale Yapay Zekâ Projesi, gazete makalelerini gözden geçirmek ve özetlemek için FRUMP (Fast Reading Understanding And Memory Program - Hızlı Okuma Anlama ve Bellek Programı) isimli tasarlanan yeni yazılımını (Hızlı Okuma Anlama ve Bellek Programı) duyurdu (deJong, 1982). FRUMP sistemi, dilbilimsel

bilgisine ek olarak "sketchy script" denilen bir yapı, çeşit pragmatik ve semantik bilgi çerçevesi kullandı. Pragmatik / semantik bilgi, yüksek düzeyde bir akıl yürütme sistemini oluşturuyordu ve düşük seviyeli (dilbilimsel) metin çözümleyicisi bu üst düzey akıl yürütmeye duyarlıydı (deJong, 1982).

1980'ler otomatik metin özetleme çalışmaları için oldukça sessizdir, ancak 1990'lar ve 2000'ler İnternet'in ticarileşmesi nedeniyle otomatik metin özetleme araştırmalarında bir patlamaya tanık olmuştur. Jones (1993), özetlemeyi bir bilgi yönetimi süreci olarak tanımladı ve özetleme görevi için insan özetleme, otomatik özetleme ve söylem yorumlama ve temsiline dayanan bir süreç yapısı tasarladı. Jones'un yaklaşımı, Retorik Yapı Teorisi (Mann & Thompson, 1988) gibi söylem analizi yaklaşımlarından yararlandı ve Halliday ve Hasan (1976), Rumelhart (1977) ve Kintsch & van Dijk (1978) gibi diğer dilsel kaynaklara atıfta bulundu. Sistem, FRUMP'da ve iletişimsel kaynaklarda özetleme yapmak için kullanılan sketchy scripts'e benzer etki alanı kaynaklarını ve dilsel kaynakları kullandı. 1995 yılında Kupiec, Pedersen ve Chen, hesaplanmış belge özetinin belirli türüne odaklandı ve istatistiksel bir çerçeve kullanarak eğitilebilir bir özetleme programı geliştirdi. Özetleyici oluşturmak için cümle uzunluğu kesme özelliği, sabit cümle parçası özelliği, paragraf özelliği, tematik sözcük özelliği ve büyük harf özelliği içeren bir özellik seti kullandılar (Kupiec, Pedersen ve Chen, 1995).

Aynı yıl McKeown ve Radev (1995) aynı olayla ilgili makaleler için ilk olarak çok belgeli bir özetleyici sistemi (SUMMONS) sundu. SUMMONS, aynı olaya dayanan birden fazla makaleyi özetleyen bir NLP özetleyici yöntemidir. Altta yatan bir bilgi tabanından bilgi seçen bir içerik planlamacısından ve seçilen bilgide yer alan kavramlara atıfta bulunmak için kelimeleri seçen ve bu kelimeleri dilbilgisi cümleleriyle birleştiren bir dilbilimsel üreticiden meydana gelmiştir. Geleneksel dil oluşturma mimarisine dayalı özetleyicinin iki ana bileşeni vardır: içerik planlayıcısı, yani ana süreç bilgi tabanına neyin dâhil edileceğini ve neyin atılacağını ve bir kavramı en iyi tanımlayan kelimeleri seçen ve tutarlı cümleler oluşturmak gibi kelimeleri düzenleyen dilsel bileşenlere dayanarak belirlemektir(McKeown, Radev, 1995).

İlk özetleme, özetlemenin cümle düzeyinde olduğu ve kelime sıklıklarına bağlı olduğu ticari haberlerde yapıldı (Brandow, Mitze ve Rau, 1995).

1997 ve 1999'da çok belgeli özetlemesi için ikinci bir yöntem ve metin özetlemesi için grafik tabanlı bir yöntem önerdi ((Mani, Gates, Bloedorn, 1999, Radev vd., 2005).

Marcu (1998), Retorik Yapı Teorisini (Mann ve Thompson, 1988) kullandı ve ağırlıklı olarak söylem belirteçlerini ve sözlük-gramer yapılarını söylemekten bahsetti. Çalışma, metinsel birimler arasındaki retorik ilişkileri varsaymak ve doğal dil metinlerini söylem ağaçlarıyla otomatik olarak eşleştirmek için retorik bir ayrıştırma algoritması geliştirmiştir.

1997 yılında skor zinciri tanıttıldı (Conroy, O'Leary, 2001). Barzilay ve Elhadad (1997), sözcük zincirlerinden türetilen metindeki konu ilerlemesini kullanarak, tam anlamsal yorumlama gerektirmeden metinleri özetlemek için bir teknik geliştirmiştir. Çalışma, nominal gruplar ve segmentasyon algoritması için WordNet, bir konuşma bölümü etiketleyicisi, sığ bir çözümleyici gibi birçok sağlam bilgi temelini birleştirdi.

Carbonell ve Goldstein (1998) sorgu alaka düzeyini bilgi yeniliği ile birleştirmiştir. İlgili yeniliğe vurgu yapan Maximal Marjinal Alaka (MMR), özellikle çok belgeli özetleme için sorgu alaka düzeyini korurken fazlalığı azaltmayı amaçlamıştır.

Witbrock ve Mittal (1999), bir eğitim kurumundan öğrenilen bir tarzda daha kısaltılmış ve daha uyumlu tutarlı özetler üretmek için seçme ve terim sıralama sürecinin istatistiksel modellerini kullanan bir özetleme yöntemi önermiştir.

Hovy ve Lin (1998; 1999) bir metni üç aşamada özetleyen bir özetleme sistemi (SUMMARIST) önermiştir:

i) Konum modülü, işaret ifade modülü, konu imza modülü, söylem yapısı modülü, konu tanımlama entegrasyon modülünü içeren konu tanımlaması;
ii) Kavram sayımı içeren konu yorumu, konu imzalarını kullanarak yorumlama;
iii) Bir Mikroplanner ve bir cümle oluşturucu içeren özet oluşturma.

Knight & Marcu (2000) yaptıkları çalışmada, önceki çalışmaları incelemiş ve bu çalışmaların yalnızca bilginin elde edilmesine odaklanmış olmasından kaynaklandıklarını ifade etmiş ve bununla birlikte, basitçe metinsel bölümlerin birleştirilmesi tutarlı çıktılar vermeyeceği sonucuna varmıştır. Be nedenle, bu çalışma cümle boyutunu azaltmak ve cümleleri sıkıştırmak için karara dayalı bir model kullanmıştır.

Osborne (2002), özete dahil edilecek cümleleri belirlerken bir sınıflandırıcı yetiştirmek için Naiv Bayes'i kullandıkları yerlerde eğitilebilir ilk yöntemin tanımlamasını yapmıştır.

Radev, Jing ve Budzikowska (2000), bir kümedeki yalnızca bir makale için değil, bir konu tespit ve takip sistemi tarafından üretilen tüm makaleler için merkezi olan kelimelerden oluşan, küme merkezlerini kullanarak özetler üreten çok belgeli bir özetleyici sunmuştur. Altı dilde göze çarpan bir çıkarımsal özetleyici sunulmuş ve MEAD özetleyicisi aracılığıyla bir çıkarımsal özet üretmiştir. Özetleyici, bir özellik çıkarıcı, bir cümle puanlayıcı ve bir cümle puanlama sisteminden oluşmuştur (Singh, Kumar, Mangal, Singhal ve Bilingual, 2016).

Bir MEAD uygulamasının uygulandığı bir diğer alan ise, çevrimiçi haberlerin özetlerini almaktır (Conroy ve O'Leary, 2001).

2001 yılında Hidden Markov Modeli kullanılarak, daha önce gözlemlenen bir sırayı temel alarak bir üretici sıranın modellenmesinde kullanılan istatistiksel bir araç olarak özetlenmiştir (Gong ve Liu, 2001.)

2001'de, bilgi edinme alanında başarıyla kullanılan LSA tabanlı bir özetleme önerisi sunularak metin özetlemesindeki uygulama başarılı olmuştur (Mihalcea ve Tara, 2004).

2002'de, Maxentrop, yani, maksimum entropi sınıflandırıcısının başarılı bir şekilde kullanılması cümle çıkarma işleminde tanıtıldı (Neto, Freitas ve Kaestner,2002).

Saggion & Lapalme (2002), girdi olarak ham bir teknik metin alan ve gösterici bilgilendirici özet üreten bir metin özetleme sistemi sunmuştur. İlk önce, SumUM, belgenin konu başlıklarını, ardından okuyucunun ilgisine göre bu başlıkların bazılarını ele aldı. 2000'li yıllarda bir dizi Belge Anlama Konferansı (DUC) otomatik metin özetleme konusunda çok verimli bir araştırma yaptı. Marcu (2001), hem tek dokümanlar hem de doküman koleksiyonları için söylem temelli bir özetleme algoritmaları sistemi (DUC-2001) önermiştir. Tek belgeli özetlemesi aşağıdaki adımları uygulamaktadır (Marcu, 2001):

i) Girdi olarak verilen metnin söylem yapısını türetmek.
ii) Giriş belgesindeki önemli cümleleri belirlemek

iii) Giriş belgesindeki tüm ortak referans bağlantılarını belirlemek
iv) Özet tutarlılığı ve kompaktlığı arttırmak
v) Özet oluşturmak

DUC-2001, belge koleksiyonlarını özetlemek için aşağıdaki gövdeleri kullanır (Marcu, 2001):

i) Koleksiyon üzerinde ön işleme
ii) Koleksiyonu özetleyen cümleleri seçmek ve sipariş etmek
iii) Üçüncü şahıs zamirlerini çözmek
iv) Sıra başlıkları
v) Özetler üretmek

2004 yılında Erkan & Radev, metinsel birimlerin göreceli önemini hesaplamak için stokastik bir grafik tabanlı yöntem kullanan LexRank'ı tanıttı. Erkan ve Radev (2004), cümlelerin grafik tabanlı merkezi puanlamalarına dayanarak cümlenin belirginliğini tanımlamak için yeni bir yaklaşım sunmuştur. Yazarlar, cümlelerin benzerlik grafiğini oluşturmanın, merkezi yaklaşımına kıyasla önemli cümlelerin daha iyi görülmesini sağladığını iddia etmişler.

Barzilay ve McKeown (2005) çok belgeli bir özetleme sistemi (MultiGen) içindeki "cümle füzyonu" yöntemini tartışmıştır. Cümle füzyonunun, bir cümledeki ortak ifadeleri bir araya getirmek için benzer bilgileri ve istatistiksel nesli ileten ifadeleri tanımlamak için aşağıdan yukarı yerel çok dilli hizalaması içerdiğini savunmuştur; bu nedenle, orijinal belgelerin hiçbirinde cümleleri içeren özet üretimi bulunmaz.

Fernández, SanJuan ve Torres-Moreno (2007), Metinsel Enerji üzerinde çalışılan sinir ağı olarak belgeleri modellemek için manyetik sistemlerin istatistiksel fiziğinden ilham alan bir sinir ağı yaklaşımı sunmuştur. Model otomatik özetleme ve Konu Segmentasyonunda iyi sonuçlar vermiştir.

Svore, Vanderwende ve Burges (2007), NetSum adı verilen sinir ağlarına dayalı otomatik tek belgeli özetlemeye yeni bir yaklaşım sunmuştur. Çalışma, Wikipedi ve haber sorgu günlüklerini kullanarak, özetleme için hem sinir ağlarını hem de özellikler için üçüncü taraf veri kümelerini kullanan ilk çalışmadır.

Saggion (2008), köklü değerlendirme araçlarıyla birlikte bir dizi uyarlanabilir özetleme bileşeni sunmaktadır. Araç seti (SUMMA), tek belgeli, çok belgeli, sorgu bazlı ve çok dilli özetleme için işlevsellik sağlamak amacıyla bir araya getirilen özetleme özelliklerinin hesaplanması için kaynakları içerir.

Filippova (2010), çok cümleli sıkıştırması adı verilen ve bir simgeleştirici ve etiketleyiciden daha fazlasını gerektiren bir sözdizimsel-yalın yöntem geliştirmiştir. Yöntem ne ayrıştırıcı ne de el yapımı dilsel kurallar gerektirmeyen sıkıştırılmış ve gramer cümleleri üreten grafik tabanlı bir yöntemdir.

Nenkova ve McKeown (2011), özetleme konusundaki 50 yıllık araştırmaya kapsamlı bir genel bakış sunmuştur. Ayrıca, dil üretimi ve daha derin semantik dil anlayışı gibi özetleme alanında açık olan zorlukları tartışmıştır. Özetleri kategorize ederek ve özetleme sistemlerinin nasıl çalıştığını açıklayarak başlamıştır. Daha sonra özetleme sürecinde kullanılan adımları ve yöntemleri incelemişler.

Torres-Moreno (2012), her cümle arasında bir belge vektöründen (metin konusu) ve bir sözcük vektöründen (cümle içindeki kelime) oluşan basit bir iç ürün hesaplayan otomatik metin özetleme için başka bir algoritma (ARTEX) sunmuştur. Daha sonra algoritma en üst sıradaki cümleleri toplayarak özetler üretmiştir. Algoritma, belgenin her cümlesinin belirgin bilgisini koruyarak herhangi bir dil bilgisi gerektirmez.

Litvak ve Vanetik (2014), çıkarımsal özeti için yeni bir model sunmakta ve bilgi kapsamını mümkün olduğu kadar koruyan bir özet elde etmeye çalışmıştır. Konularına göre verilen dokümanı tanımlayan yeni bir Tensor tabanlı gösterimi kullanmışlar. Daha sonra bu konular Tensor Ayrıştırması ve en üst sırada yer alan konuların cümlelerinden bir özet ile sıralanmıştır.

Torres-Moreno (2014), konunun temelleri ile başlayan ve en önemli kavramları, yöntemleri, sistemleri ve değerlendirme sistemlerini tartışan daha kapsamlı bir otomatik metin özetleme hesabı sunmuştur.

Duraiswamy (2014) çıkarımsal özetleme için ikili rasgele değişkenlere grafiksel modeli, Sınırlandırılmış Boltzmann makinesini, kullanmıştır. Bu, girdi, gizli ve çıktı üç katmandan oluşur. Giriş verileri, işlem için gizli katmana düzgün şekilde dağıtılmıştır. Deney yapılarak özet, farklı bilgi alanından ayarlanan üç farklı belge için oluşturulmuştur.

Rush, Chopra ve Weston (2015) özetsel cümle özetlemeye tamamen veri odaklı bir yaklaşım önerdiler. Metotlarında, giriş cümlesinde şartlandırılan özetin her bir kelimesini üreten yerel bir dikkat temelli model kullanılmıştır. Model yapısal olarak basit olsa da uçtan uca kolayca eğitilebilir ve büyük miktarda eğitim verisine ölçeklenebilir. Yazarlara göre, model, DUC-2004'teki ortak görevde, bazı güçlü taban çizgileriyle karşılaştırıldığında önemli performans kazanımları göstermiştir.

Chopra, Auli ve Rush (2016) bir girdi cümlesinin bir özetini oluşturan şartlı tekrarlayan sinir ağını (RNN) tanıtılar. Şartlandırma, kod çözücünün her nesil adımda uygun girdi kelimelerine odaklanmasını sağlayan yeni bir evrişimsel dikkat tabanlı kodlayıcı tarafından sağlanmıştır. Model yalnızca öğrenilen özelliklere dayanmıştır ve büyük veri setlerinde uçtan uca eğitilmesi kolaydır. Yaptıkları deneyler, modelin Gigaword 'ta DUC-2004 ortak görevi üzerinde rekabetçi bir şekilde performans gösteren son zamanlarda önerilen en gelişmiş yöntemden daha iyi bir performans göstermiştir.

Yousefi ve Hamey 2017 yılında Derin otomatik kodlayıcı (AE) kullanarak terim frekansı (tf) girişinden bir özellik alanı hesaplamak için çıkarımsal sorgu tabanlı tek belgeli özetleme yöntemlerini sundular. Yazarlara göre, yerel kelimeleri kullanan AE'nin açıkça daha ayırt edici bir özellik alanı sağladığını ve hatırlamayı ortalama %11,2 oranında iyileştirdiğini göstermiştir.

Paulus, Xiong ve Socher (2017) girdi üzerinde devam eden ve sürekli üretilen çıktıyı ayrı ayrı ele alan ve standart denetimli kelime tahmini ve pekiştirmeli öğrenmeyi (RL) birleştiren yeni bir eğitim yöntemi yeni bir dikkat ile sinir ağı modelini tanıtmıştır. Yazarlara göre standart kelime tahmini RL'nin dünya çapında dizi tahmin eğitimi ile birleştirildiğinde ortaya çıkan özetler daha okunaklı hale gelmiştir. Onlar modeli CNN / Günlük Posta ve New York Times veri setlerinde değerlendirdiler. Modellerinin, önceki modern modellere göre bir gelişme olan CNN / Günlük Posta veri setinde ROUGE-1 için 41,16 puan elde ettiklerini iddia etmiştir.

Klönne (2018), özetsel metin özetlemesine doğru İngilizce dili ve bağlamda verilen metinlerin doğru özetlerini oluşturmak için tekrarlayan sinir ağlarını kullanmıştır. Bir otomatik kodlayıcı yapı inşa eden Uzun Kısa Süreli Bellek Ağları (LSTM) tarafından takip edilen tekrarlayan sinir ağının kombinasyonunu hiyerarşik dikkat ile

birleştirilmiştir. Bu çalışma, metinleri otomatik olarak özetlemek için olası bir yükseltilebilir değişken göstermiştir.

Patel ve diğerleri (2018) metin özetlemeye yönelik tüm yaklaşımları ele alarak ve makine öğrenim yaklaşımı altında Encoder-Decoder adı verilen bir Mimarinin modus işleyişini sunmuştur. Bu mimari için, çeşitli makine öğrenimi ve derin öğrenme sinir ağı kütüphanelerinden oluşan Keras ve TensorFlow'da birçok yeni uygulama modeli önerdiler.

Khatri, Singh ve Parikh (2018) eBay ürün açıklama özetlemesi için diziden diziye (Seq2Seq) modelleri kullanmıştır. Özetsel ve çıkarımsal özetler için RNN'leri kullanan yeni bir Doküman Bağlamına dayalı Seq2Seq modelleri önerdiler ve daha iyi özetler elde etmek için Seq2Seq modellerinin girdileri ilk adımında bağlamsal bilgilerle başlatılması gerektiğini önermişlerdir. Ayrıca özetleri çıkarmak ve ölçekte Seq2Seq modellerin eğitmesinde kullanmak için yarı denetimli bir teknik önermiştir. Yarı denetimli modeller, insandan çıkarılan özetlere karşı değerlendirilir ve benzer etkinlik gösterdiği tespit edilir. Genel olarak, geniş belge özetlemesi için önerilen teknikleri kullanmak ve değerlendirmek için yöntem sundular. Yazara göre, bu teknikler mevcut teknikerlerden oldukça etkiliydi.

Luo, Guo ve Guo (2019) çalışmasında, transformatör ve anahtarlamalı normalizasyona dayalı bir metin özetleme modeli önerilmiştir. Normalizasyon katmanı optimize edilerek modelin doğruluğu artırılmıştır. Diğer modellerle karşılaştırıldığında, yeni model, kelimelerin anlamlarını ve ilişkilerini anlamada büyük bir avantaja sahiptir. İngilizce Gigaword veri setindeki deneysel sonuçlar, önerilen modelin yüksek (Recall-Oriented Understudy for Gisting Evaluation) ROUGE değerlerine ulaşabildiğini ve özetlemeyi daha okunaklı hale getirebildiğini göstermektedir.

Singh, Chhikara, ve Singh (2020) çalışması, Özetleme için 7 farklı Makine Öğrenimi Lojistik Regresyon, Sinir Ağı, Karar Ağacı, Rastgele Orman, Naive Bayes, XGBoost ve SVM temel modellerini kullanarak ve daha doğru bir tahmin yapmak için yedi makine öğrenimi modelinin tümünden gelen tahminleri birleştirerek daha iyi sonuçlar veren bir topluluk yaklaşımını önermiştir. Deney, BBC haberlerinin standart veri seti kullanılarak gerçekleştirilmiştir. 2004'ten 2005'e kadar BBC'nin 417 haber makalesini kullanarak her makale için beş özet verilmiştir. Çalışmada, makine öğrenimi

modellerini kullanarak çıkarımsal özetlemeyi kullanan araştırmaların baskın olduğu iddia edilmiştir. Bu çalışmada deneme, Google Collaboratory kullanılarak yapılmıştı. Yazarlar özet için en iyi cümleleri seçmek için birden çok makine öğrenimi modelinden toplanan sonuçlara dayalı oylama kullanmayı önermiştir. Sırasıyla 0,78, 0,66 ve 0,76 ortalama ROUGE-1, ROUGE-2 ve ROUGE-L puanları elde edilmiştir.

Tomer, & Kumar (2020) makalesinde, bulanık mantık kurallarını (çıkarıcı cümleleri seçen) çift yönlü uzun kısa süreli bellekle (Bi-LSTM) birleştirerek özetsel metin özetleri üretmek için bir yeni hibrid yaklaşım sunulmuştur. Bi-LSTM, ağ ağırlıklarını güncellemek için dikkat mekanizması ve adam optimizer kullanır. Önerilen yaklaşım, en uygun cümleleri bulmak için belgeden metinsel bilgileri elde etmek için bulanık ölçüler ve çıkarım kullanmıştır. Bu ilgili cümleler, önemli cümlelerin özetsel bir özetini üretmek için Bi-LSTM'ye girdi olarak verilmiştir. Önerilen FLSTM modeli, ROUGE kullanılarak değerlendirilmiştir. Deney, standart veri kümeleri (yani, DUC ve CNN/günlük posta) üzerinde gerçekleştirilmiştir. Bu çalışmanın bir diğer göze çarpan özelliği, daha iyi sonuçlar elde etmek için daha büyük bir veri seti oluşturmak üzere DUC 2003–2004, DUC 2006–2007 veri setlerinin birleştirilmesidir. FLSTM modeli, diğer son teknoloji modellerle karşılaştırılmıştır ve yazarlara göre deneysel sonuçlar, önerilen FLSTM modelinin diğer tüm modellerden daha iyi performans gösterdiğini ortaya koymuştur.

Sharmam ve diğerleri (2020) makalesinde, bulanık mantık teknikleriyle özeti daha güvenilir ve anlaşılır hale getirmeye odaklanmıştır. Yazarlara göre algoritmalarla birlikte çıkarmalı metin özetleme yöntemi ve bulanık mantık yöntemi, tek başına çalışan algoritmalarla karşılaştırılarak kanıtlanan denetimsiz öğrenmenin gücünü artırmıştır. Veri setini UC Irvine Machine Learning Repository, DUC2004, BBC haberlerinden sporla ilgili 1 ile 22 kb arasında boyutu ile 1000'den fazla metin dosyası alınmıştı.

Pattnaik, & Nayak (2020) makalesinde, Odia metin belgesi için etkili çıkarımsal tek belge otomatik özetleyici önerilmiştir. Hem istatistiksel hem de kümeleme yöntemleri uygulanmıştır ve değerlendirme için metrik F ölçümleri karşılaştırılmıştır. Deneysel belgeler haber alanına aittir. Yazarlara göre önerilen algoritmalar, dilin karmaşıklığı ile titizlikle ilgilenip ve özetleme problemini kayda değer ölçüde çözmüştür. Önerilen teknik, dili verimli bir şekilde analiz etmiş ve sistemi geliştirmek için bazı dilsel

özellikleri göz önünde bulundurarak istatistiksel yöntemler kullanmıştır. Değiştirilmiş cümle konumu özelliği değeriyle birlikte TF-IDF, %66.858'lik lider bir F puanına sahip olmuştur. Benzer şekilde, kümelemeye dayalı teknikte, F skor değeri değiştirilmiş TF-IDF modelinden daha düşük olmasına rağmen, fazlalık kontrolü söz konusu olduğunda etki göstermiştir. Dilin morfolojik zenginliği ve NLP araçlarının eksikliği, işleme yolunda bazı engeller oluşturmuştur. Ancak yine de sistem tatmin edici bir sonuç vermede başarılı olmuştur. Ayrıca yazarlara göre tartışılan yöntemler kendi açılarından etkilidir, ancak morfolojik olarak zengin olan Hint dilleri için hibrid bir yaklaşıma ihtiyaç vardır.

Zolotareva, Tashu ve Horváth (2020) çalışmalarında özetsel otomatik metin özetleme için BBC haber veri setinde T5 ve Seq2Seq modellerinden yararlanmıştır. Elde edilen sonuçlar ROUGE metrikleri ile değerlendirildiğinde T5 modeli daha iyi sonuçlar sağlamıştır.

Farahani, Gharachorloo ve Manthouri, (2021) çalışmasında önceden eğitilmiş transformer tabanlı kodlayıcı kod çözücü model kullanmışlardır. Bu makale, bu görevi ele almak için iki yöntem önermekte ve Farsça soyutlayıcı metin özetleme için pn-summary adlı yeni bir veri kümesini tanıtmıştır. Bu yazıda kullanılan modeller, mT5 ve ParsBERT modelinin kodlayıcı-kod çözücü versiyonudur (yani, Farsça için tek dilli bir BERT modeli). Bu modeller, pn-summary veri kümesinde ince ayarlanmıştır. Yazarlara göre mevcut çalışma türünün ilk örneğidir ve umut verici sonuçlara ulaşarak gelecekteki herhangi bir çalışma için bir temel oluşturabilir.

Ahuir ve diğerleri (2021) çalışmalarında, Katalan dilinde metin özetlenmesi için haber veri seti kullanarak önceden eğitilmiş ve ince ayarı yapılmış bir Transformer kodlayıcı-kod çözücü tek dilli bir model sunmuştur. Çalışmanın performansını Katalanca'dan daha yüksek kaynaklara sahip dillerde incelemek için, model ve deney İspanyol dili için tekrarlanmıştır. ROUGE değerleri incelendiğinde İspanyolca dilinde Katalanca diline göre daha iyi sonuçlar elde edildiği tespit edilmiştir.

Akhil (2022) çalışmasında, tek belge özetleme için çıkarımsal metin özetleme yaklaşımını tasarlamış ve uygulamıştır. Özeti anlamlı ve kayıpsız tutarak metinden önemli cümleler seçmek için Kısıtlı Boltzmann Makinesi ve Bulanık Mantık kombinasyonunu kullanmıştır. Özetlemede İngilizce metin belgelerinde anlamlı cümleler elde etmek için çeşitli cümle ve kelime seviyesi özellikleri kullanılmıştır. Her belge için

iki özet, Kısıtlı Boltzmann Makinesi ve Bulanık mantık kullanılarak oluşturulmuştur. Her iki özet daha sonra birleştirilmiş ve bir dizi işlem işlenerek belgenin nihai özetini elde edilmiştir. Sonuçlar, tasarlanan yaklaşımın etkili bir özet üreterek aşırı metin yükleme sorununun üstesinden geldiğini göstermiştir.

Pant ve Chopra (2022) çalışmalarında İngilizce, İspanyolca ve Yunanca olmak üzere 3 dil için özetleme yöntemleri geliştirmiştir. Çalışma İngilizce belgeler için çıkarımsal özetleyici bir RoBERTa modelinin ince ayarının yapıldığı ve İspanyolca ve Yunanca belgelerin özetlenmesi için T5 tabanlı modellerin kullanıldığı iki kısma ayrılmıştır. Bir mT5 modelinde, Yunanca ve İspanyolca için mT5 ve T5'te ince ayar yapılmıştır. ROUGE değerleri incelendiğinde Yunanca dilinde İspanyolca diline göre daha iyi sonuçlar elde edildiği tespit edilmiştir.

Fendji ve diğerleri (2021) çalışmalarında, SMS için T5 tabanlı bir Fransız Wikipedia Soyutlayıcı Metin Özetleme sistemi olan WATS-SMS'yi tanıtmıştır. Bu çalışmada önceden eğitilmiş İngilizce modeli olan T5'i, 25.000 Wikipedia sayfasında yeniden eğiterek ve ardından literatürdeki farklı yaklaşımlarla karşılaştırarak bir Fransızca metin özetleme modeli oluşturmak için kullanılmıştır. ROUGE metrikleri incelendiğinde, 500 karaktere kadar uzunluktaki makaleler için %52'lik bir değer elde edilirken, transformer-ED için %34,2 ve seq-2seq-dikkat mekanizması için %12,7'lik bir değer elde etmiş ve transformatörler-DMCA için %37'ye karşı daha büyük boyutlu ürünler için %77'lik bir değer elde etmişlerdir.

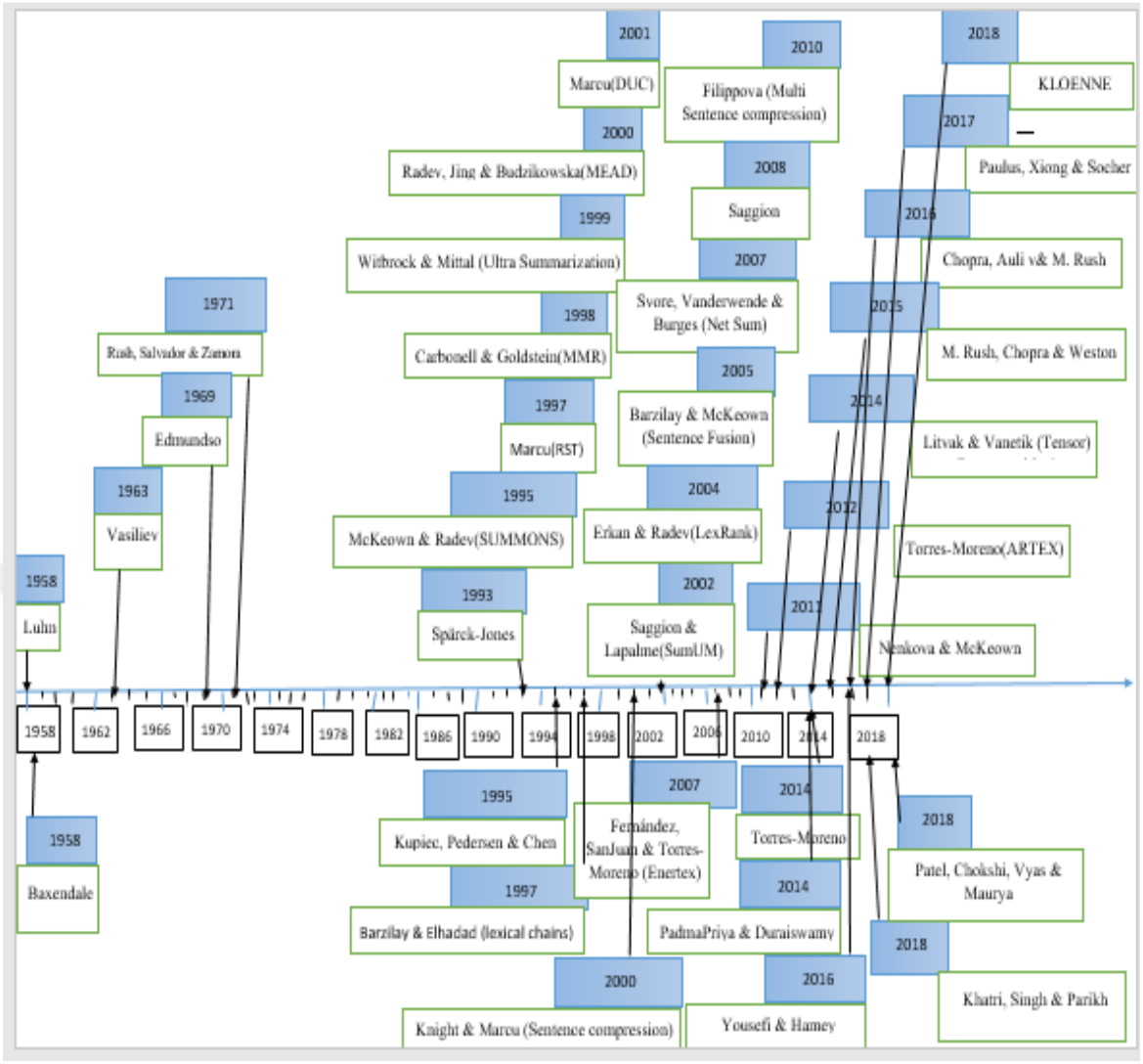
Foroutan ve diğerleri (2022) çalışmalarında çok dilli bir otomatik metin özetleme yöntemi önermiştir. Üç dilden (İngilizce, İspanyolca ve Yunanca) oluşan bir mali belge koleksiyonunda mbert ve m T5 üzerinde ince ayar yapılmıştır. Sonuçlar ROUGE değerleri ile incelendiğinde İngilizce dilinde daha iyi sonuçlar elde edilmiştir. Ayrıca genel olarak bakıldığında mT5 modeli ile mBERT modeline göre daha iyi sonuçlar elde etmiştir.

Hasan ve diğerleri (2021) yaptıkları çalışmada mT5 mimarisini kullanarak 44 dil üzerinde özetsel özetleme sistemi geliştirmiştir. Çalışmalarında XL-Sum veri seti kullanılmış ve sonuçlar ROUGE metrikleri ile incelenmiş ve İngilizce, Çince ve Hintçe dillerinde sırasıyla en iyi sonuçları elde etmişlerdir.

Bohra, Dadure ve Pakray (2022) çalışmalarında özetsel metin özetleme için T5 Transformer modelinden yararlanarak CNNDM, MSMO ve XSUM veri kümelerindeki performansını analiz etmiştir. Önerilen model, ROUGE ve BLEU puanları açısından modelin ve veri kümelerinin yeterliliğini belirlemek için veri kümeleri genelinde elde edilen çıktıyı karşılaştırdı. T5 modelinin MSMO veri seti üzerinde en iyi sonucu elde ettiği tespit edilmiştir.

Chouikhi, ve Alsuhaibani (2022) çalışmalarında çeşitli Arapça veri kümelerini uyarlayarak metin özetleme görevinde beş Son Teknoloji Arapça derin Dönüştürücü tabanlı Dil Modelini (TLM) araştırmıştır. Derin öğrenme ve makine öğrenimi tabanlı temel modellere karşı bir karşılaştırma da yapılmıştır. Deneysel sonuçlar, TLM'lerin, özellikle PEAGASUS ailesinin, birkaç kıyaslama veri setinde ortalama %90'lık bir F1 puanı ile temel yaklaşımlara karşı üstünlüğünü ortaya koymaktadır.

Otomatik metin özetleme literatürünün bir kısa gözden geçirmesinden sonra, Torres-Moreno (2014) ' dayanarak Şekil 1.2 'deki gibi otomatik metin özetleme için zaman çizelgesi çizilebilir.



Şekil 1.2. Otomatik Metin Özetleme Zaman Çizelgesi (Torres-Moreno, 2014)

1.6. TÜRKÇEDE OTOMATİK METİN ÖZETLEME ÇALIŞMALARI

NLP'nin ilk sistematik çalışmasının ve otomatik metin özetlemesinin Oflazer & Kuruöz (1994) ile başladığı, bunun içinde Türk morfolojisinin tam ölçekli iki seviyeli spesifikasyonuna dayanarak Türkçe için bir POS etiketleyici geliştirdiklerini iddia ettiler. Etiketleme aracı morfolojik belirsizlik, çok kelimeli ve deyimsel yapı tanıma, morfolojik belirsizlik, kök ve sözcüksel biçimsel derleme; ikinci ve üçüncü işlevselliklerin kurala dayalı bir alt sistem tarafından uygulandığını bütünleştirdi (Oflazer ve Kuruöz, 1994). Etiketleyici, kopyalar gibi sahte ifadeleri saptamak ve etiketlemek için sahte veya yanlış sonuçlar doğurabilecek uygun isimler gibi diğer anlamsal birleşme biçimleri gibi çok

kelimeli bir yapı işlemcisi kullandı. Yazarlar, etiketleyicinin minimum kullanıcı müdahalesiyle %98-99 doğrulukla çalıştığını iddia etmişlerdir.

Daha sonra, 2003 yılında Tür, Hakkan-Tür ve Oflazer (2003), istatistiksel dil işleme yöntemlerini kullanarak sınırsız Türkçe metinden bilgi çıkarma konusundaki çalışmalarının sonuçlarını sundu. Araştırmada geliştirilen sistem, istatistiksel modeller kullanmıştır. Çalışmanın yazarları, kelimelerin yüzey formlarını ve altta yatan formlarını, yani morfolojik yapılarını kullanan bir model inşa etmişlerdir. Bu, verileri belirten, bu belirteçleri analiz eden ve en muhtemel morfolojik analizleri veren bir ön işleme modülü kullanılarak yapıldı. Sistem, kelime tabanlı modeli ve kök tabanlı modeli bir arada kullanarak verilerdeki konu kümelerini tanımlamak için bir konu bölümleme modülü kullandı. Daha sonra sistem, adlandırılmış varlıkları, yani kişilerin adlarını, yerlerini, organizasyonlarını, parasal değerlerini, yüzdelerini, tarihlerini ve saatlerini çıkarmak için bir modül kullandı. Yazarlar, sistemlerinde %91,56 F ölçütüne ulaştığını iddia etmişlerdir.

Altan(2004) çalışmasında geliştirilen sistem, girdi olarak tek belgeli Türkçe doküman kullanmaktadır ve cümlenin konum bilgisi ve terim sıklığı bilgisi gibi özellikler kullanarak puanlanma gerçekleştirilmiştir ve bir dizi istatistiksel yöntemler kullanılarak özetler elde edilmiştir. Resmi yapıları sağlamak için tüm makalelerin içeriğini HTML belgelerine dönüştürülmüştür. Eğitim sistemi ekonomik konularda 50 farklı makaleden oluşan bir derlemi içermektedir. Belgede yüksek frekanslı bir dizi kelime aranmıştır. Başlık belirleme modül HTML biçiminde yapılmıştır. Tüm başlıklar etiketlendikten sonra dilbilimsel analiz aşamasında kullanılmak üzere kaydedilmiştir. İlk olarak noktalama işaretleri aranarak tahmini cümle sayısı elde edilmiştir. Paragraflar ve cümleler ayrıldıktan sonra başlık terimleri incelenerek, olumlu veya olumsuz cümle analizleri yapılmıştır. Kelimeleri çıkarmak için Prolog'da geliştirilen morfolojik analizör test edilmiştir, ancak her belge için önceden oluşturulmuş veri tabanından en az 250-300 kelime okuma zorunluluğu, ayrıştırma işleminin çok uzun sürmesine neden olarak iyi bir sonuç elde edilmemiştir.

Türkçede otomatik metin özetlemesi için en önemli çalışmalardan biri de Bilgin, Çetinoğlu ve Oflazer'in (2004) Türkçe için bir WordNet geliştirmeye girişimidir. Yazarlar EuroWordNet projesinin 1310 temel konseptinden oluşan bir "ilk kavramlar

seti" ile başlamıştırlar. İlk 1310 temel kavram kümesini çevirdikten sonra, yazarlar bu temel kavramlar için otomatik olarak eşanlamlıları, zıt anlamlılarını çıkarmaya çalışmış ve daha sonra, WordNet'i ikinci bir aşamada 5000 temel konseptine ve ardından üçüncü bir aşamada 8000 temel konseptine genişletmiştir. Sistem'de %68 doğruluk elde edilmiştir.

Karakaya ve Güvenir (2004), büyük metin gruplarından bilgi toplamak ve çıkarmak için metin sınıflandırma ve metin özetlemeyi bir araya getirme çalışması yaptı. ARG adlı sistem, paragrafların verilen konulara göre sınıflandırıldığı ve her konu otomatik olarak oluşturulan bir raporda özetlendiği iki aşamalı bir algoritmadır. ARG, literatürdeki diğer sistemlerden farklı olarak, tüm metnin yerine analiz birimi olarak paragraflar kullandı. Ayrıca, kullanıcıların bilgi ihtiyaçlarını ifade etmede daha iyi bir seçenek sunmalarını sağlamak için metin sınıflandırmaya bir kullanıcı denetimi kullandılar. Karakaya ve Güvenir'in (2004) ARG'de açıkladığı gibi;

I. Kullanıcı konu başlıklarını belirler.
II. Kullanıcı bir veya daha fazla makaleyi paragraflarına ayırır ve bunları konulara dağıtır.
III. Her konu ii. Adımdaki paragrafları kullanılarak endekslenir.
IV. Diğer belgeler paragraflara ayrılmıştır.
V. İv adımdaki paragraflar, ii. adımda verilen konu başlıklarına göre sınıflandırılır.
VI. Her konu özetlenir.
VII. Özetler derlenir ve rapor olarak çıkarılır.

Bununla birlikte, bu algoritma kullanıcıdan çok fazla müdahaleye ihtiyaç duyar ve büyük metin gruplarının sınıflandırılması ve özetlenmesi kullanıcı için bir yük olabilir.

Bir başka metin sınıflandırma çalışması, Amasyalı ve Diri (2006) tarafından, belgelerin yazarlarını ve yazarların cinsiyetlerini belirlemek ve belgelerin türünü sınıflandırmak için yapılmıştır. Dokümanları sınıflandırmak için n-gram model ve Naïve Bayes, Destek Vektör Makinesi, C4,5 ve Random Forest yöntemlerini kullandılar.

Başka bir çalışmada, Ercan (2006), önemli cümle ve anahtar cümle tanımlamaya odaklanan bir çıkarımsal özetleme sistemi sundu. Sistem, sözcük bağlılığı ve metinlerin konularını kullandı ve bölümlerini sözcük zincirleri aracılığıyla tanımladı. Çalışma

WordNet' teki kelimelerin özgüllüğü üzerine odaklandı ve bu özellik puanının kullanılması anahtar cümlecikleri tanımladı. Çalışma aynı zamanda aktörler, yerler ve diğer nesneler arasındaki ilişkileri, yani sözcüksel zincirler yaklaşımıyla yakalanamayan bir ilişki durumun katılımcıları arasındaki bağlantıyı yakalamak için ortak oluşum analizini kullandı. Yazar, sistemin genel olarak ümit verici sonuçlar aldığını iddia etmiştir.

Ercan ve Çiçekli (2007), kesinliği önemli ölçüde artıran sözcüksel zincir özelliklerini kullanan bir anahtar sözcük çıkarma yöntemi tanımlamıştır. Yazarlar, hâlihazırda Word Net'te gösterilmeyen terimler dışında sözcük zincirleri oluşturmanın bir yolunu bulmaya çalıştılar.

Pembe (2010) arama motorları için bilgi isteğine ve metin yapısına dayalı olarak otomatik doküman özetlenmesi önermiştir. Çalmadaki sistem, ilk aşamada her doküman, hiyerarşilerinin ilgili başlık ve alt başlıklarla birlikte ortaya çıkarmak için yapısal olarak işlenmesidir bunun için kural tabanlı bir yaklaşım oluşturulmuştur. Daha sonra, başlık ve hiyerarşi çıkarma işlemlerinin başarımına göre geliştirilen ağaç gösterimine dayalı bir makine öğrenmesi ve destek vektör makineleri ile algılayıcı algoritmalarını kullanan bir yaklaşım değerlendirilmiştir. İkinci aşamada, ilk aşamada ortaya çıkan doküman yapıları ile otomatik özetleme yöntemlerinin geliştirilmesi için cümle bazında puanlama ve bölüm bazında puanlama ile değerlendirmiştir. Yazarlara göre çalışmadan elde edilen sonuçlar Google özetleri ve bilgi isteğine yönelik özetlere göre, doğruluk açısından daha iyi sonuçlar ortaya çıkaracağını tespit etmiştir. İkinci koleksiyon (Türkçe Koleksiyon), Google'in sonuçlarından Türkçe için tanımlanmış TREC benzeri sorgular kullanılarak toplanan Türkçe Web dokümanlarını içerir (Markey, 2007). Koleksiyon, 250 belgeden rastgele seçilen 50 belgeyi içermektedir. Makine öğrenimi algoritmaları için daha büyük bir belge koleksiyonuna ihtiyaç vardı. Bu belge koleksiyonları hem yapısal işleme hem de özetleme yönteminin değerlendirilmesinde kullanılmıştır. Konum, Başlık (başlıkta geçen terimlerin cümlede geçme sıklıkları), sorgu cümlesi ve terim sıklığı (cümlede geçen terimlerin tüm dokümanda geçme sıklığından elde edilen değer) ve metotlarını kullanarak cümleleri puanladıktan sonra cümlelerin önemine göre puan verilmiş ve cümle seçimi gerçekleştirilmiştir. Türk Koleksiyonu için başlık çıkarma sonuçları Anma, Duyarlılık ve F-ölçüsü (Recall Precision F-measure) için sırayla 0,79, 0,57, 0,65 bu değerler elde edilmiştir. Yöntemlerin genel olarak Türkçe belgeler üzerinde çalışıp

çalışmadığını test etmek için bir Türk üniversite web sitesindeki belgeler üzerinde (boun.edu.tr alanındaki 50 belge) ek bir analiz gerçekleştirerek ve önerilen yaklaşımı kullanarak %71 doğruluk elde edilmiştir, Türkçe Web sayfaları için algoritmanın sağlamlığını kanıtlamıştır.

Kutlu, Çığır ve Çicekli (2010) cümle sıralaması yoluyla genel bir metin özetleme yöntemi önermiştir. Sistem, cümle puanlarını yüzey seviyesindeki özelliklerine göre hesaplamış ve orijinal belgelerden en üst sıradaki cümleleri çıkararak özetler oluşturmuştur. Sistem, terim sıklığı, anahtar ifade merkezi, başlık benzerliği ve cümle konumu gibi özellikleri kullandı. Çalışma, merkeziyet özelliğinin etkinliğini gösterdiği ve Türkçe 'de metin özetlemede anahtar ifadeleri kullanıldığını ortaya koyan birçok yönden bir ilk oldu.

Özsoy, Çiçekli ve Alpaslan (2010) iki yeni LSA temelli özetleme algoritması önermiştir. LSA'ya dayalı genel bir çıkarımsal Türkçe metin özetleme sistemi sundu. Yazarlar, araştırmada tasarlanan Cross metodun, diğer LSA yöntemlerinden daha iyi olduğunu iddia etmişlerdir.

Uzun-Per (2011), Türkçe için bir konsept çıkarma sistemi önermiştir. İlk olarak Türk alfabesinin karakterlerine ön işlem yapılmış, sonra sadece çekim morfemlerinden sıyrılan isimler kullanılmış ve devamında bu isimler K-Means algoritması ile kümelenmiş ve bir kullanıcı ara yüzü kullanılarak bu kelime kümesine kavramlar atanmıştır. Daha sonra, dokümanlar topluluğu, algoritmanın önceki aşamasında belirlenen kavramların listesine göre test edildi. Yazar, çalışmada tasarlanan sistemin, %41 kesinliğe ulaştığını ve bunun düşük olduğunu ancak literatürdeki diğer çalışmalardan daha yüksek olduğunu iddia etmiştir.

Farklı bir çalışmada, Demir ve diğerleri (2012), özetler oluşturmak için önemli bir adım olan ilk Türk eşanlamlı ifadeler derleme (paraphrase corpus) inşa etmek çabalarını sundu. Çok sayıda edebi metin çevirisinden, bir filmin iki farklı altyazısından, paralel bir derlemeden çok sayıda referans çevirisinden ve haber cümlelerinin insan tarafından yazılmış ifadelerinden çizdiler.

Güran (2013) çıkarımsal metin özetleme için gizli anlamsal analiz temelli metin özetleme yöntemlerinde kullanılabilen yeni bir ağırlık değeri önermiştir. Yazara göre önerilen yeni ağırlık değeri, dört farklı veri seti üzerinde tüm yöntem başarımlarını

arttırdığı gösterilmiştir. Bu çalışmada aynı zamanda önemli cümle çıkarımı için anlamsal ve yapısal özelliklerin birleşimini sağlayan iki farklı yaklaşım ile bir melez sistem önerisi de yapıldı. Deneysel sonuçlar, her bir özelliğin bireysel kullanımından daha iyi bir başarı elde etmesini göstermiştir.

Birant (2015) Türkçe metinlerde otomatik metin özetleme için geliştirilmiş anlam bilimsel ilişki sözlükleri bir yazılım kullanmıştır ve bu konuda yeni yöntemler kullanan yazılımları da geliştirmiştir. Yazılımın geliştirilmesi Türk Dil Kurumu tarafından onaylanmış olan Eş ve Yakın Anlamlılar ve Zıt Anlamlılar sözlüklerin sayesinde tamamlanmıştır. Bu Kural Tabanlı Yazılımının ortalama başarısını ölçmek için ROUGE-N yöntemi kullanılmıştır. Yazara göre çalışmadan ortaya çıkan veriler hem kişisel özetlere benzerlik göstermiş hem de önceki verilere yakın değerler elde etmiştir.

Baydar (2018), otomatik metin özetleme çalışmalarını, yazılımlarını, algoritmalarını, yaklaşımlarını ve metotlarını ortaya koymuştur. Bu çalışmanın amacı Türkçe metinlerde çıkarımsal özetleme yönteminin başarısının artırımıdır. Genetik algoritma kullanımıyla üç farklı yöntem sezgisel olarak belirlenen rastgele puan aralıklı değerlendirme, sabit puanlı değerlendirme ve genel aralıklı rastgele puanlama denenmiştir. Yazara göre yapılan ölçümlerde başarının arttığı tespit edilmiştir.

Abdı Omar (2018) otomatik metin özetlemesi için dört farklı algoritmanın TextRank, Sumy LSA, bir frekans olay tabanlı özetleme ve bir duygu analizi tabanlı özetleyici karşılaştırmasını ve özetleyicilerin değerlendirilmesini içerir. Değerlendirme ölçütleri ROUGE puanlarıyla karşılaştırılmıştır.

Aysu (2018) Türkçe haber metinleri üzerinde otomatik metin özetleme işlemini uygulamıştır. Çalışmanın performans testi için 20 kişiden, 50 farklı haber metninden önemli buldukları 3 cümle seçmeleri istenmiştir. Sonra kişilerden ortaya çıkan sonuç ile çalışmadan elde edilen sonuç karşılaştırılmıştır. Çalışmanın sonuçları özetleme performansını yüzde otuzun altı göstermiştir.

Özkan (2019) tez çalışmasında, Türkçe kelimeleri işlerken, kelimeleri harf harf karşılaştırarak, uyuşan harfler ve bu harflerin kelime içindeki konumlarına göre sağlıklı bir sonuca ulaşılmaya çalışmıştır. Veri seti için “f5haber.com” sitesinden, “hürriyet.com.tr” haber portalına ait Mart -Nisan 2018 tarihleri arasında yayınlanan toplam 1.005 adet haber kullanılmıştır. Veri seti oluşturduktan sonra, cümle noktalama

işaretleri normalizasyonu yapılmış, kelime sıklık listesi oluşturulmuş, cümle puanlaması yapılmış ve 3 cümle insan tarafından seçilerek referans özet oluşturmak için veri tabanına eklenmiştir. Sonuçları değerlendirmek için bir insan tarafından elde edilen özetlerle karşılaştırılıp ve değerlendirilmiştir. ROUGE-N ölçüm sonuçlarına bakıldığında anma, duyarlık ve F1 ölçümü için sırasıyla 0,5078, 0,3951 ve 0,4442 elde edilmiştir. Ayrıca, ROUGE-N Tam Puan RECALL için 23,66 ve PRECISION için 23 elde edilmiştir.

Cengiz ve diğerleri (2019) çalışmalarında tek belgeli çıkarıcı metin özetleme için denetimsiz olarak çizge tabanlı genel, cümlelere yönelik bir ağırlıklandırma önermiştir. Cümle puanlama aşaması, önemli düğümleri belirlemek için farklı düğüm ağırlıklandırma yöntemlerini kullanarak gerçekleşmiştir. Document Understanding Conference (DUC-2002) veri seti üzerindeki performansı ROUGE-2 değerlendirme metrikleri kullanılarak duyarlılık, kesinlik ve f-skor değerleri sırasıyla 0,17068, 0,15772, 0,16383 olarak hesaplanmıştır.

Erhandi'nin (2020) derin özetleme ile metin özetleme çalışmasında derin oto kodlayıcı yapısı, ara katmanlarda ise LSTM yapısı, IDE olarak PyCharm, ve Kütüphane olarak Keras kullanarak açık kaynak bir projenin üzerine çalışmanın modeli geliştirilmiştir. Türkçe için analizler yaparak iyi sonuçlar elde edilmiştir.

Afatsun (2020) çalışmasında, türkçe haber veri seti (THV) toplamak için Python dilinde bir program yardımı ile Deutsche Welle haber sitesinden 50000 üzerinde metinleri ve özetleri kullanmıştır. Bu çalışmada soyutlayıcı metin özetlemesi için dikkat katmanlı kelime gömmeleri kullanarak eğitilmiş çift yönlü bir LSTM model geliştirilmiştir. Modelin performansı hem THV'deki kelimeler kullanılarak hem de Wikipedia eğitilmiş kelime vektörleri ile ayrı ayrı değerlendirilmiştir. ROUGE-1 metriğine göre modelin THV'deki performans puanı 40,90'dır. Ayrıca Modeli doğrulamak için GigaWord ve CNN/Daily Mail veri kümeleri ile de deneyler yapılmıştır.

Karaca (2021) " derin öğrenme yöntemi ile Türkçe haber metinlerine otomatik olarak başlık üretme" çalışmasında veri seti olarak SuDer den özetlemek için uygun olan haber metinlerini, kütüphane olarak Keras kütüphanesi, eğitim için transformatör mimarisi ile soyut özetleme yöntemi kullanmıştır. Modelin 20 ve 25 dönem eğitimden sonra sırasıyla yaklaşık %75 ve %85 oranında doğruluğa ulaşarak, haber metinlerindeki bağlamı ifade etmekte yetenekli başlıklar üretebildiği elde edilmiştir.

Özkan Çelik (2021) tez çalışmasında Türk Kütüphaneciliği ve Bilgi Dünyası dergilerinden 421 bilimsel makaleyi veri seti olarak kullanarak otomatik ve yapısal yöntemlerle özetler elde etmiştir. Çalışmada Zemberek kullanılarak kök bulma işleminden sonra cümleler, kelime sıklıklarına göre derecelendirilmiştir. Çalışmaya göre ortaya çıkan otomatik yapısal özet sonuçlarından sistemin klasik özet çıktılarına göre daha iyi sonuçlar elde edilmiştir. Tüm çıktıların hesaplanan okunurluk değerleri yazar özlerinin okunurluk değerlerinden belli bir şekilde daha kolay okunurluk seviyesindedir.

Ertam (2022), Galip Aydin çalışmasında Türkçe özet özetleme için derin öğrenme yaklaşımını sunmaktadır. Önerilen yaklaşım iki aşamadan oluşmaktadır. İlk aşamada bir haber ajansından Türkçe haber içeriği toplanmıştır. Web kazıma yöntemleri ile veri toplamak için python tabanlı bir yazılım geliştirerek ve haber başlığı, kısa haber, haber içeriği ve son 5 yıldaki anahtar kelimeleri tarayarak çeşitli haber ajansı web portallarından bir Türkçe haber özetleme kıyaslama veri seti oluşturmuştur. İkinci aşamada, toplanan veriler kullanılarak Türkçe özet özetleme yapılmıştır. Çalışmada kullandığımız yaklaşım diziden diziye (Seq2Seq) modelidir. Bu teknikte, bir dizi girişini bir etiket ve dikkat değeri (tag and attention value) ile bir dizi çıkışına senkronize etmeye çalışan kodlayıcı-kod çözücü (encoder-decoder) tabanlı bir makine çevirisi yöntemidir. Amaç, özel bir belirteçle (token) çalışacak RNN veya LSTM gibi derin öğrenme yapılarını kullanmak ve önceki kümeden sonraki durum kümesini tahmin etmeye çalışmaktır. Hazırlanan veri setinden haber başlıkları ve kısa haber içerikleri kullanılarak derin öğrenme ağı eğitimi gerçekleştirilmiş ve ardından ağın kısa haber özeti olarak haber başlığını oluşturması beklenmiştir. Çalışmanın performans değerleri, her bir cümle için ve ayrıca rastgele seçilmiş 50 cümle için sonuçların ortalaması alınarak sunulmuştur. F1 ölçü değerleri ROUGE-1, ROUGE-2 ve ROUGE-L için sırasıyla 0,4317, 0,2194 ve 0,4334'tür. Yazarlara göre elde edilen sonuçlar, önerilen yaklaşımın Türkçe metin özetleme çalışmalarında etkili olduğunu göstermektedir. Gelecekte, büyük belgelerden daha iyi ve daha uzun Türkçe özetler yapmak için kendi soyut özetleme çerçevelerini oluşturmayı düşünmektedirler.

Baykara ve Güngör, (2022) çalışmalarında metin özetleme görevi için önceden eğitilmiş Seq2Seq modelini iki büyük ölçekli Türkçe veri seti olan TR-News ve MLSum üzerinde incelemiş ve sonuçları elde etmiştir. Ardından, veri kümelerindeki başlık bilgisini kullanarak ve her iki veri kümesinde de başlık oluşturma görevi için somut

temeller oluşturulmuştur. Ek olarak, çapraz veri kümeleri değerlendirmeleri, çeşitli metin oluşturma seçenekleri ve Türkçe için ROUGE değerlendirmelerinde ön işlemenin etkisi dahil olmak üzere modellerin kapsamlı analizi sağlanmıştır. Tek dilli BERT modellerinin, tüm veri kümelerindeki tüm görevlerde çok dilli BERT modellerinden daha iyi performans gösterdiği gösterilmiştir. Son olarak, üretilen özetlerin nitel değerlendirmeleri ve modellerin başlıkları verilmiştir.

İKİNCİ BÖLÜM

DERİN ÖĞRENME

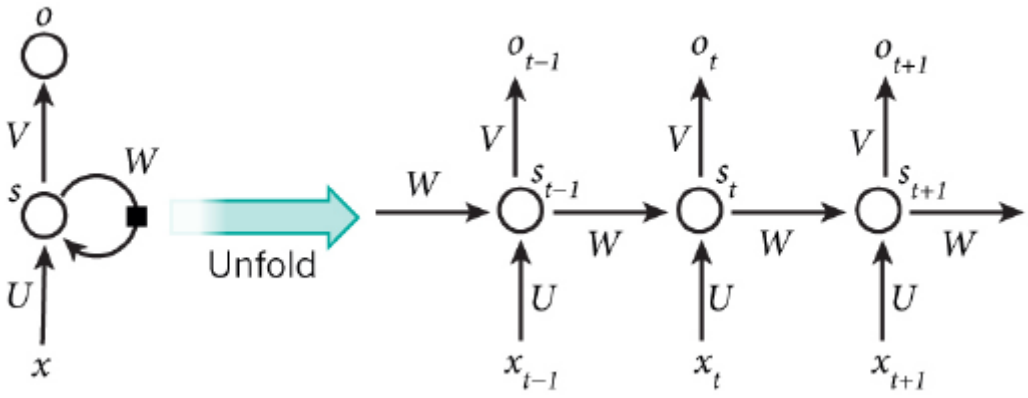
2006'dan beri derin yapılandırılmış öğrenme veya daha yaygın olarak derin öğrenme veya hiyerarşik öğrenme olarak adlandırılan, denetimli veya denetimsiz özellik çıkarımı ve dönüşümü, kalıp analizi ve sınıflandırması için birçok doğrusal olmayan bilgi işlem katmanını kullanan yeni bir makine öğrenme araştırması alanı olarak ortaya çıkmıştır (Hinton, Osindero ve Teh, 2006, Bengio, 2009, Deng ve Yu, 2014). Derin öğrenme, verilerin kendisinden öğrenerek verilerin gizli gösterimini keşfetmeyi amaçlar. Derin Öğrenme, el yapımı özellik çıkarım yöntemlerine ihtiyaç duymadan veri sunumlarını öğrenebilen sistemlere odaklanır. Yanı derin öğrenme yöntemleri, özelliklerin manuel olarak çıkarıldığı klasik makine öğreniminin aksine, ilgili özellikleri otomatik olarak ham verilerden çıkarır. Bu hiyerarşik öğrenmeye dayanarak, algoritmalar, hedef değerleri tahmin etmek için verilerdeki gizli kalıpları ve özellikleri algılayabilir. Örneğin, borsa tarihi koşulları yönlü hareketler olarak etiketlenebilir, o zaman bu modele göre değerlendirilen mevcut koşullar yarının borsa davranışını tahmin etmek için kullanılabilir. Öğrenme algoritmalarının temel kaygısı temsil ve genellemedir. Derin öğrenme sistemleri veri örneklerini girdi olarak alır ve görünmeyen veri örneklerini genelleştirmekten çok bu örnekler üzerinde değerlendirilen işlevleri temsil eder. Denetimli öğrenme, denetimsiz öğrenme, takviyeli öğrenme kategorileri altında birçok öğrenme algoritması uygulaması vardır (Karaoglu, 2018). Derin öğrenme yöntemleri, doğrudan veriden sınıflandırma gibi görevleri gerçekleştirmeyi örnek alarak öğrenir. Derin öğrenme ve veri arasındaki ilişki bütünleştiricidir, daha fazla veri genellikle sonuçların verimliliğini artırır. Ancak, en az veri kullanımıyla en iyi sonuçları elde etmek en iyi çözümdür.

2.1. TEKRARLAYAN SİNİR AĞLARI (RNN)

Geleneksel sinir ağları verilen örneklerden görevin anlaşılmasını sağlarken önceki işleri hatırlamaz. Bununla birlikte, metin özetleme gibi görevler için, giriş belgelerindeki kelimelerin sırası önemlidir. Bu durumda, modelin bir sonrakini işlerken önceki kelimeleri hatırlaması gerekir. Bunu başarabilmek için, 1980'lerde gelen benzersiz bir

sinir ağı türü, NLP alanında devrim yaratan, normal ileri beslemeli sinir ağlarının aksine geriye dönük bağlantıları olan Tekrarlayan Sinir Ağları (RNN) kullanılmak zorundadır, çünkü bunlar modelde bilginin devam edebileceği halkaları olan ağlardır (Olah, 2015; Tarwani ve Edem, 2017). RNN'ler mevcut ve geçmiş verilere dayanarak geleceği tahmin edebilen bir sinir ağı sınıfıdır. RNN'ler, grafik döngüleriyle zamanın ilerlemesini sağlayan yoğun bağlantı modellerinden oluşan gruptur (Lipton, Berkowitz, ve Elkan, 2015). Bir dizideki özetin oluşturulması, verilerle ilgili bazı önceki bilgilere ihtiyaç duyar. Böylece, Tekrarlayan Sinir Ağları, veri bağımlılığını gösteren görevler için iyi çalışır (Graves vd., 2013). Veri bağımlılığına ek olarak, bu ağ değişken uzunluk girişi için de iyi çalışır.

RNN'nin önceki verileri hatırlayabilmesi, onu dinamik bir dizi modelleme aracı haline getirmiştir. Şekil 2.1 açılmış RNN'yi göstermektedir (Iqbal, Sambyal ve Devanand, 2018).



Şekil 2.1. Açılmış RNN (Iqbal, Sambyal ve Devanand, 2018)

Tekrarlayan Sinir Ağının açılması, dizinin tamamı için yazılmış olduğu anlamına gelir. 10 kelime içeren cümle dizisi için ağ her bir kelime için açılacak veya 10 kat olacaktır. Burada hesaplamayı yapan formüller şunlardır: $S_t = F (U * X_t + W * S_{t-1})$, $O_t = \sigma (W_t * S_t)$, X_t, t zamanındaki giriş ve S_t, t adımındaki gizli durumları temsil eder ve "O_t", her zaman damgasından sonra çıkış birimlerinin değerini gösterir (giriş bir dizgelerin listesi ise, her zaman damgası bir kelimenin işlenmesi olabilir). Ek olarak, S_t, önceki zaman adımlarında hesaplanan bilgileri depolayan ağın bir hafızası olarak düşünülebilir. İşlem sırasında, RNN bilgi elemanını seçici bir şekilde dizi basamakları üzerinden geçirme gücüne sahiptir. Bağımlı olan elementlerin sırası giriş ve / veya çıkış olarak

temsil edilir. Yani Şekil 2.1 önceki zaman damgasından sonucu, sinir ağının bir yığınında gerçekleşen hesaplamanın bir kısmı için bir sonraki aşamaya geçirildiğini göstermektedir. Bu nedenle, bilgi önceki zaman damgasından alınır. Ek olarak, RNN sırayla ve zamana bağlılığı aynı anda farklı ölçeklerde modellemektedir (Lipton, Berkowitz ve Elkan, 2015, Khomenko, vd., 2016). Zaman serisi verileri, metin verileri, ses verileri, stok verileri vb. gibi çeşitli ardışık veri türlerini analiz etmek için donatılmıştır. Duygu analizi, örüntü tanıma, konuşma tanıma, metin özetleme ve sentez dili modelleme gibi doğal dil işleme sistemleri için son derece faydalıdır (Lipton, Berkowitz ve Elkan, 2015, Khomenko v.d., 2016). Bununla birlikte, pratikte, geleneksel RNN'ler genellikle, bağlanan bilgiler arasındaki artan mesafe ile bilgiyi verimli bir şekilde ezberlemezler. Her bir etkinleştirme işlevi doğrusal olmadığından, bilgi almak için yüzlerce veya binlerce işlemi geri izlemek zordur.

2.1.1. RNN'lerin Avantaj ve Dezavantajları

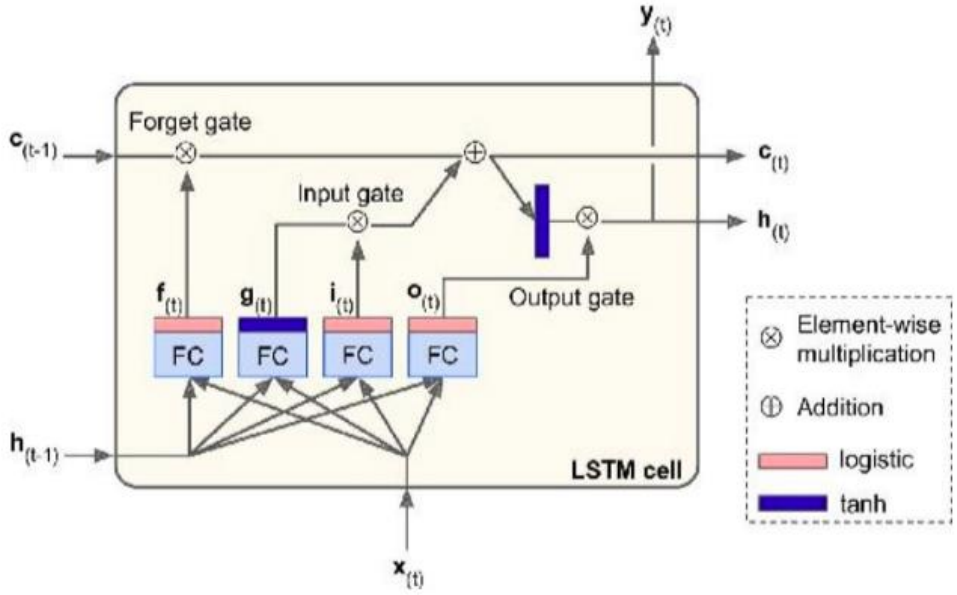
RNN'ler, diziler ve zaman serileriyle başa çıkmak için üretilmiştir. Özellikle bu amaçlar için, önceki durumlar hakkında bilgi sahibi olmak gerekebilir, örneğin bir cümlenin bir sonraki kelimesini oluşturmak için. Temel RNN'ler pratikte sadece uzun süreli bağımlılıklar denilen bir problem nedeniyle, birkaç zaman adımında sadece bir 'hafızayı' tutabiliyorlardı, Hochreiter ve Schmidhuber bir sıradaki (önceki) aşamaları hatırlamak için LSTM'leri- Uzun Kısa Süreli Hafıza Ağları- icat etti (Hochreiter vd., 2001).

2.2. UZUN KISA SÜRELİ HAFIZA (LSTM):

Uzun Kısa Süreli Bellekler (LSTM) uzun vadeli bağımlılık probleminden kaçınmak için tasarlanmıştır. LSTM, önceki modellerle uyuşmayan ardışık verilerle ilgili çeşitli öğrenme sorunları için yoğun ve genişletilebilir bir model göstermiştir. LSTM'ler genellikle uzun vadeli bağımlılıkların ele alınmasında çok yönlüdür. Dil görüntüleme ve yorumlama, söylem gösterme akustiği, söylem karışımı, protein yardımcı yapı tahmini, ses ve video bilgilerinin araştırılması gibi bazı konular için yüksek ilerlemeyi zorlayan ağ boyunca sürekli hata akışını sürdürerek geliştirme zorluklarının üstesinden gelirler. LSTM ilk önce Hochreiter ve Schmidhuber (1997) tarafından önerildi, bu noktadan sonra

ilk LSTM'de bir takım küçük değişiklikler yapıldı (Chung, Gulcehre, Cho, and Bengio, 2014). LSTM, tekrarlayan gizli katmanda bellek blokları adı verilen olağanüstü birimler içeren sıra dışı bir RNN türüdür (Sak, Senior ve Beaufays, 2014). Bu bellek blokları, ağın geçici durumunu saklayan öz ilişkilendirmelere sahip bellek hücrelerini içerir. Bellek hücrelerine rağmen, veri akışını kontrol eden Kapılar adı verilen benzersiz çarpma birimleri içerir (Sak, Senior ve Beaufays, 2014). Orijinal mimarideki her bellek bloğu bir giriş kapısı ve çıkış kapısı içerir. Geleneksel RNN'den farklı olarak, her bir LSTM hücresinin içinde, karmaşık hesaplama yapmadan verilerin taşınmasına izin veren birkaç basit doğrusal işlem vardır.

LSTM hücreleri ile temelin farkı, bir yerine iki durum vektörüne sahip olmaları ve ayrı tutulmalarıdır. LSTM hücresinin mimarisi şekil 2.2'de gösterilmektedir. LSTM hücrelerinin durumu, h ve c olarak iki vektöre ayrılır. H, kısa vadeli bir durum olarak ve c uzun vadeli durum olarak düşünülebilir. Bir LSTM hücresine sahip olmanın ana fikri, modelin neyi hatırlayacağını ve neyi unutacağını belirlemesini sağlamaktır.



Şekil 2.2. LSTM Mimarisi (Hafsari, 2020)

Zaman aşaması t için, önceki zaman aşamasının c (t-1) hücre durumu, ağı soldan sağa geçirir ve önce unutma geçidine girer, bazı hatıraları bırakır ve sonra ekleme işlemi ile bazı yeni anılar ekler. İlave işleminden sonra ortaya çıkan c (t), tanh fonksiyonundan geçirilir ve sonuç, çıkış kapısı tarafından filtrelenir. LSTM g (t) çıkışlarındaki ana katman, x (t) ve önceki kısa vadeli durumu h (t-1) girişlerini analiz etme rolüne sahiptir.

LSTM'deki ilk adım, hangi bilgileri hücre durumundan atacağımıza karar vermektir. Bu karar unutma kapısı tarafından verilir (f (t) ile kontrol edilir). Bir sonraki adım, hücre durumunda hangi yeni bilgileri depolayacağına karar vermektir. Giriş kapısı (i (t) tarafından kontrol edilir) g (t) 'nin hangi kısımlarının uzun vadeli duruma ekleneceğini kontrol eder. Son olarak, çıkış kapısı (o (t) ile kontrol edilir), uzun vadeli durumun hangi bölümlerinin okunması gerektiğini ve çıkış y (t) olarak ürettiğini kontrol eder. LSTM hücresi, önemli bir girdiyi tanıma, uzun vadeli bir durumda saklama ve gerektiğinde koruma ve çıkarmayı öğrenme yeteneğine sahiptir.

$$i(t) = \sigma(Wxi\ T.x(t) + Whi\ T.h(t-1) + bi) \quad (2.6)$$

$$f(t) = \sigma(Wxf\ T.x(t) + Whf\ T.h(t-1) + bf\ ) \quad (2.7)$$

$$o(t) = \sigma(WxoT.x(t) + WhoT\ .h(t-1) + bo) \quad (2.8)$$

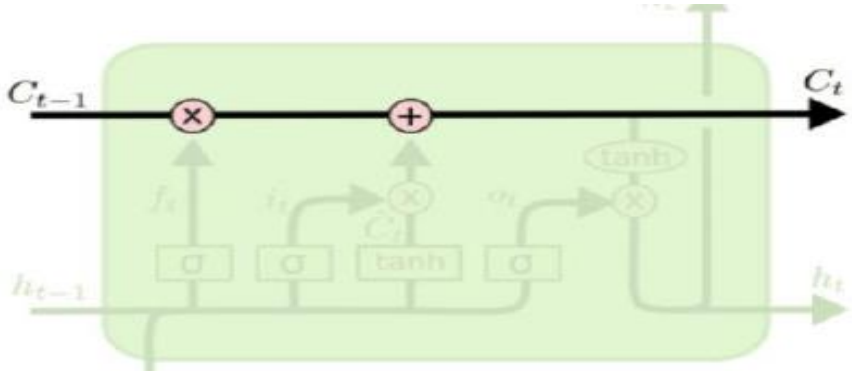
$$g(t) = \tanh(Wxg\ T.x(t) + Whg\ T.h(t-1) + bg) \quad (2.9)$$

$$c(t) = f(t) \otimes c(t-1) + i(t) \otimes g(t) \quad (2.10)$$

$$y(t) = h(t) = o(t) \otimes \tanh(c(t)) \quad (2.11)$$

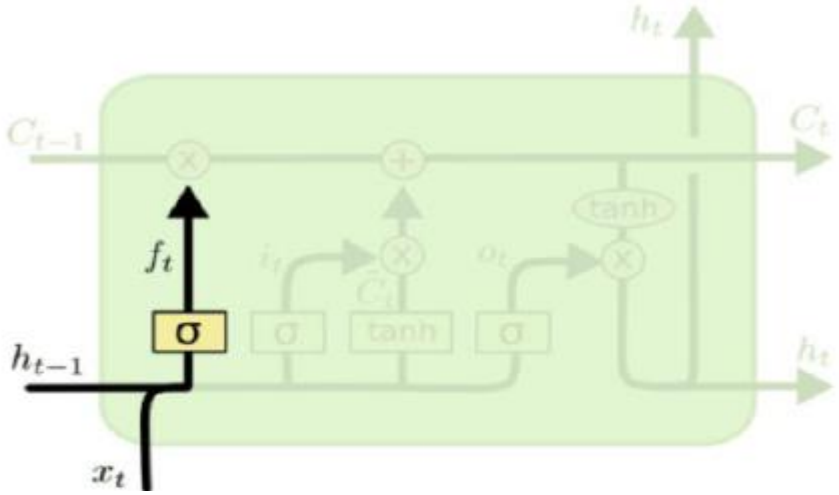
Wxi, Wxf, Wxo, Wxg, giriş katmanına x (t) bağlanması için dört katmanın her birinin ağırlık matrisleridir. Whi, Whf, Who, Whg, önceki kısa vadeli durum h (t-1) ile bağlantısı için dört katmanın her birinin ağırlık matrisidir. Ayrıca tüm eğilimler ilgili katmanlar için mevcuttur.

Şekil 2.3'te gösterildiği gibi, şu ana kadar tüm bilgileri içeren önceki hücre durumu, bazı doğrusal işlemler yaparak bir LSTM hücresinden sorunsuzca geçer. İçinde, her bir LSTM hücresi, açılan ve kapanan üç kapı vasıtasıyla hangi bilgilerin saklanacağı ve ne zaman bilginin okunmasına, yazılmasına ve silinmesine izin verileceği hakkında kararlar verir. Giriş kapısı, hafıza hücresine bilgi harekete geçirme akışını içerir. Çıkış kapısı, ağda kalanlar için hücre aktivasyonunun çıkış akışını kontrol eder. Daha sonra unutma kapısı bellek bloğuna eklendi. Unutma kapısı, hücreye eklenmeden önce hücrenin iç boyutunu ölçeklendirir (Dey ve Salem, 2017). LSTM, kaybolan ve patlayan gardiyan (vanishing and exploding gradient) sorununu aşan sistemde tutarlı hata akışına izin verir (Hochreiter ve Schmidhuber, 1997).



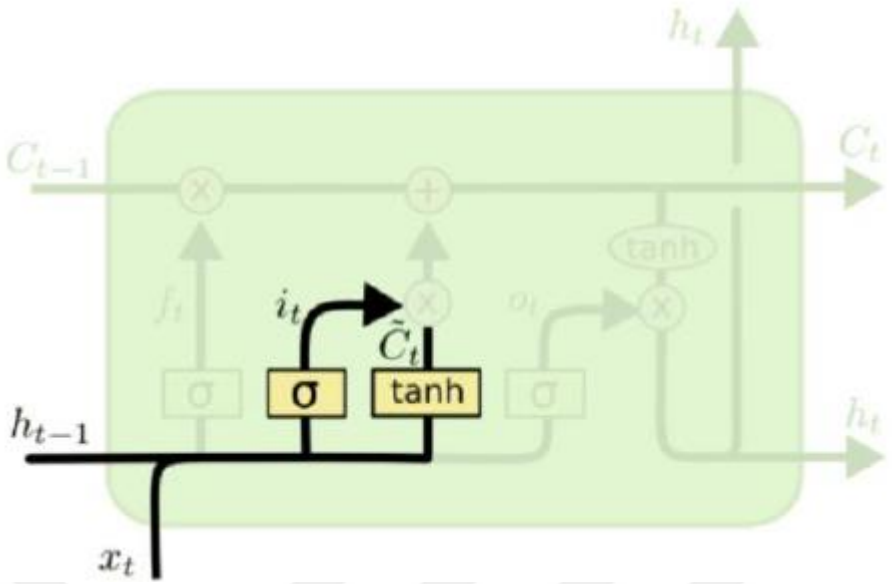
Şekil 2.3. LSTM Hücrelerinde Doğrusal İşlemler (Olah, 2015)

Şekil 2.4.'te gösterildiği gibi, birinci geçide "unutma kapısı" denilir. Önceki çıkış birimlerini h_{t-1} değerini ve mevcut x_t girişini alır ve geçen bilginin oranını belirtmek için 0 ile 1 arasında bir sayı çıkarır. 0, hiçbir bilginin geçmesine izin verme, 1 ise tüm bilgilerin geçmesine izin ver anlamına gelmektedir.



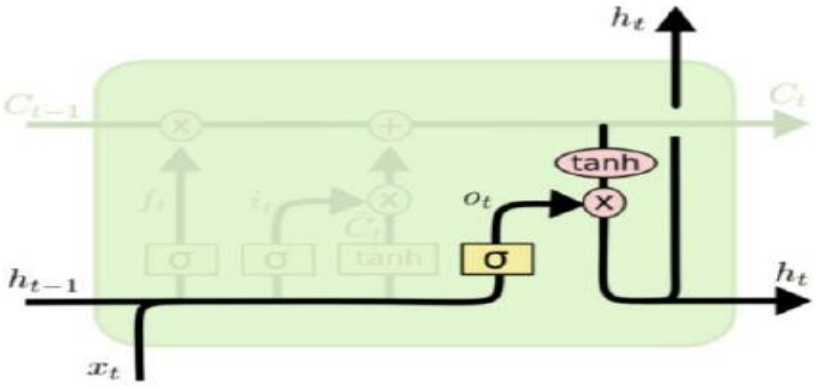
Şekil 2.4. Unutma Katmanı (Olah, 2015)

Hangi bilgilerin güncellenmesi gerektiğine karar vermek için, LSTM "giriş kapısı katmanını" içerir. Ayrıca, h_{t-1} değerinin önceki çıkış birimlerini alır ve mevcut x_t girişi ve bilgilerin hangi hücrelerin içinde güncellenmesi gerektiğini gösteren bir sayı çıkarır. Daha sonra, önceki hücre durumu yani C_{t-1} yeni durum olarak C_t'ye güncellenir.



Şekil 2.5. Giriş Kapısı Katmanı (Olah, 2015)

Son kapı, çıktının ne olması gerektiğine karar veren "çıktı kapısı katmanıdır". Şekil 2.6'da çıktı katmanında, hücre durumunun bir tanh fonksiyonundan geçtiğini ve daha sonra sigmoid fonksiyonunun ağırlıklı çıktısıyla çarpıldığını göstermektedir. Böylece, çıkış birimleri h_t değeri bir sonraki LSTM hücresine geçirilir (Olah, 2015). Basit doğrusal operatörler üç geçit katmanını birbirine bağlar. Geniş LSTM sinir ağı birçok LSTM hücresinden oluşur ve tüm bilgiler tüm hücrelerin içinden geçirilirken, ne kadar hücre olursa olsun, kritik bilgi sonuna kadar saklanır.



Şekil 2.6. Çıktı Kapısı Katmanı (Olah, 2015)

2.3. DERİN ÖĞRENMENİN ZORLUKLARI

Veri yoluyla öğrenme, derin öğrenmenin en önemli avantajlarından biri olduğu gibi, bazı dezavantajların nedeni olabilir. Bunlardan en önemlisi, eğer belirli veriler yoluyla eğitilirse, bu çözümü daha önce görülmemiş, fazla uydurma ve genelleme sorunu olarak bilinen diğer verilere devredemeyebilir. Bahsedilen sorun, veriler oldukça küçük olduğunda ortaya çıkabilir. Bununla birlikte, büyük verilerin kullanılması zaman ve kaynak açısından maliyetli olabilir. Ayrıca, gürültülü veriler bulunduğunda, temiz verilerle uğraşırken yanlış çözüm elde edilme ihtimali bulunmaktadır. Buda önyargı sorununa yol açabilir.

Derin öğrenmenin diğer dezavantajlarından bir diğeri ise, özellikle ağın çok derin olması halinde gizli temsilin belirsiz olabileceğidir. Derin bir öğrenme algoritmasının eğitimi asla kolay değildir, çünkü genellikle kaybolan gradyan ve yerel minimum gibi sorunlara neden olabilir (Şimşek, 2018).

2.4. UYGULAMALAR

Derin öğrenme yöntemleri, çeşitli alanlardaki en gelişmiş doğruluk sonuçlarına ulaşarak değerlerini kanıtlamışlardır. Bunlar, derin öğrenme yöntemlerinin dikkate değer bir başarı elde ettiği en belirgin alanlara örnektir:

- Nesne algılama ve tanıma
- Konuşma tanıma ve dil çevirisi.
- Kendi kendine sürüş arabaları.
- Doğal dil işleme (NLP).
- Tıbbi araştırma.
- Ve daha birçok şey... (Şimşek, 2018)

Yukarıda bahsedildiği gibi derin öğrenme, insan beyninde bulunan sinir ağı mimarilerini çoğaltmanın temel konsepti üzerinde çalışan, makine öğrenmesi algoritmaları ailesinin bir parçasıdır. Farklı sinir ağları sınıfları arasında, Tekrarlayan Sinir Ağları, metin özetleme görevleri için bir araç olmuştur (Bulusu, 2018). Tekrarlayan Sinir Ağlarının temel kavramları burada sunulmaktadır.

2.5. ÖLÇÜM TEKNİKLERİ

Geleneksel olarak özetlemenin değerlendirilmesi, örneğin tutarlılık, özlülük, dilbilgisellik, okunabilirlik ve içerik gibi farklı kalite ölçümlerine ilişkin insan yargılarını içerir (Mani, 2001). Bununla birlikte, Belge Anlama Konferansı'nda (DUC) (Over, 2003) olduğu gibi, birkaç dilsel kalite sorusu ve içerik kapsamı üzerinden özetlerin büyük ölçekte basit bir manuel değerlendirmesi bile, 3.000 saatten fazla insan çabası gerektirecektir. Bu çok pahalıdır ve sık sık yapılması zordur. Bu nedenle, özetlerin otomatik olarak nasıl değerlendirileceği, son yıllarda özetleme araştırmaları camiasında oldukça fazla ilgi görmüştür. Örneğin Saggion ve arkadaşları (2002b), özetler arasındaki benzerliği ölçen üç içerik tabanlı değerlendirme yöntemi önermiştir. Bu yöntemler aşağıda yazılmıştır: kosinüs benzerliği, birim örtüşmesi (yani, unigram veya bigram) ve en uzun ortak alt dizi. Ancak, bu otomatik değerlendirme yöntemlerinin sonuçlarının insan yargılarıyla nasıl ilişkili olduğunu göstermemişlerdir.

BLEU (Papineni ve diğerleri, 2001) gibi otomatik değerlendirme yöntemlerinin makine çevirisi değerlendirmesinde başarılı bir şekilde uygulanmasının ardından, Lin ve Hovy (2003), özetleri değerlendirmek için BLEU'ya benzer yöntemlerin, yani n-gram birlikte oluşum istatistiklerinin uygulanabileceğini göstermiştir. ROUGE, Gisting Assessment için Recall-Oriented Understudy'nin kısaltmasıdır ve özetler arasındaki benzerliği ölçen birkaç otomatik değerlendirme yöntemini içerir.

Özet kalitesinin değerlendirilmesi çok iddialı bir iştir. Uygun değerlendirme yöntemleri ve türleri ile ilgili ciddi sorular devam etmektedir. Özetleme sistemlerinin performansının karşılaştırılması için çeşitli olası temeller vardır. Bir sistem özeti kaynak metinle, insan tarafından oluşturulan bir özetle veya başka bir sistem özetiyle karşılaştırılabilir. Özetleme değerlendirme yöntemleri genel olarak iki kategoride sınıflandırılabilir (Jones ve Galliers, 1995). Dış değerlendirmede özet kalitesi, verilen bir görev için özetlerin ne kadar yararlı olduğuna göre değerlendirilir ve içsel değerlendirmede doğrudan özetin analizine dayanır. Sonuncusu, kaynak belgeyle bir karşılaştırmayı içerebilir, kaynak belgenin kaç ana fikrinin özet tarafından kapsandığını ölçer veya bir insan tarafından yazılmış bir özet ile içerik karşılaştırması içerebilir. Sistem özetini "ideal özet" ile eşleştirme sorunu ideal özetin oluşturulmasının zor olmasıdır. İnsan özeti, makalenin yazarı tarafından, bir özet oluşturması istenen bir yargıç

tarafından, veya cümleleri çıkarması istenen bir yargıç tarafından sağlanabilir. Belirli bir belgeyi özetleyebilecek çok sayıda özet olabilir. İçsel değerlendirmeler daha sonra genel olarak içerik değerlendirmesi ve metin kalitesi değerlendirmesine ayrılabilir. İçerik değerlendirmeleri temel konuları belirleme yeteneğini ölçerken, metin kalitesi değerlendirmeleri otomatik özetlerin okunabilirliğini, dilbilgisini ve tutarlılığını değerlendirir (Steinberger ve Jezek, 2009).

2.6. DEĞERLENDİRME ÖLÇUMLERİ

Metin kalitesi genellikle insan açıklayıcılar tarafından değerlendirilir. Her özete önceden tanımlanmış bir ölçekten bir değer atarlar. Özet kalitesinin belirlenmesine yönelik ana yaklaşım, genellikle ideal bir özet ile karşılaştırılarak yapılan içsel içerik değerlendirmesidir. Cümle alıntıları genellikle, birlikte seçilime (co-selection) ile ölçülür. Otomatik özetin kaç tane ideal cümle içerdiğini bulur. İçeriğe dayalı ölçümler, tüm cümle yerine bir cümledeki gerçek kelimeleri karşılaştırır. Avantajları, hem insan hem de otomatik alıntıları yeni yazılmış cümleler içeren insan özetleriyle karşılaştırabilmeleridir. Bir diğer önemli grup görev tabanlı yöntemlerdir. Belirli bir görev için özetlerin kullanma performansını ölçerler.

2.6.1. Metin Kalite Ölçümleri

Metin (dilsel) kalitesinin birkaç yönü vardır:

1. Dilbilgisi - metin, metinsel olmayan öğeler (yani işaretçiler) veya noktalama hataları veya yanlış kelimeler içermemelidir.
2. Artıksızlık (non-redundancy) – metin gereksiz bilgiler içermemelidir.
3. Referans netliği – isimler ve zamirler özette açıkça belirtilmelidir. Örneğin, özet bağlamında birisini kastetmesi gereken zamir.
4. Tutarlılık ve yapı (coherence and structure) – özet iyi bir yapıya sahip olmalı ve cümleler tutarlı olmalıdır. Bu otomatik olarak yapılamaz. Anlatıcılar çoğunlukla her bir özete puan verir (yani, DUC 2005'te A – çok iyi – E – çok zayıf – arasında) (Steinberger ve Jezek, 2009).

2.6.2. Birlikte Seçilim Ölçümleri

Birlikte seçimin ana değerlendirme ölçütleri duyarlık, anma ve F-ölçümüdür.

1. Duyarlık (P), hem sistemde hem de ideal özetlerde oluşan cümlelerin sayısının sistem özetindeki cümlelerin sayısına bölümüdür.
2. Anma (R) hem sistemde hem de ideal özetlerde yer alan cümlelerin sayısının ideal özetteki cümlelerin sayısına bölümüdür.
3. F-ölçümü, duyarlık ve anmayı birleştiren birleşik bir ölçüdür. F-ölçümü hesaplamanın temel yolu, harmonik bir duyarlık ve anmanın ortalamasını saymaktır (Mani vd., 1999).

$$F = \frac{2*P*R}{P+R} \tag{1}$$

Aşağıda F puanını ölçmek için daha karmaşık bir formül bulunmaktadır:

$$F = \frac{(\beta^2+1)*P*R}{\beta^2*P+R} \tag{2}$$

β, $\beta > 1$ olduğunda duyarlık ve $\beta < 1$ olduğunda anmayı destekleyen bir ağırlıklandırma faktörüdür.

2.6.3. Görece Fayda(Relative Utility)

İnsan yargıçları genellikle bir belgedeki en önemli cümlenin (yüksek p %) hangisi olduğu konusunda farklı düşünceye sahiptiler. Buda P & R ile ilgili temel sorundur. P & R'yi kullanmak, eşit derecede iyi iki özün çok farklı şekilde değerlendirilmesi olasılığına yol açmaktadır (Steinberger ve Jezek, 2009). Duyarlık ve anma sorununun üstesinden gelmek için göreceli fayda ölçüsü tanıtıldı (Radev, Jing ve Budzikowska, 2000). Göreceli fayda ile model özeti, özete dâhil edilmeleri için güven değerleri ile girdi belgesinin tüm cümlelerini temsil eder. Göreceli faydayı hesaplamak için, bir dizi yargıçtan ($N \geq 1$) bir belgedeki tüm n cümleye fayda puanları atamaları istenir.

$$RU = \frac{\sum_{j=1}^{n} \delta_j \sum_{i=1}^{N} \mu_{ij}}{\sum_{j=1}^{n} \epsilon_j \sum_{i=1}^{N} \mu_{ij}} \tag{3}$$

Fayda puanına göre en üstteki e cümleleri daha sonra e boyutunda bir cümle özü olarak adlandırılır. Daha sonra aşağıdaki sistem performansı tanımlanır.

Burada u_{ij}, i açıklayıcısından gelen j cümlesinin fayda puanıdır, ϱ_j, tüm yargıçlardan alınan fayda puanlarının toplamına göre en üstteki e cümleleri için 1'dir, aksi takdirde değeri 0'dır ve δ_j, sistem tarafından çıkarılan en üstteki e cümleler için 1'e eşittir, aksi takdirde değeri 0'dır (Radev, Jing ve Budzikowska, 2000).

2.6.4. İçerik Tabanlı Ölçümler

Birlikte-seçilim ölçümlerinde, tam olarak aynı cümleler bir eşleşme olarak sayılabilir. Buda, farklı yazılmış olsalar bile iki cümlenin aynı bilgiyi içerebileceği gerçeğini yok sayar. Ayrıca, iki farklı yorumcu tarafından yazılan özetler genel olarak aynı cümleleri paylaşmaz. Birlikte seçilim ölçümlerinin bu sorununun üstesinden gelmek için içerik tabanlı benzerlik ölçümleri kullanılır (Steinberger ve Jezek, 2009). Temel bir içerik tabanlı benzerlik ölçüsü Kosinüs Benzerliğidir (Salton ve Buckley, 1988):

$$\cos(X,Y) = \frac{\sum_i x_i * y_i}{\sqrt{\sum_i (x_i)^2} * \sqrt{\sum_i (y_i)^2}} \quad (4)$$

Burada X ve Y, vektör uzay modeline dayalı bir sistem özetinin ve referans belgesinin temsilleridir (Salton ve Buckley, 1988).

2.6.5. Birim Örtüşmesi

Bir başka benzerlik ölçüsü Birim Örtüşmesidir. Burada X ve Y, kelime veya lemma kümelerine dayalı temsillerdir (Saggion vd., 2002a).

$$\text{overlap}(X,Y) = \frac{\|X \cap Y\|}{\|X\| + \|Y\| - \|X \cap Y\|} \quad (5)$$

2.6.6. En Uzun Ortak Alt Dizi

Üçüncü içerik tabanlı ölçü, En Uzun Ortak Altdizi (LCS) olarak adlandırılır.

$$\text{lcs}(X,Y) = \frac{length(X) + length(Y) - edit_{di}(X,Y}{2} \quad (6)$$

Burada X ve Y kelime veya lemma dizilerine dayalı temsillerdir, lcs(X, Y) X ve Y arasındaki en uzun ortak dizinin uzunluğudur, uzunluk(X), X dizisinin uzunluğudur ve $edit_{di}$(X, Y), X ve Y'nin düzenleme mesafesidir (Radev vd., 2003). İki belgedeki metin

karşılaştırıldığında, bu iki belgedeki metnin cümleleri arasında normalleştirilmiş bir çift elde edilir. Birim çakışması ve kosinüs benzerliği aksine, en uzun ortak alt dizi, metinde bulunan bilginin hangi biçimde sıralandığına hassastır (Saggion vd., 2002a).

2.6.7. ROUGE-N: N-Gram Eş Oluşum İstatistikleri

DUC konferanslarının son baskısında, otomatik bir değerlendirme yöntemi olarak ROUGE (Recall-Oriented Understudy for Gisting Assessment) kullanıldı. N-gram4 benzerliğine dayanan ROUGE ölçüm ailesi ilk olarak 2003 yılında tanıtıldı (Lin ve Hovy, 2003).

Bir dizi yorumcunun referans özet seti oluşturduğu varsayılmaktadır. Bir aday özetinin ROUGE-n puanı şu şekilde hesaplanır:

$$\text{ROUGE}_{\text{-n}} = \frac{\sum_{C \epsilon RSS} \sum_{gram_n \epsilon C} Count_{match}(gram_n)}{\sum_{C \epsilon RSS} \sum_{gram_n \epsilon C} Count(gram_n)} \quad (7)$$

Burada Countmatch(gramn), bir sistem ve bir ideal özetinde eş oluşum maksimum n-gram sayısıdır ve Count(gramn) ideal özetindeki toplam n-gram sayısıdır. Ortalama n-gram ROUGE puanı bir anma ölçümü olduğuna dikkat edilmelidir. ROUGE-L ve ROUGESU4 gibi başka ROUGE puanları da vardır (Lin, 2004).

2.6.8. Piramitler

Piramit yöntemi, yeni bir yarı otomatik değerlendirme yöntemidir (Nenkova, Passonneau, 2004). Temel fikri, özetlerdeki bilgilerin karşılaştırılması için kullanılan özetleme içerik birimlerini (SCU'lar) belirlemektir. SCU'lar, bir özetler bütününün ek açıklamalarından ortaya çıkar ve bir yan tümceden daha büyük değildir. Açıklama, benzer cümlelerin tanımlanmasıyla başlar ve daha sonra tanımlamaya yol açabilecek daha dikkatli inceleme ile devam eder. Daha fazla manuel özette görünen SCU'lar daha fazla ağırlık alacaktır, bu nedenle manuel özetlerin SCU ek açıklamasından sonra bir piramit oluşturulacaktır. Piramidin tepesinde, özetlerin çoğunda görünen ve dolayısıyla en büyük ağırlığa sahip olan SCU'lar vardır. SCU piramidin içinde ne kadar düşük görünürse, ağırlığı o kadar düşük olur çünkü daha az özette yer alır. Akran özetindeki SCU'lar daha sonra, akran özeti ve manuel özet arasında ne kadar bilginin uyuştuğunu değerlendirmek

için mevcut bir piramit ile karşılaştırılır. Bununla birlikte, bu umut verici yöntem hala bazı açıklama çalışmaları gerektirmektedir.

2.6.9. Göreve Dayalı ölçümler

Görev tabanlı değerlendirme yöntemleri özetteki cümleleri analiz etmezler. Belirli bir görev için özet kullanma olasılığını ölçmeye çalışırlar. Göreve dayalı özetleme değerlendirmesine yönelik çeşitli yaklaşımlar literatürde bulunabilir. En önemli üç görevden belge sınıflandırma, bilgi erişimi ve soru cevaplamadan bahsedilmektedir (Radev vd., 2003).

ÜÇÜNCÜ BÖLÜM

TÜRK DİLİNDE DERİN ÖĞRENME İLE METİN ÖZETLEME

3.1. ARAŞTIRMANIN AMACI VE ÖNEMİ

Bu çalışmanın amacı, derin öğrenme teknikleri kullanılarak bir Türkçe metin özetleme sistemi geliştirmektir. Bu sayede bir belgenin okumaya ve zaman harcamaya değip değmediğini hızlıca değerlendirebilmek mümkün olacaktır. Bu da analistler, iş dünyası liderleri, akademik araştırmacılar için çok değerlidir. Ayrıca manuel özetlemeye göre daha az zaman ve bütçe gerektirmektedir. Ayrıca bu çalışma sayesinde henüz gelişme aşamasında olan bu konuya derin öğrenme yöntemiyle katkı sağlanacaktır. Derin öğrenme teknikleri kullanılarak Türkçe için ilk çalışmalardan biri olduğu açısından önemlidir.

3.2. ARAŞTIRMANIN METODOLOJİSİ

Metin verilerinde çoğunlukla LSTM yöntemi kullanılmaktadır. Ancak, uzun metinlerde ve Türkçe gibi sondan eklemeli dillerde bu yöntemin az katmanlı olduğundan yalnız başına başarılı olması pek mümkün görünmemektedir. Bu sebeple bu çalışmada mT5 (Xue vd., 2021) mimarisi kullanılmıştır. Bu mimari önceden eğitilmiş bir mimaridir ve 101 farklı dil için çalışmaktadır.

3.3. VERİ SETİ

Bu çalışmada iki farklı veri seti kullanılmıştır. İlk olarak hazır haber veri seti olan MLSUM (Scialom vd., 2020) kullanılmıştır. MLSUM, Fransızca, Almanca, İspanyolca, Türkçe ve Rusça olmak üzere 5 dili kapsamakta ve metin özetleme veri seti olarak bilinmektedir. MLSUM veri seti, popüler CNN/Daily mail veri setinden oluşturulmuştur. İkinci olarak, Türkçe makale veri seti bulunmadığından yazar tarafından dergipark sitesinde yaklaşık 2000 makale indirilerek makale veri seti oluşturulmuştur.

3.3.1. Makale Veri Setinin Hazırlanması

Veri setinde sosyal bilimler, fen bilimler ve sağlık bilimler olmak üzere tüm alanlardan makale bulunmasına dikkat edilmiştir. İndirilen makalelerde aşağıdaki hususlara dikkat edilmiş ve bu hususların bulunduğu makaleler veri setinde silinmiştir.

- Metinde farklı dillerin bulunması
- Makalenin uzun olması (Örneğin 40 sayfadan fazla olan makaleler)
- Türkçe dilinde yazılmaması
- Mühendislik konularında formülün çok olması
- Makalede fotoğrafın metine göre çok olması
- Edebiyat makalelerinde şiir konulu makaleler olması
- Özeti ve anahtar kelimeler olmayan makaleler
- Kopyalanmayan veya kopyalandığında harfleri değişen makaleler

Bu çerçevede indirilen makalelerden sadece 1010 makale veri setine eklenmiştir. Veri setinin oluşturulması için bir CSV dosyasında başlık, özet ve metin üç farklı başlık altında kaydedilmiştir.

3.4. ARAŞTIRMANIN KISITLARI

Çalışmanın kısıtlarına bakacak olursak ilk ve en önemli kısıtlama sistem yetersizliğidir. Metin veri setleri yüksek donanıma ihtiyaç duymaktadır ve bu çalışma Google Colab Pro Plus'ta çalışmasına rağmen çoğu durumda donanım yetersizliği hatası ile karşılanmıştır. Google Colab Pro Plus'ta 83GB RAM ve A100-SXM4-40GB GPU bulunmaktadır. Diğer önemli kısıtlamalardan biri ise Türkçe dili sondan eklemeli bir dil olduğu için doğal dil işleme çalışmaları Türkçede çok zordur. Diğer bir kısıtlama ise makale veri setinin bulunmaması ve bu veri setinin hazırlanmasının zaman alıcı olmasıdır. Çünkü makalelerin birer birer incelenmesi ve değerlendirilmesi gerekmektedir.

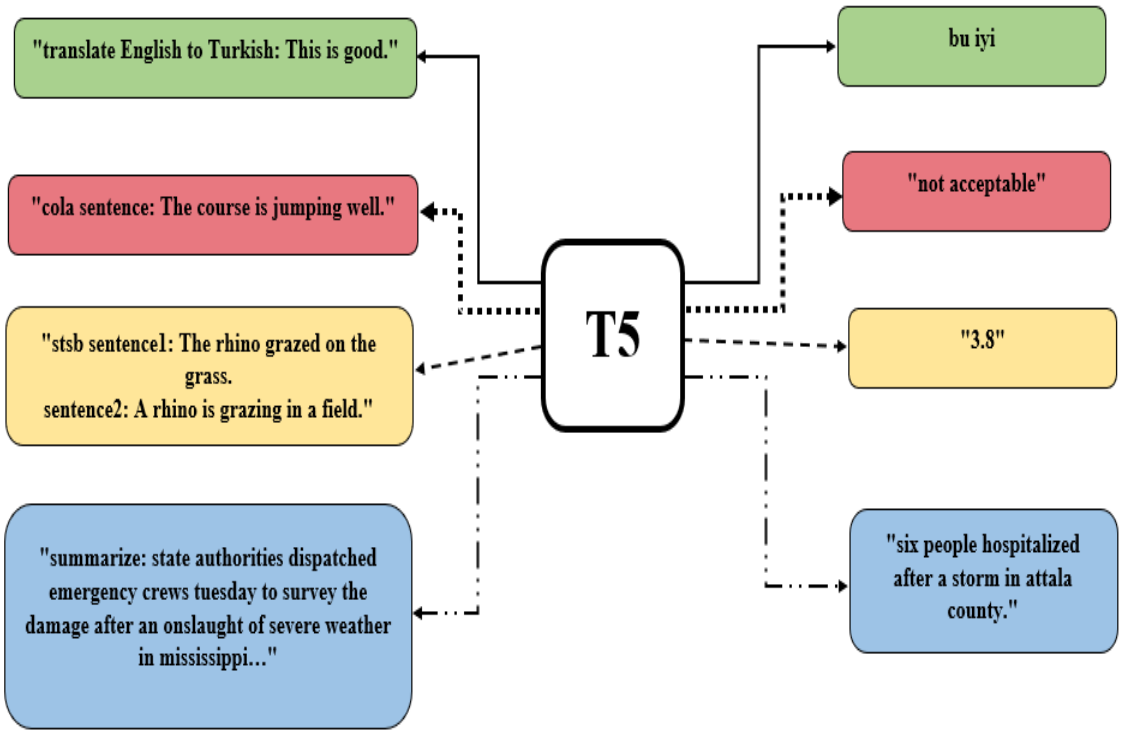
3.5. MODEL

Bu bölüm, T5 modelinin (Raffel vd., 2020) çok dilli varyantı olan mT5 (Multilingual Text-to-Text Transfer Transformer) (Xue vd., 2021) mimarisine genel bir

bakış sağlamaktadır. T5, orijinal olarak önerilen transformatör mimarisini yakından takip eden (Vaswani vd., 2017) ve tüm metin tabanlı NLP problemleri için kullanılabilen (Xue vd., 2021) önceden eğitilmiş metinden metne kodlayıcı-kod çözücü Transformer modelidir ve aşağıdaki hedefleri kapsar:

- Dil modelleme ile bir sonraki kelimeyi tahmin etme
- Karıştırmayı kaldırma (De-shuffling) ile orijinal metni yeniden tanımlama
- Corrupting Spans ile maskelenmiş kelimeleri tahmin etme (Farahani, Gharachorloo ve Manthouri, 2021).

Bu yaklaşım, bazı girdilere dayalı olarak metin oluşturmasına izin verdiği metin özetleme, soru yanıtlama ve metin sınıflandırma gibi üretken görevler için bir NLP çerçevesidir (Xue vd., 2021, Baykara ve Güngör, 2022). Böylece her görev için aynı hiper parametreler ve kayıp fonksiyonu kullanılır (Farahani, Gharachorloo ve Manthouri, 2021). Şekil 3.1 aşağı akış NLP görevleri için birleşik bir çerçeve olarak genel T5'i göstermektedir. Metinden metne formattaki her aşağı akış görevi farklı renklerle gösterilmiştir: çeviri (yeşil), dilsel kabul edilebilirlik (kırmızı), cümle benzerliği (sarı) ve metin özetleme (mavi) (Raffel vd.,2020).

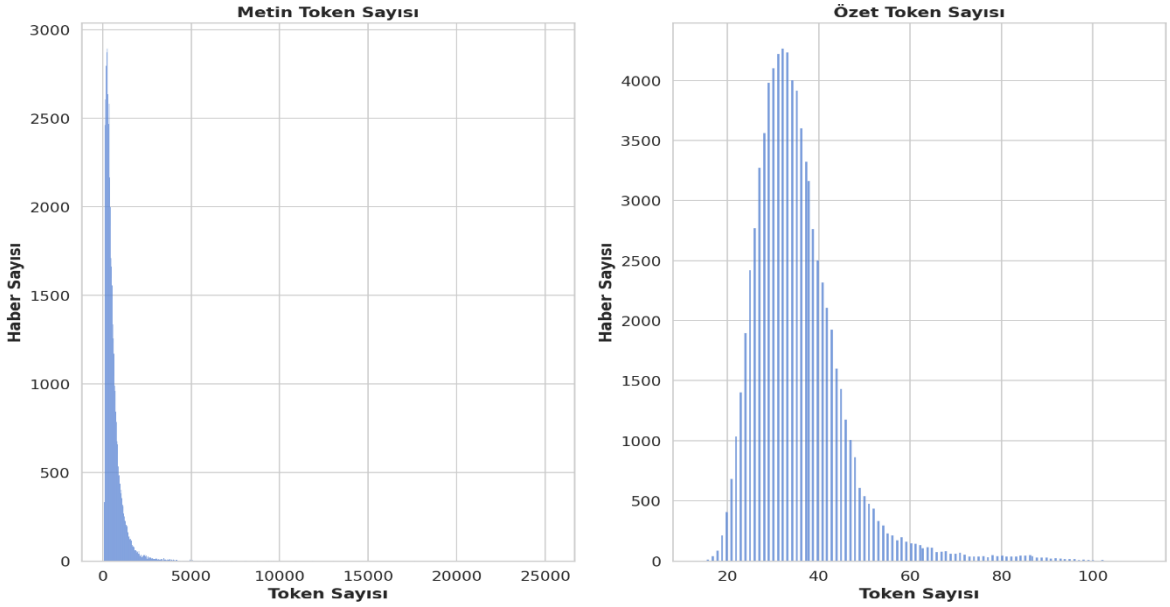


Şekil 3.1. T5 Mimarisi (Raffel vd.,2020)

T5 modeli sadece İngilizce için eğitilmiş olmasına rağmen, mT5 modeli 101 farklı dilde (Türkçe dahil) eğitilmiştir ve T5 modelinin tüm yeterliklerini kullanmaktadır. mT5, BERT, XLM-R ve çok dilli BERT (Farahani, Gharachorloo ve Manthouri, 2021) gibi diğer modellere kıyasla daha güçlü görünmektedir.

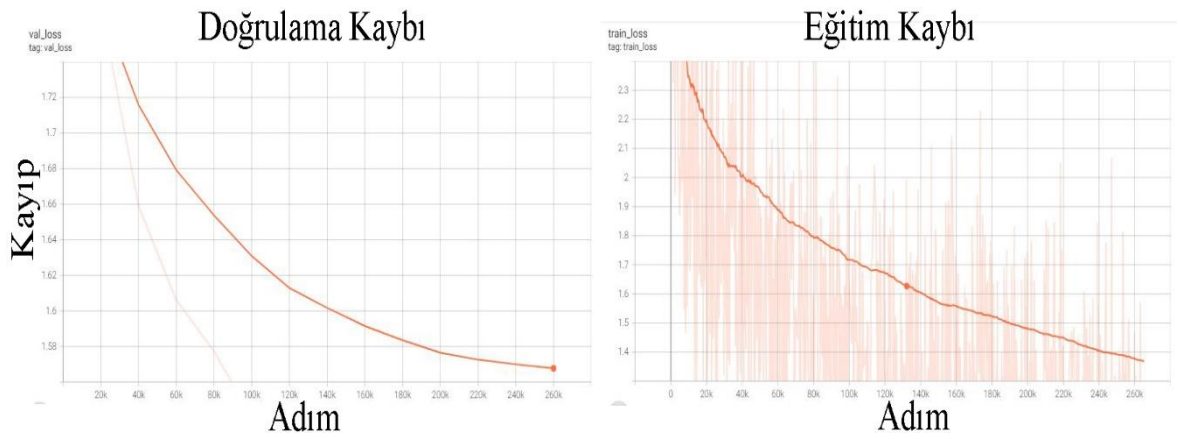
3.5.1. Haber Veri Seti

Bu yazıda Türkçe haberler ve Türkçe makalelerin özetlenmesi için mT5 Fine-tune kullanılmıştır. Hiper parametreleri olarak AdamW optimizer, öğrenme oranı için (learning rate) 0,001, batch-size için 8 ve 16 ve eğitim tekrarı için 15 kullanılmıştır. Ayrıca, uzunluk olarak veri seti incelenmiş ve şekil 3.2'de görüldüğü üzere haber metni uzunluğu için 800 token ve özet uzunluğu için 64 token seçilmiştir. Bunların seçilme sebebi ise metin uzunluğundan dolayı donanım hatasının meydana gelmemesi ve sistemin başarılı bir şekilde çalışmasıdır. Çalışmanın Python kodu Ek 1'de sunulmuştur.



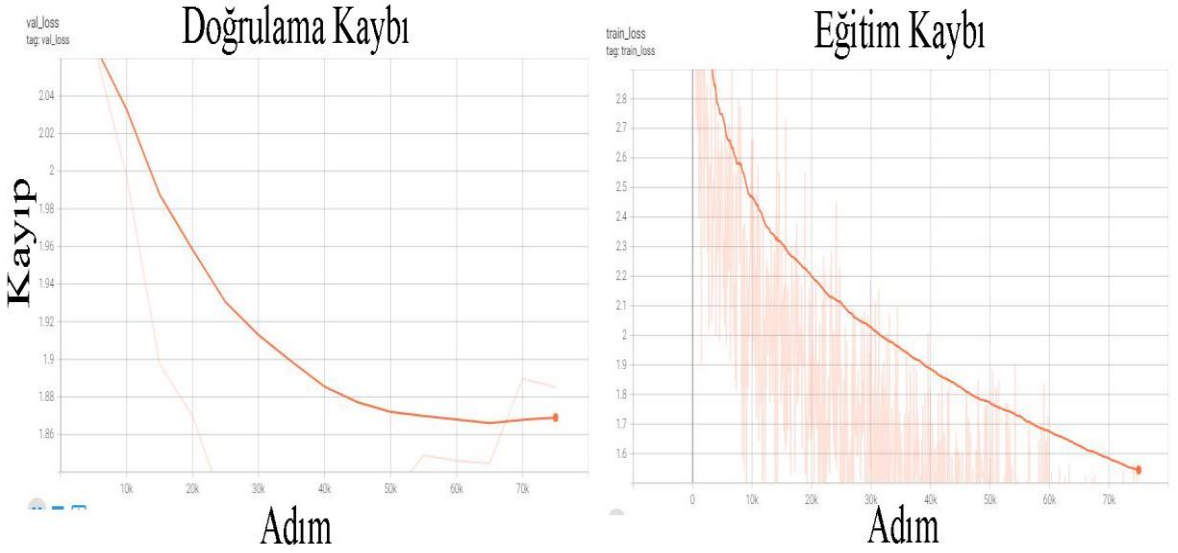
Şekil 3.2. Haber Veri Setinde Metin ve Özet Uzunluğu

Öncelikle 8 ve 16 batch-size ve 15 tekrar ile Türkçe haberler üzerine eğitim gerçekleştirilmiş ve modelin başarısının incelenmesi için, sonuçlar ROUGE metrikleri ile değerlendirilmiştir. ROUGE metrikleri, metin özetleme ve metin çevirisinin başarısını incelemek için kullanılan en yaygın yöntemlerden biridir. Çalışmamızda, ilk olarak girdi için Türkçe haber metinleri ve sonuç için Türkçe haber özeti sisteme verilmiş ve 8 batch-size ile çalıştırılmıştır. Eğitim ve doğrulama kaybı Şekil 3.3'te gösterilmiştir.



Şekil 3.3. Haber Veri Setinde 8 Batch-size için Eğitim ve Doğrulama Kaybı

Daha sonra tekrar girdi için Türkçe haber metinleri ve sonuç için Türkçe haber özeti sisteme verilmiş ancak bu sefer 16 batch-size ile çalıştırılmıştır. Eğitim ve doğrulama kaybı Şekil 3.4'te gösterilmiştir.



Şekil 3.4. Haber Veri Setinde 16 Batch-size için Eğitim ve Doğrulama Kaybı

Çalışmanın başarısını incelemek için ek 3'te sunulan Python kodu ile ROUGE-1, ROUGE-2 ve ROUGE-L değerleri elde edilmiş ve tablo 3.1'de gösterilmiştir.

Tablo 3.1. Haber Veri Setinde ROUGE Değerleri

Batch size	ROUGE-1	ROUGE-2	ROUGE-L
8	31,98	21,11	30,93
16	28,43	17,61	27,40

Tablo 3.1'de gösterildiği üzere 8 batch-size ile iyi değerler elde edilmiştir. 32 batch-size sistem yetersizliğinden dolayı denenmemiştir. Ahuir ve diğerleri. (2021), mT5 ile İspanyolca ve Katalanca özetsel özetleme üzerinde çalışmış ve ROUGE-1, ROUGE-2 ve ROUGE-L olarak aşağıdaki değerleri elde etmişlerdir.

Tablo 3.2. Diğer Çalışmalarda Elde Edilen ROUGE Değerleri

Dil	ROUGE-1	ROUGE-2	ROUGE-L
İspanyolca	30,61	12,36	23,53
Katalanca	27,00	11,28	21,27

Pant ve Chopra (2022), İspanyolca ve Yunanca metinlerde mt5 ile özetsel özetleme üzerinde çalışmış ve ROUGE-2 metriğiyle değerlendirmiştir. İspanyolca için 13,1 ve Yunanca için 13,8 başarı oranı elde etmiştir.

Elde edilen değerler bu iki çalışmanın değerleri ile karşılaştırıldığında sonuçların birinci çalışmanın değerlerinden biraz daha iyi olduğu ikinci çalışmaya göre çok daha iyi olduğu görülmektedir. Bu modelin performansını ve nasıl bir özet ürettiğini göstermek amacıyla veri kümesinden üç örnek için özet üretilmiş ve aşağıdaki tablolarda gösterilmiştir. Tablo 3.3'te 8 batch-size için üretilen özetler gösterilmiştir.

Tablo 3.3. 8 Batch-Size için Üretilen Özetler

Metin 1: Fenerbahçe kulübü, İtalya birinci futbol ligi ekiplerinden Empoli'nin orta saha oyuncusu Miha Zajc'ı kadrosuna kattığını resmen açıkladı. Sarı-lacivertli kulübün internet sitesinde yer alan açıklamada, 24 yaşındaki Zajc ile 4,5 yıllık anlaşmaya varıldığı belirtildi. Zajc transferi için Empoli'de kiralık olarak forma giyen Salih Uçanın haklarından vazgeçildiği de duyuruldu. Açıklamada, "Kulübümüz, İtalya Seri A ekiplerinden Empoli takımında forma giyen merkez orta saha ve ofansif orta saha oyuncusu Miha Zajc bonservisiyle birlikte kadromuza katmak üzere kulübüyle ve futbolcuyla anlaşmaya varmıştır. 24 yaşındaki Sloven oyuncu Miha Zajc, 4,5 sezon boyunca sarı-lacivertli forma ile mücadele edecek. Oyuncumuz Miha Zajc'a Fenerbahçeye hoş geldin diyor, çubuklu ile nice başarıla diliyoruz. Ayrıca, bu transfer kapsamında kulübümüz Salih Uçan üzerindeki haklarından vazgeçerek, Empoli ile Salih Uçan'ın anlaşmasına müsaade etmiştir" ifadeleri yer aldı.
Orijinal Özet 1: Fenerbahçe Sloven orta saha oyuncusu Miha Zajc'ı 4,5 yıllığına transfer ettiğini resme açıkladı. Bu sezon İtalya birinci futbol ligi ekibi Empoli'de 21 resmi maçta görev alan Zajc, 3 gol kaydetti.
mT5 Tarafından Üretilen Özet 1: Fenerbahçe kulübü, Empoli'nin orta saha oyuncusu Miha Zajc'ı kadrosuna kattığını resme açıkladı.
Metin 2: Muğla'nın Bodrum ilçesinde, içerisinde askeri personelin bulunduğu minibüs su kanalına düştü. Kazada minibüste bulunan 3 asker hafif şekilde yaralandı. Kaza, bugün akşam saatlerinde Bodrum-Milas karayolu üzerinde meydana geldi. Askeri personel taşıyan 48 TN 173 plakalı minibüs, Güvercinlik istikametine giderken, sağanak yağış sonrası kayganlaşan yolda sürücü direksiyon hakimiyetini kaybetti. Kontrolden çıkan minibüs, önce yol kenarında bulunan su kanalına düştü, daha sonra da kayalıklara çarparak durabildi. Kazanın ardından olay yerine gelen Muğla 911 Arama Kurtarma ekipleri hafif yaralı askerleri araçtan çıkararak sağlık ekiplerine teslim etti. Yaralan 3 askerden 2'si Bodrum Devlet Hastanesi'ne, 1 asker ise özel bir hastaneye kaldırıldı. Tedaviye alınan 3 askerin de sağlık durumu iyi olduğu öğrenildi.

Tablo 3.3. (Devamı)

Orijinal Özet 2: Muğla'nın Bodrum ilçesinde askeri personel taşıyan askeri minibüs kaza yaptı. Yol kenarında bulunan su kanalına düşen minibüste bulunan 3 asker hafif şekilde yaralandı.
mT5 Tarafından Üretilen Özet 2: Muğla'nın Bodrum ilçesinde, içerisinde askeri personelin bulunduğu minibüs su kanalına düştü. Kazada 3 asker yaralandı.
Metin 3: Dışişleri Bakanı Ahmet Davutoğlu Suriye'deki kimyasal saldırı ile alakalı MİT'in Suriye istihbaratından elde ettiği belgelerle ilgili konuştu. Açıklamasından satırbaşları... İstihbarat birimlerimiz ve uzmanlarımız bilgi edinmek için çaba sarf etti. Milli istihbarat kaynaklarımızdan sağlıklı bilgi edindik. Kimyasal saldırı yapmakta muhalefetin elinde böyle bir imkân olmadığı ortaya çıkıyor. Bu imkâna sadece rejim sahip. Sorumlu Esad rejimidir. Atım vasıtaları ve izlere bakıldığında şüphesiz rejimin sorumluluğundadır. Nihayetinde böyle bir bulgu var. Bu bizim milli istihbari değerlendirmemizdir. Bundan sonra uluslararası topluma sorumluluk düşüyor. Türkiyeyi sanki savaş istiyormuş gibi göstermek isteyenler var doğru değil. Bizim barışın bozulmaması için nasıl çaba sarfettimizi bütün dünya biliyor. Diplomasinin bütün imkânlarını kullandık. Türkiyenin güvenliğini de ilgilendiren kardeşçe tavsiyelerimiz dinlenmedi. Önce küçük çaplı operasyonlar sonra hava saldırıları düzenlendi. Sivil alanlara yapılan en ciddi katliamdı bu. Kimyasal silah kullanımı yıllar önce yasaklanmıştı. Bu bir savaş suçudur. Bir devletin kendi halkına kimyasal silah kullanması tepki verilmemesi gereken bir şeymiş gibi, savaş değilmiş gibi davranıp sanki savaş şimdi başlayacakmış gibi davranmak nedir. Türkiye savaş çağrısı yapıyor yaklaşımı haksız bir ithamdır. Elimizdeki bilgileri uluslararası toğlumla paylaştık.Şimdiye kadar BM Güvenlik Konseyi bir sonuca ulaşamadı kurallar ihlal edildi. Biz yoğun bir diplomasi yaptık. Her ülkenin kendi içinde tartışmalar oluyor bu doğaldır. Önemli olan bu eldeki bilgilerin bir an evvel değerledirilmesidir. Eğer şimdi tepki verilmezse dünyanın diğer ülkeleri de kimyasal silah kullanmada kendilerini de rahat hissedecektir. Esas mesele bütün dünyayı etkileyebilecek kitle imha silahlarının kullanımına izin verilmesidir. Karşılık verilmesi konusunda işaretler görüyoruz. Bu mesele Suriye-Türkiye meselesi değildir. Olabilecek bütün senaryoları ele alan 4 saatlik bir toplantı gerçekleştirdik. Olabilecek riskleri ele aldık alınaak tedbirleri gözden geçirdik.
Orijinal Özet 3: Ahmet Davutoğlu Birleşmiş Milletler'in Suriye'ye yönelik raporuna değerlendirmesini yaptı.
mT5 Tarafından Üretilen Özet 3: Dışişleri Bakanı Ahmet Davutoğlu Suriye'd kimyasal saldırı ile ilgili açıklama yaptı.

Tablo 3.4'te 16 batch-size için üretilen özetler gösterilmiştir. Her iki modelin de anlamlı ve kapsamlı özet ürettiği anlaşılmaktadır. Ancak, ROUGE değerleri incelendiğinde 8 batch-size gerçek ve insan tarafından yazılan özete daha yakın bir özet ürettiği söylenebilir.

Tablo 3.4. 16 Batch-Size için Üretilen Özetler

Metin 1: Fenerbahçe kulübü, İtalya birinci futbol ligi ekiplerinden Empoli'nin orta saha oyuncusu Miha Zajc'ı kadrosuna kattığını resmen açıkladı. Sarı-lacivertli kulübün internet sitesinde yer alan açıklamada, 24 yaşındaki Zajc ile 4,5 yıllık anlaşmaya varıldığı belirtildi. Zajc transferi için Empoli'de kiralık olarak forma giyen Salih Uçanın haklarından vazgeçildiği de duyuruldu. Açıklamada, "Kulübümüz, İtalya Seri A ekiplerinden Empoli takımında forma giyen merkez orta saha ve ofansif orta saha oyuncusu Miha Zajc bonservisiyle birlikte kadromuza katmak üzere kulübüyle ve futbolcuyla anlaşmaya varmıştır. 24 yaşındaki Sloven oyuncu Miha Zajc, 4,5 sezon boyunca sarı-lacivertli forma ile mücadele edecek. Oyuncumuz Miha Zajc'a Fenerbahçeye hoş geldin diyor, çubuklu ile nice başarıla diliyoruz. Ayrıca, bu transfer kapsamında kulübümüz Salih Uçan üzerindeki haklarından vazgeçerek, Empoli ile Salih Uçan'ın anlaşmasına müsaade etmiştir" ifadeleri yer aldı.
Orijinal Özet 1: Fenerbahçe Sloven orta saha oyuncusu Miha Zajc'ı 4,5 yıllığına transfer ettiğini resme açıkladı. Bu sezon İtalya birinci futbol ligi ekibi Empoli'de 21 resmi maçta görev alan Zajc, 3 gol kaydetti.
mT5 Tarafından Üretilen Özet 1: Fenerbahçe kulübü, Empoli'nin orta saha oyuncusu Miha Zajc'ı kadrosuna kattığını açıkladı.

Tablo 3.4. (Devamı)

Metin 2: Muğla'nın Bodrum ilçesinde, içerisinde askeri personelin bulunduğu minibüs su kanalına düştü. Kazada minibüste bulunan 3 asker hafif şekilde yaralandı. Kaza, bugün akşam saatlerinde Bodrum-Milas karayolu üzerinde meydana geldi. Askeri personel taşıyan 48 TN 173 plakalı minibüs, Güvercinlik istikametine giderken, sağanak yağış sonrası kayganlaşan yolda sürücü direksiyon hakimiyetini kaybetti. Kontrolden çıkan minibüs, önce yol kenarında bulunan su kanalına düştü, daha sonra da kayalıklara çarparak durabildi. Kazanın ardından olay yerine gelen Muğla 911 Arama Kurtarma ekipleri hafif yaralı askerleri araçtan çıkararak sağlık ekiplerine teslim etti. Yaralan 3 askerden 2'si Bodrum Devlet Hastanesi'ne, 1 asker ise özel bir hastaneye kaldırıldı. Tedaviye alınan 3 askerin de sağlık durumu iyi olduğu öğrenildi.
Orijinal Özet 2: Muğla'nın Bodrum ilçesinde askeri personel taşıyan askeri minibüs kaza yaptı. Yol kenarında bulunan su kanalına düşen minibüste bulunan 3 asker hafif şekilde yaralandı.
mT5 Tarafından Üretilen Özet 2: Muğla'nın Bodrum ilçesinde, içerisinde askeri personelin bulunduğu minibüs su kanalına düştü…

Tablo 3.4. (Devamı)

Metin 3: Dışişleri Bakanı Ahmet Davutoğlu Suriye'deki kimyasal saldırı ile alakalı MİT'in Suriye istihbaratından elde ettiği belgelerle ilgili konuştu. Açıklamasından satırbaşları... İstihbarat birimlerimiz ve uzmanlarımız bilgi edinmek için çaba sarf etti. Milli istihbarat kaynaklarımızdan sağlıklı bilgi edindik. Kimyasal saldırı yapmakta muhalefetin elinde böyle bir imkân olmadığı ortaya çıkıyor. Bu imkâna sadece rejim sahip. Sorumlu Esad rejimidir. Atım vasıtaları ve izlere bakıldığında şüphesiz rejimin sorumluluğundadır. Nihayetinde böyle bir bulgu var. Bu bizim milli istihbari değerlendirmemizdir. Bundan sonra uluslararası topluma sorumluluk düşüyor. Türkiyeyi sanki savaş istiyormuş gibi göstermek isteyenler var doğru değil. Bizim barışın bozulmaması için nasıl çaba sarfettimizi bütün dünya biliyor. Diplomasinin bütün imkânlarını kullandık. Türkiyenin güvenliğini de ilgilendiren kardeşçe tavsiyelerimiz dinlenmedi. Önce küçük çaplı operasyonlar sonra hava saldırıları düzenlendi. Sivil alanlara yapılan en ciddi katliamdı bu. Kimyasal silah kullanımı yıllar önce yasaklanmıştı. Bu bir savaş suçudur. Bir devletin kendi halkına kimyasal silah kullanması tepki verilmemesi gereken bir şeymiş gibi, savaş değilmiş gibi davranıp sanki savaş şimdi başlayacakmış gibi davranmak nedir. Türkiye savaş çağrısı yapıyor yaklaşımı haksız bir ithamdır. Elimizdeki bilgileri uluslararası toğlumla paylaştık.Şimdiye kadar BM Güvenlik Konseyi bir sonuca ulaşamadı kurallar ihlal edildi. Biz yoğun bir diplomasi yaptık. Her ülkenin kendi içinde tartışmalar oluyor bu doğaldır. Önemli olan bu eldeki bilgilerin bir an evvel değerledirilmesidir. Eğer şimdi tepki verilmezse dünyanın diğer ülkeleri de kimyasal silah kullanmada kendilerini de rahat hissedecektir. Esas mesele bütün dünyayı etkileyebilecek kitle imha silahlarının kullanımına izin verilmesidir. Karşılık verilmesi konusunda işaretler görüyoruz. Bu mesele Suriye-Türkiye meselesi değildir. Olabilecek bütün senaryoları ele alan 4 saatlik bir toplantı gerçekleştirdik. Olabilecek riskleri ele aldık alınaak tedbirleri gözden geçirdik.
Orijinal Özet 3: Ahmet Davutoğlu Birleşmiş Milletler'in Suriye'ye yönelik raporuna ilk değerlendirmesini yaptı.
mT5 Tarafından Üretilen Özet 3: Dışişleri Bakanı Ahmet Davutoğlu, Suriye'deki kimyasal saldırı ile ilgili MİT'in Suriye istihbaratından elde ettiği belgelerle ilgili açıklama yaptı.

Ayrıca İngilizce doğal dil işleme çalışmalarında bulunan ancak Türkçede şuana kadar yapılmamış bir diğer çalışma yazar tarafından işlenerek etkisi metin özetleme üzerinde incelenmiştir. Bu amaç doğrultusunda metinde bulunan tüm çoğul ekler ve çoğul

ekin devamında gelen ekler kelimeden silinmiş ve böylece kelime sayısı azaltılarak etkisi metin özetleme üzerinde incelenmiştir. Bu amaç doğrultusunda tablo 3.5'te gösterilen ekler metinde ve özette bulunan kelimelerden silinmiştir.

Tablo 3.5. Metinden Silinen Türkçe Ekleri

Ekler	Örnek
lar / ler	Okullar, geceler
lar/ler + hal ekleri	Kitapları (belirtme hali) Yollara (yönelme hali) Evlerden (ayrılma hali) Okullarda (bulunma hali)
lar/ler + iyelik ekleri	Kitaplarım, Kalemlerin
lar/ler + ilgi ekleri	Savaşların, Gidenlerin
lar/ler + vasıta eki	Bakanlarla
İkili veya üçlü ekler	Savaşlarından, ülkelerine

Bu eklerin silinmesi için hazırlanan kod Python programlama dilinde yazar tarafından ilk defa oluşturulmuştur. Türkçe dilinin sondan eklemeli olduğu için ve onun yanı sıra ekler silindiğinde bazen kelimenin değişmesi gerektiği için bu işlem zaman alıcı ve zordur. Örneğin fillerde çoğul eki silindiğinde aşağıdaki işlemlerin yapılması gerekmektedir.

Raflardır → raftır, dır eki tır ekine değişmiştir.

Koşmuşlardır → koşmuştur, dır eki tur ekine değişmiştir.

Bu iki örnekte de görüldüğü üzere d'nin t harfine değişmesi ve sonraki harfin önceki harflere bağlı olduğu için çoğulun tekile değiştirilmesi dikkat edilmesi gereken bir konudur. Çoğul eki silme ön işlemesi detaylı bir şekilde makale veri setinde incelediğimizde kelime sayısının değiştiği ve bu nedenle eğitimde daha kolay kullanılabileceği anlaşılmıştır. Ön işleme sonucunda oluşan değişiklikler tablo 3.6'da gösterilmiştir. Ayrıca çoğul ekinin silinmesi için yazılan Python kodunun bir kısmı ek 2'de sunulmuştur.

Tablo 3.6. Eklerin Silinmesiyle Makale Veri Setinde Oluşan Değişiklikler

Yapılan işlem	Özette bulunan toplam kelime sayısı	Metinde bulunan toplam kelime sayısı	Sadece bir kere bulunan kelime sayısı
Herhangi bir işlem olmadan	39208	314375	206900
Tüm harflerin küçük harfe dönüştürmek Durma kelimeleri silmek (acaba, ama, aslında, az, bazı, belki, biri, birkaç, bir şey, biz, bu, çok, çünkü, da, daha, de, ve benzeri)	26218 (%33,13)	153854 (%51,06)	84654 (%59,08)
Çoğul eklerin silinmesi • lar / ler • lar/ler + hal ekleri • lar/ler + iyelik ekleri • lar/ler + ilgi ekleri • lar/ler + vasıta eki • İkili veya üçlü ekler	21113 (%19,47)	124267 (%19,23)	67781 (%19,32)

Çoğul eki silindikten sonra ve sistem bu veri seti ile eğitildikten sonra sonuç, ROUGE metrikleri ile değerlendirilmiş ve Tablo 3.7'de gösterilmiştir.

Tablo 3.7. Çoğul Eklerinin Silinmesiyle Elde Edilen ROUGE Değerleri

Batch size	ROUGE-1	ROUGE-2	ROUGE-L
8	10,55	3,89	10,21

Çoğul ekler silindikten sonra bazen anlamsız özetler üretildi ve ayrıca elde edilen sonuç diğer sonuçlara göre çok düşük kalmıştır.

Çalışmanın devamında hiper parametreleri değiştirerek başarının arttırılması amaçlanmıştır. Bu amaç doğrultusunda öğrenme oranı için 0,00004 seçilmiş ve sistem haber veri seti üzerinde tekrar eğitilmiştir. Sonuç ROUGE metrikleri ile değerlendirilmiş ve Tablo 3.8'de gösterilmiştir.

Tablo 3.8. Öğrenme Oranının Değişmesiyle Elde Edilen ROUGE Değerleri

Batch size	Öğrenme Oranı	ROUGE-1	ROUGE-2	ROUGE-L
8	0,00004	58,76	52,98	58,45

Sistem tarafından üretilen 2 farklı özet Tablo 3.9'da gösterilmiştir.

Tablo 3.9. 0,00004 Öğrenme Oranı ile Üretilen Özetler

Metin 1: Kimyasal silah saldırısında görev alan Suriye Ordu birliklerinin detaylı listesinin yer aldığı bilgilere ulaşıldı. Anadolu Ajansının (AA) ulaştığı Suriye\'deki kimyasal saldırıda görev alan Suriye ordu birliklerinin detaylı listesine göre, saldırı, Şam'ın 35 kilometre kuzeyindeki Kuteyfe'deki 155'inci Füze Tugayı ile Kasyun Dağı'ndaki 4'üncü Zırhlı Tümen'e bağlı birliklerden 15-20 civarında kimyasal başlık taşıyan füze-roketle yapıldı. Kuteyfe'de FROG-7 Luna ve/veya M600 füzeleri, Kasyun'da ise 15-70 km menzilli 220 mm'lik roketler kullanıldı. İlgili istihbarat birimlerince hazırlanan ve Türk Hükümetine de ulaştırıldığı öğrenilen raporda yer alan bilgiler, saldırının, doğrudan rejim güçlerince 21 Ağustos Çarşamba saat 02.45\'te yapıldığını ortaya koyuyor. Doğu ve Batı Gota'daki Zamelka, Duma-Harasta arasındaki mıntıka ve farklı yelerşim birimlerinin hedef alındığı saldırıların, iki ayrı merkezden koordineli ve eşzamanlı yapıldığı belirtilen raporda, şu bilgilere yer veriliyor: 20 CİVARINDA KİMYASAL BAŞLIK TAŞIYAN FÜZE "-Şam'ın 35 kilometre kuzeyindeki Kuteyfe'deki 155'inci Füze Tugayı ile Kasyun Dağı'ndaki 4'üncü Zırhlı Tümen'e bağlı birliklerden 15-20 civarında kimyasal başlık taşıyan füze-roket ile yapıldı. Kuteyfe'den yapılan saldırıda FROG-7 Luna ve/veya M600 füzelerinin kullanıldığı değerlendirilmektedir. -Kasyun'dan ise 15-70 kilometre menzilli 220 mm'lik roketlerin kullanıldığı değerlendirilmektedir. Kuteyfe saldırılan bölgeye 30 kilometre, Kasyun Dağı ise 10 kilometre mesafede bulunuyor. -155'inci Füze Tugayı: 51,52,577,578,579 ve 1097'inci füze taburları ve bunu desteklemekle sorumlu Teknik Destek Taburu'ndan oluşuyor. 52 ve 577. füze tabuları El Kuteyfe'de Tugay ile aynı kışlada.
Orijinal Özet 1: AA kimyasal silah saldırısında görev alan Suriye Ordu birliklerinin detaylı listesinin yer aldığı bilgilere ulaştı.
mT5 Tarafından Üretilen Özet 1: Suriye'deki kimyasal silah saldırısında görev alan Suriye ordu birliklerinin detaylı listesinin yer aldığı bilgilere ulaşıldı.

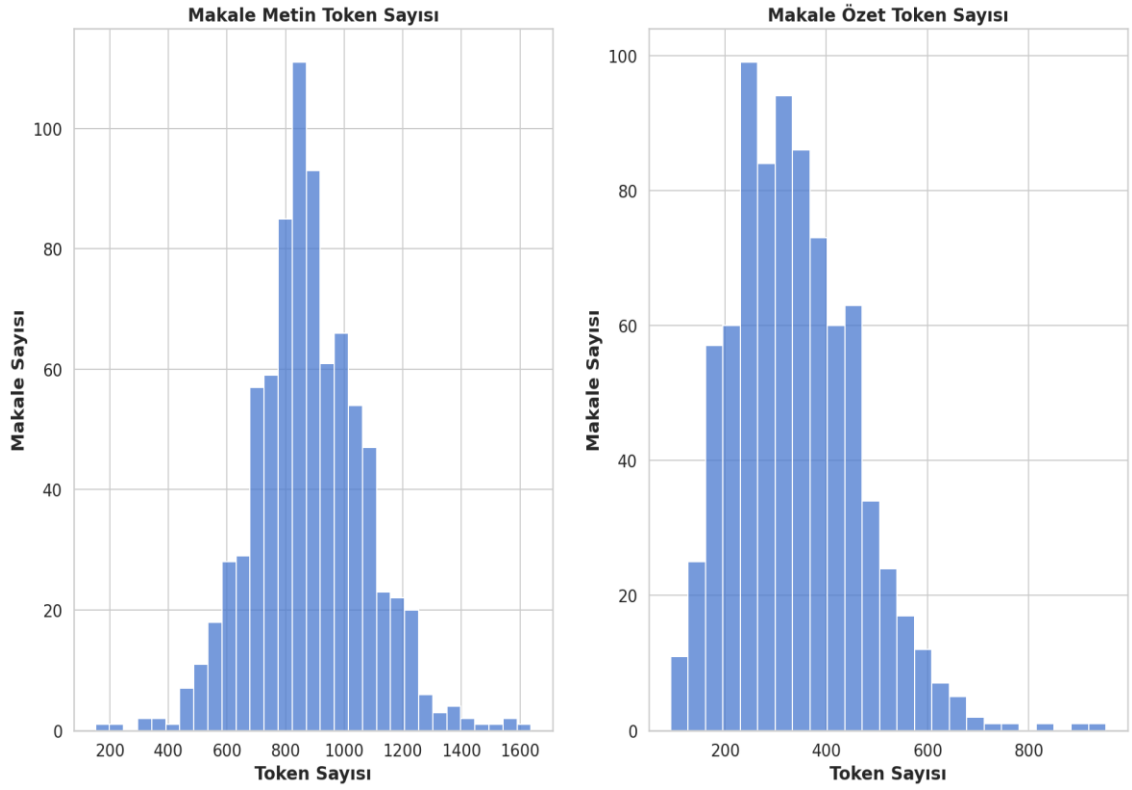
Tablo 3.9. (Devamı)

Metin 2: Madenciler sendikası NUM, yüzde 60'a varan ücret artışı talep ediyor. Bu ha başında işçiler enflasyon oranında %6'lık ücret artışı teklifini reddetmişti. Uzman dünyanın en büyük altın üreticilerinden olan Güney Afrika'da, yetersiz yatırım ve mad şirketlerinin işçilerle kötü ilişkileri yüzünden üretimin olumsuz etkilendiğini belirtiy Güney Afrika Maden Odası, altın madeni şirketlerine, grev konusunda resmi tebliga ulaştırıldığını bildirdi. G. Afrika'daki 120 bin altın madeni işçisinin %64'ü NUM örgütlü. Otomotiv, inşaat ve havacılık sektörlerinde grevle sarsılan ülkede, gelecek ha benzincilerde çalışanlar da greve gidecek. Hükümet, greve çıkan tüm işçilere barı olmaları çağrısı yaptı. Geçen yıl platin madenlerindeki grev sırasında yapılan bir gösteri polisin ateş açması sonucu . Uzmanlar, Cumhurbaşkanı Jacob Zuma'nın hem sağ hem sol partilerin baskısı altında olduğunu vurguluyor. İktidardaki Afrika Ulusal Kongr ANC, Zuma'ya yoksullukla mücadele çağrısı yaparken, işadamları da yabancı yatı çekme, bürokrasiyi azaltma ve ekonomideki yavaşlığı gidermesini talep ediyor.
Orijinal Özet 2: Güney Afrika'da altın madeninde çalışan işçilerin, ücret artışı talebi Salı gününden itibaren greve gideceği bildirildi.
mT5 Tarafından Üretilen Özet 2: Güney Afrika Maden Odası, enflasyon oranı %6'lık ücret artışı talep ediyor.

3.5.2. Makale Veri Seti

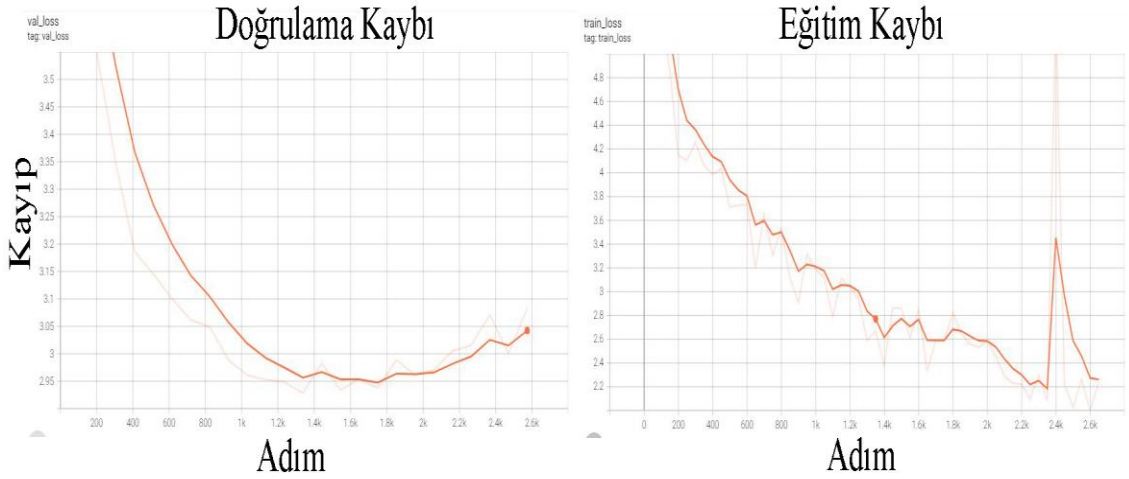
Yazar tarafından oluşturulan makale veri setinin uzun olmasından dolayı ve sistemlerde yetersiz donanım hatası verdiğinden dolayı ilk olarak makalelerde cümle sayısı en az olan makale belirlenmiştir. Bu sayı 28 olarak elde edilmiş ve daha sonra tüm makalelerin cümleleri ek 4'te Python kodu sunulan BERT çıkarımsal özetleme yöntemi ile bu sayıya indirgenmiştir. BERT, transformer mimarisini kullanan ve önceden eğitilmiş çift yönlü bir derin sinir ağıdır. BERT, maskeli dil modeli ve sonraki cümle tahmini hedefleriyle önceden eğitilmiş bir mimaridir. Maskeli dil modeli, basitçe boşlukları doldurma görevidir. BERT rastgele bir cümleden bir kelime seçer ve bu kelimeyi

içerikten tahmin etmeye çalışır. Bu maskeli kelime tahmini işlemi yapılırken, BERT kelimenin geçtiği sağ ve sol içeriği birleştirerek kelimeyi tahmin etmeye çalışır. Sonraki cümle tahmini, verilen iki cümlenin ardılını ve öncülünü bulmaya çalışan bir sınıflandırma algoritmasıdır. BERT-temel modeli, 768 gizli boyutta 12 katman ve 12 öz-dikkat başlığı içerir. BERT-large modeli, 1024 gizli boyutta 24 katman ve 16 öz-dikkat başlığı içerir (Özberk, 2022). Çalışmanın bu bölümünde girdi için yazar tarafından oluşturulan makale veri setinden makale metni ve sonuç için makale özeti sisteme verilmiş ve 8 batch-size ve 15 tekrar ile eğitilmiştir. Sonuçlar ise ROUGE metrikleri ile değerlendirilmiştir. Ancak, makalelerin uzunluğundan dolayı tekrar yetersiz donanım hatası alındığı için şekil 3.5'te görüldüğü üzere makale metni uzunluğu için 1200 token ve özet uzunluğu için 200 token seçilmiştir.



Şekil 3.5. Makale Veri Setinde Metin ve Özet Uzunluğu

Eğitim ve doğrulama kaybı Şekil 3.6'da gösterilmiştir.



Şekil 3.6. Makale Veri Setinde Eğitim ve Doğrulama Kaybı

Çalışmanın başarısını incelemek için ROUGE-1, ROUGE-2 ve ROUGE-L değerleri elde edilmiş ve aşağıdaki tabloda gösterilmiştir.

Tablo 3.10. Makale Veri Setinde ROUGE Değerleri

Veri Seti	ROUGE-1	ROUGE-2	ROUGE-L
Makale	18,34	4,62	17,63

Türkçe makale veri seti üzerinde şuana kadar herhangi bir çalışma yapılmadığından elde edilen değerlerin diğer çalışmalarla karşılaştırılması mümkün olmamıştır ve bu çalışma diğer çalışmalar için bir temel oluşturmuştur. ROUGE değerlerinin küçük olması metnin uzun olmasından kaynaklı olabilir. Üretilen özetler anlamlı olup ancak insan tarafından oluşturulan özetler ile farklılık göstermektedir. Bu modelin performansını ve nasıl bir özet ürettiğini göstermek amacıyla veri kümesinden iki örnek için özet üretilmiş ve Tablo 3.11'de gösterilmiştir.

Tablo 3.11. Makale Veri Seti için Üretilmiş Özetler

Metin 1: ağız ve diş sağlığı genel sağlıktan ayrı düşünülmemesi gereken bireyin yaşam kalitesini ve konforunu direkt olarak etkileyen önemli bir faktördür . buna karşın toplumumuzda hala ağız ve diş sağlığına gereken önem verilmemekte önleyici ve koruyucu uygulamalar gerek eğitim gerekse klinik hizmetleri aşamasında devreye sokulmamakta ve konuya yönelik eğitim ve sağlık politikaları biran önce ele alınmamaktadır . diş çürüklerinden korunmak için sorunun toplum içindeki yaygınlığının belirlenmesi ilk aşamadır . bundan sonra çürüğü etkileyen faktörler araştırılır çözüm yollan bulunur ve uygulamaya geçilir . bu araştırmanın amacı sosyo ekonomik yönden farklı iki bölgeden seçtiğimiz iki ayrı ilkokulda okuyan öğrencilerin diş sağlığı ile ilgili verileri elde ederek karşılaştınnaktır . iki ilkokul taramasında yaşlan arası değişen ted ankara koleji ilkokulu öğrencilerinden u erkek si kız toplam tandoğan ilkokulu öğrencilerinden u erkek si kız toplanı olmak üzere çocuk muayene edilmiştir . ted ilkokulunda incelenen öğrencilerin . ü erkek . si kız tandoğan ilkokulunda ise . ü erkek . sııı kız öğrenciler oluşturmuştur . sağlıklı dişlere sahip olanlar ted ilkokulunda erkekler arasında en yüksek yaşında en düşük yaşında . tandoğan ilkokulu erkek öğrencilerinde en yüksek yaşında . en düşük yaşında . olarak kız öğrencilerde ise ted de en yüksek yaşında . en düşük yaşında . tandoğan da en yüksek yaşında . en düşük yaşında . olarak bulunmuştur . toplama bakıldığında ted öğrencilerinde en sağlıklı dişlere sahip olanların . ile yaş grubunda en sağlıksız olanı ise . ile yaş grubunda tandoğan öğrencilerinde ise en sağlıklı dişlerin . ile yaşta en sağlıksız dişlerin ise . ile yaş grubundaolduğu saptanmıştır . dış çürükleri çekim veya tedavi ihtiyaçlarına göre düşünülerek ayrı ayrı değerlendirilmiştir. tandoğan ilkokulu öğrencilerinin daha fazla sayıda diş çekimine ihtiyacı olduğu ortaya çıkmıştır . tablo de görüldüğü gibi ted öğrencilerinin . sinin tandoğan öğrencilerinin de . sinin tedavi edilmesi gerekli dişi vardır . araştırma grubunda eksik dişi olan öğrencilerin yaş cinsiyet ve okullara göre dağılımı tablo de verilmiştir . bu nedenle de istatistiksel olarak anlamlı değildir . bu sonuç da istatistiksel olarak oldukça anlamlıdır . bir araştırmaya göre diş çekimlerinin primer nedeni oranında çürüktür . çeşitli afrika ülkelerinde yapılan araştırmalar ise kentsel kesimde kırsal kesime göre daha yüksek çürük prevalansını ortaya çıkarmıştır . bu durum

Tablo 3.11. (Devamı)

araştırmacılarca şehirlerde şeker tatlı çerez türü yiyeceklerin daha fazla tüketimine bağlanmıştır . farklı sosyo ekonomik düzeydeki ailelerin çocuklarının devam ettiği iki ayrı ilkokulun öğrencilerinde gerçekleştirdiğimiz araştırmamızda diş sağlığı sorunu olmayan ve tedavi görmüş dişi olan öğrencilerin dağılımının sosyo ekonomik düzeyi iyi olan ilkokulun öğrencilerinde daha iyi olduğu çekimi ve tedavisi gerekli dişi olanların ise sosyoekonomik düzeyi düşük ilkokulda oldukça fazla olduğu görülmüştür . bunun dışında gerek koruyucu gerekse tedavi hizmetlerinden yararlanabilmek belirli bir ekonomik rahatlık gerektirmektedir . ülkemizde die min verilerine göre . milyon ilkokul öğrencisi vardır . toplum ağız diş kurultayı nda belirlenmiştir . dişhekimliğinin bakanlık merkez teşkilatında genel müdürlük gibi yetkili bir birimce temsil edilememesi ağız diş sağlığı alanında gerçekçi bir politikanın saptanmasına ve uygulanmasına engel olmaktadır . ayrıca türkiye de dişhekimliği eğitiminde tedavi ağırlıklı müfredat uygulanmaktadır . gerek dişhekimliği eğitiminde gerek rutin dişhekimliği hizmetlerinde koruyucu uygulamaların yeri çok azdır .
Orijinal Özet 1: çalışmamızın umacı farklı iki sosyo ekonomik düzeydeki ailelerin çocuklarının devam ettiği iki ilkokulun öğrencilerindeki diş sağlığı durumunun değerlendirilmesidir . çalışmamız sosyo ekonoınik düzeyi iyi olarak bilinen ted ankara koleji ilk kısım öğrencileri ve sosyo ekonomik düzeyi yetersiz ailelerin çocuklarının devam ettiği tandoğarı ilkokulu öğrencilerinden toplam öğrenci üzerinde gerçekleştirilmiştir . iki ilkokulun karşılaştırılmasında sosyo ekonomik düzeyi yüksek olan ilkokulun çocuklarında diş sağlığı seviyesinin daha iyi olduğu görülmüştür . sosyo ekonomik düzeyi yüksek olan ilkokulun öğrencilerinde diş sağlığı seviyesinin daha iyi olması bu grupta eğitim öğretim düzeyinin daha yüksek olması sonucu diş sağlığına gereken önemin verilmesine uygun fırçalama ve hakim koşullarının yerine getirilmesine ve ekonomik yönden dişhekimliği hizmetlerine daha kolay ulaşabilirliğe bağlanmıştır .

Tablo 3.11. (Devamı)

mT5 Tarafından Üretilen Özet 1: bu araştırmanın amacı sosyoekonomik düzeyde dişhekimliği eğitiminde tedavi görmüş dişi olan öğrencilerin yaş cinsiyet ve okullara göre değişen faktörler araştırılmıştır. çalışmada farklı sosyoekonomik düzeydeki ailelerin çocuklarının ağız diş sağlığı ile ilgili verileri değerlendirmeye alınmıştır. araştırmanın amacı sosyoekonomik düzeyde ise sağlıklı dişlere sahip olan öğrencilerin ağız diş sağlığı sorunu olmayan ve tedavi görmüş dişi olan öğrencilerin ağız diş sağlığına gerekli uygulamaların belirlenmesidir. gereçsel analizler yapılmıştır. bulgular sonucunda ağız diş sağlığı sorunu olmayan ve tedavi görmüş dişi olan öğrencilerin ağız diş sağlığı sorunu olmayan ve tedavi görmüş dişi olan öğrencilerin ağız diş sağlığı sorunu olmayan ve tedavi görmüş
Metin 2: modern büyüme teorisinin birinci dalgası diye nitelendirilen harrod domar modeli keynes in piyasa mekanizması kendiliğinden tam istihdamı sağlamaz tezinin büyüyen bir ekonomide geçerli olup olmadığını analiz etmeyi amaçlayan bir çalışmadır . solow un yılında yayınladığı makalesinde modern büyüme teorisinin ikinci dalgası diye nitelendirilen model neoklasik iktisadın ekonomik büyüme olgusuna yönelik sonuçlarının değerlendirildiği bir çalışmadır . solow modelde kapitalist bir ekonomide uzun dönemde dengeli bir büyümenin sağlanacağını göstermiştir . içsel büyüme modeli neoklasik büyüme modelinde olduğu gibi ekonominin büyüme oranının dışsal olduğunun aksine etken ekonomik güçlerin içsel olarak belirlendiğini kabul etmektedir . buna göre uzun dönemde yaşam standardının artarak sürmesi teknolojik gelişme ve beşeri sermaye ile sağlanacaktır . beşeri sermaye teorisinin ele alınışını adam smith e kadara dayandıran teorik çalışmalar mevcuttur . uzun dönemde büyümenin insan sermayesi yoluyla sağlandığı büyüme modellerinden ilki robert lucas ın yılında yayımladığı makalesinde önerdiği modeldir . böylece beşeri sermaye büyüme hızının artması büyümeyi de hızlandıracaktır . nitelikli işgücü şeklinde ortaya çıkan nüfus beşeri sermaye stokunu oluşturmaktadır . bireyin sosyal faydasını artırma isteği toplumsal faydayıda artıracaktır . bireylerin eğitim seviyeleri yükseldikçe nitelikli işgücü ihtiyacı karşılanır işgücü verimliliği artar bilimsel ve teknolojik yenilikler hız kazanır . bu sürecin sonunda beşeri sermayeyi teşkil eden nitelikli işgücü iktisadi büyümeyi ve kalkınmayı hızlandırır . türkiyede mesleki eğitim kısa bir tarihçesi türkiye de mesleki ve teknik eğitim yoluyla eğitim kurumları itibariyle nüfusun eğitim kurumlarından yararlanmaları sonucu beşeri sermayeye katkı sağlanmaktadır .

Tablo 3.11. (Devamı)

gelişmekte olan ülkelerin içinde bulunduğu yapısal değişimin gereği olarak ortaya çıkan nitelikli işgücünü ihtiyacı eğitimin rolünü ortaya koymaktadır . bu çerçevede mesleki ve teknik eğitim bir zorunluluk olmaktadır . selçuklu ve osmanlı dönemlerde mükemmel bir şekilde işleyen gelişmesine imkan vermemiştir . cumhuriyet dönemi ile başlayan sanayileşme hamlesinde mesleki teknik öğrenim ile kalkınma arasındaki ilişki gündeme gelmiştir . ilk meslek okulu mithat paşa tarafından yılında kurulmuştur . osmanlı imparatorluğunda meslek okulları mahalli idarelerce kurulup yönetilmiştir . sayılı khk ile oluşturulan hayat boyu öğrenme genel müdürlüğü nün vizyonu farklı öğrenim ve yaş seviyesindeki bireylerin istihdam edilebilirliklerini ve sosyo kültürel gelişimlerini sağlamak amacıyla bilgi beceri ve yeterliklerini geliştirmek öğrenmeye erişimlerini artırmaktır . bu kurumlar temel ve mesleki eğitimi birlikte yürüten ve bireylerin meslek sahibi olarak iş yaşamında yer alabilmelerine imkan sağlaması bakımından önemlidir . izleyen yıllar itibariyle bu eğitim programları gelişme göstermiştir . tablo ye göre genel ortaöğretim itibariyle küçük bir oranda da olsa azalırken mesleki ortaöğretim artış göstermektedir . böylece emeğin niteliğini yükseltmenin eğitimi geliştirerek sağlanabileceği düşünülmüştür . bu gelişme beşeri sermayenin gelişerek ve oransal olarak artarak kalkınmanın sağlanmasına etki edecektir .
Orijinal Özet 2: ikinci dünya savaşı sonrası ülkelerarası gelişmişlik farklarının belirginleşmesiyle azgelişmiş veya geri kalmış ülkelerin ekonomik olarak kalkınması son derece ciddi bir sorun olarak ortaya çıkmıştır . bu süreçte gelişmişlik farlılıkları bakımından benzer olmayan ülkelerin benzer büyüme modellerini uygulamalarının mümkün olmadığı görülmüş ve kalkınma çabalarında yeni arayışlara yönelmiştir . bu dönemde oluşmaya başlayan büyüme teorilerinin özünü savaş sonrası savaştan etkilenen ekonomilerin kalkındırılması oluşturmuştur . bu doğrultuda gelişen büyüme teorileri ülkelerin gelişme çabalarında önemli rol oynamıştır . büyümenin temel belirleyicileri üzerinde yapılan değerlendirmeler ile gelişme yolunda ivme kazanılmıştır . ancak ülkelerin kalkınmalarında temel belirleyicilerden olan eğitim faktörü beşeri sermaye oluşumuna katkı sağlayarak iktisadi büyümede önemli olmaktadır . eğitim beşeri sermaye teorisinin kilit unsurlarından biridir çünkü bilgi ve beceriyi geliştirmenin birincil yolu olarak görülmektedir . buna göre eğitim düzeyi emek kalitesini ölçmenin bir yolu olarak ele alınmaktadır . nitelikli eğitim ise beşeri sermaye oluşumunun temelini oluşturmaktadır .

Tablo 3.11. (Devamı)

mT5 Tarafından Üretilen Özet 2: bu çalışmada türkiye de mesleki ve teknik eğitim kurumları itibariyle nüfus beşeri sermayenin gelişmesiyle birlikte yürütülmüştür. bu kapsamda türkiye de mesleki ve teknik eğitim kurumları itibariyle ekonomik büyüme modellerinin ortaya çıktığı bir dönemdir. türkiye de mesleki ve teknik eğitim kurumları itibariyle insan sermayesi ile sağlanmıştır. bu dönemde ekonominin sürdürülebilirliğinin artmasına yönelik sonuçlar ortaya çıkan nitelikli işgücü ihtiyacı karşılamaktadır. bu durumun sonunda beşeri sermayenin gelişmesine katkı sağladığı düşünülmektedir. bu çalışmada türkiye de mesleki ve teknik eğitim kurumları itibariyle beşeri sermayeye katkı sağlayacaktır.

Tablo 3.12'de tüm sonuçlara genel bir bakış sağlanmıştır.

Tablo 3.12. Tüm Sonuçlara Genel Bir Bakış

Veri Seti	**Haber**	**Haber**	**Haber**	**Çoğul eki silinmiş Haber**	**Makale**
Eğitim Büyüklüğü	80.000	80.000	80.000	80.000	818
Doğrulama Büyüklüğü	20.000	20.000	20.000	20.000	91
Test Büyüklüğü	2.000	2.000	2.000	2.000	101
Tekrar	15	15	15	15	15
Öğrenme Oranı	0,001	0,001	0,00004	0,001	0,001
Kullanılan Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Batch-size	8	16	8	8	8
ROUGE-1	31,98	28,43	58,76	10,55	18,34
ROUGE-2	21,11	17,61	52,98	3,89	4,62
ROUGE-L	30,93	27,40	58,45	10,21	17,63
Eğitim Süresi (Saat)	24	24	24	24	4,5
Test Süresi (Saat)	7	7	7	7	1

SONUÇ VE TARTIŞMA

Metin özetleme doğal dil işleme alanının en önemli konularından biri haline gelmiştir. Kurumlar ve kuruluşların artan verileri ve yöneticilerin hızlı bir şekilde değişen dünyada karar verebilmeleri için metin özetleme önem kazanmaktadır. Ayrıca insan sınırlı veri işleme kapasitesine sahiptir. Bu yüzden insan tarafından özetlenen metinler zaman alıcı ve maliyetli olduğu için otomatik metin özetleme daha popüler hale gelmiştir. Bu nedenle çağımızda otomatik metin özetleme araştırma konularının önemli bir parçası olmuştur.

Doğal dil işleme ve metin verilerinde çok kullanılan LSTM ve RNN hafıza sorunu ile karşı karşıya kaldığı için araştırmacılar bu sorunun üstesinde gelebilmek için son yıllarda derin öğrenmenin bir parçası olan önceden eğitilmiş seq2seq yöntemlerini kullanmaya başlamışlardır. Yapılan çalışmalara bakıldığında önceden eğitilmiş seq2seqlerin daha verimli ve daha başarılı olduğu görülmektedir. Örneğin, metinden metine aktarım dönüştürücüsü (Text-to-Text Transfer Transformer) olan T5 ve BERT algoritmalar gibi kodlayıcı kod çözücü modeller kullanılmaktadır.

Türkçe sondan eklemeli bir dildir ve onun yanı sıra diğer dillere göre çok kelime bulunmaktadır bu sebeple Türkçe için LSTM yetersiz kalmaktadır. Ayrıca otomatik metin özetleme araştırmaları daha çok İngilizce ve diğer yabancı dillerde yapılmıştır. Son zamanlarda ortaya çıkan tek dilli BERT modeli ve çok dilli önceden eğitilmiş seq2seq modellerden yardım alarak Türkçe gibi daha az kaynağı ve çalışması bulunan dillerde metin özetleme eksikliklerini gidermek için son teknoloji modellerinin kullanılmasına yol açılmıştır.

İngilizce gibi yabancı dillerde özetsel özetleme bulunurken Türkçe için özetsel metin özetlemeye odaklanmış çok az çalışma bulunmaktadır. Bu çalışmada, Türkçede özetsel metin özetleme konusu ele alanmış ve makale veri seti ilk kez yazar tarafından oluşturulmuştur.

Bu çalışmada iki farklı veri seti kullanılmıştır. İlk olarak MLSUM haber veri seti kullanılmış ve daha sonra yazar tarafından makale veri seti hazırlanmıştır. Makale veri setinin hazırlanmasında dergipark sitesinde bulunan ve Türkçe yazılan makaleler indirilmiş ve incelenerek veri setine eklenmiştir. Makaleler veri setine eklenirken 40 sayfadan uzun olmaması, yabancı kelimenin bulunmaması, kopyalama imkânının

bulunması, özetinin bulunması, Türkçe dilinde yazılması, mühendislik konularında formülün çok olmaması, makalede fotoğrafın metine göre çok olmaması ve edebiyat makalelerinde şiir konulu makaleler olmamasına dikkat edilmiştir. Böylece indirilen yaklaşık 2000 makaleden 1010 makale veri setine eklenmiştir.

Çalışmada Türkçe dâhil 101 dili destekleyen önceden eğitilmiş mT5 kullanılmıştır. İlk olarak MLSUM haber veri seti üzerinde 8 ve 16 batch-size ile eğitim gerçekleştirilmiştir. ROUGE-1, ROUGE-2 ve ROUGE-L için 8 batch-size'de sırasıyla 31,98, 21,11 ve 30,93 ve 16 batch-size'de sırasıyla 28,43, 17,61 ve 27,40 başarı elde edilmiştir. Ahuir ve diğerleri İspanyolca için sırasıyla 30,61, 12,36 ve 23,53 başarı ve Katalanca için 27,00, 11,28 ve 21,27 başarı elde etmişlerdir. Pant ve Chopra (2022), İspanyolca ve Yunanca metinlerde mt5 ile özetsel özetleme üzerinde çalışmış ve ROUGE-2 metriğiyle değerlendirmiştir. İspanyolca için 13.1 ve Yunanca için 13.8 başarı oranı elde etmiştir. Elde edilen değerler bu iki çalışmanın değerleri ile karşılaştırıldığında sonuçların birinci çalışmanın değerlerinden daha iyi ve ikinci çalışmaya göre çok daha iyi olduğu görülmektedir. Daha sonra haber veri seti üzerinde Öğrenme oranı 0,001'den 0,00004'te değiştirilmiş ve eğitim tekrar gerçekleştirilmiştir. Burada ROUGE-1, ROUGE-2 ve ROUGE-L için sırasıyla 58,76, 52,98 ve 58,46 başarı elde edilmiştir. Farahani ve diğerleri Farsça için yaptıkları çalışmada sırasıyla 42,25, 24,36 ve 35,94 başarı elde etmişlerdir. Ayrıca, Baykara ve Güngör Türkçe için yaptıkları çalışmada elde ettikleri en üst değerler sırasıyla 42,26, 27,81 ve 37,96 olmuştur. Foroutan ve diğerleri (2022) İngilizce, İspanyolca ve Yunanca dili için yaptıkları çalışmalarında mT5 ile elde ettikleri değerler sırayla 0,479, 0,267 ve 0,238 olmuştur. Bu çalışmalarla karşılaştırdığımızda yapılan çalışmanın iyi bir sonuç elde ettiği söylenebilir. Chouikhi, ve Alsuhaibani (2022) Arapça dili için yaptıkları çalışmada mT5 ile elde etikleri değerler sırayla ROUGE-1, ROUGE-2 ve ROUGE-L ANA veri seti için 83,34, 58,58 ve 76,85 AHS veri seti için 48,74, 31,22 ve 46,45 ve WikiHow veri seti için 53,16, 25,54 ve 50.66 olmuştur. Hasan ve diğerleri (2021) 44 dil üzerinde özetsel metin özetleme işlemini gerçekleştirmişlerdir. Sonuçları ROUGE metrikleriyle incelemiş ve İngilizce için en iyi sonuçlar elde edilmiştir. ROUGE-1, ROUGE-2 ve ROUGE-L değerleri için elde edilen değerler sırasıyla 36,99, 15,18 ve 29,64 olmuştur. Bu çalışmada diğer 43 dil için elde edilen sonuçlar bu değerlerin altında kalmıştır. Ayrıca haber veri setinde çoğul ekinin silinmesi ile elde edilen ROUGE değerleri sırasıyla 10,55, 3,89 ve 10,21 olmuştur. Çoğul

eki silme işlemi Türkçede ilk kez yapılmış ve sonuçları incelenmiştir. Bu sebeple herhangi bir çalışma ile karşılaştırılması mümkün değildir. Ancak gelecekteki çalışmalar için bir temel oluşturabilir. Makale veri seti ilk kez yazar tarafından oluşturulmuş ve şuana kadar makale veri seti üzerinde herhangi bir çalışma yapılmamıştır. Ancak makale veri setinde elde edilen ROUGE değerleri sırasıyla 18,34, 4,62 ve 17,63 olmuştur. Elde edilen değerler gelecekteki çalışmalar için bir temel oluşturmuş ve makale veri setinin büyütülmesi ile başarının arttırılması incelenebilir.

Çalışmanın kısıtlarına bakacak olursak Türkçede makale veri setinin bulunmaması, ayrıca çoğul ekinin silinmesi konusunda şu ana kadar herhangi bir çalışma yapılmaması ve bu iki çalışmanın yazar tarafından yapılması zaman alıcı olmuştur. Gelecekteki çalışmalarda makale veri seti büyütülerek başarının değişip değişmediği incelenebilir. Ayrıca, sistem donanımı zenginleştirilerek hiper parametreler değiştirilip ve veri seti iki katına çıkartılarak sistemin başarısı değerlendirilebilir. Bu çalışma İngilizce, Farsça ve Türkçe dahil olmak üzere farklı dillerde yapılarak başarı oranları incelenebilir. Öte yandan, metin özetlemenin yanı sıra soru çıkarma konusunun işlenmesi önemli bir araştırma olabilir.

KAYNAKÇA

Abdı Omar, N. (2018). *A User Based Comparative Study of Automatic Text Summarization.* (Master Dissertation), Kocaeli: Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü.

Afatsun, M. N. (2020). *Derin Öğrenme Yöntemleri İle Türkçe Metinlerden Anlamlı Özet Çıkarma*. (Yayımlanmamış Yüksek Lisans Tezi), Ankara: Ankara Üniversitesi, Fen Bilimleri Enstitüsü.

Ahuir, V., Hurtado, L. F., González, J. Á., & Segarra, E. (2021). "NASca and NASes: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish" *Applied Sciences*, *11*(21), 9872.

Akhil, R. (2022). Extractive Text Summarization by Deep Learning (No. 8143). *EasyChair*.

Alami, N., Meknassi, M., Ouatik, S. A., & Ennahnahi, N. (2015, November). "Arabic Text Summarization Based on Graph Theory". In *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)* (pp. 1-8). IEEE.

Altan, Z. (2004). "A Turkish Automatic Text Summarization System". In *IASTED International Conference on AIA* (pp. 16-18).

Amasyalı, M. F., & Diri, B. (2006). "Automatic Turkish Text Categorization in Terms of Author, Genre and Gender". In *International Conference on Application of Natural Language to Information Systems* (pp. 221-226). Springer, Berlin, Heidelberg.

Aysu, E. (2018). *Big Data Storage and Automated Text Summarization in Turkish Text.* (Yayımlanmamış Yüksek Lisans Tezi), İstanbul: Işık Üniversitesi.

Barzilay, R., & Elhadad, M. (1997). "Using Lexical Chains For Text Summarization". *A Workshop on Intelligent Scalable Text Summarization (ACL '97/EACL '97)*, 10–17.

Barzilay, R., & McKeown, K. R. (2005). "Sentence Fusion for Multi Document News Summarization". *Computational Linguistics*, *31*(3), 297-328.

Baxendale, P. B. (1958). "Machine-made Index for Technical Literature—an Experiment". *IBM Journal of Research and Development*, *2*(4), 354-361.

Baydar, E. (2018). *Genetik Algoritma Kullanarak Cümle Seçme Yaklaşımı ile Otomatik Metin Özetleme.* (Yayımlanmamış Yüksek Lisans Tezi), Van: Yüzüncü Yıl Üniversitesi, Fen Bilimleri Enstitüsü.

Baykara, B., & Güngör, T. (2022). "Turkish Abstractive Text Summarization Using Pretrained Sequence-to-sequence Models". *Natural Language Engineering*, 1-30.

Bengio, Y. (2009). "Learning Deep Architectures for AI". *Foundations and trends® in Machine Learning*, *2*(1), 1-127.

Bilgin, O., Çetinoğlu, Ö., & Oflazer, K. (2004). "Building a Wordnet for Turkish". *Romanian Journal of Information Science and Technology*, *7*(1-2), 163-172.

Birant, Ç. C. (2015), *Rule-based Text Summarization in Turkish* (Master dissertation), İzmir: Dokuz Eylül University.

Birant, C. C., & Aktas, O. (2018). "Rule-Based Turkish Text Summarizer (RB-TTS)". *Advances in Electrical and Computer Engineering*, *18*(3), 113-119.

Bohra, M., Dadure, P., & Pakray, P. (2022). "Comparative Analysis of T5 Model for Abstractive Text Summarization on Different Datasets". Available at SSRN: https://ssrn.com/abstract=4096413 or http://dx.doi.org/10.2139/ssrn.4096413

Brandow, R., Mitze, K., & Rau, L. F. (1995). "Automatic Condensation of Electronic Publications by Sentence Selection". *Information Processing & Management*, *31*(5), 675-685.

Bulusu, S. A. (2018). *Experiments in Text Summarization Using Deep Learning* (Doctoral dissertation), The University of North Carolina at Charlotte.

Carbonell, J., & Goldstein, J. (1998). "The use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries". In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (335-336).

Cengiz, H. A. R. K., Uçkan, T., Seyyarer, E., & Karci, A. (2019). "Metin Özetlemesi Için Düğüm Merkezliklerine Dayalı Denetimsiz Bir Yaklaşım". *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, *8*(3), 1109-1118.

Chopra, S., Auli, M., & Rush, A. M. (2016). "Abstractive Sentence Summarization With Attentive Recurrent Neural Networks". In *Proceedings of the 2016 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (93-98).

Chouikhi, H., & Alsuhaibani, M. (2022). "Deep Transformer Language Models for Arabic Text Summarization: A Comparison Study". *Applied Sciences*, 12(23), 11944.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". *arXiv preprint arXiv:1412.3555*.

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

Conroy, J. M., & O'leary, D. P. (2001). "Text Summarization via Hidden Markov Models". In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 406-407).

Darling, W.M. and Song, F. (2011). "Probabilistic Document Modeling for Syntax Removal in Text Summarization". *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, (CL' 11), ACM Press*, Stroudsburg, PA., 642-647.

Dehkordi, P.K., Khosravi, H., & Kumarci, F. (2009). "Text Summarization Based on Genetic Programming". *International Journal of Computing and ICT Research*, 3(1), 57-64.

DeJong, G. (1982). "An Overview of the FRUMP System". *Strategies for Natural Language Processing*, *113*, 149-176.

Demir, S., El-Kahlout, I. D., Unal, E., & Kaya, H. (2012, May). "Turkish Paraphrase Corpus". In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (4087-4091).

Deng, L., & Yu, D. (2014). "Deep Learning: Methods and Applications". *Foundations and trends in signal processing*, 7(3–4), 197-387.

Dey, R., & Salem, F. M. (2017). "Gate-variants of Gated Recurrent Unit (GRU) Neural Networks". *In 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)* (pp. 1597-1600). IEEE.

Dorr, B., Zajic, D., & Schwartz, R. (2003). "Hedge Trimmer: A Parse-and-trim Approach to Headline Generation". *In Proceedings of the HLT-NAACL 03 on Text Summarization Workshop, Association for Computational Linguistics*, Vol: 5, 1–8.

Duraiswamy, K. (2014). "An Approach for Text Summarization Using Deep Learning Algorithm". *Journal of Computer Science*, 10(1), 1-9.

Khatri, C., Singh, G., & Parikh, N. (2018). "Abstractive and Extractive Text Summarization Using Document Context Vector and Recurrent Neural Networks". *arXiv preprint arXiv:1807.08000*.

Edmundson, H. P. (1969). "New Methods in Automatic Extracting". *Journal of the ACM (JACM)*, *16*(2), 264-285.

Ercan, G. (2006). *Automated Text Summarization and Keyphrase Extraction* (Master Dissertation), Ankara: Bilkent University.

Ercan, G., & Cicekli, I. (2007). "Using Lexical Chains for Keyword Extraction". *Information Processing & Management*, *43*(6), 1705-1714.

Erhandı, B. (2020). *Derin Öğrenme Ile Metin Özetleme* (Master Dissertation), Sakarya: Sakarya Üniversitesi.

Erkan, G., & Radev, D. R. (2004). "Lexrank: Graph-based Lexical Centrality as Salience in Text Summarization". *Journal of Artificial Intelligence Research*, 22 (1), 457-479.

Eryiğit, G., Nivre, J., & Oflazer, K. (2008). "Dependency Parsing of Turkish". *Computational Linguistics*, 34(3), 357-389. doi:10.1162/coli.2008.07-017-R1-06-83

Ertam, F., & Aydin, G. (2022). "Abstractive Text Summarization Using Deep Learning With a New Turkish Summarization Benchmark Dataset". *Concurrency and Computation: Practice and Experience*, *34*(9), e6482.

Fattah, M. A., & Ren, F. (2008). "Automatic Text Summarization". *World Academy of Science, Engineering and Technology*, *37*(2), 192-195.

Farahani, M., Gharachorloo, M., & Manthouri, M. (2021). "Leveraging ParsBERT and Pretrained mT5 for Persian Abstractive Text Summarization". In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)* (pp. 1-6). IEEE.

Fendji, J. L. E. K., Taira, D. M., Atemkeng, M., & Ali, A. M. (2021). "WATS-SMS: A T5-Based French Wikipedia Abstractive Text Summarizer for SMS". *Future Internet*, 13(9), 238.

Fernández, S., SanJuan, E., & Torres-Moreno, J. M. (2007). "Textual Energy of Associative Memories: Performant Applications of Enertex Algorithm in Text Summarization and Topic Segmentation". In *Mexican International Conference on Artificial Intelligence* (pp. 861-871). Springer, Berlin, Heidelberg.

Filippova, K. (2010). "Multi-sentence Compression: Finding Shortest Paths in Word Graphs". In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)* (322-330).

Foroutan, N., Romanou, A., Massonnet, S., Lebret, R., & Aberer, K. (2022). "Multilingual Text Summarization on Financial Documents". *In Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022* (pp. 53-58).

Giannakopoulos, G., El-Haj, M., Favre, B., Litvak, M., Steinberger, J., & Varma, V. (2011). "TAC 2011 MultiLing Pilot Overview". *In: Text Analysis Conference (TAC) 2011, MultiLing Summarisation Pilot.* TAC, Maryland, USA.

Goldstein, J., Mittal, V. O., Carbonell, J. G., & Kantrowitz, M. (2000). "Multi-document Summarization by Sentence Extraction". In *NAACL-ANLP 2000 workshop: automatic summarization,* ACM Pres, 40-48.

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,* (19-25).

Graves, A., Mohamed, A. R., & Hinton, G. (2013). "Speech Recognition With Deep Recurrent Neural Networks. Acoustics, Speech and Signal Processing (icassp)". In *2013 ieee international conference on* (pp. 6645-6649). Ieee.

Güran, A. (2013). *Otomatik Metin Özetleme Sistemi*. (Yayımlanmamış Doktora Tezi), İstanbul: Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü.

Güran, A., Bekar, E., & Akyokus, S. (2010). "A Comparison of Feature and Semantic-based Summarization Algorithms for Turkish". In *International Symposium on Innovations in Intelligent Systems and Applications* (21-24).

Güran, A., Bayazit, N. G., & Bekar, E. (2011). "Automatic Summarization of Turkish Documents Using non-negative Matrix Factorization". In *2011 International Symposium on Innovations in Intelligent Systems and Applications* (pp. 480-484). IEEE.

Hafsari, Y. (2020). *RNN with more sophisticated memory cell and architecture (LSTM, GRU and Attention),* https://medium.com/@yyunisari158/rnn-with-more-sophisticated-memory-cell-and-architecture-lstm-gru-and-attention-528fc942d5af (Erişim Tarihi : 20/11/2020).

Hahn, U., & Mani, I. (1998). "Automatic Text summarization: Methods, systems and Evaluation". In *Tutorial, International Joint Conference on Artificial Intelligence (IJCAI 99)*.

Hakkani-Tür, D. Z., Oflazer, K., & Tür, G. (2000). "Statistical Morphological Disambiguation for Agglutinative Languages". In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*. 1-7.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English (English Language Series*) (374).

Hasan, T., Bhattacharjee, A., Islam, M. S., Samin, K., Li, Y. F., Kang, Y. B., ... & Shahriyar, R. (2021). "XL-sum: Large-scale Multilingual Abstractive Summarization for 44 Languages". *arXiv preprint arXiv:2106.13822*.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). "A fast Learning Algorithm for Deep Belief Nets". *Neural Computation*, *18*(7), 1527-1554.

Hochreiter, S., & Schmidhuber, J. (1997). "Long short-term Memory". *Neural Computation*, 9(8), 1735-1780. IEEE.

Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). "Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-term Dependencies". *In: A Field Guide to Dynamical Recurrent Neural Networks*, (pp.237-244). IEEE.

Hovy, E., & Lin, C. Y. (1998). "Automated text Summarization and the SUMMARIST system". In *Proceedings of a Workshop* (pp. 197-214).

Hovy, E., & Lin, C. Y. (1999). "Automated text Summarization in SUMMARIST". *Advances in Automatic Text Summarization*, *14*, 81-94.

Iqbal, T., Sambyal, A. S. & Devanand. (2018). "A Review of text Summarization Using Gated Neural Networks". *Int. J. Comput. Sci. Eng*, *6*(03), 51-55.

Jones, K. S. (1993). "What Might be in a Summary?". *Information Retrieval*, *93*(1), 9-26.

Kasture, N. R., Yargal, N., Singh, N. N., Kulkarni, N., & Mathur, V. (2014). "A Survey on Methods of Abstractive Text Summarization". *International Journal for Research in Emerging Science and Technology*. *Int. J. Res. Merg. Sci. Technol*, *1*(6), 53-57.

Jackson, P., & Moulinier, I. (2002). "Natural Language Processing for Online Applications". *Natural Language Processing for Online Applications*, 1-244. Philadelphia: John Benjamins.

Jones, K. S., & Galliers, J. R. (1995). "Evaluating Natural Language Processing Systems: An Analysis and Review". *In Lecture Notes in Artificial Intelligence*, No. 1083, Springer.

Karaca, A. (2021). *Derin Öğrenme ile Türkçe Haber Metinlerine Başlık Üretme.* (Yayımlanmamış Yüksek Lisans Tezi), Trakya Üniversitesi, Fen Bilimleri Enstitüsü.

Karakaya, K. M., & Güvenir, H. A. (2004). "ARG: A Tool for Automatic Report Generation". *IU-Journal of Electrical & Electronics Engineering*, *4*(2), 1101-1109.

Karaoglu, H. S. (2018). *A Stock Trading Application Using Deep Learning* (Master Dissertation), Bahçeşehir Üniversitesi, Fen Bilimleri Enstitüsü.

Khomenko, V., Shyshkov, O., Radyvonenko, O., & Bokhan, K. (2016). "Accelerating Recurrent Neural Network Training Using Sequence Bucketing and Multi-Gpu Data Parallelization". In *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)* (100-103). IEEE.

Kiani-B, A., & Akbarzadeh-T, M. R. (2006). "Automatic text summarization using hybrid fuzzy ga-gp". In *2006 IEEE International Conference on Fuzzy Systems* (pp. 977-983). IEEE.

Kintsch, W., & Van Dijk, T. A. (1978). "Toward a Model of text Comprehension and Production". *Psychological review*, *85*(5), 363-394.

Klönne, M. (2018). *Deep Recurrent Neural Networks for Abstractive Text Summarization* (Doctoral Dissertation), Universität Bielefeld.

Knight, K., & Marcu, D. (2000). "Statistics-based Summarization-step one: Sentence Compression", *17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, Austin, *AAAI/IAAI*, *2000*, 703-710.

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

Kutlu, M., Cığır, C., & Cicekli, I. (2010). "Generic Text Summarization For Turkish". *The Computer Journal*, 53(8), 1315-1323.

Lin, C. Y., & Hovy, E. (2003). "Automatic Evaluation of Summaries Using n-gram co-occurrence Statistics". In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 150-157).

Lin, C. Y. (2004). "ROUGE: A Package For Automatic Evaluation of Summaries". In *Text Summarization Branches Out* (pp. 74-81).

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). "A Critical Review of Recurrent Neural Networks for Sequence Learning". *arXiv preprint arXiv:1506.00019*.

Litvak, M., & Vanetik, N. (2014). "Multi-document Summarization Using Tensor Decomposition". *Computación y Sistemas*, *18*(3), 581-589.

Lloret, E., & Palomar, M. (2012). "Text Summarization in Progress: A Literature Review". *Artificial Intelligence Review*, *37*(1), 1-41.

Luhn, H. P. (1958). "The Automatic Creation of Literature Abstracts". *IBM Journal of Research and Development*, *2*(2), 159-165.

Luo, T., Guo, K., & Guo, H. (2019, December). "Automatic Text Summarization Based on Transformer and Switchable Normalization". In *2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)* (pp. 1606-1611). IEEE.

Mani, I. (2001a). *Automatic summarization* (Vol. 3). 1st Edn John Benjamins Publishing. Amsterdam, ISBN-10: 9027249865, (pp: 285).

Mani, I. (2001b). "Recent Developments in text Summarization". In *Proceedings of the Tenth International Conference on Information and Knowledge Management* (pp. 529-531).

Mani, I., & Bloedorn, E. (1997). "Multi-document Summarization by Graph Search and Matching". *arXiv Preprint cmp-lg/9712004, The Ninth Conference On Innovative Applications of Artificial Intelligence (AAAI/IAAI)*, 27-31.

Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., & Sundheim, B. M. (1999). "The TIPSTER SUMMAC text Summarization Evaluation". In *Ninth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 77-85).

Mani, I., Gates, B., & Bloedorn, E. (1999). "Improving Summaries by Revising Them". In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 558-565).

Mani, I., & Maybury, M. T. (1999). *Advances in Automatic Text Summarization MIT Press* (pp. 81-94).

Mann, W. C., & Thompson, S. A. (1988). "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization". *Text-interdisciplinary Journal for the Study of Discourse*, *8*(3), 243-281.

Marcu, D. (1998). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts.* (Doctoral Dissertation), University of Toronto.

Marcu, D. (2001, September). "Discourse-based Summarization in duc-2001". In *Proceedings of the Document Understanding Conference (DUC01).*

Markey, K. (2007). "Twenty-five Years of end-user Searching, Part 1: Research Findings". *Journal of the American Society for Information Science and Technology*, *58*(8), 1071-1081.

McKeown, K., & Radev, D. R. (1995). "Generating Summaries of Multiple News Articles". In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 74-82).

Meena, Y. K., & Gopalani, D. (2015). "Domain Independent Framework for Automatic Text Summarization". *Procedia Computer Science*, *48*, 722-727.

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).

Moratanch, N., & Chitrakala, S. (2016, March). "A Survey on Abstractive Text Summarization". In *2016 International Conference on Circuit, power and computing technologies (ICCPCT)* (pp. 1-7). IEEE.

Nallapati, R., Zhai, F., & Zhou, B. (2017, February). "Summarunner: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Document"s. In *Thirty-first AAAI conference on artificial intelligence.*

Nallapati, R., Zhou, B., Gulcehre, C., & Xiang, B. (2016). "Abstractive Text Summarization Using sequence-to-sequence rnns and Beyond". *arXiv preprint arXiv:1602.06023*.

Nenkova, A., & McKeown, K. (2011). "Automatic Summarization". *Foundations and Trends. in Information Retrieval*, *5*(2–3), 103-233.

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).

Neto, J. L., Freitas, A. A., & Kaestner, C. A. (2002, November). "Automatic Text Summarization Using A Machine Learning Approach". *Advances in Artificial Intelligence, 16th* In *Brazilian Symposium on Artificial Intelligence* (pp. 205-215). Springer, Berlin, Heidelberg.

Oflazer K., & Kuruöz, İ. (1994). "Tagging and Morphological Disambiguation of Turkish Text", *Proceedings of the Fourth Conference on Applied Natural Language Processing*, October 13-15, 1994, Stuttgart, Germany.

Olah, C. (2015). Understanding LSTM Networks. colah.github.io/posts/2015-08-Understanding-LSTMs. Erişim Tarihi: 10/20/2019.

Osborne, M. (2002, July). "Using Maximum Entropy for Sentence Extraction". In *Proceedings of the ACL-02 Workshop on Automatic Summarization* (1-8).

Over, P., & Liggett, W. (2002). "Introduction to duc-2002: an Intrinsic Evaluation of Generic News Text". In *Document Understanding Conference*.

Over, P. (2003). “An Introduction to DUC 2003: Intrinsic Evaluation of Generic News Text Summarization Systems”. In *Proceedings of Document Understanding Conference 2003*.

Ou, S., Khoo, C. S. G., & Goh, D. H. (2008). “Design and Development of a Concept-based Multi-Document Summarization System for Research Abstracts”. *Journal of Information Science*, *34*(3), 308-326.

Özkan, C. (2019). *İnternet tabanlı Türkçe metinler için otomatik özetleme tekniği* (Yayımlanmamış Yüksek Lisans Tezi), İstanbul: Maltepe Üniversitesi, Fen Bilimleri Enstitüsü.

Özkan Çelik, A. E. (2021). *Türkçe Akademik Yayınlar İçin Yapısal Öz Çıkarım Sistemi* (Yayımlanmamış Doktora Tezi), Ankara: Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü.

Özsoy, M., Cicekli, I., & Alpaslan, F. (2010, August). “Text Summarization of Turkish Texts Using Latent Semantic Analysis”. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)* (869-876).

Pant, M., & Chopra, A. (2022). “Multilingual Financial Documentation Summarization by Team_Tredence for fns2022”. In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022* (112-115).

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: A Method for Automatic Evaluation Of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (311-318).

Patel, M., Chokshi, A., Vyas, S., & Maurya, K. (2018). “Machine Learning Approach for Automatic Text Summarization Using Neural Networks”. *International Journal of Advanced Research in Computer and Communication Engineering*, *7*(1).

Patil, K., & Brazdil, P. (2007). “Text Summarization: Using Centrality in the Pathfinder Network”. *Int. J. Comput. Sci. Inform. Syst [online]*, *2*, (18-32).

Pattnaik, S., & Nayak, A. K. (2020). “A Simple and Efficient Text Summarization Model for Odia Text Documents”. *Indian journal of Computer Science and Engineering*, *11*(6), 825-834.

Paulus, R., Xiong, C., & Socher, R. (2017). “A Deep Reinforced Model for Abstractive Summarization”. *arXiv preprint arXiv:1705.04304*.

Pembe, F. C. (2010). *Automated Query-biased and Structure-preserving Document Summarization for web Search Tasks* (Doctoral Dissertation), İstanbul: Boğaziçi University.

Pollock, J. J., & Zamora, A. (1975). “Automatic Abstracting Research at Chemical Abstracts Service”. *Journal of Chemical Information and Computer Sciences*, *15*(4), 226-232.

Radev, D. R., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Dimitrov, S., ... & Zhang, Z. (2004). “MEAD-a Platform for Multidocument Multilingual Text Summarization”, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, 26-28.

Radev, D. R., Jing, H., and Budzikowska, M. (2000). “Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies”. *In ANLP/NAACL Workshop on Summarization*, 21-30

Radev, D., Otterbacher, J., Winkel, A., & Blair-Goldensohn, S. (2005). “NewsInEssence: Summarizing Online News Topic”s. *Communications of the ACM*, *48*(10), 95-98.

Radev, D., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Qi, H., ... & Drabek, E. F. (2003). “Evaluation Challenges in Large-scale Document Summarization”. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 375-382).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). “Exploring the Limits of Transfer Learning With A Unified Text-to-text Transformer”. *J. Mach. Learn. Res.*, *21*(140), 1-67.

Rumelhart, D. E. (1977). “Toward an Interactive Model of Reading”. In *Attention and Performance VI* (pp. 573-603). Routledge.

Rush, A. M., Chopra, S., & Weston, J. (2015). “A Neural Attention Model for Abstractive Sentence Summarization”. *arXiv preprint arXiv:1509.00685*.

Rush, J. E., Salvador, R., & Zamora, A. N. T. O. N. I. O. (1971). "Automatic Abstracting and Indexing. II. Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria". *Journal of the American Society for Information Science*, *22*(4), 260-274.

Saggion, H. (2008). "A Robust and Adaptable Summarization Tool". *Traitement Automatique des Langues*, *49*(2), 103–125

Saggion, H., & Lapalme, G. (2002). "Generating Indicative-Informative Summaries With Sumum". *Computational linguistics*, *28*(4), 497-526.

Saggion, H., Radev, D. R., Teufel, S., Lam, W., & Strassel, S. M. (2002a). "Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Cross-lingual Environment". In *LREC* (pp. 747-754).

Saggion, H., Radev, D., Teufel, S., & Lam, W. (2002b). "Meta-evaluation of Summaries in a Cross-lingual Environment Using Content-based Metrics". In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Sak, H., Senior, A. W., & Beaufays, F. (2014, september). "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling". *In Fifteenth annual Conference of the International Speech Communication Association* (338-342).

Salton, G., & Buckley, C. (1988). "T"erm-weighting Approaches in Automatic Text Retrieval". *Information Processing & Management*, *24*(5), 513-523.

Sharma, B., Katyal, N., Kumar, V., & Lathwal, A. (2020). "Automatic Text Summarization Using Fuzzy Extraction". In *International Conference on Innovative Computing and Communications* (395-405). Springer, Singapore.

Shetty A., Bajaj R. (2015). "Auto Text Summarization with Categorization and Sentiment Analysis", *International Journal of Computer Applications*, 130 (7), 57-60.

Siddharthan A. (2015). "Tutorial on Abstractive Text Summarization", *NLG Summer School*, Aberdeen [power pointslides] Retrieved from https://pdfs.semanticscholar.org/presentation/3d60/4c4c5acf3287870e7d478490aa0a26121396.pdf (Erişim Tarihi : 20/12/2021).

Singh, P., Chhikara, P., & Singh, J. (2020). “An Ensemble Approach For Extractive Text Summarization”. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)* (1-7). IEEE.

Singh, S. P., Kumar, A., Mangal, A., & Singhal, S. (2016). “Bilingual Automatic Text Summarization Using Unsupervised Deep Learning”. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (pp. 1195-1200). IEEE.

Steinberger, J. ve Jezek, K. (2009). “Evaluation Measures for text Summarization”. *Computing and Informatics*, 28(2), 251-275.

Svore, K., Vanderwende, L., & Burges, C. (2007). “Enhancing Single-Document Summarization by Combining RankNet and third-party Sources”. In *Proceedings of the 2007 joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 448-457).

Şimşek, A. (2018). *Face Recognition Via Unsupervısed Deep Learning And Auto-Encoders,* (Master Dissertation), İstanbul: İstanbul University,Cerrahpasa Institute Of Graduate Studies.

Tarwani, K. M., & Edem, S. (2017). “Survey on Recurrent Neural Network in Natural Language Processing”. *Int. J. Eng. Trends Technol*, *48*(6), 301-304.

Tomer, M., & Kumar, M. (2020). “Improving Text Summarization Using Ensembled Approach Based on Fuzzy With LSTM”. *Arabian Journal for Science and Engineering*, *45*(12), 10743-10754.

Torres-Moreno, J. M. (2012). “Artex is Another Text Summarizer”. *arXiv Preprint arXiv:1210.3312*.

Torres-Moreno, J. M. (Ed.). (2014). *Automatic Text Summarization*. John Wiley & Sons.

Tür, G., Hakkani-Tür, D., & Oflazer, K. (2003). “A Statistical Information Extraction System for Turkish”. *Natural Language Engineering*, *9*(2), 181-210.

Uzun-Per, M. (2011). *Developing a Concept Extraction System for Turkish*. (Master Dissertation), İstanbul: Boğaziçi University.

Vasiliev, A. (1963), "Automatic Abstracting and Indexing", *Automatic Documentation – Storage and Retrieval,* UNESCO/NS/WS 1163.112, 1–12.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). "Attention is all You Need". *Advances in Neural Information Processing Systems*, *30*.

Witbrock, M. J., & Mittal, V. O. (1999). "Ultra-summarization (Poster Abstract) A Statistical Approach to Generating Highly Condensed Non-extractive Summaries". *In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in İnformation Retrieval* (315-316).

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2021). "mT5: A Massively Multilingual Pre-trained text-to-text Transformer". *arXiv Preprint arXiv:2010.11934*.

Yousefi-Azar, M., & Hamey, L. (2017). "Text Summarization Using Unsupervised Deep Learning". *Expert Systems with Applications*, 68, 93-105.

Zhang, Y., Wang, D., & Li, T. (2011, September). "iDVS: an Interactive Multi-Document Visual Summarization System". In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 569-584). Springer, Berlin, Heidelberg.

Zolotareva, E., Tashu, T. M., & Horváth, T. (2020, September). "Abstractive Text Summarization using Transfer Learning". *In ITAT* (75-80).

EKLER

EK 1. Özetin Üretilmesi İçin Yazılan Python Kodu

```
[ ]  from transformers import (
         AdamW,
         MT5ForConditionalGeneration,
         T5TokenizerFast as T5Tokenizer
     )
     from tqdm.auto import tqdm
```

```
     import pytorch_lightning as pl
     from pytorch_lightning.callbacks import ModelCheckpoint
     from pytorch_lightning.loggers import TensorBoardLogger
```

```
class NewsSummaryDataset(Dataset):
 def __init__(
      self,
      data: pd.DataFrame,
      tokenizer: T5Tokenizer,
      text_max_token_len: int = 800,
      summary_max_token_len: int = 64
 ):
   self.tokenizer = tokenizer
   self.data = data
   self.text_max_token_len = text_max_token_len
   self.summary_max_token_len = summary_max_token_len
 def __len__(self):
   return len(self.data)
 def __getitem__(self, index : int):
   data_row = self.data.iloc[index]

   text = data_row["text"]
   text_encoding = tokenizer(
       text,
       max_length = self.text_max_token_len,
       padding = "max_length",
       truncation = True,
       return_attention_mask = True,
       add_special_tokens = True,
       return_tensors = "pt"
   )
```

```
        summary_encoding = tokenizer(
            data_row["summary"],
            max_length = self.summary_max_token_len,
            padding = "max_length",
            truncation = True,
            return_attention_mask = True,
            add_special_tokens = True,
            return_tensors = "pt"
        )
        labels = summary_encoding["input_ids"]
        labels[labels == 0] = -100

        return dict(
            text = text,
            summary = data_row["summary"],
            text_input_ids = text_encoding["input_ids"].flatten(),
            text_attention_mask = text_encoding["attention_mask"].flatten(),
            labels = labels.flatten(),
            labels_attention_mask = summary_encoding["attention_mask"].flatten()
        )
```

```
class NewsSummaryDataModule(pl.LightningDataModule):
    def __init__(
        self,
        train_df: pd.DataFrame,
        test_df: pd.DataFrame,
        tokenizer: T5Tokenizer,
        batch_size: int = 8 ,
        text_max_token_len: int =800,
        summary_max_token_len: int = 64
    ):
        super().__init__()
        self.train_df = train_df
        self.test_df = test_df

        self.batch_size = batch_size
        self.tokenizer = tokenizer

        self.text_max_token_len = text_max_token_len
        self.summary_max_token_len = summary_max_token_len
    def setup(self, stage = None):
        self.train_dataset = NewsSummaryDataset(
            self.train_df,
            self.tokenizer,
            self.text_max_token_len,
            self.summary_max_token_len
        )
```

```
        self.test_dataset = NewsSummaryDataset(
            self.test_df,
            self.tokenizer,
            self.text_max_token_len,
            self.summary_max_token_len
        )
    def train_dataloader(self):
        return DataLoader(
            self.train_dataset,
            batch_size = self.batch_size,
            shuffle = True,
            num_workers = 4
        )
    def val_dataloader(self):
      return DataLoader(
        self.test_dataset,
        batch_size = self.batch_size,
        shuffle = True,
        num_workers = 4
      )
```

```
[ ] MODEL_NAME = "google/mt5-small"
    tokenizer = T5Tokenizer.from_pretrained(MODEL_NAME)
```

```
num_epochs = 15
BATCH_SIZE = 8

data_module = NewsSummaryDataModule(train_df, test_df, tokenizer,batch_size = BATCH_SIZE)
```

```
[ ] class NewsSummaryModel(pl.LightningModule):
      def __init__(self):
        super().__init__()
        self.model =  MT5ForConditionalGeneration.from_pretrained(MODEL_NAME, return_dict = True)
      def forward(self, input_ids, attention_mask, decoder_attention_mask, labels = None):
        output = self.model(
            input_ids,
            attention_mask = attention_mask,
            labels = labels,
            decoder_attention_mask = decoder_attention_mask
        )

        return output.loss, output.logits

      def training_step(self, batch, batch_idx):
        input_ids = batch["text_input_ids"]
        attention_mask = batch["text_attention_mask"]
        labels = batch["labels"]
        labels_attention_mask = batch["labels_attention_mask"]
```

```
        loss, outputs = self(
            input_ids = input_ids,
            attention_mask = attention_mask,
            decoder_attention_mask = labels_attention_mask,
            labels = labels
        )
        self.log("train_loss", loss, prog_bar = True, logger = True)
        return loss
    def validation_step(self, batch, batch_idx):
        input_ids = batch["text_input_ids"]
        attention_mask = batch["text_attention_mask"]
        labels = batch["labels"]
        labels_attention_mask = batch["labels_attention_mask"]

        loss, outputs = self(
            input_ids = input_ids,
            attention_mask = attention_mask,
            decoder_attention_mask = labels_attention_mask,
            labels = labels
        )
        self.log("val_loss", loss, prog_bar = True, logger = True)
        return loss


    def test_step(self, batch, batch_idx):
        input_ids = batch["text_input_ids"]
        attention_mask = batch["text_attention_mask"]
        labels = batch["labels"]
        labels_attention_mask = batch["labels_attention_mask"]

        loss, outputs = self(
            input_ids = input_ids,
            attention_mask = attention_mask,
            decoder_attention_mask = labels_attention_mask,
            labels = labels
        )

        self.log("test_loss", loss, prog_bar = True, logger = True)
        return loss

    def configure_optimizers(self):
        return AdamW(self.parameters(), lr = 0.001)
```

```
[ ]  model = NewsSummaryModel()
```

```
[ ] !pip install tensorboard
```

```
checkpoint_callback = ModelCheckpoint(
        dirpath ="/content/drive/MyDrive/Text Summarization/FinalModels/Data/checkpoints",
        filename = "Paper-best-checkpoint-text-to-Title",
        save_top_k=1,
        verbose=True,
        monitor='val_loss',
        mode='min'
    )
```

```
trainer = pl.Trainer(
    checkpoint_callback = checkpoint_callback,
    max_epochs = num_epochs,
    gpus = 1,
    progress_bar_refresh_rate = 30
)
```

```
[ ] %load_ext tensorboard
    %tensorboard --logdir ./lightning_logs
```

```
trainer.fit(model, data_module)
```

```
[ ] model.freeze()
```

```
[ ] def summarize(text):
      text_encoding = tokenizer(
          text,
          max_length = 800,
          padding = "max_length",
          truncation = True,
          return_attention_mask = True,
          add_special_tokens = True,
          return_tensors = "pt"
      )

      generated_ids = model.model.generate(
          input_ids = text_encoding["input_ids"],
          attention_mask = text_encoding["attention_mask"],
          max_length = 64,
          num_beams = 2,
          repetition_penalty=2.5,
          length_penalty = 1.0,
          early_stopping = True
      )

      preds = [
          tokenizer.decode(gen_id, skip_special_tokens = True,clean_up_tokenization_spaces= True )
          for gen_id in generated_ids
      ]

      return "".join(preds)
```

```
[ ] text = dataset['text'][100002]
    model_summary = summarize(text)
```

EK 2. ROUGE Değerlerini Elde Etmek İçin Yazılan Python Kodu

```
!pip install rouge
```

```
[ ]  modelSummary = []
     for i in range(0,101):
       modelSummary.append(summarize(dataset2['validation']['Text'][i]))
```

```
[ ]  refSummary = []
     for i in range(0,101):
       refSummary.append(dataset2['validation']['Abstract'][i])
```

```
[ ]  from rouge import Rouge
     rouge = Rouge()
     scores = rouge.get_scores(modelSummary, refSummary, avg=True)
```

EK 3. Çoğul Ekinin Silinmesinin İçin Yazılan Python Kodunun Bir Kısımı

```
def larlersilme(text):

  #dırlar and dirler and tırlar and tirler
  text = ' '.join(word[:-3] if word[::-1][0:6] =='relrit' else word for word in text.split())
  text = ' '.join(word[:-3] if word[::-1][0:6] =='ralrıt' else word for word in text.split())
  text = ' '.join(word[:-3] if word[::-1][0:6] =='relrid' else word for word in text.split())
  text = ' '.join(word[:-3] if word[::-1][0:6] =='ralrıd' else word for word in text.split())

  #larıdır and leridir with F S T K Ç Ş H P
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdıralf' and word[::-1][8] =='o' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdıralf' and word[::-1][8] =='u' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdıralf' and word[::-1][8] =='a' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdıralf' and word[::-1][8] =='ı' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdırals' and word[::-1][8] =='o' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdırals' and word[::-1][8] =='u' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdırals' and word[::-1][8] =='a' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdırals' and word[::-1][8] =='ı' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdıralt' and word[::-1][8] =='o' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdıralt' and word[::-1][8] =='u' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdıralt' and word[::-1][8] =='a' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdıralt' and word[::-1][8] =='ı' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdıralk' and word[::-1][8] =='o' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdıralk' and word[::-1][8] =='u' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdıralk' and word[::-1][8] =='a' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdıralk' and word[::-1][8] =='ı' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdıralç' and word[::-1][8] =='o' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdıralç' and word[::-1][8] =='u' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdıralç' and word[::-1][8] =='a' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdıralç' and word[::-1][8] =='ı' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdıralş' and word[::-1][8] =='o' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdıralş' and word[::-1][8] =='u' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdıralş' and word[::-1][8] =='a' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdıralş' and word[::-1][8] =='ı' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdıralh' and word[::-1][8] =='o' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdıralh' and word[::-1][8] =='u' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdıralh' and word[::-1][8] =='a' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdıralh' and word[::-1][8] =='ı' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdıralp' and word[::-1][8] =='o' else word for word in text.split())
  text = ' '.join(word[:-7]+'tur' if len(word)>8 and word[::-1][0:8] =='rıdıralp' and word[::-1][8] =='u' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdıralp' and word[::-1][8] =='a' else word for word in text.split())
  text = ' '.join(word[:-7]+'tır' if len(word)>8 and word[::-1][0:8] =='rıdıralp' and word[::-1][8] =='ı' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirelf' and word[::-1][8] =='ö' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirelf' and word[::-1][8] =='ü' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirelf' and word[::-1][8] =='e' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirelf' and word[::-1][8] =='i' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirels' and word[::-1][8] =='ö' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirels' and word[::-1][8] =='ü' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirels' and word[::-1][8] =='e' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirels' and word[::-1][8] =='i' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirelt' and word[::-1][8] =='ö' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirelt' and word[::-1][8] =='ü' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirelt' and word[::-1][8] =='e' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirelt' and word[::-1][8] =='i' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirelk' and word[::-1][8] =='ö' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirelk' and word[::-1][8] =='ü' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirelk' and word[::-1][8] =='e' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirelk' and word[::-1][8] =='i' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirelç' and word[::-1][8] =='ö' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirelç' and word[::-1][8] =='ü' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirelç' and word[::-1][8] =='e' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirelç' and word[::-1][8] =='i' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirelş' and word[::-1][8] =='ö' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirelş' and word[::-1][8] =='ü' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirelş' and word[::-1][8] =='e' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirelş' and word[::-1][8] =='i' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirelh' and word[::-1][8] =='ö' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirelh' and word[::-1][8] =='ü' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirelh' and word[::-1][8] =='e' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirelh' and word[::-1][8] =='i' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirelp' and word[::-1][8] =='ö' else word for word in text.split())
  text = ' '.join(word[:-7]+'tür' if len(word)>8 and word[::-1][0:8] =='ridirelp' and word[::-1][8] =='ü' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirelp' and word[::-1][8] =='e' else word for word in text.split())
  text = ' '.join(word[:-7]+'tir' if len(word)>8 and word[::-1][0:8] =='ridirelp' and word[::-1][8] =='i' else word for word in text.split())
```

```
text = ' '.join(word[:-7]+'tır' if word[::-1][0:8] =='rıdıralf' else word for word in text.split())
text = ' '.join(word[:-7]+'tır' if word[::-1][0:8] =='rıdırals' else word for word in text.split())
text = ' '.join(word[:-7]+'tır' if word[::-1][0:8] =='rıdıralt' else word for word in text.split())
text = ' '.join(word[:-7]+'tır' if word[::-1][0:8] =='rıdıralk' else word for word in text.split())
text = ' '.join(word[:-7]+'tır' if word[::-1][0:8] =='rıdıralç' else word for word in text.split())
text = ' '.join(word[:-7]+'tır' if word[::-1][0:8] =='rıdıralş' else word for word in text.split())
text = ' '.join(word[:-7]+'tır' if word[::-1][0:8] =='rıdıralh' else word for word in text.split())
text = ' '.join(word[:-7]+'tır' if word[::-1][0:8] =='rıdıralp' else word for word in text.split())
text = ' '.join(word[:-7]+'tir' if word[::-1][0:8] =='ridirelf' else word for word in text.split())
text = ' '.join(word[:-7]+'tir' if word[::-1][0:8] =='ridirels' else word for word in text.split())
text = ' '.join(word[:-7]+'tir' if word[::-1][0:8] =='ridirelt' else word for word in text.split())
text = ' '.join(word[:-7]+'tir' if word[::-1][0:8] =='ridirelk' else word for word in text.split())
text = ' '.join(word[:-7]+'tir' if word[::-1][0:8] =='ridirelç' else word for word in text.split())
text = ' '.join(word[:-7]+'tir' if word[::-1][0:8] =='ridirelş' else word for word in text.split())
text = ' '.join(word[:-7]+'tir' if word[::-1][0:8] =='ridirelh' else word for word in text.split())
text = ' '.join(word[:-7]+'tir' if word[::-1][0:8] =='ridirelp' else word for word in text.split())
text = ' '.join(word[:-7]+'dur' if word[::-1][0:8] =='rıdıralo' else word for word in text.split())
text = ' '.join(word[:-7]+'dur' if word[::-1][0:8] =='rıdıralu' else word for word in text.split())
text = ' '.join(word[:-7]+'dır' if word[::-1][0:8] =='rıdıralı' else word for word in text.split())
text = ' '.join(word[:-7]+'dır' if word[::-1][0:8] =='rıdırala' else word for word in text.split())
text = ' '.join(word[:-7]+'dür' if word[::-1][0:8] =='ridirelö' else word for word in text.split())
text = ' '.join(word[:-7]+'dür' if word[::-1][0:8] =='ridirelü' else word for word in text.split())
text = ' '.join(word[:-7]+'dir' if word[::-1][0:8] =='ridireli' else word for word in text.split())
text = ' '.join(word[:-7]+'dir' if word[::-1][0:8] =='ridirele' else word for word in text.split())
text = ' '.join(word[:-7]+'dur' if len(word)>8 and word[::-1][0:7] =='rıdıral' and word[::-1][8] =='o' else word for word in text.split())
text = ' '.join(word[:-7]+'dur' if len(word)>8 and word[::-1][0:7] =='rıdıral' and word[::-1][8] =='u' else word for word in text.split())
text = ' '.join(word[:-7]+'dır' if len(word)>8 and word[::-1][0:7] =='rıdıral' and word[::-1][8] =='a' else word for word in text.split())
text = ' '.join(word[:-7]+'dır' if len(word)>8 and word[::-1][0:7] =='rıdıral' and word[::-1][8] =='ı' else word for word in text.split())
text = ' '.join(word[:-7]+'dür' if len(word)>8 and word[::-1][0:7] =='ridirel' and word[::-1][8] =='ö' else word for word in text.split())
text = ' '.join(word[:-7]+'dür' if len(word)>8 and word[::-1][0:7] =='ridirel' and word[::-1][8] =='ü' else word for word in text.split())
text = ' '.join(word[:-7]+'dir' if len(word)>8 and word[::-1][0:7] =='ridirel' and word[::-1][8] =='e' else word for word in text.split())
text = ' '.join(word[:-7]+'dir' if len(word)>8 and word[::-1][0:7] =='ridirel' and word[::-1][8] =='i' else word for word in text.split())
text = ' '.join(word[:-7]+'dır' if word[::-1][0:7] =='rıdıral' else word for word in text.split())
text = ' '.join(word[:-7]+'dir' if word[::-1][0:7] =='ridirel' else word for word in text.split())
```

```
#mişlerdir and mışlardır
text = ' '.join(word[:-6] + 'tir' if word[::-1][0:9] =='ridrelşim' else word for word in text.split())
text = ' '.join(word[:-6] + 'tır' if word[::-1][0:9] =='rıdralşım' else word for word in text.split())
text = ' '.join(word[:-6] + 'tür' if word[::-1][0:9] =='ridrelşüm' else word for word in text.split())
text = ' '.join(word[:-6] + 'tur' if word[::-1][0:9] =='rıdralşum' else word for word in text.split())


#ile
text = ' '.join(word[:-3] if word[::-1][0:4] =='alya' else word for word in text.split())
text = ' '.join(word[:-3] if word[::-1][0:4] =='elye' else word for word in text.split())
text = ' '.join(word[:-3] if word[::-1][0:4] =='elyü' else word for word in text.split())
text = ' '.join(word[:-3] if word[::-1][0:4] =='elyö' else word for word in text.split())
text = ' '.join(word[:-3] if word[::-1][0:4] =='alyu' else word for word in text.split())
text = ' '.join(word[:-3] if word[::-1][0:4] =='alyo' else word for word in text.split())
text = ' '.join(word[:-3] if word[::-1][0:4] =='elyi' else word for word in text.split())
text = ' '.join(word[:-3] if word[::-1][0:4] =='alyı' else word for word in text.split())

#lar and ler with ile
text = ' '.join(word[:-5] if word[::-1][0:5] =='elrel' else word for word in text.split())
text = ' '.join(word[:-5] if word[::-1][0:5] =='alral' else word for word in text.split())

#larının and lerinin
text = ' '.join(word[:-7] if word[::-1][0:7] =='nınıral' else word for word in text.split())
text = ' '.join(word[:-7] if word[::-1][0:7] =='ninirel' else word for word in text.split())
```

```
#larınca and lerince with F S T K Ç Ş H P
text = ' '.join(word[:-7]+'ça' if word[::-1][0:8] =='acnıralf' else word for word in text.split())
text = ' '.join(word[:-7]+'ça' if word[::-1][0:8] =='acnırals' else word for word in text.split())
text = ' '.join(word[:-7]+'ça' if word[::-1][0:8] =='acnıralt' else word for word in text.split())
text = ' '.join(word[:-7]+'ça' if word[::-1][0:8] =='acnıralk' else word for word in text.split())
text = ' '.join(word[:-7]+'ça' if word[::-1][0:8] =='acnıralç' else word for word in text.split())
text = ' '.join(word[:-7]+'ça' if word[::-1][0:8] =='acnıralş' else word for word in text.split())
text = ' '.join(word[:-7]+'ça' if word[::-1][0:8] =='acnıralh' else word for word in text.split())
text = ' '.join(word[:-7]+'ça' if word[::-1][0:8] =='acnıralp' else word for word in text.split())
text = ' '.join(word[:-7]+'çe' if word[::-1][0:8] =='ecnirelf' else word for word in text.split())
text = ' '.join(word[:-7]+'çe' if word[::-1][0:8] =='ecnirels' else word for word in text.split())
text = ' '.join(word[:-7]+'çe' if word[::-1][0:8] =='ecnirelt' else word for word in text.split())
text = ' '.join(word[:-7]+'çe' if word[::-1][0:8] =='ecnirelk' else word for word in text.split())
text = ' '.join(word[:-7]+'çe' if word[::-1][0:8] =='ecnirelç' else word for word in text.split())
text = ' '.join(word[:-7]+'çe' if word[::-1][0:8] =='ecnirelş' else word for word in text.split())
text = ' '.join(word[:-7]+'çe' if word[::-1][0:8] =='ecnirelh' else word for word in text.split())
text = ' '.join(word[:-7]+'çe' if word[::-1][0:8] =='ecnirelp' else word for word in text.split())
text = ' '.join(word[:-7]+'ça' if word[::-1][0:7] =='acnıral' else word for word in text.split())
text = ' '.join(word[:-7]+'ce' if word[::-1][0:7] =='ecnirel' else word for word in text.split())
```

```
#larca and lerce with F S T K Ç Ş H P
text = ' '.join(word[:-5]+'ça' if word[::-1][0:6] =='acralf' else word for word in text.split())
text = ' '.join(word[:-5]+'ça' if word[::-1][0:6] =='acrals' else word for word in text.split())
text = ' '.join(word[:-5]+'ça' if word[::-1][0:6] =='acralt' else word for word in text.split())
text = ' '.join(word[:-5]+'ça' if word[::-1][0:6] =='acralk' else word for word in text.split())
text = ' '.join(word[:-5]+'ça' if word[::-1][0:6] =='acralç' else word for word in text.split())
text = ' '.join(word[:-5]+'ça' if word[::-1][0:6] =='acralş' else word for word in text.split())
text = ' '.join(word[:-5]+'ça' if word[::-1][0:6] =='acralh' else word for word in text.split())
text = ' '.join(word[:-5]+'ça' if word[::-1][0:6] =='acralp' else word for word in text.split())
text = ' '.join(word[:-5]+'çe' if word[::-1][0:6] =='ecrelf' else word for word in text.split())
text = ' '.join(word[:-5]+'çe' if word[::-1][0:6] =='ecrels' else word for word in text.split())
text = ' '.join(word[:-5]+'çe' if word[::-1][0:6] =='ecrelt' else word for word in text.split())
text = ' '.join(word[:-5]+'çe' if word[::-1][0:6] =='ecrelk' else word for word in text.split())
text = ' '.join(word[:-5]+'çe' if word[::-1][0:6] =='ecrelç' else word for word in text.split())
text = ' '.join(word[:-5]+'çe' if word[::-1][0:6] =='ecrelş' else word for word in text.split())
text = ' '.join(word[:-5]+'çe' if word[::-1][0:6] =='ecrelh' else word for word in text.split())
text = ' '.join(word[:-5]+'çe' if word[::-1][0:6] =='ecrelp' else word for word in text.split())
text = ' '.join(word[:-5]+'ça' if word[::-1][0:5] =='acral' else word for word in text.split())
text = ' '.join(word[:-5]+'ce' if word[::-1][0:5] =='ecrel' else word for word in text.split())
```

EK 4. BERT Çıkarımsal Özetleme için Yazılan Python Kodu

```
[ ]  !pip install transformers
     !pip install bert-extractive-summarizer
```

```
[ ]  from summarizer import Summarizer
     bert_model = Summarizer()
```

```
[ ]  cleaned_body = []
     clean_abstract = []
```

```
[ ]  for i in range(0, 101):
       sentences=nltk.sent_tokenize(dataset['train']['text'][i])
       if len(sentences)<28:
         cleaned_body.append(dataset['train']['text'][i])
         clean_abstract.append(dataset['train']['summary'][i])
         print('%d------------------>%d' %(i, len(dataset['train']['text'][i])))
       else:
         bert_summary = ''.join(bert_model(dataset['train']['text'][i], min_length=60, num_sentences=6))
         cleaned_body.append(dataset['train']['text'][i])
         clean_abstract.append(bert_summary)
         print('%d------------------>length: %d-------------->New Length: %d' %(i, len(dataset['train']['text'][i]), len(bert_summary)))
```

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Neda ALIOUR
Doğum Yeri ve Tarihi	
Eğitim Durumu	
Lisans	
Y. Lisans	
Y. Lisans	
Bildiği Yabancı Diller	
İş Deneyimi	
Çalıştığı Kurumlar	
İletişim	
E-Posta Adresi	
Telefon	
Tarih	13.01.2023