T.C.
SÜLEYMAN DEMİREL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ KULLANARAK METİN ÖZETLEME

Gülnihal UYKUN

Danışman
Doç. Dr. Arif KOYUN

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
ISPARTA- 2024

İÇİNDEKİLER

ÖZET

Yüksek Lisans Tezi

MAKİNE ÖĞRENMESİ KULLANARAK METİN ÖZETLEME

Gülnihal UYKUN

Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Arif KOYUN

Bilgi kaynaklarının çokluğu ve bilgi içeren belgelerin büyüklüğü, metinden bilgi edinme işleminin oldukça fazla zaman almasına sebep olmaktadır. Bilişim teknolojilerindeki gelişmeler, metinden bilgiye erişmek için kullanılabilecek farklı yöntemlerin arayışını ortaya çıkarmıştır. Metinden bilgiye erişmek amacıyla, otomatik metin özetleme sistemleri geliştirilmiştir. Otomatik metin özetleme sistemleri bir belgeyi girdi olarak alır ve çıktı olarak belgenin önemli kısımlarını daha kısa, anlaşılabilir bir şekilde ortaya çıkarır. Otomatik metin özetleme işlemi, uzun metinlerden bilgiye erişim için harcanan zamanı kısaltır ve daha fazla dokümanın kısıtlı zaman aralığında incelenmesine olanak tanır.

Doğal Dil İşleme bilim dalının çalışma alanlarından olan otomatik özetleme işlemi, 1958 yılından itibaren çeşitli istatistiksel ve derin öğrenme tabanlı yöntemler ile gerçekleştirilmektedir. Transformer mimarisinin doğal dil işleme çalışmalarında kullanılmasından sonra büyük dil modelleri ortaya çıkmıştır ve bu dil modellerinin metin özetleme işlemlerinde oldukça başarılı sonuçlar elde ettiği gözlemlenmiştir. Büyük dil modellerinin yeniden eğitilmesi için güçlü donanıma sahip bilgisayarlara ihtiyaç duyulmaktadır ve bu oldukça maliyetli bir işlemdir. Bu dil modellerinin sıradan kullanıcılar tarafından kullanılabilmesi transfer öğrenme yöntemleri ile mümkün olmaktadır.

Bu tez çalışmasında Türkçe dilinde yoruma dayalı otomatik metin özetleme çalışması gerçekleştirilmiştir. Bu çalışma için, bilimsel dergileri çevrimiçi yayınlama aracı olan Dergipark üzerinden toplam 200 adet bilimsel makale toplanmıştır. Eğitim bilimleri, bilişim teknolojileri, iktisat ve işletme bilim alanlarında yayınlanmış olan dergilerden toplanan bu bilimsel makaleler ile Türkçe makale veri seti oluşturulmuştur. Veri setinde, bilimsel makalenin tüm metin içeriği, özeti, bilim alanı, makalenin adı ve anahtar kelimeleri bulunmaktadır. Büyük dil modellerinden İngilizce bir veri seti üzerinde ön eğitilmiş PEGASUS dil modeli, oluşturulan Türkçe makale veri seti ile ince ayarlama yapılarak yeni bir dil modeli elde edilmiştir. Elde edilen dil modeli ile Türkçe makale veri seti üzerinde yoruma dayalı metin özetleme çalışması gerçekleştirilmiştir.

Otomatik metin özetleme işleminin gerçekleştirilebilmesi için metnin tokenlere ayrılması gerekmektedir. Bu çalışmada veri setine özgü tokenizer oluşturulmuş ve bu tokenizer yoruma dayalı metin özetleme işleminde kullanılmıştır. Oluşturulan özetler,

dil modelinin kendi tokenizeri kullanılarak elde edilen özetler ile karşılaştırılmıştır. Karşılaştırma işlemi, makale yazarının yazmış olduğu özet referans olarak kabul edilerek otomatik metin özetleme sonrasında elde edilen özetler ile ROUGE metrikleri kullanılarak ve anlamsal bütünlüğü incelenerek gerçekleştirilmiştir. Veri setine özgü oluşturulan tokenizer ile yapılan özetleme çalışmasının ROUGE değerleri daha düşük olduğu halde makalenin içeriğine daha benzer sonuçlar elde edildiği gözlemlenmiştir.

Ön eğitimi İngilizce veri seti üzerinde yapılmış olan PEGASUS dil modeli için Türkçe belirli bir amaca özgü veri seti ile ince ayarlama işlemine literatürde rastlanmamıştır. Bu sebeple elde edilen sonuçlar, farklı çalışmalar ile kıyaslanamamıştır. Bu çalışmanın PEGASUS dil modeli ile Türkçe dilinde daha sonra yapılacak çalışmalar için bir temel oluşturacağı düşünülmektedir.

Anahtar Kelimeler: PEGASUS, yoruma dayalı metin özetleme, transformer, transfer öğrenme

2024, 57 sayfa

ABSTRACT

M.Sc. Thesis

TEXT SUMMARIZATION USING MACHINE LEARNING

Gülnihal UYKUN

Süleyman Demirel University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Assoc. Prof. Dr. Arif KOYUN

The abundance of information sources and the size of the documents containing information cause the process of obtaining information from the text to take a lot of time. Developments in information technologies have led to the search for different methods that can be used to access information from text. Automatic text summarization systems have been developed in order to access information from text. Automated text summarization systems take a document as input and output important parts of the document in a more concise, understandable way. Automatic text summarization reduces the time spent retrieving information from long texts and allows more documents to be examined in a limited time period.

Automatic summarization, which is one of the fields of study of the Natural Language Processing branch of science, has been carried out with various statistical and deep learning-based methods since 1958. After the use of Transformer architecture in natural language processing studies, large language models have emerged and it has been observed that these language models achieve very successful results in text summarization processes. Computers with powerful hardware are needed to retrain large language models, and this is a very costly process. The ability to use these language models by ordinary users is possible with transfer learning methods.

In this thesis study, a comment-based automatic text summarization study was carried out in Turkish. For this study, a total of 200 scientific articles were collected through Dergipark, an online publishing tool for scientific journals. A Turkish article data set was created with these scientific articles collected from journals published in the fields of educational sciences, information technologies, economics and business science. The data set includes the entire text content of the scientific article, its abstract, field of science, name of the article and keywords. A new language model was obtained by fine-tuning the PEGASUS language model, which was pre-trained on an English dataset from large language models, with the created Turkish article dataset. With the language model obtained, a comment-based text summarization study was carried out on the Turkish article data set.

In order for automatic text summarization to be performed, the text must be divided into tokens. In this study, a tokenizer specific to the data set was created and this tokenizer was used in the comment-based text summarization process. The created summaries were compared with the summaries obtained using the language model's

own tokenizer. The comparison process was carried out by using the summaries obtained after automatic text summarization and ROUGE metrics, and by examining their semantic integrity, considering the summary written by the article author as the reference. It was observed that the summarization study conducted with the tokenizer created specifically for the data set yielded results more similar to the content of the article, although the ROUGE values were lower.

For the PEGASUS language model, which was pre-trained on an English data set, fine-tuning with a Turkish data set specific to a specific purpose has not been found in the literature. For this reason, the obtained results could not be compared with different studies. It is thought that this study will form a basis for future studies in Turkish language with the PEGASUS language model.

Keywords: PEGASUS, comment-based text summarization, transformer, transfer learning

2024, 57 pages

TEŞEKKÜR

Bu araştırma için beni yönlendiren, karşılaştığım zorluklar karşısında anlayışla yaklaşarak bilgi ve tecrübesi ile aşmamda yardımcı olan değerli Danışman Hocam Doç. Dr. Arif KOYUN'a sonsuz saygılarımı ve teşekkürlerimi sunarım.

Tezimin her aşamasında beni yalnız bırakmayan değerli aileme ve sevgili eşime sonsuz teşekkürlerimi sunarım.

Gülnihal UYKUN
ISPARTA, 2024

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

BERT	Bidirectional encoder representations from transformers
BoW	Bag of words
DDİ	Doğal dil işleme
GAA	Gizli anlamsal analiz
GPT	Generative pre-trained transformer
GSG	Gap sentence generation
LSA	Latent semantic analysis
LSTM	Long short-term memory
NLP	Natural language processing
OMÖ	Otomatik metin özetleme
PEGASUS	Pre-training with extracted gap-sentences for abstractive summarization
RNN	Yinelemeli sinir ağı
ROUGE	Recall-oriented understudy for gisting evaluation
Seq2seq	Sequence to sequence
STSb	Semantic textual similarity benchmark
SVD	Singular value decomposition
T5	Text to text transfer transformer
TF-IDF	Term frequency - inverse document frequency

1. GİRİŞ

Bilişim teknolojilerindeki gelişmeler, bir bilgisayarın insanın konuşmasını anlayabilmesi ve insanın anlayabileceği dilde insanlarla etkileşim kurabilmesi düşüncesini ortaya çıkarmıştır. İnsan bilgisayar etkileşiminde doğal dillerin kullanılabilmesi amacıyla yapılan çalışmalar Doğal Dil İşleme (DDİ) bilim alanının ortaya çıkmasına sebep olmuştur. DDİ, yazım yanlışlarının düzeltilmesi ve metindeki bazı kelime veya cümle yapılarının bulunup değiştirilmesi gibi yazım yardımcı araçların geliştirilmesi, bilgisayarın sesli şekilde iletişim kurması, konuşmayı anlaması, soruları yanıtlaması gibi sesli etkileşim araçlarının geliştirilmesi, bilgisayarlı çeviri, doğal diller arası çeviri gibi konularla ilgilenmektedir (Adalı, 2020).

Teknolojinin hızlı gelişimi ve gündelik hayatın neredeyse her alanında kullanılması ile her gün büyük miktarlarda bilgi içeren veriye maruz kalınmaktadır. İnternet üzerindeki veri miktarı her gün artmaktadır ve çoğunlukla bu veriler metin türündedir. Spesifik bir konuda bir bilgiye erişebilmek için var olan metin kaynaklarının baştan sona okunması ve buradan bilginin yorumlanarak çıkarılması gerekir. Sürekli artış gösteren veri miktarı sebebi ile istenen belirli bir konudaki bilgiye ulaşmak zaman alıcı ve uğraştırıcı hale gelmiştir. Teknolojinin gelişimi ile metin türünde saklanan bu verilerin içerisinden istenilen bilgiye ulaşmanın daha basit ve hızlı yolları günümüzün önemli araştırma konularından birisi haline gelmiştir. Metinden bilgiye erişmek için yapılan çalışmalar Doğal Dil İşleme (DDİ) bilim alanının çalışma alanlarındandır.

Metin özetleme, orijinal metin içerisindeki temel bilgileri, içeriği ve genel anlamı korurken metnin daha kısa, akıcı bir özünü oluşturmayı amaçlar. Çevrimiçi olarak sunulan metinlerin miktarı çok fazladır ve bilgiye ulaşmak çok fazla zaman alan yorucu bir iş haline gelmiştir. Otomatik Metin Özetleme (OMÖ), bir belgeyi girdi olarak alır ve çıktı olarak büyük metinlerdeki önemli bölümlerin hızlı bir şekilde sunulmasını sağlar ve metinden bilgi çıkarmak için yapılacak olan okuma süresini azaltır. OMÖ belirli bir konuda araştırma yapılırken hangi dokümanların okunacağına karar verme sürecinde de fayda sağlamaktadır (Torres-Moreno, 2014; Güran, 2013).

OMÖ, metin formatındaki veriler için geliştirilmiş ve büyük boyutlu bir metini daha küçük boyutlu bir metin olarak ifade etmeyi amaçlamaktadır. OMÖ çalışmalarında

kaynak metnin anlamsal ve sözdizimsel bütünlüğü bozulmadan özet metinlerin oluşturulması önemlidir. Bu nedenle OMÖ bir boyut indirgeme yöntemi olmasının yanı sıra yorumlanabilir anlamlı bir metin sadeleştirme şeklidir ve gerekli kaynakların daha etkin kullanımına hizmet etmektedir (Mutlu, 2020).

Literatürde OMÖ, çıkarıma dayalı metin özetleme ve yoruma dayalı metin özetleme olmak üzere iki ana başlık altında incelenmektedir. Yoruma dayalı metin özetleme, insanların bir metni okuyup yorumlayarak, metindeki önemli bilgileri kendi cümleleri ile kullanılan dilin anlam bilimi ve biçim bilimi yapısına uygun şekilde özetlemelerine benzemektedir. Yoruma dayalı metin özetleme kullanılan dilin yapısı ile doğrudan ilişkilidir ve bilgisayar tarafından özetin ortaya çıkarılması karmaşık bir görevdir. Literatür incelendiğinde İngilizce gibi diller için çok çeşitli uygulamaları olduğu halde Türkçenin sondan eklemeli bir dil olması sebebiyle çok fazla uygulaması görülmemektedir. Çıkarıma dayalı metin özetleme ise, metni oluşturan bölümlere metindeki ana fikri taşıma kapasitelerine göre önem dereceleri atanmakta ve en önemli bölümler seçilerek özet oluşturulmaktadır. Çıkarıma dayalı metin özetleme sözcük öbeklerinin yapısı hiçbir değişime uğramadan yapılması sebebi ile dilden bağımsızdır. Türkçe OMÖ çalışmalarında sıklıkla kullanılmaktadır (Knight, 2002; Gupta, 2010; Mutlu, 2020).

Tezin kaynak özetleri bölümünde görüleceği üzere OMÖ çalışmalarında geçmişten günümüze istatiksel modeller, grafik tabanlı modeller, yapay sinir ağları tabanlı mimariler, kodlayıcı kod-çözücü, tekrarlayan sinir ağları, Transformer mimarisi kullanılmaktadır.

Bu tez çalışmasında Türkçe dilinde yoruma dayalı metin özetleme işlemini gerçekleştirmek için farklı alanlardan bilimsel makaleler seçilerek Türkçe veri seti oluşturulmuş, ön eğitilmiş PEGASUS dil modeli, oluşturulan veri seti kullanılarak ince ayarlama işlemi (fine-tunning) yapılmış oluşturulan özetlerle referans özetler karşılaştırılmıştır. Dil modeli üzerinde makale veri seti kullanılarak yapılan ince ayarlama işlemi metin özetlemenin amaca yönelik olmasına katkı sağlamıştır.

2. KAYNAK ÖZETLERİ

Metinlerin analizinde bilgisayar kullanımının Luhn (1958) ile başladığı söylenebilir. İngilizce dili üzerinde yapılmış olan çalışmada metin özetlemek için önemli cümleler, anahtar kelimeleri tespit etme ve kelimeler arasındaki ilişkileri inceleme yöntemleri ile bulunmaya çalışılmış ve bu cümleler birleştirilerek özet oluşturulmuştur.

1969 yılında Edmundson tarafından yapılan çalışmada metin belgelerinden bilgi çıkarımı için bilgiyi en iyi temsil etme potansiyeline sahip cümlelerin bilgisayar tarafından seçilmesi amaçlanmıştır. Çalışmada sadece anahtar kelimeleri içeren cümlelere odaklanılmamış, aynı zamanda ipucu kelimelerin, metnin başlığının ve başlıkta geçen kelimelerin bulunduğu cümleler, cümlelerin konum bilgisi de değerlendirilerek metinden cümle çıkarmaya odaklanılmıştır. Oluşturulan yeni sistemin sadece anahtar kelimelerin analizi ile elde edilen çıkarımlardan çok daha iyi sonuç verdiği gözlemlenmiştir.

Türkçenin morfolojik yapısı sebebi ile İngilizceden ayrıldığı ve DDİ (Doğal Dil İşleme) çalışmalarının daha karmaşık olduğu söylenebilir (Adalı, 2020). Türkçe DDİ çalışmalarının temelinin 1975 yılında Aydın Köksal tarafından yazılan "*Türkçenin Otomatik Morfolojik Analizi İçin Bilgisayarlı Modele İlk Yaklaşım*" isimli doktora tezi ile atıldığı söylenebilir.

Literatür incelendiğinde Türkçe DDİ alanında metinden duygu analizi, anlatım bozukluğu bulma, cümle ögelerini tespit etme, metinden anahtar kelime çıkarma gibi Türkçe DDİ bilim dalında önemli çalışmaların yapıldığı görülmektedir. Ancak bu bölümde tez konusunun kısıtı dahilinde otomatik metin özetleme ve uygulaması gerçeklenmiş olan büyük dil modellerinden PEGASUS ön-eğitilmiş dil modeli ile ilgili yapılmış olan çalışmalara yer verilmektedir.

Yapılan araştırma sonucunda Türkçe otomatik metin analizi ile ilgili yapılan en eski çalışmanın Oflazer ve Kuruöz tarafından 1994 yılında gerçekleştirildiği düşünülmektedir. Bu çalışmada Türkçe gibi aglutinatif dillerde birçok sözcüksel formun yapılarının morfolojik olarak belirsiz olması sebebi ile etiketlemenin önemli olduğu belirtilmiştir. Çalışmada Türkçe dili için yaklaşık 24.000 kök kelimeden oluşan

bir sözlüğe dayanan tam ölçekli iki seviyeli bir Türk morfolojisi spesifikasyonuna dayanan bir POS etiketi açıklanmaktadır. Oluşturulan etiketleyici ayrıca, hatalı morfolojik ayrışmalar, yaygın olarak kullanılan kökler gibi morfolojik analizörün istatistik derlemesi ve ince ayarlanması için işlevselliğe sahiptir. Ön sonuçlar, etiketleyicinin metinlerin yaklaşık %98-99'unu çok az insan müdahalesi ile doğru bir şekilde etiketleyebildiğini göstermektedir. Oluşturulan etiketleyicinin sadece Türkçe için değil diğer aglutinatif dillerde de kullanılabileceği belirtilmiştir (Oflazer ve Kuruöz, 1994).

Mani ve Bloedorn (1997) iki benzer konudaki dokümanın benzerliklerini ve farklılıklarını özetlemek amacıyla çizge tabanlı bir yöntem oluşturmuşlardır. Dokümanlardaki kelimeler, konuyla alakalı terimler ve özel isimler anlamsal ilişkilerine ve dokümandaki konumlarına göre çizge üzerinde düğüm olarak konumlandırılır. Sistem öncelikle her dokümanda konuyla anlamsal olarak ilişkili düğümleri keşfetmek için yayılım aktivasyon tekniği kullanmaktadır. Her dokümanın çizge üzerindeki temsilleri benzerlik ve farklılıklar açısından karşılaştırılmıştır ve ortaya çıkan çizge doğal dilde sunulmuştur.

Barzilay ve Elhadad (1997) bir metnin semantik analizini yapmadan bunun yerine metindeki konunuın gidişatını ele alarak kelime zinciri algoritması aracılığıyla bir özetleme sistemi geliştirmişlerdir. Yapılan sistem ilk olarak metni bölümlere ayırır, kelime zincirlerini oluşturur ve sonrasında güçlü zincirler tanımlanarak metinden anlamlı cümleler çıkartılır.

Jones (1998), otomatik özetleme alanındaki ilerlemenin analizini gerçekleştirmiş ve daha etkili bir yöntem geliştirmek için özetleme ve değerlendirme işlemleri üzerinde etkisi olan metnin giriş, amaç ve sonuç gibi bağlam faktörlerinin belirlenmesi gerektiğini öne sürmüştür. Bu analiz sonrasında potansiyel olarak daha verimli bir yöntemin geliştirilebileceğini iddia etmektedir.

Goldstein ve Carbonell (1998), metinden özetleme işlemini sorguya göre bilgi getirme ile birleştirerek gerçekleştirmiştir. Oluşturdukları Maksimal Marjinal İlgililik (Maximal Marjinal Relevance- MMR) kriteri metinden özet için uygun bölümleri seçerken sorguya uygunluğu koruyarak gereksiz bölümleri ayıklamayı

amaçlamaktadır. Elde ettikleri sonuçlar MMR kriterinin tek belge özetlemede fayda sağladığını göstermiştir. Çoklu belge özetlemede ve büyük belgelerin özetlenmesinde MMR kriteri ile elde edilen özetlerin açıkça başarı sağladığı iddia edilmiştir.

Witbrock ve Mittal (1999), mevcut çıkarıma dayalı metin özetleme yöntemlerini analiz etmiş ve çıkarıma dayalı metin özetleme işlemi yerine belgeyi anlamak ve sonrasında özet çıkarmanın ideal özeti oluşturacağını ifade etmişlerdir. Özetleme süreci için alternatif bir istatiksel model sunmuşlardır. Bu model ile terim seçimi ve terimlerin sıralaması süreçlerinin istatiksel modelleri birleştirilerek eğitim kümesinden öğrenilen tarzda kısa özetler oluşturulmuştur.

Knight ve Marcu (2000), metin özetleme çalışması için cümle sıkıştırmaya odaklanmışlardır. Sıkıştırılmış cümlelerin gramer yapısı düzgün cümleler olmasını ve belgedeki en önemli bilgileri korumasını hedeflemişlerdir. Bu soruna karar ağacı yaklaşımı ile çözüm getirmeye çalışmışlar ve elde ettikleri sonuçları insan muhakemesi ile sıkıştırılmış cümleler ile karşılaştırmışlardır.

2003 yılında Tür vd. sınırsız Türkçe metinden istatiksel yaklaşımla bilgi çıkarmaya yönelik bir çalışma gerçekleştirmiştir. Bu çalışmada Türkçenin morfolojik yapısı dikkate alınarak verileri tokenlere ayrıştıran ve bu tokenleri analiz eden bir ön işleme modülü kullanılmıştır. Kelime tabanlı model ve kök tabanlı model beraber kullanılarak metin verilerindeki konu kümelerini tanımlayacak bir modül oluşturulmuştur. Daha sonra gerçekleştirilen sistemde kişi adları, konumlar, kuruluşlar, parasal değerler, yüzde, tarihler ve saatleri temsil eden kelimeleri çıkarmak için bir modül kullanılmıştır. Oluşturulan sistemin %91,56'lık bir F-ölçüsüne ulaştığını iddia edilmiştir (Tür vd., 2003).

Altan, 50 tane Türkçe makale kullanarak ilk Türkçe otomatik metin özetleyiciyi gerçekleştirmiştir. Gerçeklenen sistem oluşturulacak özetin metnin yüzde kaçı olacağı bilgisini kullanıcıdan almaktadır. Çalışmada cümle ve sözcüklerin önceden tanımlanmış bir ağırlıklandırma metoduyla puanlanması ve bu puanların istatiksel analizi ile metnin özetinin oluşturulması sağlanmıştır. Metnin anlamsal bütünlüğünün sağlanabilmesi için ağırlıklandırma metodunun yanında cümle seçimi anlamsal bütünlüğü sağlayacak şekilde yapılmıştır. Altan metindeki pozitif, çok pozitif ve

negatif kelimeleri tespit etmiş, durak kelimeleri tespit ederek metinden çıkararak ön işleme gerçekleştirmiş, metnin başlığı bilgisini metinden çıkarmış ve paragrafları belirlemiş, her bir kelimeyi ayırmıştır. Daha sonra metnin başlığı bilgisini ve pozitif-negatif kelime bilgilerini kullanarak cümleleri ağırlıklandırmış, metinden önemli cümleleri seçerek özet oluşturmuştur. Bu çalışmada kullanılmış olan yöntemlerin günümüz otomatik metin özetleme çalışmalarının temelini oluşturduğu görülmektedir (Altan, 2004).

Bilgin vd., Türkçe için otomatik metin özetleme çalışmalarının gelişimi için oldukça önemli olan ilk Türkçe WordNet geliştirme girişimini başlatmışlardır. Balkanet projesinin bir parçası olarak EuroWordNet projesinin 1310 temel kavramını Türkçeye çevirmişler ve bu temel kavramların eşanlamlılarını, zıt anlamlılarını, alt kavramlarını otomatik şekilde oluşturmaya çalışmışlardır. Oluşturmuş oldukları WordNet'i 8000 temel kavrama kadar genişletmişlerdir (Bilgin vd., 2004).

Karakaya ve Güvenir, büyük metin arşivlerinden bilgi çıkarmak amacıyla metin sınıflandırma ve metin özetleme yöntemlerini beraber kullanarak yeni bir sistem oluşturmuşlardır. ARG adını verdikleri sistem tüm metni değil metin içerisindeki paragrafları ele alarak analiz işleminin gerçekleştirmektedir. Oluşturulan sistemde metin sınıflandırması işlemi insan muhakemesi yardımı ile yapılmaktadır. Sistemde, kullanıcı konu başlıklarını belirledikten sonra bir veya daha fazla dokümandaki metni paragraflara ayırıp konu başlıklarına göre sınıflandırmaktadır. Oluşturulan konu başlıklarının altındaki paragraflar özetlenerek konuya göre özetleme işlemi gerçekleştirilmektedir. Oluşturulan sistemin çok fazla kullanıcı müdahalesi gerektirmesi sebebi ile büyük metin arşivlerinde çalışmasının çok fazla zaman alacağı vurgulanmıştır (Karakaya ve Güvenir, 2004).

Tülek, yapmış olduğu çalışmada Türkçe'nin sondan eklemeli ve kurallı dil yapısını göz önüne alarak metin özetleme için istatiksel yöntemleri tanıtmış ve yazılımla gerçeklemiştir. Yapılan çalışmada gerçeklenen yöntemlerin Türkçeye olan uygunluğu tartışılmış farklı gövdeleme algoritmalarının özetleme başarımına etkisi incelenmiştir. Metindeki sözcüklerin olası kök ve eklerini üretmeyi amaçlayan biçim birimsel bir çözümleyici oluşturulmuş ve bu sözcükler farklı özetleme yöntemleri aracılığıyla

analiz edilerek özet cümlelerde yer alıp almayacağı belirlenmiştir. Üretilen sonuçla birleştirilerek özet metin oluşturulmuştur (Tülek, 2007).

Kutlu vd., cümle sıralama metodunu kullanarak bir metin özetleme yöntemi önermişlerdir. Oluşturulan sistemde cümlelere çeşitli özniteliklere göre puanlar atanarak en yüksek puana sahip cümleler metinden çıkarılarak özet metin oluşturulmaktadır. Metnin ana içeriğini yansıtan, tekrardan uzak bir özet oluşturmak amacıyla terim frekansı, anahtar kelimeler, metnin başlığı ile benzerlik ve cümlenin metindeki konumu gibi öznitelikler kullanılmıştır. Cümlelerin puan değerleri, özniteliklerin ağırlıklarını kullanan bir puanlama fonksiyonu yardımıyla hesaplanır. En iyi özellik puanı insan muhakemesi ile oluşturulan özetlerin makine öğrenmesi teknikleri ile eğitilmesi aracılığıyla belirlenmiştir. Performans analizi, 25 farklı insan tarafından metinlerin özetlenmesi sonucu elde edile özetlerle sistemin oluşturduğu özetlerin karşılaştırılması ile yapılmaktadır. Sonuçların gelecek çalışmalar için umut vadettiği iddia edilmiştir (Kutlu vd., 2010).

Güran, yapmış olduğu tez çalışmasında çıkarıma dayalı metin özetlemeye odaklanmış, bu amaçla ikisi yazar tarafından Türkçe haber metinleri ve insan muhakemesi ile özetlenerek hazırlanmış özetleri içeren, diğer ikisi İngilizce metin özetleme çalışmalarında sık kullanılan veri setlerini kullanmıştır. Yapılan çalışmada cümle çıkarımı için yapısal ve anlamsal özelliklerin birleşiminden oluşan melez bir sistem oluşturulmuş, gizli anlamsal analiz temelli metin özetleme yöntemlerinde kullanılabilecek yeni bir ağırlık değeri önerilmiştir. Başarım ölçütü olarak F-ölçüm skoru ve N-gram sayısına bağlı olan ROUGE değerlendirme paketi kullanılmıştır. Türkçe veri seti üzerinde değerlendirilen melez sistem uygulamasının sonucunda genetik algoritma tabanlı birleşim yöntemlerinin veri seti üzerinde olumlu etkileri olduğu sonucuna ulaşılmıştır (Güran, 2013).

Hatipoğlu ve Omurca, Türkçe metinlerde çıkarıma dayalı özetleme görevi için melez bir model önermişlerdir. Veri seti olarak Türkçe Wikipedia metinleri kullanılmış, cümle seçerek özetleme için istatiksel olarak cümlelerin puanlandırılması ve gizli anlam çıkarımı yöntemlerini sezgisel olarak birleştirerek yeni bir model oluşturmuşlardır. Önerilen modelin başarım değerlendirmesi için 10 farklı kişinin özetlenecek metinlerden özet için cümle seçmesi istenmiş ve bu özetler referans olarak

alınmıştır. Yapılan çalışma sonucunda önerilen modelin kısa metinlerde başarımının daha yüksek olduğu gözlemlenmiştir (Hatipoğlu ve Omurca, 2015).

Birant Türkçe metinlerde kural tabanlı metin özetleme üzerine çalışmıştır. Türkçenin kurallı bir dil olması sebebi ile Türkçe dil özelliklerini incelemiş ve bu özelliklere uygun bir algoritma geliştirerek çıkarıma dayalı bir özetleme işlemi gerçekleştirmiştir. Veri seti olarak 3 bilimsel makale ve 2 haber metni kullanmış, değerlendirme sonuçlarını yazarın kendi oluşturduğu özetle karşılaştırma yaparak elde etmiştir. Değerlendirme için ROUGE-n metriğini kullanmış bunun yanı sıra Dokuz Eylül Üniversitesi Dilbilim bölümü öğrencilerinden 10 tanesini seçerek onlardan değerlendirme yapmasını istemiştir. İnsan muhakemesi yolu ile yapılan değerlendirme sonuçlarına göre oluşturulan algoritmanın ürettiği özetin, metnin yazarının oluşturduğu özetten daha detaylı bilgi içerdiği iddia edilmektedir. Birant metin özetleme çalışması sırasında ortaya çıkan Türk Dil Kurumu tarafından onaylanan Eş ve Yakın Anlamlılar Sözlüğü ve Zıt Anlamlılar Sözlüğü ile literatüre önemli ölçüde katkı sağlamıştır. (Birant, 2015)

Şahin, yaptığı çalışmada Türkçe dili için anlamsal görev çözümlemesini konu edinmiştir. Türkçe için anlamsal görev çerçeveleri oluşturulmuş, belirli miktarda soru ve yanıt elle işaretlenmiş ve Türkçe için Önerme Veri Tabanı oluşturulmuştur. Türkçenin sondan eklemeli bir dil olması ve Türkçedeki eklerin çok sayıda olması sebebiyle tüm eylem içeren türetilmiş sözcüklerin kök çerçevesi kullanılarak karşılanmasına karar verilmiştir. Çalışmada çift yönlü LSTM birimlerinin alt sözcükleri işlemesi temeline dayanan bir yapay sinir ağı yöntemi önerilmiştir. Eğitilmiş sözcük vektörleri kullanılan yöntemlerin aksine alt sözcükler Türkçenin morfolojik yapısı dikkate alınarak çeşitli yöntemlerle birleştirilerek yeni sözcük vektörleri oluşturulmuştur. Alt sözcük vektörleri ve birleştirme fonksiyonları analiz edilerek etkileri ölçülmüştür. Önerilen birleştirme metodunun öncekilerden daha başarılı olduğu tespit edilmiştir (Şahin, 2018).

Karakoç ve Yılmaz yoruma dayalı metin özetleme işlemi için Türkçe haberlerden oluşan bir veri seti kullanarak haber metinlerinin içeriğine uygun başlık oluşturmaya çalışmışlardır. Veri setindeki kelime gömmeleri için FastText kullanılmış, eğitim tekrarlayan sinir ağları ile kodlayıcı-kod çözücü modeli kullanılarak

gerçekleştirilmiştir. Haber metinlerinin ilk cümlesi, ilk iki cümlesi ve tamamı kullanılarak ayrı ayrı eğitilerek test edilen sistemin başarısı ROUGE skoru ve anlamsal benzerlik ile ölçülmüş haber metininin tamamı ile eğitilen modelin başarımının daha yüksek olduğu görülmüştür (Karakoc ve Yilmaz 2019).

Doğan tarafından yapılmış olan tez çalışmasında bir konu hakkındaki duygu ve düşüncelerin analiz edilip kullanıcıya özet bilgi sunulması amaçlanmıştır. Bu amaç doğrultusunda Twitter verilerinden belirli bir konu hakkında, bir derin öğrenme yöntemi olan LSTM kullanılarak duygu sınıfı tespit edilip toparlanan verilerden LSA aracılığı ile metin özetlemesi yapılmıştır. Sınıflandırma yapılırken Türkçe dili için FastText ve Word2vec modelleri oluşturulmuş ve modeller karşılaştırılmıştır. Sonuç olarak FastText ile yapılan sınıflandırma işleminin daha başarılı olduğu gözlemlenmiştir (Doğan, 2019).

Gündeş, Türkçe haber metinleri kullanarak derin öğrenme temelli çıkarıma dayalı metin özetleme üzerine çalışmıştır. Bu çalışmada bilinen bir Türkçe haber web sitesinden alınan 2076 adet haber metni ile yeni bir veri seti oluşturulmuş, bu veri seti üzerindeki her bir haber metni sezgisel bir algoritma ile 0 veya 1 olarak etiketlenmiştir. Bu etiketler kullanılarak ön-eğitimli bir dil modeline, cümleler arasındaki anlamsal ilişkiyi yakalamak amacıyla transformer katmanları eklenmiş ve bu oluşturulan model kullanılarak metindeki cümleler puanlanmış, en yüksek 5 puana sahip cümleler birleştirilerek özet oluşturulmuştur. Önerilen model için ROUGE-1, ROUGE-2 ve ROUGE-L F ölçüm değerleri sırasıyla, 38.38, 26.8 ve 38.04 olarak bulunmuştur (Gündeş, 2020).

Afatsun, Türkçe metinlerde yoruma dayalı metin özetleme çalışması için derin öğrenme yöntemleri ile anlamlı özet oluşturma çalışması yapmıştır. Çalışmada kullanılan Türkçe veri seti Deutsche Welle haber sitesindeki metinler ve özetler ile oluşturulmuştur. Diziden diziye (Seq2seq) mimarisi kullanılmış, dikkat katmanlı kelime gömmeleri ile eğitilmiş çift yönlü bir LSTM modeli geliştirilmiştir. Modelin performansı Wikipedia ve oluşturulan Türkçe haber veri seti kelime vektörleri ile ayrı ayrı değerlendirilmiş, Türkçe haber veri setindeki Rouge-1 metriğine göre performans puanı 40.90 olarak hesaplanmıştır (Afatsun, 2020).

Beken Fikri vd., yoruma dayalı metin özetleme işleminde ROUGE ölçüm metriğinin yetersiz olduğunu gözlemlemiş ve bunun yerine kullanılabilecek anlamsal benzerliğe dayalı değerlendirme yöntemini önermişlerdir. Bunun için English Semantic Textual Similarity benchmark (STSb) veri setini Türkçeye çevirerek ilk Türkçe anlamsal metin benzerliği veri setini (STSb-TR) oluşturmuşlardır. Yaptıkları çalışma sonucunda elde edilen değerlendirme yönteminin BERTScore ve ROUGE skoru ile Pearson ve Spearman korelasyonları kullanılarak karşılaştırıldığında insan muhakemesi ile oluşturulan özetlere daha yakın bir değerlendirme sunduğu gözlemlenmiştir. (Beken Fikri vd., 2021)

Yüksel, Türkçe dili için yoruma dayalı metin özetleme işlemi gerçekleştirmiştir. Yapılan çalışma için haber metinlerini içeren veri seti oluşturulmuş, Dikkat Tabanlı Diziden Diziye Yapay Sinir Ağı, Pointer Generator Diziden Diziye Yapay Sinir ağı ve Diziden Diziye Yapay Sinir ağı ile Güçlendirmeli Öğrenme olmak üzere üç farklı model önerilmiştir. Her üç model de ConceptNet-Numberbatch kelime gömme hem de fastText kelime gömme ile öncesinde işlenmiştir. Modeller Rouge-1, Rouge-2 ve Rouge-L puanları ile değerlendirilmiş karşılaştırma yapılmıştır. En başarılı model fastText kelime gömme ile çalışılan Pointer Generator Diziden Diziye Yapay Sinir Ağı modeli olmuştur (Yüksel, 2021).

Yang vd. yapmış oldukları çalışmada ön-işlenmiş PEGASUS dil modelinin matematik problemlerinden oluşan bir veri seti üzerinde ince ayar yaparak maskelenmiş verileri tahmin etme gücünü incelemişlerdir. Yapılan çalışmada modelin, problemin bir kısmının maskelenmesi sonucunda problemi ne kadar doğru şekilde tahmin edebildiği ROUGE değerlendirme metrikleri ile ölçülmüştür. Kullanılan veri seti 1000, 500, 100, 50 ve 10 eğitim örneğinden oluşmaktadır ve bu sayede ön-eğitilmiş PEGASUS dil modelinin küçük veri setleri ile ince ayar yapılması sonunda maskelenmiş kelimeleri tahmin etme konusunda ne kadar başarılı olduğu gözlemlenmiştir. 10 eğitim verisi ile ince ayar yapıldığında elde edilen ROUGE değeri T5 ön eğitimli dil modelinin sonuçlarıyla karşılaştırılmış ve PEGASUS dil modelinin neredeyse iki katı daha başarılı olduğu iddia edilmiştir (Yang vd., 2021).

(Ertam & Aydin, 2022) Ertam ve Aydın yoruma dayalı metin özetleme işlemi için çeşitli haber web sitelerinden son 5 yıla ait haber başlıkları, kısa haberler, haber

içerikleri ve anahtar kelimeleri tarayarak Türkçe haber özetleme için bir veri seti oluşturmuşlardır. Çalışmalarında özetleme işlemi için Seq2seq modeli RNN ve LSTM38 derin öğrenme yapıları ile kullanılmıştır. Çalışmanın performansını değerlendirmek için hassasiyet, duyarlılık ve F1 ölçüm puanları kullanılarak Rouge-1, Rouge-2 ve Rouge-L değerleri karşılaştırılmıştır.

Kemaloğlu Alagöz, bilişim alanında yayınlanmış makaleler ile Türkçe Bilişim Literatür veri seti oluşturarak çıkarıma dayalı metin özetleme çalışması gerçekleştirmiştir. Yapılan çalışmada veri seti ön işleme adımında makale formatına uygun yeni bir ön-işlem fonksiyonu kullanılmış, özetleme adımında Derin İnanç Ağları kullanılmıştır. Veri setinde BERT Çıkarımsal Özetleyici ve Derin İnanç Ağları ile özetleme işlemi gerçekleştirildikten sonra BERTScore ile karşılaştırma yapılmış tez için geliştirilen yöntemin %88 F-Skor değeri ile özetleme işlemini gerçekleştirdiği görülmüştür (Kemaloğlu Alagöz, 2022).

Baykara ve Güngör (2022), yaptıkları çalışmada aglutinatif olarak sınıflandırılan Türkçe ve Macarca dillerinde yoruma dayalı metin özetleme çalışmasına odaklanmaktadır. Bu diller için yazarlar tarafından iki büyük ölçekli haber metinlerinden oluşan veri seti oluşturulmuştur (TR-News, HU-News). Bu diller için ön işleme metodu olarak dillerin morfolojik yapısını dikkate alarak SeperateSuffix ve CombinedSuffix tokenizasyon yöntemleri önerilmiştir. Yayınlamış oldukları veri seti öncelikli olarak metin özetleme için oluşturulmuş olsa da konu sınıflandırma, anahtar kelime çıkarımı ve başlık oluşturma işlemleri için de kullanıma uygundur. SeperateSuffix yöntemi TR-News veri setinde en yüksek ROUGE-1 skoruna ulaşmıştır.

Ertam ve Aydın (2022), çeşitli Türkçe haber web portallarından 5 yıla ait haber metinleri, haber başlıkları, kısa haberler ve anahtar kelimeleri toplayarak yeni bir veri seti oluşturmuşlardır. Hazırlanan veri setinden haber başlıkları ve kısa haber içerikleri kullanılarak Tensorflow kütüphanesi altyapısında (Seq2seq) diziden diziye yaklaşımı ile yoruma dayalı metin özetleme çalışması yapılmıştır. Bu çalışmanın performansını değerlendirmek için Rouge-1, Rouge-2 ve Rouge-L, hassasiyet, duyarlılık ve F1 ölçüm puanları kullanılmış, çalışmanın performans değerleri, her bir cümle için ve ayrıca

rastgele seçilen 50 cümle için sonuçların ortalaması alınarak sunulmuştur. F1 Ölçümü değerleri Rouge-1, Rouge-2 ve Rouge-L için sırasıyla 0,4317, 0,2194 ve 0,4334'tür.

Bal, Türkçe metinlerde çıkarıma dayalı metin özetleme için öznitelik seçme yöntemleri ve geleneksel makine öğrenmesi algoritmalarını kullanarak ön işleme tekniklerinin otomatik metin özetleme üzerindeki etkilerini analiz etmiştir. Yapılan çalışmada özetleme performansını artıran gizli anlamsal analize dayalı yeni bir ayırt edici özellik önerilmiştir. Ön işlem aşamalarından kök bulma işleminin ve durak kelimelerin atılması işleminin yüksek başarım değeri elde etmek için gerekli olduğu sonucuna ulaşılmıştır. Ayrıca derin öğrenme yöntemlerinden LSTM ve BERT modelleri kullanılarak metin özetleme işlemi gerçekleştirilmiştir. LSTM kullanarak yapılan özetleme işleminde oluşturulan özetlerin, referans özetteki kelimelerin ve kelime gruplarının çoğunu içerdiği ancak anlamlı bir bütün içermediği tespit edilmiştir (Bal, 2022).

Baykara ve Güngör (2022) tarafından sık kullanılan iki büyük Türkçe haber veri seti (MLSUM-TR ve TR-News) üzerinde ön-eğitimli diziden diziye (Seq2seq) dil modellerini kullanarak yoruma dayalı metin özetleme çalışması ve başlık üretme çalışması gerçekleştirilmiştir. Diziden diziye dil modelleri kullanılarak oluşturulan başlıkların ve metin özetlerinin başarısı ROUGE değerlendirme metrikleri ile analiz edilmiştir. Analiz sonucunda Tek dilli BERT dil modelinin çok dilli BERT modelinden daha başarılı sonuç verdiği gözlemlenmiştir.

Goyal vd. metin özetleme modellerinde öğrenme dinamiklerini araştırmışlardır. Çeşitli metin veri setlerinde (CNNDM, XSUM, MEDIASUM) ön eğitilmiş dil modellerinin ince ayar yapılması sürecinde hangi parametrelerin tekrar eğitildiği bilgisinin açık olmaması sebebiyle, metin özetlemeye odaklanarak metin üreten dil modellerinin eğitim dinamiklerini analiz etmişlerdir. Eğitim sürecinin farklı aşamalarında elde edilen özetler incelenerek girdiyi ne kadar kopyaladığı tartışılmış, eğitimi zor olan yüksek kayıplara sebep olan tokenizasyon işlemi ve eğitimi hızlı gerçekleşen düşük kayıplı tokenleri göz ardı etmek gibi iki farklı yaklaşımın özet üzerindeki etkisi incelenmiştir (Goyal vd., 2022).

Erdağı (2023) tarafından yapılan çıkarıma dayalı metin özetleme çalışmasında ulusal haber sitelerinde bulunan haber içeriklerinden oluşturduğu bir veri setini kullanılmıştır. Kontrol veri seti oluşturmak amacıyla üç farklı kişinin yapmış olduğu özetler tek bir sonuca indirgenmiş ve özet başarımının değerlendirilmesi ROUGE ve BLEU metrikleri aracılığı ile bu veri seti üzerinden gerçekleştirilmiştir. Literatürdeki cümle derecelendirme yöntemlerinden farklı olarak Türkçe dilindeki büyük ünlü uyumu ve küçük ünlü uyumu kuralları temel alınarak yeni bir hibrit model üretilmiş, geleneksel yöntemlerden daha iyi sonuç verdiği gözlemlenmiştir. (Erdağı, 2023)

Ay vd. (2023) yapmış oldukları yoruma dayalı metin özetleme çalışmasında çevrimiçi Türkçe haber kaynaklarından topladıkları haber metinleri ile oluşturdukları yeni veri setini kullanmışlardır. T5 (text to text transfer transformer) dil modelini kullanarak yaptıkları özetleme sonucunda Rouge-1, Rouge-2, Rouge-L ve Bert-score başarım değerlendirme ölçütleri 0,6913, 0,6623, 0,7528 ve 0,8718 olarak bulunmuştur.

Alıpour, yapmış olduğu çıkarıma dayalı metin özetleme çalışması için 2020 yılında Scialom ve arkadaşları tarafından hazırlanmış olan MLSUM haber veri seti (Scialom vd., 2020) ve kendi oluşturduğu 1010 makaleden oluşan makale veri seti kullanmıştır. Makale veri setinin eğitim için hazırlanması sırasında Alıpour tarafından BERT çıkarıma dayalı özetleme modelini kullanarak veri seti kısaltma işlemi yapılmıştır. mT5 (Multilingual Text-to-Text Transfer Transformer) (Xue vd., 2021) mimarisi kullanarak yapılan çalışmada sistem başarısı Rouge metrikleri ile incelenmiştir. Makale veri seti için ROUGE-1, ROUGE-2 ve ROUGE-L değerleri sırasıyla 18,34, 4,62 ve 17,63 olarak ölçülmüştür. Makale veri seti yazar tarafından oluşturulduğu için daha önce çalışılmış bir değerle karşılaştırma yapılmamıştır. Haber veri seti için elde edilen Rouge metrikleri aynı veri seti üzerinde daha önce yapılmış çalışmalar ile kıyaslandığında başarılı sonuçlar elde edildiği görülmüştür (Alıpour, 2023).

Kara, yapmış olduğu çalışmada Bilgisayar Bilimleri alanındaki Türkçe akademik yayınlardan oluşturduğu veri seti ile akademik yayınların anlamsal bütünlüğünü doğal dil işleme, derin öğrenme ve istatiksel yöntemler ile analiz etmiştir. İncelenen yayınların biçim ve anlam olarak bütünlüğünün sağlanması için gerekli olan ögeler tespit edilmiş, makalelerin özetlerinin dahil edilerek yapılan yoruma dayalı ve çıkarıma dayalı metin özetleme işlemleri kullanılarak referans özetlerinde etiketlerin

tahmin edilme oranı incelenmiş, 492 tane makaleden sadece 68 tanesinde bulunması gereken tüm ögelerin yer aldığı ifade edilmiştir. Makalelerin büyük bir kısmının referans özetinde makalenin tümü incelendiğinde özette bulunması gereken bilgilerin büyük bir kısmının olmadığı sonucuna ulaşılmıştır (Kara, 2023).

Wan ve Bansal, PEGASUS dil modelinin ince-ayarlanması sonucunda halüsinasyon görmesi konusunu ele almış, istenen metnin dışında bilgiler içermesi problemine 3 farklı yaklaşım sunarak FACTPEGASUS (Factuality-Aware Pre-training and Fine-tuning for Abstractive Summarization) dil modelini geliştirmişlerdir. PEGASUS dil modelinin eğitim öncesi cümle seçim stratejisi geliştirilmiş, referans özetlerde bulunan halüsinasyon bilgileri ortadan kaldıran bir düzeltici önerilmiş ve gerçek olmayan özetlerle gerçek olanları ayrıştırmak için karşılaştırmalı öğrenme kullanılmış ve ön-eğitim ile ince-ayarlama arasındaki boşluğu dolduracak bir bağlayıcı önerilerek yeni model geliştirilmiştir. Geliştirilen FACTPEGASUS dil modelinin yoruma dayalı metin özetleme işlemi gerçekleştirirken metindeki gerçek bilgileri daha iyi koruduğu iddia edilmiştir (Wan ve Bansal, 2022).

Sjöblom (2023) tarafından yapılmış olan tez çalışmasında metin özetleme çalışmalarında sıklıkla kullanılan CNN/DailyMAil isimli haber metinlerinden oluşan büyük veri seti kullanılarak BART ve PEGASUS ön-işlenmiş dil modelleri ince-ayarlanarak yapılan yoruma dayalı metin özetlemenin ve metin sınıflandırma ve metin özetlemenin beraber kullanılarak yapılan metin özetlemenin performansa olan etkileri incelenmiştir. PEGASUS ön-eğitilmiş dil modelinin ince-ayarlanması sonucunda metin sınıflandırma ve metin özetlemenin bir arada yapıldığı durumun en başarılı performansı gösterdiği gözlemlenmiştir.

Srivastava vd. tarafından yapılan çalışmada iki büyük veri seti ile çalışılarak (XSUM, CNNDM) yoruma dayalı metin özetlemede BART, PEGASUS, BRIO ön-işlenmiş dil modelleri kullanılarak oluşan özetlerin kalitesi bilgi kapsamı, halüsinasyon varlığı ve özetleme karmaşıklığı açısından değerlendirilmiştir. Yapılan değerlendirme sonucunda ROUGE değerlendirme ölçütünün yoruma dayalı metin özetleme için uygun olmayan bir değerlendirme metodu olduğu, mevcut özetleme dil modellerinin bilgi kapsamı konusunda yeterli olmadıkları ve halüsinasyon gördükleri, model tarafından oluşturulan özetlerin referans özetlere kıyasla az bilgi kapsadığı ve

halüsinasyon gördüğü gerekçesiyle daha iyi ve yeni referans özetlerine ihtiyaç olduğu değerlendirmesi yapılmıştır (Srivastava, 2023).

Alsuhaibani, yapmış olduğu çalışmada haber metinleri ve bunların kısa özetleri bulunan XSUM veri seti üzerinde ince-ayarlanmış PEGASUS dil modeli ile yoruma dayalı metin özetleme işlemi gerçekleştirmiştir. Çalışmada yapılan özet işlemi Arapça farklı kategorilerden oluşan NADA veri seti üzerinde test edilmiş, metinler kategorilerine göre değerlendirilmiş ve her kategorinin özetlenmesi sonucunda elde dilen özetler ROUGE metrikleri ile değerlendirilmiştir. Elde edilen ROUGE başarım ölçütünün haber metninin kategorisine göre farklılık gösterdiği gözlemlenmiştir (Alsuhaibani, 2023).

Mercan vd. LSTM ve ön-eğitilmiş T5, Pegasus, BART, BART-Large dil modellerinin açık kaynak veri setleri üzerinde ince ayarlanması ile elde edilen modeller aracılığı ile 75 adet Türkçe özgeçmişten oluşan veri setini sınıflandırma işlemi yapmışlardır. Yapılan çalışmada özgeçmiş veri setinde en yüksek başarımı ince ayarlanmış BART-Large modelinin gösterdiği bilgisi sunulmuştur (Mercan vd., 2023).

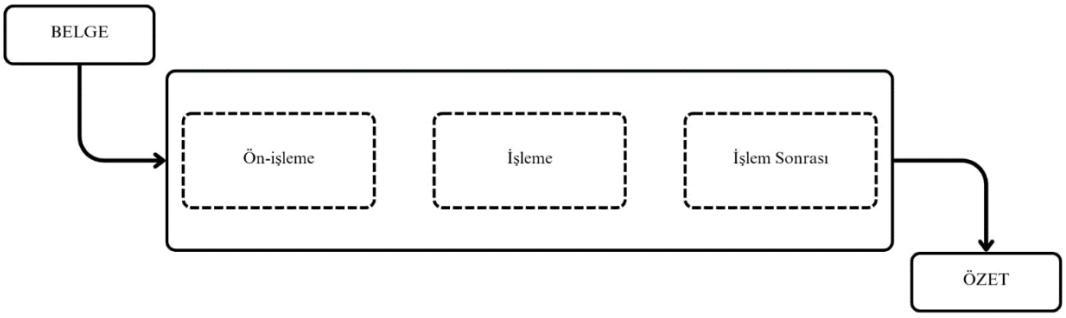
3. OTOMATİK METİN ÖZETLEME YÖNTEM BİLİMİ

İnternet üzerindeki bilgi kaynaklarının artması insanlara bilgi paylaşmak için çok büyük kolaylık sağlamaktadır. Ancak bilgi kaynaklarının çokluğu ve bilgi içeren dokümanların büyüklüğü bir sorunu ortaya çıkarmaktadır. Bilgiye erişim ve istenen bilgiye ulaşabilmek amacıyla doğru dokümanın incelenip incelenmediğinin anlaşılması için tüm dokümanın okunması gerekmektedir. Bunun çok zaman alan bir işlem olması günümüz problemlerindendir. Bu sorunu ortadan kaldırmak amacıyla metinlerin özeti çıkarılmaktadır (Adalı, 2020).

Otomatik metin özetleme ise verilen veri kümesinden en önemli bilgilerin kullanıcıya bilgisayar tarafından sunulması işlemidir (Hovy ve Lin, 1998). Hovy ve Lin'e göre metin özetleme bazı süreçlerden oluşmaktadır:

- Konunun belirlenmesi: burada dokümandaki en önemli noktaların belirlenmesi işlemi yapılır. Cümlelerin içerisindeki en sık tekrarlanan ifadeler, bazı sözcük öbeklerinden (sonuç olarak, özetle gibi) sonra gelen önemli ifadeler belirlenmektedir.
- Yorumlama: Burada önemli olarak belirlenmiş ifadeler anlamsal özelliklerine göre yeniden yorumlanır.
- Üretme: konunun belirlenmesi ve yorumlanması sonrası önemli bilgilerden yeni cümlelerin oluşturulup, birleştirilerek sunulması aşamasıdır (Gündoğdu ve Duru, 2016).

Metin özetleme işlemi sırasında işlemin gerçeklendiği dilin özelliklerine göre bazı zorluklarla karşılaşılmaktadır. Bunlara örnek olarak özetlenmek istenen dokümandaki yazıların kuralsız ve bozuk olması, metni dilimlere ayırma işleminin karmaşıklığı, sözcüklerin niteliğinin belirlenmesi, söz dizimsel belirsizliklerin giderilmesi gösterilebilir. Bu zorluklar, Türkçe dilinin morfolojik yapısının çok iyi analiz edilmesi gerekliliğini ortaya çıkarmıştır (Adalı, 2020).



Şekil 3.1. Metin özetlemenin temel adımları

Şekil 3.1'de görüldüğü üzere metin özetleme öncelikle metnin ön işleme aşaması, metnin özetlendiği işlem aşaması ve özetlenen cümlelerin birleştirilerek özetin oluşturulduğu işlem sonrası aşaması olarak 3 aşamalı bir yaklaşım gerektirir. Ön işleme adımı metnin bölümlenmesi, tokenlere ayrılması, özetleme işlemi için metne açıklamaların eklenmesini içerir. Özetlenmek istenen dokümanlar genellikle yapılandırılmamış düz metinlerdir ve özetleme işlemi için bu metinlerin standartlaştırılarak özetlemeye hazırlanması gerekmektedir. Hazırlanan bu bölümler veya tokenlerin özetleme işleminde bilgisayar tarafından okunabilir olması gerekmektedir. Özetleyici için bu birimler vektör uzayları, matrisler ağırlıklar veya grafikler gibi bilgisayarın işlem yapabileceği formatlara dönüştürülür. Özetlenmesi istenen metnin temsili oluşturulduktan sonra bu temsil özetleyici algoritma tarafından belirli özelliklere dayanarak işlenir. Daha sonra bu işlenmiş birimlerin özete dahil edilecek parçaları seçilir ve kaynak metindeki cümleler orijinal haliyle alınıp birleştirilerek veya anlamsal olarak işlenip yeni cümleler oluşturularak özet metin oluşturulur (Birant, 2015).

3.1. Metin Ön İşleme

Metin özetleme işlemi girdisi düz metin belgesidir. Çıktısı ise girdi metninin anlamsal olarak önemli kısımlarını barındıran daha kısa ve anlamlı cümlelerden oluşan metindir. Bu özetin oluşturulabilmesi için girdi olan metnin işlenebilecek hale getirilmesi gerekmektedir. Literatür incelendiğinde OMÖ çalışmaları için girdi metnin bazı ön işlemlere tabii tutulduğu gözlemlenmektedir. Bu ön işleme adımlarından bazıları aşağıda verilmiştir.

3.1.1. Metni bölümleme

Metin özetlemenin ön işlem aşaması, metnin bölümlere ayrılması ile başlamaktadır. Bu bölümler metnin uzunluğuna ve türüne göre değişmektedir. Kitap bölümleri, ana başlıklar, alt başlıklar, paragraflar, cümleler, kelimeler gibi çok uzun bölümler olabileceği gibi doküman çok kısa bölümlere de ayrılabilir. Bu bölümlere ayırma işlemi noktalama işaretlerinden faydalanarak yapılabilir ancak farklı dillerin farklı noktalama işaretleri kuralları bulunmaktadır (Woods, 2006; Birant, 2015). Cümle bölümlemesi belirlenen her cümlenin sonuna kaynak metinden ayırt edilebilen bir sembol eklenmesi ile yapılabilmektedir (Torres-Moreno, 2014).

3.1.2. Tokenler

Tokenler, metin içerisinde tek bir birim olarak ifade edilebilen yapılardır. Metin özetleme işlemi yapılırken özetleyicinin tokenleri tanıması ve tüm metni tokenlere ayırabilmesi önemlidir. Metnin benzersiz küçük yapıtaşı olan tokenler özetleyicinin yapısına uygun olacak şekilde sözcük, kök sözcük, sözcük ekleri vb. olabilir.

3.1.2.1. Tokenizasyon

Tokenizasyon, kaynak metni tokenlere ayırma işlemidir. Noktalama işaretleri dikkate alınarak yapıldığında bazı dillerde yanlış ayrıştırmalara sebep olabilmektedir. Türkçe metinlerde cümleleri kelimelere ayırarak yapılan tokenizasyon işlemi metnin analizinde kolaylık sağlamaktadır (Manning vd., 2008; Eryiğit, 2007).

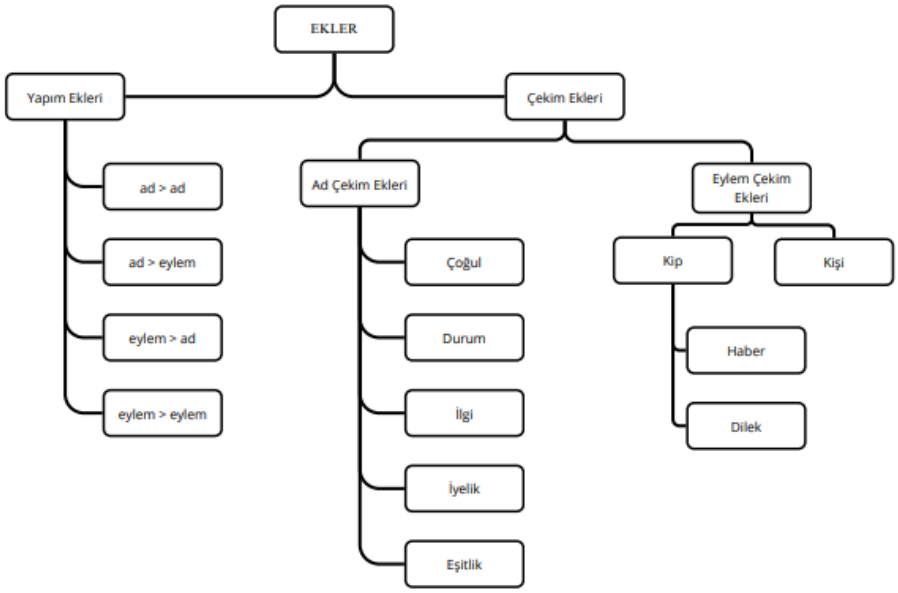
3.1.2.2. Durak kelimelerin tespiti

Durak kelimeleri metinlerde sıklıkla geçen ancak metnin analizinde anlamsal olarak hiç değere sahip olmayan veya çok az değere sahip olan kelimelerdir. Türkçe için “ve”, “ile”, “ama” gibi kelimeler durak kelimelere örnek gösterilebilir. Metinden durak kelimelerin tespit edilip çıkarılması metnin anlamsal analizi açısından daha önemi yüksek kelimelere odaklanmayı kolaylaştırmaktadır (Torres-Moreno, 2014; Kemaloğlu, 2022). Durak kelimelere dahil edilmiş bazı kelimeler metnin bütünü

incelendiğinde anlamsal değere sahip olabilir bu sebeple durak kelime seçimi yapılırken metnin anlamsal içeriğine dikkat edilmelidir.

3.1.3. Kökleme ve lemmatize etme

Kökleme, metindeki kelimelerin almış olduğu ekler kaldırılarak kelimenin kökünü bulma işlemidir. Eklemeli bir dil olması sebebi ile Türkçede sözcükler kökten sonra ek almaktadır ve alabileceği eklerin öngörülebilir bir sınırı yoktur. Türkçe sözcük türetme açısından oldukça üretken bir dildir. Eklerin çok sayıda olması ve eklendikleri sözcüğün sınıfına göre (ad-soylu veya eylem-soylu) sözcüğe çok farklı anlam katabilmeleri sebebi ile Türkçenin biçim bilimsel çözümlemesi oldukça karmaşıktır (Adalı, 2020). Türkçedeki ekler Şekil 3.2 de gösterilmiştir.



Şekil 3.2 Türkçede ekler (Adalı, 2020)

Şekil 3.2'de görüldüğü gibi Türkçede ekler yapım ekleri ve çekim ekleri olmak üzere iki temel kategoride incelenir. Çekim ekleri isim soylu kelimeler için farklı, eylem soylu kelimeler için farklı işlevlere sahiptir. Yapım ekleri ise isimden isim, isimden eylem, eylemden isim ve eylemden eylem yapan eklerdir. İsimden eylem veya eylemden isim yapan yapım ekleri sözcüklerden çıkarılırsa sözcüğün cümledeki anlamı değişecektir. Türkçede sözcüklerin tüm eklerinin temizlenmesi işlemi yani sözcüğün kökünü tespit etme işlemi metnin anlam bütünlüğünü bozmadan analizi

açısından sakıncalı olabilmektedir. Çizelge 3.1'de eylem soylu köke yapım eki eklenerek türetilen adlar görülmektedir.

Çizelge 3.1. Eylem soylu köke yapım eki eklenerek türetilen adlar (Adalı, 2020)

EK	KÖK	SÖZCÜK	EK	KÖK	SÖZCÜK
-a	doğmak	doğa	-ç	ayırmak	ayırcak
	dizmek	dize		çekmek	çekiç
-acak	içmek	içecek	-dı	uymak	uydu
	almak	alacak	-ge	bilmek	bilge
-alga	çizmek	çizelge		dizmek	dizge
-ağan	durmak	durağan	-gaç	süzmek	süzgeç
	olmak	olağan		yüzmek	yüzgeç
-amak	basmak	basamak	-gan	almak	alıngan
	kaçmak	kaçamak		kaymak	kaygan
-an	bakmak	bakan	-gı	çalmak	çalgı
	yaratmak	yaratan		dolmak	dolgu
-anak	gelmek	gelenek	-gın	bilmek	bilgin
	olmak	olanak		yanmak	yangın
-cık	öpmek	öpücük	-gıt	örmek	örgüt
	gülmek	gülücük	-ı	açmak	açı
-ıcı	akmak	akıcı		anmak	anı
	yakmak	yakıcı	-mık	kıymak	kıymık
-k	kaçmak	kaçak		kusmak	kusmuk
	konmak	konak	-nc	basmak	basınç
-l	çatmak	çatal		gülmek	gülünç
	kurmak	kural	-nti	dökmek	döküntü
-m	anlamak	anlam		söylemek	söylenti
	toplamak	toplam	-r	akmak	akar
				dönmek	döner
-ma	asmak	asma	-sak	tutmak	tutsak
	dolmak	dolma		savmak	savsak
-maca	bilmek	bilmece	-ş	açmak	açış
	çekmek	çekmece		satmak	satış
-maç	bulamak	bulamaç	-t	anmak	anıt
	dönmek	dönemeç		kanmak	kanıt
-mak	çakmak	çakmak	-tı	bağlanmak	bağlantı
	ekmek	ekmek		sallanmak	sallantı
-man	danışmak	danışman	-maz	çıkmak	çıkmaz
	göçmek	göçmen		bitmek	bitmez
-mar	yağmak	yağmur			

Çizelge 3.1 de görüldüğü üzere gülmek kelimesinin kökü gül kelimesidir ve -cık yapım ekini aldığında gülücük sözcüğü oluşmaktadır. Gülücük sözcüğü gülmek sözcüğü ile aynı anlama gelmemektedir. Metin özetleme yapılırken gülücük sözcüğü gül veya gülmek olarak temsil edilirse metnin anlam bütünlüğü bozulacaktır.

"Çocuk *gülücükler* saçıyordu." (3.1)
"Çocuk *gül* saçıyordu." (3.2)

Yukarıdaki cümleye (3.1) bakılacak olursa gülücükler kelimesinden tüm ekler çıkarıldığında kelime kökü gül kalacaktır. "gülücükler" kelimesi yerine "gül" konularak cümle yeniden kurulduğunda (3.2) çocuğun güldüğü anlamına gelen gülücük saçıyor olması değil çiçek ismi olan gül saçıyor olduğu anlamı çıkarılmaktadır.

Lemmatize etme, kelimelerin sözlük formunu bulmak için yapılan işlemdir. Türkçenin eklemeli bir dil olması sebebiyle kelimeleri lemmatize etme oldukça karmaşık ancak bir o kadar da önemlidir. (3.1) bu kapsamda tekrar incelendiğinde "gülücük" kelimesi köke indirgenmeyip anlamlı haline getirildiğinde yine "gülücük" olarak kalacaktır ve lemmatizasyon sonrası cümlenin anlam bütünlüğü bozulmayacaktır.

3.2. Metin Özetleme Türleri

Torres-Moreno (2014), özetleme türleri için giriş belge sayısına göre, giriş dili sayısına göre, giriş metni türü sınırlamasına göre, hedef kitleye göre, özetleyicinin türüne göre, metnin içeriğine göre, çıktının işlevine göre olmak üzere kapsamlı bir sınıflandırma yapmıştır. Torres-Moreno' ya göre dokümanlarda otomatik özetleme işlemi pek çok kazanç sağlamaktadır:

- Özetleme işlemi okuma süresini azaltır.
- Dokuman araştırmasında özetlerin varlığı seçim aşamasını kolaylaştırır.
- Dokümanların sınıflandırılmasını kolaylaştırır.
- OMÖ işlemi, insan muhakemesi ile yapılan özetlerden daha tarafsızdır.

- Belirli bir amaca yönelik yapılan özetleme işlemi soruların cevaplanmasında daha faydalıdır.
- OMÖ sayesinde daha fazla dokuman incelenebilmektedir.

Otomatik metin özetleme geçmişten günümüze farklı yöntem ve tekniklerle gerçekleştirilmiştir. Bu özetleme yöntemleri tez kapsamında, çıktı tipine göre çıkarıma dayalı metin özetleme ve yoruma dayalı metin özetleme; amaca göre belirtici ve bilgilendirici özetleme; girdi tipine göre tekli ve çoklu belge özetleme; sonuç içeriğine göre alana yönelik ve konuya yönelik metin özetleme; yönteme göre denetimli ve denetimsiz metin özetleme şeklinde sınıflandırılarak incelenmiştir.

3.2.1. Çıktı tipine göre metin özetleme

3.2.1.1. Çıkarıma dayalı metin özetleme

Çıkarıma dayalı metin özetleme işlemine cümle seçerek metin özetleme de denilmektedir. Metnin cümleleri anlamsal önemine göre sıralanır ve önemli olarak değerlendirilen cümleler seçilerek metnin bir alt kümesi oluşturulmuş olur. Bu tip metin özetlemede cümle yapısı değiştirilmez. Önemli cümleler frekans ve benzerlik gibi özellikler kullanılarak istatiksel algoritmalar yardımı ile tespit edilir. Seçilen cümleler birleştirilerek metin özeti oluşturulmaktadır (Alwandawybek vd., 2023).

Çıkarıma dayalı metin özetleme, semantik analizden bağımsız olması sebebi ile kullanılan dilin morfolojik özellikleri ile ilgilenmez ve özetleme işlemi dilden bağımsız olarak gerçekleştirilir. Özetleme işlemi genellikle kolay ve başarılıdır ancak çok uzun metinlerde kullanımı çok iyi sonuçlar vermeyebilir (Alwandawybek vd., 2023).

3.2.1.2. Yoruma dayalı metin özetleme

Yoruma dayalı metin özetleme, özetlenmesi istenen dokümanın tamamından anlamsal olarak bütün ve akıcı bir yeni kısa metin oluşturulmasıdır. Bu özetleme yönteminde sistem tarafından yeni cümleler oluşturulur. İnsan muhakemesi ile özetlemeye en

yakın özetleme biçimidir ve özetin ortaya çıkarılması bilgisayar için oldukça karmaşık bir görevdir. (Munot ve Govilkar, 2014).

Bu tip metin özetlemede bilgi sıkıştırma oranı yüksek olabileceği için uzun metinler çok kısa özetlere dönüştürülebilmektedir ancak özetleme işlemi çıkarıma dayalı metin özetlemeye göre daha maliyetlidir. Metnin anlamını kaybetmeden yeni cümleler ile özetlenmesi kullanılan dilin morfolojik yapısı ile yakından ilişkilidir, Türkçe gibi eklemeli dillerde bilgisayara anlamlı cümleler ürettirmek karmaşık bir görevdir (Alwandawybek vd., 2023; Adalı, 2020).

3.2.2. Amaca göre metin özetleme

3.2.2.1. Belirtici özetleme

Belirtici özetleme çoklu dokümanlardan bilgi çıkarımı amacıyla kullanılmaktadır. Özetleme işlemi sonrası dokümandaki tüm önemli bilgiler çıkarılmaz ancak uygun dokümanın seçimine katkı sağlar. Tüm dokümanlarda ilgilenilen konunun geçtiği cümleler taranır ve konu adının en çok tekrar ettiği cümleler seçilerek bir özet oluşturulur. Metnin içeriği hakkında detaylı bilgi vermeden dokümanın ne ile ilgili olduğu bilgisinin çıkarılmasını sağlar. Böylece farklı dokümanlar arasındaki farklılıkları özet olarak sunabilir. Dokümanların ana fikrini özet olarak sunarak dokümanın okumaya değer olup olmadığına hızlı karar verebilmek amacıyla kullanılabilir ve böylece zaman kazandırır (Klavans vd., 2001).

3.2.2.2. Bilgilendirici özetleme

Bilgilendirici özetleme çoklu dokümanlardan detaylı şekilde bilgilerin çıkarılarak özet oluşturulması işlemidir. Kaynak dokümanların içerdiği ortak bilgilerin sentezi ile özet elde edilmektedir ve kullanıcının geniş bilgi ihtiyacını karşılamaya yöneliktir. Belirtici özetlemeden daha uzun özetler elde edilir ve bilimsel çalışmalarda kullanımı uygundur (Klavans vd., 2001).

3.2.3. Girdi tipine göre metin özetleme

Otomatik özetleme işlemi, girdi tipine göre tek belgeli özetleme ve çok belgeli özetleme olarak iki sınıfta incelenebilir. Tek belgeli özetleme girdi olarak tek bir doküman alır ve çıktı olarak bunun özetini sunar. Çok belgeli özetleme ise birden fazla dokümanı girdi olarak kabul etmektedir. Dokümanların her birinden özet metinler çıkarıldıktan sonra bu özetler birleştirilir ve yeni bir doküman elde edilir daha sonra bu doküman özetlenerek tüm dokümanların özeti oluşturulmuş olur. Çok belgeli metin özetleme aynı konuda yazılmış birden fazla dokümandan bilgi çıkarmak için kullanılmaktadır (Mihalcea ve Tarau, 2005).

3.2.4. Yönteme göre metin özetleme

3.2.4.1. Denetimli özetleme

Bu tür özetlemede denetimli makine öğrenmesi teknikleri kullanılmaktadır. Anahtar kelime çıkarımı, cümle benzerliği tespiti gibi yöntemler aracılığı ile metin özetlemesi gerçekleştirilmektedir. Denetimli özetleme için kullanılacak algoritmanın eğitilmesi gerekmektedir ve eğitim için büyük veri setine ihtiyaç duyulmaktadır. Anahtar kelime çıkarımı için örnek verilecek olursa, eğitim için kullanılacak veri setinde metin içerisindeki anahtar kelimelerin bulunması gerekmektedir. Dokümandaki her bir cümle anahtar kelime için araştırılarak sınıflandırma işlemi gerçekleştirilir. Belirli bir alana özgü veri seti ile eğitim gerçekleştirilerek konuya özel özetleme yöntemi geliştirilebilir. Etiketlenmiş verilere ihtiyaç duyulması sebebi ile eğitim veri setinin hazırlanması maliyetli bir süreç olabilmektedir (Manesh, 2020).

3.2.4.2. Denetimsiz özetleme

Denetimsiz makine öğrenmesi yöntemleri ile gerçekleştirilen özetleme işlemidir. Etiketlenmiş eğitim verisi gerektirmemesi sebebi ile gerçekleştirmesi denetimli özetlemeye kıyasla daha kolaydır ancak denetimli metin özetlemeden daha karmaşık matematiksel işlemler gerektiren görevleri gerçekleştirmeyi sağlar. Denetimsiz özetleme için oluşturulan modeller kategorize edilmemiş metinlerden bilinmeyen paternleri bulmak konusunda başarılıdır (Manesh, 2020).

3.3. Dil Modeli Geliştirme

DDİ çalışmalarında metin özetleme, metinden bilgi çıkarma, metinle sohbet etme, metinden anahtar kelime çıkarma gibi metin madenciliği işlemlerinde bilgisayarın metin üzerinde işlem yapabilmesi için kelimelerin sayısal şekilde ifade edilmesi gerekmektedir. Metnin sayısal temsilini oluşturabilmek için kullanılan farklı yöntemler bulunmaktadır. Bunlardan Kelime Çantaları Modeli (Bag of Words/ BoW), TF-IDF Modeli (Term Frequency – Inverse Document Frequency), N-Gram Modeli incelenmiştir.

3.3.1. Kelime torbaları modeli

Kelime torbaları modeli metinlerin sayısal temsilinde kullanılan bir yöntemdir. Bu modelde metni oluşturan kelimelerin sıklıkları vektörle temsil edilmektedir. Her kelime bir sütun ile, her bir cümle satır ile ifade edilir. Cümle içerisindeki bu kelimenin sıklığı ise kelimenin bulunduğu sütuna yerleştirilir. Bu şekilde metindeki kelime dağılımını ve sıklıklarını temsil eden bir matris oluşturulmuş olur.

Kelime torbaları modeli metin içerisinde bulunan kelimelerin öncelik sırasını dikkate almadan yalnızca metindeki geçme sıklığına göre oluşturulan bir temsildir. Matriste kelimelerin kaç kez kullanıldığı bilgisine erişilebilir ancak kelimelerin her biri için aynı değerler kullanılır. Kelimeler arasında değer farkı tespit edilemez. Bu sebeple cümlenin anlamsal analizi yerine metin içerisindeki anahtar kelime çıkarımı, duygu analizi gibi sınıflandırma problemleri için daha kullanışlıdır (Zhang vd., 2010).

3.3.2. TF-IDF modeli

TF-IDF modeli iki temel kavramı ele almaktadır: terim frekansı (TF) ve ters belge frekansı (IDF). Terim frekansı bir kelimenin belirli bir cümledeki frekansını ifade ederken ters belge frekansı kelimenin tüm belge içerisindeki önemini temsil etmektedir. TF-IDF puanı bu iki temsilin çarpılması sonucunda hesaplanmaktadır.

Bir kelimenin TF-IDF değeri, belirli bir cümle içindeki sıklığı fazla olduğunda ancak tüm belgede nadir görüldüğünde yüksek hesaplanır. TF-IDF modeli belirli bir konuda

önemli ancak tüm belgelerde sıklıkla kullanılmayan kelimeleri tespit etmek için kullanılmaktadır.

TF-IDF değerlerinin hesaplaması şu şekilde yapılmaktadır:

$$TF(t,d) = \frac{\text{Terim t'nin belge } d \text{ 'de görünme sayısı}}{\text{Belge } d \text{ 'deki toplam terim sayısı}} \quad (3.3)$$

Denklem (3.3)'te *t* ile ifade edilen değer kelimeyi, *d* ile ifade edilen değer ise belgeyi temsil etmektedir.

$$IDF(t) = \log\left(\frac{\text{Toplam belge sayısı}}{\text{Terimi içeren belge sayısı}}\right) \quad (3.4)$$

Denklem (3.4)'te IDF hesaplanması gösterilmiştir. TF-IDF değeri denklem (3.3) ve denklem (3.4)'ün çarpımı sonucunda hesaplanır.

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (3.5)$$

Hesaplanan TF-IDF modeli bir kelimenin cümle içindeki frekansını dikkate aldığı gibi aynı zamanda bu kelimenin belgelerdeki önemini de dikkate alarak kelimelerin önem derecesini belirlemekte faydalıdır (Gümüş, 2019).

3.3.3. N-Gram model

N-gram modeller, doğal dil işlemede istatiksel yaklaşımla belirli dizilimlerin olasılıklarının incelendiği durumlarda kullanılmaktadır. N-gram modeli iki veya daha fazla kelimenin yan yana görülme sıklığını hesaplamaktadır. Belge içerisindeki sıklıkla yan yana görülen kelimeler, kelime grupları adaylarıdır. N-gram modeller çoğunlukla 1-gram, 2-gram, 3-gram şeklinde kullanılmaktadır. Modelin en büyük problemi kelime grupları içerisinde metnin anlamsal analizine katkı sağlamayan bazı kelimelerin (ve, veya, şey, bu, vb.) sıklıkla görülmesidir. Bunu önlemek için sıklıkla karşılaşılan kelim grupları belirlenirken sözcük türü filtreleme gibi işlemler uygulanmaktadır (Kutlugün ve Şirin, 2018).

3.4. Geçmişten Günümüze Otomatik Metin Özetleme Yöntemleri

Otomatik metin özetleme, 1958'den bu yana bilgisayar bilimcilerin araştırma konusu olmuştur (Luhn, 1958). Bilgisayar bilimi, bilişsel bilimler, dil bilimi, istatistik, yapay zeka, doğal dil işleme gibi pek çok alanın uzmanlığından yararlanan bir araştırma alanıdır. Geçmişten günümüze pek çok farklı yöntem ile otomatik metin özetleme çalışmaları gerçekleştirilmektedir. Bunlardan literatürde sıkça karşılaşılan yöntemler bu bölümde incelenmiştir.

3.4.1. Gizli anlamsal analiz (LSA- Latent semantic analysis)

Gizli Anlamsal Analiz (GAA) metindeki cümleler arasındaki benzerliği istatiksel olarak inceleyerek kelimeleri anlamlandıran bir yöntemdir. Bir cümlede hangi kelimelerin kullanıldığı bilgisini ve cümleler arasındaki ortak kullanılan kelimelerin bilgisini tutarak cümleler arasındaki ortak kelimeleri inceler ve hangi cümlelerin anlamsal olarak ilişkili olduğunu tespit etmeye çalışır.

GAA anlamsal olarak benzer kelimeleri veya cümleleri tespit etmek için kelimeler veya cümleler arasındaki ilişkiyi modelleyen Tekil Değer Ayrışımı (Singular Value Decomposition - SVD) yöntemini kullanmaktadır. GAA tabanlı özetleme yöntemleri 3 temel adımda çalışmaktadır.

1. Giriş matrisinin oluşturulması: Özetlenmesi istenen metni temsil eden seyrek bir matris oluşturulur. Hücreler, kelimelerin cümlelerdeki önemini temsil etmek için farklı yaklaşımlar kullanılarak doldurulmaktadır.
2. Tekil değer ayrışımı: kelimeler ve cümlelerin arasındaki ilişkiyi matematiksel olarak modelleyen bir yöntemdir.
3. Cümle seçimi: tekil değer ayrışımının sonuçlarını kullanarak farklı algoritmalarla giriş matrisinden cümle seçme işlemidir.

GAA yönteminin uygulanmasında giriş matrisinin oluşturulması özetleme işlemini doğrudan etkileyeceği için matrisin oluşturulmasında kullanılan farklı yöntemler bulunmaktadır. Giriş matrisinin satır boyutunu azaltmak için durak kelimelerinin çıkarılması veya kelimelerin köklerini bulma gibi işlemler kullanılmaktadır. Ayrıca

giriş matrisinin hücre değerleri doldurulurken TF-IDF yöntemi, kelimenin kök tipinin isim soylu olup olmamasına göre doldurma yöntemi, logaritmik entropi yöntemi, kelimenin ikili sayı temsili ile doldurma yöntemi vb. kullanılmaktadır (Özsoy vd., 2010).

3.4.2. Çizge teorisi tabanlı algoritmalar

Çizge teorisi farklı nesneler arasındaki ilişkilerin matematiksel bir gösterimidir. Çizge teorisi temel olarak bir problemin kenar ve düğümlerle modellenmesi ve bu modelin çizge şeklinde ifade edilmesine dayanmaktadır (Şeker, 2015).

3.4.2.1. PageRank algoritması

PageRank, Google'ın arama sonuçlarını sıralamak amacıyla kullanmış olduğu bir sıralama algoritmasıdır. Algoritma Google kurucuları tarafından web sitelerinin önemini tahmin ederek sıralamak amacıyla geliştirilmiştir. Bir sayfasının önemini tahmin etmek için sayfaya giden bağlantıların sayısını ve kalitesini hesaplar (Page vd., 1998). Denklem (3.6)'da A sayfasının PageRank değerinin formülü gösterilmektedir.

$$\mathrm{PR}\,(A) = (1-d) + d(\,\mathrm{PR(T1)}\,/\mathbf{C}(\mathbf{T}1) + \cdots + \,\mathrm{PR(Tn)}\,/\mathbf{C}(\,\mathrm{Tn}\,)) \tag{3.6}$$

Denklem (3.6)' da PR (A), A sayfasının PageRank değerini temsil etmektedir. PR(T1) A sayfasına bağlantı kuran T1 sayfalarını, C(T1) T1 sayfasından diğer sayfalara verilen bağlantıları, d ise 0-1 arasında değişen bir damping parametresini ifade etmektedir. PageRank'ın yinelemeli bir algoritma olması sebebi ile başlangıçta tüm sayfaların değeri aynıdır ve yineleme sonuçlarına göre sayfaların değerleri belirlenmektedir (Gümüş, 2019).

3.4.2.2. TextRank algoritması

TextRank algoritması Mihalcea vd. (2004) tarafından PageRank algoritmasından geliştirilmiş çizge tabanlı bir algoritmadır. PageRank algoritmasında olduğu gibi sayfalar arasındaki referans ilişkisi sayfanın önemini belirlemek için kullanılmaktadır.

Önemli olan sayfaların önemli olan diğer sayfalardan bağlantı aldığı ve bir sayfanın PageRank değerinin o sayfanın ziyaret edilme olasılığını temsil ettiği mantığına dayanmaktadır. Sayfalar yerine cümleler üzerinde çalışıldığında TextRank ile metin özetleme işlemi geçekleştirilebilmektedir. İki cümle arasındaki benzerlik PageRank algoritmasındaki sayfalar arasındaki bağlantı gibi düşünülürse cümleler arası benzerlik puanları bir matriste tutulabilir. Metnin özeti bu matristeki benzerlik değerlerine göre sıralama yapılarak oluşturulabilir (Gümüş, 2019). TextRank algoritması ile metin özetleme adımları aşağıdaki gibidir.

- Metin ön işlemeden geçirilir ve cümlelere bölünür.
- Cümleler sayısal temsillerine dönüştürülür.
- Cümle vektörlerindeki benzerlik değerleri benzerlik matrisine yazılır.
- Benzerlik matrisi çizgeye dönüştürülür ve düğümler cümleleri, kenarlar ise benzerlik değerlerini ifade eder.
- Çizge üzerindeki benzerlik değerlerine göre özette kullanılacak cümleler seçilir.

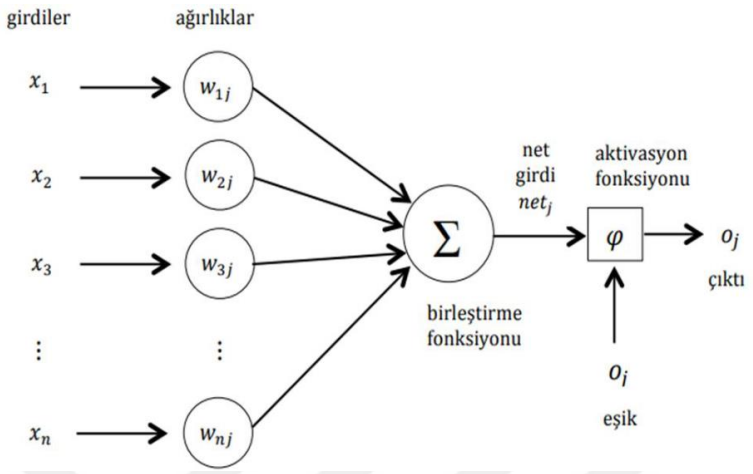
3.4.2.3. LexRank algoritması

LexRank algoritması PageRank algoritmasına dayanan çizge tabanlı bir algoritmadır. Çıkarıma dayalı metin özetlemek amacıyla kullanılmaktadır. Özet için cümle seçme prensibine dayanır. Cümlelerin önemi, metnin diğer birçok cümleye benzer olması ile belirlenmektedir. Çizge olarak ifade edilen algoritmada düğümler cümleleri, kenarlar ise benzerliği temsil eder. Kenardaki ağırlık değeri kosinüs benzerliği ile hesaplanmaktadır. Bu sayede birbirine benzer olan cümleler çizge üzerinde yakın konumlandırılacaktır. LexRank algoritması merkezilik temelli bir yaklaşımla cümle seçimini gerçekleştirmektedir. TextRank algoritmasından temel farkı kenarlara ağırlık atamak için cümlelerin merkezilik değerini hesaplamasıdır (Gümüş, 2019).

3.4.3. Derin öğrenme yaklaşımlı metin özetleme uygulamaları

Derin öğrenme, yapay sinir ağları temelli karmaşık matematiksel modeller aracılığıyla veri temsillerini öğrenen makine öğrenmesinin alt dalıdır. Yapay sinir ağları çeşitli

çıkarımsal problemlere çözüm bulmak için insan beyninin bilgiyi işleme şeklinden ilham alınarak geliştirilmiş bilgisayara bu bilgi işleme özelliklerin aktarımını amaçlamaktadır. Yapay sinir hücresinin yapısı Şekil 3.3'de gösterilmiştir.



Şekil 3.3. Yapay sinir hücresinin yapısı (Uğur, 2006)

Yapay sinir hücresinde farklı ağırlık değerlerine sahip girdiler birleştirme fonksiyonu aracılığı ile toplanarak belirli bir eşik değerine sahip aktivasyon fonksiyonundan geçirilmekte ve çıkış değeri üretilerek diğer bir yapay sinir hücresine aktarılmaktadır. Derin sinir ağları yapay sinir ağlarının çok katmanlı ve çok nöronlu durumu olarak ifade edilebilir. Kullanılan aktivasyon fonksiyonu oluşan bilgiyi özniteliklerine göre filtreleyerek öğrenme işlemine önemli katkı sağlamaktadır. Yapay sinir ağlarından en çok kullanılan aktivasyon fonksiyonları aşağıda sıralanmıştır (Kara, 2023).

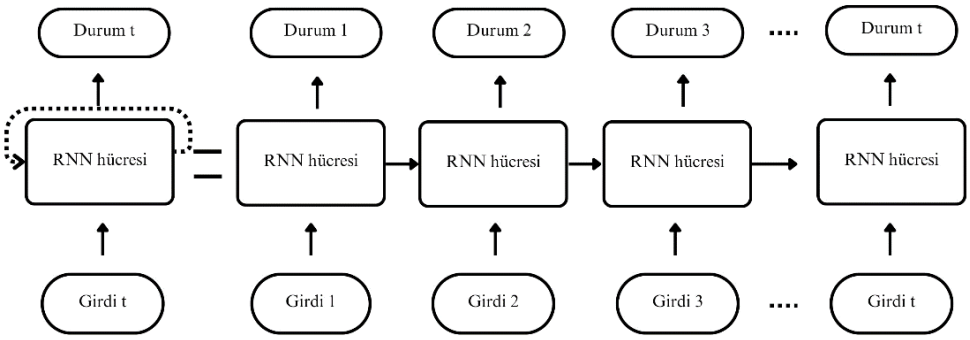
- Basamak aktivasyon fonksiyonu
- Doğrusal aktivasyon fonksiyonu
- Sigmoig aktivasyon fonksiyonu
- Hiberbolik tanjant aktivasyon fonksiyonu
- ReLU aktivasyon fonksiyonu
- Sızıntı ReLU aktivasyon fonksiyonu
- SoftPlus aktivasyon fonksiyonu
- Swish aktivasyon fonksiyonu

Transformer mimarisi, özellikle doğal dil işleme ve makine çevirisi gibi görevlerde başarıyla kullanılan bir derin öğrenme modeli mimarisidir (Vaswani vd., 2017). Transformer mimarisini ve neden büyük dil modellerinde başarılı olduğunu anlayabilmek için öncelikle kodlayıcı-kod çözücü (Encoder-Decoder) mekanizmasının incelenmesi gerekmektedir. Transformer mimarisi oluşturulmadan önce doğal dil işleme çalışmaları için LSTM gibi RNN tabanlı mimariler kullanılmaktadır.

3.4.3.1. RNN (Recurrent Neural Network) mimarisi

Tekrarlayan sinir ağları olarak da bilinen RNN'ler gizli durumlara sahip olan, önceki çıktıların girdi olarak kullanılmasına izin veren sinir ağları modelidir. Girdilerin uzunluğunun sınırı olmaması sebebi ile doğal dil işleme çalışmalarında sıklıkla kullanılmaktadır.

Tekrarlayan sinir ağlarında hesaplama sürecinde döngüler bulunmaktadır. Her bir döngüde zamansal bir gecikme bulunur ve bu gecikme sayesinde RNN hücreleri kendi çıktılarını tekrar girdi olarak alabilmektedirler. Bu sayede model durum bilgisini tutarak hafızalı olma özelliğine sahiptir.



Şekil 3.4. RNN mimari yapısı

RNN mimarisinin yapısı şekil 3.4 'te gösterilmiştir. Şekilden anlaşılacağı üzere ilk girdi birinci RNN hücresine gelir ve bir çıktı verir daha sonra bu elde edilen çıktı ile ikinci girdi beraber ikinci RNN hücresine girdi olarak gelir bu şekilde her hücreye bir önceki hücrenin çıktısı da girdi olarak alınmış olur. Buna gizli durum (hidden state)

denilmektedir. Her bir hücrede bir önceki hücrenin çıktısı da kullanıldığı için ağda bir hafıza oluşmaktadır. RNN mimarileri, zaman serileri, ses işleme gibi doğal dil işleme görevlerinde sıklıkla kullanılmaktadır. RNN mimarileri ile metin üretimi de yapılabilmektedir (Onan, 2022). Çizelge 3.2'de RNN mimarisinin avantajları ve dezavantajları verilmiştir.

Çizelge 3.2. RNN mimarisinin avantajları ve dezavantajları

Avantajlar	Dezavantajlar
Girdi olarak herhangi bir uzunlukta veri kabul eder	Hesaplama yavaştır
Girdi büyüklüğü model boyutunu etkilemez	Uzun zaman önce elde edilen bilgiye erişim zordur
Hesaplama yaparken geçmiş bilgileri dikkate alır	Mevcut durum için gelecekteki herhangi bir girdi düşünülemez

3.4.3.2. LSTM (Uzun Kısa Vadeli Bellek) mimarisi

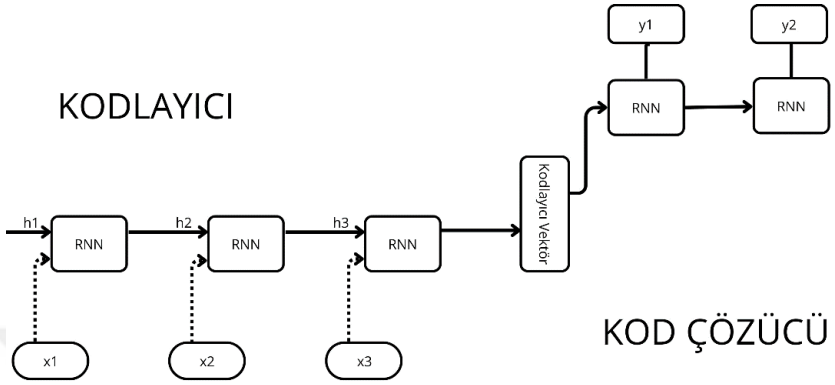
1997 yılında Hochreiter ve Schmidhuber tarafından tanıtılan çok derin eğitmesi çok zor olan çok fazla yineleme barındıran modellere çözüm getiren LSTM mimarisi doğal dil işleme alanında sıklıkla kullanılmaktadır. Özellikle zaman serileri ve doğal dil işleme gibi sıralı verilerle işlem yapılan alanlar için geliştirilen LSTM çok fazla yinelemeyi engelleyen kapılara sahiptir.

LSTM mimarisinde hücre hafızası, giriş kapısı, unutma kapısı ve çıkış kapısı bulunmaktadır. Bu birimler aracılığı ile LSTM'de önemli bilgiler uzun süreli olarak saklanmakta, gereksiz bilgiler ise unutma ve çıkış kapıları aracılığı ile hafızadan silinmektedir (Kara, 2023).

Unutma kapısı, önceki gizli durumdan gelen bilgiyi ve yeni girdiyi sigmoid fonksiyonundan geçirerek hangi bilginin sonraki hücreye aktarılacağına karar verir. Giriş kapısı hücre durumunu güncellemektedir. Önceki gizli durumdan gelen bilgi ve yeni bilgiyi hem sigmoid hem tanh fonksiyonlarından geçirerek bilginin ne kadar önemli olduğuna karar verir. Unutma kapısı hücredeki bilgilerin ne kadarının unutulacağını ve önceki bilgilerin ne kadarının korunacağını belirlemektedir.

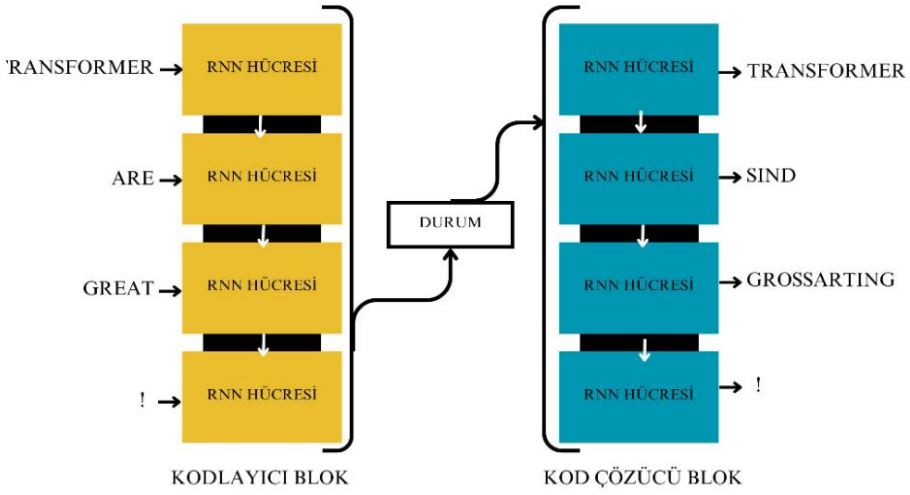
Çıkış kapısı ise hücreden çıkacak olan bilgiyi düzenlemektedir. LSTM kapıları ağırlıklı bir şekilde öğrenilir ve eğitim sırasında optimize edilir (Pattanayak, 2023).

3.4.3.3. Kodlayıcı-kod çözücü mimarisi (Encoder-Decoder)



Şekil 3.5. Kodlayıcı-kod çözücü yapısı

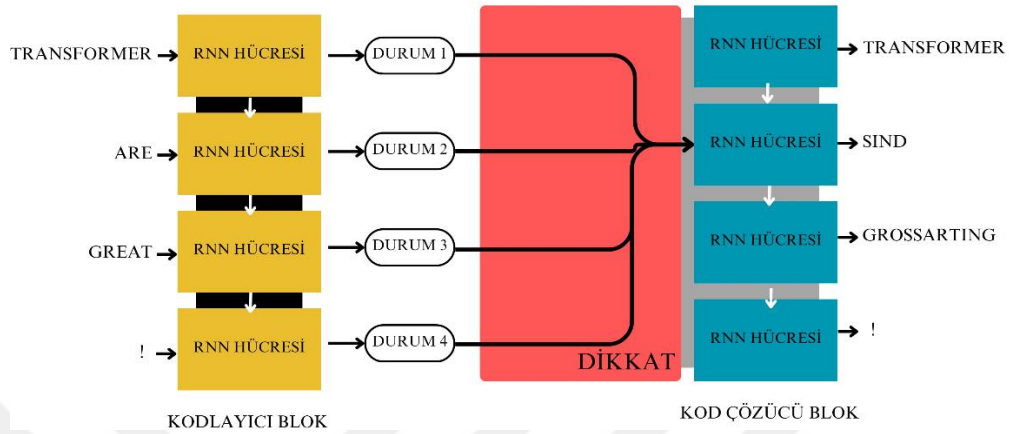
Kodlayıcı ve kod çözücünün kullanıldığı işlemlerde modele bir dizi verilir ve Şekil 3.5'te de görüldüğü üzere çıktı olarak yine bir dizi elde edilmektedir. Kodlayıcı girdi dizilerini sayısal temsillere çevirmektedir. Bu veriler kod çözücü tarafından çıktı dizisi olarak üretilmektedir.



Şekil 3.6. Kodlayıcı-kod çözücü dilden dile çeviri örneği (Tunstall vd., 2022)

Şekil 3.6'da da görüldüğü gibi her bir kelime ayrı ayrı işlenir ve encoder tarafından her kelime sayısal temsillerine çevrilir. Tüm kelimeler işlendikten sonra veriler

decoder blok tarafından çözümlenir ve bir çıktı dizisi üretilir. Bu mimarinin problemi olarak görülebilecek kısım Encoder tarafındaki son gizli state üzerinde bir darboğaz oluşmasıdır. Bu problemin çözümü için dikkat mekanizması geliştirilmiştir.

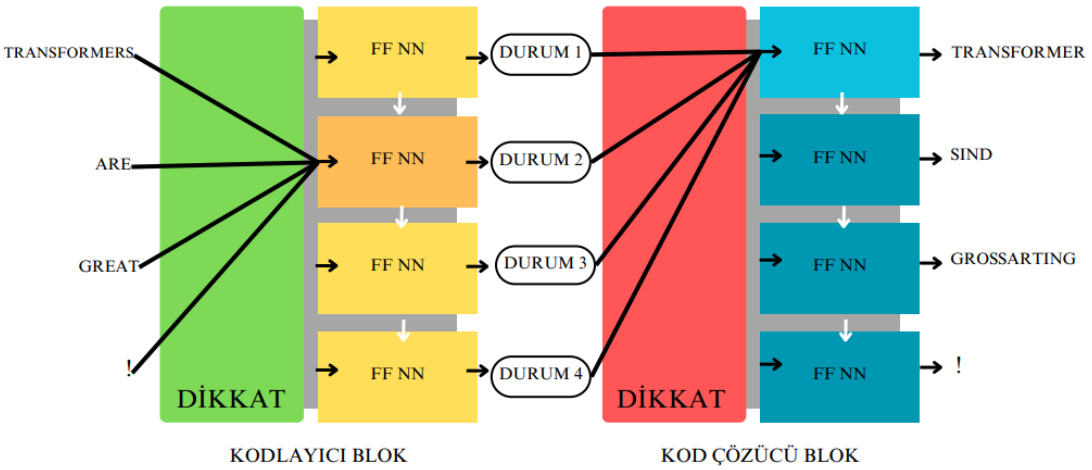


Şekil 3.7. Dikkat mekanizması eklenmiş kodlayıcı-kod çözücü (Tunstall vd., 2022)

Dikkat mekanizması her bir RNN hücresi için gizli durum üretmektedir. Bu girdi durumlarına decoder aynı anda erişmektedir. Fakat aynı zaman aralığındaki tüm durumlar decoder için büyük bir girdi oluşturabilmektedir. Bu sebeple hangi durumların daha önemli olduğunun belirlenmesi ihtiyacı oluşmaktadır. Burada dikkat mekanizması her bir gizli durum için önemine göre ağırlık atama işlemi yapmaktadır. Böylece decoder hangi gizli duruma öncelik vereceği bilgisine sahip olmuş olur (Seo vd., 2016; Ay, 2023).

Dilden dile çeviri yaparken dikkat mekanizması işlemin doğru ve hızlı yapılmasına katkı sağlamaktadır (Şekil 3.7). Ancak metin çevirisi işlem adımları incelenirse önce bir dizi oluşturulur daha sonra kelimeler tek tek işlenir dikkat mekanizması ile önceliklendirilir ve çeviri yapılarak yeni bir dizi oluşturulur ve çıktı oluşmuş olur. Ancak büyük bir metin söz konusu olduğunda işlemlerin sırayla yapılıyor olmasının süreci yavaşlatacağı görülebilmektedir. Bu diziden diziye modelinde çeviri işlemleri paralel şekilde yapılamamaktadır (Ay, 2023).

3.4.3.4. Transformer mimarisi



Şekil 3.8. Transformer Mimari Yapısı (Tunstall vd., 2022)

Transformer mimarisinde ise birçok paralel dikkat başlığı (multi-head attention), katman normalizasyonu ve tam bağlantılı besleme ileri sinir ağı (Fully Connected Feedforward Neural Network- FC-FFNN) içerir. Sinir ağları katmanlardan oluşmaktadır ve dikkat mekanizması her bir katmandaki gizli durumlar ile çalışmaktadır. Hem kodlayıcı tarafında hem kod-çözücü tarafında dikkat mekanizması bulunmaktadır (Şekil 3.8). Transformer mimarisinde RNN'lerin kullanıldığı tekrarlı sinir ağları yerine ileri beslemeli sinir ağları kullanılmaktadır (Alsuhaibani, 2023; Onan, 2022).

Çok büyük bilgisayarlar ile çok büyük veriler işlenerek Transformer mimarisi sıfırdan eğitilmiştir. Ancak birçok doğal dil işleme uygulaması için pratikte etiketli veri bulmak oldukça zordur ve büyük miktarlardaki metin verilerini işleyebilmek için çok güçlü bilgisayarlar gerekmektedir. Bu işlemin maliyet sorununu çözmek için transfer öğrenme (transfer learning) geliştirilmiştir (Alsuhaibani,2023).

3.4.3.5. Transfer öğrenme

Transfer öğrenme kısaca daha önce eğitilmiş büyük bir dil modelini kendi projemize adapte etme işlemidir. Eğitilmiş bir büyük dil modelinin (fine-tune) ince ayarlaması

ile sıfırdan model eğitmeden bu dil modelleri kullanılabilmektedir. İnce ayarlama işlemi ile büyük dil modeli yapılmak istenen göreve özgü olarak özelleştirilebilir.
Transfer Öğrenme işleminin dil modellerinde kullanımı için ULMFIT kütüphanesi geliştirilmiştir. Bu kütüphane transfer öğrenme işlemini 3 aşamada uygulamıştır (Aydoğan ve Karci, 2019).

Ön eğitim (Pretraining):
Bu adımda gelecek kelimenin tahmini önceki kelimelere dayanarak yapılmaktadır. Bu işleme dil modelleme ismi verilmiştir. Bu işlem etiketli veri ihtiyacını ortadan kaldırmıştır. İnternet üzerindeki metin verileri kullanılarak ön eğitim işlemi yapılabilmektedir.

Adaptasyon:
Dil modeli büyük ölçekli bir corpus üzerinde eğitildikten sonra spesifik bir alana adapte edilir. Örneğin dil modeli Wikipedia corpusu üzerinde eğitildikten sonra film yorumlama için IMBD corpusuna adapte edilebilmektedir. Bu aşamada da dil modelleme kullanılır ancak model hedef corpustaki gelecek kelimeleri tahmin etmektedir.

İnce Ayarlama (Fine-tuning):
Dil modeli belirli bir görev için sınıflandırma katmanı ile ince ayarlama işlemi yapılmaktadır. İnce ayarlama işleminde dil modelinin tüm parametreleri yeniden eğitilmez, küçük bir kısmı yeni veri seti üzerinde eğitilir ve bu sayede var olan dil modeli özel bir amaç için ince ayarlanmış olur.

Hugging Face tarafından geliştirilen Transformers kütüphanesi Pytorch TensorFlow ve Jax frameworklerini destekleyerek büyük dil modelleri için bir standart getirilmiştir. Tranformer mimarisi ile çalışmak için bu üç framework ünde kullanılabildiği Transformers kütüphanesi göreve özgü mimariler de sunmaktadır. Metin sınıflandırma, metin özetleme, soru cevaplama, çeviri, metin üretme özellik tanıma gibi işlemler için bu mimariler üzerinde ince ayar yapmak yeterli olmuştur.

3.4.3.6. PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) dil modeli

Literatür incelendiğinde büyük metin derlemleri üzerinde Transformer mimarisi kullanılarak ön eğitilmiş büyük dil modellerinin metin özetleme de dahil olmak üzere pek çok doğal dil işleme görevinde geleneksel yöntemlere kıyasla oldukça başarılı olduğu görülmektedir. Pegasus dil modeli Transformer mimarisi tabanlı yoruma dayalı metin özetleme için özel olarak oluşturulmuş bir dil modelidir. Kodlayıcı- kod çözücü mimari yapısını temel alır. GSG (Gap sentence generation) boşluk cümleleri oluşturma yapısını tanıtan dil modeli uzun metinlerin özetlenmesinde başarılı bulunmuştur (Zhang vd., 2020).

GSG, maskelemeli dil modellerinden esinlenilerek yoruma dayalı metin özetleme amacıyla üretilmiştir. Maskelemeli dil modelleri çıkarıma dayalı özetlemede çok yüksek başarım göstermektedir. Maskelemeli dil modellerinde cümle içerisindeki kelimeler seçilerek maskelenir ve modelin bu kelimeleri tahmin etmesi beklenir. Maskelemeli dil modelleri kelimeleri tahmin ettiğinde cümleyi kopyalamış olur ve bu çıkarımsal metin özetleme için çok başarılıdır. Ancak yoruma dayalı metin özetlemede modelin metni anlaması ve yeni cümleler üretmesi beklenmektedir. Bu sebeple GSG metindeki cümlelerden kelimeleri maskelemez, metindeki cümleleri maskeleyerek cümle oluşturması beklenir (Zhang vd., 2020).

3.5. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Değerlendirme Ölçütleri

Rouge değerlendirme ölçütleri bir özeti otomatik olarak değerlendirmek için referans özetle karşılaştırma yaparak oluşturulan özetin kalitesini belirlemektedir. Bu ölçütler bilgisayar tarafından otomatik oluşturulmuş özet ile referans özet arasındaki n-gram kelime dizileri gibi örtüşen kelime gruplarını sayar (Lin,2004).

ROUGE-n bilgisayar tarafından oluşturulmuş özet ile referans özet arasındaki n-gram örtüşme oranını ifade etmektedir. n 1'den başlayan pozitif tam sayıları temsil etmektedir. Oluşturulan özet ile referans özet arasındaki n kadar kelimenin benzerliğini ölçer.

ROUGE-1 unigram benzerliğini ifade ederken ROUGE-2 bigram benzerliğini ifade etmektedir. Yani ROUGE-1 ve ROUGE-2 değerleri sırasıyla bir kelime ve 2 kelime için benzerliği denetler.

ROUGE-L (ROUGE-Longest Common Subsequence), bilgisayar tarafından oluşturulan özet ile referans özet arasındaki en uzun ortak alt dizi benzerliğini ölçmektedir. Bu ölçüm metriği kelime sırasının benzerliğini değerlendirmektedir (Lin, 2004).

3.6. Uygulamanın Amacı

Literatür incelendiğinde Türkçe dilindeki otomatik metin özetleme çalışmalarının sayısının karmaşık biçim-bilimsel özellikleri sebebiyle az olduğu gözlemlenmiştir. Türkçe metinler üzerinde gerçekleştirilen özetleme çalışmaları çoğunlukla çıkarıma dayalı özetleme şeklindedir. Türkçe metinlerde özetleme işlemi için büyük dil modellerinin kullanımının az olması sebebiyle Türkçe büyük veri seti sayısı da azdır. Var olan Türkçe büyük veri setleri haber metinleri ve bunların özetlerinden veya Wikipedia Türkçe metinlerinden oluşmaktadır. Büyük dil modellerinin özetleme işlemi amacıyla eğitilebilmesi için metinler ve özetlerinden oluşan veri setlerine ihtiyaç vardır. Özellikle uzun metinler için özetleme çalışması yapılabilecek büyük bir Türkçe veri setine literatürde rastlanmamıştır.

Uzun dokümanlar okuması zaman alan ve özetlerine en çok ihtiyaç duyulan metinlerdir. Özellikle yapılan güncel çalışmaları takip etmek isteyen kişilerin bilimsel makale okumaya oldukça fazla zaman ayırması gerekmektedir. Zamanın daha verimli kullanılabilmesi için uzun metinlerin otomatik özetlenmesi gereklilik halini almıştır. Bu sebeple bu tez kapsamında uzun metinlerden oluşan bilimsel makaleler ile veri seti oluşturulmuş ve bu veri seti kullanılarak yoruma dayalı metin özetlemede yüksek başarım gösteren ön işlenmiş büyük dil modellerinden PEGASUS un ince ayarlanması gerçekleştirilmiş ve yoruma dayalı metin özetleme çalışması gerçekleştirilmiştir.

3.7. Kullanılan Kütüphaneler ve Geliştirme Ortamı

Çalışmanın kodlama kısmı Python programlama dili kullanılarak çeşitli Python kütüphaneleri (Pytorch, Nltk, Pandas, TensorFlow, Keras) aracılığı ile yapılmıştır.

Tez için gerçekleştirilen özetleme işlemleri Google Colab geliştirme ortamında yapılmıştır. Google Colab, Google tarafından sunulan başka geliştiriciler ile çalışma ve paylaşma imkanı sunan bulut tabanlı bir Jupyter notebook servisidir. Bu geliştirme ortamının sağlamış olduğu en faydalı özellik yüksek performans gerektiren işlemlerin gerçekleştirildiği projelerde yüksek işlem gücüne sahip GPU'lar ile çalışmaya olanak sağlamasıdır. Şekil 3.9' da Google Colab çalışma ortamından örnek bir görsel gösterilmektedir.

Şekil 3.9. Google Colab çalışma ortamı

Hugging Face, özellikle doğal dil işleme (NLP) alanında kullanılan açık kaynaklı yapay zeka ve makine öğrenimi modellerini barındıran bir platformdur. Hugging Face Transformers kütüphanesi, önceden eğitilmiş dil modellerinin paylaşılabildiği bir bulut ortamı, tokenlere ayırma işleminde kullanılan tokenizer modelleri ve doğal dil işleme alanında bilgi veri tabanı gibi hizmetler sunmaktadır. Hugging Face'in sunduğu Transformers kütüphanesi, popüler NLP modellerinin önceden eğitilmiş parametrelerini içerir ve kullanıcıların özel görevler için bu modelleri ince ayarlama yapmalarına olanak tanır. Bu çalışmada Transformers kütüphanesi kullanılmıştır.

4. ARAŞTIRMA BULGULARI VE TARTIŞMA

4.1. Veri Seti

Tez çalışması için eğitim bilimleri, bilişim teknolojileri, iktisat ve işletme olmak üzere 3 farklı alanda yazılmış, Dergipark üzerindeki dergilerden toplamda 200 adet tam metin Türkçe bilimsel makaleler rastgele seçilerek toplanmıştır. Veri setinde hangi bilim alanından kaç adet makale seçildiği bilgisi Çizelge 4.1'de gösterilmektedir.

Çizelge 4.1. Veri seti makale sayıları

	Eğitim Bilimleri	Bilişim Teknolojileri	İktisat ve İşletme
Makale sayısı	89	69	42

Veri seti oluşturulurken özetleme modelinin farklı bilim alanlarına özgü terimleri öğrenebilmesi için özellikle farklı alanlarda yazılmış bilimsel makaleler seçilmiştir.

Elektronik dergilerin çoğu gibi Dergipark makalelerin PDF (Portable Document Format) halinde indirilmesine izin vermektedir. Veri setinin oluşturulabilmesi için bilimsel makaleler düz metin biçimine dönüştürülmüştür. Daha sonra metin özetleme işlemlerinin uygulanacağı veri setine gerekli bilgiler kaydedilmiştir.

Veri setinin içerdiği bilgiler aşağıda listelenmiştir.

- Metin içeriği
- Yazar tarafından oluşturulmuş makale özeti
- Makalenin başlığı
- Yazar tarafından belirlenmiş anahtar kelimeler
- Bilim alanı

Tez kapsamında gerçekleştirilen yoruma dayalı metin özetleme işlemi için veri setindeki metin içeriği ve yazar tarafından oluşturulmuş olan özet kullanılmıştır. Veri setindeki diğer bilgiler bu çalışmada kullanılmamıştır. Ancak Türkçe bilimsel makaleler ile ilgili metin özetleme çalışmalarının az olması sebebiyle bilimsel

makalelerden oluşan bir veri seti daha önce paylaşılmamıştır. Tez kapsamında oluşturulan veri seti ileride farklı doğal dil işleme teknikleri için de kullanılabilecektir.

Veri seti oluşturulurken dergiye ait bilgiler, yazara ait bilgiler, çizelgeler, şekiller, formüller, kaynaklar ve dip notlar veri setine dahil edilmemiştir. Amacın metin özetleme olması sebebi ile düz metin içeriği üzerinde işlemler gerçekleştirilmiştir. Veri setinden noktalama işaretleri, durak kelimeleri ve kelimelerin ekleri özellikle çıkarılmamış, yoruma dayalı metin özetleme işlemi yapılacağı için dil modelinin noktalama işaretleri ile eğitilerek kurallı yeni cümleler üretebilmesi sağlanmaya çalışılmıştır. Veri seti train, test ve validation olarak üçe ayrılmıştır ve veri setinin train için ayrılan bölümünde 160, validation için ayrılan bölümünde 20, test için ayrılan bölümünde 20 adet makale bulunmaktadır.

Veri setindeki metin ve özet verilerinin uzunluğu hakkında fikir vermesi amacıyla Hugging Face Hub ortamında paylaşılmış büyük bir dil modelinin Türkçe bir veri seti üzerinde ön eğitiminde kullanılan bir tokenizer ile Türkçe makale veri setinde tokenlere ayırma işlemi gerçekleştirilmiştir. Çizelge 4.2'de oluşturulmuş üç alt kümenin metin ve özet verilerinin uzunluğu token cinsinden verilmiştir.

Çizelge 4.2. Veri setindeki metinlerin token sayıları

	En kısa üç metnin token sayıları	**En kısa üç metnin özetlerinin token sayıları**	**En uzun üç metnin token sayıları**	**En uzun üç metnin özetlerinin token sayıları**
Train	297, 222, 215	94, 97, 105	734, 770, 844	3958, 3686, 3562
Test	255, 253, 199	97, 101, 115	1360, 1405, 1517	3840, 3513, 3475
Valid	212, 210, 207	126, 141, 145	1331, 1456, 1602	3544, 3499, 3460

Tokenlere ayırma işlemlerinin görülebilmesi için veri setinden rastgele seçilen (4.1)'de gösterilen referans cümlenin tokenlere ayırma işlemi gerçekleştirilmiştir.

"Öğretmenler verilen ifadelerinden de görüldüğü gibi istenmeyen öğrenci davranışlarına karşı iş birliği içeren stratejiler kullandıklarını ifade etmişlerdir." (4.1)
Dil modeli metin özetleme işlemi yapmadan önce metni tokenlere ayırır ve sonrasında bu tokenleri sayısal temsillere dönüştürerek hafızada tutar. (4.2) 'de bu tokenlerin sayısal temsilleri gösterilmiştir.

[18060, 296, 2299, 3546, 3723, 345, 12151, 576, 17028, 3601, 3880, 886, 850, 483, 7532, 4332, 32509, 528, 4476, 1409, 18810, 18] (4.2)

['ÃκÄŁretmen', 'ler', 'Ġverilen', 'Ġifad', 'elerinden', 'Ġde', 'ĠgÃ¶rÃ¼ldÃ¼ÄŁÃ¼', 'Ġgibi', 'Ġistenmeyen', 'ĠÃ¶ÄŁrenci', 'ĠdavranÄ±ÅŁ', 'larÄ±na', 'ĠkarÅŁÄ±', 'ĠiÅŁ', 'ĠbirliÄŁi', 'ĠiÃ§eren', 'Ġstratejiler', 'Ġkullan', 'dÄ±klarÄ±nÄ±', 'Ġifade', 'ĠetmiÅŁlerdir', '.'] (4.3)

(4.3)’ te (4.2)’deki sayısal temsillerin sırasıyla cümledeki karşılıkları verilmiştir. (4.2)’de görülen ilk eleman olan “18060”, (4.3)’te “ÃκÄŁretmen” elemanını temsil etmektedir. Bu temsiller (4.1)’de verilen cümledeki “Öğretmen” karakterlerine karşılık gelmektedir. Bu gösterimde Türkçe karakterler tanımlı olmadığı için farklı karakterler ile ifade edilmiştir. (4.3) incelendiğinde ana cümledeki (4.1) bazı kelimelerin daha küçük parçalara ayrıldığı ve bu şekilde tokenler oluşturulduğu gözlemlenmektedir. Örneğin “ifadelerinden” kelimesi “Ġifad” ve “elerinden” olacak şekilde Türkçede anlamlı olmayan iki alt parçaya bölünmüştür. Aynı dil için oluşturulmuş ön eğitimli bir tokenizerın, farklı veri setinde kullanıldığında yeterli olmayabileceği görülmektedir. Tez kapsamında metin özetleme sonucunda modelin metni yorumlayarak kendi cümlelerini oluşturması istendiği için eğitim sırasında tokenlerin anlamlı kelimeler halinde kullanılması hedeflenmiştir. Bu sebeple oluşturulmuş veri seti kullanılarak yeni bir tokenizer üretilmiştir.

Tokenizer oluşturma işlemi Hugging Face’in sunduğu Transformers kütüphanesinin sağlamış olduğu hizmetlerdendir. Bir veri seti üzerinde sıfırdan tokenizer üretilebileceği gibi ön eğitilmiş bir tokenizer farklı bir veri setinde tekrar eğitilebilir. Bu çalışmada tokenizer eğitilirken tüm parametrelerin değiştirilmesi istenmediği için Türkçe dilinde ön eğitilmiş tokenizer Türkçe makale veri setinde tekrar eğitilmiştir.

Tez çalışması için oluşturulmuş tokenizer ile tokenlere ayırma işlemi gerçekleştirildiğinde (4.1)’ de verilmiş olan referans cümle (4.4)’teki gibi tokenlere ayrılmıştır.

[9685, 1857, 6066, 426, 2853, 618, 3194, 1168, 8592, 703, 465, 3200, 1968, 4955, 7817, 778, 2414, 14] (4.4)

['ÃĸÄŁretmenler', 'Ġverilen', 'Ġifadelerinden', 'Ġde', 'ĠgÃ¶rÃ¼ldÃ¼ÄŁÃ¼', 'Ġgibi', 'Ġistenmeyen', 'ĠÃ¶ÄŁrenci', 'ĠdavranÄ±ÅŁlarÄ±na', 'ĠkarÅŁÄ±', 'ĠiÅŁ', 'ĠbirliÄŁi', 'ĠiÃ§eren', 'Ġstratejiler', 'ĠkullandÄ±klarÄ±nÄ±', 'Ġifade', 'ĠetmiÅŁlerdir', '.'] (4.5)

Tokenlere ayırtma işleminde veri setine özgü tokenizer kullanıldığında aynı cümledeki token sayısının düştüğü (4.4) incelendiğinde görülmektedir. (4.5) te tokenlerin kelime karşılığı gösterilmektedir. (4.5) incelendiğinde tokenlerin kelimelerin, ek içeren, bütün halinden oluştuğu gözlemlenmiştir. Token sayısının çok olması, sözcüklerin eklerine ayrılarak tokenleştirilmesinin iyi bir yöntem olduğu gibi düşünülebilir ancak yapılmak istenen çalışma, dil modelinin kurallı ve anlamlı cümleler üreterek metin özetlemesi olması sebebiyle modelin kelimeleri ekleri ile öğrenmesi gerekli bir durumdur. Veri setine özgü oluşturulan tokenizer cümleyi kelimelerine ayırmıştır.

4.2. Yoruma Dayalı Otomatik Metin Özetleme

Bu tez çalışmasında, Hugging Face Hub ortamında Google tarafından paylaşılmış olan açık erişimli büyük dil modeli Pegasus-x-large kullanılmıştır. Uzun metinlerde yoruma dayalı metin özetleme işlemi için oluşturulmuş olan Türkçe makale veri setinde ince ayarlama işlemi gerçekleştirilmiştir.

Pegasus-x-large ön eğitilmiş dil modeli, Zhang vd. (2020) tarafından İngilizce veri setleri üzerinde eğitilmiştir. Bu modelin Türkçe metin özetleme yapabilmesi için iki yöntem uygulanabilir. Bunlardan birisi dil modelini büyük bir Türkçe veri seti kullanarak baştan eğitmektir. Ancak bu işlem büyük bir dil modeli ve büyük bir veri seti ile işlem yapmayı gerektirdiği için kullanılacak donanımın oldukça güçlü olması gerekmektedir. Yapılabilecek diğer bir işlem ise Türkçe veri seti üzerinde ince ayarlama işleminin gerçekleştirilmesidir. Bu tez çalışmasında, dil modelini baştan eğitebilecek yeterlilikte bir donanıma sahip olunmaması sebebi ile ön eğitilmiş dil modelinin, tez için oluşturulmuş olan Türkçe makale veri seti üzerinde ince ayarlaması gerçekleştirilmiştir. Yapılan çalışmanın başarısı NLTK kütüphanesinin sunmuş

olduğu ROUGE değerlendirme metrikleriyle incelenmiştir ancak ROUGE metrikleri referans özet (yazar tarafından oluşturulmuş özet) ile modelin oluşturduğu özeti n-gram örtüşme oranı ile karşılaştırır. Bu çalışmada yoruma dayalı metin özetleme gerçekleştirildiği için n-gram örtüşme oranı doğru bir değerlendirme göstermemektedir. Bu sebeple tez çalışması kapsamında gerçekleştirilen özetin başarısının incelenebilmesi için oluşturulmuş Türkçe makale veri seti içinden rastgele bir makale seçilmiştir. Şekil 4.1'de seçilmiş olan makalenin referans özeti verilmiştir. Şekil 4.2'de Pegasus dil modelinin, seçilen örnek makalenin metninden ince ayarlama öncesinde çıkarmış olduğu özet sunulmuştur.

Referans Özet

"Klasik duygu analizi yöntemlerinden farklı olarak hedef tabanlı duygu analizi birden fazla kategorinin olduğu karmaşık yapıdaki çevrimiçi tüketici geribildirimlerini değerlendirmede daha başarılı bir performans ortaya koyabilmektedir. Nitekim bir platformda yer alan tüketici geri bildirimleri bir ürüne ilişkin birden farklı hedefe atfedilebilmektedir ve standart duygu analizleri bu geribildirimleri analiz etmede yetersiz kalmaktadır. Literatürdeki gelişmeler gözden geçirildiğinde HDTA çalışmalarının duygu analizine odaklanan diğer çalışmalar içinde oldukça popüler olduğu anlaşılmaktadır. SemEval ABSA-2016 yarışmasında HTDA için 8 farklı dilde veri setleri yayınlanmış ve ekipler duygu analizi için yarışmışlardır. Yarışmada hedef terim kategori ve duygu sınıfı tespit etmek gibi farklı alt görevler bulunmaktadır. Bu alt görevlerin içindekilerden biri hedef terimin tespit edilmesidir. Türkçe dili için HTDA çalışmaları oldukça sınırlıdır. Farklı diller ve farklı kelime temsil yöntemleri kullanan çalışmalar vardır. SemEval Absa 2016 yarışması Türkçe veri seti için kelime temsil yöntemlerinin etkisini inceleyen çalışma bulunmamaktadır. Bu çalışma müşteri yorumlarındaki hedef terimlerin tespitinde farklı kelime temsil yöntemlerinin başarısının incelenmesi amacıyla gerçekleştirilmiştir. Word2Vec Glove ve Fasttext kelime temsil yöntemleri analiz kapsamında incelenmiş ve hedef terimi en başarılı tespit edebilen yöntemin Fasttext kelime temsil yöntemi olduğu görülmüştür. Çalışmada ayrıca F-1 sınıflandırma ölçütü açısından %77 başarı oranı ile Türkçe veri seti için literatürdeki en yüksek sınıflandırma başarısı elde edilmiştir."

Şekil 4.1. Referans özet

İnce Ayarlama İşlemi Öncesi PEGASUS Modelinin Oluşturmuş Olduğu Özet

"Bugn Uber yldz Amazon yldz yldz TripAdvisor yldz yldz yldz firmalar ilgili alanda faaliyet gsteren iletme ve satclara ilikin yorumlarla srasyla 100 milyon yldz 300 milyon yldz yldz ve 450 milyon yldz yldz yldz civar kullancnn youn iletiim ve iliki kurmasna imkan tanmaktadr."

Şekil 4.2. İnce ayarlama işlemi öncesi PEGASUS modelinin oluşturmuş olduğu özet

Şekil 4.2 incelendiğinde ince ayarlama işlemi öncesi PEGASUS dil modelinin oluşturduğu özetin referans özetten oldukça farklı olduğu görülmektedir. Burada dil modelinin henüz Türkçe bir veri seti ile herhangi bir eğitiminin gerçekleşmemesi sebebi ile dil modeli Türkçe karakterleri tanımamaktadır. Dil modeli özetlemesi beklenen metni doğru şekilde anlamlandıramamış, uygun şekilde eğitilmemesi sebebi ile makale içerisinde geçen bir cümleyi özet olarak göstermiştir. İnce ayarlama işlemi öncesi PEGASUS dil modelinin oluşturduğu özetin ROUGE skor değerleri Çizelge 4.3'te verilmiştir.

Çizelge 4.3. İnce ayarlama işlemi öncesi oluşturulan özetin ROUGE skor değerleri

ROUGE-1	ROUGE-2	ROUGE-L
0.2357681165992403	0.023730660848292374	0.15518827795810117

Çizelge 4.3 incelendiğinde ROUGE değerlerinin literatürdeki diğer dillerde gerçekleştirilmiş çalışmalara göre düşük olduğu görülmektedir. Ancak ROUGE değerlendirme metriğinin referans özet ve bilgisayarın oluşturduğu özeti n-gram yöntemleri ile kıyaslaması sebebiyle, uzun bir metinden yoruma dayalı özetleme işlemi sonrası yüksek değerler beklenmemektedir. Bu tez kapsamında ROUGE

değerlendirme metrikleri, yapılan işlemlerin birbirleriyle kıyaslanması amacıyla kullanılmaktadır.

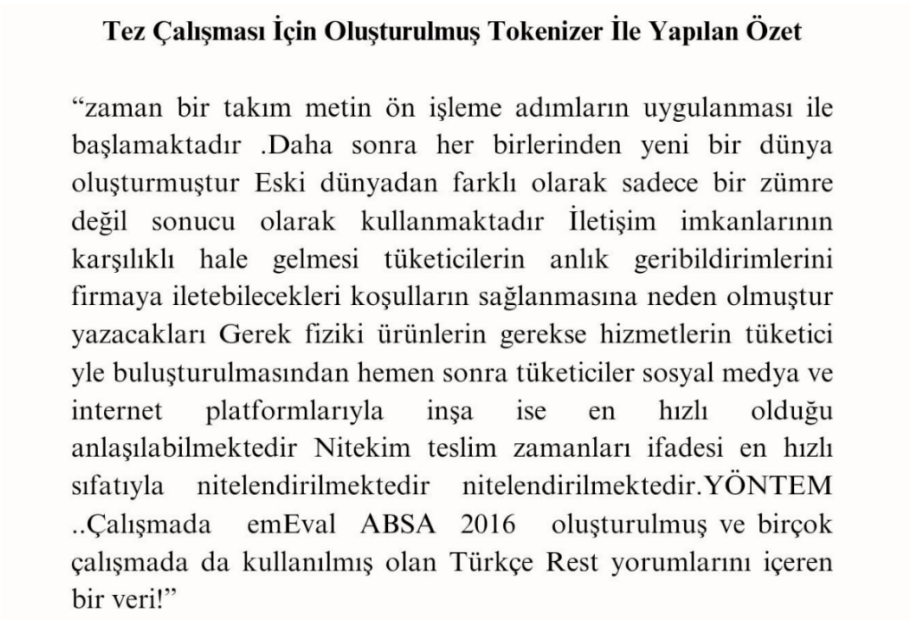
Hugging Face Hub büyük dil modellerini, büyük veri setlerini barındırdığı gibi bu dil modellerinin eğitiminde kullanılmış tokenizer modellerini de barındırmaktadır.

İnce ayarlama sonrası PEGASUS dil modelinin kendi tokenizer modeli kullanılarak gerçekleştirilen metin özetleme işlemi sonrası değerlendirme sonuçlarının ROUGE skor değerleri Çizelge 4.4'te verilmiştir.

Çizelge 4.4. İnce ayarlama sonrası modelin kendi tokenizerı ile eğitildikten sonra oluşan özetin ROUGE skor değerleri

ROUGE-1	ROUGE-2	ROUGE-L
0.23740816321413494	0.023557382861461487	0.15565327427649706

Çizelge 4.4 incelendiğinde dil modelinin makale veri seti üzerinde ince ayarlanmasının ROUGE-1 değerini artırdığı görülmektedir. Tez çalışması için oluşturulmuş tokenizer ile 1 epoch değeri için, ince ayarlanmış PEGASUS dil modelinin yaptığı özetleme işleminin ROUGE skor değerleri Çizelge 4.5'te, oluşan yeni özet ise Şekil 4.3'te verilmiştir.



Tez Çalışması İçin Oluşturulmuş Tokenizer İle Yapılan Özet

"zaman bir takım metin ön işleme adımların uygulanması ile başlamaktadır .Daha sonra her birlerinden yeni bir dünya oluşturmuştur Eski dünyadan farklı olarak sadece bir zümre değil sonucu olarak kullanmaktadır İletişim imkanlarının karşılıklı hale gelmesi tüketicilerin anlık geribildirimlerini firmaya iletebilecekleri koşulların sağlanmasına neden olmuştur yazacakları Gerek fiziki ürünlerin gerekse hizmetlerin tüketici yle buluşturulmasından hemen sonra tüketiciler sosyal medya ve internet platformlarıyla inşa ise en hızlı olduğu anlaşılabilmektedir Nitekim teslim zamanları ifadesi en hızlı sıfatıyla nitelendirilmektedir nitelendirilmektedir.YÖNTEM ..Çalışmada emEval ABSA 2016 oluşturulmuş ve birçok çalışmada da kullanılmış olan Türkçe Rest yorumlarını içeren bir veri!"

Şekil 4.3. Tez çalışması için oluşturulmuş tokenizer ile yapılan özet

Çizelge 4.5. Tez çalışması için oluşturulmuş tokenizer ile yapılan özetleme işleminin ROUGE skor değerleri (1 epoch)

ROUGE-1	ROUGE-2	ROUGE-L
0.0401192318563552	0.002684563758389261	0.02918615651257788

Çizelge 4.5 incelendiğinde ROUGE değerlerinin modelin kendi tokenizerı ile yapılan özete göre düştüğü görülmektedir. ROUGE değerlerinin düşmesi modelin daha başarısız özet oluşturduğunu ifade ediyor gibi görünse de Şekil 4.3 incelendiğinde Şekil 4.2 ye göre çok daha anlamlı cümleler ürettiği, referans özetle anlamsal olarak daha benzer olduğu görülmektedir.

Oluşturulan dil modeli 100 epoch değeri ile eğitilmiş ve elde edilen özet Şekil 4.4'te, ROUGE değerleri ise Çizelge 4.6'da gösterilmiştir.

Çizelge 4.6. Tez çalışması için oluşturulmuş tokenizer ile yapılan özetleme işleminin ROUGE skor değerleri (100 epoch)

ROUGE-1	ROUGE-2	ROUGE-L
0.1244071300549256	0.021773001581426386	0.11068085055811232

Çizelge 4.6 incelendiğinde 100 epoch değeri ile eğitildikten sonra gerçekleştirilen özetleme işleminden elde edilen ROUGE değerlerinde çizelge 4.5 ile karşılaştırıldığında artış gözlemlenmektedir.

'Bu çalışmanın amacı hedef terim tespitinde başarılı bir şekilde kullanılabileceğini göstermektedir. ROC eğrisi altında kalan alanı temsil eden AUC değerleri sırasıyla Word.Vec için 0.83 Glove için 086 ve Fasttext için ise 0.88 olarak bulunmuştur .Çalışmanın son aşamada ise 3 farklı kelime vektörü yönteminden elde edilen vektörler birleştirilerek hedef terim tespiti yapılmıştır. Her üç yöntemden elde edilen kelime temsillerinin birleştirilmesiyle 1000 boyutlu bir kelime vektörü kullanılmıştır değişikliklerin farklı kelime temsil vektörlerinin birleştirilmesinin sınıflandırma başarısı üzerinde bir iyileştirme sağlamadığı tespit edilmiştir. Bu çalışma kapsamında önerilen modelin hedef terim tespiti problemi ele alınmış olsa da kelime vektörlerinin hedef kategori ve duygu sınıfını belirlenmesi problemleri de test edilmiştir. Fakat elde edilen sonuçlar literatür de kullanılan yöntemlere oranla daha yüksek bir başarı sağlayamamıştır. Bu durum kategori sınıflarındaki örnek sayısının dengesiz olmasından kaynaklanmaktadır. Bazı kategoriler çok sayıda örnek cümle içerirken bazı kategorilerdeki örnek cümle sayısı oldukça sınırlıdır .Bu çalışmada ABSA yarışmasında oluşturulmuş olan Türkçe restoran veri kümesi üzerinde hedef terim tespiti için yeni bir model önerilmiştir Önerilen modelde sadece kökü. Bu çalışma kapsamında önerilen modelin hedef terim tespitine yönelik Türkçe Restoran veri seti üzerinde farklı çalışmaların yapıldığı ile elde şekilde tespit edilmiştir. Bu çalışma kapsamında önerilen modelin hedef terim tespiti için temel bir yöntem zaten önerilmiş ve yarışma ekiplerinden daha başarılı sonuçlar beklenmiştir ABSA tarafından önerilen temel modelde kategoriler için bir sözlük oluşturulmuştur Kategorilerde!'

Şekil 4.4. 100 epoch değeri ile eğitilmiş modelin oluşturduğu özet

Şekil 4.4'te tez için oluşturulan makale veri seti üzerinde ince ayarlanmış PEGASUS dil modelinin, 100 epoch ile eğitim sonrası oluşturmuş olduğu özet görülmektedir. Özet incelendiğinde Şekil 4.3' e kıyasla daha kurallı cümlelerden oluştuğu, bazı noktalama işaretlerinin kullanıldığı, makale hakkında fikir verdiği görülmektedir. ROUGE değerleri ile kıyaslama yapıldığında referans özetten farklı bir özet oluşturulmuş ancak referans özetin insan muhakemesi ile oluşturulduğu düşünülürse modelin özetinin tarafsız olduğu ve kısmen başarılı olduğu söylenebilir.

5. SONUÇ VE ÖNERİLER

Literatür incelendiğinde Türkçe dili için uzun metinlerden büyük dil modelleri aracılığıyla yoruma dayalı özet çalışmasına rastlanmamıştır. Bunun sebeplerinden birisi Türkçe'nin karmaşık morfolojik yapısı sebebiyle yoruma dayalı özetleme işleminin yapılabilmesi için çok büyük veri setleri ile çalışılmasının gerekliliğidir. Büyük veri setleri oluşturmak ve bunlar kullanılarak dil modeli geliştirmek oldukça maliyetli bir işlemdir.

Büyük dil modellerinin Türkçe için eğitilmesinde kullanılan büyük veri setleri incelendiğinde, verilerin genellikle bilinen haber kanallarının çevrimiçi platformlarında yayınlanmış haber metinleri ve özetlerinden oluştuğu görülmektedir. Bu veri setleri haber sayısı bakımından çok büyük veri setleri olsalar da metin uzunluğu açısından incelendiğinde yeterince büyük değillerdir. Otomatik metin özetlemenin hedeflerinden birisi olan çok uzun metinler hakkında tamamını okumadan bilgi sahibi olma fikri için, Türkçe dilinde yapılmış olan özetleme çalışmaları yeterli değildir.

Bu tezin kapsamında yapılan çalışmada, Hugging Face Hub ortamında Google tarafından paylaşılmış, açık erişimli "pegasus-x-large" ön eğitilmiş dil modeli ince ayarlanarak, Türkçe uzun metinleri özetleme amacına uygun şekilde özelleştirilmeye çalışılmıştır. Bunun için 3 farklı alanı kapsayan, rastgele seçilmiş, 200 tane akademik makalenin tüm metin içeriği, yazarlar tarafından oluşturulmuş özetleri, çalışma alanı, anahtar kelimeleri ve makalenin isminin bulunduğu Türkçe makale veri seti tez çalışması için sıfırdan oluşturulmuştur.

Çalışmanın zorlayıcı taraflarından birisi veri setinin hazırlık işlemleri olmuştur. Verilerin metin özetlemede kullanılabilmeleri için bilimsel makalelerin şekil, tablo ve formül içeriğinden arındırılması gerekmiştir. Bu oldukça zaman alan, uğraştırıcı bir işlemdir. Bu sebeple çok fazla makaleden veri toplamak oldukça zorlayıcıdır. Akademik makalelerden büyük veri seti oluşturma işleminin bireysel olarak değil, kalabalık bir ekiple önceden belirlenmiş veri seti oluşturma kurallarına uyarak yapılmasının daha kolay bir yöntem olacağı düşünülmektedir.

PEGASUS dil modeli temelinde transformer kodlayıcı kod çözücü bir mimariye sahiptir. Maskeleme işlemi yaparak boşlukların tahmin edilmesi fikri üzerine oluşturulmuş bir modeldir ancak çok iyi bir çıkarıma dayalı metin özetleme başarısına sahip BERT dili gibi sözcük ya da sözcük gruplarını maskelememekte, cümleleri maskeleyerek, maskelenmiş cümlelerin tahmini olan cümleleri birleştirerek yoruma dayalı metin özetlemeyi gerçekleştirmektedir. PEGASUS dil modeli Türkçe büyük veri setleri ile ön eğitilmemiştir. Bu tez çalışmasında PEGASUS dil modeli büyük Türkçe veri setleri üzerinde sıfırdan eğitilmeye çalışılmış ancak Google'ın sunmuş olduğu Colab Pro geliştirme ortamı donanımsal olarak bu işlem için yetersiz kalmıştır.

Tez kapsamında gerçekleştirilen metin özetleme yöntemi yoruma dayalı metin özetleme yöntemidir. Bunun için PEGASUS dil modelinin Türkçe makale veri seti üzerinde ince ayarlaması yapılmıştır. Oluşturulan yeni dil modeli ile elde dilen özetlerin başarı değerlendirmesi ROUGE ölçüm metrikleri ile gerçekleştirilmiştir. Çizelge 5.1'de elde edilen ROUGE skor değerleri gösterilmektedir.

Çizelge 5.1. ROUGE skor değerleri

Model	ROUGE-1	ROUGE-2	ROUGE-L
(A)	0.23576811659924030	0.023730660848292374	0.15518827795810117
(B)	0.23740816321413494	0.023557382861461487	0.15565327427649706
(C)	0.04011923185635520	0.002684563758389261	0.02918615651257788
(D)	0.12440713005492560	0.021773001581426386	0.11068085055811232

Çizelge 5.1'de ROUGE skor değerleri verilen modellerin açıklaması aşağıda verilmiştir.

(A): PEGASUS dil modelinin ince ayarlanması öncesi dil modeli

(B): Türkçe makale veri seti üzerinde ince ayarlanmış PEGASUS dil modeli

(C): Türkçe makale veri seti için oluşturulmuş tokenizer ile ince ayarlanmış PEGASUS dil modeli (1 epoch)

(D): Türkçe makale veri seti için oluşturulmuş tokenizer ile ince ayarlanmış PEGASUS dil modeli (100 epoch)

Çizelge 5.1 incelendiğinde dil modelinin kendi tokenizerı kullanılarak ince ayarlanma işlemi sonrasında ROUGE skor değerinin arttığı görülmektedir. Daha sonra model

Türkçe makale veri seti kullanılarak oluşturulan tokenizer ile eğitildiğinde ROUGE değerlerinde düşme gerçekleşmiştir. ROUGE skor değerleri referans özet ile modelin metinden oluşturduğu özet arasında n-gram örtüşme oranını ifade eder. ROUGE skor değerleri 0 ve 1 arasında değer alır. Örtüşmenin birebir aynı olduğu durumda ROUGE değerleri 1 olmaktadır. Oluşturulan modeller ile yoruma dayalı metin özetleme işlemi gerçekleştirildiği için yeni özetlerin modelin kendi oluşturduğu cümleler ile yapılmış olması beklenmektedir. Referans özet ise makale yazarının oluşturduğu özettir. Bu durumda modelin oluşturduğu özet ile referans özet arasında örtüşme olması beklenmemektedir. ROUGE skor değerleri ile yoruma dayalı metin özetleme karşılaştırmasının istenen bilgiyi vermediği görülmektedir. Bu sebeple veri setinde örnek bir makale için modelin oluşturduğu özet ve referans özet tezin Araştırma Bulguları başlığı altında gösterilmiştir. Bu özetler incelendiğinde (A) modeli için oluşturulan özetin makale içerisindeki rastgele bir cümle olduğu görülmektedir. (C) ve (D) modellerinin oluşturduğu özetler incelendiğinde özetteki cümlelerin yapısal olarak hatasız olmadığı halde anlamsal olarak referans özete daha benzer olduğu gözlemlenmiştir.

İnce ayarlama işleminin 200 adet makaleden oluşan küçük bir veri seti ile yapılmış olmasına rağmen Türkçe cümle kurabilen bir sistemin ortaya çıkmış olması umut vericidir. Daha büyük bir Türkçe veri seti ile ön eğitim gerçekleştirilip, sonradan bir makale veri seti üzerinde ince ayarlama yapılırsa çok daha başarılı sonuçların elde edileceği düşünülmektedir.

Literatürde PEGASUS dil modelinin Türkçe veri seti üzerinde ince ayarlama işlemine rastlanmamıştır. Bu sebeple başarı değerlendirmeleri başka bir çalışma ile karşılaştırılamamıştır ancak bu çalışmanın bundan sonra yapılacak çalışmalar için bir temel oluşturacağı düşünülmektedir.

Daha sonra yapılacak olan çalışmalarda uzun metinlerden yoruma dayalı özetlemede daha başarılı sonuçlar elde etmek için uzun metinlerin bölümlere ayrılarak öncelikle anahtar kelimeleri de kullanarak çıkarımsal özetleme gerçekleştirilmesi, oluşan görece daha kısa metin üzerinde yoruma dayalı metin özetleme çalışmaları yapılması düşünülmektedir.

KAYNAKLAR

Adalı, E., 2020. Türkçe Doğal Dil İşleme. Akçağ Yayınları, 754s, Ankara.

Afatsun, M. N. 2020. Derin Öğrenme Yöntemleri İle Türkçe Metinlerden Anlamlı Özet Çıkarma. Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 78s, Ankara.

Alıpour, N., 2023. Türk Dilinde Derin Öğrenme İle Metin Özetleme. Atatürk Üniversitesi, Sosyal Bilimler Enstitüsü, Doktora Tezi, 97s, Erzurum.

Altan, Z., 2004. A Turkish automatic text summarization system. IASTED International Conference on AIA. (pp. 16–18).

Alpaut, O., 1980. Kimyasal Termodinamik. S.D.Ü. Yayınları, A30, 558s. Isparta.

Alwandawybek, O., Serttaş, S. & Güngör, E. (2023). Metin Özetleme Yöntemlerinin İncelenmesi ve Karşılaştırılması. International Journal of Advanced Natural Sciences and Engineering Researches, 7(3), 8-23.

Ay, B., Ertam, F., Fidan, G., Aydın, G., 2023. Turkish abstractive text document summarization using text to text transfer transformer. Alexandria Engineering Journal, 68, 1-13.

Aydogan, M., Karci, A., 2019. Turkish Text Classification with Machine Learning and Transfer Learning. 10.1109/IDAP.2019.8875919.

Aydın, G., 2022. Abstractive text summarization using deep learning with a new Turkish summarization benchmark dataset. Concurrency and Computation: Practice and Experience, 34(9), e6482. https://doi.org/10.1002/cpe.6482

Barzilay, R., Elhadad, M., 1997. Using Lexical Chains for Text Summarization. In Intelligent Scalable Text Summarization.

Baykara, B., Güngör, T., 2022. Abstractive text summarization and new large-scale datasets for agglutinative languages Turkish and Hungarian. Lang Resources & Evaluation 56, 973–1007. https://doi.org/10.1007/s10579-021-09568-y

Baykara, B., Güngör, T., 2023. Turkish abstractive text summarization using pretrained sequence-to-sequence models. Natural Language Engineering, 29, 1275-1304.

Beken F., Oflazer, K., Yanikoglu, B., 2021. Semantic Similarity Based Evaluation for Abstractive News Summarization.

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. Dokuz Eylül University, Graduate School Of Natural And Applied Scıences, Thesis of PhD, 108s, İzmir

Bal, Salih., 2022. Çıkarıcı Türkçe Metin Özetleme Performansını İyileştirmek İçin Yeni Yöntemler. Eskişehir Osmangazi Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 68s, Eskişehir.

Doğan, E., 2019. Derin Öğrenme Yöntemiyle Çevrimiçi Sosyal Ağlarda Duygu Analizi ve Metin Özetleme. Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 46s, Elazığ.

Edmundson, H. P. (1969). New methods in automatic extraction, Journal of the Association for Computing Machinery, 16 (2), 264–285.

Erdağı, E., 2023. Türkçe Metinlerde Çıkarım Tabanlı Otomatik Metin Özetleme. T. C. Maltepe Üniversitesi, Lisansüstü Eğitim Enstitüsü, Doktora tezi, 91s, İstanbul.

Ertam F, Aydin G., 2022. Abstractive text summarization using deep learning with a new Turkish summarization benchmark dataset. Concurrency Computat Pract Exper. 2022;34(9):e6482. doi: 10.1002/cpe.6482

Goldstein, J., Carbonell, J., 1998. Summarization: Using MMR for Diversity- Based Reranking and Evaluating Summaries. In TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 181–195, Baltimore, Maryland, USA. Association for Computational Linguistics.

Goyal, Tanya, Xu, Jiacheng, Li, Junyi Jessy, & Durrett, G., 2022. Training Dynamics for Text Summarization Models. Findings of the Association for Computational Linguistics: ACL 2022.

Gupta, V. and Lehal, G. S. 2010. A survey of text summarization extractive techniques. Journal of Emerging Technologies in Web Intelligence. 2(3), 258-268.

Gümüş, M., 2019. An Evaluation Of Automatic Text Summarization Tecniques. Bahçeşehir University, Institute Of Science, Computer Engineering, Master's Thesis, 74, Istanbul.

Güven, Z., 2021. The Effect of BERT, ELECTRA and ALBERT Language Models on Sentiment Analysis for Turkish Product Reviews. 629-632. 10.1109/UBMK52708.2021.9559007.

Gündeş, Ö., 2020. A Deep Learning-Based Extractive Text Summarization System For Turkish News Articles. Boğaziçi University, Institute For Graduate Studies In Social Science, Master Of Arts, 102s, İstanbul.

Güran A., 2013.Otomatik Metin Özetleme Sistemi. Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 108s, İstanbul.

Hatipoğlu, A., Omurca, S.I., 2015. Türkçe Metin Özetlemede Melez Modelleme. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 17(50), 95 – 108.

Karakoc, E., Yilmaz, B., 2019. Deep Learning Based Abstractive Turkish News Summarization. 2019 27th Signal Processing and Communications Applications Conference (SIU), IEEE.

Kara, C., 2023. Bilgisayar Bilimleri Alanında Yapılan Türkçe Akademik Yayınların Doğal Dil İşleme Yöntemleri İle İncelenmesi. Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 74s, Isparta.

Karakaya K. M., & Güvenir H. A., 2004. ARG: A tool for automatic report generation, Istanbul University - Journal of Electrical & Electronics Engineering, 4 (2), 1101-1109.

Kemaloğlu Alagöz, N., 2022. Derin Öğrenme Yöntemleri İle Otomatik Metin Özetleme. Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 68s, Isparta.

Klavans, J., L., Kan, M., Y., McKeown, K., 2001. Domainspecific informative and indicative summarization for information retrieval. Proceedings of the first Document Understanding Conference. 19-26, New Orleans, USA

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. Artificial Intelligence. 139(1), 91-107.

Kutlugün, M., A., Şirin, T., 2018. Turkish meaningful text generation with class based n-gram model. 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 1-4.

Lin, C.-Y., 2004. ROUGE: A Package for Automatic Evaluation of summaries. Proceedings of the ACL Workshop: Text Summarization Braches Out. 10.

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

Mahesh, B. (2020) Machine Learning Algorithms-A Review. International Journal of Science and Research, 9, 381-386.

Mani, I., Bloedorn, E., 1997. Multi-Document Summarization by Graph Search and Matching, AAAI/IAAI.

Mercan, Ö.B., Cavsak, S.N., Deliahmetoglu, A., & Tanberk, S., 2023. Abstractive Text Summarization for Resumes With Cutting Edge NLP Transformers and LSTM. 2023 Innovations in Intelligent Systems and Applications Conference (ASYU), 1-6.

Mihalcea, R., Tarau, P., 2005. A language independent algorithm for single and multiple document summarization. In Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts.

Munot N., Govilkar, S., S., 2014. Comparative Study of Text Summarization Methods. International Journal of Computer Applications, 102(12), 33-37.

Mutlu, B. 2020. Hibrit Zeki Sistem ile Metin Özetleme. Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 155s, İstanbul.

Oflazer, K., and Kuruoz, İ., 1994. Tagging and Morphological Disambiguation of Turkish Text. In Fourth Conference on Applied Natural Language Processing, pages 144–149, Stuttgart, Germany. Association for Computational Linguistics.

Onan, A. (2022). Türkçe Metin Madenciliği için Dikkat Mekanizması Tabanlı Derin Öğrenme Mimarilerinin Değerlendirilmesi. Avrupa Bilim ve Teknoloji Dergisi (34), 403-407.

Ozsoy, M., Cicekli, I., Alpaslan, F., 2010. Text Summarization of Turkish Texts using Latent Semantic Analysis. Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference. 2. 869-876.

Pattanayak, S., 2023. Pro Deep Learning with Tensorflow 2.0: A Mathematical Approach to Advanced Artificial Intelligence in Python. Springer.

Scialom, T., Dray, P., Lamprier, S., Piwowarski, B., Staiano, J., 2020. MLSUM: The Multilingual Summarization Corpus. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 8051–8067, Online. Association for Computational Linguistics.

Seo, Minjoon & Kembhavi, Aniruddha & Farhadi, Ali & Hajishirzi, Hannaneh. (2016). Bidirectional Attention Flow for Machine Comprehension.

Sjöblom, M., 2023. Metadata Assisted Finetunning With Large Pre-Trained Language Models for Abstractive Text Summarization, Umea University, Master of Science Programme in Computing Science and Engineering, 28s, Sweden.

Srivastava, Vivek, Bhat, Savita, and Pedanekar, Niranjan. 2023. Hiding in Plain Sight: Insights into Abstractive Text Summarization. In The Fourth Workshop on Insights from Negative Results in NLP, 67–74, Dubrovnik, Croatia. Association for Computational Linguistics.

Şahin, G. G., 2018. Buildıng Of Turkish Propbank And Semantic Role Labeling Of Turkish. Istanbul Technical University, Graduate School Of Science Engineering And Technology, Ph.D. Thesis, 132p, Istanbul.

Şeker, S., 2015. Çizge Teorisi (Graph Theory), YBS Ansiklopedi, 2,2. 17-29.

Torres-Moreno, J. M., 2014. Automatic Text Summarization. John Wiley & Sons.

Tunstall, L., Werra, L.,V., Wolf, T., 2022. Natural Language Processing with Transformers: Building Language Applications with Hugging Face. Revised edition. Sebastopol: O'Reilly, 383 pp. isbn: 978-1-09-813679-6.

Tülek M., 2007. Türkçe İçin Metin Özetleme. İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 88s, İstanbul.

Tür, G., Hakkani-Tür, D., Oflazer, K. (2003). A statistical information extraction system for Turkish. Natural Language Engineering, 9, 181-210.

Uğur A., KINACI A.C., 2006. Yapay Zeka Teknikleri ve Yapay Sinir Ağları Kullanılarak Web Sayfalarının Sınıflandırılması, 11.İnternet Konferansları.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I., 2017. Attention is all you need.

Vivek Srivastava, Savita Bhat, and Niranjan Pedanekar. 2023. Hiding in Plain Sight: Insights into Abstractive Text Summarization. In The Fourth Workshop on Insights from Negative Results in NLP, 67–74, Dubrovnik, Croatia. Association for Computational Linguistics.

Wan, D., Bansal, M., 2022. FactPEGASUS: Factuality-Aware Pre-training and Fine-tuning for Abstractive Summarization. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1010–1028, Seattle, United States. Association for Computational Linguistics.

Witbrock, M., Mittal, V., 1999. Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries (poster abstract). 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.

Yang,T., H., Lu, C., C., Hsu, W., L., 2021. More than Extracting Important Sentences: the Application of PEGASUS International Conference on Technologies and Applications of Artificial Intelligence (TAAI), Taichung, Taiwan, 2021, 131-134, doi: 10.1109/TAAI54685.2021.00032.

Yüksel, Y., 2021. Tr-Sum: A Text Summarızer For Turkısh. Dokuz Eylül Unıversıty, Graduate School Of Natural And Applıed Scıences, Master Thesis, 77s, İzmir.

Zhang, Y., Jin, R., Zhou, Z., 2010. Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics. 1. 43-52. 10.1007/s13042-010-0001-0.