

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DERİN ÖĞRENME YÖNTEMLERİ KULLANILARAK TÜRKÇE
HABER METİNLERİNDEN HABER BAŞLIĞI ÜRETİLMESİ

ENİSE KARAKOÇ
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

GEBZE
2019

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DERİN ÖĞRENME YÖNTEMLERİ
KULLANILARAK TÜRKÇE HABER
METİNLERİNDEN HABER BAŞLIĞI
ÜRETİLMESİ

ENİSE KARAKOÇ
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMANI
DR. ÖĞR. ÜYESİ BURCU YILMAZ

GEBZE
2019

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**GENERATING NEWS HEADLINE FROM
TURKISH NEWS USING DEEP LEARNING
METHODS**

ENİSE KARAKOÇ
A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE
DEPARTMENT OF COMPUTER ENGINEERING

THESIS SUPERVISOR
ASSIST. PROF. DR. BURCU YILMAZ

GEBZE
2019

	YÜKSEK LİSANS JÜRİ ONAY FORMU
---	--------------------------------------

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 26/06/2019 tarih ve 2019/29 sayılı kararıyla oluşturulan jüri tarafından 22/10/2019 tarihinde tez savunma sınavı yapılan Enise KARAKOÇ'un tez çalışması Bilgisayar Mühendisliği Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE

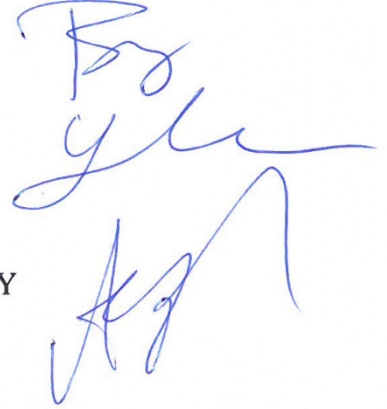
(TEZ DANIŞMANI) : Dr. Öğr. Üyesi Burcu YILMAZ

ÜYE

: Dr. Öğr. Üyesi Yakup GENÇ

ÜYE

: Dr. Öğr. Üyesi Ayşe Betül OKTAY



ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun

...../...../..... tarih ve/..... sayılı kararı.

ÖZET

Otomatik metin özetleme problemi, makine öğrenmesi ve doğal dil işleme alanlarında yaygın olarak çalışılmaktadır. Günümüzde internet üzerinden erişilebilen verilerin devasa boyutlara ulaşması sebebiyle metinleri otomatik olarak kısaltan, sadece istenilen bilgiye hızlı bir şekilde ulaştıran ve makine tarafından üretilen özet sistemlerine ihtiyaç vardır. Metin özetleme sistemlerinin günümüzde pek çok kullanım alanı bulunmaktadır. Haber başlıklarının üretilmesi bu kullanım alanlarına örnek olarak gösterilebilir.

Literatürde çıkarıma dayalı ve yoruma dayalı olmak üzere iki çeşit metin özetleme yaklaşımı mevcuttur. Çıkarıma dayalı metin özetleme yaklaşımında, girdi metnini önemli ölçüde açıklayacak kelimeler veya cümleler seçilerek metinde herhangi bir değişiklik yapılmadan özet oluşturulur. Yoruma dayalı metin özetleme yaklaşımında ise, kelimeler anlamlandırılıp yeni cümleler oluşturularak özet elde edilir. Dolayısıyla yoruma dayalı özetleme sayesinde insanlar tarafından daha anlaşılır özetler oluşturmak mümkündür.

Bu çalışmada, yoruma dayalı metin özetleme yaklaşımı ile Türkçe haber metinlerinden haber başlıkları üretilmiştir. Haber başlıkları üretilirken derin öğrenme yöntemlerinden kodlayıcı-kod çözücü mimarisi geliştirilmiştir. Sistemin eğitim aşamasında bir RNN çeşidi olan Çift Yönlü LSTM ağı dikkat mekanizması ile birlikte kullanılmıştır. Sistemin test aşamasında ise haber başlıklarının çıkarımı sırasında, kod çözücü olarak açgözlü arama ve ışıın aramaları kullanılmıştır. Deneyler sırasında sistem çeşitli sözlük boyutları, haber başlığı boyutları, haber metni boyutları ve ışıın genişlikleri ile ayrı ayrı test edilmiştir. Sistemin başarısını ölçerken ROUGE değerlendirme metriği ve anlamsal benzerlik skorları kullanılmıştır.

İki ayrı veri kümesinde gerçekleştirilen deney sonuçlarına göre dikkat mekanizması ile eğitilmiş olan Çift Yönlü LSTM ağı ile ışıın aramasının kullanıldığı modelde, en iyi parametrelerle ortalama ROUGE-1 skoru 0,27 olarak ve ortalama anlamsal benzerlik skoru 0,60 olarak elde edilmiştir.

Anahtar Kelimeler: Haber Başlığı Üretme, Yoruma Dayalı Metin Özetleme, Kodlayıcı-Kod Çözücü Modeli, Dizi Tahmin Modeli.

SUMMARY

The problem of automatic text summarization is widely studied in the field of machine learning and natural language processing. Due to size of the enormous data that is accessed over the internet, there is a need for machine-generated summary systems that automatically shorten texts, deliver only the desired information quickly. Text summarization systems can be used in many areas in our daily life. Generation of news headlines is an example of these usage areas.

There are two types of text summarization approaches in the literature: extractive and abstractive. In extractive text summarization approach, words or sentences that explain the input text better are selected and a summary is created without any changes in the original text. However, in abstractive text summarization approach, summaries are generated with new sentences that are formed by words based on their meanings. Therefore, it is possible to generate summaries that are more meaningful to the people by using abstractive text summarization approach.

In this study, news headlines have been generated from Turkish news texts by using abstractive text summarization approach. While generating news headlines, encoder-decoder architecture from deep learning methods has been developed. During training phase of the system Bidirectional LSTM network, which is a type of RNN, is used together with attention mechanism. In test phase of the system, greedy search decoder and beam search decoder approaches are used while inferencing new headlines. During the experiments, system has been tested separately on various dictionary sizes, news headline lengths, news text lengths and beam widths. The success of the system is measured by both ROUGE score and semantic similarity score.

According to the experiment results carried out in two separate news datasets, the system achieved average ROUGE-1 score of 0,27 and average semantic similarity score of 0,60 with the best parameters trained on Bidirectional LSTM network together with attention mechanism and beam search decoder.

Keywords: News Headline Generation, Abstractive Text Summarization, Deep Learning, Encoder-Decoder Model, Sequence Prediction Model.

TEŞEKKÜR

Başta, lisans ve yüksek lisans eğitimim boyunca bilgilerini, tecrübelerini, değerli zamanını benden esirgemeyerek her daim bana sabırla, anlayışla çözümler sunan, tezim sırasında bana olan inancını hiç kaybetmeyen, samimi bir arkadaş edasıyla hayatımın her alanında beni motive eden çok değerli tez danışmanım Dr. Öğr. Üyesi Burcu YILMAZ hocama sonsuz teşekkürlerimi sunarım.

Tezimin iyileştirilmesi konusunda benden bilgi ve yardımlarını eksik etmeyerek değerli katkılarda bulunan tez jürilerim Dr. Öğr. Üyesi Ayşe Betül OKTAY ve Dr. Öğr. Üyesi Yakup GENÇ hocalarıma sonsuz teşekkürlerimi sunarım.

Manevi destekleri ve bana olan inançları için aileme, değerli zamanını ayırarak tezimin yazım hatalarının giderilmesine katkıda bulunan sevgili Esra Zeynep KARAKOÇ'a teşekkürü bir borç bilirim.

Hayatımın her alanında olduğu gibi bu zorlu tez sürecimde de birçok fedakârlık göstererek bana katlanan, tezim konusunda beni cesaretlendiren ve bu tezin yazımında çokça katkıları olan sevgili eşim Necmeddin Said KARAKOÇ'a en içten teşekkürlerimi sunarım.

İÇİNDEKİLER

	<u>Sayfa</u>
ÖZET	v
SUMMARY	vi
TEŞEKKÜR	vii
SİMGELER ve KISALTMALAR DİZİNİ	x
ŞEKİLLER DİZİNİ	xi
TABLolar DİZİNİ	xiii
1. GİRİŞ	1
2. İLGİLİ ÇALIŞMALAR	5
2.1. Çıktı Türüne Göre Metin Özetleme	5
2.1.1. Çıkarıma Dayalı Metin Özetleme Çalışmaları	6
2.1.2. Yorumu Dayalı Metin Özetleme Çalışmaları	10
2.2. Girdi Türüne Göre Metin Özetleme	12
2.2.1. Tekli Doküman Özetleme Çalışmaları	13
2.2.2. Çoklu Doküman Özetleme Çalışmaları	13
2.3. Amacına Göre Metin Özetleme	13
2.3.1. Genel Metin Özetleme Çalışmaları	14
2.3.2. Alana Özgü Metin Özetleme Çalışmaları	14
2.3.3. Sorgu Tabanlı Metin Özetleme Çalışmaları	14
3. HABER BAŞLIKLARININ ÜRETİLMESİ	15
3.1. Haber Veri Kümeleri	15
3.1.1. Haber Veri Kümesi A	15
3.1.2. Haber Veri Kümesi B	19
3.2. Veri Kümelerinin Ön İşlemden Geçirilmesi	22
3.3. Sözlük Oluşturulması	24
3.4. Kelime Gömmelerinin Çıkarılması	26
3.5. Sistemin Eğitimi	30
3.5.1. Çift Yönlü LSTM	31
3.5.2. Dikkat Mekanizması	32
3.5.3. Hiper Parametre Konfigürasyonu	33

3.6. Haber Başlıklarının Üretilmesi	34
3.6.1. Açgözlü Arama	35
3.6.2. Işın Araması	35
4. DENEYLER	36
4.1. Değerlendirme Metrikleri	36
4.1.1. ROUGE Skoru	36
4.1.2. Kosinüs Benzerlik Skoru	37
4.2. Eğitim ve Test Kümelerinin Oluşturulması	37
4.3. Sonuçlar	39
5. SONUÇ VE TARTIŞMA	58
KAYNAKLAR	60
ÖZGEÇMİŞ	68
EKLER	69

SİMGELER ve KISALTMALAR DİZİNİ

<u>Simgeler ve Kısaltmalar</u>	<u>Açıklamalar</u>
RNN	: Tekrarlayan Sinir Ağı (Recurrent Neural Network)
LSTM	: Uzun Kısa Süreli Bellek (Long Short Term Memory)
LSA	: Gizli Anlamsal Analiz (Latent Semantic Analysis)
ROUGE	: Recall-Oriented Understudy for Gisting Evaluation
CBOW	: Continuous Bag of Words (Devamlı Kelime Çantası)
RSS	: Reference Summary Set (Referans Özet Kümesi)
OOV	: Out-of-Vocabulary (Sözlük Dışı Kelime)
DUC	: Document Understanding Conference (Doküman Anlama Konferansı)

ŞEKİLLER DİZİNİ

<u>Sekil No:</u>	<u>Sayfa</u>
1.1: Metin özetleme kullanım alanları. a) Haber başlığı örneği b) Film özeti örneği c) Kitap hakkında yorum örneği.	2
1.2: Sistemin çalışma yapısı.	3
2.1: Metin özetleme yaklaşımları.	5
2.2: Çıktı türüne göre metin özetleme yaklaşımlarının görselleştirilmesi.	6
3.1: Haber Veri Kümesi A haber metni dizgecik sayısı dağılımları.	17
3.2: Haber Veri Kümesi A haber başlığı dizgecik sayısı dağılımları.	17
3.3: Haber Veri Kümesi B haber metni dizgecik sayısı dağılımları.	21
3.4: Haber Veri Kümesi B haber başlığı dizgecik sayısı dağılımları.	22
3.5: Haber Veri Kümesi B'ye ait bir haber başlığından dizgecik indeks vektörünün elde edilmesi.	26
3.6: Haber Veri Kümesi B'den alınan örnek bir haber başlığına ait kelime gömmelerinin çıkarılması.	28
3.7: Haber Veri Kümesi A sözlüğünde nadir geçen "reddediyoruz" dizgeciğinin FastText'te yer alan benzer dizgecikleri.	29
3.8: Haber Veri Kümesi B sözlüğünde nadir geçen "depresif" dizgeciğinin FastText'te yer alan benzer dizgecikleri.	30
3.9: Kodlayıcı-kod çözücü mimarisi.	31
3.10: LSTM ve Çift Yönlü LSTM Mimarileri.	32
3.11: Dikkat katmanı eklenmiş kodlayıcı-kod çözücü mimarisi.	33
3.12: Işın araması algoritması.	35
4.1: Çapraz doğrulama yönteminin görselleştirilmesi ($k = 10$).	39
4.2: Haber Veri Kümesi A'ya ait pozitif bir deney sonucu.	49
4.3: Haber Veri Kümesi A'ya ait pozitif bir deney sonucu.	49
4.4: Haber Veri Kümesi A'ya ait pozitif bir deney sonucu.	50
4.5: Haber Veri Kümesi A'ya ait pozitif bir deney sonucu.	50
4.6: Haber Veri Kümesi A'ya ait pozitif bir deney sonucu.	51
4.7: Haber Veri Kümesi A'ya ait negatif bir deney sonucu.	51
4.8: Haber Veri Kümesi A'ya ait negatif bir deney sonucu.	52
4.9: Haber Veri Kümesi A'ya ait negatif bir deney sonucu.	52

4.10:	Haber Veri Kümesi B'ye ait pozitif bir deney sonucu.	53
4.11:	Haber Veri Kümesi B'ye ait pozitif bir deney sonucu.	53
4.12:	Haber Veri Kümesi B'ye ait pozitif bir deney sonucu.	54
4.13:	Haber Veri Kümesi B'ye ait pozitif bir deney sonucu.	54
4.14:	Haber Veri Kümesi B'ye ait pozitif bir deney sonucu.	55
4.15:	Haber Veri Kümesi B'ye ait negatif bir deney sonucu.	55
4.16:	Haber Veri Kümesi B'ye ait negatif bir deney sonucu.	56
4.17:	Haber Veri Kümesi B'ye ait negatif bir deney sonucu.	56

TABLÖLAR DİZİNİ

<u>Tablo No:</u>	<u>Sayfa</u>
3.1: Haber Veri Kümesi A'nın elde edildiği web siteleri ve haber sayıları.	16
3.2: Haber Veri Kümesi A'ya ait en çok haber içeren ilk 13 web sitesi, web sitelerine ait örnek haber başlıkları ve kısaltılmış örnek haber metinleri.	18
3.3: Haber Veri Kümesi B'ye ait sınıflar, sınıflara ait örnek haber başlıkları ve kısaltılmış örnek haber metinleri.	20
3.4: Haber Veri Kümesi B'ye ait istatistiksel bilgiler.	21
3.5: Haber Veri Kümesi B'ye ait ara başlık içeren örnek bir haber metni.	23
3.6: Haber Veri Kümesi A'nın sözlüğünün ilk 15 indeksi ve geçme sıklıkları.	24
3.7: Haber Veri Kümesi B'nin sözlüğünün ilk 15 indeksi ve geçme sıklıkları.	25
3.8: Sistemin eğitiminde kullanılan Hiper Parametreler.	34
4.1: Haber veri kümelerinin çapraz doğrulamanın her bir aşamasında eğitim ve test için kullanılan yaklaşık haber sayıları.	39
4.2: Deneyler esnasında geliştirilen modeller.	40
4.3: Haber Veri Kümesi A üzerinde gerçekleştirilen dört deneye ait sonuçlar.	40
4.4: Haber Veri Kümesi B üzerinde gerçekleştirilen dört deneye ait sonuçlar.	40
4.5: Haber Veri Kümesi A'ya ait örnek bir haber metninin dört farklı yöntem ile elde edilmiş deney sonuçları.	41
4.6: Haber Veri Kümesi B'ye ait örnek bir haber metninin dört farklı yöntem ile elde edilmiş deney sonuçları.	42
4.7: Haber Veri Kümesi B'de yer alan her bir kategori için Deney 4 kombinasyonu ile elde edilen başarı oranları.	43
4.8: Haber Veri Kümesi A için Deney 4 kombinasyonunda sözlük boyutunun başarı oranlarına etkisi.	44
4.9: Haber Veri Kümesi B için Deney 4 kombinasyonunda sözlük boyutunun başarı oranlarına etkisi.	44
4.10: Haber Veri Kümesi A için en fazla haber başlığı kelime sayısının başarı oranlarına etkisi.	45
4.11: Haber Veri Kümesi B için en fazla haber başlığı kelime sayısının başarı oranlarına etkisi.	45

4.12:	Haber Veri Kümesi A için en fazla haber metni kelime sayısının başarı oranlarına etkisi.	46
4.13:	Haber Veri Kümesi B için en fazla haber metni kelime sayısının başarı oranlarına etkisi.	46
4.14:	Haber Veri Kümesi A için ıřın araması algoritmasının ıřın geniřliđinin başarı oranlarına etkisi.	47
4.15:	Haber Veri Kümesi B için ıřın araması algoritmasının ıřın geniřliđinin başarı oranlarına etkisi.	47
4.16:	Haber Veri Kümesi A için deneyler esnasında haber bařlıđı kelimelerinin ekli kullanılma durumunun başarı oranlarına etkisi.	48
4.17:	Haber Veri Kümesi B için deneyler esnasında haber bařlıđı kelimelerinin ekli kullanılma durumunun başarı oranlarına etkisi.	48

1. GİRİŞ

Günümüzde, internet üzerinden erişilebilen veri miktarı her geçen gün artmaktadır. İnternet sayfaları, haberler, makaleler, bloglar, sosyal medya güncellemeleri ve çok daha fazlası yapısal veri değildir [1]. Dolayısıyla insanların sadece ilgilendiği veriye hızlı bir şekilde ulaşması büyük önem arz etmektedir. Tam da bu noktada otomatik metin özetleme sistemleri devreye girmektedir. Otomatik metin özeti oluşturma girişimleri yaklaşık 50 yıl önce başlamıştır [2]. Ancak son yıllarda teknolojinin hızlı bir şekilde gelişmesi ile birlikte internet üzerinde yer alan verinin çok yüksek miktarda artmasıyla, otomatik metin özetleme sistemlerine ihtiyaç artmış ve bu alanda sıklıkla çalışılmaya başlanmıştır [3] [4].

Metin özetleme, bir metin belgesinin en önemli noktalarını içerecek şekilde metnin kısaltılması işlemidir. Otomatik metin özetleme ise bu işlemin bilgisayarlar tarafından yapılmasıdır. İnsanlar genellikle bu tür görevlerde başarılıdırlar. Bu nedenle otomatik olarak oluşturulan metin özetlerinin görevi, elde edilen özetlerin insanlar tarafından yazılanlar kadar iyi olmasını sağlamaktır. Bir başka tanıma göre ise otomatik metin özetleme, temel bilgi içeriğini ve genel anlamı korurken, kısa ve akıcı bir özet üretme görevidir [5]. Bu tanıma göre özetlemenin ana fikri, tüm kümenin bilgisini içeren bir alt küme bulmaktır.

Özetleme her alanda olabilir, örneğin belge özetlemesi, en bilgilendirici cümleleri bularak tüm belgenin temsili bir özetini oluştururken, görüntü özetlemesinde sistem en göze çarpan görüntüleri yani tüm görüntüyü temsil edecek en iyi görüntüleri bulur.

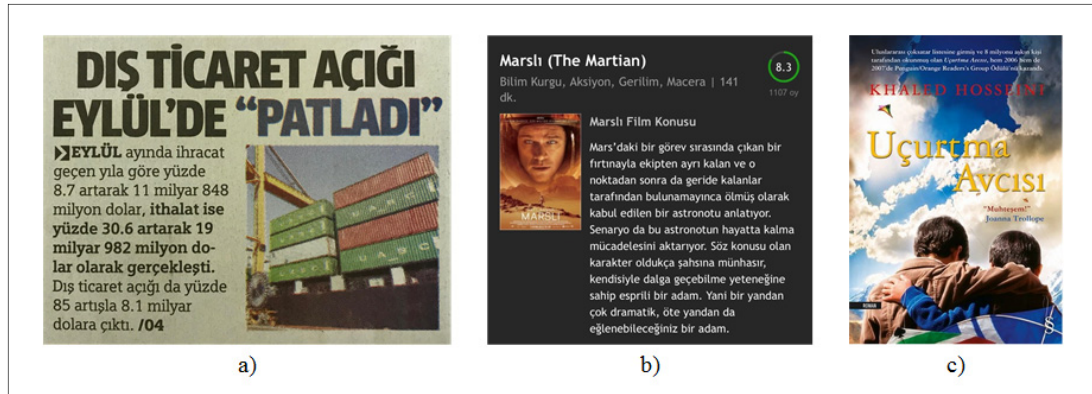
Torres-Moreno otomatik metin özetlemeye neden ihtiyaç duyulduğuna dair 6 madde sunmaktadır [1]:

- i) Özetler okuma süresini kısaltır.
- ii) Özetler belgeleri araştırırken seçme süresini kısaltır.
- iii) Otomatik özetleme indekslemenin etkinliğini artırır.
- iv) Otomatik özetleme algoritmaları, insan tarafından özetlenmiş sistemlere göre daha az önyargılıdır.
- v) Kişiselleştirilmiş özetler, kişiselleştirilmiş bilgiler sağladıkları için soru cevaplama sistemlerinde kullanışlıdır.

vi) Otomatik veya yarı otomatik özetleme sistemlerinin kullanılması, ticari hizmetlerin işleyebilecekleri metin sayısını artırmalarını sağlar.

Anlaşılabileceği üzere metni özetlemek okuma süresini kısaltır ve kişinin konu hakkında daha fazla bilgi edinmek isteyip istememesine karar vermesine büyük ölçüde yardımcı olur.

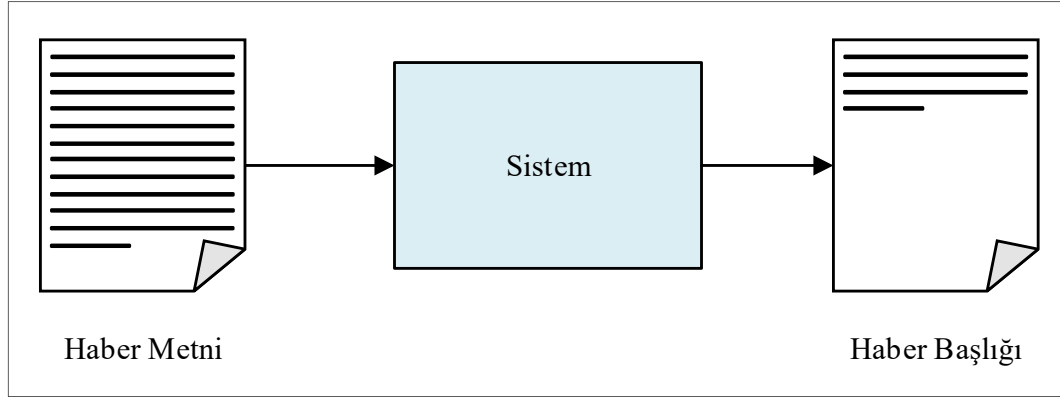
Haber başlıkları, film ön izlemeleri, bir kitap, CD veya film hakkında yapılan yorumlar, özgeçmişler veya biyografiler, hava durumu tahminleri, borsa raporları ve benzerleri doğal dil işleme alanında zor bir problem olan metin özetleme için gündelik yaşamımızdaki kullanım alanlarına örnek olarak verilebilir [6]. Örnek kullanım alanlarından birkaçı Şekil 1.1’de gösterilmiştir.



Şekil 1.1: Metin özetleme kullanım alanları. a) Haber başlığı örneği b) Film özeti örneği c) Kitap hakkında yorum örneği.

Doğal dil işleme alanında metin özetleme, bir metnin birden fazla özeti olabilmesi sebebiyle zor ve karmaşık bir problemdir. Hali hazırda zor bir problem olan metin özetleme çalışmalarında, uzun metin özetleri elde etmek için oldukça iyi donanımlara sahip güçlü makinelere ve zamana ihtiyaç vardır. Günümüzde uzun metin özetleri üretmek yerine daha basit bir problem olan haber manşetlerinin, yani haber başlıklarının üretimi üzerinde çalışmalar yoğunlaşmaktadır [7]. Başlık, aynı zamanda bir yazının en önemli bilgisini içerir ve içeriği hakkında en önemli noktayı vurgulayan yerdir. Kullanıcılar haber sitelerinde gezinirken, bir haberin detay sayfasını açıp açmama konusunda ilk olarak haberin başlığına bakarak karar verir. Bu nedenle haber başlıkları, haber verilerinin sunulduğu sitelerin ziyaretçi sayısını artırması konusunda büyük önem arz etmektedir. Dolayısıyla metin özetleme sistemleri haber başlıklarının üretiminde doğrudan kullanılabilir.

Bu tez çalışmasında, Türkçe haber metinlerinden haber başlıklarını üreten yoruma dayalı metin özetleme sistemi geliştirilmiştir [8]. Sistem Şekil 1.2’de gösterildiği gibi haber metnini girdi olarak almakta ve haberin başlığını çıktı olarak üretmektedir.



Şekil 1.2: Sistemin çalışma yapısı.

Haberlerin başlıkları üretilirken, öncelikle haber metinleri ön işlemden geçirilmiştir. Haber metinleri ön işlemden geçirilirken, kullanılan Türkçe haber veri kümelerinde tekrar eden haber metinleri ve haber başlıkları kaldırılmış, aynı manaya gelmesi için haberlerde geçen tüm sayılar ‘#’ karakteri ile yer değiştirilmiştir. Sonrasında büyük harf-küçük harf duyarlılığını ortadan kaldırmak için tüm haberler küçük harflere çevrilmiştir. Haber metinleri ön işlemden geçirildikten sonra, haberlerdeki benzeri olmayan tüm kelimeler için bir sözlük elde edilmiştir. Daha sonra sözlükte yer alan tüm kelimelerin vektör temsilinin elde edilmesi için Türkçe dili dâhil olmak üzere birçok dile ait önceden eğitilmiş kelime gömmelerini içeren FastText kütüphanesinden [9] yararlanılmıştır.

Sistemin eğitim aşamasında, derin öğrenme modellerinden literatürde ilk başta metin çeviri probleminde kullanılan kodlayıcı-kod çözücü modeli [10] kullanılmıştır. Kodlayıcı-kod çözücü modelinin, dizi tahmin problemlerinden metin çevirisi, soru cevaplama, ses tanıma ve metin özetleme gibi problemlere yüksek başarımla sahip sonuçlar ürettiği gözlemlenmiştir [11]. Kodlayıcı tarafında bir RNN [12] çeşidi olan Çift Yönlü LSTM ağı [13] ve kod çözücü tarafında ise yine bir RNN çeşidi olan LSTM ağı [14] kullanılmıştır. Metin özeti oluşturulurken bir önceki ve bir sonraki kelimeyi veya cümleyi tahmin etme konusunda başarılı olan Çift Yönlü LSTM ağları, dizi tahmin problemlerinde yaygın olarak kullanılmaktadır [15].

Sistemin test aşamasında, haber başlığının çıkarımı için kod çözücü olarak hem açgözlü arama [16] hem de ışın araması [17] kullanılmış ve sonuçlarına göre karşılaştırma yapılmıştır.

Veri kümesi olarak internet üzerindeki farklı haber sitelerinden toplanan Türkçe haber veri kümesi [18] ve Kemik Doğal Dil İşleme grubunun [19] metin sınıflandırma problemi için hazırlanmış olduğu Türkçe haber veri kümesi [20] kullanılmıştır. Eğitim ve test kümeleri hazırlanırken çapraz doğrulama yönteminden faydalanılmıştır.

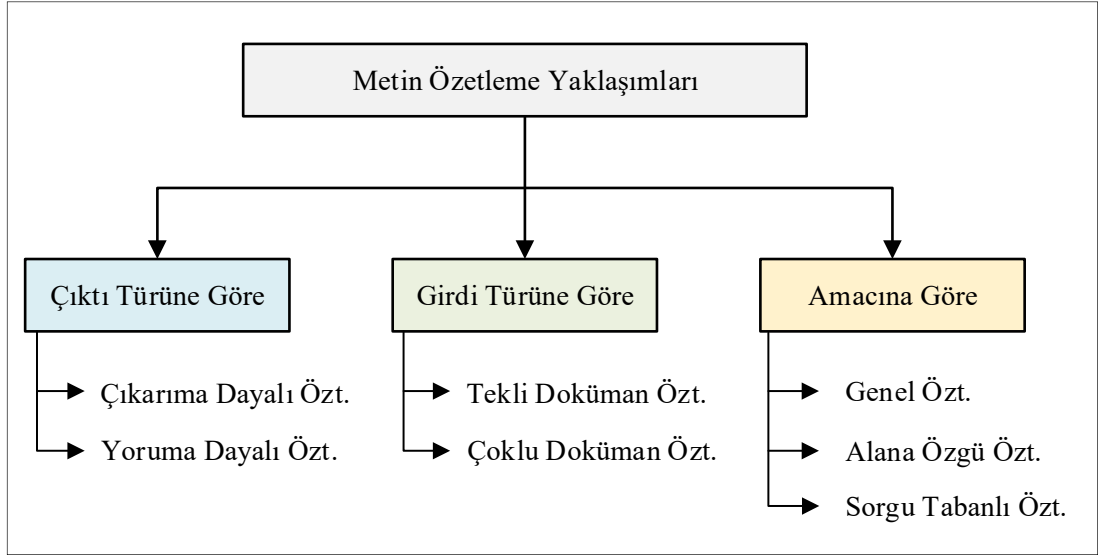
İki farklı veri kümesinin eğitimleri sırasında kodlayıcı tarafında Çift Yönlü LSTM modeli ve kod çözücü tarafında LSTM modeli kullanılmıştır. Test aşamasında ise kod çözücü olarak hem açgözlü arama hem de ışın araması denenmiştir. Bu iki deneye dikkat mekanizması da eklenerek dikkat mekanizmasının başarıya etkisi incelenmiştir. Daha sonra farklı sözlük boyutları, haber başlığı uzunlukları, haber metni uzunlukları ve ışın araması için ışın genişlikleri ile ayrı ayrı deneyler gerçekleştirilerek bu parametrelerin başarıya olan etkisi ölçülmüştür.

Sistem başarısı ölçülürken, metin özetleme alanında yaygın bir şekilde kullanılan ROUGE [21] değerlendirme metriği kullanılmıştır. Bu metriğe ek olarak tam doğru haber başlığı ile sistemin tahmin ettiği haber başlığı arasındaki anlamsal benzerlik [22] skoru da hesaplanmıştır. Anlamsal benzerlik skor hesabı yapılırken, FastText kütüphanesinde yer alan cümlelerin vektör karşılıklarından [23] faydalanılmıştır.

Bölüm 2’de metin özetleme alanında literatürde yer alan çalışmalar hakkında bilgi verilmiştir. Bölüm 3’te bu tez çalışmasında haber başlıkları üretilmesi için geliştirilen yöntemlerden, Bölüm 4’te elde edilen deney sonuçlarından bahsedilmiştir. Sonuçlar ve tartışmalar ise Bölüm 5’te yer almaktadır.

2. İLGİLİ ÇALIŞMALAR

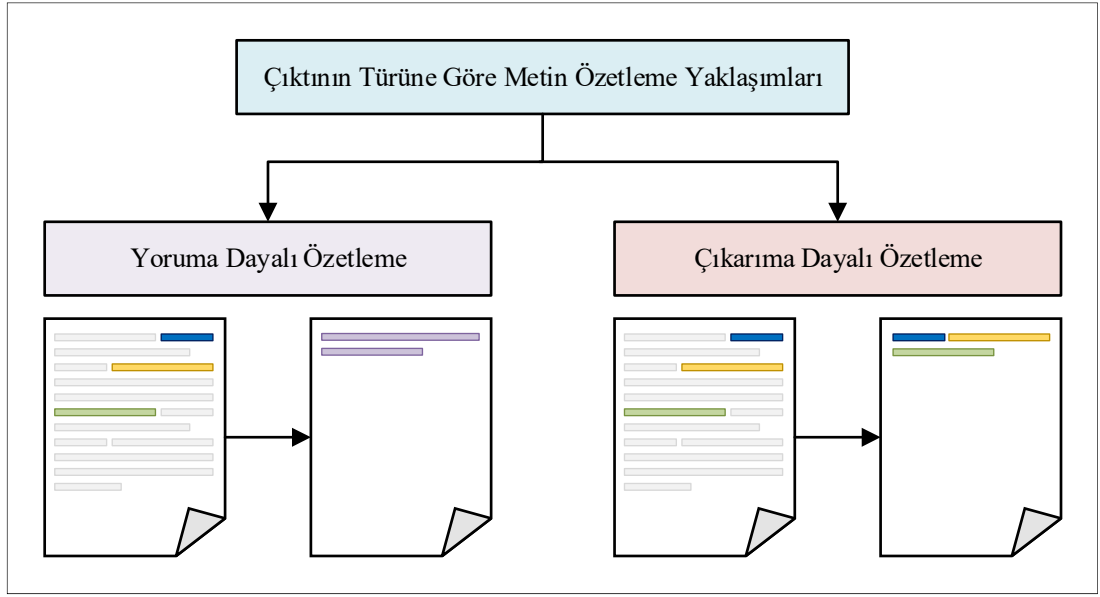
Otomatik metin özetleme yaklaşımları çıktı türüne göre, girdi türüne göre ve amacına göre olmak üzere 3 farklı şekilde kategorize edilebilir [24]. Çıktı türüne göre metin özetleme yaklaşımlarında, elde edilen özetin türüne göre çıkarıma ya da yoruma dayalı özetler üretilmektedir. Girdi türüne göre metin özetleme yaklaşımlarında, özet elde edilirken tekli ya da çoklu dokümanlar eğitilerek özetler üretilmektedir. Amacına göre özetleme yaklaşımlarında ise, özetlenecek dokümanın konusuna göre genel, alana özgü ve sorgu tabanlı özetler üretilmektedir. Metin özetleme yaklaşımları Şekil 2.1’de sunulmuştur.



Şekil 2.1: Metin özetleme yaklaşımları.

2.1. Çıktı Türüne Göre Metin Özetleme

Literatürde çıktı türüne göre metin özetleme çalışmaları çıkarıma dayalı ve yoruma dayalı olmak üzere ikiye ayrılmıştır [6]. Yoruma dayalı ve çıkarıma dayalı metin özetleme sistemleri Şekil 2.2’de görsel olarak anlatılmıştır.



Şekil 2.2: Çıktı türüne göre metin özetleme yaklaşımlarının görselleştirilmesi.

2.1.1. Çıkarıma Dayalı Metin Özetleme Çalışmaları

Çıkarıma dayalı metin özetleme sistemlerinde genellikle kelimeler, kelime grupları veya cümleler orijinal metinden herhangi bir değişiklik yapılmadan direkt olarak seçilerek özet oluşturulur [25] [26] [27] [28] [29] [30].

Bu yaklaşımda genel olarak, ilk aşamada kelimelere veya cümlelere belirli ağırlıklar verilir. Daha sonra farklı algoritmalar kullanılarak kelimeler veya cümleler birbirleri arasındaki önem ve benzerliğe göre sıralanır. Bu sıralama sonrasında ilk sıralarda geçen kelimeler veya cümleler alınarak metin özetleri elde edilir.

Çıkarıma dayalı özetleme çalışmalarında hem kelimenin hem de cümlenin özellikleri kullanılmıştır. Özete dâhil edilecek kelime veya cümleler seçilirken tartışılan bazı özellikler aşağıdaki gibidir [31] [32] [33]:

- Anahtar kelimeler, cümlenin öneminin belirlenmesinde büyük ölçüde katkı sağlar. Anahtar kelimelerden oluşan cümle, muhtemelen özete dâhil edilmiştir. Anahtar kelimeler tf-idf ölçüsüne göre belirlenir.
- Başlıkta geçen kelimeleri içeren cümle, belgenin temasını gösterdiği için nihai özette bulunma şansı yüksektir.
- Büyük harfle yazılmış kelimeler, nihai özet için cümle seçiminde önemli kelimeler olarak kabul edilir.

- İşaret grupları, belgenin akış yapısını gösteren kelimelerdir ve cümle seçiminde bir özellik olarak kullanılır. “Çünkü”, “Özetle”, “Sonuç olarak” gibi gruplar işaret gruplarına örnek olarak gösterilebilir.
- Cümlelerin konumu, özete dâhil edilecek cümlelerde etkilidir. Belgenin başındaki ya da sonundaki cümleler genellikle önemlidir.
- Cümlelerin uzunluğu, anahtar cümlelerin belirlenmesinde önemli rol oynar. Çok kısa cümleler ya da çok uzun cümleler genellikle özette yer almaz.
- Cümlelerin konumunun önemli olması gibi paragrafların konumu da cümlelerin seçiminde büyük ölçüde katkı sağlar. Mesela belgelerin başlarındaki ve sonlarındaki paragraflara daha yüksek ağırlıklar verilir.

• Çizge Tabanlı Çalışmalar

Çizgelerin dokümanları etkili bir şekilde temsil etmesi sebebiyle, çizge tabanlı modeller, doküman özetleme sistemlerinde yaygın olarak kullanılmaktadır. Sankarasubramaniam vd. Wikipedia’dan [34] elde ettiği dış bilgileri kullanarak ikili çizge çerçevesi yöntemi ile çıkarıma dayalı metin özetleme tekniği uygulamıştır [33]. Bu çalışmada önemli cümleleri seçme konusunda verimli çalışan artırılmış bir sıralama algoritması (HITS algoritmasının [35] bir çeşidi) önerilmiş ve sonuçta elde edilen özetin tutarlı olduğu görülmüştür. Bu çalışmayı diğer çalışmalardan ayıran özelliği, özetlemeyi yaparken çizge tabanlı ve kavram tabanlı yaklaşımı birlikte kullanmasıdır. Bir başka çizge tabanlı çalışma olan LexRank [36] çalışmasında ise, cümlelerin önemi öz vektörlerin merkezilik bilgisine dayanılarak belirlenir. Bu çalışmada, doküman üzerindeki cümleler bir çizge olarak ifade edilir ve cümleler arasında oluşan kenarlar ağırlıklı kosinüs benzerlik değerlerini gösterir. Cümleler benzerlik değerlerine göre gruplara ayrıldıktan sonra, PageRank algoritmasına [37] benzer şekilde LexRank skorlarına göre sıralanırlar. Önerilen bu metot diğerlerine göre nispeten daha yüksek başarımla sahipsizdir.

• LSA Tabanlı Çalışmalar

LSA yöntemi anahtar kelimeler ve bu kelimeler ile ilgili olan terimler arasındaki ilişkiyi tespit etmek için kullanılan gözetimsiz bir öğrenme yaklaşımıdır [38]. LSA algoritması çıkarıma dayalı metin özetleme çalışmalarında sıklıkla kullanılmaktadır. Bu yöntemde ilk aşamada anahtar kelimeler elde edilir. Daha sonra elde edilen anahtar kelimeler, özete katkısı olmayan cümlelerin filtrelenmesi için kullanılır. Sonrasında

ise anahtar kelimeleri içeren cümlelerden gizli anlam indeksleme metodu ve kümeleme ile sanal merkeze dayalı metot kullanılması ile özete dâhil edilecek cümleler seçilir. Ozsoy vd. çalışmalarında LSA tabanlı algoritmaları anlatmış, İngilizce ve Türkçe metinler için bu algoritmayı kullanarak çıkarıma dayalı metin özetleme sistemi gerçekleştirmişlerdir [39].

- Bulanık Mantık Tabanlı Çalışmalar

Bulanık mantık tabanlı yaklaşımlar temelde dört tane bileşenden oluşur: durulaştırıcı, bulanıklaştırıcı, bulanık bilgi tabanı ve sonuç motoru. Bulanık mantık yaklaşımında metinsel özellik olarak cümle uzunluğu, cümle benzerliği gibi bilgiler seçilerek bulanık sisteme girdi olarak verilirler [40] [41].

Ladda Suanmali vd. bulanık mantık yaklaşımını otomatik metin özetleme için önermişlerdir [42]. Metin dokümanı öznitelik çıkarımından (başlık öznitelikleri, cümle uzunluğu, cümle konumu, cümleler arası benzerlik, terim ağırlığı, özel isimler ve sayısal veriler) sonra bir ön işleminden geçirilmiştir. Metin özeti, derecelendirilmiş olan cümlelerin sıralanması ile elde edilmiştir. Bu yöntemde özetlemenin kalitesinin arttığı gözlemlenmiştir.

- İçerik Tabanlı Çalışmalar

İçerik tabanlı yaklaşımlarda, içerikler HowNet [43] ve Wikipedia [33] gibi harici bilgi sistemlerinden elde edilirler. Önerilen yöntemde [43] cümlelerin önemi kelimeler yerine HowNet üzerinden elde edilen içeriklere bağlı olarak hesaplanır. Kabaca bir özetleme ve benzerlik ölçüleri elde etmek için oluşturulan içeriksel vektör modeli, nihai özetteki fazlalık kısımları azaltmak için cümleler arasında hesaplanmıştır. İyi bir özetleme sistemi geniş kapsamlı olmaya ve fazlalıkları azaltmaya odaklanır. Ramanathan vd. çizge yapısını kullanarak özet üreten Wikipedia tabanlı bir özetleme sistemi önermiştir [33]. Önerilen metot, her cümle için içerikleri elde ederken Wikipedia'yı kullanır ve çizge tabanlı özetlemede bahsedilen cümle-içerik çizgesini oluşturur.

İçerik tabanlı özetleme sistemlerinde yer alan temel adımları şunlardır:

- i) Verilen bir metin için içerikleri harici bilgi sistemlerinden (HowNet, WordNet, Wikipedia gibi) elde edilmesi

- ii) İçerik ve cümleler arasındaki ilişkiyi gösterebilmek için içeriksel bir vektör veya çizge modeli oluşturulması
- iii) Cümleleri skorlamak için derecelendirme algoritmasının uygulanması
- iv) Elde edilen skor değerlerine göre metin özetinin oluşturulması

- Bayes Kuralı Tabanlı Makine Öğrenme Çalışmaları

Bu yaklaşımda eğitilecek dokümanlar, özetleri ile birlikte eğitim için modele girdi olarak verilirler. Makine öğrenmesi yaklaşımı, metin özetleme için öğrenmeyi sınıflandırma üzerine yapar. Bu tür yöntemlerde cümlelerden edinilen özniteliklere göre cümleler, özette yer alacak ya da almayacak şekilde sınıflandırılırlar. Sınıflandırmanın olasılığı şu Bayes kuralına [44] göre öğrenilmektedir: $P(s \in S | f_1, f_1, f_1, \dots, f_n)$. Burada s dokümanda yer alan cümleleri, f_i sınıflandırma aşamasında kullanılan öznitelikleri ve S ise özette yer alan cümleleri göstermektedir. P değeri de cümlelerin özniteliklerine göre özette yer alıp almayacağının olasılığını ifade eder.

- Sinir Ağı Tabanlı Çalışmalar

Svore vd. tarafından önerilen yaklaşımda [45], dokümanda yer alan önemli cümleler RankNet algoritması kullanılarak tespit edilmiştir. Bu yöntemde, geri yayımlı iki katmanlı sını ağı RankNet algoritması kullanılarak eğitilmiştir. İlk adımda bir makine öğrenmesi yaklaşımı kullanılarak eğitim veri kümesi etiketlenir ve eğitim ile test aşamasında kullanılacak cümlelerin öznitelikleri çıkarılır. Daha sonra eğitim veri kümesi dokümanda yer alan cümleleri derecelendirmek için sinir ağına girdi olarak verilir. Bir başka yaklaşımda [46] ise özete dâhil edilebilir veya edilemez cümlelerin karakteristiği üç katmanlı ileri eslemeli sinir ağı sayesinde öğrenilir. Bu yöntemdeki en önemli nokta, iki aşamalı olan öznitelik çıkarımıdır. İlk aşamada çok sık olmayan öznitelikler elenir ve ikinci aşamada ise cümlelerin derecelendirmeleri yapıldıktan sonra sık olan öznitelikler birleştirilerek önemli özet cümleleri tespit edilir. Hingu vd. Wikipedia makalelerini özetlemek için çıkarıma dayalı bir yöntem önermişlerdir [47]. Yöntem, metinlerin özniteliklerini çıkarır ve bir sinir ağı modeli [35] yardımıyla cümleleri skorlandırır. Yöntemde Wikipedia'da yer alan metinler dizgeciklere ayırma ve kök bulma işlemlerinin ardından sisteme girdi olarak verilir. Öznitelik çıkarımından sonra elde edilen skor değerleri sinir ağına girdi olarak verilir,

böylece ağdan çıktı olarak da tek bir skor değeri elde edilir. Sinir ağından çıktı olarak alınan skor bilgisi cümlelerin önemine ait bilgi vermektedir. Bu bilgi kullanılarak metin özeti tahmin edilmiştir.

2.1.2. Yorumaya Dayalı Metin Özetleme Çalışmaları

Yorumaya dayalı özetleme yaklaşımında; göze çarpan fikirleri yakalamak için metindeki kelimeler anlamlandırılarak orijinal metinde yer almayan kelimeler veya cümlelerle özet elde edilir [48]. Literatürde Türkçe dili için derin öğrenme tabanlı yorumaya dayalı metin özetleme çalışması mevcut değildir. Ancak İngilizce için pek çok yorumaya dayalı metin özetleme çalışmaları mevcuttur [49] [50].

Song vd. [49], önerdikleri LSTM ve Konvolüsyonel Sinir Ağı tabanlı model ile yorumaya dayalı metin özetleme sistemi geliştirmişlerdir. Önerilen yöntem iki aşamadan oluşmaktadır. İlk aşamada girdi cümlelerinden kelime grupları belirlenmekte ve ikinci aşamada ise derin öğrenme kullanılarak metin özetleme yapılmaktadır. Yöntem CNN/Daily Mail veri kümesi [51] üzerinde gerçekleştirilen deney sonuçlarına göre literatürde yer alan diğer çalışmalar arasında iyi sonuçlar elde etmiştir.

Chopra vd. [50], çalışmalarında koşullu tekrarlayan sinir ağı (RNN) kullanarak yorumaya dayalı bir metin özetleme sistemi geliştirmişlerdir. Buradaki şartlı kısım, konvolüsyonel dikkat tabanlı bir kodlayıcı tarafından sağlanmaktadır. Bu durum, kod çözücünün özet çıkarımı esnasında, her adımda doğru girdi kelimesine odaklanmasını sağlamaktadır. Önerilen yöntemin deneyleri esnasında Gigaword [52] ve DUC [53] veri kümeleri kullanılmıştır.

Derin öğrenme algoritmalarının popülerlik kazanması ile birlikte, Rush vd. 2015 yılında gerçekledikleri çalışmalarında evrimsel sinir ağlarını kullanarak DUC ve Gigaword veri kümeleri üzerinde yorumaya dayalı metin özetleme sisteminden bahsetmişlerdir [48]. Önerdikleri modelde girdi cümlesini kodlamak için evrimsel sinir ağlarını ve çıktı cümlesini yani özeti oluşturmak için ise ileri beslemeli ağları kullanmışlardır. Önerdikleri yöntemde sadece ilk cümleyi kullanarak öğrenme gerçekleştirmişlerdir.

Nallapati vd. ile Pasunuru vd. önerdikleri yorumaya dayalı özetleme sistemlerinde [11] [54] RNN kullanmışlardır. RNN tabanlı kodlayıcı-kod çözücü mimarileri, kısa girdi ve çıktı dizileri için yorumaya dayalı metin özetleme yönteminde başarılı sonuçlar vermiştir. 2017 yılında Paulus vd. yaptıkları çalışmada, CNN/Daily Mail veri

kümesinde o zamana kadar ki sonuçların üstüne çıkarak üretilen özetlerin okunabilirliğini arttıran ve daha uzun çıktılar elde edebilen bir model öne sürmüşlerdir [55]. Sistem gözetimli öğrenme ile pekiştirici öğrenmeyi birlikte kullanmıştır.

Takase vd. [56], önerdikleri yöntemde dikkat tabanlı bazal bir yapay sinir ağı modeline ek olarak sözdizimsel ve anlamsal bilgileri kullanan yeni bir model geliştirmişlerdir. Yazarlar, sistemin ürettiği başlığın konusunun veya üretilen başlıkta yer alan nesnelerin, girdi cümlesinde de karşılığı olması gerektiğini savunmaktadır. Tree-LSTM modelinin güncellenmiş bir halini kullanarak soyut anlam gösterimi (AMR) ayrıştırıcısından elde edilen sonuçlarla kodlama işlemi gerçekleştirmişlerdir. AMR, bir cümlenin anlamını kodlayan, kökü bulunan, çevrim içermeyen, yönlü bir çeşit çizgedir. Çalışmada veri kümesi olarak DUC ve Gigaword veri kümeleri kullanılmıştır. Önerilen dikkat tabanlı AMR kodlayıcı-kod çözücü modelinin, dikkat tabanlı bazal bir yapay sinir ağı modeli ile kıyaslandığında daha başarılı başlıklar üretebildiği gözlemlenmiştir.

Tan vd. [57], cümle özetleme modellerini genişleterek doküman başlığı üreten bir model üzerinde çalışmışlardır. Dokümanlar pek çok cümleden oluştuğu için, cümle özeti üreten bir model doküman başlığı üretmede oldukça düşük performans göstermektedir. Uzun girdi cümleleri ise eğitimi gittikçe daha da zorlaştırmaktadır. Önerilen çalışmada, doküman özetleme teknikleri ile ilk önce doküman içerisindeki önemli cümleler belirlenmiş ve daha sonra birden fazla cümle özetleyebilen, hiyerarşik dikkate dayalı bir model ile başlık üretilmiştir. Gerçekleştirilen deney sonuçlarına göre, önerilen yöntemin gerçek bir veri kümesi üzerinde büyük ölçüde başarılı sonuçlar verdiği gözlemlenmiştir. Deneyler esnasında New York Times (NYT) haber veri kümesi kullanılmıştır.

Liu vd. [58], pekiştirilmiş öğrenme kullandıkları çalışmalarında hem üretken hem de ayırt edici iki ayrı modeli eş zamanlı olarak eğitmişlerdir. Sistem pekiştirilmiş öğrenme için ajan olarak üretken modeli kullanmaktadır. Üretken model girdi olarak ham metin alarak yoruma dayalı özet tahmin etmektedir. Burada kodlayıcı olarak Çift Yönlü LSTM ve kod çözücü olarak dikkat tabanlı LSTM kullanılmıştır. Ayırt edici model ise, tam doğru özetten bir özet oluşturmaya çalışmaktadır. Ayırt edici model için ikili bir sınıflandırıcı kullanılmıştır. Sınıflandırma işleminin iyileştirilmesi için girdi cümlesi bir Konvolüsyonel Sinir Ağı ile kodlanmıştır. CNN/Daily Mail veri kümesi üzerinde gerçekleştirilen deney sonuçlarına göre sistem literatürde yer alan diğer çalışmalara göre daha başarılı bir performans göstermiştir.

Li vd. [59], önerdikleri yöntemde, dizi tahmin problemleri için kullanılan kodlayıcı-kod çözücü modeline ek olarak Derin Yinelemeli Üretken Kod Çözücü (Deep Recurrent Generative Decoder) kullanan yeni bir yoruma dayalı metin özetleme sistemi geliştirmişlerdir. Özetlerde yer alan gizli yapısal bilgi, yinelemeli bir gizli rastgele model sayesinde öğrenilerek özet kalitesi artırılmıştır. Önerilen yöntem sayesinde özetler yoruma dayalı olarak hem üretken gizli değişkenlerle hem de ayırt edici deterministik durumlarla elde edilmiştir. Gigaword, DUC ve LCSTS [60] veri kümeleri üzerinde gerçekleştirilen deney sonuçlarına göre sistem literatürdeki alternatif yöntemlere göre daha başarılı sonuçlar elde etmiştir.

Çıkarıma dayalı özetleme ile yüksek ROUGE skorları ve yoruma dayalı özetleme ile daha okunabilir özetler elde edilmesinden yola çıkarak, Hsu vd. [61] çalışmalarında hem çıkarıma dayalı hem de yoruma dayalı özetlemeyi birlikte kullanmışlardır. Önerilen yöntemde, cümle seviyesindeki dikkat bilgisi kullanılarak dikkatin daha az olduğu cümlelerdeki kelimelerin üretilen özete daha az dahil edildiği özetler oluşturulmuştur. Ek olarak cümle ve kelime seviyesindeki dikkat arasında oluşan tutarsızlığı cezalandıran bir kayıp fonksiyonu da geliştirilmiştir. Geliştirilen sistemin deneyleri, CNN/Daily Mail veri kümesi üzerinde Amazon Mechanical Turk (MTurk) [62] kullanılarak gerçek insanlar tarafından gerçekleştirilmiştir. Sonuçlara göre sistemin ürettiği özetler insanlar tarafından daha bilgilendirici ve okunaklı olarak bulunmuştur.

İnsanların uzun dokümanları özetleme şekillerinden yola çıkarak, Chen vd. [63] çalışmalarında ilk önce dokümanda yer alan belirgin cümleleri seçen ve daha sonra bu cümlelerden yoruma dayalı özetler çıkaran bir yöntem önermişlerdir. Önerilen yöntem yoruma dayalı özetleme için cümle düzeyinde pekiştirilmiş öğrenme kullanmıştır. CNN/Daily Mail ve DUC veri kümeleri üzerinde gerçekleştirilen deney sonuçlarına göre sistem literatüre göre başarılı sonuçlar üretmesinin yanı sıra, daha hızlı eğitim ve test sürelerine de erişmiştir.

2.2. Girdi Türüne Göre Metin Özetleme

Girdi türüne göre metin özetleme çalışmaları, tekli doküman özetleme ve çoklu doküman özetleme olmak üzere ikiye ayrılmaktadır.

2.2.1. Tekli Doküman Özetleme Çalışmaları

Tekli doküman özetleme çalışmalarında, sistem girdi olarak tek bir doküman alır ve çıktı olarak metin özetini üretir. Litvak vd. önerdikleri yöntemde, çizge tabanlı bir yaklaşım ile çıkarıma dayalı tekli doküman özetleme sistemleri için anahtar kelime çıkarımı yapmaktadır [64]. Önerilen çalışma hem gözetimli hem de gözetimsiz olarak çalışabilmektedir. Jagadeesh vd. gerçekleştirdikleri çalışmalarında cümle çıkarımına dayalı otomatik tekli doküman özetleme sistemi geliştirmişlerdir [65]. Yöntem metindeki önemli cümlelerin tanımlanmasına ve çıkarılmasına dayanarak çalışır. Özet çıkarımı için cümlelere ait özellikler çıkarılmış ve vektörel olarak ifade edilmiştir. Ayrıca cümleleri, çıkarılan özelliklerine göre sıralayan bir algoritma geliştirmişlerdir. Önerilen bu yöntem, literatürdeki diğer yöntemlere göre daha uyumlu ve okunabilir özetler üretebilmiştir. Gerçekleştirilen deneylerde cümlelere ait elde edilen özniteliklerin dokümandaki önemli bilgileri çıkarma konusunda önemli ölçüde yardımcı olduğu anlaşılmıştır.

2.2.2. Çoklu Doküman Özetleme Çalışmaları

Diğer bir yöntem olan çoklu metin özetleme çalışmalarında ise sistem birden fazla dokümanı girdi olarak alarak metin özeti üretir. Buradaki birden fazla doküman genellikle aynı konu ile alakalı olmaktadır. Metin özetleme sistemlerinin ilk çalışmaları, çoğunlukla tekli doküman özetleme ile ilgilenmiştir. Goldstein vd. gerçekleştirdikleri çalışmalarında, çıkarıma dayalı olarak tekli doküman özetleme metotları üzerine kurulmuş, dokümana ait ek bilgileri ve dokümanlar arası ilişkiyi de kullanarak çoklu doküman özetleme sistemi geliştirmişlerdir [66]. Genel olarak çoklu doküman özetleme sistemlerinde yaşanan sıkıştırma, hız, fazlalık ve yazı bloklarının seçilmesi gibi problemlerin üzerine yoğunlaşmışlardır. Bu tür problemleri alandan bağımsız teknikler kullanarak istatistiksel yöntemlerle çözmüşlerdir.

2.3. Amacına Göre Metin Özetleme

Amacına göre metin özetleme çalışmaları genel, alana özgü ve sorgu tabanlı olmak üzere üçe ayrılmaktadır.

2.3.1. Genel Metin Özetleme Çalışmaları

Genel metin özetleme çalışmalarında amaç, tüm metni konusundan ya da alanından bağımsız olarak özetlemektir. Yani dokümanla ilgili herhangi bir önceden kabullenme yapılmamakta ve tüm metinler homojen olarak kabul edilmektedir. Kutlu vd. gerçekleştirdikleri çalışmalarında çıkarıma dayalı Türkçe metin özetlerini oluşturan bir yöntem önermişlerdir [67]. Yöntemde ilk önce cümleler skorlanarak derecelendirilmiştir. Cümle skorları yüzey seviyesi öznitelikleri kullanılarak hesaplanmıştır. Sistem en yüksek dereceye sahip olan cümleleri seçerek metin özetini oluşturmuştur.

2.3.2. Alana Özgü Metin Özetleme Çalışmaları

Alana özgü metin özetlemede, sadece belirli bir konu ile alakalı metinler seçilerek özet oluşturulmaktadır. Bu sayede alternatif yöntemlere kıyasla daha gürbüz özetler elde edilmektedir. Mesela sadece biyomedikal belgelerin özetlenmesi veya e-postaların özetlenmesi [68] bu tür çalışmalara örnek olarak gösterilebilir.

2.3.3. Sorgu Tabanlı Metin Özetleme Çalışmaları

Metin belgelerinin bir sorguya ilişkin özetini oluşturan sistemler sorgu tabanlı metin özetleme sistemleri olarak tanımlanmaktadır. Bu tür sorgular, genellikle doğal dil soruları veya belirli bir konuyla ilgili anahtar kelimelerdir. Örneğin; arama motorları tarafından arama sonucunda bulunan kısımları vurgulama işlemi sorgu tabanlı uygulamalara bir örnektir.

3. HABER BAŞLIKLARININ ÜRETİLMESİ

Haber başlıklarının üretilmesi için ilk önce tez çalışması kapsamında kullanılan iki farklı haber veri kümesi ön işleminden geçirilerek başlık üretimine uygun hale getirilmiştir. İkinci aşamada ise ön işleminden geçirilmiş olan veri kümelerinin sözlükleri çıkarılmıştır. Çıkarılan sözlük veri kümelerinde yer alan eşsiz kelimelerin indekslenmesi ile oluşturulmuştur. Daha sonra sözlükte yer alan kelimeleri vektörel hale getirebilmek için kelime gömmeleri çıkarılmıştır. Kelime gömmeleri çıkarıldıktan sonra sistemin eğitimi gerçekleştirilmiş ve eğitilen model sayesinde haber başlıkları tahmin edilmiştir.

3.1. Haber Veri Kümeleri

Literatürde haber metinlerini ve haber başlıklarını içeren açık kaynaklı Türkçe veri kümesi bulunmamaktadır. Bu sebeple tez çalışmasında internet üzerinden elde edilen Haber Veri Kümesi A ve metin sınıflandırma problemi için hazırlanmış olan Haber Veri Kümesi B olmak üzere iki farklı Türkçe haber veri kümesi kullanılmıştır.

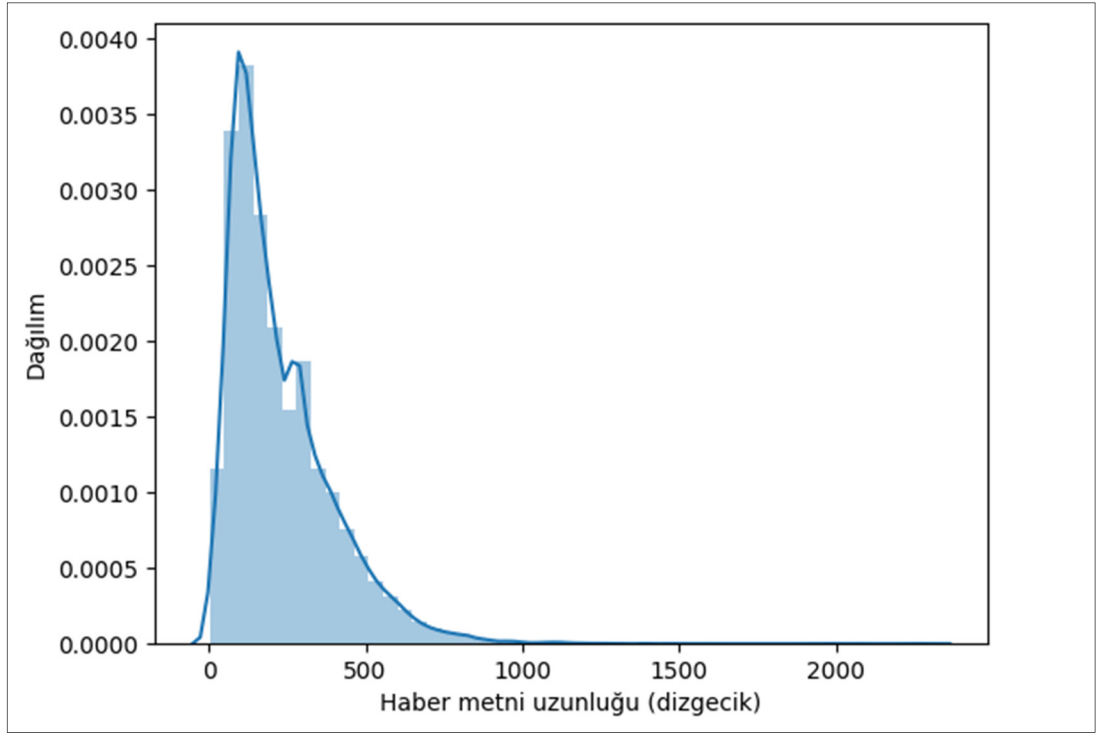
3.1.1. Haber Veri Kümesi A

Haber Veri Kümesi A, internet üzerinden çeşitli haber web siteleri kullanılarak elde edilmiş bir veri kümesidir [18]. Veri kümesi 8 Şubat 2019 ile 11 Mart 2019 tarihleri arasında toplanmış haberleri içermektedir. Toplamda 31 farklı web sitesinden elde edilen veri kümesinde toplamda 72,110 tane Türkçe haber bulunmaktadır. Veri kümesindeki haber kaynaklarının istatistikleri Tablo 3.1’de verilmiştir. İstatistiklere göre en çok haber 14,049 haber ile “haberturk.com” sitesinden ve en az haber de 27 haber ile “evrensel.net” sitesinden alınmıştır.

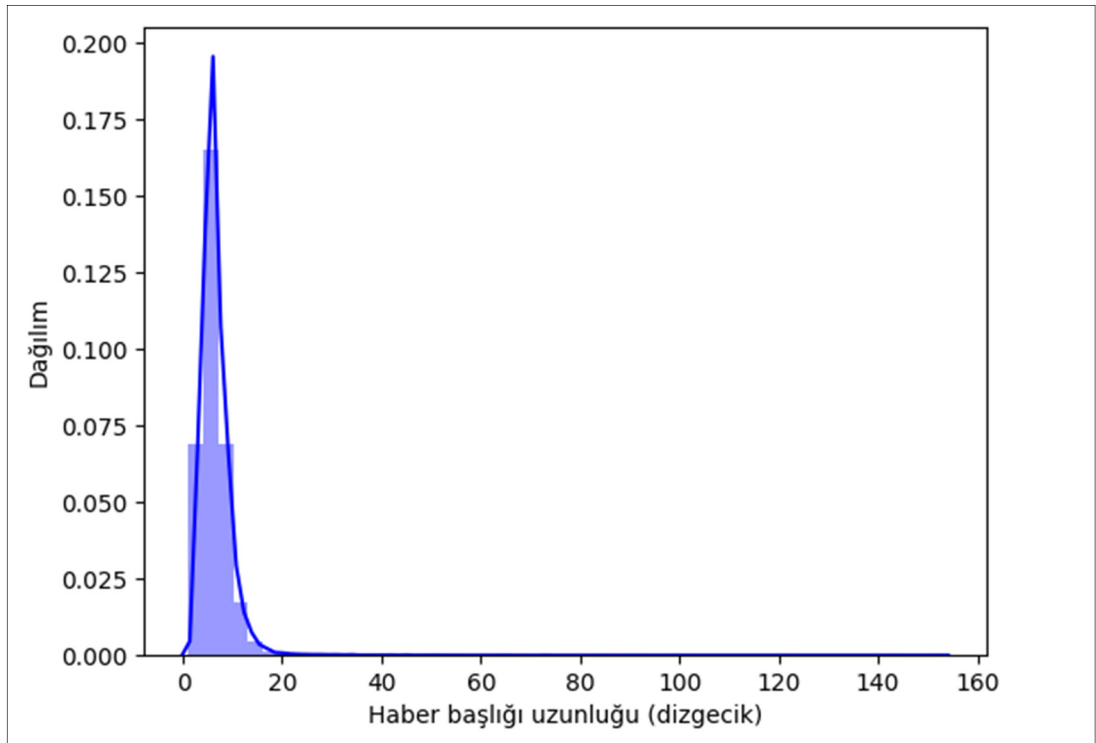
Tablo 3.1: Haber Veri Kümesi A'nın elde edildiği web siteleri ve haber sayıları.

Web Sitesi	Haber Sayısı	Web Sitesi	Haber Sayısı
haberturk.com	14049	medyatava.com	417
haber7.com	7100	birgun.net	391
hurriyet.com.tr	6742	milliyet.com.tr	372
turkiyegazetesi.com.tr	6392	posta.com.tr	299
sabah.com.tr	6072	odativ.com	296
gercekgundem.com	5992	fanatik.com.tr	123
cumhuriyet.com.tr	5523	internethaber.com	102
ntv.com.tr	4446	aksam.com.tr	95
fotomac.com.tr	2947	gunes.com	94
ahaber.com.tr	2912	t24.com.tr	78
trthaber.com	1841	nayn.co	66
takvim.com.tr	1641	tgrthaber.com.tr	58
sozcu.com.tr	1390	ntvspor.net	57
diken.com.tr	1059	abcgazetesi.com	33
cnnturk.com	882	evrensel.net	27
yenisafak.com	613		

Haber Veri Kümesi A'nın haber metinleri için kelime sayısı frekansına göre dağılımlarını gösteren grafik Şekil 3.1'de ve haber başlıkları için kelime sayısı frekansına göre dağılımlarını gösteren grafik Şekil 3.2'de verilmiştir. Elde edilen dağılımlara göre haber başlıkları en sık 6 tane kelimedenden ve haber metinleri en sık 280 tane kelimedenden oluşmaktadır. Haber Veri Kümesi A'da yer alan haber başlıkları ortalama 7 kelime, haber metinleri ise ortalama 228 kelimedenden oluşmaktadır.



Şekil 3.1: Haber Veri Kümesi A haber metni dizgecik sayısı dağılımları.



Şekil 3.2: Haber Veri Kümesi A haber başlığı dizgecik sayısı dağılımları.

Haber Veri Kümesi A’da yer alan web sitelerinden en çok haberin yer aldığı ilk 13 web sitesi, her web sitesinden örnek bir haber başlığı ve kısaltılmış örnek bir haber metni Tablo 3.2’de gösterilmiştir.

Tablo 3.2: Haber Veri Kümesi A'ya ait en çok haber içeren ilk 13 web sitesi, web sitelerine ait örnek haber başlıkları ve kısaltılmış örnek haber metinleri.

Web Sitesi	Örnek Haber Başlığı	Kısaltılmış Örnek Haber Metni
haberturk.com	Van Persieden hat-trick	Fenerbahçede bekleneni veremeyen Robin van Persie, Feyenoordta gollerine devam ediyor. Van Persie,(...)
haber7.com	Muğlada korkutan deprem	Muğlanın Dalaman ilçesinde 4,0 büyüklüğünde deprem meydana geldi. Afet ve Acil Durum Yönetimi (...)
hurriyet.com.tr	İhracata 8 milyar dolar garanti	İhracatta rekor üstüne rekor kırılırken, Eximbanktan sonra ihracatçıya bir destek de faktoring şirketlerinden (...)
turkiyegazetesi.com.tr	Japonyada deprem	Japonyanın kuzeyindeki Hokkaido Adasında 5,7 büyüklüğünde deprem olduğu bildirildi. Japon haber ajansı(...)
sabah.com.tr	İşte yeni askerlik sistemi	Yeni askerlik sisteminde de bedelli sürekli hale getirilecek. Bedelliden yararlananlar bir aylık zorunlu (...)
gercekgundem.com	Silivride fabrika yangını	İstanbulun Silivri ilçesinde bir otomotiv yedek parça fabrikasında yangın çıktı. Yangın, itfaiye ekiplerinin (...)
cumhuriyet.com.tr	Fuel Oile zam geliyor	Enerji Petrol Gaz İkmal İstasyonları İşveren Sendikası, kalorifer yakıtı, Fuel Oil ve Yüksek Kükürlü Fuel Oile (...)
ntv.com.tr	Kartalda riskli binaların yıkımı sürüyor	Anadolu Ajansı çöken Yeşilyurt Apartmanının bitişiğindeki 10 katlı Yunus Apartmanının yıkım (...)
fotomac.com.tr	Gaziantep'te puanlar paylaşıldı	TFF 1.Lig Haberleri Gaziantep Spor Toto 1inci Ligin 23üncü haftasında Gazişehir , sahasında konuk ettiği (...)
ahaber.com.tr	Türkiye, o ülkede altın ve volfram arayacak	MTA, Özbekistanda iki altın ve bir volfram sahasında arama ruhsatına ilişkin Özbekistan Jeoloji ve (...)
trthaber.com	Myanmada fabrika yangını: 6 ölü	Myanmada bir fabrikada çıkan yangında 6 kişi hayatını kaybetti, 13 kişi yaralandı. Şinhuada yer alan (...)
takvim.com.tr	Uçakta uyku konforu	TÜRK Hava Yolları (THY), kıtalar arası uçuşlarda Business Class sınıfında servis ettiği uyku setini yeniledi. (...)
sozcu.com.tr	Potada milli mesai başlıyor	2019 FIBA Dünya Kupası Avrupa Elemelerinde 2. tur I Grubundan çıkmayı garantileyerek Çin vizesi (...)

3.1.2. Haber Veri Kümesi B

Haber Veri Kümesi B için Kemik Doğal Dil İşleme Grubu'nun [19] 2013 yılında hazırlamış olduğu veri kümesi [20] kullanılmıştır. Bu veri kümesi sınıflandırma problemi için hazırlanmış olup, 13 farklı sınıfa ait olmak üzere toplamda 41992 tane Türkçe haber içermektedir. Haber Veri Kümesi B'de yer alan tüm sınıflar, her sınıftan örnek bir haber başlığı ve kısaltılmış örnek bir haber metni Tablo 3.3'de gösterilmiştir.

Türkçe metin özetleme alanında herhangi bir veri kümesi bulunmadığı için metin sınıflandırma problemi için hazırlanan Kemik haber veri kümesi, metin özetleme veri kümesi olmaya uygun hale getirilmiştir.

Haber Veri Kümesi B, 13 farklı haber sınıfından oluşmaktadır. Farklı sınıflara ait klasörlerdeki her bir metin dosyasına bakıldığında haberlerin ilk satırları haberin ana başlığını yani özetini içermektedir. Bu nedenle, her bir haber için haber metninin ilk satırı o haberin başlığı ve kalan satırları da o haberin metni olacak şekilde etiketlenmiştir. Daha sonra haber metni NLTK [69] kütüphanesinin yardımıyla cümlelerine ayrıştırılmıştır. Cümlelerine ayrıştırma işleminde NLTK kütüphanesinin dil seçeneği Türkçe seçilmiştir. Böylece her bir haber için bir başlık ve haber metnine ait bir cümle listesi elde edilmiştir. Haber Veri Kümesi B'de yer alan farklı haber sınıflarına ait istatistiksel bilgiler Tablo 3.4'de verilmiştir.

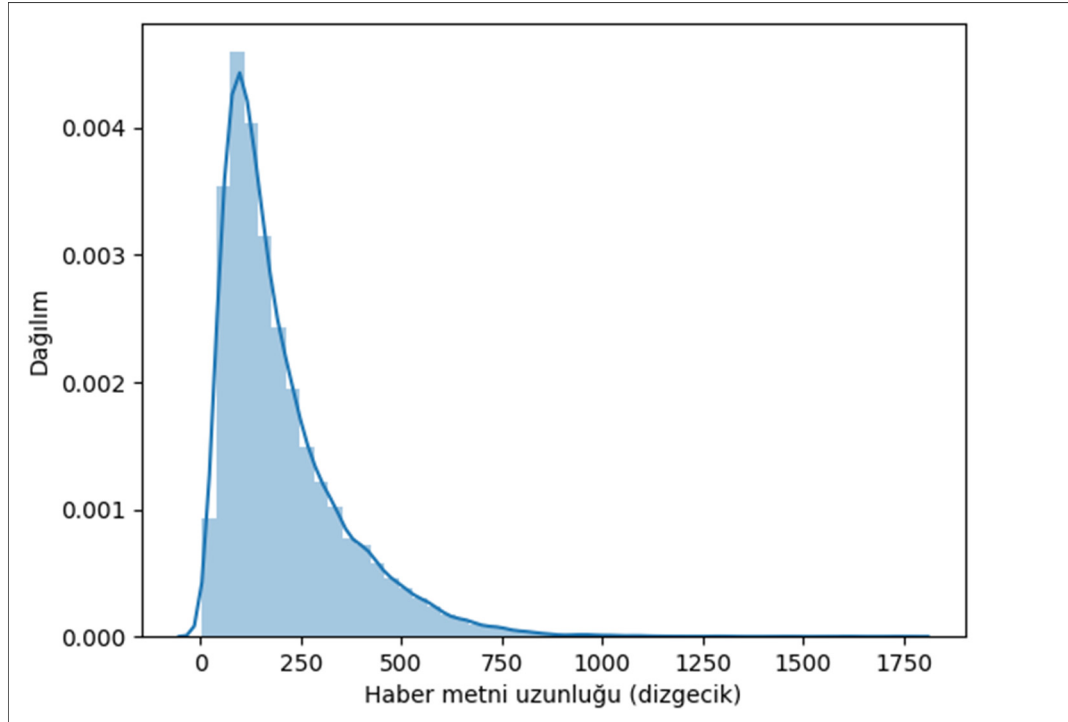
Haber Veri Kümesi B'nin haber metinleri için kelime sayısı frekansına göre dağılımlarını gösteren grafik Şekil 3.3'de ve haber başlıkları için kelime sayısı frekansına göre dağılımlarını gösteren grafik Şekil 3.4'de verilmiştir. Elde edilen dağılımlara göre haber başlıkları en sık 4 tane kelimedenden ve haber metinleri en sık 84 tane kelimedenden oluşmaktadır. Haber Veri Kümesi B'de yer alan haber başlıkları ortalama 5 kelime, haber metinleri ise ortalama 202 kelimedenden oluşmaktadır.

Tablo 3.3: Haber Veri Kümesi B'ye ait sınıflar, sınıflara ait örnek haber başlıkları ve kısaltılmış örnek haber metinleri.

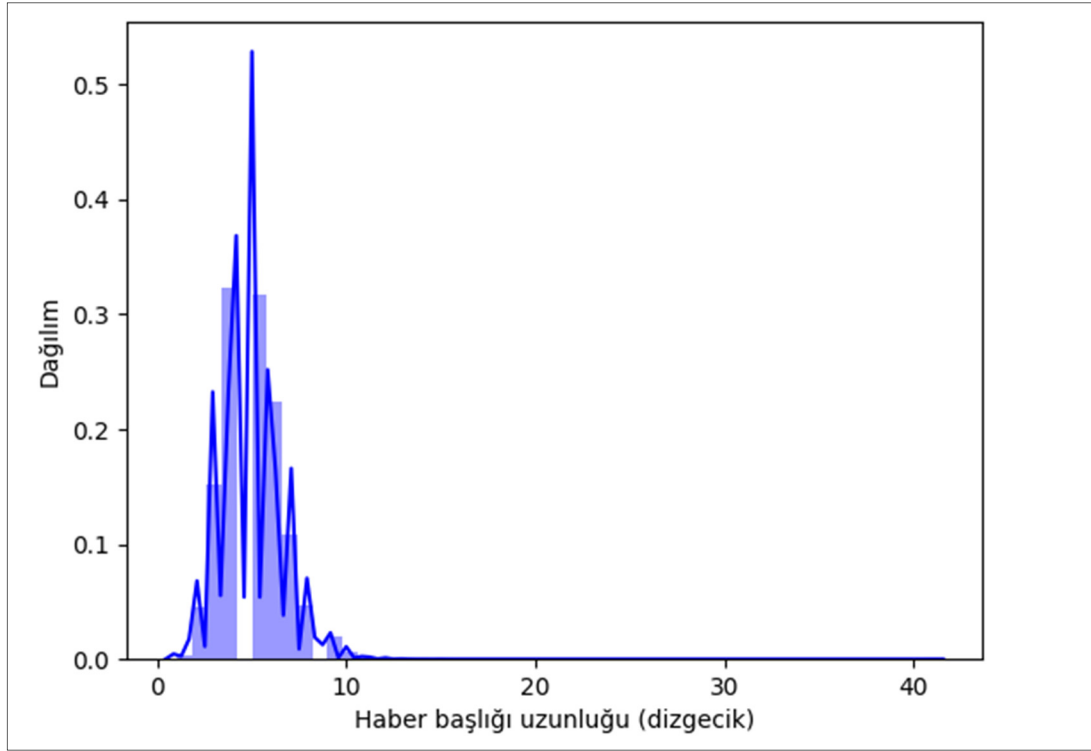
Sınıf İsmi	Örnek Haber Başlığı	Kısaltılmış Örnek Haber Metni
Dünya	Cameron ile Salmond referandumu görüşecek	İngiltere Başbakanı David Cameron ile İskoçya'daki bölgesel hükümetin Başbakanı Alex Salmond gelecek (...)
Ekonomi	Maliye 400 vergi müfettiş yardımcısı alacak	Maliye Bakanlığı Vergi Denetim Kurulu Başkanlığı 400 vergi müfettiş yardımcısı alacak. Konuya ilişkin (...)
Genel	Gözünden 10 santimetrelilik parazit çıktı	Tarsus ilçesinde, emekli bir öğretmenin gözünden solucana benzer 10 santimetre uzunluğunda parazitin (...)
Güncel	'Doğalgaz faciasının' 5. bilirkişi raporu	Ankara'da, 7 üniversite öğrencisinin öldüğü doğalgaz faciasıyla ilgili davada alınan son bilirkişi raporu (...)
Kültür Sanat	'Ateşin Düştüğü Yer' Oscar'ı yakacak	Türkiye'nin Oscar adayı "Ateşin Düştüğü Yer" filminin yapımcısı Baran Seyhan, 85. Akademi Ödülleri'nde (...)
Magazin	'Televizyona veda ediyorum'	'Babalar ve Evlatları' dizisi erken final yapan Mahsun Kırmızıgül, yeni dizisi 'Benim İçin Üzülme'nin son (...)
Planet	Atletizmin yıldızı Pistorius kız arkadaşını öldürdü	Güney Afrikalı ampute atlet Oscar Pistorius'un sevgilisini silahla vurarak öldürdüğü bildirildi. Pistorius (...)
Sağlık	Sivas'tan dünyanın gözlerine katarakt lens	Sivas'ta faaliyet gösteren Türkiye'nin ilk lens fabrikasında üretilen göz içi katarakt lensler, Avrupa'dan Asya (...)
Siyaset	Gensoru gündeme alınmadı	TBMM Genel Kurulu'nda, CHP'nin, Gıda Tarım ve Hayvancılık Bakanı Mehdi Eker hakkında verdiği (...)
Spor	Galatasaray'dan Alex açıklaması	Galatasaray Kulübü, Fenerbahçe'yle yollarını ayıran Alex de Souza'yla ilgili bir açıklama yaptı. Sarı-kırmızılı (...)
Teknoloji	iOS 6 güncelleniyor!	iOS 6, iPhone, iPad ve iPod touch için yayınlanalı bir aydan fazla bir süre oldu. Son raporlar ise Apple'ın (...)
Türkiye	Çevre kirliliğini protesto için soyundular	Antalya'da bir grup çevreci, 'Yaklaşan belediye seçimlerinde çevre kirliliğini değil, iklimi seçeceğiz' diyerek (...)
Yaşam	Önce dövdüler, sonra gasbettiler	Eskişehir'de 76 yaşındaki bir kişiyi darp ettikten sonra cebindeki 2 bin lirayı gasbettiği öne sürülen 3 (...)

Tablo 3.4: Haber Veri Kümesi B'ye ait istatistiksel bilgiler.

Sınıf İsmi	Haber Sayısı	Haber Başlığı Ortalama Kelime Sayısı	Haber Metni Ortalama Kelime Sayısı	Haber Metni Ortalama Cümle Sayısı
Dünya	3724	5,14	198,15	10,18
Ekonomi	3265	5,18	336,23	16,63
Genel	6673	4,87	312,55	19,17
Güncel	5847	5,21	257,56	15,11
Kültür Sanat	1155	5,30	307,52	14,66
Magazin	2792	4,49	180,70	13,12
Planet	1953	4,75	220,94	11,76
Sağlık	1383	5,07	368,69	19,89
Siyaset	1849	5,59	382,16	22,73
Spor	9997	4,65	240,42	13,75
Teknoloji	771	4,90	241,72	11,88
Türkiye	1939	5,13	276,65	15,43
Yaşam	644	5,05	209,95	12,02



Şekil 3.3: Haber Veri Kümesi B haber metni dizgecik sayısı dağılımları.



Şekil 3.4: Haber Veri Kümesi B haber başlığı dizgecik sayısı dağılımları.

3.2. Veri Kümelerinin Ön İşlemden Geçirilmesi

Haber metnlerinin önışlemden geçirilmesi aşamasında, bazı haber metinleri geçersiz karakterler içerdiği için öncelikle geçersiz karakter içeren kelimeler temizlenmiştir. Bunun için karakter eğer “ıİöÖüÜşŞğĞçÇ” listesindeki karakterlerden biri değilse veya Unicode [70] karakteri temsil eden tamsayı 0 ile 127 arasında değilse, o karakterin bulunduğu kelime silinmiştir. Sonrasında genellikle haberlerin en başında ya da en sonunda yer alan haber ajanslarının isimleri (AA, NTV vb.) kaldırılmıştır.

Haber metnlerinin bazıları ara başlıklar da içermektedir. Bu ara başlıklar çoğu zaman 3 kelime ya da daha az kelimeden oluşmaktadır. Ara başlıkların haber metninde yer almaması için, 3 kelime veya 3 kelimeden daha az olan ve tamamı büyük harf olan satırlar (ara başlıklar) haber metnlerinden çıkarılmıştır. Tablo 3.5’te Haber Veri Kümesi B’ye ait örnek bir ara başlık içeren haber metnini oluşturan satırlar gösterilmiştir. Burada 4. ve 6. satırlar ara başlıklardır ve haber metninden kaldırılmıştır. Sonrasında ise Beautiful Soup kütüphanesi [71] yardımıyla haberlerde bulunan linkler kaldırılmıştır.

Tablo 3.5: Haber Veri Kümesi B’ye ait ara başlık içeren örnek bir haber metni.

Satır No	Haber İçeriği
1	Türkiyeden lüks saate know-how
2	Armani, Burberry, DKNY, Michael Kors gibi markaların saat lisansörü olan, 3 milyar dolar cirolu Fossil Group, 50 ülkeden 200 distribütörünü İstanbul’da ağırladı.
3	Türkiye’de Saat&Saat’in organizasyon yapısından etkilenen distribütörler, Türk yetkililerden know-how talep etti. Fossil Group’un CEO’su Robert Christiano, Doğu ve Batı’nın buluşmasının simge (...)
4	TÜRKİYE ÖRNEK OLDU
5	Fossil Avrupa, Ortadoğu ve Avrupa Başkanı Dieter Waldner ise Türkiye Doğu Avrupa, Ortadoğu ve Afrika içinde satışta en büyük pazarımız. Ama aynı zamanda diğer distribütörlerimiz Türkiye’yi örnek alıyor (...)
6	CENNETTEN BİR PARÇA
7	Fossil Group’un CEO’su Robert Christiano, ilk 9 aylık dönemde dünya genelinde yüzde 18 büyümeye yaşadıklarını, Avrupa’da bu oranın tek haneli olduğunu hatırlatarak şöyle konuştu: Büyümemiz devam (...)

Haber metni içerisinde yer alan sayılar genellikle benzer şeyleri ifade etse de (miktar ya da bir numaralandırma gibi) değişen değerlerdir. Örneğin “2010 yılında” kelime grubu “2017 yılında” kelime grubu ile benzer bir anlam taşısa da içerlerinde yer alan sayılar farklı olduğu için derin öğrenme modeli bu iki gruba farklı anlamlar katar. Aynı zamanda sayıları kullanmak eşsiz dizgecik sayısını da artıracığı için, tüm bu nedenler dili öğrenmeyi zorlaştırmaktadır. Bu sebeple haberlerde yer alan tüm sayılara ortak anlam yüklemek amacıyla tüm sayılar ‘#’ karakteri ile değiştirilmiştir.

Son olarak haber metinlerinden harf ve ‘#’ karakteri haricindeki karakterlerin tamamı silinmiştir. Haber metinlerinde hem büyük hem de küçük harflerin bulunması öğrenme işlemini zorlaştıracığı için haber metinlerinin tamamı küçük harflere çevrilmiştir. En nihayetinde haber metinleri dizgeciklerine ayrılarak haber metinlerinin ön işlemden geçirilme aşaması tamamlanmıştır.

3.3. Sözlük Oluşturulması

Haber başlıkları ve haber metnlerinin derin öğrenme modeline verilebilmesi için haberlerde geçen her bir eşsiz kelimenin karşılığı tamsayı olacak şekilde bir sözlük oluşturulmuştur. Sözlükte yer alan haber başlıkları ve haber metnlerinde toplam Haber Veri Kümesi A’da toplamda 328787 tane ve Haber Veri Kümesi B’de toplamda 275664 tane eşsiz dizgecik gözlenmiştir. Sözlük oluşturulurken dizgeciklerine ayrılan haberler dizgeciklerin geçme sıklığına göre en sık geçen dizgecik 4’ten başlayacak şekilde indekslenmiştir. İlk indeks, sağdan sıfır dolgu için, ikinci indeks sözlükte yer almayan kelimeler için, üçüncü indeks cümlelerin başlangıcı için ve dördüncü indeks ise cümlelerin bitişi için kullanılmıştır. Haber Veri Kümesi A için oluşturulan sözlüğün ilk 15 indeksi Tablo 3.6’da ve Haber Veri Kümesi B için oluşturulan sözlüğün ilk 15 indeksi Tablo 3.7’de sunulmuştur. Tablolardan görüldüğü üzere, durak kelimelerin (ve, bir, da, de, ile, veya vb.) frekansı diğer kelimelere göre daha yüksek çıkmıştır. Durak kelimelerin frekansının yüksek olması, bu kelimelerin dilbilgisi için önemli olduğunu göstermektedir. Bu sebeple, durak kelimeler haber metnlerinden çıkarılmamıştır. Durak kelimeler sayesinde üretilen haber başlıkları dilbilgisi kurallarına daha uygun olarak tahmin edilmiştir.

Tablo 3.6: Haber Veri Kümesi A’nın sözlüğünün ilk 15 indeksi ve geçme sıklıkları.

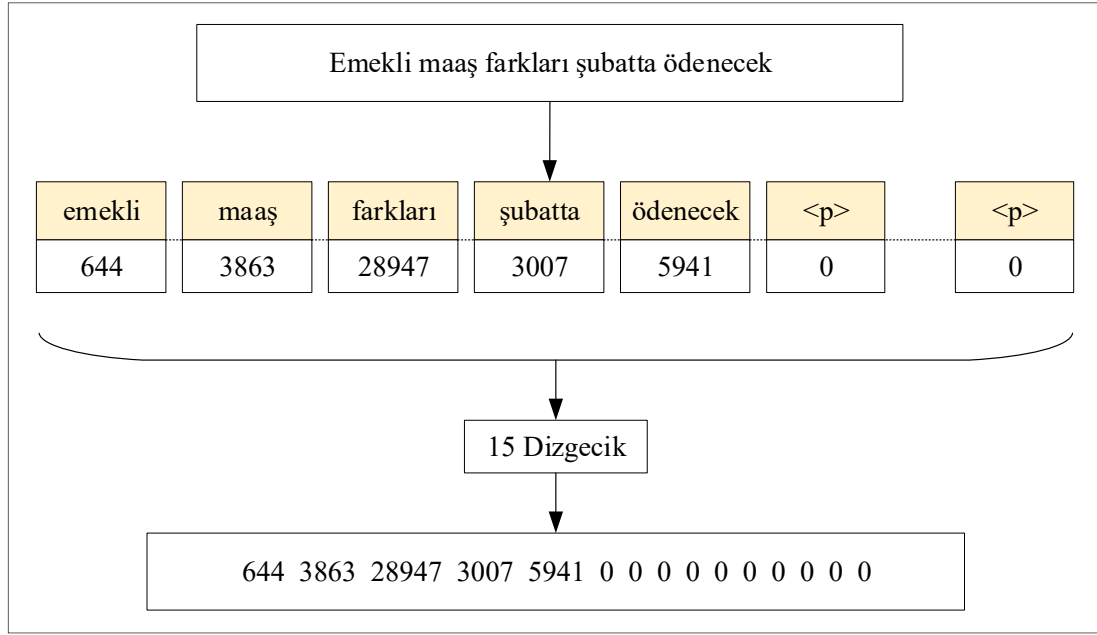
Dizgecik	İndeks	Geçme Sıklığı
< pad > (padding)	0	-
< unk > (unknown word)	1	-
< s > (sentence start)	2	-
</s > (sentence end)	3	-
#	4	647578
ve	5	326517
bir	6	208912
bu	7	122767
ile	8	91585
için	9	90543
da	10	89249
de	11	88031
çok	12	50254
olarak	13	50102
en	14	41999

Tablo 3.7: Haber Veri Kümesi B'nin sözlüğünün ilk 15 indeksi ve geçme sıklıkları.

Dizgecik	İndeks	Geçme Sıklığı
< <i>pad</i> > (padding)	0	-
< <i>unk</i> > (unknown word)	1	-
< <i>s</i> > (sentence start)	2	-
</ <i>s</i> > (sentence end)	3	-
#	4	354398
ve	5	153476
bir	6	112352
bu	7	70442
da	8	68036
de	9	62032
ile	10	47682
için	11	46737
olarak	12	27671
çok	13	26452
ise	14	24105

Sözlük oluşturulduktan sonra haber başlıkları ya da haber metnlerinin her biri farklı sayıda kelimelerden oluştuğundan dolayı aynı boyuta gelmesi için sağ taraftan sıfır dolgu ile ekleme yapılmıştır. Haber başlıkları için bu boyut 15 iken, haber metinleri için bu boyut 50 olarak belirlenmiştir.

Sonuç olarak tüm haber başlıkları ve haber metinleri için önceden tanımlanmış boyutlarda dizgecik indeks vektörleri elde edilmiştir. Haber Veri Kümesi B'den alınmış örnek bir haber başlığı için, dizgecik indeks vektörünün oluşturma aşamaları Şekil 3.5'de gösterilmiştir. Haber başlıkları ve haber metinlerindeki tüm eşsiz kelimeler için dizgecik indeks vektörleri oluşturulduktan sonra, sözlükteki her bir kelimenin vektör temsili için FastText kütüphanesinin [72] Türkçe kelime vektörleri [73] kullanılmıştır.



Şekil 3.5: Haber Veri Kümesi B'ye ait bir haber başlığından dizgecik indeks vektörünün elde edilmesi.

3.4. Kelime Gömmelerinin Çıkarılması

Doğal dil işleme alanında kelimelerle çalışılırken, kelimeleri bilgisayarların anlayabilmesi için sayısal biçimde kodlamak gerekmektedir. Burada kullanılan yöntemlerden biri Mikolov vd. tarafından geliştirilen kelime gömmesi [74] yöntemidir. Kelime gömmesi, benzer anlamı olan kelimelerin özellik uzayında birbirine yakın şekilde temsil edilmesini sağlayan bir sayısal gösterim türüdür. Derin öğrenme alanında önemli bir gelişme olan kelime gömmesi yöntemi ile doğal dil işleme çalışmalarındaki zor problemlerde kayda değer yollar kat edilmiştir [75]. Bu teknik ile her bir kelime, çok boyutlu ve gerçek değerli bir vektör ile temsil edilir.

Literatürde metin özetleme alanında yayınlanan çalışmalara bakıldığında, metin verilerinden kelime gömmesini elde etmek için Word2Vec [76], GloVe [77] ve FastText [72] olmak üzere 3 farklı yöntemin yaygın olarak kullanıldığı görülmektedir.

- Word2Vec Tekniği

Word2Vec [76], metin kümesinden kelime gömmelerinin öğrenilmesi için kullanılan bir algoritmadır. Word2Vec modeli, temel olarak bir adet girdi katmanı, bir adet gizli katman ve bir adet çıktı katmanından oluşan bir yapay sinir ağıdır. Word2Vec, girdi olarak bir metin kümesi alır ve çıktı olarak bir vektör uzayı üretir. Bu vektör uzayı genellikle yüzlerce boyuttan oluşur ve her bir eşsiz kelimeye bir

vektör atanacak şekilde elde edilir. Vektör uzayı birbiri ile ilişkili olan kelimeler için uzayda birbirine yakın olan vektörlerden oluşmaktadır.

Bu algoritma, kelimeleri vektör uzayında ifade ederken gözetimsiz ve tahmine dayalı olarak çalışmaktadır. Word2Vec yönteminin CBOW (devamlı kelime çantası) ve Skip-Gram olmak üzere iki farklı öğrenme mimarisi vardır. CBOW mimarisinde, kelimelerin vektörleri elde edilirken pencere genişliği, gömme boyutu gibi hiper parametreler kullanılmaktadır. Skip-Gram mimarisinde ise model, kelime vektörlerini elde ederken, sadece şu anki kelimeyi kullanmaktadır. Gerçekleştirilen deney sonuçlarına göre CBOW mimarisi hızlı çalışırken, Skip-Gram mimarisi daha yavaş ancak seyrek geçen kelimeler için daha iyi sonuçlar elde edebilmektedir.

- GloVe Tekniği

GloVe [77], Word2Vec yönteminin devamı niteliğinde olan ve kelimelerin vektör gösterimlerinin gözetimsiz bir öğrenme algoritması sayesinde elde edildiği yöntemdir. Vektör gösterimleri, LSA gibi matrisi çarpanlarına ayırma yöntemleri kullanılarak elde edilir. Matrisi çarpanlarına ayırma işlemi, istatistiksel açıdan iyi sonuçlar verebilse de, Word2Vec kadar anlamsal gösterimler oluşturamamaktadır.

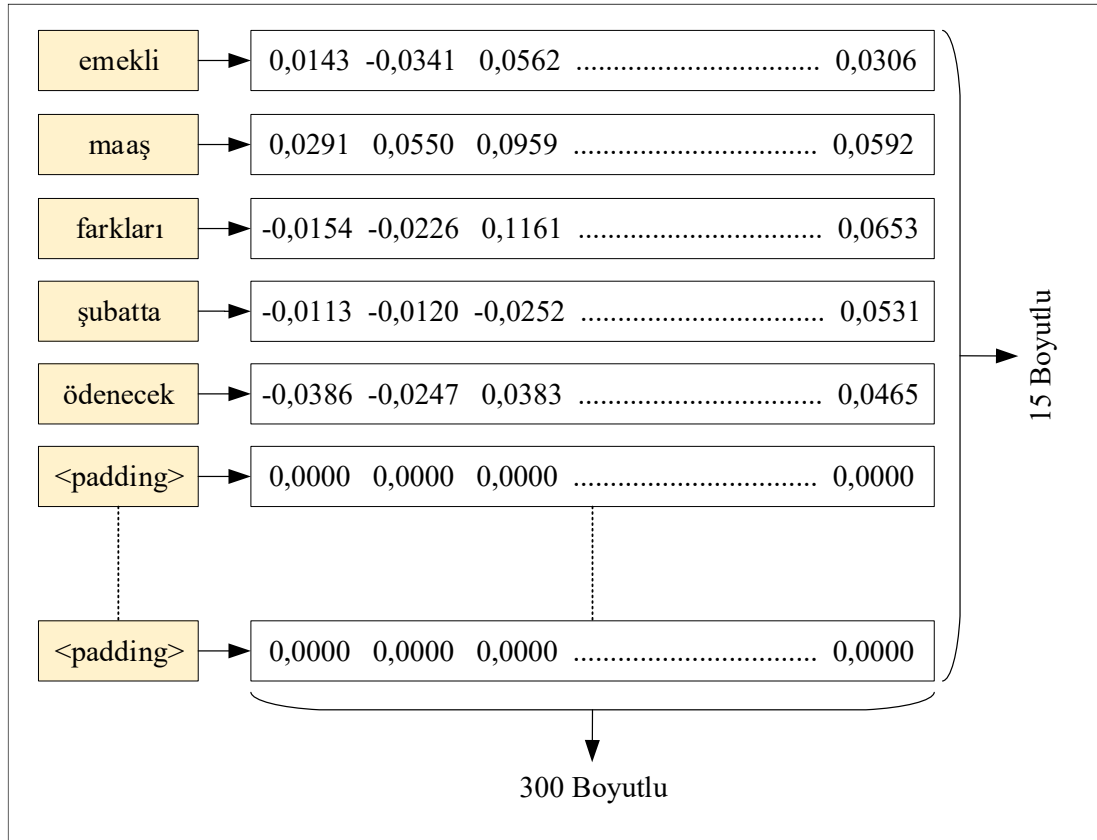
- FastText Tekniği

Bu motorlarda motorun stator ve rotor sargılarında direkt akımlar akmaktadır. Matematiksel analizleri nispeten daha kolay olmakla beraber, üretimleri daha karmaşıktır. Net tork üretimi için akım yönünü değiştirmek amacı ile komütatöre ihtiyaç duyarlar. Tez çalışması kapsamında kullanılan FastText [72] yöntemi, Facebook AI Research tarafından metin sınıflandırma problemi için geliştirilen [78] ve yine Word2Vec yönteminin uzantısı olan bir yöntemdir. FastText, Wikipedia ve internet üzerinden elde edilen metinler kullanılarak, içerisinde Türkçe dilinin de yer aldığı 157 farklı dil için [9] önceden eğitilmiş kelime vektörlerini içermektedir. Bu vektörler herkese açık olarak paylaşılmıştır. Modeller, CBOW metodu ile 300 boyutunda 5 uzunluğunda karakter n-gram'ı ile eğitilmiştir.

FastText, kelimeleri tek tek yapay sinir ağına vermek yerine, karakter tabanlı olarak n-gram halinde parçalar. FastText karakter tabanlı olduğu için kelime tabanlı Word2Vec ve GloVe vektörlerine göre daha avantajlıdır. Bu avantajları aşağıdaki gibi sıralamak mümkündür:

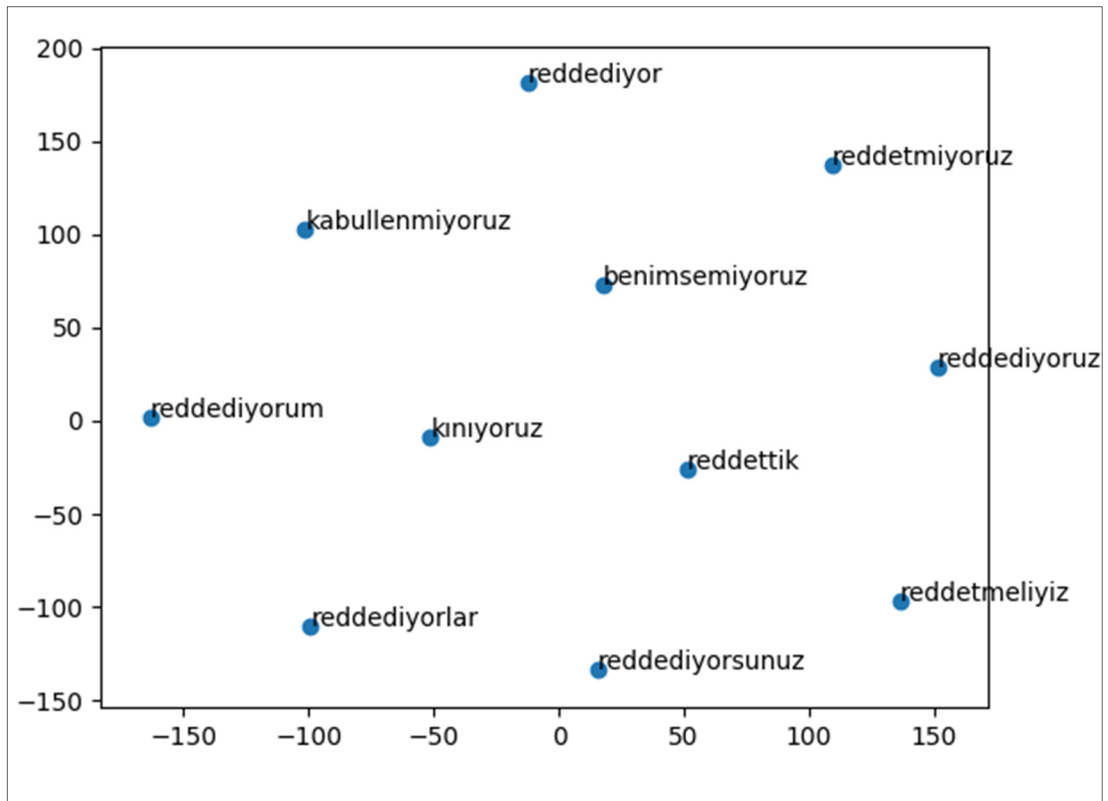
- Çok nadir geçen kelimelerin n-gram'larının ortaya çıkma ihtimalinin düşük olmasından dolayı, FastText sayesinde artık bu kelimeler daha doğru bir şekilde temsil edilebilir.
- Sözlük dışındaki bir kelime (OOV) eğitimde görünmese bile, karakterin gram cinsinden vektör karşılığı bilinebileceğinden bu kelimeler için de vektör temsili oluşturulabilir.

Bu tür avantajlar göz önünde bulundurularak, tez çalışması kapsamında sözlükteki her bir kelime için FastText kütüphanesi yardımıyla vektör karşılıkları elde edilerek, 300 boyutlu kelime gömmeleri oluşturulmuştur. Kelime gömme karşılığı çıkarılacak olan kelimenin vektör karşılığı FastText kütüphanesinde yer almadığı durumda, vektör karşılığı olarak 300 boyutlu 0 dizisi kullanılmıştır. En nihayetinde hazırlanan kelime gömmeleri sistemin eğitimi için, kodlayıcı modele girdi olarak verilmiştir. Haber Veri Kümesi B'ye ait örnek bir haber başlığından FastText aracılığı ile elde edilmiş kelime gömmeleri Şekil 3.6'da gösterilmiştir.

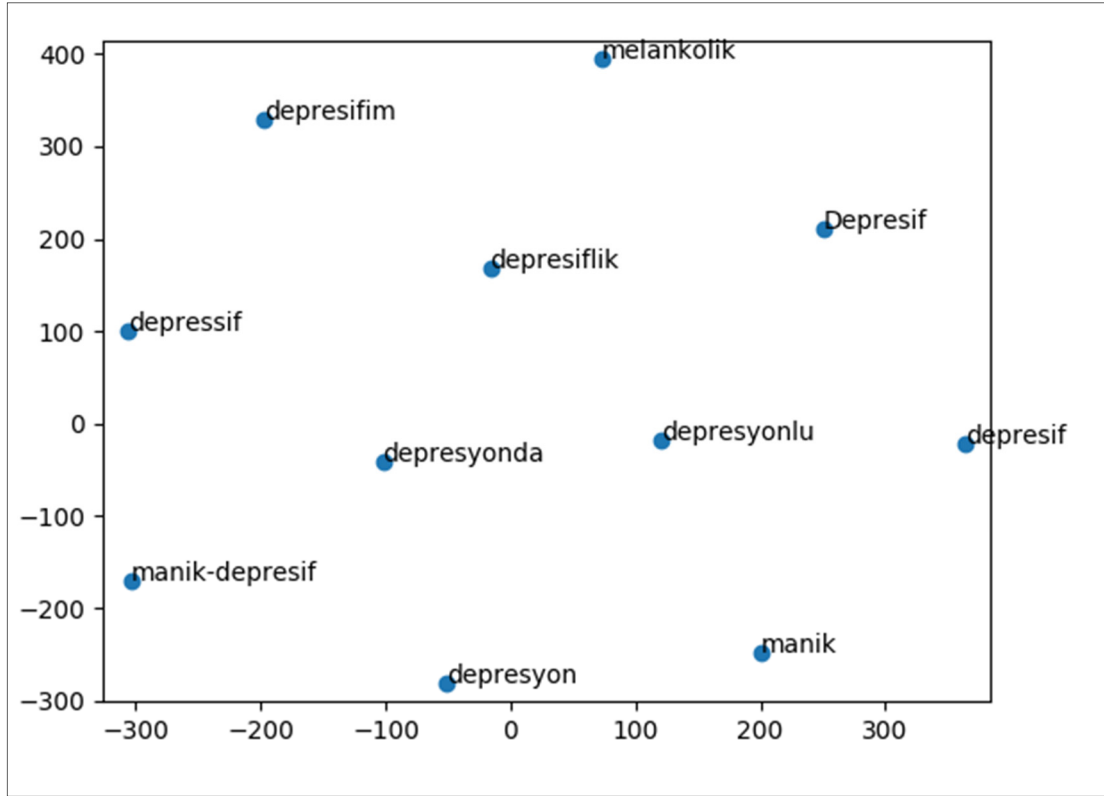


Şekil 3.6: Haber Veri Kümesi B'den alınan örnek bir haber başlığına ait kelime gömmelerinin çıkarılması.

Haber Veri Kümesi A için oluşturulan sözlükte nadir geçen kelimelerden (yaklaşık 50000 sıralarında) “reddediyoruz” kelimesinin FastText aracılığı ile elde edilmiş benzer kelimeleri Şekil 3.7’de gösterilmiştir. Haber Veri Kümesi B için oluşturulan sözlükte, yine nadir geçen kelimelerden (yaklaşık 30000 sıralarında) “depresif” kelimesinin FastText aracılığı ile elde edilmiş benzer kelimeleri ise Şekil 3.8’de gösterilmiştir. Görsellerde, kelimelerin iki boyutlu uzaydaki konumları, kelime vektörlerine göre belirlenmiştir. Şekillerden anlaşılacağı gibi kelimeler her ne kadar nadir geçiyor olsa da, FastText bu kelimeleri başarılı bir şekilde vektöre dönüştürebilmiş ve eş anlamlı kelimelerini bulabilmiştir.



Şekil 3.7: Haber Veri Kümesi A sözlüğünde nadir geçen "reddediyoruz" dizgeciğinin FastText'te yer alan benzer dizgecikleri.

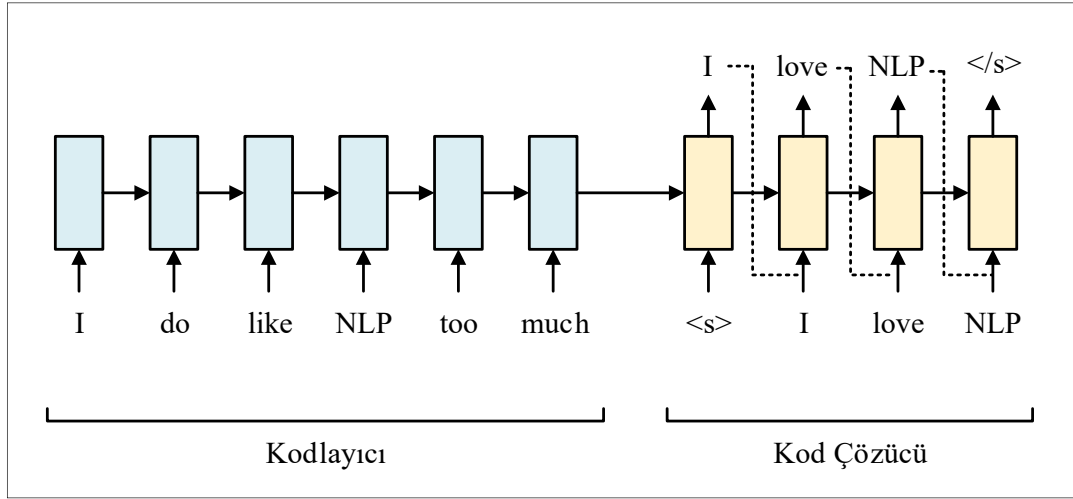


Şekil 3.8: Haber Veri Kümesi B sözlüğünde nadir geçen "depresif" dizgeciğinin FastText'te yer alan benzer dizgecikleri.

3.5. Sistemin Eğitimi

Sistemin eğitim aşamasında, literatürde doğal dil işleme alanında yaygın olarak kullanılan kodlayıcı-kod çözücü modeli [79] geliştirilmiştir. Kodlayıcı-kod çözücü modeli, sıradan sıraya tahmin problemlerine çözüm getirmek için tasarlanmış bir RNN modelidir. İlk olarak makine çevirisi problemine çözüm sunmak için geliştirilen kodlayıcı-kod çözücü modelinin, soru cevaplama ve metin özetleme gibi doğal dil işleme problemlerinde de oldukça başarılı sonuçlar verdiği gözlemlenmiştir.

Bu modelde, kodlayıcı giriş dizisinin tamamını okur ve sabit uzunluklu olan bağlam vektörüne kodlar. Kod çözücü ise kodlayıcı tarafından kodlanmış giriş dizisini alır ve çıkış dizisini oluşturur. Model, dizi sonu belirtecini çıkardıktan sonra tahmin yapmayı durdurur. Temel olarak kodlayıcı-kod çözücü mimarisi Şekil 3.9'da gösterilmiştir.



Şekil 3.9: Kodlayıcı-kod çözücü mimarisi.

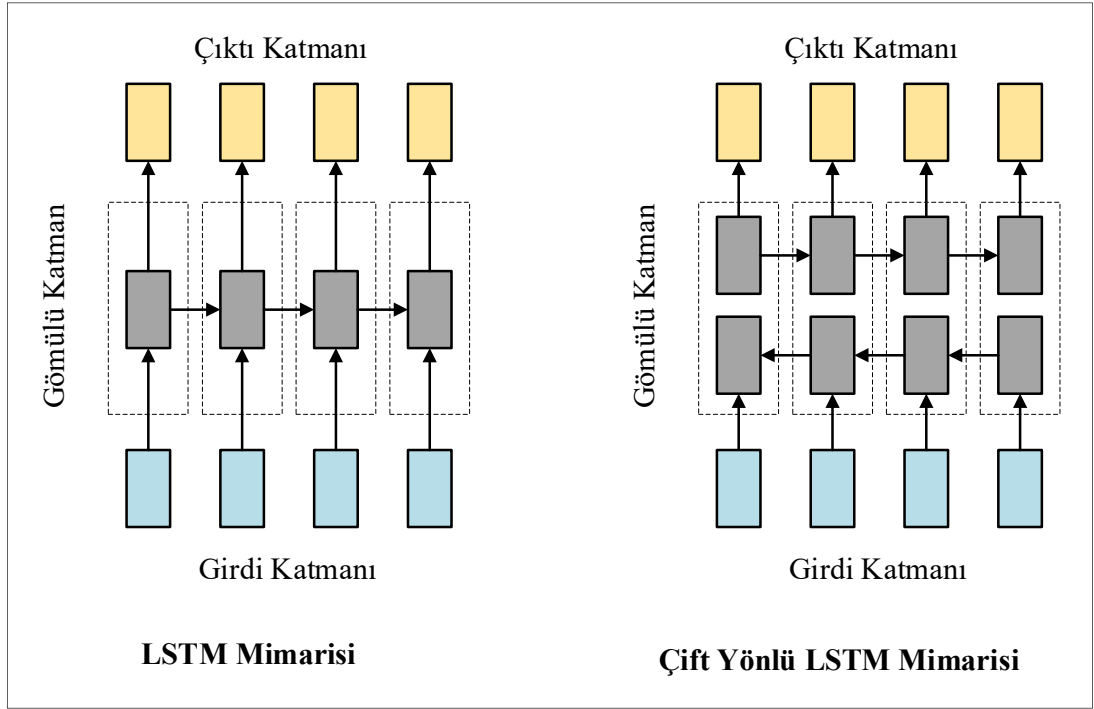
Bu çalışma kapsamında kodlayıcı-kod çözücü mimarisi geliştirilmiştir. Kodlayıcı tarafında Çift Yönlü LSTM ağı kullanılırken, kod çözücü tarafında LSTM ağı kullanılmıştır. Kodlayıcı girdi olarak haber metinlerini almaktadır. Kod çözücü ise, girdi olarak haber başlıklarını almaktadır. Kod çözücünün başlangıç durumu kodlayıcının son tekrarlayan katmandaki son durumudur. Eğitim aşamasının tamamı TensorFlow [80] kütüphanesi kullanılarak geliştirilmiştir.

3.5.1. Çift Yönlü LSTM

LSTM ağları, dizi tahmin problemlerinde sıra bağımlılıklarını öğrenmede kullanılan bir RNN türüdür. LSTM çoğunlukla makine çevirisi, metin özetleme, ses tanıma gibi karmaşık alanlarda kullanılmaktadır.

LSTM; uzun vadeli bağımlılıkları öğrenme yeteneğine sahip RNN çeşididir. Çift Yönlü LSTM ise LSTM'in bir uzantısı olarak geliştirilmiştir. Çift Yönlü LSTM modeli, dizi tahmin problemlerinde daha iyi performans göstermektedir. Çift Yönlü LSTM modelinde girdi dizisinde bir yerine iki LSTM modeli eğitilmektedir. İlk model LSTM modelinin aynısı, ikincisi ise, dizinin tersten sıralı halini kullanmaktadır. LSTM sadece geçmişteki bilgilerden yararlanırken, Çift Yönlü LSTM hem geçmişteki hem de gelecekteki bilgilerden yararlanarak, doğal dil işleme problemlerinde daha iyi performans göstermektedir. LSTM ve Çift Yönlü LSTM mimarileri kabaca Şekil 3.10'da verilmiştir.

Sistemin eğitimi sırasında kelime gömmeleri kodlayıcı modeline girdi olarak verildikten sonra, Çift Yönlü LSTM ile eğitilmiştir.

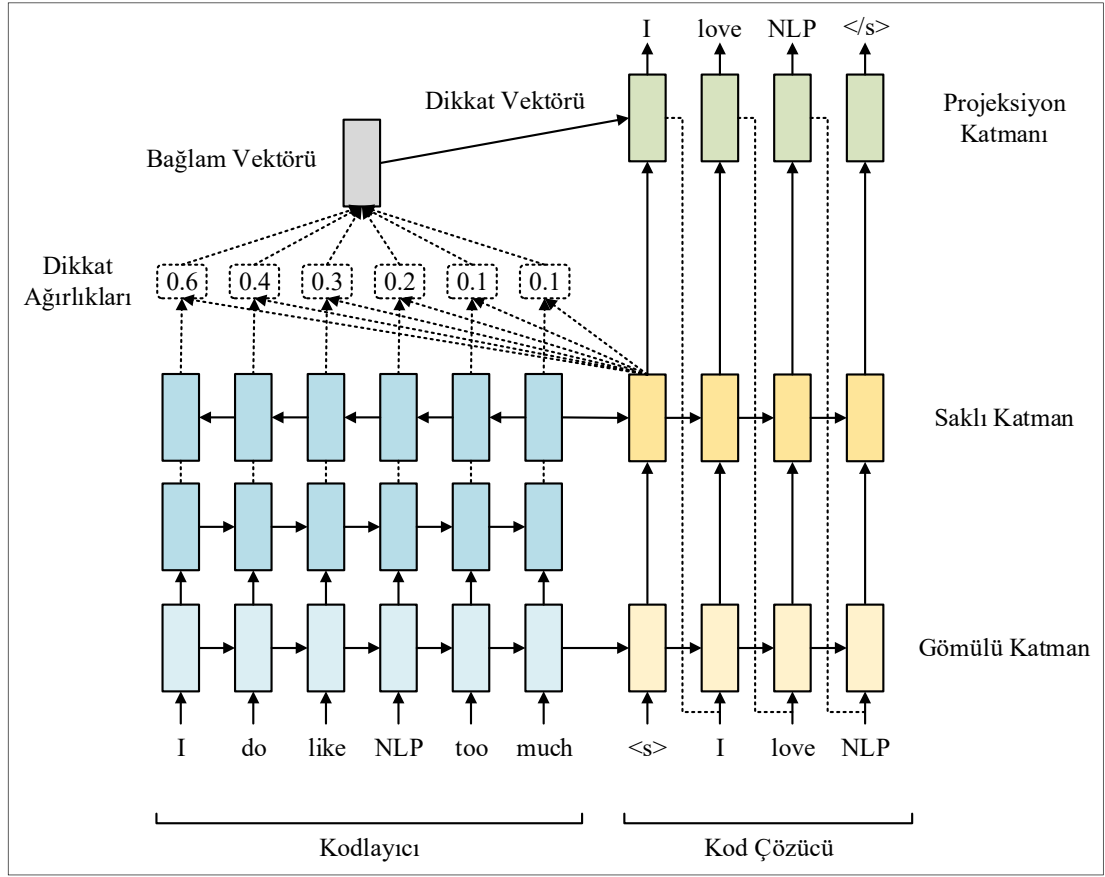


Şekil 3.10: LSTM ve Çift Yönlü LSTM Mimarileri.

3.5.2. Dikkat Mekanizması

Dikkat katmanı [81], kodlayıcı-kod çözücü RNN modelinin makine çevirisi problemindeki performansını iyileştirmek için geliştirilmiş bir mekanizmadır.

Kodlayıcı-kod çözücü modeline dikkat mekanizması eklenmiş mimari Şekil 3.11’de gösterilmiştir. Şekil 3.11’e göre her bir girdi kelimesine dikkat mekanizması tarafından bir ağırlık verilmektedir. Bu ağırlık, kod çözücü tarafında haber başlığı oluşturulurken, bir sonraki kelimeyi tahmin etmede kullanılmaktadır.



Şekil 3.11: Dikkat katmanı eklenmiş kodlayıcı-kod çözücü mimarisi.

3.5.3. Hiper Parametre Konfigürasyonu

Hiper Parametre, makine öğrenmesi modellerinin tasarımları esnasında belirlenmiş, farklı problemlere veya farklı veri boyutlarına göre değiştirilmesi gereken parametrelerdir. Bu parametrelerin belirlenmesi, sistemin eğitimi gerçekleştiren kullanıcıya bırakılmıştır. Parametre değerleri, makine öğrenmesi probleminin uygulandığı alana ve eğitimin gerçekleştirileceği verinin büyüklüğüne göre değişiklik göstermektedir. Bu bağlamda, sistemin eğitiminde metin özetleme problemi için yaygın olarak kullanılan parametreler kullanılmıştır. Kelime gömme boyutu; FastText'ten gelen vektör boyutunu ifade etmektedir. Sistemin eğitiminde kullanılan Hiper Parametre'lerin tamamı ve bu parametrelerin değerleri Tablo 3.8'de verilmiştir.

Tablo 3.8: Sistemin eğitiminde kullanılan Hiper Parametreler.

Hiper Parametre Adı	Hiper Parametre Değeri
Eğitim Tur Sayısı (Number of Epochs)	30
Yığın Boyutu (Batch Size)	64
Gizli Katman Sayısı (Number of Hidden Layers)	150
Katman Sayısı (Number of Layers)	2
Kelime Gömme Boyutu (Word Embedding Size)	300
Öğrenme Hızı (Learning Rate)	0,003

3.6. Haber Başlıklarının Üretilmesi

Doğal dil işleme alanında başlık üretme, metin özetleme ve metin çevirisi gibi problemlerde, elde edilmeye çalışılan çıktı, bir cümle dizisidir. Bu tür problemlerde kullanılan modeller, genellikle çıktı olarak sözlükte yer alan her kelimeye bağlı bir olasılık dağılımı vermektedir. Bundan sonraki aşamada ise, bu olasılıkları çıktı olarak elde edilecek sıralı kelime dizisine çevirme işlemi, kod çözücüyü bırakılır. Bu durum metnin çıktı olarak üretildiği metin madenciliği ve doğal dil işleme problemlerinde aynıdır. Sinir ağı modelinin son katmanında, sözlükte yer alan her bir kelime için bir nöron ve olasılığın elde edilmesi için bir softmax aktivasyon fonksiyonu kullanılır. En uygun kelime dizisini çıktı olarak verebilmek için, tüm olası çıktı dizilerinde olasılıklarına göre bir arama yapmak gerekmektedir. Sözlükler bazen yüzlerce, bazen binlerce ve hatta bazen de milyonlarca kelime içerebilmektedir. Bu sebeple, bu arama problemi çıktıda elde edilen kelime dizisinin uzunluğuna göre üstel bir hale gelebilmektedir. Dolayısı ile burada bir arama algoritmasına ihtiyaç duyulmaktadır.

Literatürde açgözlü arama ve ışıın araması olmak üzere iki çeşit kod çözücü algoritması kullanılmaktadır. Tez çalışması kapsamında, haber başlıklarının çıkarımı aşamasında, en iyi çıkış dizisini seçerken, kod çözücü algoritmalarından hem açgözlü arama hem de ışıın araması kullanılmış ve karşılaştırma yapılmıştır.

3.6.1. Açgözlü Arama

Açgözlü arama [16] kod çözücü algoritması, çıktı sırasındaki her adımda en büyük olasılıklı kelimeyi seçen algoritmadır. Açgözlü arama algoritması hızlıdır, ancak çıkış dizisi her zaman en uygun (optimal) dizi olmayabilir.

3.6.2. Işın Araması

Işın araması [17] kod çözücü algoritması, açgözlü arama algoritmasının üzerine geliştirilmiş olan ve en olası çıktı kelime dizisini üreten, literatürde oldukça popüler olarak kullanılan bir algoritmadır. Bu algoritma çıktı dizisini açgözlü olarak oluşturmak yerine, en fazla k uzunluğunda olan tüm olası çıktı dizilerinde arama yaparak oluşturur. Buradaki k değeri, kullanıcı tarafından belirlenir ve aramadaki ışın sayısını ya da paralel arama sayısının da belirlenmesinde kullanılır. Işın arama algoritmasının sözde kodu Şekil 3.12’de verilmiştir.

Algoritma 1 Işın Araması

```
1:  $beams = \{\emptyset\}$ ;
2:  $scores(\emptyset, 0) = 1$ ;
3: for  $t = 1 \dots T$  do
4:    $bestbeams = bestbeams(beams, BW)$ ;
5:    $beams = \{\}$ ;
6:   for  $b \in bestBeams$  do
7:      $beams = beams \cup b$ ;
8:      $scores(b, t) = calcScore(mat, b, t)$ ;
9:     for  $c \in alphabet$  do
10:       $\hat{b} = b + c$ ;
11:       $scores(\hat{b}, t) = calcScore(mat, \hat{b}, t)$ ;
12:       $beams = beams \cup \hat{b}$ ;
13:   end
14: end
15: end
16: return  $bestbeams(beams, 1)$ ;
```

Şekil 3.12: Işın araması algoritması.

4. DENEYLER

Bu bölümde, tez çalışmasında gerçekleştirilen deneyler anlatılmıştır. İlk önce deneyler gerçekleştirilirken kullanılan değerlendirme metrikleri (ROUGE skoru ve kosinüs benzerliği) anlatılmış, daha sonra kullanılan veri kümelerinin eğitim ve test aşaması için nasıl gruplandığı anlatılmış ve en son ise elde edilen deney sonuçlarından bahsedilmiştir.

4.1. Değerlendirme Metrikleri

Sistemin başarısı ölçülürken, literatürde metin özetleme alanında oldukça yaygın olarak bilinen ROUGE değerlendirme metriği [21] kullanılmıştır. Bu metrik, skor hesabı yaparken yalnızca kelime eşleşmesini dikkate almaktadır. Bu nedenle, sistemin ürettiği özet ile tam doğru olan özet anlamsal olarak benziyor olsa da kelime farklılığından dolayı düşük skor elde edebilmektedir. Bu problemin önüne geçebilmek için; ek olarak anlamsal benzerlik skorlama metriği [82] de kullanılmıştır.

4.1.1. ROUGE Skoru

Chin-Yew Lin tarafından geliştirilen ROUGE metriği [21], ilk önce 2001 yılından itibaren yıllık olarak yapılan DUC konferansında [53] kullanılmıştır. Bu konferans metin özetleme sistemlerinin karşılaştırılması için ortak bir zemin ve bu alanda kullanılabilecek veri kümeleri sağlar. ROUGE değerlendirme metriği, son zamanlarda metin özetleme alanında en yaygın olarak kullanılan metriklerden biridir. ROUGE metriğinin çalışma mantığı, temelde kelimelerin eklerine bakılmaksızın birebir eşleşmesine dayanır. ROUGE skoru 0 ile 1 arasında bir değer almaktadır. Tam doğru metin ile tahmin edilen metin arası kelime eşleşmesi ne kadar fazla ise, ROUGE skoru da 1'e o kadar yakın çıkmaktadır. ROUGE değerlendirme metriği, ROUGE-N, ROUGE-L, ROUGE-S, ROUGE-W ve ROUGE-SU olmak üzere toplamda beş tane değerlendirme skoruna sahiptir.

Sistemin ROUGE skoru ile değerlendirmesi yapılırken, ROUGE-N tekniği kullanılmış ve $N = 1$ olarak alınmıştır. ROUGE-N skoru, tam doğru özet ile sistemin ürettiği özet arasındaki N-gram benzerliğe bakmaktadır. ROUGE-N skorunun hesaplanma formülü Denklem 4.1'de verilmiştir.

$$ROUGE - n = \frac{\sum_{C \in RSS} \sum_{gram_n \in C} Count_{match}(gram_n)}{\sum_{C \in RSS} \sum_{gram_n \in C} Count(gram_n)} \quad (4.1)$$

Denklem 4.1’de RSS ; tam doğru özet kümesi, C ; tam doğru özet kümesinden herhangi bir özet, $gram_n$; N-gram, $Count(gram_n)$; tam doğru olan özetteki N-gram sayısıdır. $Count_{match}$ fonksiyonu ise, tam doğru olan özet ile sistemin ürettiği özet arasındaki N-gram kelime eşleşmesini hesaplamaktadır. N-gram kelime eşleşmesi, N uzunluğunda sıralı olarak eşleşen tüm kelime gruplarının toplam sayısına denktir.

4.1.2. Kosinüs Benzerlik Skoru

Kosinüs benzerlik skoru, tam doğru haber başlığı ile sistemin ürettiği haber başlığının anlamsal açıdan ne kadar birbirine benzediğini ölçmek için kullanılmıştır. Bu skor değeri oluşturulurken, ilk önce tam doğru haber başlığı ile sistemin tahmin ettiği haber başlığının vektörel gösterimleri, FastText kütüphanesi yardımıyla elde edilmiştir. Daha sonra ise, bu iki vektör arasındaki açının kosinüs değeri alınarak 0 ile 1 arasında skor değeri elde edilmiştir. Bu skor değeri 1’e yaklaştıkça, benzerliğin arttığı, 0’a yaklaştıkça benzerliğin azaldığı anlaşılmaktadır. Kosinüs benzerlik skorunun hesaplanma formülü Denklem 4.2’de verilmiştir.

$$f_{x,y} = \cos \theta = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|} \quad (4.2)$$

Burada $\vec{x}$, tam doğru haber başlığının FastText üzerinde hesaplanan vektör gösterimini ve $\vec{y}$ ise sistemin ürettiği haber başlığının FastText üzerinde hesaplanan vektör gösterimini ifade etmektedir.

4.2. Eğitim ve Test Kümelerinin Oluşturulması

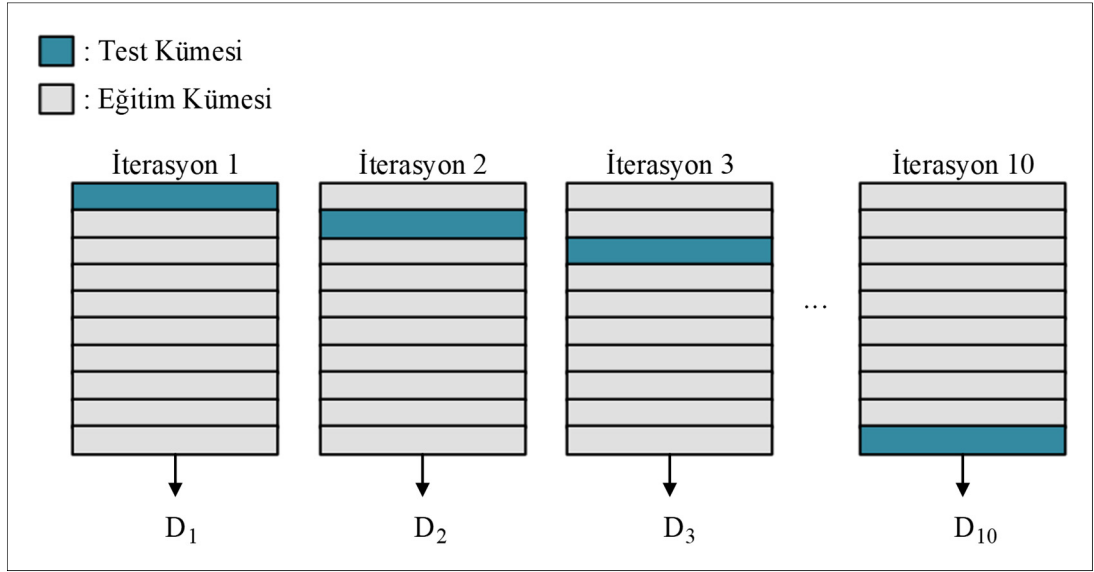
Makine öğrenmesi modellerinin iyileştirilebilmesi için en fazla çaba istatistiksel yöntemlere harcansa da, insanlar genellikle test veri kümesini yanlış şekilde seçerek, modellerinin başarılarını geçersiz hale getirmektedirler [83]. Bu sebeple, veri kümesinin eğitim ve test için en doğru şekilde ayrılması en iyi model parametrelerini seçebilmek için oldukça önemlidir. Bu tez çalışmasında, eğitim ve test kümelerinin oluşturulması için çapraz doğrulama yöntemi kullanılmıştır. Çapraz doğrulama,

makine öğrenmesi modellerinin başarısını ölçmek için kullanılan bir istatistiksel yöntemdir. Bu yöntem, kullanımın oldukça basit olması, gerçekleşmesinin kolay olması ve diğer yöntemlere göre yanlılık değerinin düşük olması sebebi ile makine öğrenmesi alanında oldukça yaygın bir şekilde kullanılmaktadır. Yöntemin k isminde tek bir parametresi bulunmaktadır. k değeri, veri kümesinin bölüneceği grup sayısını temsil etmektedir. Bu sebeple yöntem aynı zamanda k -çapraz doğrulama olarak da adlandırılmaktadır. Yöntemin adımları sırası ile şu şekildedir:

- Veri kümesini rastgele olarak karıştır
- Veri kümesini k tane gruba ayır
- Her bir grup için:
 - Grubu test kümesi olarak belirle
 - Kalan grupları eğitim kümesi olarak belirle
 - Eğitimi eğitim kümesi üzerinde gerçekleştir
 - Test aşamasını test kümesi üzerinde çalıştır ve başarıyı hesapla
- Elde edilen tüm başarı oranlarının ortalamasını alarak, tüm veri kümesine ait tek bir başarı oranı hesapla

Çapraz doğrulama yöntemi sayesinde, veri kümesindeki her bir verinin eğitim için $k - 1$ kere kullanıldığı ve test için 1 kere kullanıldığı garanti edilmiştir. k değerinin seçilmesi dikkat gerektirmektedir. Yanlış seçilen bir k değeri, yüksek varyansa veya yüksek yanlılığa sebep olabilmektedir [84]. Her ne kadar kural olarak tanımlı olmasa da k değeri genellikle 5 veya 10 olarak seçilmektedir. k değeri arttıkça, eğitim kümesi ile yeniden örneklenen alt kümeler arasında fark azalmakta ve yanlılık değerinin daha düşük olmasına neden olmaktadır [85]. k -çapraz doğrulama yöntemi Şekil 4.1’de görsel olarak anlatılmıştır. Çapraz doğrulama sonucunda doğruluk değerinin elde edilmesi Denklem 4.3’te verilmiştir.

$$D = \frac{1}{10} \sum_{i=1}^{10} D_i \quad (4.3)$$



Şekil 4.1: Çapraz doğrulama yönteminin görselleştirilmesi ($k = 10$).

Tez çalışmasında çapraz doğrulama yöntemi için $k = 10$ olarak alınmıştır. Kullanılan haber veri kümelerinin çapraz doğrulama yönteminin her bir adımı için eğitim ve test aşamasında kullanılan yaklaşık haber sayıları Tablo 4.1’de verilmiştir.

Tablo 4.1: Haber veri kümelerinin çapraz doğrulamanın her bir aşamasında eğitim ve test için kullanılan yaklaşık haber sayıları.

Veri Kümesi	Eğitim	Test
Haber Veri Kümesi A	58396	6488
Haber Veri Kümesi B	34998	3889

Çapraz doğrulama yönteminin gerçekleşmesi için, literatürde çok sık kullanılan veri madenciliği ve analizi için basit ve verimli araçlar barındıran Scikit-Learn kütüphanesinden faydalanılmıştır [86].

4.3. Sonuçlar

Sistemin deneyleri esnasında dikkat mekanizmasının ve ışın aramasının Çift Yönlü LSTM ağı üzerindeki etkilerini görebilmek için, dört farklı kombinasyon geliştirilmiştir. Geliştirilen kombinasyonlar Tablo 4.2’de verilmiştir. Işın aramasının kullanıldığı kombinasyonlarda ışın genişliği $k = 10$ olarak kullanılmıştır. Dört kombinasyonda da sözlük boyutu en fazla 30000, haber başlıkları en fazla 15 kelime ve haber metinleri en fazla 50 kelime olacak şekilde sınırlandırılmıştır.

Tablo 4.2: Deneyler esnasında geliştirilen modeller.

Deneyler	Kullanılan Kombinasyon
Deney 1	Çift Yönlü LSTM + Ağgözlü Arama
Deney 2	Çift Yönlü LSTM + Işın Araması ($k = 10$)
Deney 3	Çift Yönlü LSTM + Dikkat Mekanizması + Ağgözlü Arama
Deney 4	Çift Yönlü LSTM + Dikkat Mekanizması + Işın Araması ($k = 10$)

Haber Veri Kümesi A ve Haber Veri Kümesi B üzerinde gerçekleştirilen bu dört deneyde elde edilen ROUGE ve anlamsal benzerlik skorları sırası ile Tablo 4.3’de ve Tablo 4.4’de verilmiştir. Deney sonuçlarına göre her iki veri kümesinde de Deney 4 kombinasyonu (dikkat mekanizması ile birlikte eğitilen Çift Yönlü LSTM ve ışın araması) en yüksek başarıyı elde etmiştir.

Tablo 4.3: Haber Veri Kümesi A üzerinde gerçekleştirilen dört deneye ait sonuçlar.

Deney	ROUGE-1	Anlamsal Benzerlik
Deney 1	0.202	0.561
Deney 2	0.214	0.554
Deney 3	0.229	0.574
Deney 4	0.236	0.570

Tablo 4.4: Haber Veri Kümesi B üzerinde gerçekleştirilen dört deneye ait sonuçlar.

Deney	ROUGE-1	Anlamsal Benzerlik
Deney 1	0.109	0.428
Deney 2	0.111	0.414
Deney 3	0.115	0.441
Deney 4	0.118	0.435

Haber Veri Kümesi A’ya ait haber metni “Kayseride polis ekipleri çevreye rahatsızlık veren şahısların üzerinde yaptığı aramada farklı seriden çoğaltılmış 54 adet 100 TL’lik sahte para ele geçirdi. Edinilen bilgiye göre, bugün akşam saatlerinde

Kocasinan ilçesine bağlı İstasyon Mahallesi Depo caddesi üzerinde alkollü şahısların çevreyi rahatsız ettiği ihbarına giden Kocasinan Devriye Ekipler Amirliği ekipleri, Mustafa D. ve Sebahattin T.nin birbirlerine yüklü miktarda para verdiğini gördü. Bunun üzerine şahıslarda arama yapan ekipler, iki farklı seriden çoğaltılmış 54 adet 100 TL'lik sahte para olduğunu tespit etti. Sahte paralara el konulurken, Mustafa D. ve Sebahattin T. gözaltına alındı. Etiketler : Kayseri,sahte para” olan ve tam doğru haber başlığı “Polis ekiplerinin dikkati sahte parayı ortaya çıkardı” olan bir haberin dört farklı kombinasyon ile gerçekleştirilen modeller kullanılarak elde edilmiş deney sonuçları Tablo 4.5’de verilmiştir.

Tablo 4.5: Haber Veri Kümesi A'ya ait örnek bir haber metninin dört farklı yöntem ile elde edilmiş deney sonuçları.

Deney	Sistemin Ürettiği Haber Başlığı	ROUGE-1	Anlamsal Benzerlik
Deney 1	kablo hırsızları polis # kişiyi böyle aldı	0,14	0,65
Deney 2	kablo hırsızları polis gözaltına alındı	0,16	0,64
Deney 3	polis ekiplerinin <unk> # tl sahte para geçirdi	0,40	0,66
Deney 4	polis ekiplerinin dikkati sahte parayı bile <unk>	0,71	0,87

Haber Veri Kümesi B'ye ait haber metni “Ocak ayında Tüketici Fiyatları Endeksi (TÜFE) yüzde 1.65 artarken, Üretici Fiyatları Endeksi (ÜFE) yüzde 0.18 geriledi. Yıllık enflasyon ise TÜFE'de yüzde 7.31, ÜFE'de yüzde 1.88 oldu.” olan ve tam doğru haber başlığı “Enflasyon rakamları açıklandı” olan bir haberin dört farklı kombinasyon ile gerçekleştirilen modeller kullanılarak elde edilmiş deney sonuçları Tablo 4.6’da verilmiştir.

Tablo 4.6: Haber Veri Kümesi B'ye ait örnek bir haber metninin dört farklı yöntem ile elde edilmiş deney sonuçları.

Deney	Sistemin Ürettiği Haber Başlığı	ROUGE-1	Anlamsal Benzerlik
Deney 1	enflasyon şubatta düşük yüzde #	0,25	0,59
Deney 2	enflasyon şubatta yüzde #	0,28	0,55
Deney 3	enflasyon son # gün şubatta açıklandı	0,44	0,66
Deney 4	enflasyon rakamları açıklandı	1,00	1,00

13 Farklı kategoriden oluşan Haber Veri Kümesi B için en başarılı sonucu veren Deney 4 kombinasyonu (dikkat mekanizması ile birlikte eğitilen Çift Yönlü LSTM ve ışın araması) kullanılarak her bir kategori için ayrı ayrı elde edilen başarı oranları Tablo 4.7’de verilmiştir.

Tablo 4.7: Haber Veri Kümesi B'de yer alan her bir kategori için Deney 4 kombinasyonu ile elde edilen başarı oranları.

Sınıf İsmi	ROUGE-1	Anlamsal Benzerlik
Dünya	0.153	0.479
Ekonomi	0.126	0.439
Genel	0.090	0.399
Güncel	0.137	0.459
Kültür Sanat	0.073	0.372
Magazin	0.091	0.379
Planet	0.096	0.415
Sağlık	0.084	0.393
Siyaset	0.096	0.442
Spor	0.126	0.446
Teknoloji	0.120	0.430
Türkiye	0.131	0.460
Yaşam	0.123	0.432

Deneyleri gerçekleştirilen dört kombinasyonda sözlük boyutu en fazla 30000 olacak şekilde sınırlandırılmıştır. Sözlük boyutunun deney sonuçlarına olan etkisini görebilmek amacı ile en yüksek sonucu elde eden Deney 4 (dikkat mekanizması ile birlikte eğitilen Çift Yönlü LSTM ve ışın araması) kombinasyonu sözlük boyutu en fazla 30000, 40000, 50000 ve 60000 olacak şekilde yeniden eğitilerek başarı oranları hesaplanmıştır. Farklı sözlük boyutları için Haber Veri Kümesi A üzerinde elde edilen deney sonuçları Tablo 4.8'de ve Haber Veri Kümesi B üzerinde elde edilen deney sonuçları Tablo 4.9'da verilmiştir. Sonuçlara göre her iki haber veri kümesi için de en başarılı sonuç, 60000 sözlük boyutu ile elde edilmiştir.

Tablo 4.8: Haber Veri Kümesi A için Deney 4 kombinasyonunda sözlük boyutunun başarı oranlarına etkisi.

Sözlük Boyutu	ROUGE-1	Anlamsal Benzerlik
30000	0.236	0.570
40000	0.246	0.588
50000	0.247	0.593
60000	0.249	0.596

Tablo 4.9: Haber Veri Kümesi B için Deney 4 kombinasyonunda sözlük boyutunun başarı oranlarına etkisi.

Sözlük Boyutu	ROUGE-1	Anlamsal Benzerlik
30000	0.1178	0.435
40000	0.1180	0.440
50000	0.1180	0.442
60000	0.1183	0.447

Deneyleri gerçekleştirilen dört kombinasyonda haber başlıkları en çok 15 kelime ve haber metinleri en çok 50 kelime olacak şekilde sınırlandırılmıştır. Haber başlığı için belirlenen bu sınırın başarıya olan etkisini ölçebilmek için, sistem haber başlığı kelime sayısı sınırı en fazla 10, 15 ve 20 kelime olacak şekilde ayrı ayrı eğitilmiştir. Eğitim esnasında en yüksek başarıyı elde eden kombinasyon olan Deney 4 (dikkat mekanizması ile birlikte eğitilen Çift Yönlü LSTM ve ışıın araması) üzerine en yüksek başarıyı elde eden sözlük boyutu (her iki haber veri kümesi için de 60000) kullanılmıştır. Elde edilen deney sonuçları, Haber Veri Kümesi A için Tablo 4.10’da ve Haber Veri Kümesi B için Tablo 4.11’de verilmiştir. Sonuçlara göre Haber Veri Kümesi A için en yüksek başarı, en fazla 20 kelimedenden oluşan haber başlıkları ile ve Haber Veri Kümesi B için en yüksek başarı, en fazla 15 kelimedenden oluşan haber başlıkları ile elde edilmiştir. Deney sonuçlarından Haber Veri Kümesi A’nın ortalama haber başlığı kelime sayısının Haber Veri Kümesi B’ye göre daha fazla olduğu anlaşılmıştır.

Tablo 4.10: Haber Veri Kümesi A için en fazla haber başlığı kelime sayısının başarı oranlarına etkisi.

En Fazla Haber Başlığı Kelime Sayısı	ROUGE-1	Anlamsal Benzerlik
10	0.244	0.594
15	0.249	0.596
20	0.250	0.599

Tablo 4.11: Haber Veri Kümesi B için en fazla haber başlığı kelime sayısının başarı oranlarına etkisi.

En Fazla Haber Başlığı Kelime Sayısı	ROUGE-1	Anlamsal Benzerlik
10	0.1179	0.4469
15	0.1183	0.4471
20	0.1155	0.4460

Haber metinleri için belirlenen en çok 50 kelime sınırının başarıya olan etkisini ölçebilmek için, sistem haber metni kelime sayısı sınırı en fazla 50 ve 100 olacak şekilde ayrı ayrı eğitilmiştir. Eğitim esnasında en yüksek başarıyı elde eden kombinasyon olan Deney 4 (dikkat mekanizması ile birlikte eğitilen Çift Yönlü LSTM ve ışın araması) üzerine en yüksek başarıyı elde eden sözlük boyutu (her iki haber veri kümesi için de 60000) üzerine en yüksek başarıyı elde eden haber başlığı en fazla kelime boyutu (Haber Veri Kümesi A için 20 ve Haber Veri Kümesi B için 15) kullanılmıştır. Elde edilen deney sonuçları Haber Veri Kümesi A için Tablo 4.12’de ve Haber Veri Kümesi B için Tablo 4.13’te verilmiştir. Sonuçlara göre Haber Veri Kümesi A için en yüksek başarı, en fazla 50 kelimeden oluşan haber metinleri ile ve Haber Veri Kümesi B için en yüksek başarı, en fazla 50 kelimeden oluşan haber metinleri ile elde edilmiştir.

Tablo 4.12: Haber Veri Kümesi A için en fazla haber metni kelime sayısının başarı oranlarına etkisi.

En Fazla Haber Metni Kelime Sayısı	ROUGE-1	Anlamsal Benzerlik
50	0.250	0.599
100	0.243	0.593

Tablo 4.13: Haber Veri Kümesi B için en fazla haber metni kelime sayısının başarı oranlarına etkisi.

En Fazla Haber Metni Kelime Sayısı	ROUGE-1	Anlamsal Benzerlik
50	0.118	0.447
100	0.115	0.444

Her iki haber veri kümesi için de en yüksek başarının elde edildiği kombinasyonların test aşamasında ışın araması kullanılmıştır. Işın aramasının ışın genişliği $k = 10$ olarak alınmıştır. Her ne kadar $k = 10$, yaygın olarak kullanılıyor olsa da, farklı k değerlerinin başarıya etkisini görebilmek için $k = 5$ ve $k = 10$ olacak şekilde ayrı ayrı deneyler gerçekleştirilmiştir. Deneylerin eğitimi esnasında en yüksek başarıyı elde eden kombinasyon olan Deney 4 (dikkat mekanizması ile birlikte eğitilen Çift Yönlü LSTM ve ışın araması) üzerine en yüksek başarıyı elde eden sözlük boyutu (her iki veri kümesi için de 60000) üzerine en yüksek başarıyı elde eden haber başlığı en fazla kelime boyutu (Haber Veri Kümesi A için 20 ve Haber Veri Kümesi B için 10) üzerine en yüksek başarıyı elde eden haber metni en fazla kelime boyutu (her iki haber veri kümesi için de 50) kullanılmıştır. Gerçekleştirilen deneylere ait sonuçlar Haber Veri Kümesi A için Tablo 4.14'te ve Haber Veri Kümesi B için Tablo 4.15'te verilmiştir. Sonuçlara göre Haber Veri Kümesi A için ışın genişliğinin $k = 5$ olduğu durumda ve Haber Veri Kümesi B için $k = 5$ olduğu durumda en yüksek başarı oranları elde edilmiştir.

Tablo 4.14: Haber Veri Kümesi A için ışıın araması algoritmasının ışıın genişliğinin başarı oranlarına etkisi.

İşin Araması İşin Genişliği	ROUGE-1	Anlamsal Benzerlik
$k = 5$	0.24992	0.600
$k = 10$	0.24987	0.599

Tablo 4.15: Haber Veri Kümesi B için ışıın araması algoritmasının ışıın genişliğinin başarı oranlarına etkisi.

İşin Araması İşin Genişliği	ROUGE-1	Anlamsal Benzerlik
$k = 5$	0.1184	0.4476
$k = 10$	0.1183	0.4471

Türkçe dilinin yapısı gereği, veri kümelerinde yer alan kelimelerde pek çok ek bir arada bulunabilmektedir. Türkçe’de aynı kelime farklı eklerle kullanılabilmekte ve benzer anlamlara gelebilmektedir. ROUGE değerlendirme metriği, başarı oranını hesaplarken, doğrudan kelime eşleşmesine bakmaktadır. Bu sebeple, iki ayrı cümlede aynı kelimeler farklı eklerle kullanılmış olsa da başarı oranını oldukça düşük hesaplamaktadır. Kelimelerde yer alan eklerin başarı oranlarına etkisini ölçebilmek için, sistemin test aşamasında haber başlıklarında yalnızca kelimelerin kökleri kullanılarak deneyler gerçekleştirilmiştir. Deneylerin eğitimi esnasında en yüksek başarıyı elde eden kombinasyon olan Deney 4 (dikkat mekanizması ile birlikte eğitilen Çift Yönlü LSTM ve ışıın araması) üzerine en yüksek başarıyı elde eden sözlük boyutu (her iki haber veri kümesi için de 60000) üzerine en yüksek başarıyı elde eden haber başlığı en fazla kelime boyutu (Haber Veri Kümesi A için 20 ve Haber Veri Kümesi B için 15) üzerine en yüksek başarıyı elde eden haber metni en fazla kelime boyutu (her iki haber veri kümesi için de 50) kullanılmıştır. Test aşamasında ise her iki haber veri kümesi için de ışıın genişliği $k = 5$ olarak kullanılmıştır. Elde edilen deney sonuçları Haber Veri Kümesi A için Tablo 4.16’da ve Haber Veri Kümesi B için Tablo 4.17’de verilmiştir. Sonuçlara göre her iki haber veri kümesi için de en yüksek başarı sadece kök içeren haber başlıkları ile elde edilmiştir. Deney sonuçlarına göre, sadece kök kullanılması durumunda ROUGE skoru artsa da cümlelerin anlamında meydana gelen değişimden dolayı anlamsal benzerliğin düştüğü görülmüştür.

Tablo 4.16: Haber Veri Kümesi A için deneyler esnasında haber başlığı kelimelerinin ekli kullanılma durumunun başarı oranlarına etkisi.

Haber Başlığı Kelimelerinin Ekli Kullanılma Durumu	ROUGE-1	Anlamsal Benzerlik
Kök ve Ekler	0.250	0.600
Sadece Kök	0.272	0.560

Tablo 4.17: Haber Veri Kümesi B için deneyler esnasında haber başlığı kelimelerinin ekli kullanılma durumunun başarı oranlarına etkisi.

Haber Başlığı Kelimelerinin Ekli Kullanılma Durumu	ROUGE-1	Anlamsal Benzerlik
Kök ve Ekler	0.118	0.448
Sadece Kök	0.124	0.421

Gerçekleştirilen tüm deneylerin sonuçlarına göre, sistem Haber Veri Kümesi A için en yüksek başarı eğitim için Deney 4 kombinasyonu (dikkat mekanizması ile birlikte eğitilen Çift Yönlü LSTM ve ışın araması), 60000 sözlük boyutu, en fazla 20 kelime uzunluğunda haber başlığı, en fazla 50 kelime uzunluğunda haber metni ile elde edilirken; test için ise ışın araması ışın genişliği $k = 5$ olarak ve sadece kök içeren haber başlıkları ile elde edilmiştir. Haber Veri Kümesi B için ise en yüksek başarı, eğitim için Deney 4 kombinasyonu (dikkat mekanizması ile birlikte eğitilen Çift Yönlü LSTM ve ışın araması), 60000 sözlük boyutu, en fazla 15 kelime uzunluğunda haber başlığı, en fazla 50 kelime uzunluğunda haber metni ile elde edilirken; test için ışın araması ışın genişliği $k = 5$ olarak ve sadece kök içeren haber başlıkları ile elde edilmiştir. Elde edilen deney sonuçlarına göre, en başarılı sonuçların elde edildiği model ile Haber Veri Kümesi A üzerinde ROUGE-1 için ortalama 0,27 skor değeri, anlamsal benzerlik için ise ortalama 0,60 skor değeri ve Haber Veri Kümesi B üzerinde ROUGE-1 için ortalama 0,12 skor değeri ve anlamsal benzerlik için ise ortalama 0,44 skor değeri elde edilmiştir.

En iyi başarı oranına sahip model kombinasyonu kullanılarak elde edilen Haber Veri Kümesi A'ya ait bazı pozitif deney sonuçları Şekil 4.2, Şekil 4.3, Şekil 4.4, Şekil 4.5 ve Şekil 4.6'da, bazı negatif deney sonuçları ise Şekil 4.7, Şekil 4.8 ve Şekil 4.9'da verilmiştir.

Haber Metni	Aydında polis ekipleri tarafından yapılan operasyonda 400 yıllık olduğu tahmin edilen Tevrat ele geçirildi. İHA Edinilen bilgiye göre, Aydın Emniyet Müdürlüğü Kaçakçılık ve Organize Suçlarla Mücadele Şube Müdürlüğü ekipleri Efeler ilçesinde G.D isimli şahsın ikametinde 400 yıllık Tevrat olduğu ihbarı üzerine harekete geçti. Polis ekipleri tarafından G.Dnin ikametinde yapılan aramada, beyaz poşette dışı metal süslemeli üzerinde bordo renkli taşlar bulunan, içerisinde İbranice olduğu değerlendirilen deri sayfalar bulunan eski kitap bulundu. Ele geçirilen ve Tevrat olduğu tahmin edilen tarihi kitap muhafaza altına alınarak, Aydın Müze Müdürlüğüne teslim edilirken, kitabı evinde bulunduran G.D isimli şahıs gözaltına alındı. Müze Müdürlüğü yetkilileri tarafından teslim alınan tarihi kitabın incelenmesinin ardından içeriğinin belirleneceği öğrenildi.
Tam Doğru Haber Başlığı	Aydında 400 yıllık Tevrat ele geçirildi
Sistemin Ürettiği Haber Başlığı	aydında # yıllık tevatat polisler tarafından ele geçirildi
ROUGE-1	0,85
Anlamsal Benzerlik	0,89

Şekil 4.2: Haber Veri Kümesi A'ya ait pozitif bir deney sonucu.

Haber Metni	Hazine ve Maliye Bakanlığınca yarın 2 ihale gerçekleştirilecek. AA Ekonomi Haberleri Hazine ve Maliye Bakanlığının internet sitesinde yer alan duyuruya göre, yarınki ilk ihaleyle 11 ay (336 gün) vadeli, kuponsuz devlet tahvili yeniden ihraç edilecek. İkinci ihalede ise 2 yıl (728 gün) vadeli kira sertifikasının doğrudan satışı yapılacaktır.
Tam Doğru Haber Başlığı	Hazine yarın 2 ihale düzenleyecek
Sistemin Ürettiği Haber Başlığı	hazine # ihale düzenleyecek
ROUGE-1	0,88
Anlamsal Benzerlik	0,92

Şekil 4.3: Haber Veri Kümesi A'ya ait pozitif bir deney sonucu.

Haber Metni	Brent petrolün varili uluslararası piyasalarda 63,12 dolara yükseldi. Dün 61,57 dolardan güne başlayan ve gün içinde 63,41 dolara kadar çıkan Brent petrolün varil fiyatı, bu sabah 62,51 dolardan açıldı. Brent petrolün varili saat 09.20 itibarıyla yüzde 0,99 artışla 63,12 dolardan işlem görüyor. Aynı dakikalarda Batı Teksas türü (WTI) ham petrolün varili 53,62 dolardan alıcı buldu. AAnın haberine göre; Suudi Arabistan Enerji Bakanı Halid El-Falihin, martta ülkenin günlük petrol üretimini yaklaşık 500 bin varil azaltacaklarını açıklamasıyla küresel petrol fiyatları yükselişe geçti. (...)
Tam Doğru Haber Başlığı	Petrol fiyatı 63 doların üzerinde
Sistemin Ürettiği Haber Başlığı	petrol fiyatları # doların üzerini gördü
ROUGE-1	0,90
Anlamsal Benzerlik	0,94

Şekil 4.4: Haber Veri Kümesi A'ya ait pozitif bir deney sonucu.

Haber Metni	Mersinde yüzü maskeli bir kişi soyamak için girdiği kuyumcуда bulunan iş yeri sahibi ile bir kişiyi öldürdü. Bir miktar altın alan maskeli soyguncu izini kaybettirdi. Olay, Akdeniz ilçesi Hastane Caddesi üzerinde meydana geldi. Edinilen bilgiye göre, akşam saatlerinde yüzü maskeli bir kişi kuyumcuya girdi. Zanlı, kuyumcуда bulunan para ve altınları almak istedi. Ancak kuyumcu sahibi direnince yüzü maskeli kişi tabancayla iş yeri sahibi ve bir kişiyi tabancayla vurarak yaraladı. Silah sesleri üzerine vatandaşlar durumu polis ve sağlık ekiplerine bildirdi. Yüzü maskeli soyguncu ise bir miktar altın alarak izini kaybettirdi. Olay yerine gelen sağlık ekipleri iş yeri sahibi ve vurulan diğer kişinin de öldüğünü belirledi. Olayla ilgili soruşturma devam ediyor.
Tam Doğru Haber Başlığı	Mersinde kanlı soygun: 2 ölü
Sistemin Ürettiği Haber Başlığı	mersinde kanlı kuyumcu soygunu # ölü
ROUGE-1	0,90
Anlamsal Benzerlik	0,95

Şekil 4.5: Haber Veri Kümesi A'ya ait pozitif bir deney sonucu.

Haber Metni	Şampiyon Rus boksör Alexander Kostromin, evinin yakınında ölü olarak bulundu. Rus boksör Alexander Kostromin, öldürüldü. Rus boksörün cansız bedeni dün akşam saatlerinde Rusyanın Belgorod şehrindeki evinin hemen yakınında bulundu. Yerel basın, boksörün 8 kurşun ile vurulduğunun belirlendiğini; cinayeti işleyen kişinin ise araba ile olay yerinden kaçtığının tespit edildiğini yazdı. Rus boksör Alexander Kostromin ismini ilk kez 2005teki müsabakada rakibini yenerek duyurmuştu. 2014te Kırımda düzenlenen uluslararası boks şampiyonasında altın madalya kazanmıştı. 81 kiloda yarışan boksör, 16 Şubatta Letonyanın başkenti Rigada düzenlenecek karşılaşmaya hazırlanıyordu. (...)
Tam Doğru Haber Başlığı	Şampiyon Rus boksör Alexander Kostromin öldürüldü
Sistemin Ürettiği Haber Başlığı	şampiyon rus boksör ölü bulundu
ROUGE-1	0,54
Anlamsal Benzerlik	0,77

Şekil 4.6: Haber Veri Kümesi A'ya ait pozitif bir deney sonucu.

Haber Metni	TEKİRDAĞ'ın Çorlu ilçesinde sabah saatlerinden itibaren etkili olan yağmur, hayatı olumsuz etkiledi. Hava sıcaklığının 10 derece olduğu Çorluda sabah saatlerinde başlayan yağmur, etkisini sürdürüyor. Yağış nedeniyle rögarların taşmasıyla bazı cadde ve sokaklarda su birikintileri oluştu. Su dolan yollarda, belediye ekipleri iş makineleri ile su tahliyesi yaptı. Yağmur nedeniyle araç sürücülerini zor anlar yaşadı. Yağmura hazırlıksız yakalanan vatandaşlar ise saçak altlarına sığındı. Yetkililer, olası su baskınlarına karşı vatandaşları uyardı. Yağmurun aralıklarla devam edeceği bildirildi.
Tam Doğru Haber Başlığı	Çorluda yağmur, hayatı olumsuz etkiledi
Sistemin Ürettiği Haber Başlığı	ada namazı anlar yıkılma dolu yaşattı
ROUGE-1	0,00
Anlamsal Benzerlik	0,31

Şekil 4.7: Haber Veri Kümesi A'ya ait negatif bir deney sonucu.

Haber Metni	Zonguldak K�lt�r ve Eēitim Vakfı (ZOKEV) 2018 yılı K�lt�r �d�l� yazar Hamit Kalyoncuya, Eēitim �d�l� de �aycuma Őehit Hasan Yaēlı İlkokulu �ēretmenlerinden Zeynep ve Ali Hasırcıya Zonguldak Maden M�hendisleri Odası Lokalinde yapılan geniŐ katılımlı bir t�renle verildi. Ali Ayaroēlu 13:26 �d�l�n, Hamit Kalyoncuya, kentin sosyal, k�lt�rel geliŐimine yaptığı eŐsiz hizmetler; Zeynep ve Ali Hasırcıya da olumsuz hava koŐulları, hastalık dahil hi�bir mazeretin arkasına sığınmadan her �ērenciye ulaŐarak bilim �ēretmek i�in harcadıkları olaēan�st� �abalar nedeniyle verildi. T�rene CHP Zonguldak Belediye BaŐkan Adayı Őenol Őanalda katıldı. (...)
Tam Doēru Haber BaŐlıēı	ZOKEV �d�lleri sahiplerini buldu
Sistemin �rettiēi Haber BaŐlıēı	<unk> bayrak eēitimi
ROUGE-1	0,00
Anlamsal Benzerlik	0,02

Őekil 4.8: Haber Veri K mesi A'ya ait negatif bir deney sonucu.

Haber Metni	Őırnak Silopi Belediyesinde g�venliēi saēlamakla g�revli polis memuru, belediye bah�esinde doēum yaptıktan sonra �len k�peēin yavrularına sahip �ıktı. YaklaŐık 1 ay �nce doēum yapan ve doēumun ardından �len k�peēin geride kalan 5 yavrusu i�in polis memuru Cevdet Tun�un talebi �zerine Silopi Belediyesinin bah�esinde barınak yapıldı. Barınakta kalan k�pek yavrularına őefkat g�steren Tun�, yavruları biberonla, bebe bisk�visi ve s�tle hazırladıēı mamalarla besliyor. Tun�, g�n i�erisinde mesaisinden arta kalan zamanlarda k�pek yavrularını barınaktan �ıkararak, onlarla yakından ilgileniyor. K�peklerin g�nl�k ihtiya�larını karŐılamada, belediye �alıŐanları da destek veriyor. (...)
Tam Doēru Haber BaŐlıēı	K�pek yavrularına polis őefkati
Sistemin �rettiēi Haber BaŐlıēı	�atıdan a� oldu
ROUGE-1	0,00
Anlamsal Benzerlik	0,16

Őekil 4.9: Haber Veri K mesi A'ya ait negatif bir deney sonucu.

En iyi başarı oranına sahip model kombinasyonu kullanılarak elde edilen Haber Veri Kümesi B'ye ait bazı pozitif deney sonuçları Şekil 4.10, Şekil 4.11, Şekil 4.12, Şekil 4.13 ve Şekil 4.14'te, bazı negatif deney sonuçları ise Şekil 4.15, Şekil 4.16 ve Şekil 4.17'de verilmiştir.

Haber Metni	Ermenistan cumhurbaşkanı adaylarından Paruir Ayrakyan silahlı saldırıya uğradı. Ermenistan'da 18 Şubat'ta yapılacak Cumhurbaşkanlığı seçimlerinde yarışacak adaylardan Milli İrade Birliği Başkanı Paruir Ayrakyan'ın Erivan'da uğradığı silahlı saldırı sonucu vücuduna bir kurşun isabet ettiği ve hastaneye kaldırıldığı bildirildi. Kaldırıldığı Aziz Gregory Lusavoriç Hastanesi'nden yapılan açıklamada, Ayrakyan'ın bilincinin açık, durumunun ise orta ağır olduğu belirtildi. (...)
Tam Doğru Haber Başlığı	Cumhurbaşkanı adayına silahlı saldırı!
Sistemin Ürettiği Haber Başlığı	ermenistan cumhurbaşkanı adayına silahlı saldırı
ROUGE-1	0,88
Anlamsal Benzerlik	0,94

Şekil 4.10: Haber Veri Kümesi B'ye ait pozitif bir deney sonucu.

Haber Metni	Çin ordusunun ve partinin başkanı seçilen Cumhurbaşkanı Yardımcısı Xi Jinping, orduya gerçek savaşa hazır olmaları emrini verdi. Parti başkanlığına geldikten sonra geçtiğimiz hafta Guangdong eyaletindeki dört günlük ilk denetleme gezisinde Xi, askerlerin hazırlıklı olması için gerçek savaşın farkındalığını artırma emrini verdi. Çin ordusu 2.3 milyonla dünyanın en kalabalık ordusu konumunda buluyor. Xi, ordudaki asker ve subaylara gerçek savaş kriterlerini benimsemeleri için eğitimlerde söz konusu farkındalığı artırmalarını istedi. (...)
Tam Doğru Haber Başlığı	Xiden Çin ordusuna: Gerçek savaşa hazır olun!
Sistemin Ürettiği Haber Başlığı	<unk> çin ordusuna gerçek savaşa hazır olun
ROUGE-1	0,75
Anlamsal Benzerlik	0,75

Şekil 4.11: Haber Veri Kümesi B'ye ait pozitif bir deney sonucu.

Haber Metni	Çin'de yeni Devlet Başkanı Şi Cinping oldu. Ulusal Halk Kongresi'nde delegelerin devlet başkanı olarak belirlediği Şi Cinping, Çin Merkezi Askeri Komisyonu Başkanı olarak da seçildi. Çin Komünist Partisi (ÇKP) Merkez Komitesi Politbüro Üyesi Li Yüençao da devletbaşkan yardımcılığına seçildi. Ülkede geçen yıl yapılan ÇKP 18. Ulusal Halk Kongresi'nin ardından devlet başkanı yardımcısı olan Şi Cinping, ÇKP Genel Sekreterliği'ne getirilmişti. Böylece, Şi'nin devlet başkanı olarak seçileceği kesinleşmişti.
Tam Doğru Haber Başlığı	Çin'in yeni Devlet Başkanı Şi Cinping oldu
Sistemin Ürettiği Haber Başlığı	çin in yeni devlet başkanı belli oldu
ROUGE-1	0,79
Anlamsal Benzerlik	0,91

Şekil 4.12: Haber Veri Kümesi B'ye ait pozitif bir deney sonucu.

Haber Metni	Trafiğe kaydı yapılan motorlu taşıt sayısı Ekim ayında bir önceki aya göre yüzde 1,8 artarak 86 bin 333 oldu. Ocak-Ekim döneminde trafikteki toplam taşıt sayısı 812 bin 257 artarken, trafiğe kayıtlı toplam araç sayısı ise 16 milyon 901 bin 785 oldu. Türkiye İstatistik Kurumu (TÜİK), Ekim ayına ait Motorlu Kara Taşıtları İstatistikleri'ni yayınladı. Ekim ayında trafiğe kaydı yapılan toplam 86 bin 333 taşıt içinde otomobil yüzde 54,7 pay ve 47 bin 250 adet ile ilk sırada yer alırken; otomobili yüzde 17,7 pay ve 15 bin 294 adet ile kamyonet, yüzde 16,4 pay ve 14 bin 177 adet ile motosiklet, yüzde 4,8 pay ve 4 bin 130 adet ile traktör takip etti. Ekim ayında trafiğe kaydı yapılan taşıtların yüzde 6,3 pay ve 5 bin 482 adedini ise minibüs, otobüs, kamyon ve özel amaçlı taşıtlar oluşturdu. (...)
Tam Doğru Haber Başlığı	Trafiğe kayıtlı araç sayısı 16,9 milyonu aştı
Sistemin Ürettiği Haber Başlığı	trafikteki araç sayısı ekimde arttı
ROUGE-1	0,49
Anlamsal Benzerlik	0,78

Şekil 4.13: Haber Veri Kümesi B'ye ait pozitif bir deney sonucu.

Haber Metni	Türk Hava Yollarının (THY) Şanghay- İstanbul seferini yapan TK-27 sefer sayılı uçağı Hazar Denizi üzerindeyken kabin basıncının düşmesi nedeniyle İstanbul Atatürk Havalimanı yerine Ankara Esenboğa Havalimanına zorunlu olarak indi. Yolcular 20 dakika süreyle oksijen maskesiyle yolculuk yaptı. Yolculardan bazıları da o anları cep telefonu ile çekti. THYden alınan bilgiye göre içinde 272 yolcusu bulunan TC- JJK kuyruk tescilli, Boeing 777-ER tipi 'Akdeniz' adlı uçağın Hazar Denizi üzerindeyken kabin içi basıncı düştü. Oksijen maskeleri otomatik olarak açıldı ve yolcular kabin görevlilerinin kontrolünde maskeleri taktılar. Pilot, yolculara uçağın teknik arıza nedeniyle Ankaraya ineceği bilgisini verdi. 20 dakika süreyle yolcular maskelerle yolculuk yaptı, uçak normal oksijenin alınabildiği irtifaya inince maskeler çıkartıldı. (...)
Tam Doğru Haber Başlığı	Uçağın kabin basıncı düştü, uçak Ankara'ya indi
Sistemin Ürettiği Haber Başlığı	thy uçağı acil iniş
ROUGE-1	0,18
Anlamsal Benzerlik	0,69

Şekil 4.14: Haber Veri Kümesi B'ye ait pozitif bir deney sonucu.

Haber Metni	Her yıl Oscar Ödüllerinden iki hafta önce dağıtılan BAFTA ödülleri, geçtiğimiz akşam Londradaki Royal Opera Houseda düzenlenen törenle sahiplerini buldu. Sunuculuğunu İngiliz komedyen ve aktör Stephen Fryın yaptığı gecede İngiliz şarkıcı Paloma Faith de sahne aldı. Törene Hugh Jackman, George Clooney, Daniel Day-Lewis, Jessica Chastain, Ben Affleck, Helen Mirren, Judi Dench, Anne Hathaway, Sarah Jessica Parker, Sally Field ve Samuel L. Jackson gibi ünlü isimler katıldı. (...)
Tam Doğru Haber Başlığı	Oscar provası
Sistemin Ürettiği Haber Başlığı	<unk> sahiplerini buldu
ROUGE-1	0,00
Anlamsal Benzerlik	0,07

Şekil 4.15: Haber Veri Kümesi B'ye ait negatif bir deney sonucu.

Haber Metni	Çağlar: Eczacılar, Medula'ya alternatif çözüm istiyor. Bölge Trabzon Eczacı Odası Başkanı Sabih Tekin Çağlar, Sosyal Güvenlik Kurumu'nun (SGK), Medula sistemini kesintiye uğratmadan çalışmasını sağlaması ve sistem arızalandığında da devreye girebilecek bir yedek sistem kurması gerektiğini söyledi. Çağlar, bunların Medula'nın eksiklerinin giderilmesine kadar da zorunlu e-reçete sistemine geçişin ertelenmesi gerektiğini savundu. Sabih Tekin Çağlar, yaptığı yazılı açıklamada, SGK'nın başlattığı uygulama ile 1 Temmuz tarihinden itibaren elektronik reçete sistemine geçildiğini hatırlattı. (...)
Tam Doğru Haber Başlığı	Eczacılar dertli
Sistemin Ürettiği Haber Başlığı	çalıntı <unk> için tatil
ROUGE-1	0,00
Anlamsal Benzerlik	0,08

Şekil 4.16: Haber Veri Kümesi B'ye ait negatif bir deney sonucu.

Haber Metni	İZMİRde 3 yıl önce 3üncü kattan kendini boşluğa bırakıp yaralı kurtulan 24 yaşındaki Ayşegül Aşkın, bu kez bir otelin 5inci katının balkonundan atladı. Otelin giriş kapısındaki camlı bölümün üzerine düşen Aşkın, hayatını kaybetti. Ayşegül Aşkın, 27 Ocak 2010da, erkek arkadaşı ile yaşadığı tartışmanın ardından, kendini üçüncü kattan boşluğa bıraktı. Kameralarca görüntülenen o anların ardından hastaneye kaldırılan Aşkın sağ bacağında kalıcı hasar meydana geldi. Hastanedeki tedavisi tamamlanan Ayşegül Aşkın, itfaiye görevlileri hakkında, olay anında yavaş davrandıkları ve hava yastığını geç açtıkları iddiasıyla suç duyurusunda bulundu ancak dava açılmasına yer olmadığına karar verildi. (...)
Tam Doğru Haber Başlığı	3 yetmedi 5 öldürdü
Sistemin Ürettiği Haber Başlığı	kanlı toprak kayması ağır
ROUGE-1	0,00
Anlamsal Benzerlik	0,05

Şekil 4.17: Haber Veri Kümesi B'ye ait negatif bir deney sonucu.

Şekil 4.2, Şekil 4.3, Şekil 4.4 ve Şekil 4.5'te yer alan örnek sonuçlardan görülebileceği üzere, sistem sayı geçebilecek yerleri başarılı bir şekilde tahmin edebilmiştir. Eğitilen model, haberlerin ön işlemden geçirilmesi sırasında sayılara ortak anlam yüklemek adına sayıların # karakteri ile yer değiştirilmesi sayesinde, haber başlıklarında "ölü", "yaralı" gibi kelimelerden önce sayı geldiğini öğrenmiştir. Eğitimde kullanılan haber veri kümelerinde görüldüğü üzere haberler en çok sayı dizgeciğinden oluşmaktadır. Dolayısıyla sayılara ortak anlam yüklenmesi, tahmin edilen özetlerin daha anlamlı ve dil bilgisine uygun halde üretilmesine katkı sağlamıştır.

Şekil 4.11'de "Xiden" kelimesinin nadir geçen kelime olmasından kaynaklı sözlükte yer almamasından dolayı bilinmeyen kelime (<unk>) olarak işaretlendiği gözlemlenmiştir.

Şekil 4.14'teki gibi bazı deney sonuçlarından görülebileceği üzere, tam doğru haber başlığı ile sistemin tahmin ettiği haber başlığı arasındaki ROUGE-1 skoru oldukça düşük çıkmıştır. ROUGE-1 skoru direkt olarak tam doğru haber başlığı ve sistemin tahmin ettiği haber başlığı arasındaki kelimelerin birebir eşleşmesine bakmaktadır. Bu sebeple, tahmin edilen haber başlığı tam doğru haber başlığı ile anlamsal olarak benzese de birebir eşleşmemesinden dolayı, ROUGE-1 skoru oldukça düşük çıkabilmektedir. Ancak, sistemin ürettiği haber başlığına ve orijinal haber metnine bakıldığında, üretilen haber başlığının habere uygun bir özet olduğu görülmektedir. Bu sonuç, ikinci bir değerlendirme metriği olarak kullanılan anlamsal benzerlik metriğinin gerekliliğini göstermektedir.

Şekil 4.7, Şekil 4.8, Şekil 4.9, Şekil 4.15, Şekil 4.16 ve Şekil 4.17'de görüldüğü üzere sistemin her zaman haber içeriğine ve dilbilgisi kurallarına uygun cümle kuramadığı gözlemlenmiştir.

5. SONUÇ VE TARTIŞMA

Bu tez çalışmasında, derin öğrenme yöntemleri kullanılarak Türkçe haber metinlerinden haber başlığı üreten bir sistem geliştirilmiştir. Sistem gerçekleştirirken, metin özetleme yaklaşımlarından yoruma dayalı metin özetleme yaklaşımı kullanılmıştır. Sistemin eğitim aşamasında, LSTM modeli ve Çift Yönlü LSTM modeli dikkat mekanizması ile birlikte kullanılmış, farklı sözlük boyutları, haber başlığı boyutları ve haber metni boyutları ile ayrı ayrı eğitim gerçekleştirilmiştir. Test aşamasında ise hem açgözlü arama, hem de ışın araması yöntemleri kullanılmıştır. İki ayrı ölçüm metriği kullanılan test aşamasında doğrudan kelime eşleşmesine bakan ROUGE metriğinin yanı sıra, anlamsal benzerliği de ölçen yeni bir metrik önerilmiştir. Sistemin deneyleri, iki ayrı veri kümesi üzerinde gerçekleştirilmiştir. Elde edilen sonuçlara göre sistem, her iki veri kümesinde de, eğitim için Çift Yönlü LSTM modelinin dikkat mekanizması ile birlikte kullanıldığı durumda ve test için ışın aramasının kullanıldığı durumda en başarılı sonucu üretmiştir.

Literatürde yer alan derin öğrenme tabanlı önceki çalışmalara bakıldığında, her ne kadar İngilizce dili için epey çalışma mevcut ise de, Türkçe dili için metin özetlemenin bir alt kümesi olan haber başlığı üreten çalışmalar mevcut değildir. Bunun nedeni, Türkçe dilinin sondan eklemeli olmasından dolayı zor bir dil olmasıdır. Ayrıca Türkçe dilinde, metin özetleme için derin öğrenme algoritması ile iyi sonuçlar elde edebilecek kadar kapsamlı bir veri kümesi de bulunmamaktadır. Önerilen yöntem sayesinde, Türkçe haber metinleri için haber başlıkları elde edilmiş ve literatüre yeni bir çalışma alanı kazandırılmıştır. Ayrıca, uygulanan anlamsal benzerlik metriği ile literatüre haber başlığı üretme alanında yeni bir metrik kazandırılmıştır.

Çalışma kapsamında uygulanan yöntem, gerçekleştirilen deneylerden elde edilen çıktılara göre başarılı sonuçlar vermiştir. Deneylerin sonuçlarından anlaşılaacağı üzere; anlamsal benzerlik skorlarının ROUGE skorlarına göre daha yüksek çıktığı görülmektedir. Sistemin ürettiği özetler arasında ROUGE skoru sıfıra yakın olup (yani negatif sonuçlar), dilbilgisi ve anlamsal benzerlik bakımından doğru özetler bulunmakla birlikte (yani yanlış-negatif sonuçlar), anlamsal benzerlik açısından da farklı olan sonuçlar (yani doğru-negatif) da gözlenmektedir. Bu bağlamda, ROUGE skoru doğrudan kelime eşleşmesine bakılarak hesaplandığı için, tek başına yeterli bir değerlendirme metriği olamamaktadır. Bu durum metin özetleme alanında yeni değerlendirme metriklerinin tanımlanması gerektiğini açıkça ortaya koymaktadır. Bu

nedenle bu alanda yeni metrik tanımlamaları üzerinde çalışmalar yapılması planlanmaktadır. Ayrıca, gelecek çalışmalarda Türkçe haber başlığı üretme sistemleri için veri kümesi oluşturulması da planlanmaktadır.

Metin özetleme, bir metnin birden fazla özetinin olabilmesi sebebiyle, diğer doğal dil işleme problemlerine göre daha zor bir problemdir. Günümüzde halen yoruma dayalı metin özetleme, doğal dil işleme alanında tamamen çözülebilmüş bir problem değildir. Ancak, yapay sinir ağları ve derin öğrenme algoritmalarının gelişmesiyle birlikte, yoruma dayalı metin özetleme alanında halen sıklıkla çalışılmaya devam edilmektedir.

KAYNAKLAR

- [1] Torres-Moreno J.-M., (2014), “Automatic Text Summarization”, 1st Edition, Wiley-ISTE.
- [2] Luhn H. P., (1958), “The Automatic Creation of Literature Abstracts”, IBM Journal of Research and Development, 2 (2), 159-165.
- [3] Alemany L. A., Castellón I., Roca S. C., Fort M. F., Padró L., Rodríguez H., (2004), “Approaches to Text Summarization: Questions and Answers”, Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial, 8 (22), 79-102.
- [4] Hovy E., Lin C.-Y., Zhou L., Fukumoto J., (2006), “Automated Summarization Evaluation with Basic Elements”, Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, May 24-26.
- [5] Allahyari M., Pouriyeh S. A., Assefi M., Safaei S., Trippe E. D., Gutierrez J. B., Kochut K., (2017), “Text Summarization Techniques: A Brief Survey”, International Journal of Advanced Computer Science and Applications, 8, 397-405.
- [6] Mani I., Maybury M. T., (1999), “Advances in Automatic Text Summarization”, 1st Edition, Cambridge: MIT Press.
- [7] Lopyrev K., (2015), “Generating News Headlines with Recurrent Neural Networks”, arXiv preprint arXiv:1512.01712.
- [8] Karakoç E., Yılmaz B., (2019), “Deep Learning Based Abstractive Turkish News Summarization”, 27th Signal Processing and Communications Applications Conference (SIU), Sivas, Turkey, April 24-26.
- [9] Grave E., Bojanowski P., Gupta P., Joulin A., Mikolov T., (2018), “Learning Word Vectors for 157 Languages”, Proceedings of the International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan, May 7-12.
- [10] Sutskever I., Vinyals O., Le Q. V., (2014), “Sequence to sequence learning with neural networks”, Advances in Neural Information Processing Systems, 2, 3104-3112.
- [11] Nallapati R., Zhou B., Santos C. d., Gulcehre C., Xiang B., (2016), “Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond”, Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany, August 11-12.

- [12] Jain L. C., Medsker L. R., (1999), “Recurrent Neural Networks: Design and Applications”, 1st Edition, CRC Press.
- [13] Graves A., Schmidhuber J., (2005), “Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures”, Neural Networks, 18 (5-6), 602-610.
- [14] Hochreiter S., Schmidhuber J., (1997), “Long Short-Term Memory”, Neural Computation, 9 (8), 1735-1780.
- [15] Gehring J., Auli M., Grangier D., Yarats D., Dauphin Y. N., (2017), “Convolutional Sequence to Sequence Learning”, Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, August 6-11.
- [16] Feo T. A., Resende M. G., (1995), “Greedy Randomized Adaptive Search Procedures”, Journal of Global Optimization, 6 (2), 109-133.
- [17] Freitag M., Al-Onaizan Y., (2017), “Beam Search Strategies for Neural Machine Translation”, Proceedings of the First Workshop on Neural Machine Translation, Vancouver, July 30-August 4.
- [18] Web 1, (2019), <https://www.kaggle.com/suleymancan/turkishnews70000>, (Erişim Tarihi: 22/10/2019).
- [19] Web 2, (2019), <http://www.kemik.yildiz.edu.tr/>, (Erişim Tarihi: 22/10/2019).
- [20] Yıldırım O., Atık F., (2013), “Kişisel Gazete”.
- [21] Lin C.-Y., (2004), “ROUGE: A Package for Automatic Evaluation of Summaries”, Text Summarization Branches Out, Association for Computational Linguistics, 74-81.
- [22] Steinberger J., Ježek K., (2012), “Evaluation Measures for Text Summarization”, Computing and Informatics, 28 (2), 251-275.
- [23] Joulin A., Grave E., Bojanowski P., Mikolov T., (2016), “Bag of Tricks for Efficient Text Classification”, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, April 3-7.
- [24] Yogan J. K., Goh O. S., Halizah B. N., Choon N. H., Suppiah P. C., (2016), “A Review on Automatic Text Summarization Approaches”, Journal of Computer Science, 12 (4), 178-190.
- [25] Gambhir M., Gupta V., (2017), “Recent automatic text summarization techniques: a survey”, Artificial Intelligence Review, 47 (1), 1-66.

- [26] Verma R., Lee D., (2017), "Extractive summarization: Limits, compression, generalized model and heuristics", *Computacion y Sistemas*, 21 (4), 787-798.
- [27] Bhatia N., Jaiswal A., (2016), "Automatic text summarization and it's methods-a review", 6th International Conference-Cloud System and Big Data Engineering (Confluence), Uttar Pradesh, Noida, India, January 14-15.
- [28] Lloret E., Palomar M., (2012), "Text summarisation in progress: a literature review", *Artificial Intelligence Review*, 37 (1), 1-41.
- [29] Saggion H., Poibeau T., (2013), "Automatic text summarization: Past, present and future", *Multi-source, multilingual information extraction and summarization*, Springer, 3-21.
- [30] Das D., Martins A. F., (2007), "A survey on automatic text summarization", Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD), Qingdao, China., July 30-August 1.
- [31] Kiyomarsi F., Esfahani F. R., (2011), "Optimizing Persian Text Summarization Based on Fuzzy Logic Approach", 2011 International Conference on Intelligent Building and Management, Sydney, Australia, May 2-4.
- [32] Chen F., Kesong H., Guilin C., (2002), "An Approach to Sentence-Selection Based Text Summarization", IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering. TENCOP'02. Proceedings., Beijing, China, October 28-31.
- [33] Sankarasubramaniam Y., Ramanathan K., Ghosh S., (2014), "Text Summarization Using Wikipedia", *Information Processing & Management*, 50 (3), 443-461.
- [34] Web 3, (2019), <https://www.wikipedia.org/>, (Erişim Tarihi: 22/10/2019).
- [35] Kleinberg J. M., (1999), "Authoritative sources in a hyperlinked environment", *Journal of the ACM (JACM)*, 46 (5), 604-632.
- [36] Erkan G., Radev D. R., (2004), "Lexrank: Graph-Based Lexical Centrality as Salience in Text Summarization", *Journal of Artificial Intelligence Research*, 22, 457-479.
- [37] Page L., Brin S., Motwani R., Winograd T., (1999), "The PageRank citation ranking: Bringing order to the web", Stanford InfoLab.
- [38] Mashechkin I., Petrovskiy M., Popov D., Tsarev D. V., (2011), "Automatic Text Summarization using Latent Semantic Analysis", *Programming and Computer Software*, 37 (6). 299-305.

- [39] Ozsoy M. G., Alpaslan F. N., Cicekli I., (2011), "Text Summarization Using Latent Semantic Analysis", *Journal of Information Science*, 37 (4), 405-417.
- [40] Kyoomarsi F., Khosravi H., Eslami E., Dehkordy P. K., Tajoddin A., (2008), "Optimizing Text Summarization Based on Fuzzy Logic", *International Conference on Computer and Information Science*, Portland, OR, USA, May 14-16.
- [41] Suanmali L., Binwahlan M. S., Salim N., (2009), "Sentence Features Fusion for Text Summarization Using Fuzzy Logic", *International Conference on Hybrid Intelligent Systems*, Shenyang, China, August 12-14.
- [42] Suanmali L., Salim N., Binwahlan M. S., (2009), "Fuzzy logic based method for improving text summarization", *International Journal of Computer Science and Information Security*, 2 (1), 4-10.
- [43] Wang M., Wang X., Xu C., (2005), "An approach to concept-obtained text summarization", *15th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Abu Dhabi, United Arab Emirates, December 7-10.
- [44] Neto J. L., Freitas A. A., Kaestner C. A., (2002), "Automatic text summarization using a machine learning approach", *Brazilian Symposium on Artificial Intelligence*, Porto de Galinhas/Recife, Brazil, November 11-14.
- [45] Svore K., Vanderwende L., Burges C., (2007), "Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources", *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, June 28-30.
- [46] Kaikhah K., (2004), "Automatic text summarization with neural networks", *2nd International IEEE Conference on Intelligent Systems. Proceedings (IEEE Cat. No. 04EX791)*, Varna, Bulgaria, Bulgaria, June 22-24.
- [47] Hingu D., Shah D., Udmale S. S., (2015), "Automatic text summarization of Wikipedia articles", *International Conference on Communication, Information & Computing Technology (ICCICT)*, Mumbai, India, January 15-17.
- [48] Rush A. M., Chopra S., Weston J., (2015), "A Neural Attention Model for Abstractive Sentence Summarization", *Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, October 31–November 4.
- [49] Song S., Huang H., Ruan T., (2019), "Abstractive text summarization using LSTM-CNN based deep learning", *Multimedia Tools and Applications*, 78 (1), 857-875.

- [50] Chopra S., Auli M., Rush A. M., (2016), “Abstractive Sentence Summarization with Attentive Recurrent Neural Networks”, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, June 12-June 17.
- [51] See A., Liu P. J., Manning C. D., (2017), “Get To The Point: Summarization with Pointer-Generator Networks”, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, July 30-August 4.
- [52] Napoles C., Gormley M., Durme B. V., (2012), “Annotated Gigaword”, AKBC-WEKEX '12 Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, Montreal, Canada, June 7-8.
- [53] Over P., Dang H., Harman D., (2007), “DUC in context”, Information Processing & Management, 43 (6), 1506-1520.
- [54] Pasunuru R., Bansal M., (2018), “Multi-reward Reinforced Summarization with Saliency and Entailment”, The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, June 1-6.
- [55] Paulus R., Xiong C., Socher R., (2018), “A Deep Reinforced Model for Abstractive Summarization”, International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, April 30-May 3.
- [56] Takase S., Suzuki J., Okazaki N., Hirao T., Nagata M., (2016), “Neural Headline Generation on Abstract Meaning Representation”, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, USA, November 1-5.
- [57] Tan J., Wan X., Xiao J., (2017), “From Neural Sentence Summarization to Headline Generation: A Coarse-to-Fine Approach”, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), Melbourne, Australia, August 19-25.
- [58] Liu L., Lu Y., Yang M., Qu Q., Zhu J., Li H., (2018), “Generative Adversarial Network for Abstractive Text Summarization”, The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, Louisiana, USA, February 2-7.
- [59] Li P., Lam W., Bing L., Wang Z., (2017), “Deep Recurrent Generative Decoder for Abstractive Text Summarization”, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, September 7-11.

- [60] Hu B., Chen Q., Zhu F., (2015), “LCSTS: A Large Scale Chinese Short Text Summarization Dataset”, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, September 17–21.
- [61] Hsu W.-T., Lin C.-K., Lee M.-Y., Min K., Tang J., Sun M., (2018), “A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss”, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, July 15-20.
- [62] Web 4, (2019), <https://www.mturk.com/>, (Erişim Tarihi: 22/10/2019).
- [63] Chen Y.-C., Bansal M., (2018), “Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting”, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, July 15-20.
- [64] Litvak M., Last M., (2008), “Graph-based keyword extraction for single-document summarization”, Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, Manchester, United Kingdom, August 23-23.
- [65] Jagadeesh J., Pingali P., Varma V., (2005), “Sentence Extraction Based Single Document Summarization”, International Institute of Information Technology, 5.
- [66] Goldstein J., Mittal V., Carbonell J., Kantrowitz M., (2000), “Multi-document summarization by sentence extraction”, Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization, Seattle, Washington, USA, April 30-30.
- [67] Kutlu M., Cigir C., Cicekli I., (2010), “Generic Text Summarization for Turkish”, The Computer Journal, 53 (8), 1315-1323.
- [68] Nenkova A., McKeown K., (2011), “Automatic Summarization”, Foundations and Trends in Information Retrieval, 5, 103-233.
- [69] Loper E., Bird S., (2004), “NLTK: The Natural Language Toolkit”, Proceedings of the ACL Interactive Poster and Demonstration Sessions, Barcelona, Spain, July 21-26.
- [70] Costello A., (2003), “Punycode: A Bootstring Encoding of Unicode for Internationalized Domain Names in Applications (IDNA)”, RFC Editor.
- [71] Web 5, (2019), <https://www.crummy.com/software/BeautifulSoup/>, (Erişim Tarihi: 22/10/2019).

- [72] Bojanowski P., Grave E., Joulin A., Mikolov T., (2017), “Enriching Word Vectors with Subword Information”, Transactions of the Association for Computational Linguistics, 5, 135-146.
- [73] Web 6, (2019), <https://fasttext.cc/>, (Erişim Tarihi: 22/10/2019).
- [74] Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., (2013), “Distributed Representations of Words and Phrases and Their Compositionality”, Advances in Neural Information Processing Systems, 3111-3119.
- [75] Wang Y., Liu S., Afzal N., Rastegar-Mojarad M., Wang L., Shen F., Kingsbury P., Liu H., (2018), “A Comparison of Word Embeddings for the Biomedical Natural Language Processing”, Journal of Biomedical Informatics, 87, 12-20.
- [76] Goldberg Y., Levy O., (2014), “Word2Vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method”, arXiv preprint arXiv:1402.3722.
- [77] Pennington J., Socher R., Manning C., (2014), “Glove: Global Vectors for Word Representation”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, October 25–29.
- [78] Joulin A., Grave E., Bojanowski P., Douze M., Jégou H., Mikolov T., (2017), “Fasttext. zip: Compressing text classification models”, International Conference on Learning Representations, Toulon, France, April 24-26.
- [79] Cho K., Merrienboer B. v., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y., (2014), “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, October 25–29.
- [80] Abadi M., Agarwal A., Barham P., Brevdo E., Chen Z., Citro C., Corrado G. S., Davis A., Dean J., Devin M., Ghemawat S., Goodfellow I., Harp A., Irving G., Isard M., Jia Y., Jozefowicz R., Kaiser L., Kudlur M., Levenberg J., Mane D., Monga R., Moore S., Murray D., Olah C., Schuster M., Shlens J., Steiner B., Sutskever I., Talwar K., Tucker P., Vanhoucke V., Vasudevan V., Viegas F., Vinyals O., Warden P., Wattenberg M., Wicke M., Yu Y., Zheng X., (2015), “TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems”, Software available from tensorflow.org.
- [81] Bahdanau D., Cho K., Bengio Y., (2015), “Neural Machine Translation by Jointly Learning to Align and Translate”, 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, May 7-9.
- [82] Gomaa W. H., Fahmy A. A., (2013), “A survey of text similarity approaches”, International Journal of Computer Applications, 68 (13), 13-18.

- [83] Russell S., Norvig P., (2009), “Artificial Intelligence: A Modern Approach”, 3rd Edition, Pearson.
- [84] James G., Witten D., (2013), “An Introduction to Statistical Learning: with Applications in R”, 1st Edition, Springer.
- [85] Kuhn M., Johnson K., (2013), “Applied Predictive Modeling”, 1st Edition, Springer.
- [86] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É., (2011), “Scikit-learn: Machine Learning in {P}ython”, Journal of Machine Learning Research, 12, 2825-2830.

ÖZGEÇMİŞ

Enise Karakoç 1989 yılında, İzmit'te doğdu. 2010 yılında başladığı Gebze Teknik Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü'nü 2015 yılında başarıyla tamamlayarak, 2016 yılında Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda Yüksek Lisans eğitimine başladı. 2018 yılı Şubat ayından bu yana Kuveyt Türk Katılım Bankası'nda Veri Bilimcisi olarak çalışmaktadır.

EKLER

Ek A: Tez Çalışması Kapsamında Yapılan Yayınlar

Karakoç E., Yılmaz B., (2019), “Deep Learning Based Abstractive Turkish News Summarization”, 27th Signal Processing and Communications Applications Conference (SIU), Sivas, Turkey, April 24-26.