İNTERNET TABANLI TÜRKÇE METİNLER İÇİN OTOMATİK ÖZETLEME TEKNİĞİ

Cem Özkan
161402107

YÜKSEK LİSANS TEZİ
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Yüksek Lisans Programı
Danışman: Dr. Öğr. Üyesi Volkan Tunalı

İstanbul
T.C. Maltepe Üniversitesi
Fen Bilimleri Enstitüsü
Eylül, 2019

İNTERNET TABANLI TÜRKÇE METİNLER İÇİN OTOMATİK ÖZETLEME TEKNİĞİ

Cem Özkan
161402107
Orcid: 0000-0003-1894-1098

YÜKSEK LİSANS TEZİ
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Yüksek Lisans Programı
Danışman: Dr. Öğr. Üyesi Volkan Tunalı

İstanbul
T.C. Maltepe Üniversitesi
Fen Bilimleri Enstitüsü
Eylül, 2019



JÜRİ VE ENSTİTÜ ONAYI

CEM ÖZKAN'ın "İNTERNET TABANLI TÜRKÇE METİNLER İÇİN OTOMATİK ÖZETLEME TEKNİĞİ" başlıklı tezi 27.09.2019 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Maltepe Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliği" nin ilgili maddeleri uyarınca Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans/Doktora tezi oy birliğiyle/~~oy çokluğuyla~~, başarılı/~~başarısız~~ olarak kabul edilmiştir.

	Unvanı. Adı ve Soyadı	İmza
Üye (Tez Danışmanı)	Dr. Öğr. Üyesi Volkan TUNALI	.....
Üye	Dr. Öğr. Üyesi Mehmet Ali Aksoy TÜYSÜZ	.....
Üye	Dr. Öğr. Üyesi Selim BAYRAKLI	.....

Prof. Dr. İlter BÜYÜKDIĞAN
Enstitü Müdürü

Dr. Öğr. Üyesi Erdal Güvenoğlu

	ŞEKİL ONAY SAYFASI	Doküman No	FR-105
		İlk Yayın Tarihi	20.12.2017
		Revizyon Tarihi	10.12.2018
		Revizyon No	01
		Sayfa	1/2

ŞEKİL ONAY SAYFASI

31/10/2019

FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE,

Aşağıda bilgileri bulunan lisansüstü öğrencinin tezi şekil yönünden tarafımca incelenmiş ve Enstitüye teslim edilmesi uygun bulunmuştur.

Anabilim Dalı Başkanı
Prof. Dr. Ahmet Mesut RAZBONYALI
İmza

ÖĞRENCİ BİLGİLERİ	
ADI SOYADI	CEM ÖZKAN
ÖĞRENCİ NUMARASI	161402107
ANABİLİM DALI	BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
PROGRAMI	(X) YÜKSEK LİSANS ( ) DOKTORA ( ) SANATTA YETERLİK
DANIŞMANI	DR. ÖĞR. ÜYESİ VOLKAN TUNALI
TEZ BAŞLIĞI	İNTERNET TABANLI TÜRKÇE METİNLER İÇİN OTOMATİK ÖZETLEME TEKNİĞİ
SAVUNMA TARİHİ	27/09/2019
e-posta	scemozkan@gmail.com

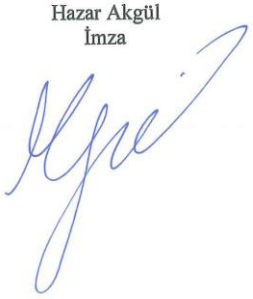
İç Kapak	☑ Var ☐ Yok
Jüri Onay Sayfası	☑ Var ☐ Yok
Etik İlke ve Kurallara Uyum Beyanı	☑ Var ☐ Yok
İntihal Raporu	☑ Var ☐ Yok
Teşekkür Sayfası	☑ Var ☐ Yok
Öz (Başlık-Öz-Anahtar Sözcükler)	☑ Var ☐ Yok
Abstract (Title-Abstract-Key Words)	☑ Var ☐ Yok
İçindekiler	☑ Var ☐ Yok
Çizelgeler Listesi	☐ Var ☑ Yok
Şekiller Listesi (varsa)	☐ Şekil yok ☑ Uygundur ☐ Uygun Değildir
Kısaltmalar Listesi	☑ Var ☐ Yok
Tablolar Listesi (varsa)	☐ Tablo yok ☑ Uygundur ☐ Uygun Değildir
Ekler Listesi (varsa)	☑ Ek yok ☐ Uygundur ☐ Uygun Değildir
Özgeçmiş	☑ Var ☐ Yok
Sayfa Genişliği	☑ Uygundur ☐ Uygun Değildir
Yazı Tipi	☑ Uygundur ☐ Uygun Değildir

Hazırlayan: İlgili Birim	**Onaylayan:** Kalite Yönetim Koordinatörlüğü

	ŞEKİL ONAY SAYFASI	Doküman No	FR-105
		İlk Yayın Tarihi	20.12.2017
		Revizyon Tarihi	10.12.2018
		Revizyon No	01
		Sayfa	2/2

Referans Kullanımı	☑ Uygundur ☐ Uygun Değildir
Kaynakça Yazımı	☑ Uygundur ☐ Uygun Değildir
Ekler (varsa)	☑ Ek yok ☐ Uygundur ☐ Uygun Değildir

Hazar Akgül
İmza



Hazırlayan: İlgili Birim	**Onaylayan:** Kalite Yönetim Koordinatörlüğü

Revizyon Takip Tablosu

REVİZYON NO	TARİH	AÇIKLAMA
00	01.03.2018	İlk yayın.

ETİK İLKE VE KURALLARA UYUM BEYANI

27/09/2019

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarından bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilmeyen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; çalışmanın Maltepe Üniversitesinde kullanılan "bilimsel intihal tespit programı" ile tarandığını ve öngörülen standartları karşıladığını beyan ederim.

Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçlara razı olduğumu bildiririm.

Cem Özkan

İnternet Tabanlı Türkçe Metinler İçin Otomatik Özetleme Tekniği

ORIJINALLIK RAPORU

%7	%6	%1	%7
BENZERLIK ENDEKSI	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCI ÖDEVLERI

BIRINCIL KAYNAKLAR

1	Submitted to The Scientific & Technological Research Council of Turkey (TUBITAK) Öğrenci Ödevi	%5
2	www.icens.eu İnternet Kaynağı	%1
3	Submitted to Üsküdar Üniversitesi Öğrenci Ödevi	<%1
4	Submitted to Harran Üniversitesi Öğrenci Ödevi	<%1
5	Submitted to Batman University Öğrenci Ödevi	<%1
6	sbe.maltepe.edu.tr İnternet Kaynağı	<%1
7	Submitted to Anadolu University Öğrenci Ödevi	<%1
8	Submitted to Tobb University of Economics & Technology Öğrenci Ödevi	<%1

Dr. Öğr. Üyesi Volkan TUNALI

TEŞEKKÜR

Uzun soluklu bu tez çalışmam boyunca anlayışı ve desteğiyle yardımcı olan ve bu sayede bu tezin tamamlanmasını sağlayan Sayın Dr. Öğr. Üyesi Volkan Tunalı'ya teşekkürlerimi sunuyorum.

Ayrıca bu tez çalışmasında kullanılan haber veritabanına erişimimi sağlayan Sayın M. Sait Bilgin'e ve özellikle haber özetleme aşamasında en büyük yardımcım olan sevgili eşim Hülya Özkan'a teşekkür ediyorum.

Cem Özkan
Eylül, 2019

ÖZ

İNTERNET TABANLI TÜRKÇE METİNLER İÇİN OTOMATİK ÖZETLEME TEKNİĞİ

Cem Özkan
Yüksek Lisans Tezi
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Yüksek Lisans Programı
Danışman: Dr. Öğr. Üyesi Volkan Tunalı
Maltepe Üniversitesi Fen Bilimleri Enstitüsü, 2019

Günümüzde internet ve sosyal medya kullanımı giderek daha da yaygınlaşmaktadır. Dolayısıyla üretilen Türkçe içerik de doğru orantılı olarak artmakta ve Türkçe veri madenciliği alanında çalışmalar yapılmasını bir gereklilik haline getirmektedir.

Türkçe metinleri işlerken sağlıklı bir sonuç elde edebilmek için kelime kelime çözümleme yapılması gerekmektedir. Bu oldukça zahmetli bir süreçtir, çünkü sondan eklemeli bir dil olan Türkçedeki kelimeleri işleyebilmek için önce etimolojik ve/veya morfolojik bir ayıklama yapmak, kelimeleri köklerine, eklerine ayrımak ve ardından işleme tabii tutmak gerekmektedir.

Bu tez çalışmasında, Türkçe metinleri ayıklama (extraction) yöntemi ile özetlemek için farklı bir yaklaşım önerilmiş, Türkçe kelimeleri işlerken kök ve ek ayıklaması yapmak yerine, kelimeleri harf harf karşılaştırarak, uyuşan harfler ve bu harflerin kelime içindeki konumlarına göre sağlıklı bir sonuca ulaşılmaya çalışılmıştır.

Çalışma sonucunda elde edilen çıktılar, sağlama yapabilmek amacıyla gerçek bir insan tarafından çıkarılan özetlerle karşılaştırılmış ve buna göre bir değerlendirmede bulunulmuştur.

Anahtar Sözcükler: Türkçe Haber Özetleme, Metin Özetleme, Rouge-N, NLP.

ABSTRACT

AUTOMATIC SUMMARY TECHNIQUE FOR INTERNET BASED TURKISH TEXTS

Cem Özkan
Master Thesis
Department of Computer Engineering
Computer Engineering Programme
Advisor: Asst. Prof. Volkan Tunalı
Maltepe University Graduate School of Science and Engineering, 2019

Today, the use of internet and social media is becoming more and more widespread. Therefore, the Turkish content produced increases in direct proportion and makes it necessary to carry out studies in the field of Turkish data mining.

In order to obtain a healthy result when processing Turkish texts, word to word analysis is required. This is a very troublesome process because in order to be able to process the words in Turkish, which is a suffix language, it is necessary to make an etymological and/or morphological sorting, to separate the words to their roots and suffixes and then to process them.

In this thesis, a different approach is tried to summarize the Turkish texts by extraction method. Instead of making root and suffix extraction while processing Turkish words, it is tried to reach a healthy result according to the matching letters and their positions in the word by comparing them letter by letter.

The results obtained from the study are compared with the abstracts prepared by a real person in order to provide checks and an evaluation is made accordingly.

Keywords: Turkish Text Summarizing, Text Summarizing, Rouge-N.

İÇİNDEKİLER

JÜRİ VE ENSTİTÜ ONAYI ii
ETİK İLKE VE KURALLARA UYUM BEYANI ..... HATA! YER İŞARETİ TANIMLANMAMIŞ.
İNTİHAL RAPORU Vi
TEŞEKKÜR ..... Vİİi
ÖZ ..... Vİİİi
ABSTRACT ix
İÇİNDEKİLER X
TABLOLAR LİSTESİ ..... XİİI
ŞEKİLLER LİSTESİ ..... Xİİİi
KISALTMALAR ..... XİV
ÖZGEÇMİŞ XV
BÖLÜM 1. GİRİŞ 1
BÖLÜM 2. İLGİLİ ÇALIŞMALAR 4
2.1. Yabancı Diller İçin Yapılan Çalışmalar 4
2.2. Türkçe İçin Yapılan Çalışmalar 5
BÖLÜM 3. YÖNTEM VE UYGULAMA 7
3.1. Metin Özetleme 7
3.1.1. Özet Nedir? 7
3.1.2. Metin Özetleme Mimarisi 7
3.1.3. Metin Özetleme Türleri (Kategorileri) 9
3.2. Doğal Dil İşleme (NLP) 11
3.2.1. Doğal Dil İşleme (NLP) Nedir? 11
3.2.2. Doğal Dil İşleme Yapısı ve Çalışması 11
3.2.3. Doğal Dil İşleme Kütüphanleri 13
3.2.4. Zemberek NLP 16
3.3. Ölçüm Teknikleri 18
3.3.1. Metin Özetleme Başarı Ölçümü 18
3.3.2. Başarı Ölçümü Değerlendirme Araçları 20
3.3.3. ROUGE Ölçüm Metriği 22

3.4. Veriler ve Uygulama 23
3.4.1 Verilerin Toplanması 23
3.4.2. Frekans Listesi Oluşturma Algoritması 27
3.4.3. Zemberek NLP Uygulaması 30
3.4.4. Özetlerin Ölçümlenmesi 30
BÖLÜM 4. BULGULAR VE YORUMLAR 37
4.1. Bulgular 37
4.2. Yorumlar 39
BÖLÜM 5. SONUÇ 41
KAYNAKÇA 43

TABLOLAR LİSTESİ

Tablo 1 - Tez Çalışmasına Konu Haberlerin Veritabanına Kayıt Detayları 23

Tablo 2 - Zemberek Nlp Rouge-N Ölçüm Sonuçları (Ana Veri Seti) 31

Tablo 3 - Matematiksel Yöntem Rouge-N Ölçüm Sonuçları (Ana Veri Seti)..... 31

Tablo 4 - Zemberek Nlp Rouge-N Tam Puan (Ana Veri Seti) 32

Tablo 5 - Matematiksel Yöntem Rouge-N Tam Puan (Ana Veri Seti) 32

Tablo 6 - Zemberek Nlp Rouge-N Ölçüm Sonuçları (Kontrol Veri Seti) 33

Tablo 7 - Matematiksel Yöntem Rouge-N Ölçüm Sonuçları (Kontrol Veri Seti)..... 33

Tablo 8 - Diğer Yöntem Rouge-N Ölçüm Sonuçları (Kontrol Veri Seti) 34

Tablo 9 - Zemberek Nlp Rouge-N Tam Puan Sonuçları (Kontrol Veri Seti)..... 35

Tablo 10 - Geliştirilen Yöntem Rouge-N Tam Puan Sonuçları (Kontrol Seti) 35

Tablo 11 - Diğer Yöntem Rouge-N Tam Puan Sonuçları (Kontrol Veri Seti) 36

ŞEKİLLER LİSTESİ

Şekil 1 - Metin Özetleme Mimari Yapısı 8

Şekil 2 – Doğal Dil İşleme Çalışma Akışı 12

Şekil 3 – Zemberek Nlp, Akış Diyagramı 16

Şekil 4 – Özet Değerlendirme Ölçüm Sistematiği 19

Şekil 5 – SEE Ölçümleme Oturumundan Bir Ekran Görüntüsü 21

Şekil 6 - Referans Özeti Oluşturma Arayüzüne Ait Ekran Görüntüsü 25

Şekil 7 - Ana Veri Seti Cümle-Kelime Dağılımı 26

Şekil 8 - Kontrol Veri Seti Cümle-Kelime Dağılımı 27

Şekil 9 - Ana Veri Seti Rouge-N Recall Karşılaştırma Grafiği 37

Şekil 10 - Kontrol Veri Seti Rouge-N Recall Karşılaştırma Grafiği 38

KISALTMALAR

BLEU : Bilingual Evaluation Understudy (İkili Dilli Ölçümleme Metodu)

DUC : Document Understanding Conferences (Dokuman Anlama Konferansları)

IDF : Inverse Term Frequency (Ters Doküman Sıklığı)

ISI : Information Science Institute (Güney Kaliforniya Üniversitesine bağlı "Bilgi Bilimleri Enstitüsü")

NLP : Doğal Dil İşleme (Natural Language Processing)

ROUGE : Recall-Oriented Understudy for Gisting Evaluation (Geri Çağırma Odaklı, Özet Ölçümleme Metodu)

TF : Term Frequency (Terim Sıklığı)

ÖZGEÇMİŞ

Cem Özkan

Bilgisayar Mühendisliği Anabilim Dalı

Eğitim

Y.Ls.	2019	Maltepe Üniversitesi, Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı
Ls.	2013	Anadolu Üniversitesi İşletme Fakültesi, İşletme Bölümü
Lise	1996	Ankara Keçiören Kanuni Lisesi

İş/İstihdam

Yıl	*Görev*
2011 -	Kurucu Yönetici, Eksibir Interactive
2009 - 16	Program Yapımcısı, Web Proje Direktörü, Doğuş Medya Grubu
2009 - 11	Web Görsel Yönetmen, İpek Medya Grubu
2006 - 08	Program Yapımcısı, Web/IT Uzmanı. N.A.R. Group
2004 - 06	Program Yapımcısı, Power Group
2001- 03	Multimedya Tasarımcısı, Noktalar A.Ş.

Kişisel Bilgiler

Doğum yeri ve yılı : Kırıkkale, 1977 Cinsiyet: Erkek
Yabancı diller : İngilizce
GSM / e-posta : 0 532 485 95 00 / scemozkan@gmail.com

BÖLÜM 1. GİRİŞ

Günümüzde bilgi, çok hızlı üretilmekte ve tüketilmektedir. Özellikle internetin gündelik hayata iyiden iyiye girdiği son 15 yıl içerisinde, internet tabanlı bilgi ve kaynaklar sayılamayacak kadar çoğalmıştır [1].

İnternet tabanlı bu büyük veri havuzu, doğaldır ki veri madenciliği çalışmalarını kaçınılmaz kılmış ve bu çalışmalar için çeşitli teknikler geliştirilmiştir. Bu tekniklerin temeline indiğimizde, büyük metni işleyebilmek için -doğal olarak- önce en küçük parçaları yani kelimeleri işleme gerekliliği bulunmaktadır.

Kelime işleme amacıyla geliştirilen birçok farklı yöntem bulunsa da bunlardan çok azı özellikle Türkçe dilini hedeflemektedir.

Türkçe için geliştirilen bu az sayıda yöntem, daha çok NLP (Natural Language Processing – Doğal Dil İşleme) kütüphanesi olarak tasarlanmakta ve Türk dilinin yapısı gereği kelimeleri, türleri, cümle içindeki yerleri ve aldıkları çeşitli ekler ile değerlendirmeye çalışmaktadır. Bunun için de önce bir kök, ek, kelime sözlüğü yaratılması ve bu sözlüğün güncelliğinin de mutlaka zaman içinde kontrol edilmesi gerekmektedir. Bu tür bir kütüphane ile çalışmak yazılımsal olarak bazı zorluklar getirmektedir, çünkü üzerinde çalışılan metin sadece 50 kelime dahi olsa, bu 50 kelimenin teker teker binlerce sözlük kaydı ile karşılaştırılması gerekmektedir.

Çok daha hızlı olunması gereken ve görece daha kolay bir iş olan metin içerisinden özet ayıklama işlemleri için her ne kadar bahsedilen bu NLP kütüphanelerini kullanmak mümkün olsa da daha basit, aynı oranda başarılı ve sadece özet ayıklama amacına hizmet eden daha hızlı bir yöntemin eksikliği görülmektedir.

Bu tez çalışması ile direkt olarak Türk dilini ve internet tabanlı haber metinlerini hedef alan, daha hızlı ve kolay bir kelime işleme ve sonrasında da özet ayıklama yöntemi geliştirilmiştir.

Herhangi bir dildeki bir metnin özetini çıkarmak için günümüzde genelde Exctraction (Ayıklama) ve Abstraction (Özet Yazma) olmak üzere iki farklı yöntem benimsenmektedir [2].

NLP kütüphaneleri, "Abstraction" yöntemi için bir gerekliliktir, çünkü metinden bağımsız, anlamlı ve dilin kurallarına uygun yeni cümlelerin yaratılması ve bu cümleler ile metnin özetlenmesi gerekmektedir [3]. Bir bakıma bilgisayarın/yazılımın insanlarla anlamlı cümleler kurarak konuşması sağlanmaya çalışılmaktadır. Bu sebeple de üzerinde çalışılan dilin cümle yapısı ile ilgili tüm kurallar bilinmeli ve uygulanmalıdır. Hatta çoğu zaman "Yapay Zeka" ve "Makine Öğrenimi" ile de yöntemin geliştirilmesi ve desteklenmesi gerekmektedir.

Öte yandan, bu tez çalışmasına konu olan "Extraction" yöntemi ise, değil "Yapay Zeka", NLP kütüphanelerinin bile kullanılmasına gerek duymayacak kadar yalın ve sadedir. Daha doğrusu, öyle olmalıdır. Çünkü amacı, üzerinde çalışılan metnin içerisinden, o metnin tümünü en iyi özetleyen cümleleri bulmak ve onları ayıklamaktır. Böylelikle büyük metnin, küçük bir özeti elde edilmiş olur. Bu yöntem tabii ki romanlar, hikayeler, ve benzeri çok sayfalı yazınlar için uygun değildir ama internet tabanlı haber metinleri ve benzeri anlam bütünlüğü içeren kısa metinler için biçilmiş kaftandır [4].

Bu tez çalışmasıyla, internet tabanlı haber metinleri üzerinde bir çalışma yapılarak, Türkçe dilinin yapısını hedef alan, kısa ve anlam bütünlüğü olan metinlerin, kök, ek, kelime karşılaştırmaları yapmaya gerek duymadan ve bir sözlük ya da kütüphane kullanmadan hızlı ve etkin bir biçimde, özetlerinin çıkarılması amaçlanmaktadır.

Bu amaç için yine de mutlaka kelime işleme çalışmaları yapmak kaçınılmazdır. Bu sebeple, Türk dilinin özelliklerine göre kelimeleri harf harf işleyen, kısa ve hızlı bir kelime işleme algoritmasının geliştirilmesi de bu tez çalışmasının amaçları içerisinde yer almaktadır ki böylelikle özet olabilecek cümleler metinlerin içerisinden başarıyla ayıklanabilsin.

Bu tez çalışmasında geliştirilen, hızlı ve verimli çalışan, güncelleme gerektirmeyen ve beklentiler doğrultusunda işe yarar, kabul edilebilir sonuçlar üreten bir özet ayıklama yöntemi, Türkçe veri madenciliği çalışmalarında birçok alanda kullanılabilir. Burada anahtar öneme sahip olan özellikler, basitlik, verimlilik ve hızlılıktır.

Bu sayede çok daha kapsamlı bir çalışma yapmadan önce, örneğin "Abstraction" yöntemini kullanmadan önce, bu tez çalışmasında geliştirilen yöntem ile bir ön sonuç alınabilir ve buna göre sonraki aşamaya geçilebilir ya da sonuçlar karşılaştırılabilir.

Metin özetleme çalışmaları yerine metin kategorize etme çalışmaları, metin tarama çalışmaları için de kullanılabilecek olan bu yöntem, özellikle sosyal medya verilerine de uyarlanabilir.

Ayrıca bu tez çalışmasında geliştirilen yönteme, sektörel eklentiler yapılarak ya da sektörel özel kurallar yazılarak, farklı sektörlere de uyarlanabilir. Örneğin e-ticaret ya da turizm sitelerinde ürün açıklamaları, yorum karşılaştırmaları ve/veya özetlemeleri için kullanılabilir.

Kısaca, bu tez çalışmasında ortaya koyulan yöntem her ne kadar Türkçe haber metinlerinin özetlenmesi için kullanılmış olsa da, daha spesifik çalışmalar için geliştirmeler yapılarak bir ön süzgeç görevi görebilir ve metin özetleme özelliği yerine sadece kelime işleme özelliği kullanılabilir. Böylelikle farklı sektörlerde, farklı amaçlar için yapılan veri madenciliği çalışmalarına katkı sağlayabilir.

BÖLÜM 2. İLGİLİ ÇALIŞMALAR

Bu bölümde Türkçe ve diğer diller için daha önce yapılan önemli metin özetleme çalışmalarına ve detaylarına yer verilmiştir.

2.1. Yabancı Diller İçin Yapılan Çalışmalar

İlk metin özetleme çalışmaları 1950'lerde yapıldı ve bugün kullanılan birçok terim bu ilk çalışmalarda ilk kez yayınlandı. Hans Peter Luhn'un 1957 yılına ait çalışmasında ortaya attığı "terim sıklığı (Term Frequency - TF)" kavramı günümüzdeki metin özetleme çalışmalarının hâlâ en önemli parçası ve en çok kullanılan ağırlık ölçümleme şemasıdır [5].

Luhn çalışmasında, "Bir terim ya da terim öbeği, metin içerisinde ne kadar çok tekrarlanıyorsa, aynı oranda yazar o terime önem veriyordur [6]" fikrini ileri sürmüş ve "terim sıklığı" kavramını getirmiştir. Öte yandan kısa zaman içerisinde bu yaklaşımın yanlış/yetersiz sonuçlar verebileceği görülmüştür. Çünkü bu yöntemle, dilin yapısı gereği sık tekrarlanan bağlaçlar, edatlar vb. cümle ögeleri metin içerisinde anlamsal bir öneme sahip olmasalar da terim sıklığı açısından en üste çıkabilmektedirler.

Bu noktada metin özetleme çalışmaları adına başka öncü bir çalışma K. S. Jones tarafından 1972 yılında yayınlanmıştır. Jones, IDF (Inverse Document Frequency – Ters Doküman Sıklığı) kavramını bu çalışmasında detaylıca anlatmış [7] ve Luhn'un metodunda ortaya çıkan sıklık listesinin aksine, normalize edilmiş daha sağlıklı bir sıklık listesi oluşturulmasını sağlamıştır. Jones çalışmasında, en sık tekrarlanan terimlerin ağırlık değerlerini düşürürken, en az tekrarlanan terimlerin ağırlık değerlerini yükseltmiş ve daha homojen bir sıklık/ağırlık şeması oluşturmuştur [8].

Luhn ve Jones'un yöntemleri daha sonra sıkça birlikte kullanılmış ve metin özetleme çalışmaları içerisinde TFxIDF olarak birlikte yer edinmişlerdir.

2000'li yıllara kadar daha birçok önemli çalışma yapılmış ve her birinde farklı yöntemler denenmiştir, öne çıkanlar:

P. B. Baxendale (1958): Cümlenin metin içerisindeki pozisyonunun önemi üzerine bir çalışma yayınlamıştır. Metnin ilk ve son cümlelerinin daha kıymetli olduğunu vurgulamıştır, tıpkı bir kitabın ilk 80 ve son 40 sayfasının okunması gibi.

H. P. Edmundson (1969): Pozisyon ve frekansın yanı sıra ipucu kelimeleri işlemeye çalışmıştır (özellikle, imkansız, mutlaka vb. ipucu kelimeler). Ayrıca metin yapısını dikkate almış ve başlık ya da ilk cümlelerin kıymetli olduğunu vutgulamıştır.

G. DeJong (1979 - 1982): FRUMP adı verilen çalışma, yapay zeka çalışmaları kapsamında geliştirilmiştir ve bilgiye dayanan ilk çalışmadır. "Sketchy Script" isimli önceden hazırlanmış yaklaşık 60 adet şablon ile, UPI (United Press International) üzerindeki haberler işlenmiş ve haberler içerisinden çıkarılan cümleler önceden hazırlanan bu şablonlara yerleştirilmiştir, oldukça başarılı sonuçlar elde etmiştir.

J. Paice (1990): Metin içerisinden çıkarılan özetlerin denge ve uyum eksikliklerini gidermek üzere bir çalışma yapmıştır. Bunun için artgönderimler (Fr: anaphore) ve retorik bağlantılar ile ilgilenmiştir.

Ayrıca, R. Brandow et al. (1995), P. Kupiec et al. (1995), ilk çoklu doküman çalışması SUMMONS adı ile K. R. McKeown ve D. R. Radev (1995) tarafından yapılan çalışmalar da başarılı sonuçlar elde etmiştir.

2.2. Türkçe İçin Yapılan Çalışmalar

Tüm dillerde olduğu gibi Türkçe için de metin özetleme çalışmaları yapılmaktadır. Özellikle 2000'li yıllar sonrasında yapılan bu çalışmalar, alanda çalışan birçok araştırmacı için hem bilgi hem de veri seti birikimini artırmaktadır [9].

E. Uzundere, E. Dedja, B. Diri, M.F. Amasyalı (2008): Türkçe haber metinleri üzerine yapılan bu çalışmada özetleme işlemi, cümlelerin puanlandırılmasıyla yapılmıştır. 10 farklı metin, aynı zamanda 15 kişi tarafından da özetlenmiş ve sonuç karşılaştırmasında, %55 başarı yakalandığı gözlenmiştir [10].

A. Güran, S. N. Arslan, E. Kılıç, B. Diri (2014): Bu çalışmada da yine 20 adet haber metni kullanılmıştır. Ayrıca ilk kez "Özel İsim Tanıma" metodu kullanılarak, sonuç cümlelerin en iyi olanları listelenmiştir. Bu çalışmanın ana amacı aday cümleler içerisinden en iyi olanlarının seçilebilmesidir. Bu çalışmada sonuçlar eşit cinsiyet dağılımlı 30 kişilik bir grup tarafından analiz edilmiştir [11].

M. Çakır, E. Çelebi (2011): Bu çalışmada dil bağımsız bir özetleme yöntemi geliştirilmiştir. C3M (Cover Coefficient-Based Clustering Methodology) metodolojisi ile birlikte TF ve özellik ayıklama yöntemleri kullanılmış ve sonuçlar, Re-Call, Precision, Rouge-N metrikleriyle değerlendirilmiştir. İnsan analizi için ise 10 kişilik bir grubun özetleri kullanılmıştır. Çıkan sonuçlarda, benzer birçok sistemden daha başarılı olduğu görülmüştür [12].

M. V. Sami, B. Diri (2010): Bu çalışma için web tabanlı metinler kullanılmış ve Türkçe siteler içerisinden belirlenen kriterlere göre özetleme yapılması sağlanmıştır. Yine cümle puanlama yönteminin kullanıldığı çalışma, %59 oranında başarılı bulunmuştur [13].

Alanda yapılan daha birçok önemli çalışma bulunmaktadır. Bu çalışmaların her biri hem yöntem eksikliklerini gidermeyi amaçlamakta, hem de Türkçe metinler için başarılı özetleme tekniklerinin oluşturulmasını sağlamaya çalışmaktadır: A. Güran "Otomatik Metin Özetleme Sistemi" (2013), F. C. Pembe "Automated Query-Biased And Structure-Preserving Document Summarization For Web Search Tasks" (2010), M. Y. Nuzumlalı "Analyzing Stemming And Sentence Simplification Methodologies For Turkish Multi-Document Text Summarization" (2010).

BÖLÜM 3. YÖNTEM VE UYGULAMA

Bu bölümde metin özetleme çalışmaları, türleri, tanımları ve süreçleri ile ilgili detaylı bilgilere yer verilmiştir. NLP kütüphanleri, kelime frekans listeleri, çeşitli başarı ölçüm metrikleri ve bu tez çalışmasında kullanılan ölçüm metriği "Rouge" ile ilgili detaylı bilgilere ve uygulama detaylarına yer verilmiştir.

Metin özetleme sürecinin başlangıcında bu kavramların incelenmesi ve anlaşılması önemli bir gerekliliktir.

3.1. Metin Özetleme

3.1.1. Özet Nedir?

Özetin sözlük karşılığı, "Bir şey hakkında temel gerçekleri veya fikirleri veren kısa ve net bir açıklama [14]" olarak geçmektedir. Bu çok genel bir tanımdır.

Günümüzde veri çok hızlı üretilmekte ve tüketilmektedir. Bu büyük verinin hem işlenmesi hem de saklanması için kolay yönetilir küçük parçalara bölünmesi veri yönetimi için bir gerekliliktir. Bu sebeple veri madenciliği açısından "Özet" tanımının çok daha geniş bir karşılığı vardır. Veri madenciliğinde, sıkıştırma, etiketleme, ilişkilendirme vb. yöntemlerin tümü özet olarak değerlendirilebilmektedir [15].

Metin özetleme çalışmaları için ise özet, daha net bir şekilde tanımlanır: "Özet, kaynak metni, bilgi içeriğini ve genel anlamını koruyarak daha kısa bir hale dönüştürmek [16]". Metin özetleme çalışmaları için "özet"in tanımında ayrıca bazı kıstaslar da bulunmaktadır, örneğin bir metnin özet olarak kabul edilebilmesi için, ait olduğu kaynak metnin uzunluğunun yarısını geçmemesi gerekmektedir [17].

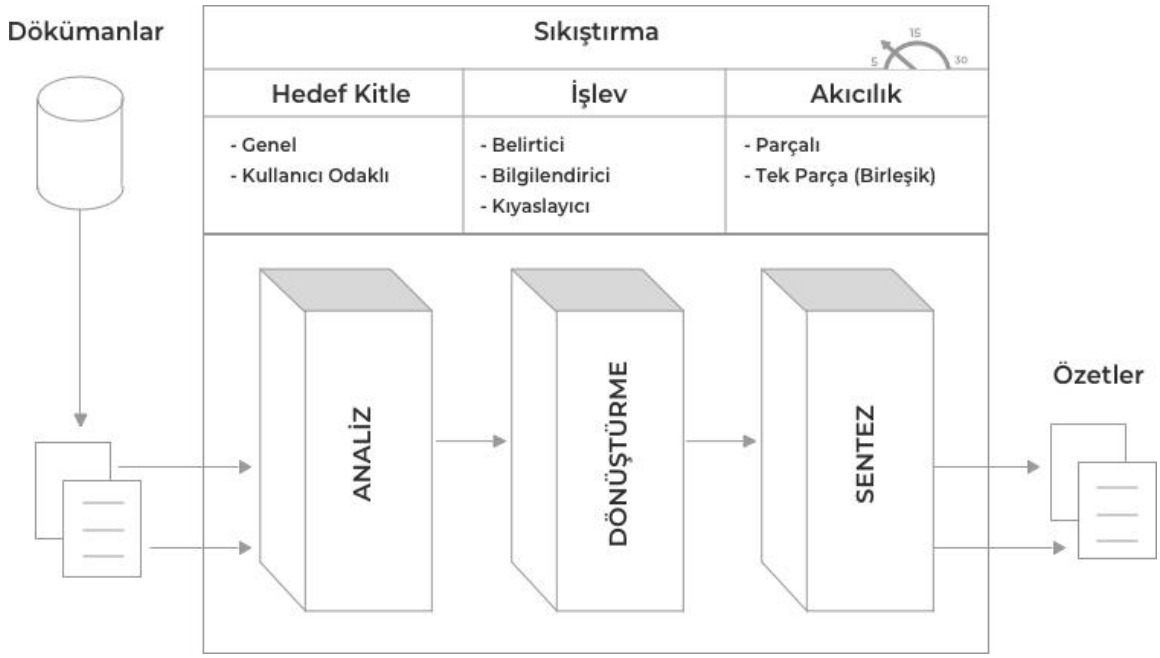
3.1.2. Metin Özetleme Mimarisi

Metin özetleme çalışmaları, hangi yöntem kullanılırsa kullanılsın mimari açıdan aynı yapıya sahiptir. Günlük hayatta birçok örneği bulunan metin özetleme işlemi, tekli

ya da çoklu dokümanın, bir grup kişi ya da bir görev için süzülmesi işlemidir ve kapsamı çok geniştir [18]:

- Film konusunu anlatan sinopsisler,
- Haber başlıkları ve özetleri
- Biyografiler,
- Kitap eleştirileri, önsözleri,
- Müzik albümü eleştirileri vb.

Genel olarak metin özetleme konusunda insanlar oldukça başarılıdır ve aynı başarıyı bilgisayarlı otomatik özetleme sistemleri için de yakalayabilmek, Şekil 1'de gösterilen farklı kriterleri ve süreçleri takip etmeyi gerektirir.



Şekil 1 - Metin Özetleme Mimari Yapısı [18]

Doküman(lar): Özetlenecek kaynak metin tek bir doküman ya da ilişkili ve/veya ilişkisiz birçok doküman olabilir.

Sıkıştırma: Çıkarılacak özetlerin, kaynak metne oranını belirler, tekli dökümanlar için bu oran daha düşük iken, çoklu dokümanlarda doküman başına çok daha fazla sıkıştırma gerekir, bu sebeple doküman başına daha yüksek oran belirlenir. (tekli doküman %30 oranında sıkıştırılırken, çoklu dökümanda her bir doküman %5 oranında sıkıştırılabilir)

Hedef Kitle: Çıkarılacak özet kim için kullanılacak? Özel amaç için çıkarılacak olan özetler ve genel amaçlı özetler farklılıklar gösterebilir.

İşlev: Özetin hangi amaç için çıkarıldığı oldukça önemlidir. Bilgilendirici bir özet, konunun tüm hatlarını içermeliyken, belirtici bir özet sadece ana hatlarını içerebilir. (bir haber özeti ya da bilimsel bir yayının özeti gibi)

Akıcılık: Çıkarılacak olan özetlerin yapısını belirler. Özet içerisinde yer alan cümlelerin ya da paragrafların bir birleriyle uyumlu olup olmamasını belirler.

3.1.3. Metin Özetleme Türleri (Kategorileri)

Metin özetleme çalışmaları, birçok farklı amaç için yapılabileceği gibi, farklı şekillerde de yapılabilir. Kullanılan yöntemler ve istenilen çıktılara göre kategorize edildiğinde genel hatlarıyla aşağıdaki listede belirtilen başlıklara göre ayrıştırılabilir [19]:

Kaynak Türüne Göre:

Tekli Doküman (Single Document): Bir tek doküman içerisinde yapılan özetleme çalışmalarıdır.

Çoklu Doküman (Multi Document): Bir birleriyle ilişkili olan ya da olmayan birden fazla doküman ile yapılan özetleme çalışmalarıdır.

Yönteme Göre: Çıkarılacak özetlerin ne şekilde oluşturulacağına bağlı olarak [16]:

Ayıklayıcı (Extractive): Kaynak metin içerisinden cümlelerin seçilmesi ile yapılan özetleme

Oluşturucu (Abstractive): Kaynak metnin konusunu inceledikten sonra, özetin sistem tarafından oluşturulan yeni cümleler ile yapılmasıdır.

Amaca Göre: Çıkarılacak özetlerin ne amaçla kullanılacağına bağlı olarak:

Belirtici (Indicative): Kaynak metinden konunun önemli sayılan kısımları bir araya getirilerek oluşturulan özetler. Kaynakta yer alan tüm bilgi, özette yer almayabilir.

Bilgilendirici (Informative): Kaynak metnin içerisindeki tüm bilgiyi aktarmaya yöneliktir. Özellikle bilimsel çalışmaların özetlemeleri için tercih edilir.

Sonuç İçeriğine Göre: Çıkarılacak özetlerin neler içericeğine bağlı olarak (daha çok, çoklu döküman özetlerinde kullanılır, ve istenilen bilginin dökümanlar içerisinden ayıklanması ve özete alınması sağlanır):

Alana Yönelik (Domain): Bir alana yönelik sonuç getiren özetler.

Konuya/Türe Yönelik (Genre Specific): Sadece belirli bir konuya yönelik olarak yapılan özetleme çalışmaları.

Bağımsız (Indipendent): Alan ya da konu bağımsız olarak metne yönelik özetleme çalışmaları.

Sorguya Göre: Çıkarılacak özetlerin nasıl sorgulanacağına, görüntüleneceğine bağlı olarak:

Sorgu Tabanlı (Query Based): Yapılan sorgu ile uyuşan özetlerin oluşturulması. Genellikle arama motorlarında kullanılan özetleme biçimidir.

Genelleştirilmiş (Generalized): Bir sorgu olmaksızın, genelleştirilmiş özetlerin görüntülenmesidir.

3.2. Doğal Dil İşleme (NLP)

3.2.1. Doğal Dil İşleme (NLP) Nedir?

Üzerinde fikir birliğine varılmış net ve tek bir tanımı olmayan "Doğal Dil İşleme (NLP)" kısaca şu şekilde tarif edilebilir: "Doğal Dil İşleme, bir dizi görev ya da uygulama için, doğal şekilde oluşan metinleri, insana en yakın biçimde dilbilimsel olarak analiz ve temsil etmek için teorik mativasyonlu hesaplama tekniklerinin tümüdür [20]".

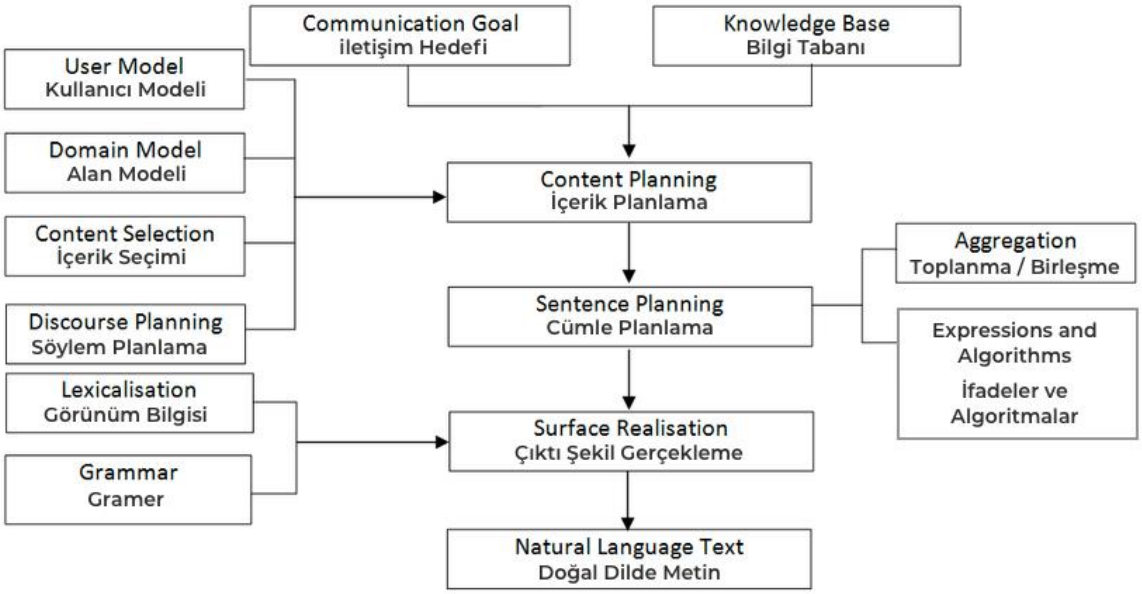
Bu tanımdan yola çıkarak, günlük hayatta, hayatın olağan akışı ile oluşan tüm metinlerin, sistemler tarafından, bir insan gibi anlaşılması ve/veya oluşturulması için kullanılan ve kullanılabilecek olan tüm yöntem ve teknikler, Doğal Dil İşleme sürecinin parçasıdır. Doğal Dil İşleme, sistemlerin, konuşulan dili anlaması ve bir insan gibi rahat bir şekilde o dili konuşmasını sağlamayı amaçlamaktadır.

Günümüzde üzerinde en çok emek harcanan çalışmalar, özellikle yapay zeka ve makine öğrenimi çalışmaları ile paralel olarak, doğal dil işleme çalışmalarıdır.

Bu çalışmalara örnek verilmesi gerekirse: Metin özetleme çalışmaları, tıp alanı çalışmaları, meteoroloji alanında otomatize edilmiş hava raporu çalışmaları, sesli asistanlar vb. çalışmalar.

3.2.2. Doğal Dil İşleme Yapısı ve Çalışması

Doğal dil işlemenin, tanım olarak net bir sözlük tanımı olmasa da süreç ile ilgili Şekil 2'de görülen gibi net bir mimarisi ve akışı vardır.



Şekil 2 – Doğal Dil İşleme Çalışma Akışı [21]

Dil için kullanılan "Dilin Katmanları" tanımı, doğal işleme süreci ve mimarisi hakkında da oldukça açıklayıcı bir tanımdır ve Şekil 2'de gösterildiği gibi 3 aşaması vardır [21]:

İçerik Planlama: Bu süreç, doğal dil işlemede ilk adımdır ve hem genel olarak "iletişim hedefleri" ve "mevcut bilgi tabanından" etkilenir, hem de istekte bulunan kullanıcının seçeceği kriterlerden etkilenir ve bunlara göre cümleye dökülecek içeriğe karar verir.

Cümle Planlama: Bu süreç "cümle birleştirmeleri (birden fazla mesajın tek cümleye indirgenmesi)", "cümle bağlantılarının oluşturulmaları" ve "referans ifadeler"in belirlendiği ve gerçekleştirildiği süreçtir. Bu işlemlerin kombinasyonu ile ilgili ortak bir uzlaşı bulunmamaktadır, örneğin Matthiessen (1991) "cümle bağlantıları"nın bu aşamada yer almaması gerektiğini savunmaktadır [22].

Şekil 2'de de tıpkı Matthiessen (1991)'in savunduğu gibi, bu işlem (lexicalisation) bir sonraki adımın parçası olarak gösterilmiştir.

Çıktı Şekil Gerçekleme (Dilbilimsel Gerçekleme): Bu süreç sözdizimsel, morfolojik ve ortografik işlemenin gerçekleştiği süreçtir. Doğal dil işleme katmanları içerisinde,

çıktıya ulaşılan son katmandır. Sözdizimsel işlemler tamamlanır ve anlamlı cümleler çıktı olarak elde edilir.

3.2.3. Doğal Dil İşleme Kütüphanleri

Doğal dil işleme kütüphaneleri, yapılandırılmamış veriyi, bir önceki başlıkta açıklanan işlem süreçleriyle yapılandırıken yardımcı olarak kullanılan araçların tümüdür [23].

Doğal dil işleme çalışmalarında hem "cümle planlama" hem de "çıktı şekil gerçekleme" aşamalarında, üzerinde çalışılan dilin tüm özelliklerinin bilinmesi ve kurallarının uygulanması gerekmektedir. Bu sebeple belli bir dile yönelik hazırlanan ticari, açık kaynak ya da özel amaçlı bu kütüphaneler, çalışılan dilin gramer, etimolojik ve morfolojik özelliklerini içerir. Bu sayede hem girdi cümlelerin sistem tarafından sağlıklı işlenmesi sağlanır, hem de çıktı cümlelerin konuşulan dilin kurallarına uygun olması sağlanır. Bu kütüphaneler kullanılarak, cümle bölüntüleme, ögelerine ayırma, kelime köklerini ve ekleerini bulma, bağlaç – edat - zamir gibi cümle ögelerini ayırt etme vb. işlemler rahatlıkla gerçekleştirilebilir. Her bir kütüphane, aynı işlem için farklı oranlarda başarıya sahiptir. Bazen, bir kütüphane kök bulma konusunda iyiyken, başka bir kütüphane cümle sonu bulma konusunda daha iyi olabilir ve birlikte kullanılması gerekebilir.

Bu kütüphaneler içerisinde:

Türkçe Harici Diller İçin Kütüphaneler:

NLTK (Natural Language Toolkit): Açık kaynaklı olan bu kütüphane, 2001 yılında Pennsylvania Üniversitesinde öğretim görevlisi olan Steven Bird tarafından yaratılmıştır. Eski bir kütüphane olması dolayısıyla hakkında çok fazla kaynak ve kitap vardır. Birçok işletim sistemi için versiyonları oluşturulmuştur. Aynı zamanda doğal dil işleme çalışmalarının en çok tercih edilen araçlarından birisidir ve birçok makalede, tezde, projede kullanılmış, üzerine çalışmalar yapılmıştır [24].

CoreNLP: Stanford Üniversitesi tarafından oluşturulan bu kütüphane, kullanıma hazır ve hızlı yapısıyla olgunlaşmış kütüphanelerden birisidr. İlk olarak Java ile geliştirilen kütüphanenin günümüzde farklı programlama dilleri için de versiyonu bulunuyor. Öne çıkan özellikleri: PoS (Part-of-Speech, Konuşma Parçası) işaretlemesi yapabilmesi, öğrenim döngüsü ayrıştırması, özel isim ayrıştırması, hızlı olması ve görece daha kesin sonuçlar üretmesi [25].

TextBlob: NLTK üzerine ek olarak inşa edilen bu kütüphane, NLTK'in kullanım zorluğunu bertaraf etmek ve hızlıca proje geliştirmek için tasarlanmıştır. Python ile geliştirilen TextBlob, halen açık kaynak olarak eklentiye ve geliştirmeye açıktır. Doğal dil işleme süreçlerine derinlemesine dalmadan, hızlıca çalıştırılabilecek bir kütüphanedir. Tıpkı CoreNLP gibi PoS işaretlemesi yapabilen kütüphane, isim ifade ayrıştırması ve duyarlılık çözümlemesi yapabilmektedir [26].

Gensim: Saydığımız diğer kütüphaneler kadar özellikli ve başarılı olmasa da kendi alanında öne çıkan bir kütüphanedir. Özellikle çoklu döküman benzerliklerini çözümlemede ve konu modelleme alanlarında diğer kütüphanelere göre daha başarılıdır. Bünyesinde LDA (Latent Dirichlet Allocation) algoritmasını barındırıan kütüphane, konu ayrıştırma alanında oldukça başarılı sonuçlar elde etmektedir [27].

SpaCy: Cython ile geliştirilen bu yeni kütüphane, NLTK ya da diğerleri gibi tam donanımlı bir kütüphane olmak yerine onlarla birlikte çalışan ve bir tek konuya odaklanan, hızlı ve hafif bir kütüphane. İstatistiki model konusunda oldukça başarılı olan SpaCy, diğer kütüphanelerle birlikte çalışarak doğal işleme projelerinde kendi alanında öne çıkmaktadır [28].

TensorFlow: Google tarafından açık kaynak haline getirilen TensorFlow, tam olarak bir doğal dil işleme kütüphanesi olmaktan öte, makine öğrenimi – yapay zeka ekosistemidir. Dil işleme süreçlerinde, sistemin cümle kurmayı öğrenmesini sağlamak için ihtiyaç duyulan, sınıflandırma, algılama, anlama, keşfetme, tahmin etme ve oluşturma gibi becerilerin sağlanması için yaygınca kullanılmaktadır [29].

Doğal dil işleme kütüphaneleri sadece burada listelenenlerden ibaret değildir ve yapay zeka çalışmaları arttıkça yenileri de eklenecektir. Yukarıda sayılan

kütüphaneler birçok dili desteklemektedir ama Türkçe için direkt Türkçeyi hedef alan kütüphaneler daha başarılı olmaktadır.

Türkçe İçin Geliştirilen Kütüphaneler:

Zemberek: Açık kaynaklı olan bu kütüphane, şu anda Türkçe doğal dil işleme çalışmaları için en aktif olarak geliştirilen kütüphanelerden birisidir [30]. Uzun bir süre yeni versiyonu çıkmayan Zemberek, bu aranın ardından yeni sürümüyle ve elden geçirilmiş yeni kod tabanıyla tekrar kullanılır hale gelmiştir. Ahmet A. Akın tarafından geliştirilen kütüphanede, bölüntüleme ve cümle sonu bulma, morfolojik çözümleme, düzeltme / önerme (spell checker / word suggestion), özel isim çözümleme, metin sınıflandırma, dil tanımlama vb birçok modül bulunmaktadır. Bu tez çalışmasında sonuçların kıyaslanması için kullanılan kütüphane de Zemberek NLP'dir.

ITU Türkçe Doğal Dil İşleme Yazılım Zinciri: İTÜ tarafından geliştirilen bu uygulama web tabanlı olarak çalışmaktadır. Ana hatlarıyla 3 modülden oluşmaktadır [31]: Prof. Dr. K. Oflazer tarafından geliştirilen iki aşamalı morfolojik çözümleyici, H. Sak, M. Saraçlar and T. Güngör tarafından geliştirilen morfolojik bozukluk çözümleme modeli, G. Eryiğit, J. Nivre ve Prof. Dr. K. Oflazer tarafından geliştirilen veri tabanlı bağlılık çözümleyici [32].

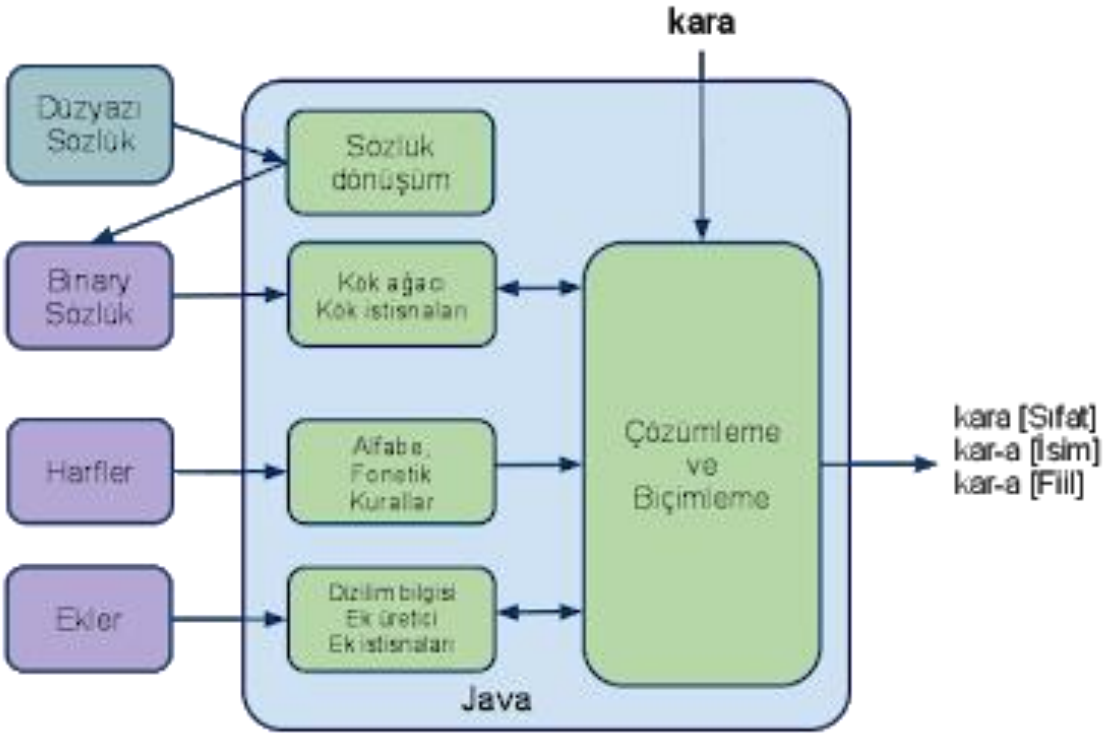
Nûve: H. R. Zafer tarafından geliştirilen ve yönetilen bu kütüphane, morfolojik çözümleme, morfolojik oluşturma, kök-ek ayrımı, cümle sonunu belirleme ve N-gram ayıklama özelliklerine sahiptir [33].

Türkçe doğal dil işleme çalışmaları için geliştirilmekte olan birçok kütüphane ve çalışma bulunmaktadır. Doğal dil işleme çalışmalarında birden fazla kütüphane kullanımı da mümkün olmaktadır, adı geçen diğer kütüphanelerden bazıları: "TRmorph", kelime kökü bulmak için "seq2seq", O. T. Yıldız tarafından geliştirilen "NLP Toolkit" vb.'dir.

3.2.4. Zemberek NLP

Yukarıda ismi geçen Türkçe doğal dil işleme kütüphaneleri içerisinde en olgunlaşmış olanı Zemberek NLP'dir. Bu tez çalışmasında da geliştirilen yöntem, Zemberek NLP sonuçları ve altın özetlerle kıyaslanmıştır.

Kütüphanenin yapısı iki ana parçadan oluşmaktadır; "Dil yapı bilgisi" ve "Dil işleme çekirdeği" [34]. Kütüphanenin çekirdek kodları Türkçe için yazılmış olsa da diğer diller için adaptasyona açıktır. Bu iki yazılım bölümünün nasıl çalıştığı Şekil 3'te gösterilmiştir.



Şekil 3 – Zemberek NLP, Akış Diyagramı [35]

Zemberek NLP birçok görev modülünü barındırmaktadır:

Çekirdek: Özelleştirilmiş veri yapılarını ve yardımcı nesneleri barındırır.

Morfoloji Modülü: Bu modül morfolojik çözümleme ve oluşturma işlemlerini üstlenir. Morfolojik anlamda zengin olan Türkçe'nin, karmaşık kelime yapısını

çözmek için sözlük kuralları tanımlanmıştır. Ayrıca belirsizlik çözümü için de bir tahmin ve sınıflandırma mekanizması bulunmaktadır.

Bölüntüleme (Cümle Sonu Bulma) Modülü: Doğal dil işleme süreçlerinin en önemli parçası, cümleleri başarıyla ayırt edebilmektir. Zemberek NLP, bu işlem için kural tabanlı nesneler yanısıra, tahmin ve çözümlemeye dayalı bir model kullanarak cümle sonlarını bulmaktadır.

Normalizasyon Modülü: Bu modül, yazım hatalarını bulmak ve düzeltmek için gerekli olan nesneleri barındırır. Hata düzeltme ve yerine kelime önerme gibi işlemleri gerçekleştirir.

Özel İsim Tanımlama Modülü: Kaynak içerisinden özel isimleri ayırt etme görevini üstlenen bu modül, kullanıcı tarafından kullanıcı ihtiyaçlarına göre şekillendirilmesi (eğitilmesi) gereken bir modüldür. Zemberek NLP'nin şu anki sürümünde, bu modüle entegre edilmiş hazır bir model bulunmamaktadır.

Sınıflandırma Modülü: Doğal dil işleme çalışmaları için önemli bir görev olan sınıflandırma, Zemberek NLP bünyesinde her ne kadar bir modül olarak yer alsa da, aslında bu modül, Java ile yazılmış başka bir sınıflandırma uygulaması olan "fastText"in kodlarına geçiş sağlamakta ve kullanmaktadır. Bu modülün de önce kullanıcı tarafından şekillendirilmesi (eğitilmesi) gerekmektedir.

Dil Tanımlama Modülü: Temel dil tanımlama işlemleri için geliştirilmiştir. Harf bazında n-gram modelini kullanan modül, 62 farklı dili tespit edebilmektedir.

Dil Modelleme Modülü: Dil model kütüphanesinin sıkıştırma algoritmalarını içeren modüldür. Geliştirilen "SmoothLM" kütüphanesini kullanan bu modül, yüksek sıkıştırma oranına ulaşabilmek için "Minimal Perfect Hash Functions" (MPHF) kullanmaktadır [36].

3.3. Ölçüm Teknikleri

3.3.1. Metin Özetleme Başarı Ölçümü

Metin özetleme çalışmalarında, aşılması gereken bir diğer engel, çıktı olarak elde edilen özetlerin başarılarını belirleyebilmektir. 1960'lardan itibaren, metin özetleme çalışmaları ile birlikte bu alanda da çalışmalar yapılmaktadır [37]. Günümüze kadar oy birliği ile üzerinde anlaşılmış tek bir çözüm bulunmasa da, sürecin nasıl işleyeceğine dair ana hatlar netleşmiştir. Özet başarılarını ölçümlemek için bazı zorlukların üstesinden gelinmesi gerekmektedir [38]:

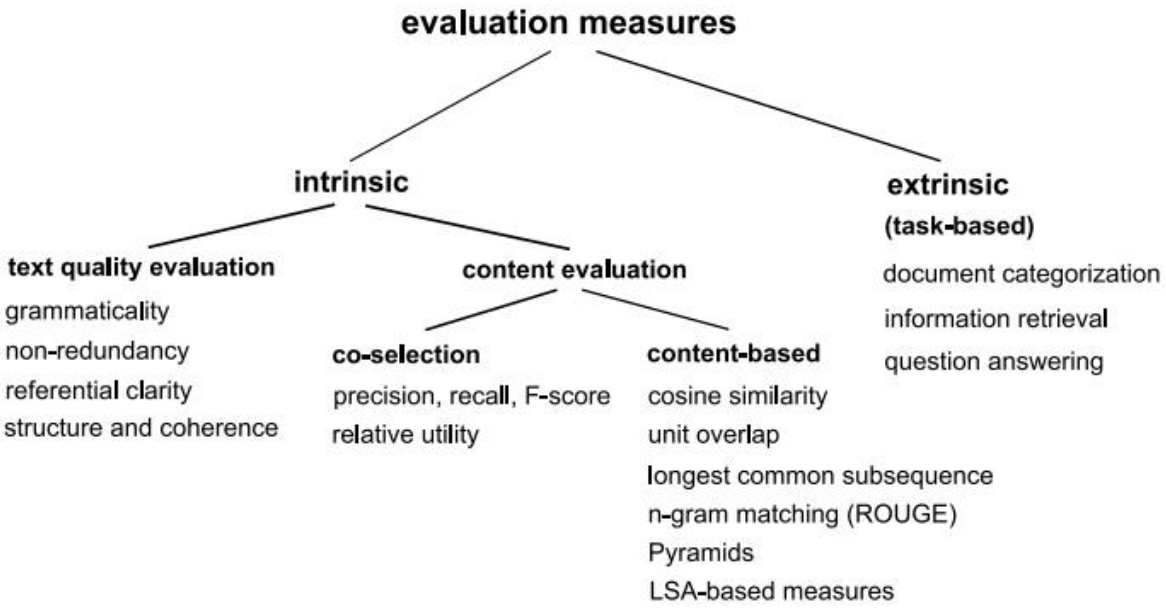
- Özet çıktılar sistemler tarafından oluşturulmaktadır ve doğal dil akışı ile ilintilidir. Bazı durumlarda çıktı olarak elde edilen özet, aranan sorunun cevabı olsa da, daha iyi ifade edilebilmeye gerek duyabilir.

- Oluşturulan özetler, nihayetinde insan tarafından değerlendirileceği için, bu zorlu muhakemeyi başarıyla geçebilmesi için geliştirilen yöntemler kaynak dostu olmayabilir, vakit ve kaynak israf edebilir.

- Özetleme, aynı zamanda sıkıştırma ile ilişkilidir, farklı sıkıştırma oranlarındaki özetlerin değerlendirilmesi, sürecin karmaşıklığının ve boyutunun artmasına neden olabilir.

- Özetler farklı amaçlar için oluşturulduğundan, başarısını belirleyebilmek için amacın ne olduğuna dair kriterlerin ölçüm sürecine dahil edilmesi gerekir, bu da değerlendirme süreç tasarımının karmaşıklaşmasına yol açabilir.

Metin özetleme çalışmalarını ve doğal dil işleme çalışmalarını değerlendirme metotları, kabaca iki başlık altında toplanabilir [39]: "İçsel Değerlenirme Metotları (Intrinsic)" ve "Dışsal Değerlendirme Metotları (Extrinsic)".



Şekil 4 – Özet Değerlendirme Ölçüm Sistematiği [40]

İçsel Değerlendirme Metotları: Bu metotlar, çıktı olarak elde edilen özetlerin kendi içlerinde, kendi başarı kıstaslarıyla değerlendirilmesini sağlamaktadır ve aşağıdaki listede belirtilen başarıları ölçümler:

- **Özet Uyumu:** Özeti oluşturan cümlelerin uyumu, aralarında kopukluk olup olmadığı.

- **Özet Bilgi Mahiyeti:** Özette yer alan bilginin yararlı ve yeterli olup olmadığı.

Dışsal Değerlendirme Metotları: Bu metotlar ise çıkan özetlerin, ilişkili oldukları diğer görev ve süreçlere sağladıkları katkı ile başarılarını ölçmeye çalışmaktadır. Sınıflandırması doğru mu, bilgi aktarımı yeterli mi, isteğe cevap veriyor mu?

Metin özetleme çıktılarını değerlendirmede insanın yeri çok önemlidir. "İçsel Değerlendirme" yapılırken, başarı, insanlar tarafından oluşturulan altın özetlerle ölçümlenir. Nadir durumlarda ise çıktı olarak elde edilen özetlerin, kaynak girdiyle ölçümlenmesi ve değerlendirilmesi istenir.

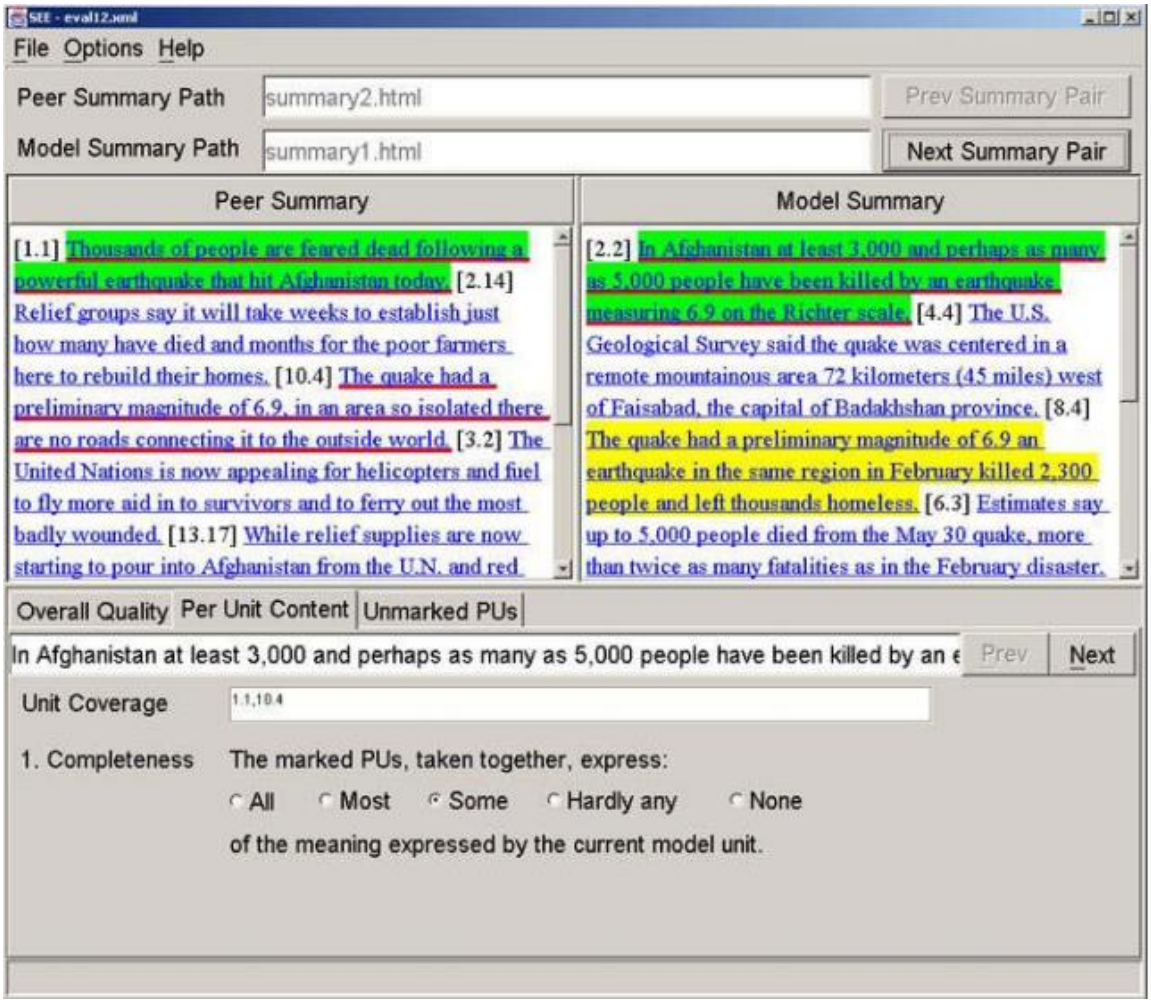
Öte yandan, "Dışsal Değerlendirme" yapılırken, özetin görevini ne kadar başarıyla tamamladığı, parçası olduğu sistemin modülleriyle değerlendirilebilir. Aynı

zamanda bu sürece “gerçek” insan da dahil edilebilir, bu tamammen sistem kurucusunun tercihidir.

3.3.2. Başarı Ölçümü Değerlendirme Araçları

Metin özetlerinin başarısının ölçümlenmesi sürecinde, detaylı ve tekrarlanabilir bir kıyaslama prosedürü yaratılması ve bu işlemlerin bir kısmının otomatize edilmesi için kaynak metnin, çıkan özetin ve referans özetin bir arada bulunduğu ve erişildiği bir yapı oluşturmak oldukça faydalıdır [41]. Bu amaçla geliştirilmiş birçok araç/yöntem geliştirilmiştir:

Summary Evaluation Environment (SEE): C. Lin tarafından ilk kez 2001 yılında geliştirilen uygulamada, kullanıcının girdiği iki farklı metin yan yana kıyaslanarak başarı ölçümü yapılmaktadır. Girilen metin özetlerinden birisi, referans özet iken diğeri emsal/aday özettir. Uygulama, metinleri ön-işleme tabii tutarak kıyaslama öncesi cümle seviyesinde bölüntüleme yapmakta ve kullanıcının seçeceği kriterlere göre ölçüm yapmaktadır.



Şekil 5 – SEE Ölçümleme Oturumundan Bir Ekran Görüntüsü [42]

MEADeval (ex. LexRank): MEADeval (Winkel, Radev), 2002 yılında geliştirilmiştir. MEAD ölçüm sistemini kullanan ve DUC (Document Understanding Conferences) tarzı ayıklama özetleri oluşturan bu uygulama, Perl ile yazılmış ve geniş bir framework'tür (çatı sistem). Özet oluşturma ve ölçümleme yapabilmektedir. MEAD, cümleleri puanlarken 3 farklı kıstası değerlendirir: Cümle uzunluğu, ağırlık merkezi ve cümle pozisyonu. Referans ve emsal özetler arasındaki cümle örtüşmeleri üzerinden değerlendirme sonucu oluşturur.

ROUGE (ISI): IBM'in BLEU (Bilingual Evaluation Understudy) ölçüm metriğinin güncellenmiş / günümüze adapte edilmiş versiyonudur. DUC-2002'deki çalışmaların ölçümlenmesi yapılırken kullanılan bu yöntem, şaşırtıcı derecede yüksek başarı yakalamış ve insan değerlendirmesine yakın sonuçlar üretmiştir. BLEU "hassisyet

odaklı" iken, aksine ROUGE "geri-çağırım" odaklıdır. Emsal ve referans özetleri kıyaslarken, n-gram'lar (n adet kelime örtüşümü) kullanarak değerlendirme sonucu üreten ROUGE, BLEU'nun aksine "uzunluk" hatası vermez. Bu tez çalışmasında da kullanılan sonuç değerlendirme yöntemidir [43].

3.3.3. ROUGE Ölçüm Metriği

DUC çalışmalarında olduğu gibi günümüz birçok metin özetleme çalışmasında da, insan değerlendirmesine en yakın sonuçlar ürettiği için kullanılan ROUGE, bu tez çalışmasında da kullanılmıştır. ROUGE sistem tarafından oluşturulan özet ile insan tarafından oluşturulan referans (altın) özet arasındaki kelime çakışımlarını sayarak, 0 ve 1 arasında yüzdelik bir sonuç döndürür. "Recall (aynı zamanda BLEU değeri olarak da adlandırılır)" ve "Precision (aynı zamanda Rouge değeri olarak da adlandırılır)" ölçümleri yapılabilir. Recall ölçümü yapılırken sistem özeti ile referans özetin çakışan kelime sayıları Denklem 1'deki yöntemle hesaplanırken, precision (hassaslık) ölçümü Denklem 2'deki yöntemle hesaplanır [44].

$$Rouge\ Recall = \frac{Çakışan\ Kelime\ Sayısı}{Referans\ Özetteki\ Toplam\ Kelime\ Sayısı} \quad (1)$$

$$Rouge\ Precision = \frac{Çakışan\ Kelime\ Sayısı}{Sistem\ Özetindeki\ Toplam\ Kelime\ Sayısı} \quad (2)$$

Rouge ölçüm metriğinde recall ve precision değerleri elde edildikten sonra, bu iki değer üzerinden yeni bir hesaplama daha yapılarak "Rouge F1 Score" sonuçlarına ulaşılır, ulaşılan bu değer, tek başlarına bir birleriyle uyumsuz görünen ve birarada anlamlandırılamayan recall ve precision değerlerini anlamlandırarak, sonucu tek bir

ölçüm çıktısına çevirmeye yarar ve Denklem 3'teki yöntemle hesaplanır, elde edilen sonuçların harmonik ortalamasıdır.

$$Rouge\ F1\ Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (3)$$

ROUGE, özetler arası çakışımları hesaplarken, n-gram kullanır. Burada "N", çakışımları kontrol edilecek kelime zincirinin uzunluğunu belirtmektedir ve ROUGE-N olarak gösterilir. N, 1, 2, 3 veya 4 olabilir. Örtüşen en uzun kelime zincirleri de ayrıca ROUGE ile hesaplanabilir.

Bu tez çalışmasında hesaplamalar, ROUGE-1, ROUGE-2 ve ROUGE-3 sonuçlarına göre değerlendirilmiştir.

3.4. Veriler ve Uygulama

3.4.1 Verilerin Toplanması

Bu çalışmada, "f5haber.com" sitesine ait veri tabanından, içerik ortağı olduğu "hürriyet.com.tr" haber portalına ait Mart 2018 – Nisan 2018 tarihleri arasında yayınlanan toplam 1.005 adet haber özetlenmiştir. Bu haberler, 5.000 adetlik bir veri havuzundan teker teker seçilmiş, ve gündelik haber akışına uygun kategorilerde olmalarına özen gösterilmiştir. Bu sayede, özetleme sonuçlarının değerlendirilmesinde, haberin kategorisinin ve içeriğinin sürece etki etmemesi sağlanmıştır.

Haberler sisteme, başlık, özet ve detay bölümleri ayrıştırılmış şekilde eklenmiş ve özetleme çalışmaları sadece detay metinleri içerisinde yapılmıştır.

Tablo 1 - Tez Çalışmasına Konu Haberlerin Veritabanına Kayıt Detayları

id	baslik	ozet	detay

1	'Cicişler' Etiler'deki evlerinde bir…	"Cicişler" olarak tanınan Esra Ers…	Beşiktaş, Tepecik Yolu cadde...
2	İran tehdit etti! 'İhtiyacımız olan h…	İran Cumhurbaşkanı Hasan Ruha…	İran'ın Ulusal Ordu Günü dol...
3	Ankara'da kritik görüşmede son d…	Cumhurbaşkanı Recep Tayyip Er…	MHP Genel Başkanı Bahçeli,…
4	Trabzonspor'da Demir Grup Sivas…	Trabzonspor, Spor Toto Süper Li…	Mehmet Ali Yılmaz Tesisleri'…
5	Okul servisinde bu kez bıçaklı ka…	Adana'da lise öğrencisi iki kardeş,…	Merkez Sarıçam ilçesi Mümin...
6	Sosyal medyadan canlı yayınladık…	GAZİANTEP- Şanlıurfa yolunda,…	Kaza, dün gece, Gaziantep- Ş…
7	Ibrahimovic Dünya Kupası için …	ABD'nin Los Angeles Galaxy eki…	ABD'de yayınlanan Jimmy K…

Veriler veritabanına girildikten sonra, cümle sonu bölüntülemesi için normalizasyon yapabilmek amacıyla cümlelerdeki cümle sonu anlamına gelmeyen noktalar geliştirilen bir web uygulaması ile teker teker insan tarafından temizlenmiştir.

Bu cümle normalizasyonu çalışması sonrasında, ROUGE ölçümü yapabilmek için her bir haberin içerisinden, yine insan tarafından 3'er adet (referans özeti oluşturacak) cümle seçilmiş ve Şekil 6'daki ara yüze sahip, geliştirilen bir web uygulaması ile veritabanına eklenmiştir.



Şekil 6 - Referans Özeti Oluşturma Arayüzüne Ait Ekran Görüntüsü

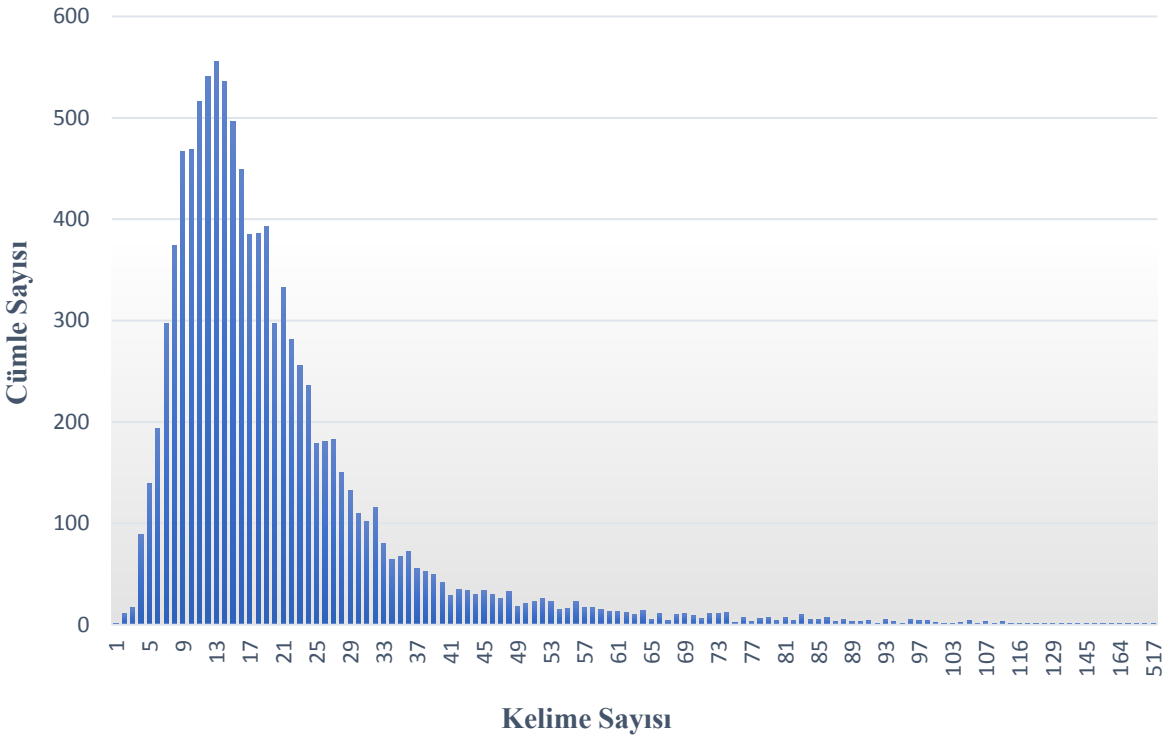
Bütün bu özetleme süreci ön işlemleri ardından, veri seti istatistiği incelendiğinde

Toplam Haber Sayısı: 1.005

Toplam Cümle Sayısı: 10.116

Toplam Kelime Sayısı: 206.232

sonucu elde edilmiştir. Bu veri setine ait "cümle – kelime sayısı" grafiği Şekil 7'de görülebilir.



Şekil 7 - Ana Veri Seti Cümle-Kelime Dağılımı

Bu veri seti, üzerinde çalışmak için yeterli bir veri seti olsa da, geliştirilen tekniğin benzer diğer çalışmaların sonuçlarıyla kıyaslanabilmesi için internet ortamında ücretsiz olarak ve metin özetleme çalışmalarına katkı sağlaması amacı ile yayınlanan bir diğer veri seti ile de çalışma yapılmıştır [45]. Bir önceki veri setinde olduğu gibi, haberler veritabanına aktarıldıktan sonra, cümle noktalama işaretleri normalizasyonu yapılmış, referans özet oluşturmak için insan tarafından seçilen 3'er adet cümle veritabanına eklenmiştir.

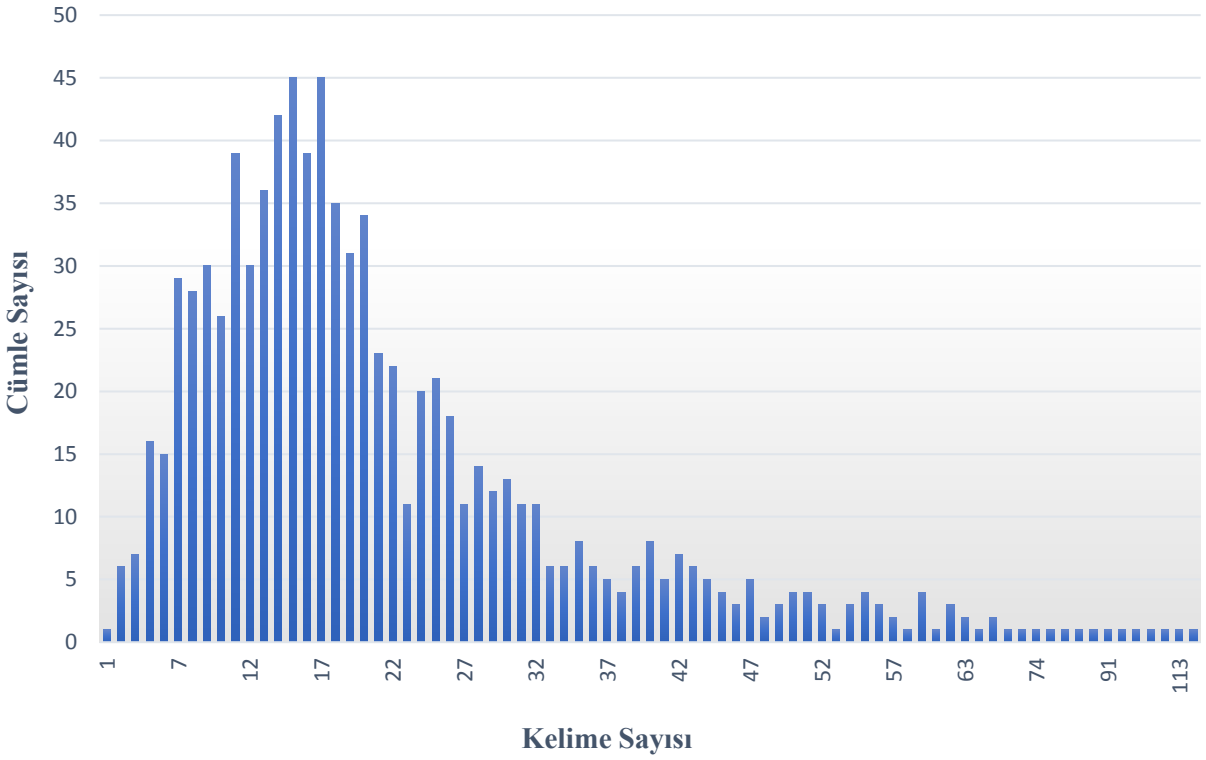
Bu 100 haberlik ek veri setinin istatistikleri ise şu şekildedir;

Toplam Haber Sayısı: 100

Toplam Cümle Sayısı: 862

Toplam Kelime Sayısı: 18.769

Bu veri setine ait "cümle – kelime sayısı" grafiği Şekil 8'de görülebilir.



Şekil 8 - Kontrol Veri Seti Cümle-Kelime Dağılımı

3.4.2. Frekans Listesi Oluşturma Algoritması

Bu tez çalışmasının ana amacı, NLP kütüphanesi kullanmadan, terim sıklık listesi üzerinden cümle puanlaması ile seçilen cümlelerden oluşturulacak özetlerin başarısını test etmektir.

Bu sebeple geliştirilen algoritma, uygulama aşamasında aşağıda belirtilen adımları takip etmekte ve sonuç üretmektedir. Algoritmanın çıktıları, Zemberek NLP

kullananılan bir diğer özetleme süreci ve referans özetlerle ROUGE-N metriği kullanılarak kıyaslanmıştır.

Kelime Sıklık Listesi Oluşturma Algoritma Adımları:

Geliştirilen algoritma ilk olarak bir kelime sıklık listesi oluşturmaktadır, bu daha sonra cümle puanlamasında kullanılacaktır:

- **Adım 1:** Girdi olarak gelen kaynak metin, oluşturulan bir bağlaç-edat listesindeki girdiler ile kıyaslanır, ve özete değer katmayacağı düşünülen bu bağlaç ve edatlar, kaynak metinden atılır.

- **Adım 2:** Bağlaç ve edatlardan temizlenmiş kaynak metin sonrasında harf-sayı ve "."(nokta) haricinde kalan tüm karakterlerden arındırılır. (tek tırnak, çift tırnak, tire vb. harf ya da sayı olmayan tüm karakterler silinir). Aynı zamanda birden fazla boşluk art arda gelmişse, onlar da tek bir boşluk haline dönüştürülür.

- **Adım 3:** Ön-işleme sürecinde temizlenen kaynak metin, cümle sonunu belirten (".") nokta işaretlerinden cümlelerine bölünür.

- **Adım 4:** Her bir cümle sırasıyla işleme alınır ve o an üzerinde işlem yapılan cümle kelimelerine ayrılır (kelimeler arası boşluklar kullanılarak). Oluşan bu kelime listesindeki her bir kelime kaynak metindeki kelimeler ile teker teker karşılaştırılır. Bu karşılaştırma, harf harf yapılır ve aranan benzerlik, karşılaştırılan iki kelimenin de ilk harflerinden itibaren uyuşmasıdır. Bu uyuşma (benzer harf zinciri), ilk harf ile başlar ve eğer son harfe kadar devam ederse kelimeler bire bir aynı demektir. Eğer bu benzer harf zinciri bir noktada koparsa, o zaman da kelimelerden birisi diğerini içeriyor ve kısa olan kelime, uzun olan kelimenin (muhtemel) kökü olabilir demektir. Örn: AYAK ve AYAKKABI kelimeleri 4. harften itibaren benzerlik zinciri kopuyor ama öte yandan uzun olan kelime kısa olanı içeriyor ve "AYAK" kelimesi "AYAKKABI" kelimesinin muhtemel kökü olarak işaretleniyor. Böyle bir "muhtemel kök" ve "muhtemel türetilmiş" kelime eşi yakalandığında, kaynak metin içerisinde muhtemel

türetilmiş kelimenin tüm kopyaları da, muhtemel kök olarak işaretlenmiş kelimeyle değiştirilir. Biraz önceki örnek göz önüne alındığında, bu benzerlik yakalndığında kaynak metindeki tüm "AYAKKABI" kelimeleri "AYAK" haline getirilir. Bu değişikliklerin ardından benzer bu kelimelerden muhtemel kök olan kısa versiyonu "kelime sıklık listesi"ne eklenir.

Kelime sıklık listesi oluşturulduktan sonra, süreç tekrar başa döner bu sefer cümle puanlaması yapmak için gerekli işlemleri yapar.

Cümle Puanlama Algoritma Adımları:

- **Adım 1:** İlk cümleden itibaren, bu sefer o cümleyi oluşturan kelimeler teker teker "kelime sıklık listesi" ile karşılaştırılır. Bir önceki süreçte olduğu gibi "muhtemel kök" ve "muhtemel türetilmişlik" ilişkisi bu kıyaslamada da yakalanırsa, bu sefer frekans listesinde "türetilmiş" kelime, muhtemel kök yerine geçecek kısa versiyonu ile değiştirilir, böylece ileriye dönük sadeleştirme yapılır. Her bir cümle, "kelime sıklık listesindeki" kelimelerden hangilerini içerdiklerine, yine "sıklık listesinde" saklanan kelime puanları ile puanlanır. En sık tekrarlanan kelimeler ne kadar içerdiğine bağlı olarak, bir cümle o kadar yüksek puan alır.

- **Adım 2:** Tüm cümleler puanlandıktan sonra, en yüksek puan alan 3 cümle, sistem özeti olarak seçilir çıktı verilir.

Geliştirilen bu teknikle, Türkçe'nin sondan eklemeli bir dil olması, matematiksel bir avantaja dönüştürülmeye çalışılmıştır. Tabii ki bir birini içeren her Türkçe kelime arasında kök-ek-türeme ilişkisi bulunmamaktadır. Ama üzerinde çalışılan haber metinleri ve bu metinleri oluşturan kelimeler açısından değerlendirildiğinde, ortalama üstü bir başarı hedeflenmiştir.

3.4.3. Zemberek NLP Uygulaması

Bir bakıma Zemberek NLP kütüphanesi taklit etmeye çalışan ve kelimeleri (muhtemel) kökleri ile ifade eden, yukarıda bahsedilen matematiksel yöntemin çıktılarını değerlendirebilmek için veri setindeki tüm haberler, aşağıdaki kod adımlarını / iş akışını takip ederek aynı zamanda Zemberek NLP kütüphanesi aracılığıyla da bir kelime sıklık listesi oluşturulmuştur:

Kelime Sıklık Listesi Oluşturma Zemberek NLP Adımları:

Bu süreçte, sıklık listesi oluşturmadan önce yapılan işlerin tümü, matematiksel kök bulma tekniği (algoritması) ile aynıdır. Süreç sıklık listesi oluşturma aşamasına geldiğinde ayrışır.

- **Adım 1:** Girdi olarak gelen kaynak metin, Zemberek NLP kütüphanesinin kök bulma yöntemi kullanılarak sadeleştirilir, kaynak metindeki tüm kelimeler sadece kökleri kalacak şekilde sadeleştirilir.

- **Adım 2:** Kaynak metinde sadece kelimelerin kökü bırakıldıktan sonra, her bir kelime kaynak metnin tümü ile kıyaslanır, tekrar sayısına göre bir puan alır ve sıklık listesi oluşturulur.

- **Adım 3:** Oluşturulan bu sıklık listesi kullanılarak her bir cümle puanlamaya tabii tutulur.

- **Adım 4:** en çok puan alan ilk 3 cümle kaynak metnin özeti olarak işaretlenir ve çıktı alınır.

3.4.4. Özetlerin Ölçümlenmesi

Veri setindeki her bir haber için, hem geliştirilen matematiksel yöntem, hem de Zemberek NLP işleme yöntemi ile oluşturulan üçer cümlelik özetler, veri tabanına kaydedildikten sonra, sonuçların değerlendirilmesi ve her iki yöntemin bir birleriyle ve

referans özetlerle kıyaslanması ve sonuçların değerlendirilmesi için yeterince veri sağlanmıştır.

Oluşturulan özetler, ROUGE-N metriği kullanılarak referans özetlerle kıyaslanmıştır. ROUGE-1, ROUGE-2 ve ROUGE-3 için her bir özet teker teker puanlanmıştır:

Tablo 2 - Zemberek Nlp Rouge-N Ölçüm Sonuçları (Ana Veri Seti)

ZEMBEREK NLP ROUGE-N ÖLÇÜM SONUÇLARI			
N-GRAM	**RECALL**	**PRECISION**	**F1 SCORE**
ROUGE-1	0,6384	0,6281	0,6330
ROUGE-2	0,5950	0,5840	0,5893
ROUGE-3	0,5764	0,5653	0,5707
ORTALAMA	**0,6032**	**0,5924**	**0,5976**

Bu tez çalışmasında geliştirilen matematiksel yöntem ile oluşturulan özetlerin Rouge-N ölçüm sonuçları:

Tablo 3 - Matematiksel Yöntem Rouge-N Ölçüm Sonuçları (Ana Veri Seti)

MATEMATİKSEL YÖNTEM ROUGE-N ÖLÇÜM SONUÇLARI			
N-GRAM	**RECALL**	**PRECISION**	**F1 SCORE**
ROUGE-1	0,5731	0,4356	0,4948
ROUGE-2	0,4882	0,3842	0,4298
ROUGE-3	0,4622	0,3657	0,4082
ORTALAMA	**0,5078**	**0,3951**	**0,4442**

Elde edilen bu sonuçlar kıyaslandığında Zemberek NLP kütüphanesinin referans özetlere çok daha yakın sonuçlar ürettiği ama öte yandan geliştirilen matematiksel yöntemin de kabul edilebilir sınırlarda başarılı olduğu görülmektedir.

Matematiksel yöntem, ortalama Recall değerleri üzerinden hesaplandığında, %83,35 oranında Zemberek NLP ile aynı başarılı sonuçlarını yakalamıştır.

Elde edilen sonuçlar teker teker haber bazında incelendiğinde ise, Zemberek NLP kütüphanesinin geliştirilen matematiksel yönteme göre çok daha başarılı olduğu Tablo 4 ve Tablo 5'te görülmektedir:

Tablo 4 - Zemberek Nlp Rouge-N Tam Puan (Ana Veri Seti)

ZEMBEREK NLP TAM PUAN ÖLÇÜM SONUÇLARI		
N-GRAM	**RECALL**	**PRECISION**
ROUGE-1	165	153
ROUGE-2	134	131
ROUGE-3	133	131
ORTALAMA	**144**	**138,33**
GENEL ORTALAMA	**141,16**	

Aynı ölçüm matematiksel yöntem için yapıldığında ise:

Tablo 5 - Matematiksel Yöntem Rouge-N Tam Puan (Ana Veri Seti)

MATEMATİKSEL YÖNTEM TAM PUAN ÖLÇÜM SONUÇLARI		
N-GRAM	**RECALL**	**PRECISION**
ROUGE-1	49	49
ROUGE-2	11	10
ROUGE-3	11	10

ORTALAMA	**23,66**	**23**
GENEL ORTALAMA	**23,33**	

Kontrol Veri Seti Sonuç Ölçümleri:

Bu ölçümlerin ardından, geliştirilen matematiksel yöntem, 100 haberlik -başka bir frekans tabanlı istatistiki yöntemle özetlenmiş- veri seti ile karşılaştırılmıştır. Aynı zamanda bu kontrol veri seti Zemberek NLP ile de özetlenmiş ve sonuçları karşılaştırma tablosuna eklenmiştir:

Bu kontrol veri seti, internet üzerinden elde edilmiştir ve veri seti içerisinde -detayları açıklanmayan başka - bir istatistiki yöntemle çıkarılan özetler gömülü olarak gelmiştir.

Tablo 6 - Zemberek Nlp Rouge-N Ölçüm Sonuçları (Kontrol Veri Seti)

ZEMBEREK NLP ROUGE-N ÖLÇÜM SONUÇLARI			
N-GRAM	**RECALL**	**PRECISION**	**F1 SCORE**
ROUGE-1	0,7738	0,7675	0,7705
ROUGE-2	0,7521	0,7429	0,7474
ROUGE-3	0,7407	0,7302	0,7354
ORTALAMA	**0,7555**	**0,7468**	**0,7511**

Geliştirilen Yöntem Sonuçları:

Tablo 7 - Matematiksel Yöntem Rouge-N Ölçüm Sonuçları (Kontrol Veri Seti)

MATEMATİKSEL YÖNTEM ROUGE-N ÖLÇÜM SONUÇLARI			
N-GRAM	**RECALL**	**PRECISION**	**F1 SCORE**
ROUGE-1	0,7028	0,5958	0,6448

ROUGE-2	0,6459	0,5549	0,5969
ROUGE-3	0,6246	0,5371	0,5775
ORTALAMA	**0,6577**	**0,5626**	**0,6064**

Diğer İstatiski Yöntem Sonuçları:

Tablo 8 - Diğer Yöntem Rouge-N Ölçüm Sonuçları (Kontrol Veri Seti)

DİĞER YÖNTEM ROUGE-N ÖLÇÜM SONUÇLARI			
N-GRAM	**RECALL**	**PRECISION**	**F1 SCORE**
ROUGE-1	0,6254	0,6864	0,6544
ROUGE-2	0,5859	0,6415	0,6124
ROUGE-3	0,5689	0,6221	0,5943
ORTALAMA	**0,5934**	**0,5626**	**0,6203**

Sonuçlar değerlendirildiğinde, bu tez çalışmasında geliştirilen matematiksel yöntem, detayları bilinmeyen benzer bir istatistiki yöntemden Recall ortalama sonuçlarına göre %10,24 oranında daha başarılı özetler üretmiştir. Öte yandan Zemberek NLP'den %12,49 oranında daha az başarıya ulaşmıştır.

Bu kontrol veri setinde de haber bazında inceleme yapıldığında ;

Zemberek NLP Sonuçları:

Tablo 9 - Zemberek Nlp Rouge-N Tam Puan Sonuçları (Kontrol Veri Seti)

ZEMBEREK NLP TAM PUAN ÖLÇÜM SONUÇLARI		
N-GRAM	**RECALL**	**PRECISION**
ROUGE-1	44	36
ROUGE-2	37	36
ROUGE-3	37	36
ORTALAMA	**39,3**	**36**
GENEL ORTALAMA	**37,65**	

Geliştirilen Yöntem Sonuçları:

Tablo 10 - Geliştirilen Yöntem Rouge-N Tam Puan Sonuçları (Kontrol Seti)

MATEMATİKSEL YÖNTEM TAM PUAN ÖLÇÜM SONUÇLARI		
N-GRAM	**RECALL**	**PRECISION**
ROUGE-1	16	16
ROUGE-2	2	2
ROUGE-3	2	6
ORTALAMA	**6,66**	**6,66**
GENEL ORTALAMA	**6,66**	

Diğer İstatiski Yöntem Sonuçları:

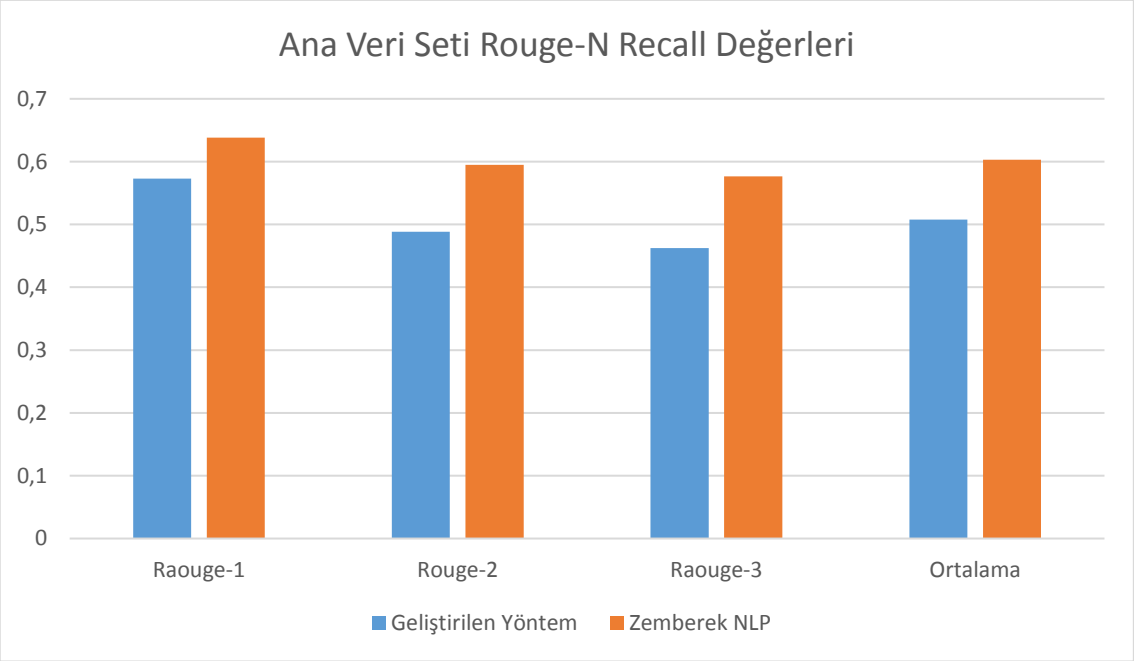
Tablo 11 - Diğer Yöntem Rouge-N Tam Puan Sonuçları (Kontrol Veri Seti)

DİĞER YÖNTEM TAM PUAN ÖLÇÜM SONUÇLARI		
N-GRAM	**RECALL**	**PRECISION**
ROUGE-1	19	27
ROUGE-2	19	22
ROUGE-3	19	22
ORTALAMA	**19**	**23,66**
GENEL ORTALAMA	**21,33**	

BÖLÜM 4. BULGULAR VE YORUMLAR

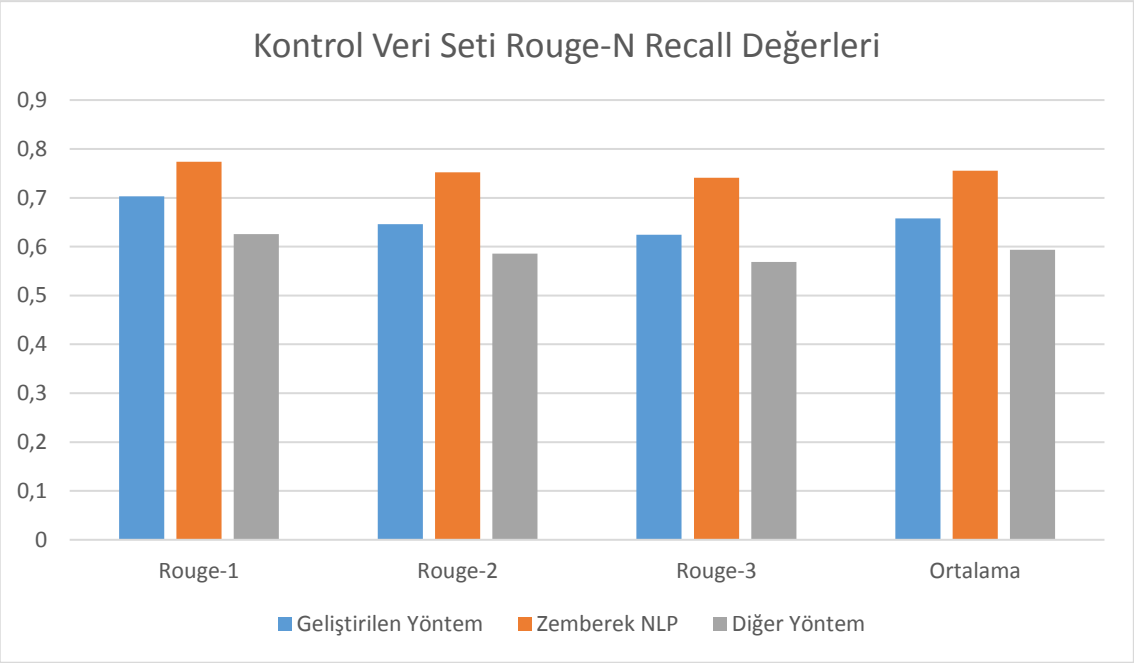
4.1. Bulgular

Bu tez çalışması kapsamında, geliştirilen matematiksel yöntem ile Zemberek NLP kütüphanesi ve bir diğer istatistiki yöntem kıyaslanmıştır. Bu kıyaslama neticesinde elde edilen sonuçlar ışığında; Zemberek NLP kütüphanesi beklendiği gibi insan özetlerine en yakın sonuçları üretmiş ve en yüksek ölçüm puanlarını almıştır. Karşılaştırma grafiği Şekil 9'da gösterilmiştir.



Şekil 9 - Ana Veri Seti Rouge-N Recall Karşılaştırma Grafiği

Öte yandan bu çalışma kapsamında geliştirilmiş olan matematiksel yöntem, genel açıdan kabul edilebilir sınırlar içerisinde sonuçlar üretmiş ve kıyaslandığı diğer teknikten yaklaşık %10 oranında daha başarılı olmuştur.



Şekil 10 - Kontrol Veri Seti Rouge-N Recall Karşılaştırma Grafiği

Elde edilen ölçüm sonuçları derinlemesine incelendiğinde ise Zemberek NLP kütüphanesinin başarısı devam ederken, hem diğer tekniğin hem de bu çalışmada geliştirilen matematiksel yöntemin başarı oranları düşmüştür.

Bu sonuçlar ışığında, genel bir değerlendirme yapıldığında (özellikle Zemberek NLP ve bu çalışmada geliştirilen yöntem arasında) Zemberek NLP kütüphanesi, çok sayıda - insana yakın - başarılı özet oluşturmuş ama öte yandan başarısız olduğu özetlerde de çok daha düşük puan almıştır. Tıpkı derslerinin yarısından tam puan alıp, diğer yarısından geçer puan alan bir öğrenci gibi ve bu sayede genel ortalamada yüksek ölçüm değerlerine ulaşmıştır. Bu durum tam puan alan ROUGE-N sonuçlarında da görülmüştür ve Tablo 4'te yer almaktadır.

Diğer yandan bu çalışmada geliştirilen yöntem, tam puan alma konusunda Zemberek NLP kadar başarılı olmasa da genel olarak haber bazında ortalama bir başarı skalası yakalamış ve ROUGE-N tam puan özet sayısı az olsa da genel ortalama

puanında Zemberek NLP'nin yakaşık %10-12 gerisinde kalmıştır. Bu durum Tablo 7 ve Şekil 9'da görülmektedir.

Bütün bunlar bir arada değerlendirildiğinde, kullanım amacı ve hassasiyet ihtiyacına göre, Zemberek NLP kadar başarılı sonuçlar hedeflenmeyen projelerde, bu çalışmada geliştirilen yöntem kullanılmaya değerdir.

Çünkü %10-12'lik başarı kaybını tolere edebilecek avantajları bulunmaktadır. Hiçbir kütüphaneye veya ek kuruluma ihtiyacı yoktur, istenilen programlama dilinde yeniden yazılabilir, bu sayede javascript ortamına aktarılması halinde, sunucu tarafında değil, istemci tarafında çalışabilir ve metin özetlemesini basit bir kod dizisiyle web sitelerinin, mobil uygulamaların bir parçası haline getirebilir, bir in-page özelliğe dönüştürebilir.

Öte yandan algoritmada yapılacak geliştirmelerle, ön-işlem sürecine eklenecek kural ve kriterlerle çıktıların daha yüksek kalitede olması sağlanabilir.

4.2. Yorumlar

Metin özetleme çalışmaları için, bir NLP kütüphanesini, istatistiki bir matematiksel yöntemle değiştirmek çıktı olarak elde edilen özetleri direkt olarak etkilyecek bir karardır ve amaç kaliteli (altın özetlere en yakın) özetleri çıkarmak ise tercih edilecek bir yöntem değildir. Yine de bu çalışma ile ulaşılan sonuçlar değerlendirildiğinde, eğer metin özetleme çalışması, sistemin merkezinde değil ise, ek hizmet veya bir ek özellik olarak projede yer alıyorsa, değerlendirilebilir bir alternatiftir.

Bu tez çalışmasında geliştirilen yöntem bir arama motoru oluşturmak için uygun olmasa da, bir site içerisinde kullanıcılara sunulacak ek bir hizmet olarak oldukça yeterli ve uygundur.

Ayrıca bu yöntem, hızlı ve kolay uygulanabilir olması sayesinde metin özetleme çalışmalarında özet çıkarma görevini üstlenmese bile, bir ön eleme-işleme süreci olarak kullanılabilir. Çok daha yüksek sistem kaynağı ve zaman gerektirecek işlemlerin

öncesinde, kaynak ve zaman tasarrufu sağlamak adına ihtiyaçlara göre şekillendirilerek sürece entegre edilebilir.

BÖLÜM 5. SONUÇ

Bu tez çalışmasında, harhangi bir NLP kütüphanesine ihtiyaç duymadan, Türkçe'nin sondan eklemeli bir dil olmasının avantajını, matematiksel olarak uygulamaya çalışan bir yöntem geliştirilmiş ve böylece metin özetleme çalışmalarında kolay kullanılır, hızlı ve her platformda ve dilde çalışabilir bir alternatif yaratılmaya çalışılmıştır. Geliştirilen yöntem ROUGE-N Recall metriği kullanılarak, hem Zemberek NLP kütüphanesi sonuçlarıyla, hem de benzeri diğer bir istatistiki yöntemle kıyaslanmıştır.

Çıkan sonuçlar analiz edildiğinde, beklendiği gibi Zemberek NLP'nin bu üç yöntem içerisinde en başarılı sonuçları ürettiği görülmüştür. Diğer yandan, bu çalışmada geliştirilen yöntemin de kabul edilebilir sınırlar içerisinde sonuçlar ürettiği ve geliştirmeye açık olduğu farkedilmiştir.

Bu yöntem mevcut hali ile, başarı beklentisi yüksek olmayan uygulamalarda kullanılabilir sonuçlar üretmiştir. Üzerinde çalışılmaya devam edilirse başarı oranının daha da yükseleğini görülmüştür. Yüksek beklentili metin özetleme çalışmaları için ise, tek başına özetleme yükünü üstlenemese bile, yardımcı bir araç olarak sistemlere entegre edilebilir olduğunu ispatlamıştır.

Kullanılan veri seti ve gerçekleştirilen işlemler ışığında, geliştirilen bu yöntem ile oldukça önemli sonuçlara ulaşılmıştır. Türkçe'nin dil bilgisi ve kelime yapısı göz önüne alındığında, bu yapıyı metin özetleme çalışmaları için matematiksel bir avantaja dönüştürmenin mümkün olduğu görülmüştür. Üzerinde çalışılan veri seti detaylıca incelendiğinde, özellikle bir başkasına ait tırnak içine alınmış pasajların metin özetleme çalışmalarının aşması gereken önemli bir sorun olduğu tespit edilmiştir. Bu tespit bile, yapılan bu çalışmanın gerekliliğini ortaya çıkarmaktadır. Karmaşık haber metinlerinde, ana metnin özetlenmesi işlemi başka bir kütüphane/kod tarafından yapılırken, bu yöntem ile, metin içi pasaj özetlemelerinin yapılmasının ve ana metin özetleyicisinin işini kolaylaştırmanın mümkün olduğu sonucuna ulaşılmıştır.

Daha da önemlisi bu çalışma sonucunda elde sonuçlar kabul edilebilir sınırlar içerisinde kalmış ve muadili diğer tekniklere nazaran tercih edilebilir bir yöntem olduğunu göstermiştir. Ayrıca geliştirilen yöntemin geliştirmeye açık olduğu da ortaya çıkmış ve bu geliştirmeler yapıldığı takdirde metin özetleme çalışmalarına katkı sağlayabileceği anlaşılmıştır.

Bu çalışma internet tabanlı haber metinleri ile gerçekleştirilmiştir. Seçilen veri seti çalışmanın sonuçlarında doğrudan etki eden bir faktördür, bu yüzden geliştirilen yöntem hakkında daha detaylı analizler yapabilmek için, farklı içerik şekillerine de uyarlanabilir. Haber metinleri yerine çok daha kısa olan ve daha az karmaşık olan sosyal medya paylaşımlarına, ya da forumlara, site kullanıcı yorumlarına uyarlanabilir.

Yapılacak geliştirmeler ve sektörel eklentiler ile, bu yöntem ile farklı amaçlara yönelik veri madenciliği yapılabilir. Sadece özet çıkarmak için değil, benzerlik ilişkisi kurmak için kullanılabilir. Bir diğer yandan yöntem üzerinde yapılacak geliştirmeler ve eklenecek kural kütüphaneleri ile başarı oranı artırılabilir. Daha da önemli IoT cihazlara kolaylıkla aktarılabilir ve farklı cihazlarda farklı amaçlar için kullanılabilir.

KAYNAKÇA

[1] TÜİK, "Türkiye Hanehalkı Bilişim Teknolojileri Kullanım İstatistiği," Ankara, 2019.

[2] Ö. E. Gündoğdu and N. Duru, "Türkçe Metin Özetlemede Kullanılan Yöntemler," in *18. Akademik Bilişim Konferansı - AB'16*, Aydın, 2016.

[3] N. Munot and S. S. Govilkar, "Comparative Study of Text Summarization Methods," *International Journal of Computer Applications*, vol. 102, no. 12, pp. 33-37, 2014.

[4] R. A. Garcia-Hernandez et al., "Text Summarization by Sentence Extraction Using Unsupervised Learning," vol. 5317, pp. 133-143, 2008.

[5] J. Bael, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey," *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305-338, 2016.

[6] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM Journal of Research and Development*, vol. 1, no. 4, pp. 309-317, 1957.

[7] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11-21, 1972.

[8] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *Journal of Documentation*, vol. 6, no. 5, pp. 503-520, 1972.

[9] Muhaz, URL: http://muhaz.org/otomatik-metin-ozetlemede-kullanlan-ve-one-ckan-yontemler.html (Erişim Zamanı; Eylül 2019)

[10] E. Uzundere, E. Dedja, and M. F. Amasyalı, "Türkçe Haber Metinleri İçin Otomatik Özetleme," in *Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu*, Isparta, 2008.

[11] A. Güran, S. N. Arslan, E. Kılıç, and B. Diri, "Sentence selection methods for text summarization," in *22nd Signal Processing and Communications Applications Conference (SIU)*, Trabzon, 2014.

[12] M. Çakır and E. Çelebi, "Kapsama katsayisi tabanli kümeleme ile belge özetleme," in *IEEE 19th Signal Processing and Communications Applications Conference (SIU)*, Antalya, 2011.

[13] M. V. Sami and B. Diri, "Web Tabanlı Otomatik Özet Çıkarma Sistemi," in *ASYU 2010 - Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu*, Kayseri, Kapadokya, 2010.

[14] Cambridge Dictionary, URL: https://dictionary.cambridge.org/dictionary/english/summary (Erişim Zamanı; Eylül 2019)

[15] B. Mirkin, *Core Concepts in Data Analysis: Summarization, Correlation and Visualization*.: Springer, 2011.

[16] V. Gupta and G. S. Lehal, "A Survey of Text Summarization Extractive Techniques," *Journal of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, pp. 258-268, 2010.

[17] R. D. Radev, E. Hovy, and K. McKeown, "Introduction to the Special Issue on Summarization," *Computational Linguistics*, vol. 28, no. 4, pp. 399-408, 2002.

[18] I. Mani and M. T. Maybury, *Advances in Automatic Text Summarization*.: MIT Press Cambridge, 1999.

[19] S. R. Patil and S. M. Mahajan, "Optimized Summarization of Research Papers as an Aid for Research Scholars Using Data Mining Techniques," in *International Conference on Radar, Communication and Computing (ICRCC)*, Tiruvannamalai, 2012.

[20] E. D. Liddy, *Natural Language Processing*.: In Encyclopedia of Library and Information Science, 2001.

[21] E. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural Language Processing: State of The Art, Current Trends and Challenges," *ArXiv*, vol. abs/1708.05148, 2017.

[22] E. Reiter and R. Dale, "Building Applied Natural Language Generation Systems," *Natural Language Engineering*, vol. 3, 2002.

[23] Dev.to, URL: https://dev.to/adalycoder/top-3-natural-language-processing-

libraries-hf9 (Erişim Zamanı; Eylül 2019)

[24] E. Loper and S. Bird, "*NLTK: The Natural Language Toolkit*," arXiv, 2002.

[25] C. Maning et al., "Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations," *Association for Computational Linguistics*, pp. 55-60, 2014.

[26] S. Loria, *TextBlob Documentation*.: TextBlob, 2018.

[27] R. Rehurek and P. Sojka, *Gensim - Statistical Semantics in Python*.: EuroScipy, 2011.

[28] B. Sirinivasa-Desikan, *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*.: Packt Publishing, 2018.

[29] N. Shukla, *Machine Learning with TensorFlow*.: Manning, 2018.

[30] Zemberek NLP, URL: https://github.com/ahmetaa/zemberek-nlp (Erişim Zamanı; Eylül 2019)

[31] G. Eryiğit, "The Impact of Automatic Morphological Analysis & Disambiguation on Dependency Parsing of Turkish," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, İstanbul, 2012.

[32] G. Eryiğit, J. Nivre, and K. Oflazer, "Dependency Parsing of Turkish," *Computational Linguistics*, vol. 34, no. 3, pp. 357-389, 2008.

[33] Nûve, URL: http://hrzafer.com/nuve-ile-turkce-cumle-sonu-tespiti (Erişim Zamanı; Eylül 2019)

[34] A. A. Akın and M. D. Akın, "Zemberek, an open source NLP framework for Turkic Languages," *Structure*, vol. 10, pp. 1-5, 2007.

[35] Zemberek NLP, URL: http://zembereknlp.blogspot.com (Erişim Zamanı; Eylül 2019)

[36] A. A. Akın and C. Demir, "Smoothlm: A language model compression library," in *22nd Signal Processing and Communications Applications Conference (SIU)*, Trabzon, 2014.

[37] H. P. Edmundson, "New Methods in Automatic Abstracting," *Journal of The Association for Computing Machinery*, vol. 16, no. 2, pp. 264-285, 1969.

[38] K. S. Jones and J. Galliers, *Evaluating Natural Language Precessing Systems: An Analysis and Review*.: Springer, 1996, Lecture Notes in Artificial Intelligence 1083.

[39] I. Mani, *Summarization Evaluation: An Overview*.: NTCIR, 2001.

[40] J. Steinberger, J. Jezek, and K. Jezek, "Evaluation Measures for Text Summarization," *Computing and Informatics*, vol. 28, pp. 251-275, 2009.

[41] M. Hassel, Evaluation of Automatic Text Summarization, 2004, Lisans Tezi.

[42] E. Lloret, L. Plaza, and A. Aker, "The challenging task of summary evaluation: an overview," *Language Resources and Evaluation*, vol. 52, pp. 101-148, 2018.

[43] C. Lin, *Looking for a Few Good Metrics: ROUGE and its Evaluation*. Tokyo, Japonya: National Institute of Informatics, 2004.

[44] C. Lin, "Rouge: a package for automatic evaluation of summaries," *Association for Computational Linguistics*, pp. 25-26, 2004.

[45] Hakan.io, URL: https://hakan.io/makine-ogrenmesi-turkce-haber-metinleri-veri-seti/ (Erişim Zamanı; Eylül 2019)