
GAZİ ÜNİVERSİTESİ
Gazi
1926

HİBRİT ZEKİ SİSTEM İLE METİN ÖZETLEME

Begüm MUTLU

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

HAZİRAN 2020

Begüm MUTLU tarafından hazırlanan "HİBRİT ZEKİ SİSTEM İLE METİN ÖZETLEME" adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi Bilgisayar Mühendisliği Ana Bilim Dalında DOKTORA TEZİ olarak kabul edilmiştir.

Danışman: Prof. Dr. M. Ali AKCAYOL

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum. ……..……...

İkinci Danışman: Prof. Dr. Ebru SEZER

Bilgisayar Mühendisliği Ana Bilim Dalı, Hacettepe Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum. ...…….

Başkan: Prof. Dr. Hasan OĞUL

Bilgisayar Mühendisliği Ana Bilim Dalı, Başkent Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum. ...…….

Üye: Prof. Dr. O. Ayhan ERDEM

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum. ...…….

Üye: Prof. Dr. Şeref SAĞIROĞLU

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum. ...…….

Üye: Doç. Dr. Hacer KARACAN

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum. .....…….

Üye: Doç. Dr. Oktay YILDIZ

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum. ...…….

Tez Savunma Tarihi: 29/06/2020

Jüri tarafından kabul edilen bu çalışmanın Doktora Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

……..…….

Prof. Dr. Sena YAŞYERLİ

Fen Bilimleri Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

……...

Begüm MUTLU

29/06/2020

HİBRİT ZEKİ SİSTEM İLE METİN ÖZETLEME

(Doktora Tezi)

Begüm MUTLU

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Haziran 2020

ÖZET

Metin özetleme, kaynak belgenin ana içeriğini ve genel anlamını koruyarak metnin özet adı verilen daha kısa bir sürüm halinde yapılandırılmasıdır. Çıkarmalı metin özetlemede, cümleler önem düzeylerine göre derecelendirilmekte ve ardından özete dâhil edilmeye değer cümleler belirlenmektedir. Bu tez çalışmasında, çıkarmalı metin özetlemede, cümlelere önem derecesi atamada başlıca etken olan öznitelikler sözdizimsel ve anlamsal olmak üzere incelenmiş ve sözdizimsel ve anlamsal özniteliklerin birlikte kullanımını öneren yeni bir cümle derecelendirme yöntemi geliştirilmiştir. Geliştirilen bu yöntem, insanlar tarafından hazırlanmış hedef özetleri içeren yeni bir veri kümesi üzerinde ve literatürde mevcut olan bir karşılaştırma veri kümesi üzerinde deneysel çalışmalarla değerlendirilmiştir. Ayrıca bu veri kümeleri üzerinde bulanık çıkarsama sistemi, yapay sinir ağı ve bir gözetimsiz metin özetleme yönteminin gerçekleştirimi yapılmış ve ortaya çıkan model özetleri önerilen yöntemden elde edilen özetlerle karşılaştırılmıştır. Değerlendirmelerden elde edilen çıkarmalı özetlerin başarısı incelendiğinde, sözdizimsel ve anlamsal özniteliklerin birlikte kullanımının ayrı ayrı kullanıma göre, kaynak metnin karakteristik özelliklerini yansıtma ve kaynak metni temsil etme kriterleri açısından daha başarılı olduğu görülmüştür. Geliştirilen mimari ile literatürdeki otomatik metin özetleme alanındaki benzer çalışmalar deneysel olarak karşılaştırılmış ve geliştirilen mimarinin otomatik metin özetlemede literatürdeki çalışmalardan daha başarılı olduğu gösterilmiştir.

Bilim Kodu	:	92432
Anahtar Kelimeler	:	Metin Madenciliği, Makine Öğrenmesi, Otomatik Metin Özetleme
Sayfa Adedi	:	155
Danışman	:	Prof. Dr. M. Ali AKCAYOL
2. Danışman	:	Prof. Dr. Ebru SEZER

TEXT SUMMARIZATION BY HYBRID INTELLIGENT SYSTEM

(Ph. D. Thesis)

Begüm MUTLU

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

June 2020

ABSTRACT

Text summarization is generating a shorter version of a document while preserving its main content and general meaning. In extractive text summarization, the sentences are scored according to their level of importance, and then summary-worthy sentences are determined to be included in the summary. In this thesis study, the features, the main factors in sentence scoring during summarization, have been examined as syntactic and semantic approaches, and a new sentence scoring method has been developed that proposes the joint use of these syntactic and semantic features. This method was evaluated by experimental studies on a new dataset containing human-generated goal summaries, and a benchmark dataset available in the literature. Additionally, a fuzzy system, artificial neural network, and an unsupervised text summarization method were performed on these datasets, and the resulting model summaries were compared with the summaries obtained from the proposed method. Once the success of the resulting summaries obtained from the evaluations was examined, it was concluded that the joint use of syntactic and semantic features was more successful in reflecting the characteristics of the source document and representing it, compared to the individual use of these features. The proposed method has been empirically compared with similar studies in the field of automatic text summarization in the literature, and the results showed that it was more successful than the studies in the literature in automatic text summarization.

Science Code	:	92432
Key Words	:	Text Mining, Machine Learning, Automatic Text Summarization
Page Number	:	155
Supervisor	:	Prof. Dr. M. Ali AKCAYOL
Co Supervisor	:	Prof. Dr. Ebru SEZER

TEŞEKKÜR

Doktora sürecimin her aşamasında bana yol gösteren, kıymetli bilgi ve deneyimlerini aktaran, değerli görüşleri ve önerileriyle tez çalışmamın şekillenmesine ve geliştirilmesine büyük katkılar veren danışman hocam Sayın Prof. Dr. M. Ali AKCAYOL'a, bilgi birikimi ve tecrübesi ile lisansüstü öğrenim hayatımın tüm zorlu aşamalarında bana her açıdan destek olan, kıymetli görüşleri ve çok yönlü bakış açısı ile çalışmama yön veren ve bana her koşulda yanımda olduğunu hissettiren, kendisini tanımaktan ve öğrencisi olmaktan büyük onur ve mutluluk duyduğum eş danışman hocam ve fahri ailem Sayın Prof. Dr. Ebru SEZER'e, doktora öğrenimim süresince kıymetli fikir ve önerileri ile çalışmama değer katan tez ileme komitesi üyelerim Sayın Prof. Dr. O. Ayhan ERDEM ve Sayın Prof. Dr. Hasan OĞUL'a, doktor unvanı alarak akademik hayata adım atıyor olduğum bu dönemde gelecek çalışmalarım üzerine değerli görüşlerini benimle paylaşan tez savunma sınavı jüri üyelerim Sayın Prof. Dr. Şeref SAĞIROĞLU, Sayın Doç. Dr. Hacer KARACAN ve Sayın Doç. Dr. Oktay YILDIZ'a, çalışmamın derlem oluşturma aşamasında özveriyle çalışmış olan değerli okurlara, çalışmalarım sırasında manevi desteğini her zaman hissettiğim, kaygılarıma ortak olan sevgili arkadaşlarım Esra ŞAHİN, Emine YÜKSEL, İpek BAŞGÜN, Dr. Feyza YILDIRIM OKAY, Ebru AYDOĞAN DUMAN ve Meryem BÖCEK'e öğrenim hayatım boyunca beni yürekten destekleyen, tüm kaynaklarını çocuklarını topluma yararlı bireyler olarak yetiştirmek için harcayan, varlığımı ve sahip olduklarımı tümüyle borçlu olduğum sevgili annem A. Belma MUTLU ve babam İsmail H. MUTLU'ya, hayattaki koruyucu kalkanım olan ablam Sayın Uzm. Dr. Bengü MUTLU SARIÇİÇEK'e ve cennetteki yeğenim Ece Mutlu SARIÇİÇEK'e, hayatına dokunduğum ve hayatıma dokunan herkese beni dönüştürdükleri kişi için içtenlikle teşekkür ederim.

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	**Açıklamalar**
$\boldsymbol{A}$	Anma
$\boldsymbol{D}$	Duyarlık
$\boldsymbol{F}$	F-ölçüt
$\boldsymbol{S}$	Cümle
ç-	Çift yönlü
$\boldsymbol{d}$	Özyerleşik boyu
$\boldsymbol{n}$	Öznitelik adedi
t	Zaman bilgisi
$\boldsymbol{t}$**-**	Tek yönlü
$\boldsymbol{y}$	Çıkış parametresi
$\boldsymbol{x}$	Giriş parametresi
σ	Sigmoid aktivasyon fonksiyonu
Δ	Referans özetle model özeti arasındaki boy farkı

Kısaltmalar	**Açıklamalar**
ç-ÖSA	Çift Yönlü Özyineli Sinir Ağı
DUC	Doküman Anlama Konferansı
E2E	End-to-End Hierarchical Fuzzy Inference Solution
ESA	Evrişimli Sinir Ağı
GRU	Geçitli Tekrarlayan Birimler
LSTM	Uzun-Kısa Dönem Bellek
LSTM-NN	Uzun-Kısa Dönem Bellek-Sinir Ağı
ÖSA	Özyineli Sinir Ağı
ROC	Receiver Operating Characteristics
ROC-AUC	ROC eğrisinin altında kalan alan

Kısaltmalar	Açıklamalar
ROUGE	Özet Değerlendirmede Anma Odaklı Yardımcı Metot
SIGIR	Special Interest Group on Information Retrieval
TF	Terim Frekansı
TF-ISF	Terim Frekansı-Ters Cümle Frekansı
***t*-ÖSA**	Tek Yönlü Özyineli Sinir Ağı
YSA	Yapay Sinir Ağı

1. GİRİŞ

İnternette, her geçen gün hızla artmakta olan çok büyük boyutta veri bulunmaktadır. Verinin boyutu sadece hacmiyle değil, türü, kaynak çeşitliliği, içeriği, yapısı, kalitesi ve ifade çeşitliliği ile de ilgilidir. Günümüzdeki verinin boyutu göz önüne alındığında, büyük boyutlu veri içerisinden hedeflenen bilgileri manuel olarak algılamak ve çıkarmak birçok uygulama alanı için hem zaman hem de maliyet açısından uygulanabilir değildir. Bu yüzden büyük boyutlu veriden bilgi çıkarımı işlemleri bilgi erişim sistemleriyle otomatik olarak gerçekleştirilmeye çalışılmaktadır. Ancak, verideki her geçen gün devam eden hızlı artış, yüksek teknolojiye sahip makineler için bile bilgi erişiminin etkin bir şekilde ve kabul edilebilir süre içerisinde gerçekleştirilememesine neden olmaktadır. Bu nedenle, bilgi erişiminin süresini kısaltmak ve daha etkin hale getirmek için üzerinde çalışılan veride boyut indirgeme sıklıkla başvurulan bir yaklaşımdır. Ancak, söz konusu veri, metin dosyalarından oluşan ve gerek yapısı gerekse anlamsal özellikleri açısından bütünlüğü olan bilgi kümelerinden oluştuğunda, yapısal veri üzerindeki öznitelik indirgeme ve öznitelik seçme süreçleri metnin yapısal ve anlamsal bütünlüğünü olumsuz etkilemektedir. Bu nedenle, metin verisinde boyut indirgeme sürecinin insan tarafından da anlamlandırılabilir ve yorumlanabilir nitelikte olması gerekmektedir. Literatürde yapılan çalışmalarda, boyut indirgeme ve öznitelik seçmeden daha önemli olarak metnin yerine kullanılabilecek daha küçük boyutlu özetlerinin türetilmesi temel problem olarak ele alınmaktadır [1].

Metin özetleme, kaynak belgenin ana içeriğini ve genel anlamını olabildiğince koruyarak metnin özet adı verilen daha kısa bir sürüm halinde yeniden yapılandırılması sürecidir [1]. Günümüzde hızla artmakta olan veri göz önüne alındığında, metnin özetini elle çıkarmak, başka bir deyişle insan yapımı özetler üretmek hem çok zor ve hem de çok zaman gerektiren bir süreçtir. Ayrıca, metnin özetinin insan tarafından oluşturulabilmesi için ilgili alanda uzmanlığının olması da gerekli olabilir. Kaynak metnin boyutunun fazla olması halinde, özetin farklı formatlardaki kaynaklardan elde edilmesinin gerekli olması halinde, kaynaktan özellikle objektif/sübjektif bilgilerin çıkarılmasına ihtiyaç olması halinde veya hangi oranda küçültmeye gidilmesi gerektiğinin bilinmediği durumda metin özetleme çok daha zor bir hale gelmektedir.

Otomatik metin özetleme, metin formundaki büyük veriler için geliştirilmiş olan ve genellikle büyük boyutlu metnin içerdiği temel bilgiyi daha küçük boyuttaki bir metin içinde

otomatik olarak ifade etmeyi amaçlamaktadır. Otomatik metin özetleme çalışmalarının tümünde doğal dilin kurallarına uygun, kaynak metindeki sözdizimsel ve anlamsal bütünlüğü mümkün olduğunca koruyan özetlerin türetilmesi hedeflenmektedir. Çıktısı bu özelliklere sahip bir boyut indirgeme sürecinin yorumlanabilirlik niteliğine sahip olması gereklidir. Bu nedenle otomatik metin özetleme bilgi erişim teknikleri açısından bir boyut indirgeme yöntemi olarak düşünülmesinin yanı sıra, insan için de anlamlı, akla yatkın ve yorumlanabilir bir metin sadeleştirme şeklidir. Sonuç olarak, otomatik metin özetleme hangi amaca yönelik yapılıyor olursa olsun (insan için özet, bilgi erişim sistemleri için özet), o amaç için gerekli kaynakların (insan, zaman, hafıza vb.) daha etkin kullanımına olanak sağlamaktadır.

Literatürde otomatik metin özetleme yaklaşımları soyutlamalı ve çıkarmalı metin özetleme olmak üzere iki temel kategoride ele alınmaktadır [2]. Soyutlamalı metin özetlemede bir metnin içerisindeki temel kavramların ve kavramlar arasındaki anlamsal ilişkilerin çıkarılarak dilin kurallarına ve yapısına uygun yeni bir özet metnin üretilmesi amaçlanmaktadır. Bu tip metin özetleme yaklaşımı daha çok doğal dil işleme için kullanılan yöntemlerle ilişkilidir ve bu yönüyle üzerinde çalışılan doğal dilin kurallarına bağımlıdır. Bu yaklaşım, insanların bir metni okuyup, içerisindeki temel bilgiyi algılayarak özeti kendi cümleleri ile yeniden oluşturmalarına benzer olmakla birlikte; büyük boyutlu bir dokümanın tümüyle anlaşılıp özetinin bu şekilde çıkarılması henüz tam anlamıyla uygulanabilir bir yöntem değildir. Bunun yerine soyutlamalı özetleme, bir metin için başlık/anahtar söz/vurgu üretme [3–5], cümle sıkıştırma [6–8], cümle birleştirme [9], otomatik soru cevaplama [10, 11] gibi daha dar kapsamlı olarak uygulanmaktadır. Çıkarmalı özetlemede ise metni oluşturan bölümler için önem dereceleri atanmakta ve bu önem dereceleri dikkate alınarak bölümler özet metne doğrudan aktarılmaktadır. Metinden seçilen bölümlerin içeriğinde değişiklik yapılmadığından ortaya çıkan özetler doğal olarak dil yapısına uygun ve anlamlıdır. Bölümlere önem dereceleri atama, öneme göre bölüm seçme ve seçilen bölümleri özete dâhil etme süreçleri üzerinde çalışılan dilden bağımsızdır ve uzun metinler için uygulanabilirliği yüksektir.

Bu tez çalışmasının amacı, kaynak metnin yerini mümkün olduğunca tutacak otomatik metin özetlerini oluşturmaktır. Bu amaç doğrultusunda, çıkarmalı metin özetlemede cümle bölümlerinin önem derecelerinin belirlenmesinde kullanılan öznitelikler, önem derecelendirme yöntemleri ve metin özetlemedeki sıkıştırma oranının başarıya etkisi detaylı

bir şekilde incelenmiştir. Hem literatürde kullanılan veri kümeleri hem de bu tez çalışması kapsamında oluşturulan yeni veri kümesi kullanılarak geliştirilen metin özetleme yaklaşımı ile literatürdeki metin özetleme yaklaşımları yapılan deneysel çalışmalar ile detaylı bir şekilde karşılaştırılmıştır.

Genel amaçlı otomatik metin özetlemede temel alınan yaklaşım, orijinal metindeki önemli bilginin özet metinde mümkün olduğunca yer alması ve özet metnin orijinal metnin yerini içerdiği bilgi açısından mümkün olduğunca tutmasıdır. Bu durum insanlar tarafından çıkarılan özetlerin de başlıca hedeflerinden birisidir. Başka bir deyişle, otomatik metin özetleme yöntemleri insanlardaki okuma, anlama, ilişki kurma ve özet çıkarma sürecini taklit etmeye çalışmaktadır. Bu açıdan bakıldığında, otomatik metin özetlemenin amacı, bir metnin uzman kişi tarafından okunup elle özetlenmiş formuna mümkün olduğunca yakın özetleri üretmektir. Ancak, bu noktada insan davranışının özetlenen metnin özelliklerine ve özetleme niyetine göre farklılık göstereceği ve önem tanımının bu açılardan değişebileceği açıktır. Örneğin, bir haber makalesindeki cümlelerin önem tanımı ile bir elektronik postadaki cümlelerin önem tanımı farklı özniteliklere bağlıdır. Ayrıca, özetlemeyi yapan kişinin ortaya çıkacak metin özetinden beklentisi de önem tanımının değişmesine sebep olabilmektedir. Bir metin, içerdiği bilgiyi sıkıştırmak ve tüm bilgiyi daha küçük boyutta yeniden sunmak için özetleniyor olabileceği gibi; cümleler içerisinde belirli bir sorunun cevabı aranıyor da olabilir. Bu durumda, değişen özetleme amacına göre önem tanımının da değiştiği gibi önemi etkileyecek özellikler de farklılık gösterecektir.

Çıkarmalı metin özetlemede bir cümlenin öneminin tanımlanabilmesi için iki tip özellikten faydalanılmaktadır. Bunlar metni sözdizimsel öznitelikler üzerinden ifade etmek ve metni anlamsal öznitelikler üzerinden ifade etmektir. Sözdizimsel öznitelikler, metindeki cümlelerin önceden tanımlanmış çeşitli özniteliklerin metnin tümünden veya bir bölümünden elde edilmesiyle oluşturulur ve genel olarak özetlemede dikkate alınan belirli unsurlar üzerinden önem tanımı yapılmaya çalışılır. Öte yandan, anlamsal öznitelikler cümleyi oluşturan terimlerin anlamlarına ve birbirleriyle ilişkilerini göz önüne alır. Bu iki yaklaşım, literatürde otomatik metin özetleme çözümlerinin temelini oluşturmuş ve farklı özetleme yöntemlerinin geliştirilmesinde kullanılmıştır.

Sözdizimsel veya anlamsal özniteliklerle hem insan hem de otomatik özetleyiciler ile yapılan özetleme çalışmalarında, özetlenen dokümanın karakterine ve özetleme niyetine

uygun öznitelikler kullanılmaktadır [5, 11–15]. Bunun için otomatik özetleme yöntemleri sözdizimsel veya anlamsal yöntemlerden birisinin kullanımına yoğunlaşır ve amaca uygun özetler üretir. Ancak, genel bir otomatik metin özetlemede kaynak metnin türünden kaynaklanan yazım şekli, kullanılan kelimeler, metnin karakteristik özellikleri ve özetlemedeki amaç dikkate alınmaz. Sözdizimsel veya anlamsal özniteliklerin tekli kullanımında ise, özetlemede belirleyici olması muhtemel olan özelliklerden bir grubunun dışarda bırakılmış olması riski ortaya çıkmaktadır. Dolayısıyla, önerilen çözümün metin özetlemedeki amacı ve doküman türünden kaynaklanan farklılıkları dikkate alabilmesi için öğrenebilen yöntemlerle birlikte kullanılması gerekir. Bu tez çalışmasında, otomatik metin özetleme için sözdizimsel ve anlamsal özniteliklerin birlikte kullanımını sağlayan ve cümlelerin önem derecesini bu iki kanaldan elde ederek bütünleştiren bir çözüm mimarisi geliştirilmiştir.

Son yıllarda, hem soyutlamalı hem de çıkarmalı metin özetleme problemi için geliştirilen çözümlerin önemli bir kısmında, metinden elde edilen öznitelikler kullanılarak önemli cümlelerin öğrenilmesini amaçlayan makine öğrenmesi yöntemleri kullanılmıştır. Ancak, öğrenmeyi sağlamak amacıyla kullanılan veri kümelerinde, çıkarmalı metin özetlemenin uygulanabilirliğini azaltan bir takım eksiklikler söz konusudur. Bu eksiklikler yazılı belge ihtiyacı ve belgeye ait hedef özetlerin tanımlanmış olması gerekliliğidir. Hem bilgi çıkarımı hem de metin madenciliğinin birçok problemi için oluşturulmuş çok sayıda veri kümesi bulunmaktadır. Çıkarmalı metin özetlemenin amaçlandığı bir çalışmada kullanılacak veri kümesinin bazı temel gereksinimleri karşılaması gereklidir. Öncelikle, özetlemede kullanılacak veri kümesi belirli bir konuyla ilgili kapsamlı olarak hazırlanmış yazılı belgeleri içermelidir. Ancak, günümüzde kullanılan veri kümelerinin çoğu ürün veya sunulan servise ait yorum ve sosyal medya paylaşımı gibi yapısal olmayan kısa metinler kullanılarak oluşturulmuştur [16, 17]. Ayrıca, özetleme için kullanılacak veri kümesinin kaynak metinden elde edilen hedef özetleri içermesi gereklidir. Bu denetimli öğrenmeye dayanan metin özetleme yöntemleri için gerekli olmakla birlikte, denetimsiz öğrenmeye dayanan yöntemler için de elde edilen özetlerin başarısının değerlendirilebilmesi için gereklidir. Bu hedef özetler özetleme türü dikkate alınarak oluşturulmalı, soyutlamalı ve çıkarmalı özetleme çalışmaları için özelleştirilmiş olmalıdır. Ayrıca, bu metin özetleri insanlar tarafından kaynak metinlerin okunup anlaşılmasıyla üretilmeli ve otomatik modelin insan davranışını ne ölçüde taklit ettiği ölçülebilmelidir. Ancak, çok sayıda metin için insanlar tarafından soyutlamalı ve çıkarmalı özetler oluşturulması hem süre hem de maliyet açısından

bakıldığında oldukça zordur. Buna ek olarak, tek bir okuyucu tarafından çıkarılan özetler üzerinden bir veri kümesi oluşturmak, tutarlılık ve doğruluk açısından yeterli görülmeyebilir. Aslında, çok sayıda insan tarafından oluşturulan özetler için de elde edilen özetlerin doğrulanması ve geçerlenmesi şarttır. Tüm bu zorluklardan dolayı, literatürde çok az sayıda özetleme veri kümesi bulunmaktadır ve metin özetleme çalışmalarının tamamına yakını özetleme başarılarını bu veri kümelerini kullanarak değerlendirmektedir.

Bu amaçla, bu tez çalışmasında otomatik metin özetleme problemi için özelleştirilmiş çok sayıda akademik yayın ve bu yayınlara ait soyutlamalı ve çıkarmalı metin özetlerini barındıran yeni bir veri kümesi (SIGIR2018) üretilmiştir. Bunun için 2018 yılında gerçekleşen 41. Uluslararası ACM (Association for Computing Machinery, Hesaplama Makineleri Birliği) SIGIR (Special Interest Group on Information Retrieval, Bilgi Erişimi Özel İlgi Grubu) 2018 yılında yapılan Bilgi Erişiminde Araştırma ve Geliştirme Konferansındaki (Conference on Research and Development in Information Retrieval 2018) açık erişimli 125 bildirinin tam metni kullanılmıştır. İlk olarak, hedef soyutlamalı özetler yayının kendi yazarları tarafından hazırlanmış olan özet bölümlerinden elde edilmiştir. Hedef çıkarmalı metin özetleri ise üç gönüllü okurun metinlerde özete dâhil etmeye değer buldukları cümleleri seçmesi ile oluşturulmuştur. Çalışmadaki okurların biri bilgisayar mühendisliği lisans derecesine sahiptir ve diğer ikisi lisans eğitimine devam eden öğrencilerdir. Okurlar, 125 bildirinin tamamının giriş bölümlerini dikkatle okumuş ve birbirlerinin önemli olarak seçtikleri cümlelerden haberdar olmadan kendi çıkarmalı özetlerini oluşturmuşlardır. Her okurun seçtiği cümleler o okurun aday cümle kümesi olarak değerlendirilmiştir. Bu sürecin sonunda tüm dokümanlar için üç aday cümle kümesi meydana gelmiştir. Ardından, her bir doküman için aday cümle kümelerinin kesişimi ve birleşimi üzerinden iki hedef çıkarmalı özet elde edilmiştir.

Hedef çıkarmalı metin özetlerinin geliştirilmesinde okurlar tarafından gerçekleştirilen cümle seçme sürecinin doğrulanması ve geçerlenmesi için elde edilen özetler, rastgele üretilmiş çıkarmalı metin özetleri ile karşılaştırılmıştır. Buna ek olarak, farklı metin özetleme yaklaşımlarından elde edilen çıkarmalı özetlerin de okurlar tarafından üretilen özetlerle karşılaştırılması yapılmıştır. Yapılan karşılaştırma sonucunda SIGIR2018 veri kümesi için oluşturulmuş olan çıkarmalı özetlerin rastgele olmaktan uzak ve mevcut yaklaşımlardan belirgin bir oranda daha başarılı olduğu görülmüştür.

Tez çalışması kapsamında önerilmiş olan SIGIR2018 metin özetleme veri kümesine, çıkarmalı metin özetleme problemi için bulanık kural tabanlı sistem, yapay sinir ağı, derin öğrenme yaklaşımı olmak üzere üç denetimli ve cümlelerin birbirleri ile ilişkileri üzerinden gruplanmasına dayanan bir denetimsiz öğrenme yaklaşımı geliştirilerek uygulanmıştır. Bulanık sistemler ve yapay sinir ağı sözdizimsel öznitelikler üzerinden çözüm geliştirirken, uygulanan derin öğrenme yaklaşımı ile denetimsiz öğrenme çözümünde anlamsal ilişkiler üzerinden çıkarmalı metin özetleme yapılmıştır. Bu dört modelden elde edilmiş olan özet değerlendirme skorları elde edilmiş ve kendi içlerinde karşılaştırılmıştır. Ayrıca bu tez çalışmasında çıkarmalı metin özetleme problemi için özete dâhil edilmeye değer cümlelerin seçimini sağlayan özyineli sinir ağı ile yapay sinir ağının birlikte kullanımıyla geliştirilen LSTM-NN (Long-Short Term Memory-Neural Network) mimarisi sunulmuştur. Bu mimari üzerinde öncelikle kelime özyerleşiklerine dayalı anlamsal özniteliklerin tek başına kullanılmasının ve ardından yine bu çalışmada önerilmiş olan geniş bir sözdizimsel öznitelik listesinin tek başına kullanılmasının metin özetleme başarısına etkisi incelenmiştir. Bu iki yaklaşımla elde edilen performans değerleri literatürdeki mevcut çalışmalarla aynı oranda başarı düzeyine sahiptir.

Özniteliklerin tekil kullanımının SIGIR2018 üzerinde gerçekleştirilmesinin ardından, LSTM-NN sözdizimsel ve anlamsal özniteliklerin birlikte kullanımına imkân verecek hibrit bir yaklaşımla genişletilmiş ve ortaya çıkan çıkarmalı özetlerin performansı gerek özet değerlendirme ölçütleri gerekse standart sınıflandırma performans değerlendirme ölçütleri kullanılarak ölçülmüştür. Performans ölçümleri sonucunda, önerilen hibrit öznitelik kullanımının özniteliklerin tekil kullanımına göre özetleme performansını yüksek oranda artırdığı görülmüştür.

Özetle, bu tez çalışmasında üzerinde çalışılmış olan araştırma problemleri ve bu araştırma problemleri için yapılan faaliyetler ve edinilen temel bulgular aşağıda listelenmiştir:

Çıkarmalı metin özetlemede kullanılan veri kümeleri ve bu veri kümelerinin özellikleri araştırılmıştır. Metin özetleme problemi için SIGIR2018 ismiyle yeni bir veri kümesi oluşturulmuştur.

Çıkarmalı otomatik metin özetlemede cümle önem derecesinin tanımlanmasında kullanılması gereken özellikler incelenmiştir. Cümle önem derecesinin belirlenmesi için

kullanılan öznitelikler sözdizimsel ve anlamsal öznitelikler olmak üzere iki tür olarak incelenmiştir. Çıkarmalı metin özetleme problemi için kapsamlı bir sözdizimsel öznitelik listesi oluşturulmuş ve bu sözdizimsel özniteliklerin özetleme başarısına etkisi incelenmiştir. Çıkarmalı metin özetleme probleminde anlamsal ve sözdizimsel özniteliklerin kullanımının deneysel karşılaştırması yapılmıştır.

Çıkarmalı metin özetleme yöntem araştırması yapılmıştır. Sözdizimsel veya anlamsal öznitelikler üzerinden çözüm sunan makine öğrenmesi yöntemlerinin çıkarmalı metin özetleme problemi üzerinde karşılaştırmalı analizleri sunulmuştur. Metin özetleme probleminin yalnızca anlamsal argümanlarla gerçekleştirilmesinin yetersizliği iki farklı veri kümesindeki deneylerle gösterilmiştir. Öte yandan, bulanık sistemlerin kullanılması ile yapılan deneylerde bazı sözdizimsel özniteliklerin metin özetleme başarısına olumlu etkisinin olduğu gözlenmiştir. Bu bulgular çerçevesinde çıkarmalı metin özetleme problemine anlamsal ve sözdizimsel özniteliklerin birlikte kullanımını temel alan hibrit bir çözüm mimarisi geliştirilmiştir. Yapılan deneysel çalışmalarda bu özniteliklerin birlikte kullanımının tekli kullanımına olan üstünlükleri gösterilmiştir.

2. OTOMATİK METİN ÖZETLEME ALAN BİLGİSİ

2.1. Otomatik Metin Özetleme

Metin özetleme, bir belgenin içerisinde barındırdığı esas ve önemli bilgileri mümkün olduğunca içerecek daha kısa bir sürümünü üretme sürecidir. Okur açısından düşünüldüğünde, insanlar bir konudaki çok sayıda belgenin içeriklerinin tümünü okumak ve bunların içinden kendilerine gerekli bilgiyi çıkarmak zorunda kalmadan, mevcut belgelerin bütünleştirilmiş ve tek bir özete indirgenmiş sürümlerinden kendileri için gerekli olan bilgiyi kolaylıkla çıkarabilirler. Öte yandan, bilgi erişim sistemleri açısından düşünüldüğünde, metin özetleme işlemi, etkili bir şekilde uygulanırsa, iyi bir boyut küçültme aracıdır. Bu durum bilgi erişim sistemlerinin doğruluğunu arttırırken bilgi erişim maliyetini gerek hafıza gereksinimi gerekse hesaplama karmaşıklığı açısından düşürmektedir.

Literatürdeki otomatik metin özetleme yaklaşımları çeşitli açılardan kategorize edilmiş ve bu kategoriler Çizelge 2.1'de sunulmuştur.

Çizelge 2.1. Metin özetleme kategorizasyonu

Kıstas	Kategori Adı
Çözüm stratejisi	Soyutlamalı metin özetleme
	Çıkarmalı metin özetleme
Kaynak çeşitliliği	Tek belge özetleme
	Çoklu belge özetleme
Özetin hedefi	Genel metin özetleme
	Sorgu tabanlı metin özetleme
Özetin kapsamı	Gösterge niteliğinde metin özetleme
	Bilgilendirici metin özetleme
Özetin dili	Tek dil üzerinden metin özetleme
	Çok dilli metin özetleme

Otomatik metin özetleme çalışmaların tümü göz önüne alındığında çözüm stratejisi üzerine yapılan çalışmaların yoğunlaştığı görülmektedir. Çalışmaların tümü metin özetleme problemine soyutlamalı ve çıkarmalı metin özetleme olmak üzere iki faklı strateji ile yaklaşmakta, çözüm yöntemleri de bu iki stratejiye göre çeşitlilik göstermektedir. Bu

nedenle çözüm stratejisine bağlı metin özetleme yaklaşımları Bölüm 2.1.1 ve Bölüm 2.1.2'de ayrıntılı bir şekilde ele alınmıştır. Diğer bir yoğun çalışma alanı ise kaynak çeşitliliği üzerinedir. Burada özetleme tek belge üzerinden gerçekleştirilebileceği gibi aynı konudaki çok sayıda belgedeki bilginin öncelikle birleştirilmesi ve sonra özetlenmesi şeklinde de yapılabilmektedir. Çoklu belge özetlemede metin tipinde çok sayıda belge işlenebileceği gibi farklı formatlardaki belgeler üzerinden de metin formunda özetler elde edilebilmektedir.

Özetin hedefi açısından metindeki ana temayı genel hatlarıyla özete yansıtmaya çalışmak veya kaynak metindeki bir veya birkaç sorgu alanında önemli kabul edilen bölümleri özet içerisinde bulundurmak yaygın kullanılan iki yaklaşımdır. Genel metin özetleri ise kaynak dokümanda vurgulanan bilginin genel hatlarını hiçbir özel alana alana odaklanmadan sunmayı amaçlar. Öte yandan, sorgu tabanlı metin özetlemede belirli sorgu veya temaların çevresindeki bilginin metin içerisinden çıkarılmasına odaklanılır ve alan bağımlıdır [14].

Özetin kapsamı göz önüne alındığında ise gösterge niteliğinde özetler veya bilgilendirici özetler meydana gelebilmektedir. Gösterge niteliğindeki bir özette kaynak metnin içeriğinin tümünün özette yer alması beklenmez, ancak metnin ana konusunun ve alanının yer alması sağlanır. Bilgilendirici bir özette ise metnin ana hatları mümkün olduğunca, ayrıntılı bilgiler ise özet boyutundaki kısıtların elverdiği ölçüde yer alır. Bilgilendirici bir özet okunduğunda, okuyucu metnin hangi konuda olduğunun yanı sıra kapsamına ve metinde vurgulanmış bilgiye de sahip olur.

Metin özetleme problemi için de üzerinde çalışılan dil çok önemlidir. Soyutlamalı metin özetlemenin dil bağımlılığı çıkarmalı metin özetlemeye göre çok daha fazladır. Çünkü soyutlamalı metin özetlemede üzerinde çalışılan dile uygun yeni metin elde edilmeye çalışılırken, çıkarmalı metin özetlemede dili önemli kılan tek süreç metin üzerinde özetleme sürecinden önce uygulanan önişleme aşamasıdır. Bu açıdan, metin özetleme alanında tam olarak dil bağımsızlığından söz edilemez. Bu nedenle, literatürdeki bazı çalışmalar çoklu dil desteğini sunan metin özetleme süreçleri tanımlamışlardır [18, 19]. Ancak bu çalışmalarda sunulan çözüm ya belirli sayıdaki dil için ayrı ayrı model geliştirmek [18] ya da kaynak metni ilk olarak sabit bir dile veya standarda çevirmek, ardından bu dil üzerinde metin özetleme sürecini gerçekleştirmek, ve ortaya çıkan özeti istenen dile dönüştürerek sunmaktır [19].

2.1.1. Soyutlamalı metin özetleme

İnsanlar, bir metni okuduklarında ve bütünüyle anladıklarında o metindeki temel bölümleri ve bu bölümler arasındaki neden-sonuç ilişkilerini çıkarabilirler. Özet çıkarmak için bu bölümleri ve ilişkileri kullanarak yeni bir metin meydana getirirler. Soyutlamalı metin özetleme, bir metnin insan tarafından elle üretilmiş özetine benzer şekilde, doğal görünümlü özetler çıkarmayı hedefler [6, 20–24]. Farklı şekillerde uygulamaları mevcut olsa da soyutlamalı özetlemede ilk olarak kaynak dokümandaki temel kavramlar çıkarılır ve anlamsal ilişkiler tanımlanır. Ardından bu kavramlar ilgili doğal dilin gramer kuralları ve kısıtlamalarına uygun olarak özet metinde yeniden sunulur [22, 24, 25].

Soyutlamalı özetleme, insanın elle yaptığı özet çıkarma sürecine yakınlığı ve doğallığı nedeniyle araştırmacılar tarafından büyük ilgi görmektedir [22]. Metni oluşturan bölümlerin anlamsal ilişkilerinin çıkarılması, çıkarmalı metin özetlemede karşılaşılan anafor çözümleme gereksiniminin doğal olarak üstesinden gelinmesini sağlamaktadır. Ancak soyutlamalı özetlemede, çıkarmalı özetlemeye göre önemli bazı güçlükler bulunmaktadır.

Soyutlamalı metin özetleme yöntemleri doğal dile bağımlıdır [2]. Kaynak dokümandan özet metin üretilirken üzerinde çalışılan doğal dilin kurallarına ve kısıtlarına uymak zorunludur.

Soyutlamalı özetleme üzerine yapılmış çalışmalar genel olarak metin bölümlerinin daha kısa biçimde ifade edilmesine yöneliktir. Bu çalışmalar literatürde metin bölümü sıkıştırma ismiyle ele alınır ve bu çalışmalarda genel amaç cümleleri gereksiz ifadelerden arındırarak kısaltmaktır [6]. Bu özelliği ile soyutlamalı özetleme genellikle kısa metinlerin özetlenmesinde kullanılmaktadır [2]. Büyük boyuttaki metinler için temel kavramların çıkarılması, ilişkilendirilmesi ve doğru biçimde yeniden bir metin olarak düzenlenmesi oldukça zordur. Bu yüzden büyük boyuttaki metinlerin özetlenmesinde çıkarmalı özetlemenin kullanımı daha yaygındır [2].

Soyutlamalı özetleme yöntemlerinin yetenekleri, dili temsil yeteneği ve yeni yapılar oluşturma kabiliyetleri ile sınırlıdır [2]. Başka bir deyişle, özetleyici önceden tanımlanan yapısal sınırlamaların dışına çıkamaz. Kısıtlı alanlarda, genel amaçlı olmayan, sorguya dayalı özetlerin çıkarılmasında bu yapıları tasarlamak mümkün olabilir, ancak genel amaçlı bir çözüm kapsamlı bir anlambilim analizini gerektirmektedir. Gerçek hayatta doğal dilde

yazılmış metni tümüyle idrak edebilen ve özet çıkarabilen sistemler, günümüz teknolojisinin sahip olduğu yeteneklerle başarılabilecek olgunlukta değildir.

2.1.2. Çıkarmalı metin özetleme

Metin özetlerindeki mevcut çalışmaların büyük kısmı, metin içerisindeki bölümlerin önem derecelerini hesaplamaya ve bu bölümler içerisinden amaçlanan sıkıştırma oranına göre en önemlilerini özete dâhil etmeye dayanır [25]. Bu metin bölümleri, cümleler, kelime öbekleri, paragraflar veya başka herhangi bir metin parçası olabilir. Literatürde bu alanda en sık kullanılan yaklaşım cümleler üzerinden önem derecesinin hesaplanması ve özete önemli cümlelerin yerleştirilmesidir.

Çıkarmalı metin özetleme, soyutlamalı metin özetlemeye kıyasla uygulaması daha kolay bir yaklaşımdır. Metin içindeki bölümlerin orijinal yapısına müdahale etmediğinden doğal dile uygun özetler sunarken, bu bölümler hali hazırda insan tarafından yazılmış olduğundan anlamsal değişimler ortaya çıkmaz. Metnin sözdizimsel ve anlamsal bütünlüğünü korumak daha kolaydır. Bununla birlikte, çıkarmalı metin özetlemede dikkate alınması gereken bazı zorluklar da mevcuttur.

Çıkarmalı metin özetlemede, özetleme yöntemleri genellikle uzun metin bölümlerini önemli metin olarak etiketlemeye daha çok yatkındır [2]. Çünkü bu bölümler kendi içerisinde önemli alt parçalar bulundurma potansiyeline sahiptir. Bu durumda elde edilen özet, çok uzun cümlelerden ve paragraflardan oluşacaktır. Bu özet üzerinde aslında önemli olmayan alt parçaların ayrıştırılması gerekebilir.

Eğer önemli bölüm ve parçalar doküman içerisinde homojen dağıtılmış durumdaysa bunların tümünü içerecek özetin oluşturulması güçtür [2]. Bunun yanında aynı anlamı içeren çakışmalı ifadeler de problem teşkil edebilmektedir [2]. Özet metinde aynı anlamı içeren çok sayıda cümlenin yer alması yerine farklı anlamı içeren cümlelerin yer alması tercih edilir.

Çıkarmalı metin özetlemenin bir diğer sorunu ise anafor çözümleme ihtiyacıdır [2]. Metin içerisinde öncelikle anaforların atıfta bulundukları ifadelerle yer değiştirmesi ve ardından oluşan cümleler üzerinden önem derecelerinin hesaplanması daha tutarlı sonuçlar

vermektedir. Anafor çözümlemeye yönelik otomatik metin özetleme çalışmaları literatürde çok az sayıdadır [26].

2.2. Çıkarmalı Metin Özetlemede Öznitelikler

Metin özetleme probleminde özete alınmaya değer cümlelerin seçilmesinde en etkili bileşenler öğrenilen veri ile özetleme yöntemidir. Verinin içerdiği bilginin dengeli dağılımı, gürültünün azlığı ve kaliteli oluşu sonucu doğrudan etkilemektedir. Ancak bu alanda daha önemli olan verinin modele girdi olarak sunulmak üzere nasıl temsil edildiğidir. Metin özetleme çalışmalarında metni temsil etmek için sözdizimsel öznitelikler ile temsil ve anlamsal öznitelikler ile temsil olarak iki tür yöntem yaygın kullanılmaktadır. Sözdizimsel öznitelikler, cümlelerin önem derecelerini belirlemede önceden tanımlanmış öznitellikleri ifade eder. Anlamsal öznitelikler ise, metindeki cümlelerin sahip oldukları kelimelerin anlamsal ilişkilerinin belirlenmesinde kullanılır. Bu bölümün devam eden kısmında, metin özetleme problemi için metni temsil etmede kullanılan bu iki tür yöntem ayrıntılı bir şekilde sunulmuştur.

2.2.1. Sözdizimsel öznitelikler

Çıkarmalı metin özetlemede sözdizimsel öznitelikler belirli tanımlamalara göre elle üretilmiş ölçütler ve metotlar üzerinden metni temsil etmeye yarayan özelliklerdir. Bu özniteliklerin tanımlanmasında veya hesaplanmasında katı sınırlamalar olmadığından literatürdeki çalışmalarda farklı öznitelikler farklı şekillerde tanımlanmıştır [1]. Bu çalışmalarda bir cümlenin öneminin nelere bağlı olabileceği ile ilgili araştırmalar yapılmış ve çok sayıda metin özetleme özniteliği tanımlanmıştır.

Bu tez çalışmasında literatürde metin özetleme çalışmalarında kullanılagelmiş olan sözdizimsel özniteliklerin tümünün değerlendirlmesi ve bu öznitelikler üzerinden bir sözdizimsel öznitelik listesinin oluşturulması hedeflenmiştir. Bu nedenle, bu bölümün alt başlıklarında, tez çalışması kapsamında üzerinde çalışılmış olan sözdizimsel özniteliklerin tanımına ve çok sayıdaki hesaplama şekillerine yer verilmiştir. Burada sunulmuş olan tüm özniteliklerin ve öznitelik değeri hesaplama çeşitlerinin birleştirilmesi ile kapsamlı bir sözdizimsel öznitelik listesi oluşturulmuştur.

Terim Frekansı

Metin madenciliğinde, doğal dil işlemede ve bilgi erişim teknolojilerinde sıklıkla kullanılan terim sıklığı (Term Frequency-TF), bir terimin bir dokümanda hangi sıklıkta yer aldığını gösterir. Her belgenin uzunluğu farklı olduğundan, bir terimin uzun belgelerde daha kısa olanlara göre daha sık yer alma olasılığı yüksektir. Bu nedenle terim sıklığı, normalleştirme yöntemi olarak genellikle belgedeki toplam terim sayısına bölünür ve TF değeri elde edilir.

Terim frekansının metin özetleme probleminde öznitelik olarak kullanılması için her cümleye terim frekansı öznitelik değerinin $F_{TF}(S)$ atanması gerekecektir. Bunun için her terime ait TF değerinin cümlenin TF değerini yansıtacak şekilde hesaplanması gerekir. Bu hesaplamada literatürde Eş. 2.1 ve Eş 2.2'den faydalanılmıştır.

$$F_{TF}^{1}(S) = \sum_{w \in S} \mathrm{TF}(\mathrm{w}) \tag{2.1}$$

$$F_{TF}^{2}(S) = \frac{\sum_{w \in S} \mathrm{TF}(\mathrm{w})}{N_S} \tag{2.2}$$

Eş. 2.1'de cümle içerisindeki tüm terimlerin TF değerlerinin toplamı hesaplanmaktadır [27, 28]. Eş. 2.2'de ise hesaplanan bu toplam değer ilgili cümledeki tekil kelime sayısına (N_S) bölünmektedir [29].

Terim Frekansı-Ters Cümle Frekansı

Metin özetleme problemindeki terim frekansı-ters cümle frekansı (Term Frequency-Inverse Sentence Frequency-TF-ISF), metin madenciliği, doğal dil işleme ve bilgi erişim uygulamalarında yaygın kullanılan terim frekansı-ters doküman frekansı (Term Frequency-Inverse Document Frequency-TF-IDF) ölçütünün doküman yerine cümle tabanlı ele alınması ile geliştirilmiş bir özniteliktir. TF-IDF, yalnızca TF değerinin kullanımının tüm dokümanlarda görülen ortak terimlere daha yüksek değer atamadan kaynaklanan olumsuzluğu ortadan kaldırmak için geliştirilmiş bir çözümdür. Buna göre toplam doküman sayısının, söz konusu terimi içeren doküman sayısına bölümünün logaritması ile IDF değeri hesaplanır. Tüm dokümanlarda görülen terimlerin IDF değeri sıfırdır. İlgili terim diğer

dokümanlarda ne kadar az görülürse IDF değeri o kadar büyür. Bu sayede, hesaplamalar ilgili doküman için önem arz eden ayırt edici terimler üzerinden yapılabilir.

TF-ISF hesabında, TF-IDF hesabındaki dokümanların yerini cümleler alır. Yani bir terimin birden fazla cümlede sıklıkla geçmesi, o terimin ISF değerini yani ayırt edicilik derecesini küçültecektir.

TF-ISF vektörü üzerinden öznitelik değerinin hesaplaması Eş. 2.3'te verilmiştir.

$$w_{ij} = TF_{ij} \times ISF = TF_{ij} \times \log \frac{n}{|n_i|}$$

$$TF-ISF = \begin{matrix} S_1 \\ \vdots \\ S_n \end{matrix} \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & \cdots \end{bmatrix} \tag{2.3}$$

$$F_{TF-ISF}\left(S_j\right) = \frac{\sum_{i=1}^{k} w_i\left(S_j\right)}{\max \sum_{i=1}^{k} w_i\left(S_j\right)}; j = 1,..,n$$

TF_{ij}, i teriminin j cümlesi (S_j) içindeki frekansı, n dokümandaki toplam cümle sayısı ve $|n_i|$ ise i terimini içeren cümle sayısı olmak üzere dokümandaki cümlelerin TF-ISF vektörü üzerinden öznitelik değerini göstermektedir [30, 31]. İlk olarak cümlelerdeki her bir terim için TF-ISF değeri hesaplanır. Bu değerler, her bir satırı bir cümleyi, sütunları ise sözlükteki kelimeleri işaret eden bir cümle-terim matrisinde birleştirilir. Ardından bu matristeki her bir satır için bir TF-ISF öznitelik derecesi Eş. 2.3'teki şekilde elde edilir.

Başlık

Başlık özniteliği cümle (S) ve dokümanın başlığı (T) arasındaki kelime kesişimleri cümlenin önemi konusunda fikir verebileceğinden metin özetlemede sıkça kullanılmaktadır. Cümlelerin başlık özniteliğiyle derecelendirilmesinde sıkça rastlanan eşitlikler Eş. 2.4, Eş. 2.5 ve Eş. 2.6'da sunulmuştur.

$$F^{1}_{başlık} = \frac{S \text{ ve } T \text{ içindeki ortak terimlerin sayısı}}{T \text{ içindeki terim sayısı}} \tag{2.4}$$

$$F^{2}_{başlık} = \text{jaccard}\left(\vec{S},\vec{T}\right) = \frac{S \text{ ve } T \text{ ortak terimleri sayısı}}{\text{S ve } T \text{ içindeki tüm terimlerin sayısı}} \tag{2.5}$$

$$F^{3}_{başlık} = \text{kosinüs}\left(\vec{S},\vec{T}\right) = \frac{\sum_{i=1}^{n} S_i T_i}{\sqrt{\sum_{i=1}^{n} S_i^{\,2}}\sqrt{\sum_{i=1}^{n} T_i^{2}}} \tag{2.6}$$

Eş. 2.4'te cümlede ve başlıkta geçen ortak kelimelerin sayısı, başlıktaki kelime sayısına bölünmektedir [30, 31]. Eş. 2.5'te başlık ve cümle arasındaki Jaccard benzerliği ölçülmektedir [32]. Eş. 2.6'da ise başlık ve cümle vektörlerinin arasındaki kosinüs benzerliğine göre öznitelik değeri hesaplanmaktadır [28].

Anahtar Sözcükler

Dokümanda tanımlanmış anahtar sözcükler, tematik sözcükler veya söz öbekleri ile belirli bir alanda tanımlanmış özel terimler cümlenin önem derecesi ile ilişkili olabilir [33]. Bunun için otomatik metin özetleme probleminde de bu ifadelerin varlığı halinde ilgili cümle ile benzerliği hesaplanarak cümlenin öznitelik derecesi hesaplanmaktadır. Anahtar sözcükler için cümleye öznitelik değeri atamasında Eş. 2.6'dan yararlanılabilmektedir. Buna göre ilgili cümle ile anahtar sözcük dizisi arasında kosinüs veya jaccard gibi benzerlik ölçüm yöntemleri kullanılarak cümleye öznitelik değeri atanabilir.

Göreceli paragraf pozisyonu

Göreceli paragraf pozisyonu, paragrafın metindeki konumunu ifade etmektedir. Bu öznitelik tek paragraflık metinler için uygulanabilir olmamakla birlikte çok sayıda paragrafın varlığı halinde ayırt edici olabilmektedir.

Göreceli cümle pozisyonu

Göreceli cümle pozisyonu, metnin belirli bir paragrafındaki cümlenin paragrafa göre konumunu belirtmek için tanımlanan bir özniteliktir [34]. Çok paragraflı metinlerde göreceli paragraf pozisyonu ile birlikte kullanımı ayırt edici bir öznitelik sunmaktadır.

Cümle pozisyonu

Cümlenin paragraftaki ve metnin tümündeki konumu metindeki vurgunun yoğunlaştığı bölgelerle ilişkilidir. Örneğin düzenli bir yazının yazımında paragrafın ilk cümlesinin paragrafın bütünü hakkında bilgi taşıması, devam eden cümlelerin de baştaki yargıyı açıklayıcı ve destekleyici nitelikte olması sık karşılaşılan bir yazma şeklidir [35]. Bu önerme, tüm metin türleri için genelleştirilemez olsa da bu durum otomatik metin özetleme için pozisyon özniteliğinin tanımlanmasında kullanılabilmektedir.

Otomatik metin özetlemede pozisyon bilgisinin öznitelik olarak cümle derecelendirilmesinde kullanılması genel kabul görmüş bir yaklaşımdır. Pozisyon, cümlenin tüm metindeki yerine karşılık gelebileceği gibi, ait olduğu paragraftaki yerine de karşılık gelecek şekilde değerlendirilebilmektedir [36].

Ancak pozisyon özniteliğinin hesaplanmasında literatürde çok sayıda farklı görüş bulunmaktadır. Cümlelerin pozisyon özniteliğiyle derecelendirilmesinde sıkça rastlanan eşitlikler Eş. 2.7, Eş. 2.8 ve Eş. 2.9'da sunulmuştur.

$$F^{1}_{pozisyon}(\mathrm{S}) = \begin{Bmatrix} \dfrac{6-i}{5} & ; & i = \{1,2,3,4,5\} \\ 0 & ; & diğer \end{Bmatrix} \tag{2.7}$$

$$F^{2}_{pozisyon}(\mathrm{S}) = \begin{Bmatrix} 1-\dfrac{i-1}{n} & ; & \text{metnin1. yarısı} \\ i/n & ; & \text{metnin2. yarısı} \end{Bmatrix} \tag{2.8}$$

$$F^{3}_{pozisyon}(\mathrm{S}) = \begin{Bmatrix} 1-\dfrac{i-1}{n} & ; & \mathrm{i}>1 \\ 1 & ; & \mathrm{i}=1 \end{Bmatrix} \tag{2.9}$$

Çalışmalarda sıklıkla başvurulan yöntemde, Eş. 2.7'de görüldüğü gibi metnin ilk beş cümlesi doğrusal bir fonksiyonla derecelendirilmekte, diğer cümlelere ise pozisyon özniteliği derecesi sıfır olarak atanmaktadır [28]. Eşitlik 2.8'de metnin ilk yarısı ve ikinci yarısı için ayrı pozisyon fonksiyonları tanımlanmıştır [33]. Buna göre vurgu ilk ve son cümleler üzerinde yoğunlaşırken metnin iç kısımlarına doğru azalmaktadır. Öte yandan Eş. 2.9'da yer alan hesaplama şeklinde ilk cümlelerin önemli olduğuna vurgu yapılmıştır ve özniteliğin

değeri ilk cümle için maksimum iken metnin ilerleyen kısımlarında doğrusal olarak azalmaktadır [37].

Uzunluk

Metin içerisinde çok sayıda tekil terim barındıran cümlelerin daha önemli olabileceği hipoteziyle tanımlanan uzunluk özniteliği genel olarak uzun cümlelere daha büyük öznitelik derecesi atanmasını sağlarken çok kısa cümlelerin özette yer almasını engellemeye çalışır. Uzunluk öznitelik değerinin çeşitli hesaplama şekillerine Eş. 2.10, Eş. 2.11 ve Eş. 2.12'de yer verilmiştir.

$$F^{1}_{uzunluk}(\mathrm{S}) = \frac{t(S)}{\bar{t}} \tag{2.10}$$

$$F^{2}_{uzunluk}(\mathrm{S}) = \frac{t(S)}{\max\limits_{0<j<n}\left(t\left(S_j\right)\right)} \tag{2.11}$$

$$F^{3}_{uzunluk}(\mathrm{S}) = \ln\left(\bar{t} - \left|\frac{\bar{t} - t(S)}{\sigma}\right|\right) \tag{2.12}$$

Eş. 2.10'da görüldüğü gibi cümle (S) içinde geçen tekil terim sayısı cümle uzunluğu $t(S)$ olarak alınmakta ve bu değer bütün dokümandaki cümlelerin ortalama uzunluğuna bölünmektedir [30]. Benzer şekilde, Eş. 2.11'de cümle uzunluğu dokümandaki maksimum uzunluğa sahip cümlenin uzunluğuna bölünmektedir [37]. Bu yöntemlere göre cümle uzunluğu öznitelik derecesini doğru orantılı olarak etkilemektedir. Eş. 2.12'de ise bu yaklaşımdan farklı olarak cümle uzunluğu öznitelik derecesi, ilgili cümlenin uzunluğu, dokümandaki ortalama cümle uzunluğuna yaklaştıkça artar, uzaklaştıkça azalır [33, 38]. Eş. 2.12'deki $\bar{t}$ ve σ sırasıyla dokümandaki ortalama cümle uzunluğunu ve standart sapmayı ifade eder. Bu yöntemle çok uzun ve çok kısa cümlelerin önem derecelerinin düşürülmesi hedeflenir.

Tamlamalar

Otomatik metin özetlemede terimlerin yanında tamlamalar da önemlidir. Bir cümlenin isim tamlaması, sıfat tamlaması, eylem öbeği, edat ve isimden oluşan söz öbekleri gibi

tamlamalar içermesi o cümlenin önem derecesi hakkında bilgi verebilir. Bu yüzden bu tamlamalar metin özetleme öznitelikleri arasında değerlendirilmektedir. Cümlelerin tamlama öznitelik derecelerinin hesaplanmasına yönelik çok sayıda yaklaşım mevcuttur [1, 30]. Eş. 2.13, Eş. 2.14 ve Eş. 2.15'te bu yaklaşımlardan yaygın olanlar sunulmuştur.

$$F_p^1(\mathrm{S}) = \frac{\text{cümledeki } p \text{ sayısı}}{t(S)} \tag{2.13}$$

$$F_p^2(\mathrm{S}) = \frac{\text{cümledeki } p \text{ sayısı}}{\text{dokümanda bir cümledeki ortalama } p \text{ sayısı}} \tag{2.14}$$

$$F_p^3(\mathrm{S}) = \frac{\text{cümledeki } p \text{ sayısı}}{\text{dokümanda bir cümledeki maksimum } p \text{ sayısı}} \tag{2.15}$$

Eş. 2.13, Eş. 2.14 ve Eş. 2.15'te, p sembolü $P = [NP, AP, VP, PP]$ tamlamalar dizisindeki her elemanı ifade etmektedir ve ilgili cümleye ait öznitelik derecesi isim tamlaması (NP), sıfat tamlaması (AP), eylem öbeği (VP) ve edat öbeği (PP) olmak üzere dört tamlama çeşidi için ayrı ayrı hesaplanıp değerlendirilmiştir.

Özel isimler

Özel isimler, kurum adları, kısaltmalar gibi ifadeler de tamlamalar gibi cümlelerin önemi hakkında bilgi vermektedir. Bunun için bu tip terim ve tamlamalar da otomatik metin özetlemede cümlelerin derecelendirilmesinde kullanılan öznitelikler arasında yer almaktadır. Cümleye ait özel isim öznitelik derecesinin hesaplanmasında Eş. 2.13, Eş. 2.14 ve Eş. 2.15'teki yöntemler özel isimleri dikkate alacak şekilde özelleştirilip kullanılmaktadır [1, 32, 33].

Özel isimlere benzer şekilde hesaplanabilecek diğer bir öznitelik türü de tematik kelimelerdir. Bu kelimeler, belirli bir alandaki özel terimlerin yer aldığı bir sözlükten elde edilebileceği gibi dokümanlarda sık kullanılan terimler de alınarak oluşturulabilir.

Cümle-cümle ilişkisi

Metin özetleme alanındaki çalışmaların çoğunda cümlenin dokümandaki diğer cümleleri ile ilişkisinin cümlenin önemi hakkında bilgi verebileceği gösterilmiştir [39]. Burada,

dokümandaki diğer cümlelerle yüksek ilişkisi olan cümlelerin önem derecelerinin de yüksek olacağı varsayılmakta ve bu cümleler özete değer nitelikte kabul edilmektedir. Literatürde, bir cümle için bu öznitelik değerinin hesaplanmasında çeşitli yaklaşımlar uygulanmıştır [29, 31, 33] ve bu yaklaşımlar Eş. 2.16 ve Eş. 2.17'de sunulmuştur.

$$F_{ss}^{1,\mathrm{n}}(\mathrm{S}) = \frac{\sum \text{benzerlik}_{\mathrm{n}}\left(S, \dot{S}\right)}{\text{maks}_{S \in D}\left(\sum \text{benzerlik}_{\mathrm{n}}\left(S, \dot{S}\right)\right)} \tag{2.16}$$

$$F_{ss}^{2,\mathrm{n}}(\mathrm{S}) = \frac{S \text{ ve } \dot{S} \text{ içindeki ortak terimlerin sayısı}}{\text{sözlük boyutu}} \tag{2.17}$$

Eş. 2.16'da ilgili cümlenin (S) dokümandaki diğer tüm cümlelerle ($\dot{S}$) benzerliği ölçülür ve toplam benzerlik, dokümandaki her cümle değerlendirildiğinde meydana gelen maksimum toplam benzerliğe bölünür [31]. Buradaki benzerlik ölçütü kosinüs ve Jaccard gibi metin benzerliğini ölçmede kullanılan ölçütler üzerinden hesaplanabilir. Eş. 2.17'de ise cümlenin diğer cümlelerde geçen terimleriyle kesişim kümesindeki terim sayısı, sözlükteki toplam terim sayısına bölünmektedir [29, 33]. Cümle-cümle ilişkisi özniteliği bir kelimelik terimlerin yanında n kelimelik söz öbekleri için de uygulanabilir.

Şebeke bağlantı sıklığı

Şebeke bağlantı sıklığı özniteliğinde kaynak metin bir şebeke olarak ele alınır. Burada metindeki her cümle bir düğümü işaret eder ve iki düğüm arasındaki benzerlik belirli bir eşik değerin üzerindeyse o iki düğüm arasında bağlantı tanımlanır. Bu şekilde elde edilen bir şebeke yapısında çok sayıda bağlantısı olan düğümlerin önemli bilgiyi içermesi dolayısıyla daha önemli olabileceği sonucuna varılmaktadır. Bir cümle için şebeke bağlantı sıklığı, o cümle düğümünden çıkan bağlantıların sayısı alınarak belirlenmektedir [32].

Toplam benzerlik

Toplam benzerlik, şebeke bağlantı sıklığına benzer şekilde şebeke tabanlı bir özniteliktir. Toplam benzerlik özniteliğinin tanımlanması için metindeki her cümle bir düğümü işaret edecek şekilde bir şebeke yapısı tasarlanır ve iki düğüm arasındaki benzerlik belirli bir eşik değerin üzerindeyse, şebekede o iki düğüm arasına bağlantı oluşturulur. Bir cümle için

şebeke bağlantı sıklığı, o cümlenin işaret ettiği düğümünden çıkan bağlantıların benzerlik değerlerinin toplamı alınarak hesaplanır [30, 33].

2.2.2. Anlamsal öznitelikler

Bir metindeki anlamın çıkarılması, metindeki kelime, kelime kümeleri, işaretler ve semboller gibi birimlerin neyi ifade ettiklerinin ve kullanımlarındaki anlamsal ilişkilerin çıkarılabilmesi ile mümkün olur. Bu noktada bu birimlerin anlamsal özellikleri, kelimelerin metinde geçme sıklıkları, birliktelikleri ve bağlam bilgisi gibi farklı şekillerde tanımlanabilmektedir. Metin özetleme probleminde ise metne ait anlamsal öznitelikler, metindeki kelimelerin özyerleşikleri (embedding) temel alınarak tanımlanmış ve kullanılmıştır. Bu bölümde kelime özyerleşikleri ile ilgili temel alan bilgisi sunulmuştur ve bu çalışmada kullanılmış olan iki özyerleşik çeşidine yer verilmiştir.

Bir metni makine öğrenmesi veya bilgi erişim görevlerinde kullanabilmek için onu sayısal bir forma dönüştürmek gerekir. Bu sayısal formun metni temsil düzeyi amaçlanan görevin başarısını etkileyen en önemli faktörlerden birisidir.

Metni temsil etmede kullanılan en basit yaklaşımlardan birisi metindeki kelimeleri veya kelime öbeklerini kategorik değişkenler olarak ele almak ve bunlar üzerinde tek dereceli kodlama uygulamaktır. Bu durumda her bir kelime, sözlükteki kelime sayısı kadar elemanı olan bir vektörle ifade edilir ve bu vektörde ilgili kelimenin sözlükte bulunduğu sıradaki değer 1, vektörün diğer elemanları 0 olarak atanır. Örneğin, S = [“ece”, “kraliçe”, “mutlu”, “bebek”, “matem”] olmak üzere S sözlüğü için kelime vektörleri sırasıyla [1, 0, 0, 0, 0], [0, 1, 0, 0, 0], [0, 0, 1, 0, 0] ve [0, 0, 0, 0, 1] olmaktadır.

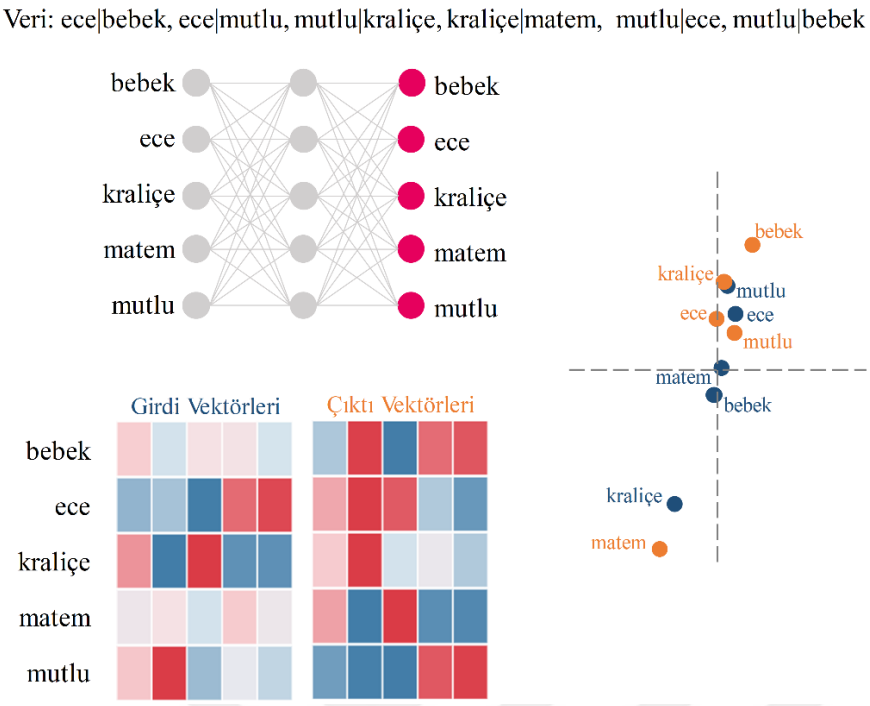
İkili değer alan vektörlerle kelime temsili, söz konusu görevin gerçekleştirilmesi için uygulanabilir olmakla birlikte önemli bazı dezavantajları vardır. Öncelikle sözlükteki kelime sayısı kadar elemana sahip olan ancak sadece bir değeri anlamlı değer taşıyan vektörlerin bir araya gelmesiyle oluşan seyrek matrisler hesaplama maliyetini önemli oranda artırmaktadır. Daha da önemlisi, bu gösterim şekli kelimelerdeki anlamsal benzerliği hesaba katmamaktadır ve bu yüzden tüm kelimelerin birbiriyle eşit derecede ilişkili/ilişkisiz olmasına dayalı modeller geliştirilmesine yol açmaktadır. Aslında “ece” ve “kraliçe” kelimeleri eşanlamlı olduklarından [1, 0, 0, 0, 0] ve [0, 1, 0, 0, 0] şeklinde ayrı iki vektörle

temsilleri kelimelerdeki bu anlamsal bağlılığı ortadan kaldırmaktadır. Öte yandan, bu iki kelime her durumda birbiri yerine kullanılamayacak olduğundan tek bir vektörle temsilleri de çözüm olmayacaktır. Bu yüzden bu kelimelerin anlam uzayında birbirine yakın vektörlerle temsil edilmesi ve model geliştirilirken bu anlamsal ilişkinin de göz önüne alınması gerekmektedir.

Metni sayısal olarak ifade etmek için sıklıkla kullanılan bir diğer yaklaşım ise TF veya TF-IDF tabanlı kelime kümesi üretmektir. Farklı problemler için farklı uygulama alanlarına sahip bu yöntemde kelime kümesi, doküman-terim veya metin özetleme problemi için Eş.2.3'te görüldüğü gibi cümle-terim matrisini ifade etmektedir.

Bir doküman-terim matrisinin satırları dokümanlara ve sütunları ise sözlükteki terimlere karşılık gelmektedir. Matristeki sütun sayısı yine sözlükteki terim sayısı kadardır ve matrisin elemanları TF veya TF-IDF değerleridir. Burada terimler, bir vektörle değil dokümana özgü bir değerle ifade edilmiş olur. Bu yaklaşım, uzun yıllar çok sayıda makine öğrenmesi modelinin metin madenciliğinde uygulanmasına imkân sağlamış ve bilgi erişim sistemlerinin temel metotları arasında yaygın kullanılmış olmakla birlikte, matris elemanları terimlerin yalnızca terim frekansıyla ilgili bilgi içerdiğinden anlamsal bağlantılara sahip değildir.

Metni temsil etmede daha yeni yaklaşım olan özyerleşik, metin madenciliğinde bir kelime veya kelime grubunun anlam bilgisinin gerçek değerli vektörlerle temsiline verilen addır [40]. Özyerleşikler kelimelere, öbeklere, cümlelere, paragraflara ve hatta dokümana ait olmak üzere oluşturulabilirler ve söz konusu metnin vektörel gösterimini sağlarlar. Özyerleşiklerin tanımlanması temelde bir yapay sinir ağına dayanır. Şekil 2.1'de beş kelimeden oluşan bir sözlükteki kelimelere ait vektörlerin hesaplanabilmesi için örnek bir mimari sunulmuştur.



Şekil 2.1. Beş kelimelik bir sözlük için özyerleşik örneği

Şekilde 2.1'de eğitimin sonucunda elde edilen ve anlamsal ilişkilerin benzer renklerle gözlenebildiği kelime vektörleri de yer almaktadır. Oluşan vektörlerin temel bileşen analizi (principal component analysis) ile iki boyutlu bir düzleme yerleştirilmesi sonucunda kelimelerin birbirleriyle anlamsal ilişkileri net bir şekilde görülebilmektedir. Sarı renk ile gösterilen çıktı kelime vektörleri incelendiğinde, bu veri için "ece" ve "kraliçe" kelimeleri uzayda birbirlerine yakın olmakla birlikte "matem" bu kelimelerden çok uzaktır.

Özyerleşiklerin literatürde uzun bir geçmişi olmasına rağmen, bu konudaki çalışmalara yön vermiş olan en etkili çalışma Mikolov ve diğerleri tarafından 2013'te yayınlanmıştır [34, 41]. Bu çalışmada, önceki karmaşık mimarilere alternatif olarak basit, hesaplama maliyetini önemli ölçüde azaltan ve büyük veri kümelerinde daha yüksek ölçeklenebilirliğe sahip iki farklı mimari önerilmiş ve word2vec ismiyle anılan bir kelime temsil yöntemi geliştirilmiştir. Burada üzerinde durulan başlıca prensip, ilk olarak Harris tarafından önerilmiş olan "benzer bağlamlara sahip kelimeler benzer anlama sahiptir" önermesidir [42]. Word2vec modelin önerilmesinin ardından, kelime özyerleşikleri, metin madenciliğinin, doğal dil işlemenin ve bilgi çıkarım teknolojilerinin her alanında sıklıkla kullanılan bir metin temsil yöntemi haline gelmiştir [43,44].

Word2vec özyerleşikleri iki farklı mimari ile geliştirilebilir. Bunlar sürekli kelime kümesi (Continuous Bag-of-Words-CBOW) modeli ve sürekli atlamalı-gram (Continuous Skip-Gram) modelidir. Her iki durumda pencere tabanlı bir veri yapısından faydalanılır ve pencere boyutuna göre kelime ve bağlam tanımlanır. Bağlam, kelimenin pencere boyutuna göre belirlenen komşu kelimelerinden oluşmaktadır. CBOW modelde özyerleşikler, bağlam dikkate alınarak hedef kelimenin tahmin edilebileceği şekilde öğrenilir. Atlamalı-gram modelde ise kelimeden bağlam tahmin edilmeye çalışılır.

Literatürde sıklıkla kullanılan bir diğer popüler özyerleşik türü ise kelime gösterimi için küresel vektörler (Global Vectors for Word Representation: GloVe) olarak adlandırılmaktadır [45]. Bu yaklaşımın da temelinde "benzer bağlamlara sahip kelimeler benzer anlama sahiptir" önermesi vardır [42]. Ancak word2vec modelden farklı olarak GloVe, ağırlıklı en küçük kareler modeli üzerinden küresel kelime-kelime birlikteliği sayısını eğiterek özyerleşiklerin tanımlanmasına yönelik istatistiksel bir yaklaşım sunar.

2.3. Otomatik Metin Özetleme Yöntemleri

Otomatik metin özetlemede en temel aşama önemli ve vurgulanmış ifadelerin tespit edilmesidir. Ardından çıkarmalı metin özetlemede bu önemli ifadeler birleştirilir. Soyutlamalı metin özetlemede ise anlamsal ilişkilere göre yeniden cümleler kurularak özet metin üretilir. Bu yüzden hangi ifadelerin metin özetlerine dâhil edilmek için seçileceğini belirleyen önemlilik derecesinin tespiti özetleme sonucunda elde edilen metnin doğruluğunu etkileyen en önemli aşamadır.

2.3.1. Frekans tabanlı terim ağırlıklandırma yaklaşımı

Çıkarmalı metin özetlemede metin bölümlerinin önem derecelerinin belirlenmesinde kullanılmış olan yaklaşımlar arasında en basit ve eski olanı frekans tabanlı terim ağırlıklandırma yaklaşımlarıdır. Bu yaklaşımlar temelde terimin metinde görülme sıklığını esas alan TF, TF-IDF gibi terim ağırlıklandırma ölçütlerine dayanır. Burada genel yaklaşım bir dokümanda görülme sıklığı yüksek olan terimlerin önem derecelerinin büyük olduğu ve buna bağlı olarak bu terimleri içeren ifadelerin de önemli olduğu varsayımına dayanmaktadır [27].

2.3.2. Şebeke tabanlı yaklaşımlar

Şebeke tabanlı yaklaşımlar içerisinde en yaygın kullanılan yöntem TextRank algoritmasıdır [46]. Bilgi çıkarım metotlarında, özellikle arama motorlarında sıklıkla başvurulan PageRank yönteminin [47] metin özetleme problemine uygulanması ile elde edilmiş olan bu algoritma, metin özetlemeye denetimsiz bir yaklaşım sunar [46, 48]. Burada metin içerisindeki cümle, paragraf gibi bölümler düğümleri, ilişkiler de kenarları ifade eder. Düğümler arasındaki benzerlik derecesine göre kenarlar oluşturulur ve ağırlıklandırılır. Şebeke içerisinde bir düğümün önem derecesi komşu düğümler ile arasındaki ilişkiyi ifade eden kenarların ağırlıkları kullanılarak hesaplanır. Şebeke tabanlı yaklaşımlar, bir metin içindeki bölümler arasındaki ilişkinin bir şebeke yapısıyla ifade edilmesine dayanır [48–50]. Metin içindeki bölümlerin birbirleri ile benzerlikleri modellenerek genel özetleme yapılabilir. Ayrıca, belirli sorgulara göre modellenen şebekelerle sorgu tabanlı özetleme yapmak da mümkündür [51].

2.3.3. Saklı anlam analizi

Otomatik metin özetleme alanında ilk yıllarda en sık karşılaşılan yaklaşım metin içerisindeki cümle ve paragraf gibi bölümlere saklı anlam analizi uygulamaktır [52–57]. Saklı anlam analizi, ortak kelimeler bulundurmuyor olsalar dahi metin bölümleri arasında gizlenmiş anlamsal ilişkilerin tanımlanmasını sağlamaktadır [2, 57]. Metin bölümleri olarak cümlelerin seçildiği düşünülürse bir doküman özetinin çıkarılması için ilk olarak dokümanın (D) cümle kümesinden ($D=\left[S_1,S_2,\ldots,S_s\right]$) oluşan terim-cümle şeklindeki bir girdi matrisi $A=\left[A_1,A_2,\ldots,A_n\right]$ olarak ifade edilmesi gerekir. Burada A_s, s indisli cümlenin ağırlıklandırılmış terim frekanslarını ifade eder. Bu matris üzerine tekil değer ayrıştırma uygulanarak matris, $r=\mathrm{rank}\left(A\right)$ adet bağımsız taban vektörüne ayrıştırılır. Dokümandaki tüm cümleler bu vektörlerden birinin alt kümesi niteliğindedir. Böylelikle, metni birbiri ile ilişkili cümleler aynı taban vektöründe olmak üzere anlamsal olarak kümelere ayrıştırmış olur. Bu kümelerin içerisindeki en yüksek indeks değerine sahip cümle özet metinde yer almaya değer olarak belirlenir.

2.3.4. Makine öğrenmesi

Bilgisayar sistemlerinde öğrenme, veriden öğrenme ve uzman görüşüne dayalı öğrenme olmak üzere iki şekilde yapılabilmektedir. Herhangi bir problem için yeterli ve kaliteli veri olması halinde veriden öğrenmeye dayalı makine öğrenmesi yöntemleri iyi düzeyde çıkarsama ve tahmin başarısı sunabilmektedir. Buna göre eldeki verinin kalitesi, doğruluğu ve örnekleme başarısı kullanılacak olan makine öğrenmesi tekniğinin başarısını doğrudan etkilemektedir. Öte yandan, veri olmaması, verinin tutarlı ve kaliteli olmaması veya olası tüm durumları kapsayamaması halinde uzman görüşünden yararlanan öğrenme teknikleri kullanılmaktadır. Otomatik metin özetleme probleminde metindeki cümlelerin önem derecelerinin belirlenmesinde ve önemli cümlelerin seçiminde bu iki öğrenme yaklaşımından da yararlanılmaktadır.

Otomatik metin özetleme problemi, metne ait hedef çıkarmalı özetlerin olması halinde denetimli öğrenme yöntemleri ile modellenebilir. Bu durumda problem metin parçalarının sınıflandırması olarak değerlendirilir. Bunun için, metin parçalarının cümleler olduğu düşünüldüğünde, bir cümle ($\vec{S} = F_1, F_2, \cdots, F_n$) için bir öznitelik vektörünün tanımlanması ve bu cümlenin özete alınmaya değer olup olmadığının model tarafından öğrenilmesi gerekmektedir.

Sınıflandırıcı modelin eğitiminin ardından, yeni bir dokümanın içindeki cümlelerin özete alınmaya değer olma derecesi hesaplanır ve daha önce belirlenmiş bir değerin üzerinde olan cümleler özet metne alınarak çıkarmalı özetleme yapılmış olur.

Çıkarmalı metin özetleme, cümle sınıflandırma problemi olarak ele alındığında Naive Bayes [58–62], destek vektör makineleri [62–65], karar ağaçları [58, 62, 66–68], yapay sinir ağları [36, 58, 62, 69] ve özellikle son yıllarda derin öğrenme yöntemleri [70–74] kullanılarak gerçekleştirilebilmektedir.

Naive Bayes sınıflandırıcı

Naive Bayes, koşullu olasılığa bağlı bir istatistiksel sınıflandırma yöntemidir. Sınıflandırma için kullanılacak her bir niteliğin istatistiksel açıdan birbirinden bağımsız olduğu varsayımına dayanarak istatistiksel çıkarsama yapar. $P\left(x|C_j\right)$ ifadesinin j sınıfındaki bir

örneğin x olma olasılığı, $P(C_j)$ ifadesinin j sınıfının olasılığı, $P(x)$ ifadesinin herhangi bir örneğin x olma olasılığı olduğu düşünüldüğünde x olan bir örneğin j sınıfına ait olma olasılığı Eş. 2.18'deki gibi hesaplanmaktadır.

$$P(C_j|x) = \frac{P(x|C_j)P(C_j)}{P(x)} = \frac{P(x|C_j)P(C_j)}{\sum_k P(x|C_k)P(C_k)} \tag{2.18}$$

Yeni gelen bir giriş için tüm sınıflar üzerinden Eş. 2.18'deki şartlı olasılık değeri hesaplandığında, yeni giriş maksimum $P(C_j|x)$ 'e sahip olan sınıfa atanmaktadır.

Karar ağacı

Karar ağaçları, bir veri kümesindeki kayıtları, niteliklerin aldığı değerler üzerinden tanımlanan bir dizi kuralla sınıflandırır. Sınıflandırma problemlerinin çözümü açısından ele alındığında, yeni bir giriş için oluşturulan kurallar yeniden kullanılarak ait olduğu sınıf bulunur. Karar ağaçlarının uygulanmasında en önemli aşama girişlerdeki niteliklerin en iyi şekilde sıralanarak sınıflandırmanın yapılmasıdır. Niteliğin kategorik değere sahip olmaması halinde farklı yöntemler ile karar ağaçları oluşturulabilmektedir. Karar ağaçlarının oluşturulmasında entropiye dayalı yaklaşıma sahip olan ID3 veya C4.5 ile regresyona dayalı yaklaşıma sahip olan Twoing veya Gini yaygın kullanılan algoritmalardır.

Destek vektör makinesi

Destek vektör makinesi, Vapnik-Chervonenkis istatistiksel öğrenme teorisine dayalı bir sınıflandırma yöntemidir [75]. Çok boyutlu bir veri kümesindeki kayıtları, sınıflandırma doğruluğunu en aza indirecek şekilde bir hiper düzlem ile ayrıştırabilmeyi amaçlar ve bu düzlemi tanımlayan fonksiyonu öğrenir. Sınıflandırma problemini kareli en iyileştirme problemine dönüştürüp çözer ve bu özelliğiyle sınıflandırma performansı, hesaplama karmaşıklığı ve çok boyutlu veride uygulanabilirlik açısından önemli avantajları vardır.

Destek vektör makinesi yüksek genelleme kapasitesine sahiptir ve çok boyutlu uzaylarda etkili sınıflandırma yapılabildiğinden doğası gereği karmaşıklık içeren metin verilerinin üzerinden yapılan çıkarsamalarda sıklıkla kullanılan bir yöntemdir [76–78].

Yapay sinir ağları

Yapay sinir ağları, insan beynindeki sinir hücrelerinin etkileşimlerini taklit ederek beynin öğrenme, hatırlama, genelleme yapma gibi işlevlerini makine öğrenmesi alanında gerçekleştirebilmeyi amaçlayan yapay öğrenme modelleridir. Klasik bir yapay sinir ağında ortak olan bileşenler sinir hücrelerine karşılık gelen düğümler, aksonlara karşılık gelen bağlantılar, sinapslara karşılık gelen girişler, dentritlere karşılık gelen toplama fonksiyonu ve hücre çekirdeğine karşılık gelen aktivasyon fonksiyonlarıdır. Alınan bir girdiden çıktı üretilmesi için girdiler ağırlık değerleriyle çarpılır, birbiri ile toplanır, varsa bias değeri eklenir ve sonuç aktivasyon fonksiyonundan geçirilerek istenen aralıktaki çıktı elde edilmiş olur. Bu yapının doğru bir çıkış değeri üretebilmesi için eğitilmesi gerekmektedir. Eğitim sürecinde, eğitim veri kümesi ile düğümlerin ağırlık değerlerinin toplam hatayı minimum yapacak şekilde ayarlanması gereklidir.

Derin öğrenme

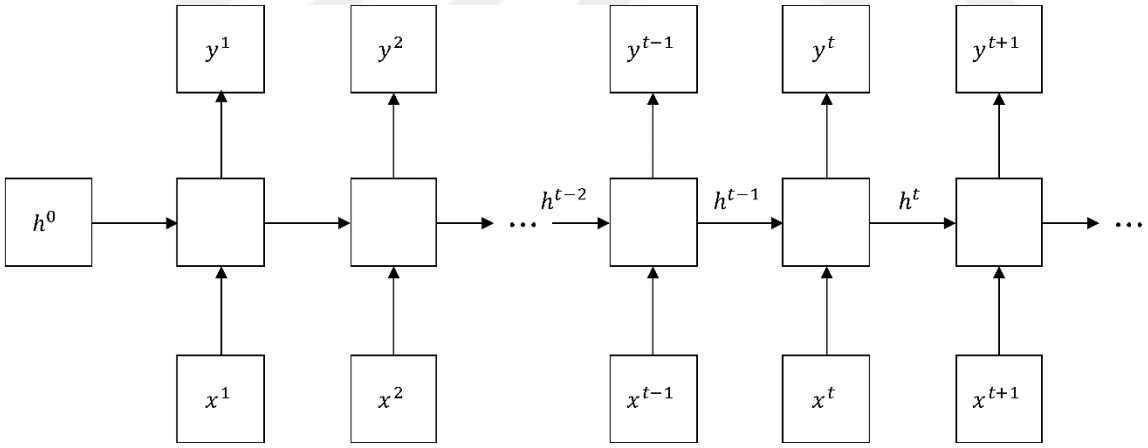
Geleneksel yapay sinir ağları, çok büyük boyuttaki veriler üzerinde çalışırken hem hesaplama maliyeti hem de çıkarsama ve genelleme başarısı açısından olumsuz yönde etkilenir. Bu tür bazı problemlerin çözümünde giriş verisi için öznitelik silme, öznitelik seçme, öznitelik birleştirme gibi boyut indirgeme yapılmakta ve bu süreçler sonucunda elde edilen öznitelikler yapay sinir ağının eğitiminde kullanılmaktadır. Bu durumun sonucunda sinir ağlarının gerçekleştiriminin bir ön aşaması olarak kapsamlı bir öznitelik analizine ihtiyaç duyulur. Bu tür büyük boyutlu veriler için öznitelik çıkarımına ihtiyaç duymayan ve ham veriler üzerinde öznitelik çıkarımını da kendi yapısı içerisinde elde eden derin öğrenme yöntemleri kullanılabilir. Derin öğrenme bir temsil öğrenme sürecine sahiptir. Girdi vektörünün, çıkış katmanına gelmeden önce çok sayıda doğrusal olmayan işlemlerden geçirilmesi ile öznitelik öğrenme işlemi gerçekleştirilmiş olur.

Basit özniteliklerin olduğu veriler için çok sayıda makine öğrenmesi yöntemlerinden birisi uygulanabilir. Verilerin karmaşık hale gelmesi durumunda özniteliklerin çıkartılması gereksinimi doğmaya başladığında derin ağlar karmaşık veriler için daha iyi performans sergilemeye başlar. Verinin karmaşıklığı ve boyutu daha da arttıkça, sinir ağı çok daha fazla gizli katmana ve düğümlere sahip olma eğilimi gösterir ve bu durum eğitim maliyetini önemli ölçüde arttırır. Görüntü ve metin gibi çok karmaşık öznitelikler içeren verilerde, derin

sinir ağları neredeyse en uygun seçenektir. Çünkü böyle bir problem için amaç fonksiyonunu modelleyebilmek için gerekli katman ve düğüm sayısı geleneksel sinir ağlarının uygulanamaz hale gelmesine sebep olmaktadır. Bu noktada, derin sinir ağları ile daha kolay gerçekleştirim yapılabilmekte ve başarılı sonuçlar elde edilebilmektedir [7, 79–84].

Metin türündeki verilerde yaygın kullanılan derin öğrenme modelleri arasında evrişimli sinir ağı (ESA) ve özyineli sinir ağı (ÖSA) sayılabilir. ESA, genellikle görüntü veriler üzerine yapılan çalışmalarda sıklıkla kullanılan bir veri filtreleme ve ham veriden öznitelik keşfetme modelidir. Görüntü sınıflandırma, nesne tanıma, kümeleme gibi denetimli ve denetimsiz birçok çözümün bir parçası olarak kullanılabilmektedirler. Metin verisi için ESA'lar, genellikle metindeki kelime özyerleşiklerinin ham giriş vektörünün, sınıflandırma yapmak adına yeniden yapılandırılması görevini üstlenmiştir [24, 85].

ÖSA, sıralı bir veri üzerindeki bir girişe ait çıkarsamada önceki girişlerin değerlerini veya çıkarsama sonuçlarını da dikkate alarak sonuç üreten derin sinir ağı çeşididir. Genel bir ÖSA gösterimine Şekil 2.2'de yer verilmiştir.



Şekil 2.2. Çok giriş-çok çıkış mimarisinde tek yönlü bir ÖSA'nın şematik gösterimi

Şekil 2.2'de görüldüğü gibi bir t anında bir ÖSA hücresinden elde edilen çıktının belirlenmesinde t anındaki girdi (x^t) ile $t-1$ anında üretilen çıktı (h^{t-1}) birlikte değerlendirilmektedir.

Eş. 2.19 ve Eş. 2.20'de bir ÖSA hücresinde bir önceki katmandan gelen h^{t-1} gizli çıktının, ilgili katmanın girdisiyle (x^t) birlikte hücrenin çıktısının (y^t) ve gizli çıktısının (h^t) nasıl hesaplandığı sunulmuştur.

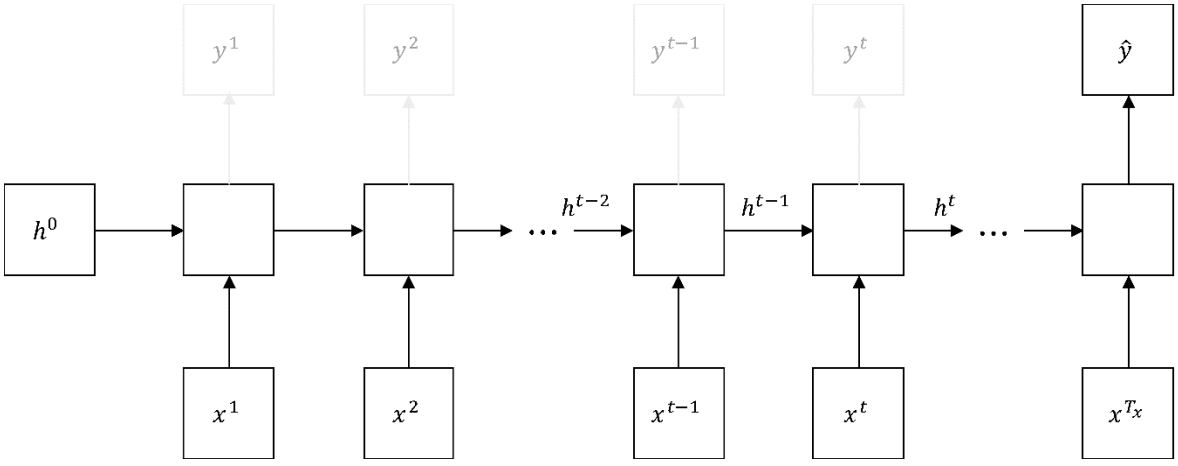
$$h^t = g_1\left(W_{hh}h^{t-1} + W_{xh}x^t + b_h\right) \tag{2.19}$$

$$y^t = g_2\left(W_{yh}h^t + b_y\right) \tag{2.20}$$

Eş. 2.19 ve Eş. 2.20'de temel olarak mevcut gizli çıktının önceki gizli çıktı ve mevcut girişin bir fonksiyonu olduğu görülmektedir. Burada g_1 ve g_2 aktivasyon fonksiyonları; W ve b ise fonksiyonlara ait katsayılardır.

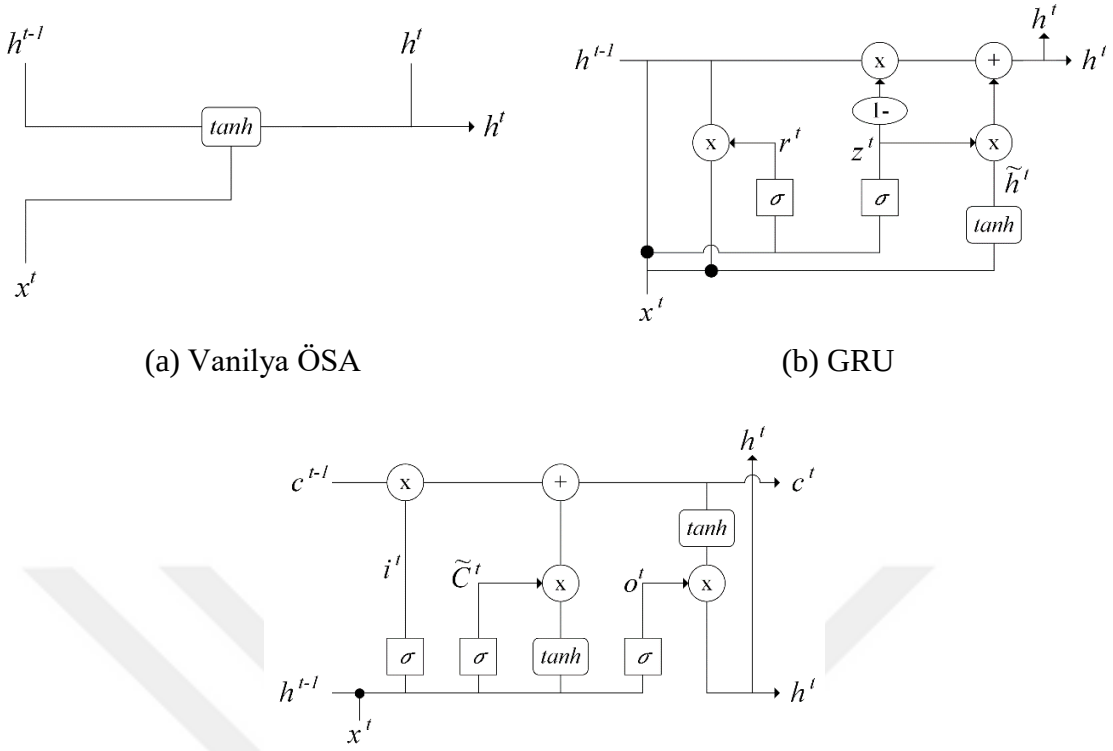
Şekil 2.2' yer alan ÖSA, tek yönlü ve standart ÖSA yapısıdır (t-ÖSA). Bunun yanında daha karmaşık yapıdaki çift yönlü ÖSA (ς-ÖSA) yapısı da farklı problem türleri ve çözüm mimarilerinde kullanılmaktadır [34]. Bu durumda t anındaki hücreden elde edilecek sonuç, mevcut durumdan önceki ÖSA hücresinin sonucuna bağlı olduğu gibi sonraki hücrenin de durumuna bağlı olarak hesaplanmaktadır.

ÖSA'ların giriş parametrelerinin sayısına (T_x) ve gizli çıkış sayısı (T_y) olmak üzere bir giriş-bir çıkış ($T_x = 1,\ T_y = 1$), bir giriş-çok çıkış ($T_x = 1,\ T_y > 1$), çok giriş-bir çıkış ($T_x > 1,\ T_y > 1$) ve çok giriş-çok çıkış ($T_x > 1,\ T_y > 1$) olmak üzere dört temel mimariye sahiptir. Otomatik metin özetleme problemi, cümle sınıflandırma problemi olduğundan Şekil 2.3'te verilen çok giriş-bir çıkış mimarisindeki ÖSA yaygın olarak kullanılmaktadır [34].



Şekil 2.3. Çok giriş-tek çıkış mimarisinde tek yönlü bir ÖSA'nın şematik gösterimi

Literatürde sıklıkla kullanılan ÖSA hücresi çeşitlerine Şekil 2.4'te yer verilmiştir.



(a) Vanilya ÖSA (b) GRU

(c) LSTM

Şekil 2.4. ÖSA türlerinin sinir hücrelerinin içyapısı

Şekil 2.4'te en basit yapıdaki ÖSA olarak değerlendirilen ve Şekil 2.4 (a)'da gösterilen vanilya ÖSA hücresinde bir önceki durum girdisi olan h^{t-1}, şimdiki girdi verisi x^t ile bir aktivasyon fonksiyonuna girdi olarak verilir ve hücrenin çıktısı üretilir. Bu çıkış değeri hem (bir aktivasyon fonksiyonundan geçirilerek) dışarı verilir hem de bir sonraki duruma girdi olarak sunulur. Giriş ve çıkış durumların aynı mantıkla birleştirildiği ancak standart ÖSA'daki kaybolan gradyan (vanishing gradient) probleminin çözümü olarak hücrelere ilave bazı yeteneklerin daha getirildiği iki tip ÖSA hücre yapısı daha mevcuttur. Bunlardan geçitli tekrarlayan birimler (gated recurrent units: GRU) tipi ÖSA'da bir sıfırlama ve bir güncelleme bloğu bulunur. Bir GRU hücresi Şekil 2.4 (b)'de sunulmuştur. Uzun-kısa dönem bellekli (long short-term memory: LSTM) ÖSA ise Şekil 2.4 (c)'deki giriş birimi, unutma birimi ve çıkış birimi olmak üzere üç temel bileşenden oluşmaktadır.

Kural tabanlı uzman sistemler

Otomatik metin özetlemeye kural tabanlı uzman sistemlerle çok sayıda çözüm geliştirilmiştir. Bu sistemler makine öğrenmesi yöntemlerinden farklı olarak veriden değil uzman bilgisinden öğrenirler ve çıkarsama sürecini uzmandan edinilen kuralları kullanarak

gerçekleştirirler. Bu alanda yoğun olarak çalışılmış uzman sistemler bulanık çıkarsama sistemleridir [1, 19, 40–43]. Bulanık çıkarsama sistemleri, uzmandan edindiği kuralları, bulanık küme teorisini de hesaba katarak modelleyen ve çıkarsama sürecinde, belirli bir girdi kümesi için yine bu kurallar üzerinden karar verilmesini sağlayan sistemlerdir. Otomatik metin özetlemede sözdizimsel özniteliklerin birbirleriyle ve sistem çıktısı olan cümle önemi ile olan ilişkisi tanımlanarak bulanık kurallar elde edilir. Ardından Mamdani, Takagi-Sugeno, Tsukamoto gibi çeşitli bulanık sistem mimarileri üzerinden cümlelerin önem derecesi belirlenmektedir.

Bulanık çıkarsama sistemleri arasında insan düşünce yapısına ve doğal dile yakınlığı dolayısıyla en çok tercih edilen Mamdani tipi bulanık çıkarsamada kuralların hem öncül hem de ardıl kısımları doğal dil ile doğrudan ilişkilidir ve bulanık değerlere sahiptir. Sisteme keskin bir girdi kümesi verildiğinde, girişlerin her biri ilgili giriş sözel değişkeni ile eşleştirilir ve bulanıklaştırılır. Bulanıklaştırma, ilgili sözel değişkenin o an almış olduğu keskin değerin sözel değişkene ait bulanık kümelere aidiyet derecesini hesaplama işlemidir. Bulanıklaştırma sonrası her bir kural için çıkarsama yapılır ve bu çıkarsama sonuçları ardıl kısmı aynı olan kurallar için birleştirilir. Sonuçta çıkış sözel değişkeninde yer alan bulanık küme sayısı kadar bulanık küme içeren ve bu kümelerin farklı derecelerle aktif olduğu bir bulanık sonuç elde edilir. Sistemin keskin bir sonuç üretebilmesi için bu sonucun keskin bir değere dönüştürülmesi gerekir. Bunun için ağırlık merkezi, alanın merkezi, maksimumun merkezi gibi farklı yöntemlerle durulaştırma yapılır ve sistemin keskin çıktısı elde edilir.

Takagi-Sugeno tipi bulanık çıkarsamada, bulanık kuralların öncül kısımları bulanıkken ardıl kısımları girdi sözel değişkenlerine bağımlı keskin bir fonksiyondur. x ve y giriş, z çıkış sözel değişkeni ve s_i bu sözel değişkenlere ait bulanık kümeler olmak üzere Takagi-Sugeno çıkarsamanın kural yapısı "EĞER $x:s_1$ ve $y:s_1$ İSE $z=f\left(x,y\right)$" şeklindedir. Nihai çıktının değerinin elde edilmesinde genellikle ağırlıklı ortalama yöntemi kullanılır. Tsukamoto bulanık çıkarsamada ise, kuralların sonuç kısımları monoton fonksiyonlardır ve kural yapısı sabit c değeri için "EĞER $x=s_1$ ve $y=s_1$ İSE $z=c$" şeklindedir. Bir başka deyişle bu tür çıkarsama için kullanılan kuralların öncül kısmı bulanık iken ardıl kısımları keskin değerdir. Her bir kuralın sonucunun ağırlıklı ortalaması alınarak nihai değer hesaplanır.

Bulanık sistem gerçekleştiriminin en önemli aşaması bulanık kuralların tanımlanmasıdır. Bu aşama, en iyi durumda uzman tarafından yapılır, ancak uzman görüşünün elde edilemediği durumlarda veya çok karmaşık öznitelik havuzundan faydalanılması gereken problemlerde mevcut veriden kural üretmek de mümkündür. Bu durum, bulanık sistemlere veri bağımlılığı getiriyor olsa da çok sayıda girdi sözel değişkenin birbiriyle ve çıktı sözel değişkeni ile ilişkisinin insanlar tarafından elle tanımlanması ve bu ilişkilerin bulanık kurallar olarak ifade edilmesi çok zor ve bazı durumlarda imkânsızdır. Bu nedenle veri üzerinden bulanık kuralları oluşturan bir yönteme ihtiyaç duyulmaktadır.

Literatürde veriden bulanık kuralları öğrenen çok sayıda çalışma bulunmaktadır. Ancak bu yöntemler bulanık kuralları oluştururken verinin tümü üzerinden en iyileştirme yöntemleriyle, tamamen veri bağımlı çözümler sunarlar. Böyle bir durumda çıkarsamayı veri üzerinden gerçekleştirmemeyi ve aynı problem için farklı veriler üzerinde genel çözümler sunabilmeyi en büyük avantaj olarak değerlendiren bulanık kural tabanlı sistemin uygulanmasında çelişki doğmaktadır. Çünkü veriden öğrenme noktasında çok sayıda makine öğrenmesi yönteminden doğrudan yararlanılabilir. Böyle bir durumda veriden doğrudan faydalanmak yerine kural tabanlı sistemin bir bileşeni olan kuralları üretmek amacıyla yararlanmanın etkinliği ve verimliliği her zaman beklenen düzeyde olmamaktadır. Gerçekte, kural tabanlı bir sistemde veri bağımlılığı olmamalı veya minimum düzeye indirgenmelidir. Bu nedenle, bu çalışmada üzerinde çalışılan probleme ait verinin büyüklüğünden çok az etkilenen, en iyileştirme ve öğrenme süreçleri içermeyen, kuralları verideki kayıtlar üzerinde bir kez gezerek çıkarmaya çalışan bir yöntem olan Wang-Mendel kural üretme algoritmasından (WM) [86] yararlanılmıştır. Ayrıca aynı amacı gerçekleştirmek için veri kümesini her bir kaydı doğrudan kullanmayan yeni bir kural üretme yaklaşımı önerilmiştir.

WM yönteminde bulanık kuralların tanımlanması için verideki her kayıt bir aday bulanık kural olarak değerlendirilir ve bu aday kurala bir ağırlık değeri belirlenir. Ağırlık değerinin belirlenmesinde tüm girdi sözel değişkenlerinin o kayıtta karşılık geldiği bulanık kümeye aitlik derecesi hesaplanır. Bulanık küme teorisinde, bir giriş birden fazla bulanık kümeye belirli derecede aittir. WM yönteminde, bu durumlarda ilgili giriş en yüksek derece ile ait olduğu bulanık kümeye ait olarak kabul edilir. Bulanık kuralın çıktı sözel değişkeni için de aynı şekilde bir aidiyet derecesi hesaplanmaktadır. Tüm sözel değişkenler için bu aidiyet dereceleri hesaplanıp birbiriyle çarpılır ve elde edilen değer aday bulanık kuralın ağırlık

değerini ifade eder. Sonraki kayıtlar içerisinde bu kuralın öncül kısmı ile örtüşen başka bir aday bulanık kural olursa ve kural ağırlığı önceki aday bulanık kural ağırlığından daha yüksekse, kuralın ağırlığı ve gerekirse ardıl kısmındaki sonuç bulanık küme güncellenir.

Bu tez çalışmasında, veriden minimum yararlanarak bulanık kural üretimini hedefleyen uçtan uca bulanık çıkarım yapan bir kural üretme yöntemi E2E önerilmiştir [74]. Orijinal E2E bir hiyerarşik bulanık sistemin kural üretme, kural bölme ve hiyerarşik bulanık çıkarsamayı gerçekleştirme adımlarının tümünü ele alan bir yaklaşım olarak önerilmiştir. Bu tez çalışması kapsamında E2E yönteminin kural üretme işlemi kullanılmıştır. E2E'de veri kümesindeki her bir öznitelik bir sözel değişken olarak tanımlanmaktadır. Bu sözel değişkenlere ait bulanık kümelerin tanımlanmasında WM yaklaşımı kullanılabileceği gibi özelleştirilmiş bulanık küme tanımları da yapılabilmektedir. Şekil 2.5'te E2E ye ait kaba kod sunulmuştur.

procedure SKORAYARLA($\vec{C}, fv$)
  for all bulanık küme $fs \in fv \rightarrow$bulanık kümeler **do**
    $s_p \leftarrow$ önceki bulanık küme ile kesişim noktası
    $e_p \leftarrow$ sonraki bulanık küme ile kesişim noktası
    $fs \leftarrow$SKORHESAPLA($\vec{C}$, s_p, e_p)
  end for
  $fv \leftarrow fv$'deki her bulanık küme skorunu normalize et
  return fv
end procedure
procedure SKORHESAPLA($\vec{C}$, min, max)
  $nbP \leftarrow C = [min, max]$ aralığında pozitif sınıf etiketli örnek sayısı
  $nbT \leftarrow C = [min, max]$ aralığında toplam örnek sayısı
  return nbP/nbT
end procedure
procedure KURALSETIOLUSTUR($Veri$)
  for all $c \in Veri \rightarrow$ öznitelikler
    $\vec{C} \leftarrow c$ için öznitelik değerleri listesi **do**
    $fv \leftarrow c$ için sözel değişken
    $fv =$ SKORAYARLA($\vec{C}, fv$)
  end for
  2D öncül kural matrisini oluştur R^o
  1D ardıl kural matrisini tanımla R^a
  $R = [R^o R^a]$
  for all $r \in R^o_r$ **do**
    $R^a_r = \frac{\sum_{i=1}^{n} \overline{p_{i,j}}^r}{\sum_{i=1}^{n} max(p_i)}$
    Sonuç sözel değişkeninin bulanık kümelerini $\overline{R^a_r}$ değeriyle tetikle
    $R^o_r \leftarrow$ Sonuç sözel değişkeninin maksimum tetiklenen bulanık kümesi
  end for
  return R
end procedure

Şekil 2.5. E2E kural üretme sözde kodu

Şekil 2.5'te 16. satırda sözel değişkenlere ait özelleştirilmiş bulanık küme tanımları yapılmaktadır. Satır 17'de $\vec{C}$, öznitelik c için verideki tüm kayıtları göstermektedir. Ardından özniteliği gösteren sözel değişken (fv) ile $\vec{C}$, SkorAyarla() fonksiyonuna girdi parametresi olarak sunulmakta ve burada ilgili sözel değişkenin tüm bulanık kümelerine, o bulanık kümenin sonuç sınıf etiketine etkisini belirten bir skor değeri atanmaktadır.

Tüm sözel değişkenler için bulanık kümelere ait skor ataması yapıldıktan sonra Şekil 2.5'te 21. satırda bulanık kuralların öncül kısmını gösteren 2 boyutlu R^o matrisi oluşturulur ve sözel değişkenlerin alabilecekleri tüm bulanık küme kombinasyonlarına göre değer atanır. Bu matrisin satır sayısı toplam öznitelik sayısına sütun sayısı toplam kural sayısına eşittir. Üretilecek kuralların sayısı sözel değişkenlerdeki bulanık kümelerin sayılarının çarpımına eşittir.

Şekil 2.5'te 22. satırda bu kuralların ardıl kısımlarında yer alması gereken bulanık küme için R^a tanımlanmıştır. Sonuç olarak R^a dizisinin elemanlarının tanımlanması amaçlanmaktadır. Bunun için ilk olarak R^o matrisinin her satırının karşılık geldiği bulanık kümelerin skor değerleri toplanmakta ve maksimum skor değerleri toplamına bölünmektedir. Bu işlemin sonucunda elde edilen değer ile sonuç sözel değişkeni için tanımlanmış bulanık kümelerin tümü aktif hale gelir ve maksimum tetiklenen bulanık küme R^a, dizinin ilgili elemanı olarak eklenerek R^o matrisinin o satırı için bir kural tanımı yapılmış olur. R^o matrisindeki tüm satırlar için bu döngünün tekrar edilmesi ile R^a dizisi ve sonuçta bu iki dizinin birleşimi olan kural kümesi R elde edilmiş olur.

E2E yaklaşımında sözel değişkenlerin bulanık kümelerine skor ataması için Şekil 2.5'teki SkorAyarla() fonksiyonu kullanılmıştır. Buna göre ilk olarak ilgili sözel değişkenin her bulanık kümesi için skor hesaplaması Şekil 2.5'te 10. Satırdaki gibi yapılır. Burada bir bulanık kümenin öncesindeki ve sonrasındaki bulanık kümelerle olan kesişim noktalarının arasında kalan aralığa denk gelen veri noktalarının pozitif sınıfa ait olma olasılığı skor değeri olarak değerlendirilmektedir. Bu değerler, ilgili sözel değişkenin tüm bulanık kümeleri için hesaplandıktan sonra toplamı 1 değerine eşit olacak şekilde normalize edilir (satır 7) ve ilgili bulanık kümeye ait nihai skor değerleri elde edilir.

E2E kural üretme yöntemi verideki kayıtları tek tek ele almak yerine belirli aralıklardaki istatistiksel dağılımı bulanık küme skorlarının hesaplanmasında kullanarak, veri üzerinden yüksek maliyetli en iyileştirme süreçleri gerçekleştirmeyen basit bir kural oluşturma çözümü sunmaktadır. Buradaki skor değerlerinin uzman tarafından oluşturulması halinde ise bu yöntem veriden tamamen bağımsız hale gelmektedir. Bu tez çalışmasında skor değerleri veriden hesaplanarak elde edilmiştir.

2.4. Metin Özetlemede Kullanılan Veri Kümeleri

Otomatik metin özetleme, hangi şekilde kullanılıyor olursa olsun kaynak metne ve hedeflenen bir özet metne ihtiyaç duymaktadır. Bu özet metnin insan tarafından hazırlanmış olması beklenir. Bu ihtiyaç hem denetimli bir öğrenme için metin özetleme yöntemlerinin uygulanabilirliği açısından hem de probleme denetimsiz bir yaklaşım sunuluyor bile olsa ortaya çıkan özet metnin değerlendirilmesi açısından gerekmektedir. Ancak böyle bir derlem oluşturmak önemli miktarda işgücüne ihtiyaç duyduğundan oldukça maliyetlidir. Bu yüzden literatürde, bu ihtiyacı dikkate alarak derlenmiş yeterli sayıda metin özetleme veri kümesi bulunmamaktadır. Ancak, hedef özetleri bulundurmayan veri kümelerinde bir takım varsayımlarla hedef metin özetleri oluşturularak bu özet metinlere yakın özetler üretecek model çalışmaları yapılmıştır [87, 88]. Ancak varsayımlarla oluşturulmuş hedef özetlerin kaynak metni ne derece temsil ettiğine yönelik doğrulama çalışmaları yeterli düzeyde yapılmamıştır.

2.4.1. DUC (Document Understanding Conference)

Document Understanding Conference (DUC) veri kümesi, Amerika Birleşik Devletleri, Ulusal Standartlar ve Teknoloji Enstitüsü, Bilgi Teknolojileri Laboratuvarı desteğiyle 2001’den 2007’ye kadar her yıl çeşitli özetleme çalışmaları için hazırlanmış veri kümeleri bütünüdür [89]. DUC veri kümeleri belirli konularda yazılmış İngilizce haber makalelerinden oluşur.

Çizelge 2.2’de her yıla ait DUC verisinde bulunan dokümanların çeşitli özellikleri sunulmuştur.

Çizelge 2.2. Metin özetlemede kullanılan DUC veri kümelerinin temel özellikleri

Veri kümesinin ismi	Metin özetleme görevi	Veri kümesindeki başlık sayısı	Başlıktaki doküman sayısı	Hedef soyutlamalı özet	Hedef çıkarmalı özet
DUC2001	Genel metin özetleme	60	~10	50,100,200,400 kelime	100 kelime
DUC2002	Genel metin özetleme	59	~10	10,50,100,200 kelime	200,400 kelime
DUC2003	Çok kısa metin özetleme	30	~10	10 kelime	-
DUC2003	Kısa metin özetleme	30	~22	100 kelime	-
DUC2004	Çok kısa metin özetleme	50	~10	75 karakter	-
DUC2004	Kısa metin özetleme	50	~10	665 karakter	-
DUC2005	Soru cevaplama	25-50	-	-	250 kelime
DUC2006	Soru cevaplama	50	~25	-	250 kelime
DUC2007	Soru cevaplama	45	~25	-	250 kelime
DUC2007	Güncelleyici özetleme	10	~25	100 kelime	250 kelime

Burada ilk sütun veri kümesinin yayınlandığı yıla bağlı olarak tanımlanmış olan ismidir. Ardından ilgili veri kümesinin hangi metin özetleme görevi için tasarlandığı bilgisi ikinci sütunda yer almaktadır ve birden fazla görev için uygulanabilir olan veri kümelerine ait özellikler ayrı satırlarda gösterilmektedir. DUC veri kümelerinin en güçlü özelliği çoklu doküman özetlemeye uygun derlenmiş olmasıdır. Bunun için her veri kümesinde değişen sayıda başlık bulunur ve bu başlık altında tanımlanmış çok sayıda doküman üzerinden özetleme görevi yerine getirilir. Çizelge 2.2'de ilgili veri kümesinin barındırdığı başlık sayısı ve her başlıkta ortalama kaç dokümanın yer aldığı bilgisine de yer verilmiştir.

DUC, metin özetleme probleminin en yaygın kullanılan veri kümesidir. Bu durumun sebebi, kaynak dokümanların yanında çok sayıda katılımcı tarafından hazırlanmış soyutlamalı ve çıkarmalı hedef özetleri de barındırıyor olmasıdır. Çizelge 2.2'de bu özetlerin kelime veya karakter tabanlı uzunluklarına yer verilmiştir. Çizelgede sunulduğu gibi çıkarmalı metin özetleme problemi için en uygun DUC veri kümesi DUC2002 ve DUC2001'dir. DUC2002 veri kümesi, iki farklı boyutta çıkarmalı hedef özet barındırdığından literatürde çıkarmalı metin özetlemede en çok başvurulan veri kümesidir. DUC2005, DUC2006 ve DUC2007 veri kümeleri için ise, önceden tanımlı sorulara cevap olacak en uygun cümleyi işaretlemeye dayalı çıkarmalı hedef özetleri geliştirilmiş olduğu için genel metin özetlemede değil, sorgu tabanlı metin özetleme ve soru cevaplama araştırma alanlarında sıklıkla kullanılmaktadır [90–92].

2.4.2. CNN/DailyMail

CNN/DailyMail, DUC veri kümesinden sonra metin özetlemede sıklıkla kullanılmakta olan büyük boyutlu bir özetleme veri kümesidir [93, 94]. Bu veri kümesinde, kaynak dokümanlar ortalama 39 cümleden oluşan haber makaleleridir. Hedef özet dokümanlar yaklaşık 56 kelime ve ortalama 3.75 cümleden oluşan vurgu cümleleridir. Bu vurgu cümleleri kaynak dokümandan seçilen cümleler olmadıkları ve yeni birkaç cümleden oluştukları için soyutlamalı hedef özetler olarak değerlendirilmektedir. CNN/DailyMail veri kümesi 287226 eğitim, 13368 doğrulama ve 11490 test olmak üzere toplam 312084 adet doküman-özet çiftine sahip olması nedeniyle özellikle soyutlamalı özetleme uygulamaları için makine öğrenmesi yöntemlerinin geliştirilmesinde sıklıkla kullanılmaktadır [3, 93–95].

2.4.3. SciSumm

SciSumm, hesaplama dilbilimi alanında akademik yayınlar içeren bir özet veri kümesidir. Bu veri kümesinin, yayınlanma yılına göre SciSumm2016, SciSumm2017, SciSumm2018 olarak üç sürümü vardır ve her sürüm kendisinden önceki sürümleri de kapsamaktadır. Veri kümesinde, kaynak doküman olarak değerlendirilen akademik makaleler, her bir referans yayın için yaklaşık 10 adet olmak üzere alıntı makaleler ve hedef özetler bulunmaktadır. SciSumm2016 için 20, SciSumm2017 için 40 ve SciSumm2018 için 60 adet kaynak makale toplanmıştır. Veri kümelerinde, makalenin yazarları tarafından yazılan soyutlamalı özet, alıntı makalelerin kaynak makaleye yaptığı atıf cümlelerinden oluşan topluluk özeti ve

okuyucu soyutlamalı özetler olmak üzere üç tür hedef özet bulunmaktadır. Bu üç tür özet, makalelerin önemli bilgilerini içeriyor olsa da, çıkarmalı özetlemede her zaman başarılı sonuç vermeyebilir. Çünkü bu özetlerin üçü önemli bilgilerin harmanlanarak yeni bir metin yazılmasıyla oluşturulmuştur. Bunun yanında, topluluk özetleri makaledeki önemli cümleleri seçmekten çok atıfta bulunan makalelerin önem verdikleri yargıları desteklemeye odaklandığı için genel bir hedef özet niteliğinde değil, sorgu tabanlı soru cevaplama şeklindeki çalışmalarda kullanılabilecek yapıdadır.

2.4.4. WikiHow

WikiHow veri kümesi, sanat, eğlence, bilgisayar, elektronik gibi çeşitli alanlarda yordamsal bir görevi anlatan yazılar içermektedir. Her yazı "nasıl yapılır…" ile başlayan bir başlık ve makalenin kısa bir tanımından oluşur. WikiHow içerisinde iki tür yazı vardır: ilki, tek yöntemli görevleri açıklarken, ikincisi, bir görev için farklı yöntemlerin birden çok adımını açıklamaktadır. Veri kümesi oluşturulurken 20 farklı başlık altında 142783 adet metin elde edilmiştir. Bu yazılar gazeteciler tarafından yazılmakta ve gazetecilik stilini kullanmaktadır. Gazeteciler genellikle bir metnin açılış paragraflarında bir hikâyenin en önemli, ilginç veya dikkat çekici öğelerinden bahsedip, daha sonra detaylar ve arka plan bilgilerini ekleyerek ters piramit stilini [89] takip ederler. Bu yazım stilinden yola çıkılarak hazırlanan WikiHow yazılarının hedef özetleri cümlenin yazıdaki pozisyonuna bağlı olarak kural tabanlı oluşturulmuş ve alt başlıkların ilk cümlelerinden ve yazının ilk üç cümlesinden olmak üzere iki tür çıkarmalı hedef özet elde edilmiştir.

2.4.5. Açıklamalı Enron konu satırı

Açıklamalı Enron konu satırı veri kümesi, Enron veri kümesinin otomatik metin özetleme için genişletilmiş bir sürümüdür [96]. Özgün Enron veri kümesi, 150 kullanıcıya ait 517401 elektronik posta iletisini içermektedir. Bu veri kümesi konu satırı hedef özet olarak alınarak metin özetleme probleminde kullanılabilir hale getirilmiştir. Konu satırının tek cümleden veya kelime kümesinden oluşması uygulanabilirliği açısından kısıtlayıcı bir özellik olarak değerlendirilmiştir. Bu nedenle, açıklamalı Enron veri kümesinde, elektronik postalara ait üçer adet konu satırı bağımsız üç okur tarafından oluşturulmuş ve içerisinde dört cümle (veya söz öbeği) bulunan hedef özetler elde edilmiştir. Ancak, bu özetlerin de kaynak metinden doğrudan seçilen cümlelerden oluşmaması çıkarmalı metin özetlemeye

uygulanabilirliğini güçleştirmektedir. Bu veri kümesi genel olarak başlık ve vurgu cümlesi oluşturmayı hedefleyen çalışmalarda kullanılmaktadır.

2.4.6. Mühendislik bilgisi veri kümesi

Mühendislik bilgisi veri kümesi, 21 adet mühendislik dergisinden toplanmış olan akademik yayınlara ait toplam 188 doküman-özet çiftini içermektedir. Hedef özet metinleri, okuyucular tarafından soyutlamalı özet tipinde hazırlanmıştır ve ortalama üç cümle içermektedir [97].

2.4.7. BigPatent

BigPatent veri kümesi Amerika Birleşik Devletler Patent ve Marka Ofisi'ndeki patent dokümanlarının bir alt kümesi kullanılarak derlenmiştir [98]. Veri kümesi, patent dokümanlarını ve insan tarafından yazılmış soyutlamalı özetleri içermektedir. Veri kümesinde 1 341 362 adet doküman mevcuttur ve insanlar tarafından yazılmış olan soyutlamalı özetlerde ortalama 3.5 cümle bulunmaktadır. Bu veri kümesi soyutlamalı metin özetleme çalışmalarına uygundur ve cümlelerin önem dereceleriyle veya özete alınmaya değer olup olmadıklarıyla ilişkili bilgi sunmamaktadır.

2.4.8. Endonezya haber sitesi

Endonezya haber sitesinden alınmış 100 doküman-özet çiftinden oluşan bu veri kümesinde her dokümanda en az 6, en çok 28 olmak üzere ortalama 15 cümle bulunmaktadır [61]. İnsanlar tarafından hazırlanmış özetlerde ise en az 3, en fazla 7 olmak üzere ortalama 5 cümle yer almaktadır. Bu veri kümesi, insan özetlerinin çıkarmalı metin özetlemeye uygun olarak hazırlanmıştır. Ancak Endonezya dilinde olması nedeniyle performans karşılaştırmalarında kullanılamadığından literatürde kullanımı yaygın değildir.

2.4.9. TeMario

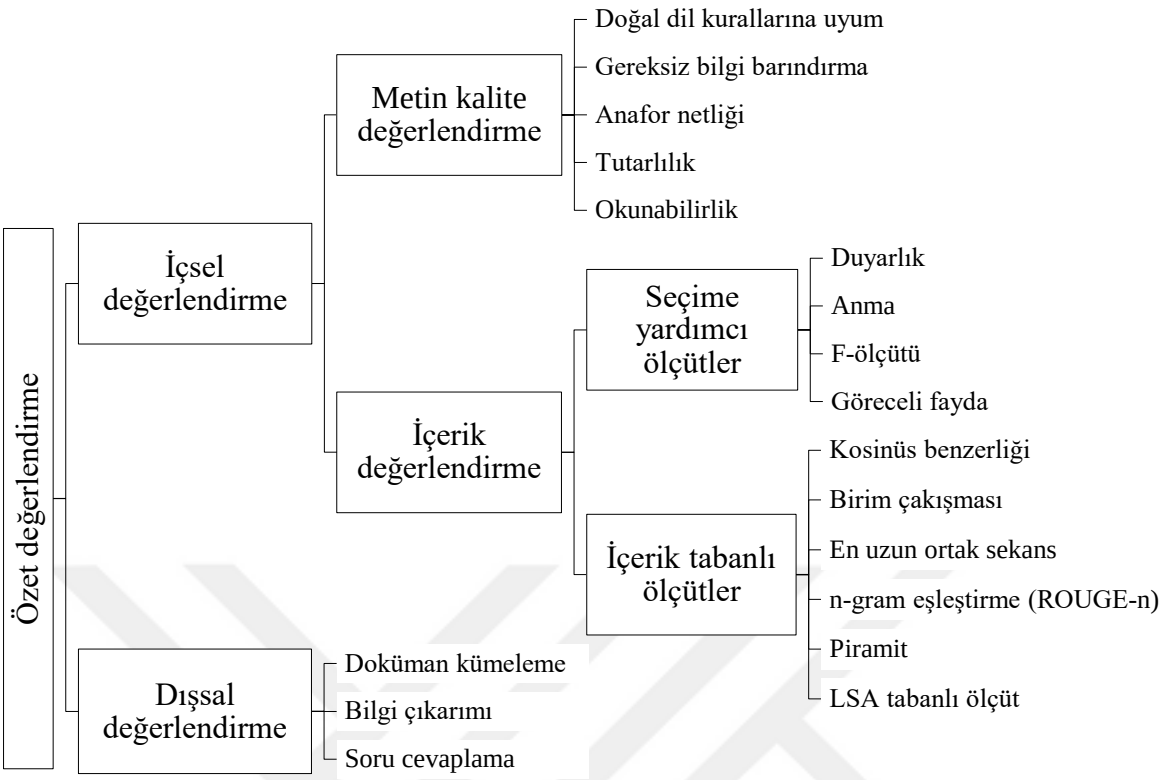
TeMario veri kümesi, 100 köşe yazısından oluşmaktadır ve ortalama 614 kelime olmak üzere toplam 61 412 kelime içermektedir [88]. Bu yazıların 60'ı Folha de São Paulo Gazetesi'nden, 40'ı ise Brazil Gazetesi'nden temin edilmiştir. Veri kümesinde, bu dokümanlara ait soyutlamalı özetler ve ideal çıkarmalı özetler de bulunmaktadır.

Dokümanların soyutlamalı özetleri, ortalama 193 kelimeden oluşmaktadır ve insan tarafından yazılmıştır. Bunun yanı sıra ideal çıkarmalı özetler ise, ortalama 232 adet kelimeden oluşan özet ile dokümanın kosinüs benzerliğini en yüksek oranda sağlayan dokümanlardır.

2.5. Özet Metin Değerlendirme Yöntemleri ve Kullanılan Ölçütler

Bir özet metni değerlendirmek için içsel ve dışsal değerlendirme olmak üzere iki değerlendirme türünden birisi kullanılır [99]. İçsel değerlendirme metnin içeriği ile ilgilenirken dışsal değerlendirme metnin metin madenciliği uygulamalarındaki veya metinden bilgi erişimi uygulamalarındaki davranışını ölçmeye çalışır. Bir dokümanın orijinali ile özetinin değerlendirme işlemlerine ayrı ayrı girdi olarak verilmesi durumunda elde edilen sonuçlar birbirine mümkün olduğu kadar benzer olmalıdır. Steinberger tarafından metin özetlerini değerlendirmeye yönelik yaklaşımların derlenmiş taksonomisi Şekil 2.6'da sunulmuştur [99].

İçsel değerlendirme yöntemlerini, metin kalite değerlendirme ve içerik değerlendirme olmak üzere iki kısımda incelemek mümkündür. Metin kalite değerlendirme yöntemleri, otomatik elde edilen özetin doğal dilin yapısına ve kurallarına uygunluğunu, gereksiz bilgi içermemesini ve ifadelerin doğruluğunu insan gözlemine dayalı olarak belirlemektedir.



Şekil 2.6. Özet değerlendirme yöntemleri

İçerik değerlendirme bir metnin orijinalinin özeti ile benzerliğine odaklanır. Eğer özetleme cümle çıkarma şeklinde yapılırsa ve problem cümle temelli bir sınıflandırma problemi olarak ele alınırsa, yaygın kullanılan duyarlılık (precision), anma (recall) ve f-ölçütü başarıyı değerlendirmede kullanılabilmektedir. Ayrıca, cümle yerine terim odaklı bir önem derecesi belirlenerek yapılan çıkarmalı özetlemede, dokümanın özete benzerliği terim ve/veya n-gram temelli olarak ölçülebilmektedir.

Özet değerlendirmede en sık kullanılan yöntem, Eş. 2.21'de verilen hedef özetin otomatik özetlerle n-gram eşleştirmesine dayanan özet değerlendirmede anma odaklı yardımcı metodudur (Recall-Oriented Understudy for Gisting Evaluation-ROUGE).

$$ROUGE-n = \frac{\sum_{M \in H} \sum_{g_n \in M} C_m(g_n)}{\sum_{M \in H} \sum_{g_n \in M} C(g_n)} \tag{2.21}$$

ROUGE, kelime kümelerinin (n-gram) içerdiği kelime sayısına göre ROUGE-n olarak isimlendirilir. Buna göre ROUGE-1 kelime (1-gram) kesişimini, ROUGE-2 iki kelimeli sözcük öbeklerinin (2-gram) kesişimini, ROUGE-L en uzun ortak sıralı söz öbeğinin

kesişimini, ROUGE-S atlamalı-sözcük çiftlerinin ve ROUGE-SU atlamalı sözcükleri ve sözcük çiftlerinin birlikteliğini inceler.

Eş. 2.21'de hedef özetlerde yer alan cümleler dizisi (H) olmak üzere, g_n n kelimeli sözcük kümesi, $C(g_n)$ hedef özette g_n kelime kümesinin bulunma sıklığı ve $C_m(g_n)$ kelime kümesi g_n'in hedef özette ve otomatik özette bulunma sıklığıdır.

Hedef özet 1: *The dog was under the bed in the baby's room.*
Hedef özet 2: *The dog was under the baby's bed.*
Otomatik özet: The dog was under the bed.

Hedef özetteki ikili söz öbekleri:
Hedef özet 1: *The dog, dog was, was under, under the, the bed, bed in, in the, the baby's, baby's room.*
Hedef özet 2: *The dog, dog was, was under, under the, the baby's , baby's bed.*
Otomatik özet: *The dog, dog was, was under, under the, the bed.*

$$ROUGE-2 = \frac{5+4}{9+6} = 0.60$$

Şekil 2.7. ROUGE hesaplama örneği

Eş. 2.21'deki ROUGE değerlendirme ölçütünden yola çıkılarak, iki hedef özetten ve bir otomatik özetten oluşan bir doküman için ROUGE-2 değerinin hesaplamasına ait bir örneğe Şekil 2.7'de yer verilmiştir. Burada, $n = 2$ olmak üzere hedef özetlerin ikili kelime kümeleri ile otomatik özetin ikili kelime kümeleri içerisinden eşleşenler farklı renklerle gösterilmiştir. Buna göre hedef özetlerin birincisinin ve ikincisinin otomatik özetle kesişim sayısı sırasıyla 5 ve 4'tür. ROUGE ölçümünde anma değeri üzerinden bir hesap yapıldığından eşitliğin paydası hedef özetteki ikili kelime kümelerinin sayısının toplamıdır. Bu değer ilk hedef özet için 9, ikinci hedef özet için ise 6'dır. ROUGE metriğinde ilgili değerlerin yerine konulması sonucunda bu örnek için elde edilen ROUGE-2 değeri 0.60 olarak hesaplanmıştır.

3. LİTERATÜRDE YER ALAN METİN ÖZETLEME ÇALIŞMALARI

Çıkarmalı metin özetleme problemi, metnin içeriği için önemli olan ve metin içeriğini mümkün olduğunca yansıtan cümlelerin seçilmesini amaçlar. Bunun için metindeki cümlelerin önem derecelerinin belirlenmesi gerekmektedir. Ancak önem kavramı göreceli bir kavram olup, cümlenin önemini belirtmek için farklı bakış açılarıyla farklı değerlendirmeler yapılabilir.

Okuyucu için çıkarılan özetlerde beklenti, kaynak metinlerin içerisindeki gereksiz bilginin elenip okuyucuya metnin ana hatlarını doğru bir akışla sunmaktır. Bilgi çıkarımı açısından değerlendirildiğinde ise özetin bilgi çıkarım veya metin madenciliği uygulamalarında kaynak metnin davrandığı şekilde davranması beklenmektedir. Bu iki temel nokta göz önüne alındığında, çıkarmalı metin özetleme için cümlenin önemi, kaynak metindeki bilginin o cümlede ne oranda yer aldığıyla ilgilidir. Çünkü gerek okuyucular açısından gerekse bilgi çıkarım süreçleri açısından özetteki bilginin kaynak metindeki bilgiyi mümkün olduğunca kapsaması gerekmektedir.

Metinlerin türü ve karakteristik özellikleri ile metni özetlemedeki niyetin çeşitliliğine göre önem tanımı da değişmektedir. Genel bir özetin çıkarılabilmesi için bu faktörlerin etkisizleştirilmesi gerekmektedir. Mevcut metin özetleme çalışmaları göz önünde bulundurulduğunda özette yer alması gereken cümlelerin önem açısından derecelendirilmesi için mevcut modellerin üzerinde çalıştığı iki tip öğrenme kanalı vardır. Bunlar sözdizimsel ve anlamsal özniteliklerdir. Ayrıca, bu çalışmada vurgulanmak istendiği şekilde önem derecelendirme noktasında bu iki tip özniteliğin bütünleştirilmesini bir oranda ele almış az sayıda çalışma mevcuttur.

Bu bölümün alt başlıklarında bu üç öznitelik tipi üzerinden şimdiye kadar yapılmış olan çıkarmalı metin özetleme çalışmalarına, çalışmaların hangi veri kümeleri üzerinde ne oranda başarılı özetler ürettiklerine yer verilmiştir.

3.1. Sözdizimsel Öznitelikler ile Çıkarmalı Metin Özetleme Çalışmaları

İnsanlar tarafından bir metindeki önemli cümlelerin seçilmesi istendiğinde kişiler tarafından özellikle dikkat edilen bazı hususlar vardır. Muhakkak cümlelerin neyi ifade ettikleri ve

metni ne oranda yansıttıkları üzerinden özetler üretilir. Bunun yanında, bazı şekilsel ve sözdizimsel özellikler de metinde vurgulanan cümlelerin bulunmasına yardımcı olur. Aslında bu durum, özeti çıkaran okurun değil, metni yazan kişinin kontrolündedir ve yazım stili ile bağlantılıdır.

Metnin karakteristik özelliklerinin metni meydana getiren cümlelerin önem derecelerine bir etkisinin olup olmadığı tartışmaya değer bir konudur. Örneğin bir haber makalesinde ilk cümleler, yaşanan olayla ilgili vurgulu cümleleri içerecek şekilde yazılır ancak bu durum tüm metin tipleri için genel bir yargı oluşturamaz; elektronik postalar gibi yapısal olmayan metinlerde bu önerme geçerli olmamaktadır. Buna göre sözdizimsel öznitelikler, metindeki karakteristik özellikleri, metin özetleme problemi için anlam ifade edecek özellikler üzerinden tanımlanmasından elde edilir. Sözdizimsel öznitelikleri matematiksel olarak tanımlayan, cümlelere bu sözdizimsel özniteliklere göre derecelendirme yapan ve bu özniteliklerin birlikteliğiyle cümlelerin önemine karar veren makine öğrenmesi yöntemleri, çıkarmalı metin özetlemede en yaygın kullanılan yöntemler olmuştur.

Literatürde yer alan makine öğrenmesi yaklaşımları, bu yaklaşımların kullandıkları sözdizimsel özniteliklerin listesi, kullandıkları veri kümeleri, geliştirilen özetleri hangi ölçütlerle değerlendirip ne oranda başarı elde ettikleri, Naive Bayes sınıflandırıcı, karar ağaçları, destek vektör makineleri, yapay sinir ağları ve bulanık kural tabanlı uzman sistemler için sırasıyla Çizelge 3.1, Çizelge 3.2, Çizelge 3.3, Çizelge 3.4 ve Çizelge 3.5'te sunulmuştur.

Çizelge 3.1. Naive Bayes metodu ile çıkarmalı metin özetleme

Ref.	Kullanılan sözdizimsel öznitelikler	Kullanılan veri kümesi	Değerlendirme ölçütü	Başarım oranı
[97]	Uzunluk, pozisyon, özel isimler, tamlamalar	Mühendislik bilgisi veri kümesi	Anma	%43.0
[100]	TF-IDF, pozisyon, ilk cümleye benzerlik	Çalışmaya özel veri kümesi (20 doküman)	Anma Duyarlılık Tutarlılık Okunabilirlik	%83.0 %85.0 %90.0 %100.0
[58]	Pozisyon, özel isimler, TF, sebep-sonuç ilişkisi, cümle-cümle ilişkisi	Çalışmaya özel veri kümesi (9 patent dokümanı)	Anma Duyarlılık	%52.3 %55.0

Çizelge 3.1. (devam) Naive Bayes metodu ile çıkarmalı metin özetleme

Ref.	Kullanılan sözdizimsel öznitelikler	Kullanılan veri kümesi	Değerlendirme ölçütü	Başarım oranı (%)
[61]	Pozisyon, uzunluk, TF, anahtar kelime, tamlamalar, başlık, cümle-cümle ilişkisi, nümerik veri, özel isimler, TF-IDF	Endonezya haber sitesi	Anma Duyarlılık	71.0 73.0
[66]	TF-ISF, uzunluk, pozisyon, başlık, anahtar kelime, cümle-cümle ilişkisi, ağaç derinliği, kök düğüme uzaklık, ana konsepte uyum, özel isimler, anaforlar, tamlamalar	TIPSTER [101] (30 dokümanın insan tarafından tanımlanmış çıkarmalı özet metni vardır)	Duyarlılık/ Anma	37.5
[102]	Pozisyon, tematik kelimeler, uzunluk, alıntı bulundurma sıklığı, özel isimler	HOLJ (House of Lords)	F-ölçütü	51.7

Çizelge 3.1'deki çalışmalar otomatik metin özetlemenin makine öğrenmesi yöntemleri arasındaki öncül çalışmalarındandır. Burada problem standart sınıflandırma problemi olarak değerlendirilmiş ve ortaya çıkan özetlerin başarısı anma ve duyarlılık ölçütlerine göre ölçülmüştür.

Çizelge 3.1'de Naive Bayes sınıflandırıcının otomatik metin özetlemeden elde edebildiği en yüksek anma değeri %83.0 olduğu görülmektedir. Ancak bu çalışmanın kullandığı veri kümesi çalışmaya özgü olduğundan, buradaki bulgu adil bir kıyaslama yapabilmek için yeterli değildir.

Sözdizimsel öznitelik kümesinin otomatik metin özetlemede kullanımında karar ağacı ile cümle sınıflandırmaya dayanan çalışmalar, burada kullanılan veri kümeleri, ortaya çıkan özetleri değerlendirme ölçütleri ve bu ölçütlere göre oluşan başarı yüzdeleri Çizelge 3.2'de sunulmuştur.

Çizelge 3.2. Karar ağaçları ile çıkarmalı metin özetleme

Ref.	Kullanılan sözdizimsel öznitelikler	Kullanılan veri kümesi	Değerlendirme ölçütü	Başarım oranı (%)
[58]	Pozisyon, özel isimler, TF, sebep-sonuç ilişkisi, cümle-cümle ilişkisi	Çalışmaya özel veri kümesi (9 patent dokümanı)	Anma Duyarlılık	54.8 39.0
[102]	Pozisyon, tematik kelimeler, uzunluk, alıntı bulundurma, özel isimler	HOLJ (House of Lords)	F-ölçütü	59.7
[68]	Uzunluk, başlık, tematik kelimeler, nümerik veri, cümle-cümle ilişkisi, isim tamlamaları	DUC2002	Sıkıştırma oranı Anma Duyarlılık F-ölçütü	(%10-30) 53.9-58.8 67.7-68.9 60.0-63.4
[66]	TF-ISF, uzunluk, pozisyon, başlık, anahtar kelime, cümle-cümle ilişkisi, ağaç derinliği, kök düğüme uzaklık, ana konsept bilgisine uyum, özel isimler, anaforlar, tamlamalar	TIPSTER [101] (30 dokümanın insan tarafından tanımlanmış çıkarmalı özet metni vardır)	Duyarlılık/ Anma	28.01

Çizelge 3.2'de karar ağacı algoritmalarının metin özetlemede cümle sınıflandırma başarısı sunulmuştur. Burada kullanılan özet değerlendirme yöntemleri standart sınıflandırıcı değerlendirme yöntemleridir ve çalışmaların çoğu az sayıda doküman içeren çalışmaya özel veri kümeleri üzerinden değerlendirilmiştir.

Çizelgede, Hannah tarafından yapılan çalışmada karar ağacının DUC2002 veri kümesine uygulanmasında %10 sıkıştırma oranına göre F-ölçütünün %60.0'a eriştiği görülmektedir [68]. Bu çalışmada kullanılan sözdizimsel öznitelikler, cümle uzunluğu, başlık, tematik kelimeler, nümerik veri, cümle-cümle ilişkisi ve isim tamlamalarıdır. Çalışmanın ROUGE analiz sonuçları bulunmamaktadır.

Çizelge 3.3. Destek vektör makinesi ile çıkarmalı metin özetleme

Ref.	Kullanılan sözdizimsel öznitelikler	Kullanılan veri kümesi	Değerlendirme ölçütü	Başarım oranı (%)
[102]	Pozisyon, tematik kelimeler, uzunluk, alıntı bulundurma, özel isimler	HOLJ (House of Lords)	F-ölçütü	60.6
[103]	Pozisyon, dokümandaki cümle sayısı, kümedeki cümle sayısı, söylem işaretçisi, özel isimler, TF (kelime olasılığı tabanlı)	DUC2005	ROUGE-2 (A)	7.80
[104]	Pozisyon, uzunluk, büyük-küçük harfle yazım, cümle-cümle ilişkisi, logaritmik-olasılık	DUC2001	ROUGE-2 (A) ROUGE-1 (F)	44.0 38.5
[105]	Pozisyon, uzunluk, büyük-küçük harfle yazım, cümle-cümle ilişkisi, logaritmik-olasılık	DUC2001	ROUGE-2 (A) ROUGE-1 (F)	46.0 39.4
[106]	Sorgu bağımlı öznitelikler: Sorgu eşleşmesi (TF), özel isim eşleşmesi Bağımsız öznitelikler: TF-IDF, özel isim, pozisyon	DUC2005 DUC2006 DUC2007	ROUGE-1 / ROUGE-2	36.9/7.0 39.2/8.8 43.4/11.1

Çizelge 3.3'te destek vektör makinesinin çıkarmalı metin özetleme başarısı çoğunlukla ROUGE ölçümleri açısından sunulmuştur. Burada DUC veri kümesinin farklı sürümlerinin uygulamalarından elde edilen ROUGE değerlerine göre ROUGE-2 anma değerleri %7.80, %46.0 aralığında değişmektedir. Wu ve diğerlerininin DUC2001 veri kümesi üzerinde yaptıkları çalışmada pozisyon, uzunluk, büyük-küçük harfle yazım, cümle-cümle ilişkisi ve logaritmik-olasılık özniteliklerinin cümleler için hesaplanması sonucunda destek vektör makineleri ile elde edilen özetlerin ROUGE-1 F-ölçütü %39.4 ve ROUGE-2 anma değeri %46.0 olarak ölçülmüştür [105].

Çizelge 3.4. Yapay sinir ağları ile çıkarmalı metin özetleme

Ref.	Kullanılan sözdizimsel öznitelikler	Kullanılan veri kümesi	Değerlendirme ölçütü	Başarım oranı (%)
[58]	Pozisyon, özel isimler, TF, sebep-sonuç ilişkisi, cümle-cümle ilişkisi	Çalışmaya özel veri kümesi (9 patent dokümanı)	Anma Duyarlılık	53.0 50.5
[88]	Uzunluk, pozisyon, edat cümleleri, TF, TF-ISF, tematik kelimeler	TeMario	Anma Duyarlılık	40.3 43.8
[1]	Dokümandaki cümle sayısı, uzunluk, özel isimler, tamlamalar (isim tamlaması, sıfat tamlaması, eylem öbeği, edat), pozisyon, şebeke bağlantı sıklığı, cümle-cümle ilişkisi, TF-ISF, TF, başlık	DUC2002	ROUGE-1 (A) ROUGE-2 (A) ROUGE-L (A)	6.3 1.6 6.6

Çizelge 3.4'te yapay sinir ağlarının metin özetleme başarısı sunulmuştur. Yapay sinir ağları, sözdizimsel öznitelikler üzerinde yapılan metin özetleme çalışmalarında yaygın kullanılan bir yöntem olmamakla birlikte, bu yöntemlerden elde edilmiş olan en yüksek standart anma değeri %53.0 olarak rapor edilmiştir. DUC2002 veri kümesindeki deneylerden elde edilmiş olan ROUGE-1 anma değeri ise %6.3 olarak sunulmuştur.

Bu tez çalışmasında, makine öğrenmesi yöntemlerinin metin özetleme başarısının yanında bulanık kural tabanlı sistemler üzerine yapılan metin özetleme çalışmaları da sunulmuştur. Çalışmaların genelinde, sistemlerin karar vermede kullandığı bulanık kurallar uzmanlar tarafından elle üretilmiştir. Kuralların üretilmesindeki bir diğer alternatif yöntem ise veri kümesindeki belirli özellikleri kullanmak ve sisteme genelleştirme kabiliyeti kazandırmaktır. Ancak her iki durumda da sistemler, veri üzerinden eğitim ve en iyileştirme süreçlerine maruz kalmamaktadır. Bu noktada veri kümesinden öğrenmeyen bu yaklaşımların metin özetleme problemine sunacağı çözüm makine öğrenmesi yöntemlerinden farklılık gösterir ve değerlidir. Çizelge 3.5'te bulanık kural tabanlı sistemlerin metin özetleme çalışmalarında kullanımına örnekler sunulmuştur.

Çizelge 3.5. Bulanık kural tabanlı sistemler ile çıkarmalı metin özetleme

Ref.	Kullanılan sözdizimsel öznitelikler	Kullanılan veri kümesi	Değerlendirme ölçütü	Başarım oranı (%)
[107]	Cümle-cümle ilişkisi, başlık, TF-ISF, anahtar kelimeler, uzunluk, pozisyon, özel isimler, nümerik veri, ilk cümleye benzerlik	DUC2002	ROUGE-1 (A)	46.6
[108]	Cümle-cümle ilişkisi, uzunluk, isim tamlaması, eylem öbeği	DUC2002	ROUGE-1 (A)	12.8
[1]	Dokümandaki cümle sayısı, uzunluk, özel isimler, tamlamalar (isim tamlaması, sıfat tamlaması, eylem öbeği, edat), pozisyon, şebeke bağlantı sıklığı, cümle-cümle ilişkisi, TF-ISF, TF, başlık	DUC2002	Sıkıştırma oranı: ROUGE-1 (A) ROUGE-2 (A) ROUGE-L (A) Sıkıştırma oranı: ROUGE-1 (A) ROUGE-2 (A) ROUGE-L (A)	(%95) 51.6 41.4 49.5 (%90) 66.3 59.1 66.7
[31]	Başlık, uzunluk, TF-ISF, pozisyon, cümle-cümle ilişkisi, özel isimler, tematik kelimeler, nümerik veri	DUC2002	ROUGE-1 (A)	45.7
[109]	Başlık, uzunluk, TF-ISF, cümle-cümle ilişkisi, özel isimler, tematik kelimeler, nümerik veri	DUC2002	ROUGE-1 (A) ROUGE-1 (D)	49.2 47.3
[33]	TF-ISF, cümle-cümle ilişkisi, pozisyon, uzunluk	Çalışmaya özel veri kümesi (20 cümlelik 63 Portekizce doküman)	Sıkıştırma oranı ROUGE-1 (A) ROUGE-1 (D) Sıkıştırma oranı ROUGE-1 (A) ROUGE-1 (D)	(%80) 39.8 41.7 (%70) 49.6 36.6

Literatürde çıkarmalı metin özetlemede en yaygın kullanılan DUC2002 veri kümesi dikkate alındığında, çalışmalardaki sınıflandırma başarısı açısından en yüksek anma değerinin %53.0 ile yapay sinir ağı çözümünden elde edildiği görülmektedir. Öte yandan, metin özetlerinin esas alınan değerlendirme ölçütü olan ROUGE analizleri dikkate alındığında en

yüksek ROUGE-1 anma değerinin %51.6 olduğu Çizelge 3.5'te rapor edilmiştir ve bu değer bulanık çıkarsama sistemlerinden elde edilmiştir.

3.2. Anlamsal Öznitelikler ile Çıkarmalı Metin Özetleme Çalışmaları

Metindeki cümlelerin içerdiği kelimelerin anlamlarının özete etkisini belirleyebilmek için kelime özyerleşikleri kullanılmaktadır. Kelime özyerleşikleri kelimelere ait vektörel bir temsildir ve çok büyük veri kümelerinden kelimelerin anlamsal ilişkilerinin öğrenilmesinin bir sonucu olarak elde edilir. Bu özyerleşiklerin çıkarmalı metin özetlemede kullanımına ilişkin çalışmalara Çizelge 3.6'da yer verilmiştir.

Çizelge 3.6. Anlamsal öznitelikler ile çıkarmalı metin özetleme çalışmaları

Ref.	Öznitelik kaynağı	Yöntem	Kullanılan veri kümesi	Değerlendirme ölçütü	Başarım oranı (%)
[71]	GloVe	ESA & LSTM	DUC2002	ROUGE-1 (A) ROUGE-2 (A)	39.2 10.7
[72]	Word2vec	ESA & ÖSA	DUC2002	ROUGE-1 (A) ROUGE-2 (A) ROUGE-L (A)	47.4 23.0 43.5
			CNN/DailyMail	ROUGE-1 (A) ROUGE-2 (A) ROUGE-L (A)	21.2 8.3 12.0
[110]	Kelime torbası	kısıtlı Boltzmann makineleri otokodlayıcı	elektronik posta sorgu tabanlı özetleme veri kümesi	ROUGE-1 (A) ROUGE-SU (A)	50.34 25.37
[111]	Word2vec	YSA	DUC2002 (15 dokümanlı altkümesi)	ROUGE-1 (A) ROUGE-2 (A)	55.1 22.6
[112]	Word2vec	ESA	DUC2005 DUC2006 DUC2007	ROUGE-1 / ROUGE-2	37.0/6.9 40.9/9.4 43.9/11.5
[113]	Word2vec CBOW	ileri beslemeli aşırı öğrenme makinesi	DUC2006 DUC2007	ROUGE-1 (A) / ROUGE-2 (A)	34.4/5.1 37.3/1.2
[114]	Word2vec CBOW	ESA ve (cümle seçiminde) en iyileştirme modeli	DUC2002 DUC2004	ROUGE-1 (A) / ROUGE-2 (A)	51.0/26.9 40.9/10.7

Çizelge 3.6'da, anlamsal özniteliklerin hangi kelime özyerleşik türünden kaynaklandığı, cümle önem derecelendirme veya cümle sınıflandırma için başvurulmuş olan makine öğrenmesi yöntemi, kullanılan veri kümeleri, geliştirilen özetlerin hangi ölçütlerle değerlendirilip ne oranda başarı elde edildiği görülmektedir. Bu çalışmalarda, metin içerisindeki cümlelerin çok boyutlu vektörlerle ifade edilmesi sonucunda elde edilen karmaşık öznitelik kümesi üzerinde derin öğrenme yaklaşımları kullanılmaktadır.

Kelime özyerleşiklerinden cümlelerin vektörleri tanımlanmakta ve ortaya çıkan karmaşık doküman vektörlerindeki cümleler sınıflandırma problemi olarak ilgili modele girdi olarak sunulmaktadır. Bu alanda üzerinde en sık çalışılmış olan derin öğrenme yöntemleri karmaşık öznitelik uzayını yeniden yapılandıran ESA ve ÖSA'dır.

Literatürde çıkarmalı metin özetlemede en yaygın kullanılan DUC2002 veri kümesi üzerindeki çalışmalar incelendiğinde en yüksek ROUGE-1 anma değerinin %51 olduğu Çizelge 3.6'da görülmektedir [114].

3.3. Hibrit Öznitelikler ile Çıkarmalı Metin Özetleme Calışmaları

Bu tez çalışmasında, metin özetleme problemi için sözdizimsel özniteliklerin ve anlamsal özniteliklerin birlikte kullanımı önerilmiştir. Literatürde, bu tez çalışmasında önerildiği şekilde uygulanmış olan kapsamlı bir öznitelik tartışması mevcut değildir, ancak anlamsal özniteliklere birkaç sözdizimsel özniteliğin de eklenmesine yönelik öncü yaklaşımlar mevcuttur.

Anlamsal özniteliklere sözdizimsel özniteliklerin desteğiyle çıkarmalı metin özetleme performansını arttıran ilgili çalışmalara Çizelge 3.7'de yer verilmiştir. Çizelgede, bu yöntemlerin hangi kelime vektörleri üzerinde anlamsal öznitelik oluşturdukları ve hangi sözdizimsel özniteliklerle özetleme modellerini destekledikleri sunulmaktadır. Ayrıca modelin geliştirilmesindeki derin öğrenme algoritması ve performans ölçümleri de Çizelge 3.7'de sunulmuştur.

Çizelge 3.7. Hibrit öznitelik uzayı üzerinden çıkarmalı metin özetleme çalışmaları

Ref.	Sözdizimsel Öznitelikler	Yöntem	Kullanılan veri kümesi	Değerlendirme ölçütü	Başarım oranı (%)
[85]	Pozisyon Ortalama TF Ortalama küme frekansı	ESA & Maksimum örnekleme	DUC2001 DUC2002 DUC2004	ROUGE-1(A)/ ROUGE-2(A)	35.9/7.8 36.6/8.9 38.9/10.0
	Özyerleşik [115]				
[116]	TF-ISF Uzunluk Pozisyon Cümle-cümle ilişkisi Derinlik Tematik kelimeler Özel isimler Önemsiz ifadeler	Yapay sinir ağı	DUC2002	ROUGE-1(A) ROUGE-2(A) ROUGE-L(A)	38.2 22.5 27.4
	GloVe				
[15]	Pozisyon	Otokodlayıcı-doğrusal cümle derecelendirme fonksiyonu	DUC2002	ROUGE-1(A) ROUGE-2(A) ROUGE-L(A)	51.7 27.5 44.6
	Atlamalı-düşünce vektörü (Word2vec tabanlı) [117]				
[34]	Mutlak cümle pozisyonu Göreceli cümle pozisyonu	ç-GRU	DUC2002	ROUGE-1 (A) ROUGE-2 (A) ROUGE-L (A)	46.6±0.8 23.1±0.9 43.03±0.8
	Word2vec (İçerik, ilgi ve özgünlük ölçümü)		CNN/ DailyMail	ROUGE-1 (A) ROUGE-2 (A) ROUGE-L (A)	42.0±0.2 16.9±0.4 34.1±0.3

Cao ve diğerleri, cümledeki kelime özyerleşiklerinden elde edilen anlamsal özniteliklere pozisyon, ortalama terim sıklığı ve cümlelerin kümelenmesi sonucunda elde edilen ortalama küme sıklığı olmak üzere üç tip sözdizimsel özniteliğin eklenmesi ile hibrit bir cümle önem derecelendirme modeli önermişlerdir [85]. Bu modelin önemli olarak etiketlediği cümleler özete dâhil edilmeden önce o ana kadar oluşmuş olan özet dokümanına benzerliğine göre yeniden değerlendirilmekte ve özette yer almayan özgün bilginin özete dâhil edilmesi sağlanmaktadır. Jain ve diğerleri ise TF-ISF, cümle uzunluğu, cümlenin pozisyonu, cümle-cümle ilişkisi, cümlelerin kümelenmesi sonucunda elde edilen hiyerarşik yapının derinliği, tematik kelimeler, özel isimler ve belirli bağlaçlarla önceden tanımlanmış olan önemsiz

ifadelerin sözdizimsel öznitelikler olarak değerlendirildiği çalışmalarına, cümlelerin içerdiği kelimelere ait özyerleşikler üzerinden bir anlamsal özellik eklentisi yaparak öznitelik uzayını genişletmişlerdir [116]. Burada cümlelere ait anlamsal öznitelik değeri sayısal bir vektördür ve kelime özyerleşiklerinin vektörel toplamının cümledeki kelime sayısına bölümü alınarak hesaplanmaktadır. Genişletilmiş bu öznitelik kümesi, üç katmanlı bir yapay sinir ağına girdi olarak verilmekte ve sınıflandırma başarısı ölçülmektedir. Bu modelin ürettiği çıkarmalı özetlerin DUC2002 veri kümesi için ROUGE-1 anma değeri %38.2 olarak hesaplanmıştır.

Joshi ve diğerleri tarafından 2019 yılında sözdizimsel özniteliklerin anlamsal özniteliklere desteği ile çıkarmalı metin özetleme üzerine bir diğer çalışma yapılmıştır [86]. Burada özetleme görevi için anlamsal öznitelikler, word2vec kelime özyerleşikleri tabanlı atlamalı-düşünce vektörlerinin üzerinden, içerik ve özgünlük olmak üzere iki özelliği ölçmek için kullanılmaktadır. İçerik özniteliğinin tanımlanmasında kodlayıcı, temsil ve çözümleme için otokodlayıcı geliştirilmiş ve otokodlayıcının temsil katmanından elde edilen vektör içeriğe ait öznitelik olarak kullanılmıştır. Bunun yanı sıra özgünlük özniteliği, cümle özyerleşiklerinden elde edilmektedir. Pozisyon sözdizimsel özniteliği ile içerik ve özgünlük anlamsal özniteliklerinin doğrusal bir fonksiyona giriş olarak verilmesi ile cümleye ait önem derecesi elde edilmiştir ve yüksek önem derecesine sahip cümleler özete dâhil edilmiştir. DUC2002 veri kümesinde bu modelin çıktısı olan çıkarmalı özetlerden elde edilen ROUGE-1 anma değeri ise 51.7 olarak hesaplanmıştır.

Nallapati ve diğerleri tarafından 2017 yılında önerilmiş olan SummaRuNNer çıkarmalı metin özetlemede yaygın kullanıma sahip bir yöntemdir [34]. SummaRuNNer, ÖSA tabanlı hiyerarşik bir sıra modelidir. Bu modelde, cümlenin pozisyonu ile ilişkili sözdizimsel öznitelikler anlamsal özniteliklerle birlikte cümle önem derecelerinin belirlenmesinde kullanılmaktadır. Bu özelliğiyle SummaRuNNer, bu tez çalışmasında önerilen yaklaşıma en yakın modeldir. SummaRuNNer mimarisi, en altta giriş katmanı, kelime katmanı, cümle katmanı ve en üstte sınıflandırma katmanından oluşmaktadır.

Giriş katmanındaki her cümle, içerisinde barındırdığı tüm kelimelerin word2vec tabanlı kelime özyerleşiklerinin birleşiminden elde edilir ve sabit bir sıra boyutundan küçük olan cümleler sıfır vektörü ile doldurulur. Kelime katmanı kelime düzeyinde çalışır ve geçerli kelime özyerleşiklerine ve bir önceki gizli duruma bağlı olarak her bir kelime konumundaki gizli durum gösterimlerini sırayla hesaplar. Kelime katmanında ayrıca, son sözcükten ilk

sözcüğe kadar geriye doğru çalışan başka bir ÖSA (ς -GRU) kullanılır. Cümle katmanı, cümle düzeyinde çalışır ve kelime düzeyinden elde edilen çıkışın ortalamaya göre örneklenmesi sonuçlarının birleşimini giriş olarak kabul eden ς -GRU modellerinden oluşan ÖSA'dır. Bu katmandan elde edilen gizli durumlar, belgedeki cümlelerin vektörel temsiliyle cümle özyerleşiklerini göstermektedir. Tüm belgenin temsili daha sonra iki yönlü cümle düzeyinde ÖSA'nın birleştirilmiş gizli durumlarının ortalamaya göre örnekleminin doğrusal olmayan bir dönüşümü olarak modellenir. Son olarak sınıflandırma katmanında ise, her cümle ikinci bir geçişte ardışık olarak yeniden ziyaret edilmektedir. Burada bir lojistik katman bu cümlenin özete ait olup olmadığı konusunda ikili bir karar vermektedir. Bunun için özyerleşiklerden elde edilen anlamsal özniteliklerin yanında mutlak ve göreceli cümle pozisyonu da değerlendirmeye alınmaktadır. Çizelge 3.7'de SummaRuNNer'ın DUC2002 ve CNN/DailyMail veri kümelerindeki özetleme başarısı ROUGE ölçütleriyle sunulmuştur. Buna göre ROUGE-1 anma değerleri DUC2002 için %46.6 ± 0.8, CNN/DailyMail için %42.0 ± 0.2 olarak elde edilmiştir.

4. TEZ KAPSAMINDA OLUŞTURULAN VERİ KÜMESİ: SIGIR2018

4.1. SIGIR2018 Veri Kümesinin Oluşturulması ve Değerlendirilmesi

Bu tez çalışması kapsamında akademik yayınlar kullanılarak metin özetleme için yeni bir veri kümesi oluşturulmuştur. Bu amaçla, 2018 yılında gerçekleşen 41. Uluslararası ACM SIGIR Bilgi Erişiminde Araştırma ve Geliştirme Konferansı'ndan açık erişimli 125 bildirinin tam metni elde edilmiştir. SIGIR2018 konferansındaki 77 yayın 9-10 sayfalık uzun bildirilerin tümünü kapsamaktadır. Veri kümesi için toplanan diğer 48 yayın ise 4 sayfalık kısa bildirilerden oluşmaktadır.

SIGIR2018 dokümanlarını otomatik metin özetleme veri kümesine dönüştürmek için ilk olarak dokümanlardaki bölümlerde kısıtlamaya gidilmiştir. Bir akademik yayında giriş bölümünde çalışmanın üzerinde durduğu problem, amacı, kapsamı tanıtılır; geliştirilen yönteme dair ayrıntılar ve yapılan testlerin detaylarına yer verilir; son olarak da çalışmadan elde edilen çıktı ve bulgular sunulur [35]. SIGIR2018 yayınları da bu yapıyı koruyacak şekilde düzenlenmiş olan dokümanlardır.

Yayınların giriş bölümlerinde SIGIR konferanslarında kullanılan yapının ve metin akış şeklinin korunmuş olması nedeniyle dokümanların giriş bölümleri kaynak metin olarak değerlendirilmiş ve dokümanlardaki diğer kısımlar değerlendirilmeye alınmamıştır. Giriş bölümünün kaynak doküman olarak değerlendirildiği bu veri kümesinde yayınların orijinal özetleri hedef soyutlamalı özetler olarak değerlendirilmek üzere toplanmıştır. Bunun yanı sıra, yayınlar için tanımlanan anahtar kelimeler ve yayınları konferans yönetim sistemine yüklerken seçilen alan kategorileri (konsept) de veri kümesine dahil edilmiş olan yardımcı argümanlardır.

Şekil 4.1'de Wu ve diğerlerinin SIGIR2018 yayınlarına ait giriş bölümünün bir parçası örnek kaynak doküman niteliğinde sunulmuştur [118].

< GİRİŞ
DOKÜMAN= p495-wu
BAŞLIK = Learning Contextual Bandits in a Non-stationary Environment>
<s no=1 p=1 c=1 d=3> Multi-armed bandit algorithms provide a principled solution to the explore exploit dilemma, which exists in many important real-world applications such as display advertisement, recommender systems, and online learning to ran. **</s>**
<s no=2 p=1 c=2 d=2> Intuitively, bandit algorithms adaptively designate a small amount of traffic to collect user feedback in each round while improving their model estimation quality on the fly. **</s>**
<s no=3 p=1 c=3 d=0> In recent years, contextual bandit algorithms have gained increasing attention due to their capability of leveraging contextual information to deliver better personalized online services. **</s>**
<s no=4 p=1 c=4 d=0> They assume the expected reward of each action is determined by a conjecture of unknown bandit parameters and given context, which give them advantages when the space of recommendation is large but the rewards are interrelated. **</s>**
<s no=5 p=2 c=1 d=1> Most existing stochastic contextual bandit algorithms assume a fixed yet unknown reward mapping function. **</s>**
⋮
<s no=29 p=5 c=7 d=3> Extensive empirical evaluations on both a synthetic dataset and three real-world datasets for content recommendation confirmed the improved utility of the proposed algorithm, compared with both state-of-the-art stationary and non-stationary bandit algorithms. **</s>**
</ GİRİŞ >

Şekil 4.1. SIGIR2018 veri kümesinden bir örnek kaynak doküman

Şekil 4.1'de her cümle etiketler arasında sunulmuş ve cümlenin tekil numarası (no), bulunduğu paragrafın numarası (p), paragraftaki sırası (c) ve kaç okuyucu tarafından özete değer olarak seçildiği (d) özellikleri ile birlikte düzenlenmiştir. Yayınlardaki anahtar kelimeler ve alan kategorileri Şekil 4.2'de sunulmuştur.

< ANAHTAR KELİMELER
DOKÜMAN= "p495-wu"
<s no="1">Non-stationary Bandit**</s>**
<s no="2">Recommender Systems**</s>**
<s no="3">Regret Analysis**</s>**
</ ANAHTAR KELİMELER >

Alan Kategori 1: Information systems → Recommender systems
Alan Kategori 2: Theory of computation → Online learning algorithms
Alan Kategori 3: Theory of computation → Regret bounds
Alan Kategori 4: Computing methodologies → Sequential decision making

Şekil 4.2. SIGIR2018 kaynak dokümana ait anahtar kelimeler ve alan kategori listesi

Kaynak dokümanlarda uzun bildiriler için ortalama 37.25, kısa bildiriler için ortalama 21.79 olmak üzere ortalama 31.31 cümle bulunmaktadır. Ayrıca her bir kaynak doküman, ortalama 443.13 tekil kelime barındırmaktadır.

SIGIR2018'in metin özetleme problemi için uygun bir veri kümesi olabilmesi için referans alınabilecek özet dokümanlara sahip olması gerekmektedir. Bu bağlamda ilk olarak soyutlamalı özetler, yayının kendi yazarları tarafından hazırlanmış olan özet bölümlerinden elde edilmiştir. Şekil 4.3'te Wu ve arkadaşlarının SIGIR2018 yayınlarına ait soyutlamalı özet örneği sunulmuştur [118].

< SOYUTLAMALI ÖZET
DOKÜMAN= "p495-wu"
BAŞLIK = "Learning Contextual Bandits in a Non-stationary Environment">
<s no="1" p="1" c="1"> Multi-armed bandit algorithms have become a reference solution for handling the explore/exploit dilemma in recommender systems, and many other important real-world problems, such as display advertisement. **</s>**
<s no="2" p="1" c="2"> However, such algorithms usually assume a stationary reward distribution, which hardly holds in practice as users' preferences are dynamic. **</s>**
<s no="3" p="1" c="3"> This inevitably costs a recommender system consistent suboptimal performance. **</s>**
<s no="4" p="1" c="4"> In this paper, we consider the situation where the underlying distribution of reward remains unchanged over (possibly short) epochs and shifts at unknown time instants. **</s>**
<s no="5" p="1" c="5"> In accordance, we propose a contextual bandit algorithm that detects possible changes of environment based on its reward estimation confidence and updates its arm selection strategy respectively. **</s>**
<s no="6" p="1" c="6"> Rigorous upper regret bound analysis of the proposed algorithm demonstrates its learning effectiveness in such a non-trivial environment. **</s>**
<s no="7" p="1" c="7"> Extensive empirical evaluations on both synthetic and real-world datasets for recommendation confirm its practical utility in a changing environment. **</s>**
</ SOYUTLAMALI ÖZET >

Şekil 4.3. SIGIR2018 soyutlamalı özet dokümanı örneği

Yayınların orijinal yazarları tarafından hazırlanmış özetlerin veri kümesine dâhil edilmesi ile veri kümesi soyutlamalı metin özetleme için de uygulanabilir hale gelmiştir. Yayınların özet bölümlerinde ortalama 7.73 cümle, 115.86 tekil kelime bulunmaktadır.

SIGIR2018 veri kümesini mevcut veri kümelerinden ayıran en önemli özelliği, özete girmeye değer cümle kümelerinin kaynak dokümanlar içerisinden okurlar tarafından seçilmiş olmasıdır. Aday cümle kümelerinin seçimi bilgisayar mühendisliği lisans derecesine sahip bir okur ve lisans eğitimine devam eden iki okur tarafından gönüllü olarak yapılmıştır. Katılımcılar, 125 bildirinin tamamı için giriş bölümlerini dikkatle okumuş ve diğer okurların seçtikleri cümlelerden haberdar olmadan kendi çıkarmalı özetlerini oluşturmuşlardır. Bu çalışma sürecinde okurlara iki kısıt konulmuştur. Birinci kısıt, ortaya çıkacak çıkarmalı özetin metinde aktarılmak istenen bilginin mümkün olduğunca büyük bir

kısmını içerecek şekilde cümleler içermesidir. Bu kısıt genel bir özetleme görevinin sahip olması gereken bir özelliktir ve metindeki genel bilginin özet dokümanda da mümkün olduğunca korunmasını gerektirir.

Okurlara koyulan ikinci kısıt metinlerdeki cümle sayıları dikkate alındığında, ortaya çıkan çıkarmalı özetin boyunun kaynak metnin boyuna oranının yaklaşık olarak 1/3 olmasıdır. Bir başka ifadeyle, okurların metin özetlerini seçerken sıkıştırma oranının 2/3 olması beklenmiştir ve okurlardan kaynak metindeki cümlelerin yalnızca 1/3'ünü muhafaza ederek kaynak metni iyi temsil edecek özetler üretmeleri istenmiştir. Okurlar, bu kısıtlara göre özete girmeye değer cümleleri seçerek kendi aday cümle kümelerini oluşturmuşlardır. Bu sürecin sonunda tüm dokümanlar için üç aday cümle kümesi meydana gelmiştir.

Okurlar, cümle seçiminde birbirlerinden bağımsız karar vermiş olduklarından ve okurlara konulan kısıtlar metni ne şekilde özetleyeceklerine göre bir yönlendirme içermediğinden, aday cümle kümeleri birbirinden tamamıyla bağımsız oluşturulmuştur ve birbirinden büyük oranda farklılık göstermektedir. Bu noktada, her doküman için çıkarmalı metin özetlerini bu aday cümle kümelerinden elde etme ihtiyacı ortaya çıkmıştır. Bunun için birleşim ve kesişim olmak üzere iki yaklaşım kullanılmış ve her bir doküman için iki farklı boyutta çıkarmalı metin özetleri elde edilmiştir. Ortaya çıkan tüm metinler için ortalama cümle ve gereksiz kelimelerin atılmasından sonra kalan tekil kelime sayısını içeren istatistiklere Çizelge 4.1 ve Çizelge 4.2'de yer verilmiştir.

Çizelge 4.1. SIGIR2018 veri kümesindeki dokümanların cümle sayıları

	Kaynak metin		$Ext_{\cup}$		$Ext_{\cap}$		Soyutlamalı özet	
	μ	σ	μ	σ	μ	σ	μ	σ
Tüm dokümanlar	31.31	11.84	15.18	5.96	7.46	3.53	7.73	2.42
Uzun bildiriler	37.25	9.82	17.96	5.50	9.19	3.20	8.53	2.38
Kısa bildiriler	21.79	7.96	10.71	3.42	4.67	1.85	6.46	1.89

Çizelge 4.2. SIGIR2018 veri kümesindeki dokümanların tekil kelime sayıları

	Kaynak metin		$\text{Ext}_{\cup}$		$\text{Ext}_{\cap}$		Soyutlamalı özet	
	μ	σ	μ	σ	μ	σ	μ	σ
Tüm dokümanlar	443.13	160.33	250.69	93.30	132.82	61.05	115.86	32.80
Uzun bildiriler	528.48	126.67	300.60	76.00	166.03	50.48	128.95	30.70
Kısa bildiriler	306.21	103.89	170.63	54.74	79.56	31.76	94.88	24.17

Aday cümle kümelerinin birleşiminden elde edilen çıkarmalı özetler, okuyucuların bir doküman için tanımladıkları aday cümle kümelerinin birleşim kümesidir ($\text{Ext}_{\cup}$). Diğer bir ifadeyle bu çıkarmalı özetler, kaynak dokümandaki cümlelerin en az bir okuyucunun özete dâhil etmeye değer bulduğu cümlelerden oluşmaktadır. Buna göre $\text{Ext}_{\cup}$ hiçbir okuyucunun özete değer bulmadığı cümleler hariç tüm cümleleri içermektedir. Çizelge 4.1'de görüldüğü gibi ortalama 31.31 cümleden oluşan kaynak dokümanlardan elde edilen $\text{Ext}_{\cup}$ özetlerinin ortalama cümle sayısı 15.18'dir. Bu durumda bu özetler kaynak dokümandaki cümlelerin neredeyse yarısını içerir uzunluktadır.

Aday cümle kümelerinin kesişiminden elde edilen çıkarmalı özetler, okuyucuların bir doküman için tanımladıkları aday cümle kümelerinin kesişim kümesidir ($\text{Ext}_{\cap}$). Burada, kaynak dokümandaki cümlelerin içerisinden tüm okuyucuların özete dâhil etmeye değer buldukları cümleler kullanılarak çıkarmalı özet $\text{Ext}_{\cap}$ elde edilmiştir. Çizelge 4.1'de görüldüğü gibi ortalama 31.31 cümleden oluşan kaynak dokümanlardan elde edilen $\text{Ext}_{\cap}$ özetlerinin ortalama cümle sayısı 7.46'dır. Bu durumda bu özetler kaynak dokümandaki cümlelerin yaklaşık olarak %23'ünü içermektedir. Bu oran dokümanlara ait soyutlamalı metin özetlerinin boyutunun kaynak metnin boyutuna oranıyla yaklaşık olarak aynıdır.

Daha önce ifade edildiği gibi toplanan veri kümesi aşağıdaki bileşenlerden oluşmaktadır:

- Kaynak doküman: bildirilerin giriş bölümleri
- Abs: bildirilerin özet bölümleri

- $\text{Ext}_{\cup}$: Aday cümle kümelerinin birleşiminden elde edilen çıkarmalı metin özetleri
- $\text{Ext}_{\cap}$: Aday cümle kümelerinin kesişiminden elde edilen çıkarmalı metin özetleri

Bildirilere ait soyutlamalı metin özetleri (Abs), bildirilerin yazarları tarafından bildirinin vurgulanmak istenen bilgisini mümkün olduğunca aktaracak şekilde hazırlanmıştır. Ancak veri kümesinin oluşturulabilmesi için oluşturulan çıkarmalı özetler ($\text{Ext}_{\cup}$ ve $\text{Ext}_{\cap}$) metni en iyi bilen ve anlayan kişiler olan yazarlar tarafından değil, çalışma kapsamındaki okuyucuların önemli olarak değerlendirdiği cümle kümelerinin birleşim ve kesişiminden elde edilmiştir. Bu nedenle, okuyucuların seçmiş olduğu cümlelerin değerlendirilmesi ve doğrulanması gereklidir. Bunun için özetler ve kaynak dokümanlar arasında metin özetlerini değerlendirme yöntemlerinden en yaygın kullanılan ROUGE analizi yapılmış, özetlerin kaynak dokümandaki bilgiyi ne oranda koruyabildiği ölçülmüştür. Çizelge 4.3'te bu ölçümün sonuçları sunulmuştur.

Çizelge 4.3. Soyutlamalı ve çıkarmalı özetlerin kaynak dokümana göre ROUGE değerleri

	ROUGE-1			ROUGE-2			ROUGE-L		
Özet	F	D	A	F	D	A	F	D	A
Abs	34.9	66.0	23.7	11.8	26.1	7.6	21.9	41.3	14.9
$\text{Ext}_{\cup}$	78.7	100.0	64.9	71.5	97.1	56.5	78.7	100.0	64.9
$\text{Ext}_{\cap}$	55.5	100.0	38.4	45.5	96.6	29.8	55.5	100.0	38.4

ROUGE, en yaygın olarak anma değerleri üzerinden değerlendirilmektedir. Çizelge 4.3'teki ROUGE anma değerlerine göre kaynak dokümana en yakın özetler aday cümle kümelerinin birleşimi olan $\text{Ext}_{\cup}$ özetleridir. Bu özetlerin duyarlılık değeri %100 seviyesindedir, çünkü özetteki tüm cümleler referans alınan kaynak dokümanda mevcut olan cümlelerdir. Anma değeri ise %64.9 olarak hesaplanmıştır. Kaynak dokümana göre hesaplanan ROUGE anma değerini özetin cümle sayısının diğer özetlerden büyük olması etkilemektedir. $\text{Ext}_{\cap}$ özetleri ise içerisinde barındırdığı cümlelerin sayısı soyutlamalı özetlerle (Abs) aynı olmasına rağmen, soyutlamalı özetlere göre çok yüksek ROUGE anma değerine sahiptir. Elde edilen sonuçlara göre metinden çıkarmalı olarak elde edilen özetler, kaynak metni yazarlar tarafından hazırlanan özetlerden çok daha iyi yansıtmaktadır.

Çizelge 4.3'ten elde edilen sonucu desteklemek amacıyla kaynak doküman ve özetler arasında kosinüs benzerliği hesaplanarak bir analiz daha yapılmıştır. Bu analizin uygulanmasında her bir metin çeşidinin aynı çeşitteki diğer metinlerle benzerliği hesaplanmıştır. Yani kaynak dokümanların birbiriyle, soyutlamalı Abs özetlerin birbiriyle, çıkarmalı $\text{Ext}_{\cup}$ özetlerin birbiriyle ve çıkarmalı $\text{Ext}_{\cap}$ özetlerin birbiriyle kosinüs benzerliği ölçülmüş ve 125×125'lik dört benzerlik matrisi elde edilmiştir. Ardından bu benzerlik matrislerinin birbirlerine Öklid uzaklığı ölçülmüş, ortaya çıkan uzaklık değerleri [0,1] aralığında normalleştirilmiş ve Çizelge 4.4'teki değerler elde edilmiştir.

Çizelge 4.4. Doküman-doküman benzerlik matrislerinin normalize edilmiş uzaklık değerleri

	Kaynak metin	Abs	$\text{Ext}_{\cup}$	$\text{Ext}_{\cap}$
Kaynak metin	-	1.00	0.52	0.95
Abs	1.00	-	0.77	0.70
$\text{Ext}_{\cup}$	0.52	0.77	-	0.62
$\text{Ext}_{\cap}$	0.95	0.70	0.62	-

Çizelge 4.4'te görüldüğü gibi kaynak dokümanların birbirine benzerliğine en yakın benzerlik çıkarmalı $\text{Ext}_{\cup}$ özetlerinin birbirine benzerliğinden elde edilmiştir. Bunu $\text{Ext}_{\cap}$ özetlerin benzerliği izlemektedir. Kaynak dokümanların benzerliğine en uzak benzerlik Abs soyutlamalı özetlerinin benzerliğinden elde edilmiştir. Elde edilen sonuçlara göre, kaynak dokümanın benzerlik tabanlı bir bilgi çıkarım görevinde yerini en iyi tutacak olan özet $\text{Ext}_{\cup}$ daha sonra $\text{Ext}_{\cap}$ ve kaynak dokümanın yerini en düşük seviyede tutacak olan özetler ise soyutlamalı Abs özetlerdir. Gerçekleştirilmiş olan bu benzerlik analizi, ROUGE tabanlı değerlendirmeyi destekleyen nitelikte ölçüm sonuçları içermektedir.

Kaynak dokümanların okuyucular tarafından çıkarmalı özetleri oluşturmak amacıyla etiketlenmesini doğrulamak için yapılan bir diğer analiz ile de ortaya çıkan çıkarmalı özetlerin okuyucular tarafından rastgele seçilip seçilmediği ölçülmüştür. Bunun için üç farklı sıkıştırma oranının dikkate alınmasıyla iki farklı sayıda cümle içeren rastgele özetler elde edilmiş ve ROUGE tabanlı değerlendirmeye tabi tutulmuştur.

Rastgele özetlerin içerdikleri cümle sayısı küçük özetler için 8, büyük özetler için 16'dır. Bu değerler, okuyucular tarafından üretilmiş çıkarmalı özetlerdeki ortalama cümle sayılarının

yukarı yuvarlanmış tamsayı değerlerine karşılık gelmektedir. Bu sayede, geliştirilmiş olan çıkarmalı özetler, kendileriyle eş sayıda ve bazı durumlarda kendilerinden daha fazla sayıda cümle içeren rastgele özetlerle adil bir şekilde karşılaştırılabilmiştir. Küçük ve büyük rastgele özetlerin her biri beşer kez üretilmiş ve ortalama ROUGE değerleri Çizelge 4.5 ve Çizelge 4.6'da sunulmuştur.

Çizelge 4.5. Küçük rastgele özetlerin soyutlamalı ve çıkarmalı özetlerle karşılaştırılması

	ROUGE-1 (%)			ROUGE-2 (%)			ROUGE-L (%)		
Özet	F	D	A	F	D	A	F	D	A
Abs	34.9	66.0	23.7	11.8	26.1	7.6	21.9	41.3	14.9
$Ext_{\cap}$	55.5	100.0	38.4	45.5	96.6	29.8	55.5	100.0	38.4
$Rastgele_{8}$	42.4	100.0	26.9	31.4	94.6	18.8	27.9	66.5	17.6

Çizelge 4.6. Büyük rastgele özetlerin çıkarmalı özetlerle karşılaştırılması

	ROUGE-1 (%)			ROUGE-2 (%)			ROUGE-L (%)		
Özet	F	D	A	F	D	A	F	D	A
$Ext_{\cup}$	78.7	100.0	64.9	71.5	97.1	56.5	78.7	100.0	64.9
$Rastgele_{16}$	74.4	100.0	59.2	63.3	93.8	47.8	38.0	51.2	30.2

Çizelge 4.5 ve Çizelge 4.6'da rastgele üretilmiş olan özetlerin ROUGE değerlerinin okuyucular tarafından üretilen çıkarmalı özetlere nazaran belirgin derecede düşük olduğu açıkça görülmektedir. Elde edilen sonuçlar, okuyucular tarafından oluşturulan özetlerin rastgele oluşturulmadığını göstermektedir.

4.2. SIGIR2018 Üzerinde Metin Özetleme Çalışmaları

Önerilmiş olan metin özetleme veri kümesi SIGIR2018, literatürde sıklıkla çalışılmış olan metin özetleme yaklaşımlarına tabi tutulmuştur. Bunun için şebeke tabanlı denetimsiz öğrenme yaklaşımı, bulanık kural tabanlı sistem yaklaşımı, sığ makine öğrenmesi yöntemi ve derin öğrenme modelinin gerçekleştirimi yapılmıştır. Modellerden elde edilen çıkarmalı özetlerin kaynak dokümanı temsil etme derecesi ROUGE ölçütleri kullanılarak ölçülmüştür. Bu yöntemlerden elde edilen ROUGE sonuçları tez çalışmasında geliştirilen metin özetleme yöntemi için de bir referans noktası olarak da değerlendirilmiştir.

Gerçekleştirilen makine öğrenmesi yaklaşımları içerisinde bulanık kural tabanlı sistemler ve yapay sinir ağlarının uygulamasında sözdizimsel özniteliklerden, derin öğrenme yaklaşımında ise anlamsal öznitelik olan kelime özyerleşiklerinden yararlanılmıştır. Öte yandan, denetimsiz yaklaşım olarak uygulanmış olan TextRank algoritması metinlere ait cümlelerin ham halini girdi olarak kabul eder ve cümleler arasındaki benzerlik üzerinden önemli cümleleri seçebilmeyi hedefler.

Bulanık kural tabanlı sistemlerin ve yapay sinir ağlarının gerçekleştiriminde ortak olarak izlenen süreç sözdizimsel özniteliklerin tüm cümleler için toplanması ve modelin giriş veri kümesinin oluşturulmasıdır. Bu bölümde, bulanık sistemin ve yapay sinir ağının gerçekleştirimindeki bu ortak adımlara ait detaylara yer verilmiştir.

Bu tez çalışmasında kullanılan sözdizimsel öznitelikler, literatürdeki metin özetleme çalışmalarının yararlandığı sözdizimsel özniteliklerin bir araya getirilmesiyle oluşturulmuş ve geniş bir özellik listesi oluşturularak Çizelge 4.7'de sunulmuştur.

Çizelge 4.7. SIGIR2018 veri kümesinden elde edilen sözdizimsel öznitelikler listesi

Sıra numarası	Öznitelik	Eşitlik	Önem Değeri
1	TF	Eş. 2.1	972.27
2	Uzunluk	Eş. 2.10	79.37
3	Göreceli cümle pozisyonu	-	40.16
4	Uzunluk	Eş. 2.11	36.80
5	Tamlamalar (PP)	Eş. 2.15	35.28
6	Tamlamalar (NP)	Eş. 2.15	32.32
7	TF-ISF (4-gram)	Eş. 2.3	23.94
8	TF-ISF (3-gram)	Eş. 2.3	19.86
9	Tamlamalar (AP)	Eş. 2.15	18.68
10	TF-ISF (2-gram)	Eş. 2.3	16.44
11	Tamlamalar (AP)	Eş. 2.14	15.94
12	Tamlamalar (VP)	Eş. 2.15	13.40
13	TF-ISF (1-gram)	Eş. 2.3	12.74
14	Özel isimler	Eş. 2.14	12.51
15	Şebeke bağlantı sıklığı	-	11.73
16	Tamlamalar (PP)	Eş. 2.14	11.34

Çizelge 4.7. (devam) SIGIR2018 veri kümesinden elde edilen sözdizimsel öznitelikler listesi

Sıra numarası	Öznitelik	Eşitlik	Önem Değeri
17	Özel isimler	Eş. 2.15	10.65
18	Tamlamalar (NP)	Eş. 2.14	9.73
19	Başlık	Eş. 2.6	8.45
20	Başlık	Eş. 2.4	8.32
21	Toplam benzerlik	-	7.93
22	Cümle-cümle ilişkisi (1-gram)	Eş. 2.16	7.93
23	Pozisyon	Eş. 2.7	7.23
24	Tamlamalar (VP)	Eş. 2.14	6.64
25	Anahtar sözcükler	Eş. 2.6	4.19
26	Cümle-cümle ilişkisi	Eş. 2.17	3.73
27	Göreceli paragraf pozisyonu	-	3.00
28	Pozisyon (metne göre)	Eş. 2.9	2.22
29	Tamlamalar (VP)	Eş. 2.13	2.19
30	Cümle-cümle ilişkisi (2-gram)	Eş. 2.16	2.01
31	Pozisyon (paragrafa göre)	Eş. 2.8	1.72
32	Başlık	Eş. 2.5	1.69
33	Tamlamalar (PP)	Eş. 2.13	0.50
34	Cümle-cümle ilişkisi (3-gram)	Eş. 2.16	0.49
35	Pozisyon (metne göre)	Eş. 2.9	0.47
36	Cümle-cümle ilişkisi (4-gram)	Eş. 2.16	0.32
37	Pozisyon (paragrafa göre)	Eş. 2.8	0.29
38	Tamlamalar (AP)	Eş. 2.13	0.23
39	Özel isimler	Eş. 2.13	0.18
40	TF	Eş. 2.2	3.32E-04
41	Tamlamalar (NP)	Eş. 2.13	1.05E-04
42	Uzunluk	Eş. 2.12	1.47E-05

Çizelge 4.7'de görüldüğü gibi bir cümlenin metin özetlemede özete değer cümle olarak seçilmesinde etkili olacak çok sayıda sözdizimsel öznitelik vardır. Bu tez çalışmasında bu öznitelikler TF, TF-ISF, uzunluk, pozisyon, başlık, anahtar sözcükler, cümle-cümle ilişkisi, şebeke bağlantı sıklığı, toplam benzerlik, tamlamalar, özel isimler olarak ele alınmış ve bunlara ait farklı hesaplama şekilleri ile toplam 42 sözdizimsel öznitelik tanımlanmıştır. Bu özniteliklerle ilgili ayrıntılı bilgiye ve hesaplama şekillerine Bölüm 2.2.1'de yer verilmiştir.

Bu çalışmada, bir sözdizimsel özniteliğin birden çok hesaplama şekli kullanılmıştır. Çünkü metindeki bir cümlenin belirli bir öznitelikten alacağı değer o metnin türüyle ve yazım şekliyle doğrudan bağlantılıdır ve bir öznitelik türünün tek bir eşitlik üzerinden katı bir yaklaşımla uygulanması, ayırt edici olabilecek bazı yaklaşımların yok sayılması sonucunu çıkarabilmektedir. Bu nedenle, bu çalışmada bir özniteliğin birden çok hesaplama şekline yer verilerek toplam 42 sözdizimsel öznitelik tanımlanmıştır. Tüm cümleler için Çizelge 4.7'de referans gösterilmiş olan öznitelik hesaplama ölçütlerine göre öznitelik değerleri atanmıştır. Oluşan veri kümesinin üç cümleden oluşan bir parçasına Şekil 4.4'te yer verilmiştir.

Kayıt Numarası**;** Göreceli cümle pozisyonu[-]**;** Göreceli paragraf pozisyonu[-]**;** TF[Eş. 2.1]**;** TF[Eş. 2.2]**;** TF-ISF(1-gram)[Eş. 2.3]**;** TF-ISF(2-gram)[Eş. 2.3]**;** TF-ISF(3-gram)[Eş. 2.3]**;** TF-ISF(4-gram)[Eş. 2.3]**;** Cümle-cümle ilişkisi(1-gram)[Eş. 2.13]**;** Cümle-cümle ilişkisi[Eş. 2.14]**;** Toplam benzerlik[-]**;** Cümle-cümle ilişkisi(2-gram)[Eş. 2.13]**;** Cümle-cümle ilişkisi(3-gram)[Eş. 2.13]**;** Cümle-cümle ilişkisi(4-gram)[Eş. 2.13]**;** Şebeke bağlantı sıklığı[-]**;** Başlık[Eş. 2.4]**;** Başlık[Eş. 2.5]**;** Başlık[Eş. 2.6]**;** Anahtar sözcükler[Eş. 2.6]**;** Pozisyon[Eş. 2.7]**;** Pozisyon(paragrafa göre)[Eş. 2.8]**;** Pozisyon(metne göre)[Eş. 2.9]**;** Pozisyon(paragrafa göre)[Eş. 2.8 Pozisyon(metne göre)[Eş. 2.9 Uzunluk[Eş. 2.10]**;** Uzunluk[Eş. 2.11]**;** Uzunluk[Eş. 2.12]**;** Tamlamalar(VP)[Eş. 2.13]**;** Tamlamalar(VP)[Eş. 2.14]**;** Tamlamalar(VP)[Eş. 2.15]**;** Tamlamalar(NP)[Eş. 2.13]**;** Tamlamalar(NP)[Eş. 2.14]**;** Tamlamalar(NP)[Eş. 2.15]**;** Tamlamalar(PP)[Eş. 2.13]**;** Tamlamalar(PP)[Eş. 2.14]**;** Tamlamalar(PP)[Eş. 2.15]**;** Tamlamalar(AP)[Eş. 2.13]**;** Tamlamalar(AP)[Eş. 2.14]**;** Tamlamalar(AP)[Eş. 2.15]**;** Özel isimler[Eş. 2.13]**;** Özel isimler[Eş. 2.14]**;** Özel isimler[Eş. 2.15]**;** Özete-Dâhil[Çıktı]

0; 1; 1; 20; 0.0039525692000000005; 0.7050643139; 0.6185055527000001; 0.5986692921; 0.5890150894; 0.7673530591000001; 0.0769230769; 0.0; 0.8106419621000001; 0.9113492222; 0.9345180353; 0; 0.1111111111; 0.0357142857; 0.1170257574; 0.0; 1.0; 1.0; 1.0; 1.0; 1.0; 14.163.090.129; 0.3846153846; 26.030.836.315; 0.1; 0.0429184549; 0.4; 0.55; 0.0643776824; 0.5; 0.2; 0.0708154506; 0.33333333330000003; 0.30000000000000004; 0.3540772532; 0.5454545455; 0.1; 0.1573676681; 0.25; [0]

1; 2; 1; 17; 0.0039525692000000005; 0.6458970167; 0.5844658561; 0.5641367138; 0.5530880987; 0.9230870711; 0.0653846154; 0.0; 0.9404717508; 0.9710247913000001; 1.0; 0; 0.0; 0.0; 0.1672135036; 0.0; 0.8; 0.8333333333; 0.8; 0.9696969697000001; 0.9687500000000001; 12.746.781.116; 0.3461538462; 26.184.791.772.000.000; 0.1666666667; 0.057939914200000005; 0.6000000000000001; 0.5; 0.0474053843; 0.4090909091; 0.1666666667; 0.0478004292; 0.25; 0.0555555556; 0.053111588; 0.0909090909; 0.0; 0.0; 0.0; [0]

2; 3; 1; 44; 0.0039525692000000005; 1.0; 1.0; 1.0; 1.0; 0.7014863646; 0.16923076920000002; 0.0; 0.5954300929; 0.7826320651; 0.9345180353; 0; 0.1111111111; 0.0192307692; 0.1480253007; 0.0; 0.6000000000000001; 0.0; 0.6000000000000001; 0.9393939394; 0.9375; 36.824.034.335; 1.0; 23.177.525.504; 0.0384615385; 0.1115879828; 0.4; 0.4230769231; 0.3347639485; 1.0; 0.2307692308; 0.552360515; 1.0; 0.2115384615; 16.877.682.403; 1.0; 0.0; 0.0; 0.0; [1]

Şekil 4.4. Örnek 3 cümle için sözdizimsel öznitelik değerleri

Şekil 4.4'te, her cümle için toplanmış olan öznitelikler ve bu cümlenin özete dâhil olmaya değer olup olmadığı bilgisi yer almaktadır.

Sözdizimsel özniteliklerin veri kümesi, özniteliklerin insanlar tarafından seçilen cümleler üzerindeki etkisinin ölçülebilmesi amacıyla bir öznitelik seçme algoritmasına sunulmuştur. Öznitelik seçme sürecinin çıktısı veri kümesine bağımlıdır. Çizelge 4.7'de SIGIR2018 için her bir özniteliğin Chi kare (χ^2) testine göre önem değerleri hesaplanmış ve öznitelikler önem değerine göre sıralanmıştır. Metin özetleme sürecinde farklı büyüklükteki sözdizimsel öznitelik kümeleri üzerinde yapılan deneysel çalışmalarda elde edilen bu sıralamadan faydalanılmaktadır. Buna göre n adet sözdizimsel öznitelik üzerinde yapılan deneylerde bu listedeki ilk n öznitelik kullanılacaktır.

4.2.1. Bulanık kural tabanlı sistemin SIGIR2018 veri kümesinde uygulanması

Bulanık kural tabanlı uzman sistemler, bulanık küme teorisini uzman sistemin çıkarsama süreçlerine dâhil eden, çıkarsamada girdi sözel değişkenlerinin tümünü çıktı sözel değişkenleri ile örtüştüren kuralları kullanan sistemlerdir. SIGIR2018 kümesinden bulanık kural tabanlı sistemlerle metin özetlemede aşağıdaki adımlar uygulanmıştır:

- Kaynak metinlerden sözdizimsel özniteliklerin çıkarılması
- Girdi matrisinin oluşturulması
- Bulanık kuralların üretilmesi
- Cümle önem derecelendirmesinde bulanık kural tabanlı sistem gerçekleştirimi
- Elde edilen özetlerin değerlendirilmesi

Kaynak metinlerden sözdizimsel özniteliklerin çıkarılması

Bu tez çalışmasında bu öznitelikler TF, TF-ISF, uzunluk, pozisyon, başlık, anahtar sözcükler, cümle-cümle ilişkisi, şebeke bağlantı sıklığı, toplam benzerlik, tamlamalar, özel isimler olarak ele alınmış ve bunlara ait farklı hesaplama şekilleri ile toplam 42 sözdizimsel öznitelik tanımlanmıştır. Bu özniteliklerle ilgili ayrıntılı bilgiye ve hesaplama şekillerine Bölüm 2.2.1'de yer verilmiştir. Çalışmanın bu aşamasında, kaynak dokümanlardaki her bir cümle için Bölüm 2.2.1'de verilmiş olan sözdizimsel özniteliklere ait cümle öznitelik değerleri hesaplanmıştır.

Girdi matrisinin oluşturulması

SIGIR2018 veri kümesi için geliştirilmiş olan çıkarmalı özetler aday cümle kümelerinin kesişimi ve birleşimi olmak üzere iki türdür. Aday cümle kümelerinin birleşimi olan $\mathrm{Ext}_{\cup}$ çıkarmalı özetler, dokümandaki cümlelerin yaklaşık yarısının özete değer olarak etiketlendiği geniş çıkarmalı özetlerdir. Öte yandan, aday cümle kümelerinin kesişimi olan $\mathrm{Ext}_{\cap}$ çıkarmalı özetler, yazarlar tarafından hazırlanmış olan soyutlamalı özetler kadar cümle içermekte ve içerdikleri cümle sayısı, kaynak dokümandaki cümle sayısının yaklaşık %23'üne karşılık gelmektedir. Bu özetlerdeki cümleler tüm okuyucular tarafından özete değer olarak seçilmiş cümlelerdir. Bu nedenle, bu tez çalışmasında öznitelikler çıkarıldıktan sonra girdi matrisinin oluşturulmasında $\mathrm{Ext}_{\cap}$ tipi özetlerin etiketleri üzerinden çalışılmıştır.

Bulanık kural tabanlı sistemler üzerinde yapılan metin özetleme çalışmasında değişken öznitelik vektörü üzerinden çok sayıda deney yapılmıştır. Öznitelikler, χ^2 değerlerine göre Çizelge 4.7'deki gibi sıralanmış ve bulanık kural tabanının girişleri, girdi özniteliklerin sonuca etkisini araştırmak adına en önemli 4 öznitelikten başlanarak 21 özniteliğe kadar arttırılarak değiştirilmiştir. Bu sayede her bir satırı, dokümanlardaki cümlelere denk gelen, cümlelerin özete ait olma bilgisinin ise çıkış parametresini oluşturduğu 19 farklı veri kümesi oluşturulmuştur.

Modelin geliştirilmesinde maksimum öznitelik sayısının 21 ile sınırlandırılmış olması bulanık kural tabanlı sistemin çok sayıda girdi sözel değişkeninin tanımlandığı karmaşık problemlerde yeterli verimlilikte uygulanabilir olmayışından kaynaklanmaktadır. Artan girdi sayısı, tüm girdiler ile tüm çıktılar arasındaki ilişkinin doğru şekilde tanımlanarak doğru kural üretilmesi sürecini zorlaştırmakta, tek bir kuralın yapısal karmaşıklığını arttırmakta ve bunun sonucunda modelin başarısını düşürmektedir. Bu nedenle girdi sözel değişkenleri olan özniteliklerin sayısı maksimum χ^2 değerine sahip 4 öznitelikle 21 öznitelik arasında sınırlandırılmıştır.

Bulanık kuralların üretilmesi

Özniteliklerin her biri bulanık sistemin girdi sözel değişkenini oluşturmuştur ve her bir sözel değişken üç üçgen bulanık küme ile gösterilmiştir. Bulanık küme parametrelerinin ve

ardından bulanık kuralların üretiminde WM kural üretme yönteminden faydalanılmıştır. Buna göre, girdi sözel değişkenleri “düşük (D)”, “orta (O)” ve “yüksek (Y)” olmak üzere üç üçgen bulanık küme ile ifade edilmiştir. Bulanık kümelere ait sınır parametreleri veri kümesindeki ilgili özniteliğin değer aralığına bağlı olarak WM algoritmasının önermiş olduğu şekilde tanımlanmıştır. Her bir özniteliğin değer aralığı üç eş aralığa ayrılarak tanımlanmıştır. Çıkış sözel değişkeni ise “özete değer” ve “özete değer değil” olarak iki üçgen bulanık küme ile ifade edilmiş ve bu bulanık kümelerin sınır parametreleri (0.0, 1.0, 1.0) ve (0.0, 0.0, 1.0) olarak tanımlanmıştır.

Kullanılan sözdizimsel özniteliklerin sayısı n ve $4 \leq n \leq 21$ olmak üzere değişken öznitelik sayısına sahip 19 veri kümesi WM algoritmasına kural üretimi için verilmiştir. Bunun için beş-katlı çapraz geçerleme uygulanmış ve çapraz geçerlemenin her iterasyonunda tüm veri kümesinin rastgele seçilen %80’i kuralların üretilmesi amacıyla kullanılmıştır. Değişken sayıda özniteliğe sahip tüm bu veri kümeleri için meydana gelen kuralların ortalama sayısını ve birisi özete dâhil edilmeye değer (“1”), birisi özete dâhil edilmeye değer olmayan (“0”) ardıl kısma sahip örnek kuralları içeren liste Çizelge 4.8’de verilmiştir.

Çizelge 4.8. Değişken sayıda öznitelik için elde edilen örnek bulanık kurallar

n	Sayı	Örnek kural
4	22	1: D & 2: D & 3: D & 4: O; 0 1: O & 2: O & 3: O & 4: Y; 1
5	50	1:D & 2: D & 3: D & 4: D & 5: D; 0 1: D & 2: Y & 3: Y & 4: Y & 5: Y; 1
6	92	1: D & 2: O & 3: D & 4: D & 5: O & 6: O; 0 1: D & 2: O & 3: O & 4: Y & 5: O & 6: Y; 1
7	116	1: D & 2: O & 3: O & 4: D & 5: D & 6: O & 7: O; 0 1: D & 2: O & 3: Y & 4: O & 5: Y & 6: Y & 7: Y; 1
8	137	1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: O & 8: O; 0 1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: O & 8: D; 1
9	274	1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: O & 8: O & 9: O; 0 1: D & 2: O & 3: Y & 4: Y & 5: O & 6: Y & 7: Y & 8: Y & 9: Y; 1
10	315	1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: O & 8: O & 9: O & 10: O; 0 1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: O & 8: O & 9: D & 10: D; 1

Çizelge 4.8. (devam) Değişken sayıda öznitelik için elde edilen örnek bulanık kurallar

n	Sayı	Örnek kural
11	328	1: D & 2: O & 3: Y & 4: Y & 5: Y & 6: O & 7: O & 8: Y & 9: Y & 10: D & 11: D; 0 1: D & 2: O & 3: Y & 4: O & 5: O & 6: D & 7: O & 8: Y & 9: Y & 10: D & 11: Y; 1
12	549	1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: O & 8: O & 9: O & 10: D & 11: D & 12: O; 0 1: D & 2: O & 3: O & 4: O & 5: O & 6: D & 7: D & 8: D & 9: O & 10: D & 11: D & 12: D; 1
13	665	1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: O & 8: O & 9: O & 10: O & 11: O & 12: D & 13: O; 1 1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: D & 8: D & 9: D & 10: O & 11: D & 12: D & 13: D; 1
14	687	1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: O & 8: O & 9: O & 10: O & 11: O & 12: D & 13: D & 14: D; 0 1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: O & 8: D & 9: D & 10: D & 11: O & 12: D & 13: D & 14: D; 1
15	867	1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: O & 8: D & 9: O & 10: O & 11: O & 12: O & 13: D & 14: D & 15: D; 0 1: D & 2: Y & 3: Y & 4: Y & 5: Y & 6: Y & 7: D & 8: Y & 9: Y & 10: D & 11: Y & 12: Y & 13: Y & 14: Y & 15: D; 1
16	889	1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: D & 8: O & 9: O & 10: O & 11: O & 12: D & 13: O & 14: D & 15: D & 16: D; 0 1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: D & 8: O & 9: O & 10: O & 11: O & 12: D & 13: O & 14: D & 15: O & 16: D; 1
17	1146	1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: D & 8: O & 9: O & 10: O & 11: O & 12: D & 13: O & 14: D & 15: O & 16: D & 17: O; 0 1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: D & 8: O & 9: O & 10: O & 11: O & 12: D & 13: O & 14: D & 15: D & 16: D & 17: D; 1
18	1177	1: D & 2: O & 3: Y & 4: Y & 5: Y & 6: D & 7: O & 8: O & 9: D & 10: D & 11: O & 12: O & 13: D & 14: D & 15: D & 16: D & 17: D & 18: Y; 0 1: D & 2: O & 3: Y & 4: Y & 5: Y & 6: O & 7: D & 8: O & 9: O & 10: O & 11: D & 12: O & 13: D & 14: Y & 15: D & 16: Y & 17: D & 18: D; 1
19	1331	1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: D & 8: D & 9: O & 10: O & 11: O & 12: D & 13: O & 14: D & 15: O & 16: D & 17: O & 18: D & 19: O; 0 1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: D & 8: D & 9: O & 10: O & 11: O & 12: D & 13: O & 14: D & 15: O & 16: D & 17: O & 18: D & 19: D; 1

Çizelge 4.8. (devam) Değişken sayıda öznitelik için elde edilen örnek bulanık kurallar

n	Sayı	Örnek kural
20	1654	1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: D & 8: D & 9: D & 10: O & 11: O & 12: O & 13: D & 14: O & 15: D & 16: O & 17: D & 18: O & 19: D & 20: O; 0 1: D & 2: O & 3: O & 4: O & 5: O & 6: O & 7: D & 8: D & 9: D & 10: O & 11: O & 12: O & 13: D & 14: O & 15: D & 16: O & 17: D & 18: D & 19: D & 20: D; 1
21	1391	1: O & 2: O & 3: Y & 4: Y & 5: Y & 6: Y & 7: D & 8: D & 9: D & 10: D & 11: O & 12: O & 13: O & 14: O & 15: O & 16: O & 17: Y & 18: D & 19: O & 20: D & 21: D; 0 1: D & 2: O & 3: Y & 4: Y & 5: O & 6: O & 7: D & 8: D & 9: D & 10: D & 11: O & 12: O & 13: O & 14: D & 15: O & 16: O & 17: D & 18: D & 19: O & 20: D & 21: D; 1

Çizelge 4.8'de ilk sütunda söz konusu veri kümesinin kaç öznitelikten oluştuğu sunulmaktadır. İkinci sütun oluşturulan bulanık kuralların sayısıdır. Kullanılacak özniteliklerin listesi Çizelge 4.7'de belirtildiği şekilde alınmıştır. Örneğin, 4 adet özniteliğe sahip veri kümesinde kullanılan öznitelikler Çizelge 4.7'de $n=\{1,2,3,4\}$ sıra numaralarına sahip olan özniteliklerdir.

Çizelge 4.8'de bir kuralın tanımlanmasında bu sıra numaralarına göre öznitelikler ve özniteliklere ait bulanık kümeler kullanılmıştır. Çizelgedeki bir kural olan "1: D & 2: D & 3: D & 4: Y; 1" ifadesinin karşılık geldiği bulanık kuralın sözel ifadesi Eş. 4.1'deki gibidir.

[EĞER $TF_{\text{Eş. 2.1}}$ Düşük & $Uzunluk_{\text{Eş. 2.10}}$ Düşük & *Göreceli Cümle Pozisyonu* Düşük & $Uzunluk_{\text{Eş. 2.11}}$ Yüksek İSE cümle özete dâhil edilmeye değerdir.] (4.1)

Cümle önem derecelendirmesinde bulanık kural tabanlı sistem gerçekleştirimi

Giriş ve çıkış sözel değişkenleri için bulanık kümelerin tanımlanmasının ve bunlar arasındaki ilişkinin WM ile modellenmesi ile bulanık kuralların oluşturulmasının ardından Mamdani tipi bir bulanık çıkarsama sisteminin gerçekleştirilmiştir. Mamdani bulanık çıkarsama sisteminin kural tetikleme aşamasında minimum, kural birleştirme aşamasında ise maksimum operatörü kullanılmıştır. Durulaştırma adımında bölgenin merkezi kullanılarak bulanık çıkarsama sürecinin [0, 1] aralığında keskin çıktısı üretilmiştir.

Elde edilen özetlerin değerlendirilme

SIGIR2018 üzerinde sözdizimsel özniteliklerin artırımlı kullanımı ile oluşturulan veri kümeleri ile bulanık sistem gerçekleştirimi beş-katlı çapraz geçerleme ile yapılmıştır. Aslında, bir bulanık sistemin geliştirilmesi için veriden öğrenme söz konusu olmadığından eğitim ve test süreçleri genellikle ayrı ayrı gerçekleştirilmez. Ancak bu çalışmada kural üretme aşamasında veriden öğrenen bir yöntem kullanıldığından çapraz geçerleme yapılmıştır. Çapraz geçerlemede eğitim kümesi kural üretiminde, test kümesi ise üretilen kuralların entegre edildiği bulanık sistemin metin özetlemede kullanımında ve özet başarısının ölçülmesinde kullanılmıştır.

Beş-katlı çapraz geçerlemede toplam 125 yayından oluşan SIGIR2018'in rastgele seçilen 100 dokümanı kural kümelerinin oluşturulmasında eğitim kümesi olarak değerlendirilmiş; kalan 25 doküman test kümesi olarak kullanılarak özetin başarısı ölçülmüştür.

Özet başarısının değerlendirilmesinde temel olarak ROUGE yöntemi kullanılmıştır. ROUGE analizi, modelin ürettiği otomatik özetlerle referans alınan doküman arasındaki n-gram örtüşmesini ölçmektedir. Bu çalışmada referans alınan dokümanlar, kaynak metnin tümü ve aday cümle kümelerinin kesişimi olan $\mathrm{Ext}_{\cap}$ çıkarmalı özetleri olmak üzere iki türdür.

Çalışmada kaynak dokümanların da ROUGE hesaplarında referans alınmasının sebebi metnin otomatik özetleri oluşturulduğunda kaynak dokümandaki bilginin ne ölçüde korunabildiğinin araştırılmasıdır. Öte yandan, $\mathrm{Ext}_{\cap}$ özetlerin referans alınmasının nedeni ise otomatik özetlerin, insanlar tarafından elle üretilen özetlere ne oranda yaklaşabildiğinin gözlenebilmesidir.

Çizelge 4.9 ve Çizelge 4.10'da otomatik üretilmiş olan özetlerin sırasıyla kaynak dokümana ve $\mathrm{Ext}_{\cap}$ çıkarmalı özetlere göre ROUGE değerleri eğitim ve test için beş-katlı çapraz geçerlemenin ortalama değerlerini içerecek şekilde sunulmuştur.

Çizelge 4.9. Bulanık sistemlerin kaynak metne göre metin özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
n	D	A	D	A	D	A	D	A	D	A	D	A
4	99.4	48.5	96.4	38.9	99.4	48.5	99.2	49.7	96.2	39.8	99.2	49.7
5	100.0	48.6	97.1	39.1	100.0	48.6	100.0	48.5	97.0	39.0	100.0	48.5
6	100.0	51.2	97.1	41.3	100.0	51.2	100.0	50.9	97.1	41.3	100.0	50.9
7	100.0	49.4	96.9	39.4	100.0	49.4	100.0	49.5	96.8	39.8	100.0	49.5
8	100.0	44.2	96.8	34.6	100.0	44.2	100.0	44.2	96.8	34.8	100.0	44.2
9	99.6	43.8	96.3	34.2	99.6	43.8	100.0	44.7	96.7	35.1	100.0	44.7
10	99.8	43.2	96.5	33.7	99.8	43.2	100.0	43.6	96.7	34.1	100.0	43.6
11	99.8	44.7	96.6	35.0	99,8	44.7	100.0	44.8	96.8	35.3	100.0	44.8
12	99.4	39.8	96.2	30.8	99.4	39.8	100.0	40.0	96.8	30.9	100.0	40.0
13	99.4	39.5	96.2	30.5	99.4	39.5	100.0	39.2	96.7	30.2	100.0	39.2
14	99.6	39.1	96.3	30.1	99.6	39.1	99.2	37.5	95.8	28.5	99.2	37.5
15	99.4	38.9	96.1	29.9	99.4	38.9	100.0	37.6	96.5	28.7	100.0	37.6
16	99.8	39.7	96.5	30.6	99.8	39.7	99.2	37.4	95.8	28.4	99.2	37.4
17	99.8	39.7	96.5	30.8	99.8	39.7	98.4	36.8	95.0	28.2	98.4	36.8
18	99.6	39.2	96.3	30.3	99.6	39.2	99.2	36.6	95.7	27.9	99.2	36.6
19	99.0	37.8	95.7	29.0	99.0	37.8	100.0	35.6	96.5	27.0	100.0	35.6
20	99.0	38.2	95.7	29.3	99.0	38.2	100.0	36.0	96.5	27.4	100.0	36.0
21	99.0	38.3	95.6	29.4	99.0	38.3	100.0	36.2	96.5	27.5	100.0	36.2

Çizelge 4.9 ve Çizelge 4.10 için edinilen ortak bulgu, sözdizimsel özniteliklerin artışının başta ROUGE değerlerinde olumlu sonuçlar doğurduğu, 6 ve 7 özniteliğin kullanımı ile bu değerlerde bir olgunluk düzeyine erişildiği ve 8 öznitelik ve sonrası için artan öznitelik sayısının WM kurallarındaki doğruluğu düşürdüğünden otomatik özetlerin de başarısını belirgin bir şekilde düşürdüğüdür.

Çizelge 4.10. Bulanık sistemlerin hedef çıkarmalı özetlere göre özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
n	D	A	D	A	D	A	D	A	D	A	D	A
4	49.4	63.6	38.0	51.0	45.0	58.2	48.9	64.5	37.2	51.3	44.2	58.7
5	50.4	63.9	39.4	51.5	46.4	58.8	50.5	63.1	39.4	50.5	46.3	57.8
6	50.8	67.4	40.2	55.0	47.3	62.7	50.3	66.7	39.3	54.2	46.6	61.9
7	50.0	64.3	39.0	51.4	46.1	59.1	49.7	64.3	38.5	51.2	45.8	59.2
8	51.0	58.6	39.4	45.7	46.5	53.5	50.0	56.9	38.1	43.6	45.3	51.6
9	52.6	60.6	41.2	48.0	48.3	55.7	49.0	57.2	36.7	43.4	43.8	51.3
10	54.1	61.0	42.9	48.7	50.0	56.3	49.4	56.0	37.0	42.5	44.3	50.4
11	53.7	62.2	42.5	49.6	49.6	57.4	49.4	57.7	37.2	43.9	44.3	51.8
12	59.0	61.2	48.3	49.6	54.9	56.9	50.4	52.9	37.7	39.7	45.1	47.5
13	61.2	62.6	51.0	51.8	57.5	58.9	51.6	52.7	39.2	39.9	46.5	47.6
14	62.2	63.1	52.2	52.4	58.6	59.4	50.0	49.2	36.8	35.6	44.4	43.6
15	63.4	64.5	54.0	54.5	60.1	61.1	50.5	49.1	37.0	35.2	44.7	43.3
16	64.4	66.6	55.1	56.9	61.2	63.4	51.1	49.7	37.9	35.9	45.5	43.9
17	69.1	71.1	60.6	62.6	66.3	68.4	51.0	49.5	38.2	36.8	45.8	44.4
18	69.2	70.3	60.8	61.6	66.5	67.6	51.0	49.1	37.8	36.1	45.5	43.9
19	71.7	71.3	64.2	63.3	69.3	69.0	50.7	47.8	37.0	34.8	44.7	42.3
20	71.8	72.1	64.5	64.2	69.5	69.9	50.1	48.0	36.2	34.9	44.1	42.5
21	72.0	72.4	64.8	64.7	69.9	70.3	49.9	48.1	36.0	34.9	43.8	42.5

SIGIR2018 üzerinde yapılan bulanık sistem gerçekleştiriminde kaynak dokümana göre ROUGE anma değeri (%49.5, %50.9) iken $\text{Ext}_{\cap}$ çıkarmalı özetlere göre ROUGE-1 anma değeri (%64.3, %66.7) olarak ölçülmüştür. Bu değerlere erişilmesini sağlayan öznitelikler, TF (Eş. 2.1), uzunluk (Eş. 2.10), göreceli cümle pozisyonu, uzunluk (Eş. 2.11), tamlamalar (PP) (Eş. 2.15), tamlamalar (NP) (Eş. 2.15) ve TF-ISF (4-gram) (Eş. 2.3) olarak tespit edilmiştir.

Metin özetlerinin değerlendirilmesinde temel kabul edilen ROUGE ölçümlerinin yanında problem cümle sınıflandırma problemi olarak ele alıp standart sınıflandırma başarısı da ölçülmüştür. Elde edilen sonuçlar Çizelge 4.11'de sunulmuştur.

Çizelge 4.11. Bulanık sistemlerin cümle sınıflandırma başarısı

	Eğitim başarısı (%)				Test başarısı (%)			
n	D	A	Doğruluk	ROC-AUC	D	A	Doğruluk	ROC-AUC
4	36.0	45.8	67.8	61.4	35.0	46.6	67.6	61.1
5	36.8	45.1	68.6	61.9	35.1	43.6	68.0	61.1
6	37.8	47.5	69.0	62.5	36.0	45.8	68.3	61.5
7	35.0	45.4	66.9	60.8	33.5	44.6	66.4	59.7
8	33.5	38.3	67.2	58.9	31.4	36.4	66.5	57.6
9	35.6	42.7	68.0	60.4	29.6	37.5	64.8	56.5
10	36.9	43.1	68.9	61.0	30.3	36.8	65.5	56.9
11	36.7	43.8	68.6	61.0	30.7	37.9	65.6	57.2
12	41.4	43.2	71.9	63.2	32.3	33.7	68.0	57.6
13	44.6	45.3	73.5	65.2	32.9	33.6	68.2	57.5
14	45.2	46.3	73.7	65.8	31.4	31.6	67.6	56.4
15	48.4	49.9	75.3	68.1	31.2	32.0	67.3	56.3
16	49.2	51.3	75.7	69.0	31.3	31.8	67.4	56.7
17	53.3	56.6	77.8	72.6	32.5	33.1	68.0	56.8
18	54.0	56.1	78.0	72.6	32.1	32.0	67.9	56.7
19	57.1	57.9	79.6	74.4	31.0	30.1	68.0	56.0
20	57.4	58.7	79.8	74.7	30.7	30.1	67.8	55.9
21	57.8	59.3	80.0	75.2	30.5	30.1	67.7	55.9

Çizelge 4.11'de görüldüğü gibi duyarlık (D), anma (A), doğruluk ve sınıflandırma eşik değerinden etkilenmeyen ROC (Receiver Operating Characteristics) eğrisinin altında kalan alan (ROC-AUC) ölçülmüştür. Ancak bu ölçütler içerisinde anma ve ROC-AUC sonuçları daha çok öneme sahiptir. ROUGE ölçümlerinin sonuçlarına benzer şekilde bu ölçüm sonucunda da beş-katlı çapraz geçerleme testinde en yüksek anma ve ROC-AUC değerleri 6, 7 öznitelik kullanılan bulanık sistemlerden elde edilmiştir.

4.2.2. Yapay sinir ağının SIGIR2018 veri kümesinde uygulanması

SIGIR2018 veri kümesine çok katmanlı bir yapay sinir ağı uygulanarak cümle önem derecesinin belirlenmesi ve ardından önemli cümlelerin özete dâhil edilmesi ile otomatik metin özetlemenin gerçekleştirilmesi sürecinde sırasıyla aşağıdaki adımlar uygulanmıştır:

- Kaynak metinlerden sözdizimsel özniteliklerin çıkarılması
- Girdi matrisinin oluşturulması
- Cümle önem derecelendirmesinde yapay sinir ağı gerçekleştirimi
- Elde edilen özetlerin değerlendirilmesi

Kaynak metinlerden sözdizimsel özniteliklerin çıkarılması

Bu tez çalışmasında bu öznitelikler TF, TF-ISF, uzunluk, pozisyon, başlık, anahtar sözcükler, cümle-cümle ilişkisi, şebeke bağlantı sıklığı, toplam benzerlik, tamlamalar, özel isimler olarak ele alınmış ve bunlara ait farklı hesaplama şekilleri ile toplam 42 sözdizimsel öznitelik tanımlanmıştır. Bu özniteliklerle ilgili ayrıntılı bilgiye ve hesaplama şekillerine Bölüm 2.2.1'de yer verilmiştir. Çalışmanın bu aşamasında, kaynak dokümanlardaki her bir cümle için Bölüm 2.2.1'de verilmiş olan sözdizimsel özniteliklere ait cümle öznitelik değerleri hesaplanmıştır.

Girdi matrisinin oluşturulması

Aday cümle kümelerinin kesişimi olan $\mathrm{Ext}_{\cap}$ çıkarmalı özetler, yazarlar tarafından hazırlanmış olan soyutlamalı özetler kadar cümle içermekte ve kaynak dokümandaki cümle sayısının yaklaşık %23'üne karşılık gelmektedir. Bu özetlerdeki cümleler tüm okuyucular tarafından özete değer olarak seçilmiş cümlelerdir. Tüm öznitelikler çıkarıldıktan sonra girdi matrisinin oluşturulmasında $\mathrm{Ext}_{\cap}$ tipi özetlerin etiketleri kullanılmıştır.

Yapay sinir ağı ile metin özetleme çalışmasında da bulanık sistem gerçekleştiriminde olduğu gibi değişken öznitelik vektörü üzerinden çok sayıda deney yapılmıştır. Öznitelikler χ^2 değerlerine göre Çizelge 4.7'deki gibi sıralanmış ve yapay sinir ağının girişleri, girdi özniteliklerin sonuca etkisini araştırmak adına en önemli 4 öznitelikten 42 özniteliğe kadar genişletilmiştir.

Cümle önem derecelendirmesinde yapay sinir ağı gerçekleştirimi

Dört öznitelikten başlanmak üzere, 42 özniteliğe kadar toplam 39 veri kümesi, her bir satır cümleleri her bir sütun ise özniteliklerin aldıkları değerleri gösterecek şekilde oluşturulmuştur. Bu 39 veri kümesinin her biri bir giriş, iki ara, bir çıkış katmanından oluşan

yapay sinir ağının eğitiminde kullanılmıştır. Veri kümesindeki öznitelik sayısı n olmak üzere yapay sinir ağının gerçekleştirimine ilişkin detaylı bilgi Çizelge 4.12'de sunulmuştur.

Çizelge 4.12. Yapay sinir ağı sistem mimarisi ve parametreleri

Katmanlar	Giriş Katmanı	Ara Katman 1	Ara Katman 2	Çıkış Katmanı
Sinir hücresi sayısı	n	$2.n$	n	1
Üyelik fonksiyonları	RELU	RELU	RELU	Sigmoid
Başlangıç ağırlık vektörü	Xavier uniform [119]	Xavier uniform [119]	Xavier uniform [119]	Xavier uniform [119]
Başlangıç önyargı vektörü	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$
Tür			Geri besleme	
Kayıp fonksiyonu			İkili capraz entropi	
En iyileştirme			Adam [120]	
En iyileştirme ölçütleri			Doğruluk, ROC-AUC	
Erken durdurma minimum hata farklı			1.00E-07	
Erken durdurma beklenen iterasyon sayısı			10	

Yapay sinir ağından elde edilen değer cümlenin önemli kabul edilme derecesi olarak alınmış ve önem derecesi 0.5 sınıflandırma eşik değerinden yüksek olan cümleler özete dâhil edilmiştir.

Elde edilen özetlerin değerlendirilmesi

Bölüm 4.2.1'de yer alan bulanık sistem gerçekleştiriminde olduğu gibi özet değerlendirme beş-katlı çapraz geçerleme ile yapılmıştır. Benzer şekilde eğitim ve test kümeleri cümle tabanlı değil doküman bazlı ayrıştırılmıştır. Toplam 125 yayından oluşan SIGIR2018'in rastgele seçilen 100 dokümanı eğitim kümesi olarak değerlendirilmiş, kalan 25 doküman test kümesi olarak özetlemeye tabi tutulup özetin başarısının ölçülmesinde kullanılmıştır. Özet başarısının değerlendirilmesinde temel alınan ölçüm yöntemi ROUGE analizidir. ROUGE ölçümlerinde, kaynak metnin tümü ve aday cümle kümelerinin kesişimi olan $\text{Ext}_{\cap}$ çıkarmalı özetleri olmak üzere iki tür referans metin üzerinde çalışılmış ve sonuçlar Çizelge 4.13 ve Çizelge 4.14'te sunulmuştur.

Çizelge 4.13. Yapay sinir ağının kaynak metne göre metin özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
n	D	A	D	A	D	A	D	A	D	A	D	A
4	52.4	9.9	51.9	7.0	52.4	9.9	44.0	8.8	43.7	6.4	44.0	8.8
5	62.4	12.4	61.8	8.8	62.4	12.4	58.4	12.5	57.8	9.0	58.4	12.5
6	70.8	15.8	69.8	11.4	70.8	15.8	79.2	16.6	78.3	12.1	79.2	16.6
7	65.6	12.2	65.0	8.7	65.6	12.2	63.2	11.6	62.7	8.4	63.2	11.6
8	62.4	12.5	61.6	8.9	62.4	12.5	63.2	14.1	62.3	10.3	63.2	14.1
9	65.4	14.2	64.6	10.3	65.4	14.2	68.0	15.0	67.2	11.1	68.0	15.0
10	65.0	14.1	64.1	10.1	65.0	14.1	69.6	14.0	68.8	10.3	69.6	14.0
11	80.6	17.6	79.4	12.6	80.6	17.6	79.2	18.1	78.1	13.2	79.2	18.1
12	79.2	20.4	77.8	15.0	79.2	20.4	84.0	21.9	82.6	16.2	84.0	21.9
13	77.4	16.2	76.3	11.6	77.4	16.2	73.6	16.3	72.5	11.7	73.6	16.3
14	84.8	19.5	83.4	14.1	84.8	19.5	83.2	17.9	81.9	12.7	83.2	17.9
15	78.8	18.6	77.4	13.5	78.8	18.6	78.4	20.1	76.9	15.0	78.4	20.1
16	94.0	23.3	92.0	17.1	94.0	23.3	95.2	24.7	93.0	18.2	95.2	24.7
17	91.8	22.0	90.0	16.2	91.8	22.0	88.8	21.9	86.9	16.3	88.8	21.9
18	91.6	22.1	89.7	16.4	91.6	22.1	93.6	23.2	91.5	17.4	93.6	23.2
19	97.8	27.6	95.1	20.5	97.8	27.6	97.6	27.2	95.0	20.2	97.6	27.2
20	88.8	25.7	86.6	19.3	88.8	25.7	86.4	23.5	84.3	17.7	86.4	23.5
21	97.8	29.2	95.2	22.1	97.8	29.2	97.6	27.9	95.0	21.0	97.6	27.9
22	98.2	29.0	95.5	21.8	98.2	29.0	96.8	27.8	94.2	21.1	96.8	27.8
23	98.8	29.7	95.9	22.2	98.8	29.7	97.6	29.5	94.6	22.1	97.6	29.5
24	99.2	32.7	96.2	24.8	99.2	32.7	99.2	34.8	96.2	26.7	99.2	34.8
25	99.2	34.7	96.0	26.7	99.2	34.7	99.2	33.4	96.0	25.4	99.2	33.4
26	99.4	33.5	96.2	25.5	99.4	33.5	99.2	33.4	96.3	25.5	99.2	33.4
27	99.2	35.3	96.1	27.0	99.2	35.3	100	36.9	97.0	28.5	100	36.9
28	100	32.0	96.9	24.1	100	32.0	99.2	32.7	96.1	24.7	99.2	32.7
29	100	35.9	96.8	27.6	100	35.9	99.2	33.3	96.3	25.6	99.2	33.3
30	99.4	36.4	96.2	28.0	99.4	36.4	100	37.9	96.8	29.4	100	37.9
31	99.8	34.5	96.6	26.3	99.8	34.5	100	32.6	96.9	24.8	100	32.6
32	100	38.5	96.6	29.9	100	38.5	100	40.1	97.0	31.7	100	40.1
33	100	35.7	96.7	27.3	100	35.7	100	34.4	97.0	26.7	100	34.4
34	100	35.9	96.7	27.5	100	35.9	97.6	33.3	94.6	25.4	97.6	33.3

Çizelge 4.13. (devam) Yapay sinir ağının kaynak metne göre metin özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
n	D	A	D	A	D	A	D	A	D	A	D	A
35	100	36.8	96.6	28.3	100	36.8	98.4	35.0	95.2	27.0	98.4	35.0
36	100	37.2	96.6	28.7	100	37.2	100	37.0	96.8	29.1	100	37.0
37	100	37.0	96.6	28.5	100	37.0	99.2	36.0	96.0	28.0	99.2	36.0
38	100	37.9	96.6	29.3	100	37.9	99.2	35.4	96.0	27.4	99.2	35.4
39	100	38.3	96.6	29.7	100	38.3	99.2	35.7	96.0	27.8	99.2	35.7
40	100	37.8	96.6	29.2	100	37.8	100	36.7	96.6	28.5	100	36.7
41	100	38.1	96.6	29.5	100	38.1	100	35.6	96.9	27.4	100	35.6
42	100	36.8	96.6	28.3	100	36.8	100	34.3	97.0	26.6	100	34.3

Çizelge 4.13'te SIGIR2018 veri kümesinde sözdizimsel öznitelikler üzerinden yapılan metin özetlemede yapay sinir ağlarının ROUGE analiz sonuçları kaynak metnin referans alınması ile elde edilmiştir. Burada elde beş-katlı çapraz geçerlemeden elde edilen maksimum ROUGE-1 anma değeri %40.1, ROUGE-2 anma değeri %31.7 ve ROUGE-3 anma değeri %40.1 olarak ölçülmüştür. Çizelge 4.14'te yapay sinir ağının metin özetleme performansı, aday cümle setlerinin kesişimi olan çıkarmalı özetlerin referans alınmasıyla yapılan ROUGE ölçümleri ile sunulmuştur.

Çizelge 4.14. Yapay sinir ağının hedef özetlere göre metin özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
n	D	A	D	A	D	A	D	A	D	A	D	A
4	34.9	16.2	29.4	12.0	32.6	15.0	25.1	13.2	20.1	9.4	23.2	12.1
5	41.1	20.2	34.7	15.1	38.5	18.9	36.3	20.4	29.7	15.4	33.9	18.9
6	47.0	25.9	39.7	19.8	44.1	24.2	48.2	25.7	39.5	19.1	44.5	23.7
7	44.4	20.7	37.7	15.7	41.7	19.3	39.9	18.8	33.0	14.0	37.1	17.4
8	41.4	20.5	34.7	15.3	38.6	19.1	37.1	21.6	30.1	15.8	34.5	19.8
9	43.1	23.0	36.4	17.4	40.4	21.5	41.4	22.9	34.4	17.2	38.4	21.1
10	42.3	23.0	35.2	17.4	39.6	21.5	41.9	21.8	34.4	16.0	38.7	20.0
11	53.2	29.0	44.7	22.0	49.5	26.9	50.0	28.8	42.0	22.0	46.6	26.7
12	52.7	33.7	44.7	26.4	49.4	31.6	48.7	32.8	39.2	24.6	44.7	30.1

Çizelge 4.14. (devam) Yapay sinir ağının hedef özetlere göre metin özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
n	D	A	D	A	D	A	D	A	D	A	D	A
13	52.9	28.4	45.1	22.0	49.9	26.8	46.0	25.2	37.2	18.6	42.3	23.0
14	58.1	33.2	49.8	26.0	54.9	31.3	50.9	27.0	41.7	19.7	47.4	25.0
15	56.7	33.1	49.3	26.5	53.9	31.5	46.9	31.0	38.1	23.7	43.4	28.6
16	70.6	44.7	62.6	36.9	67.4	42.7	57.0	38.3	44.8	28.8	52.0	35.0
17	74.5	44.8	68.6	38.3	72.5	43.6	54.1	33.3	44.0	25.5	49.8	30.6
18	70.3	42.8	62.8	35.8	67.7	41.2	57.3	35.8	46.0	27.0	52.5	32.9
19	82.4	59.8	76.9	53.2	80.9	58.7	58.8	41.9	47.2	31.8	54.2	38.5
20	71.5	51.8	65.3	45.3	69.6	50.4	48.8	35.1	37.8	26.4	43.9	31.7
21	80.2	60.3	74.0	53.5	78.4	59.0	58.2	43.1	46.2	33.6	53.2	39.6
22	78.9	60.0	72.5	53.3	77.1	58.7	59.2	42.9	47.8	33.6	54.7	39.8
23	88.2	68.5	84.1	62.9	87.4	67.9	55.1	42.5	43.4	31.9	50.0	38.6
24	88.0	74.6	84.2	69.5	87.2	74.1	56.2	49.8	44.4	38.4	51.4	45.6
25	89.6	80.8	86.4	76.9	89.1	80.4	54.4	47.3	41.9	35.7	48.9	42.5
26	89.9	78.5	86.6	73.9	89.4	78.1	53.7	46.8	41.0	35.4	48.7	42.6
27	86.9	79.3	83.3	74.6	86.3	78.8	56.2	52.9	44.8	41.2	51.4	48.5
28	93.8	77.7	91.4	73.4	93.5	77.4	57.4	48.0	45.8	37.1	52.6	43.9
29	92.2	85.3	89.6	81.7	91.9	85.0	56.0	47.7	43.7	36.4	51.3	43.8
30	90.6	85.5	88.1	82.1	90.2	85.1	56.3	54.3	44.4	42.6	51.4	49.6
31	95.5	85.3	93.7	82.2	95.3	85.2	52.9	45.4	40.5	34.2	47.8	41.2
32	93.2	93.0	91.2	90.9	93.0	92.9	51.5	53.6	39.1	41.7	46.9	48.9
33	95.1	87.7	93.3	84.7	95.0	87.5	55.4	48.9	43.3	38.0	50.6	44.7
34	98.6	91.7	97.8	89.9	98.6	91.7	53.3	46.8	41.1	35.6	48.7	42.7
35	97.7	93.4	96.8	91.9	97.6	93.4	54.8	50.3	42.6	39.0	49.8	45.8
36	97.8	94.8	97.1	93.7	97.8	94.8	54.9	51.7	42.6	39.9	49.6	46.7
37	98.8	95.0	98.2	94.0	98.7	95.0	54.6	51.5	42.1	40.0	49.6	47.0
38	96.8	95.3	96.0	94.2	96.8	95.3	54.6	50.0	42.1	38.1	49.5	45.3
39	97.2	96.7	96.4	95.8	97.2	96.6	54.2	49.6	41.4	37.9	49.0	45.0
40	97.4	95.6	96.6	94.5	97.3	95.5	54.7	51.9	42.3	40.3	49.4	47.0
41	98.9	98.0	98.5	97.5	98.9	98.0	55.1	51.2	42.9	39.4	50.3	46.7
42	98.6	94.5	97.9	93.2	98.5	94.4	54.8	48.3	42.4	37.4	49.9	44.0

Bulanık sistemlerde 6 öznitelik ile erişilen ROUGE değerlerine yapay sinir ağında söz konusu parametre kümesi için erişilememiştir. Bu bağlamda doğru kural kümelerinin tanımlanması durumunda özet çıkarmada bulanık sistem başarısının daha yüksek olduğu söylenebilir. Genel olarak, yapılan tüm deneyler için oldukça düşük ROUGE değerlerine sahip özetler üretilmiş olmakla birlikte, uygulanan bu yapay sinir ağ modeli için en yüksek başarı 30 ve 32 özniteliğin kullanımı ile elde edilmiş ve ROUGE-1 anma değerleri kaynak dokümana göre (%37.9, %40.1) ve $\text{Ext}_{\cap}$ çıkarmalı özetlerine göre (%54.3, %53.6) olarak ölçülmüştür.

Çizelge 4.15'te de uygulanan yapay sinir ağının standart cümle sınıflandırma performans ölçümleri sunulmuştur. ROUGE ölçüm sonuçları gibi bu değerlendirmenin sonuçları da bulanık sistemden elde edilen performans değerlendirme değerlerinin belirgin oranda gerisindedir.

Çizelge 4.15. Yapay sinir ağının cümle sınıflandırma başarısı

	Eğitim başarısı (%)				Test başarısı (%)			
n	D	A	Doğruluk	ROC-AUC	D	A	Doğruluk	ROC-AUC
4	57.2	7.1	76.3	73.0	37.6	5.5	76.1	72.6
5	56.5	8.8	76.4	73.0	46.3	9.2	76.4	72.6
6	54.4	12.3	76.5	73.5	50.5	12.1	76.1	72.8
7	56.1	9.5	76.6	73.8	50.9	8.7	76.7	72.9
8	57.1	9.6	76.5	73.8	60.7	10.4	75.9	72.7
9	55.3	10.8	76.5	73.8	49.8	11.2	75.9	72.6
10	0.0	11.0	76.5	74.4	0.0	10.4	76.1	72.3
11	54.6	14.3	76.7	74.4	50.6	14.2	76.5	72.4
12	55.3	18.4	76.9	74.3	46.4	16.1	75.4	72.3
13	58.5	15.4	77.3	76.0	46.5	12.1	76.2	71.9
14	59.1	18.5	77.3	75.7	45.5	13.2	76.2	72.1
15	60.9	19.6	77.9	76.8	49.7	16.6	75.6	71.2
16	64.4	29.3	79.4	79.9	45.0	22.2	75.4	70.5
17	74.0	31.9	81.1	82.4	47.3	20.0	75.9	68.4
18	68.7	29.0	79.9	80.6	46.8	20.5	75.2	69.7
19	78.1	46.7	84.2	88.3	43.1	26.1	74.7	67.8
20	72.6	39.0	81.9	84.8	42.5	21.6	74.3	69.9

Çizelge 4.15. (devam) Yapay sinir ağının cümle sınıflandırma başarısı

	Eğitim başarısı (%)				Test başarısı (%)			
n	D	A	Doğruluk	ROC-AUC	D	A	Doğruluk	ROC-AUC
21	73.7	46.6	82.9	86.9	43.9	26.8	74.2	68.0
22	73.5	46.6	83.4	86.5	44.9	27.6	74.7	68.7
23	85.7	58.1	87.7	92.3	38.8	25.9	73.1	65.0
24	84.9	65.2	88.9	93.8	39.2	32.1	71.8	66.3
25	86.4	74.0	91.0	95.7	36.0	28.8	71.6	64.0
26	87.1	69.2	90.2	94.8	37.4	29.0	71.9	64.7
27	83.4	71.2	89.6	94.8	38.5	35.3	71.7	65.7
28	91.7	71.1	91.5	96.7	39.9	31.4	73.0	67.1
29	88.7	78.5	92.4	97.2	38.8	32.2	71.9	65.3
30	88.2	80.3	92.6	97.1	37.4	35.2	71.0	66.0
31	93.5	79.7	93.8	97.9	35.9	28.4	71.4	64.2
32	90.6	89.8	95.2	98.8	33.7	35.3	68.7	63.8
33	92.8	82.5	94.3	98.2	37.9	32.3	71.5	65.0
34	97.9	89.3	97.0	99.6	36.1	30.1	71.2	63.6
35	96.8	91.0	97.2	99.5	36.8	32.2	70.9	64.9
36	96.7	93.4	97.6	99.7	36.3	35.5	70.3	63.6
37	98.4	93.0	98.0	99.7	37.7	34.4	71.4	63.7
38	96.0	93.9	97.6	99.6	36.4	33.9	70.7	63.6
39	96.5	95.5	98.1	99.7	36.2	33.5	70.5	64.6
40	96.8	94.1	97.8	99.7	36.1	33.9	70.5	64.6
41	98.4	97.2	99.0	99.9	37.6	34.2	71.5	64.6
42	98.2	92.5	97.8	99.6	36.6	30.8	70.6	64.0

4.2.3. Derin öğrenme yönteminin SIGIR2018 veri kümesinde uygulanması

SIGIR2018 veri kümesi üzerinde sözdizimsel özniteliklerin kullanımı ile gerçekleştirilen bulanık kural tabanlı sistem ve yapay sinir ağının yanı sıra anlamsal özniteliklerin kullanımına yoğunlaşmış olan ve cümle derecelendirme için derin öğrenme yaklaşımlarından faydalanan üç tane ÖSA tabanlı modelin gerçekleştirimi de yapılmıştır. Bu ÖSA modellerinin temel aldıkları mimari, 2017 yılında geliştirilmiş olan ve çıkarmalı metin özetlemede çoğu modelden daha yüksek özetleme başarısına sahip olan SummaRuNNer isimli modeldir [34]. SummaRuNNer, özyerleşiklere dayanan anlamsal özniteliklerin

yanında cümle önem derecelerinin belirlenmesinde sözdizimsel öznitelikler olan mutlak ve göreceli iki pozisyon özniteliğini kullanmaktadır. Bu nedenle SIGIR2018 veri kümesinden elde edilecek SummaRuNNer özetlerinin başarısı, tez çalışmasında referans kabul edilmiştir.

Bölüm 3.3'te SummaRuNNer mimarisinin ayrıntılarına yer verilmiştir. Bu mimaride giriş katmanı, kelime katmanı, cümle katmanı ve sınıflandırma katmanı olmak üzere hiyerarşik bir yapı bulunmaktadır. Orijinal SummaRuNNer mimarisinde kelime ve cümle katmanlarında çift yönlü ÖSA (ς-GRU) kullanılmaktadır. Ancak tez bu çalışmasında, bu mimari üzerinde bazı ek düzenlemeler yapılmıştır. Kelime katmanındaki ÖSA'nın yerine ESA ve odak tabanlı ÖSA olmak üzere iki farklı model yerleştirilerek toplamda üç ayrı modelin gerçekleştirimi yapılmıştır.

ESA & ς-GRU modelinde kelime katmanında art arda iki evrişimli katmandan oluşan bir ESA çalışmaktadır. Cümle katmanında yine ς-GRU uygulanmakta ve ardından sınıflandırma işlemi aynı şekilde yapılmaktadır. Öte yandan, odak tabanlı hiyerarşik ÖSA'da her bir ς-GRU katmanı arasında bir odak katmanı eklenmiştir. Kelime katmanının ardından eklenen odak katmanı, her bir cümledeki kelimelerin, cümle katmanının ardına eklenen dikkat katmanı ise dokümandaki cümlelerin ağırlıklarını hesaplamaktadır. Ardından ileri besleme aşamasında kelime katmanından elde edilen ς-GRU çıktısı olan cümle özyerleşikleri bu cümledeki kelimelerin ağırlıkları ile çarpılarak kelime katmanına iletilmektedir. Benzer şekilde, cümle katmanındaki ς-GRU'dan elde edilen gizli durumların birleşimi ile cümle ağırlıkları çarpımından doküman temsili elde edilmektedir. Doküman temsili ve cümle katmanının çıktısı alınarak sınıflandırma katmanı orijinal SummaRuNNer modelindeki gibi sınıflandırma amacıyla kullanılmaktadır.

Bu üç modelin de girdisi her biri önceden eğitilmiş word2vec kelime özyerleşikleridir. Her bir kelime özyerleşiği 100 elemandan oluşan bir dizidir. Modellerin SIGIR2018 veri kümesinde testi sonucunda her bir cümlenin bir önem derecesi elde edilir ve bu değerin 0.5 eşik değerinin üstünde olması halinde cümle özete dâhil edilir.

SummaRuNNer'ın beş-katlı çapraz geçerleme ile gerçekleştirimi sonucunda ortaya çıkan özetlerin eğitim ve test ROUGE değerleri Çizelge 4.16'da sunulmuştur.

Çizelge 4.16. SIGIR2018 verisi üzerinde SummaRuNNer'in metin özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
Model	D	A	D	A	D	A	D	A	D	A	D	A
Odak tabanlı ς-GRU & ς-GRU	83.2	92.9	80.4	90.5	83.2	92.9	53.5	60.7	41.6	48.9	49.1	56.1
ESA & ς-GRU	82.9	92.1	80.5	90.3	82.9	92.1	52.6	59.2	40.2	46.3	47.6	54.1
ς-GRU & ς-GRU	83.6	92.9	80.6	90.6	83.6	92.9	53.4	62.6	42.5	51.0	49.5	58.3

Çizelge 4.16'da görüldüğü gibi en başarılı özetlerin SummaRuNNer'ın orijinal mimarisinden (ς-GRU & ς-GRU) elde edildiği görülmektedir. SummaRuNNer orijinal mimarisinden, SIGIR2018 veri kümesinin üzerinden elde edilen maksimum ROUGE-1 anma değeri % 62.6, duyarlılık değeri ise %53.4 olarak ölçülmüştür. ROUGE-2 için bu değerler sırasıyla %51.0 ve %42.5; ROUGE-3 için ise %58.3 ve %49.5'tir. SummaRuNNer, metin özetleme problemine kelime özyerleşiklerine dayalı anlamsal özniteliklerin haricinde pozisyon bilgisinin üzerine kurgulanmış birkaç sözdizimsel özniteliği de dâhil etmesi sebebiyle bu tez çalışmasındaki performans karşılaştırmalarında temel alınan yöntemlerden birisidir.

4.2.4. TextRank algoritmasının SIGIR2018 veri kümesinde uygulanması

TextRank algoritmasına göre metindeki cümleler şebekedeki düğümlere, cümleler arası benzerlik ise olası kenarlara karşılık gelmektedir. Bir düğüm, benzerlik değerine göre ağırlık değeri atanmış olan kenarlarının toplam ağırlığı kadar önem derecesine sahiptir. Algoritmanın uygulanmasında model çıktısı olan özet metnin içerisinde yer alması istenen cümle sayısı 4, 8, 12 ve 16 olmak üzere dört değişken değer almaktadır. 8 cümleli özetler, aday cümle kümelerinin kesişimi olan çıkarmalı özetlerin, 16 cümleli özetler ise aday cümle kümelerinin birleşimine karşılık gelen çıkarmalı özetlerin içerdikleri ortalama cümle sayısının yukarı yuvarlanmış tamsayı değeriyle ilişkilidir. Bu özelliğiyle TextRank, kaynak metnin referans alındığı ROUGE değerleri açısından avantajlı konumdadır. Çizelge 4.17'de

n cümle içeren TextRank özetlerinin, kaynak metne, aday cümle kümelerinin birleşimi ve kesişimi olan çıkarmalı özetlere göre ROUGE değerleri sunulmuştur.

Çizelge 4.17. TextRank ile elde edilen metin özetlerinin ROUGE analizi sonuçları

		ROUGE-1			ROUGE-2			ROUGE-L		
Referans	Özet metin	F	D	A	F	D	A	F	D	A
Kaynak metin	$TextRank_4$	36.7	100.0	22.5	27.8	98.2	16.2	36.7	100.0	22.5
	$TextRank_8$	52.2	100.0	35.3	43.9	98.0	28.3	52.2	100.0	35.3
	$TextRank_{12}$	65.0	100.0	48.2	58.4	98.2	41.6	65.0	100.0	48.2
	$TextRank_{16}$	75.0	100.0	60.0	70.1	98.4	54.5	75.0	100.0	60.0
$Ext_{\cup}$	$TextRank_4$	44.2	85.7	29.8	33.2	75.2	21.3	43.5	84.3	29.3
	$TextRank_8$	57.8	81.6	44.8	47.5	71.3	35.6	57.1	80.6	44.2
	$TextRank_{12}$	66.8	78.1	58.4	57.7	68.3	49.9	66.1	77.3	57.8
	$TextRank_{16}$	72.6	75.4	70.0	64.6	66.2	63.0	72.2	74.9	69.6
$Ext_{\cap}$	$TextRank_4$	40.6	52.9	33.0	23.6	32.2	18.6	38.1	49.7	30.9
	$TextRank_8$	48.6	49.3	47.9	32.6	32.6	32.6	46.4	47.1	45.7
	$TextRank_{12}$	52.2	45.8	60.6	37.6	31.6	46.3	50.7	44.5	58.9
	$TextRank_{16}$	54.0	43.3	71.6	40.8	30.9	60.1	53.1	42.6	70.5

Çizelge 4.17'de $Ext_{\cup}$, aday cümle kümelerinin birleşimi olan çıkarmalı özetleri, $Ext_{\cap}$ aday cümle setlerinin kesişimi olan çıkarmalı özetleri, $TextRank_n$ ise *n* cümle içeren TextRank otomatik özetlerini temsil etmektedir.

Özette yer alan cümlelerin sayısındaki artış ROUGE değerlerini arttırmaktadır. Çizelge 4.17'de dikkat edilmesi gereken başlıca husus, aday cümle kümelerinin kesişimi olan çıkarmalı özetlerin muadili olan 8 cümleli özetlerin kaynak dokümana göre ROUGE-1 anma değerinin %35.3 olduğudur. Bu değer aday cümle kümelerinin birleşimi olan çıkarmalı özetlerin muadili olan 16 cümleli TextRank özetleri için %60.0'dır.

5. GELİŞTİRİLEN METİN ÖZETLEME ÇÖZÜM MİMARİSİ

Çıkarmalı metin özetleme, metinde özete dâhil edilmeye değer önemli cümlelerin belirlenmesi sürecidir. Bunun için metinden elde edilen çeşitli özelliklerin kullanımı ile cümle önem derecelendirilmesinin yapılması gerekmektedir. Ardından önem derecesi tanımlanmış cümleler belirli bir eşik değere göre veya metindeki belirli sayıdaki en önemli cümlelerin seçimi ile özet dokümanı meydana getirir.

Bu tez çalışmasında metin özetleme problemi cümle derecelendirme ve cümle sınıflandırma olarak ele alınmıştır. Bu problemin, bir öznitelik uzayını hesaba katarak çözümü üzerine çok sayıda yöntem geliştirilmiş ve bu yöntemlere ait genel bilgilere ve özetleme başarılarına sırasıyla Bölüm 2.3 ve Bölüm 3'te yer verilmiştir.

Özetleme metodunun belirli yöntemlerle sınırlı kaldığı ancak eğitimde kullanılan verinin ve bu veriden elde edilen özniteliklerin farklılığa sahip olduğu bilinmektedir. Bu yöntemlerin tamamı sınıflandırma yapmak için metindeki cümleleri farklı özniteliklerle temsil etmektedir. Ancak, günümüzde halen metin özetleme problemi için kullanılması gereken giriş öznitelik uzayının nelerden oluşması gerektiği kesin olarak belirlenebilmiş değildir. Bu durumdan yola çıkılarak bu tez çalışmasında özetleme görevinde kullanılan öznitelikler sözdizimsel ve anlamsal olmak üzere iki grupta ele alınmıştır.

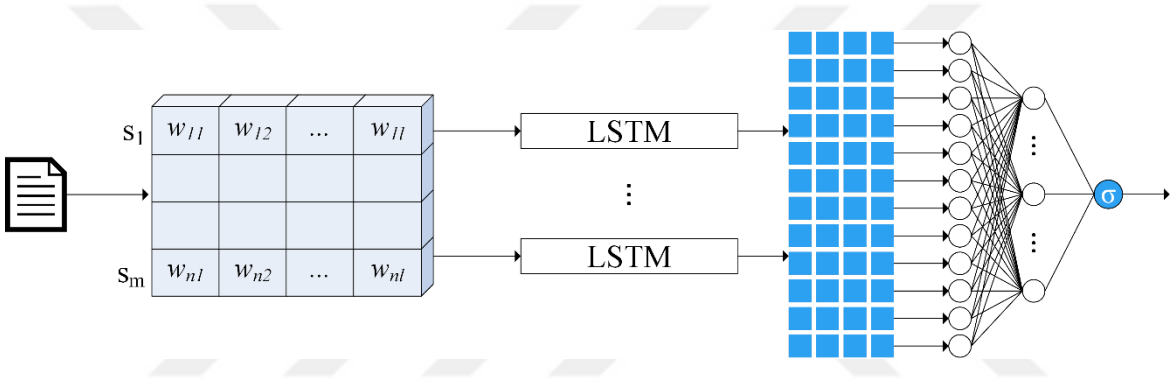
Bu bölümde, literatürdeki çalışmalardan derlenen tüm bu özniteliklerin ayrı ayrı ve birlikte kullanımının SIGIR2018 veri kümesinin özetleme başarısına etkisi incelenmiştir. Bir başka ifadeyle, cümle önem derecelendirmede sözdizimsel özniteliklerin, anlamsal özniteliklerin ve bu iki öznitelik vektörünün birlikte kullanımını ifade eden hibrit özniteliklerin, metin özetleme başarısına etkisi, ayrı başlıklar altında değerlendirilmiştir. Burada özetleme aşamasında cümle önem derecelendirme modeli olarak sıralı dizi modellemede LSTM ve cümle sınıflandırmada yapay sinir ağı altyapısına sahip bir modelin gerçekleştirimi yapılmıştır. Bu model bir LSTM katmanı ve bir de yapay sinir ağından (NN) oluştuğundan LSTM-NN olarak anılacaktır.

SIGIR2018 verisinde toplam 125 konferans bildirisi, cümlelerine göre ayrıştırılmış olarak bulunmaktadır. Bölüm 4.1'de açıklandığı gibi bu cümlelerin tümü özete dâhil edilmeye değer veya özete dâhil edilmeye değer değil olarak üç okur tarafından etiketlenmiş ve bu

etiketleme sürecinin sonucunda aday cümle kümelerinin birleşimi ve kesişimi olmak üzere iki tip çıkarmalı özet meydana getirilmiştir. Bu çalışmada aday kümelerinin kesişiminden oluşan $\text{Ext}_{\cap}$ çıkarmalı metin özetleri, performans ölçümlerinde referans alınacak çıkarmalı özet olarak değerlendirilmiştir.

5.1. Anlamsal Öznitelikler Üzerinden LSTM-NN

Anlamsal öznitelikler üzerinden yapılan metin özetlemede önceden eğitilmiş GloVe ve word2vec kelime özyerleşiklerinden yararlanılmıştır. Şekil 5.1'de bir dokümandaki cümlelerin metin formundan vektörel temsile nasıl geçirildiğinin görseli sunulmuştur.



Şekil 5.1. Anlamsal öznitelikler üzerinde LSTM-NN mimarisi

Şekil 5.1'de görülen üç boyutlu anlamsal özyerleşik kümesinde her satır cümleleri ifade etmektedir ve her sütun ilgili satıra karşılık gelen cümledeki her bir tekil kelimeye karşılık gelmektedir.

Kelimeler $[1,d]$ boyutlu kelime özyerleşikleri ile ifade edildiğinden söz konusu giriş matrisi d derinliğine sahiptir. i ve j birer tamsayı, $0<i\leq m$ ve m yığın boyutu ve $0<j\leq l$ ve l maksimum dizi boyu olmak üzere bir kelimenin kelime özyerleşiği w_{ij} ile ifade edilmektedir. Anlamsal öznitelik kümesi LSTM yığınına giriş olarak verilmektedir ve ardından her cümlenin anlamsal temsili ortaya çıkmaktadır. Bu temsil vektörü çok katlı algılayıcı ağa giriş olarak verilmekte ve cümle sınıflandırma yapılmaktadır.

GloVe özyerleşiklerinin $d=50$, $d=100$, $d=200$ ve $d=300$ elemanlı önceden eğitilmiş kelime vektörleri bulunmaktadır. Öte yandan önceden eğitilmiş word2vec kelime vektörleri $d=100$ elemanlıdır. GloVe özyerleşikleri ile word2vec özyerleşiklerinin eş modellere girdi

olarak sunulabilmesi amacıyla gerekli olduğunda ilgili özyerleşik boyuna lineer bir katmanla indirgenmesi veya genişletilmesi söz konusudur. Bu sayede orijinalinde 100 elemanlı olan word2vec özyerleşikleri $d=50$, $d=100$, $d=200$ ve $d=300$ olarak ölçeklenebilmiştir.

Anlamsal öznitelik matrisi, cümlelerin kelimelerden oluşan sıralı dizileri ifade edeceği şekilde tanımlanmıştır. Burada maksimum sıralı dizinin boyunun belirlenmesinde çeşitli yaklaşımlar üzerine çalışılmıştır. Bunlar, yığındaki minimum tekil kelime sayısı, yığındaki ortalama tekil kelime sayısı, yığındaki ortalama tekil kelime sayısı ile kelime sayılarının standart sapmasının toplamı ve yığındaki maksimum tekil kelime sayısıdır. Bu yaklaşımların her biri için özetleme başarısı ölçülmüş ve maksimum sıralı dizi boyunun yığındaki bir cümlede yer alan maksimum tekil kelime sayısı kadar olmasına karar verilmiştir.

Maksimum sıralı dizi boyundan daha az sayıda kelime içeren cümleler için sıralı dizinin boş kalan kısmının doldurulmasında kelime özyerleşikleri ile eşit boyda sıfır vektörleri kullanılmıştır. Sıfır vektörünün oluşturulmasında üç strateji kullanılmıştır. Bunlar sıfır vektörünün sıralı dizinin önüne yerleştirilmesi, sıfır vektörünün sıralı dizinin iki yanına eşit olarak yerleştirilmesi, sıfır vektörünün sıralı dizinin ardına yerleştirilmesidir. Tüm bu yaklaşımların özet çıkarmadaki başarısı ölçülmüş; en başarılı özetleme sıfır vektörünün sıralı dizinin ardına yerleştirilmesinden elde edilmiştir.

Hazırlanmış olan üç boyutlu anlamsal özyerleşik matrisi özyerleşik boyu kadar LSTM hücresine sahip katmana giriş olarak verilmiştir. Ardından bu katmandaki gizli durumlar, ilk katmanında özyerleşik matrisi boyu kadar sinir hücresi içeren çok katmalı bir yapay sinir ağına verilmiştir. Sınıflandırmanın son katmanında kullanılan üyelik fonksiyonu sigmoid, diğer katmanlardaki üyelik fonksiyonu RELU olarak belirlenmiştir.

Modelin eğitiminde en iyileştirilecek ölçütler doğruluk ve ROC-AUC'tur. En iyileştirme yöntemi Adam [120] ve kayıp fonksiyonu ikili çapraz entropi olarak seçilmiştir. Minimum hata farkının üst üste 10 kez 1.00E-07 değerinin altına düşmesi halinde eğitim erken durdurulmaktadır.

Anlamsal öznitelikler üzerinden gerçekleştirilen LSTM-NN'nin özetleme başarısı, SIGIR2018 veri kümesi üzerinde beş-katlı çapraz geçerleme ile ölçülmüştür. Bunun için beş-katlı çapraz geçerlemenin her bir adımında 100 doküman eğitimde kullanılmış, 25 doküman

ise modelin başarısının ölçülmesinde test kümesi olarak değerlendirilmiştir. Çizelge 5.1'de ve Çizelge 5.2'de elde edilen otomatik özetlerin ROUGE değerleri sunulmuştur.

Çizelge 5.1. Anlamsal özniteliklerle LSTM-NN'nin kaynak metne göre özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)					
Kelime vektörü	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
	D	A	D	A	D	A	D	A	D	A	D	A
G_{50}	100	38.4	96.6	29.8	100	38.4	99.2	38.5	96.1	30.3	99.2	38.5
G_{100}	100	38.4	96.6	29.8	100	38.4	100	41.9	96.8	33.6	100	41.9
G_{200}	100	38.4	96.6	29.8	100	38.4	99.2	40.1	96.2	31.9	99.2	40.1
G_{300}	100	38.4	96.6	29.8	100	38.4	99.2	41.6	96.0	33.6	99.2	41.6
W_{50}	100	38.7	96.6	30.1	100	38.7	100	39.6	97.0	31.7	100	39.6
W_{100}	100	38.5	96.6	29.8	100	38.5	100	38.4	96.9	30.4	100	38.4
W_{200}	100	38.4	96.6	29.8	100	38.4	100	41.4	97.0	33.2	100	41.4
W_{300}	100	38.4	96.6	29.8	100	38.4	100	39.0	96.9	31.1	100	39.0

Çizelge 5.1'de ilk sütun kullanılan özyerleşiğin türünü ve boyunu ifade etmektedir. G sembolü, GloVe özyerleşiklerini W sembolü ise word2vec özyerleşiklerini temsil etmektedir. Özyerleşiklere ait alt indisle gösterilen değerler özyerleşik boyunu ifade etmektedir.

Çizelge 5.1'de yer alan ROUGE değerleri kaynak dokümanının referans alınarak oluşturulmuştur. Otomatik özetlerin kaynak dokümanın yerini tutma derecesinin ölçülmesi amacıyla hesaplanmıştır. Buna göre kaynak dokümandaki bilgiyi en çok koruyan özetler GloVe vektörlerinin $d = 100$ ve $d = 300$ özyerleşik boyuyla kullanımından elde edilmiştir. Bunun yanında word2vec özyerleşiklerinin $d = 200$ boyunda ölçeklenmesiyle elde edilen model özetleri de yüksek ROUGE değerlerine sahiptir. Kaynak metnin referans alındığı beş-katlı çapraz geçerleme test sonuçlarına göre kelime özyerleşiklerine dayalı anlamsal özniteliklerden elde edilen maksimum ROUGE-1 anma değeri %41.9 olarak ölçülmüştür.

Çizelge 5.2'da yine GloVe ve word2vec özyerleşiklerinin değişken özyerleşik boylarının kullanımı ile elde ettikleri özetleme başarılarının aday cümle kümelerinin kesişimi olan $\text{Ext}_{\cap}$ çıkarmalı özetlerine göre ROUGE değerleri sunulmuştur.

Çizelge 5.2. Anlamsal özniteliklerle LSTM-NN'nin hedef özete göre özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)						
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L		
	D	A	D	A	D	A	D	A	D	A	D	A	Δ
G_{50}	99.7	99.4	99.6	99.3	99.7	99.4	54.8	54.1	41.9	42.0	49.4	49.3	2
G_{100}	99.9	99.7	99.9	99.7	99.9	99.7	53.8	58.3	41.2	46.6	49.0	53.6	2
G_{200}	99.9	99.8	99.9	99.8	99.9	99.8	55.9	57.4	43.8	45.8	51.3	52.8	3
G_{300}	100	99.8	100	99.8	100	99.8	55.8	60.5	44.4	50.2	51.8	56.8	3
W_{50}	99.0	99.6	98.8	99.4	99.0	99.6	54.7	55.3	42.4	44.1	50.3	51.1	1
W_{100}	99.8	99.8	99.7	99.7	99.8	99.8	57.9	57.4	46.2	46.8	53.6	53.4	1
W_{200}	99.9	99.6	99.8	99.6	99.9	99.6	54.4	58.3	42.1	47.0	49.8	53.8	2
W_{300}	100	99.8	99.9	99.8	100	99.8	58.1	57.3	46.7	46.6	54.0	53.3	3

Çizelge 5.2'de ilk sütun kullanılan özyerleşiğin türünü ve boyunu ifade etmekte, G, GloVe özyerleşiklerini W ise word2vec özyerleşiklerini temsil etmektedir. Özyerleşiklere ait alt indisle gösterilen değerler özyerleşik boyunu ifade etmektedir. Son sütundaki Δ değeri otomatik özetteki cümle sayısı ile referans alınan $\text{Ext}_{\cap}$ çıkarmalı özetlerinin içerdiği cümlelerin sayısının farkını göstermektedir.

Çizelge 5.2'de yer alan verilere göre en başarılı model GloVe'un $d = 300$ kelime özyerleşiklerinden elde edilmiştir. Burada erişilen ROUGE-1 anma değeri %60.5, Δ değeri ise 3'tür. Burada yüksek ROUGE değerine erişilmiş ancak hedef çıkarmalı özetlere göre daha fazla cümle barındıran özetler geliştirilmiştir. Öte yandan, GloVe'un $d = 100$ olan kelime özyerleşikleri ile word2vec'in $d = 200$ kelime özyerleşikleri daha küçük Δ değeri ile yüksek ROUGE değerlerine erişimi sağlamışlardır. Bu modellerden elde edilen ROUGE-1 değeri %58.3 seviyesindedir ve Δ değeri 2'dir.

Çıkarmalı metin özetleme problemi, cümle önem derecelendirme ve önemli cümleyi seçme olarak tanımlandığından performansı, cümle sınıflandırma açısından da değerlendirilmiştir. Çizelge 5.3'te anlamsal özniteliklerin kullanımı ile oluşturulan otomatik özetlerin sınıflandırma başarısına olan etkisi sunulmuştur.

Çizelge 5.3. Anlamsal özniteliklerle LSTM-NN'nin cümle sınıflandırma başarısı

	Eğitim başarısı (%)				Test başarısı (%)			
	D	A	Doğruluk	ROC-AUC	D	A	Doğruluk	ROC-AUC
G_{50}	99.6	99.4	99.8	99.7	36.8	37.6	70.0	64.3
G_{100}	99.9	99.7	99.9	99.9	35.8	40.9	69.2	62.3
G_{200}	100.0	99.8	99.9	99.9	37.5	40.4	70.6	66.4
G_{300}	100.0	99.9	100.0	100.0	39.4	44.7	70.8	68.3
W_{50}	98.7	99.5	99.6	99.8	37.1	40.2	70.2	65.4
W_{100}	99.8	99.9	99.9	99.9	39.1	40.3	71.5	66.0
W_{200}	99.8	99.7	99.9	99.9	37.9	42.9	70.2	67.1
W_{300}	99.9	99.9	100.0	99.9	39.8	41.8	71.3	68.1

Sınıflandırma eşik değeri 0.5 olarak atanmış ve en yüksek anma değeri, ROUGE tabanlı değerlendirmeye benzer şekilde GloVe'un $d = 300$ boyundaki özyerleşiklerinden elde edilmiştir. Benzer bir sonuç ROC-AUC değerleri için de elde edilmiştir.

Sonuç olarak, SIGIR2018 veri kümesinden anlamsal özniteliklerin tek başına kullanımı ile elde edilen maksimum ROUGE değeri %60.5 olarak bulunmuştur. Bu değer literatürdeki benzer çalışmalarla yaklaşık olarak aynı düzeydedir. Ancak tek başına kelime özyerleşiklerinin kullanımı üzerine yoğunlaşmış olan bu yaklaşıma dokümanın bazı yapısal özelliklerinin ve sözdizimsel cümle özelliklerinin eklenmesi anlamlı sonuçlar doğurabilir. Bu nedenle ilerleyen bölümlerde sırasıyla bu sözdizimsel özniteliklerin tek başına ve anlamsal özniteliklerle birlikte kullanımının metin özetlerine etkisi üzerine yapılan çalışmaların sonuçları da değerlendirilmiştir.

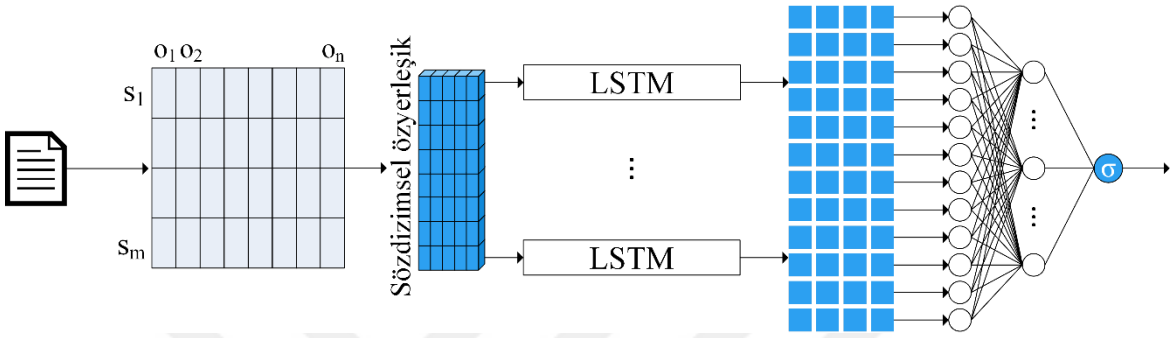
5.2. Sözdizimsel Öznitelikler Üzerinden LSTM-NN

Bölüm 4.2.1 ve Bölüm 4.2.2'de SIGIR2018 veri kümesi üzerinde metin özetleme probleminin sözdizimsel öznitelikle çözümünde sırasıyla bulanık sistemlerden ve yapay sinir ağlarından yararlanılmıştır. Bu yöntemlerin kullanmış olduğu sözdizimsel öznitelikler LSTM-NN modeli için de uygulanmış ve bu modelin özetleme başarısı ölçülmüştür.

Kaynak dokümanlardaki her bir cümle için sözdizimsel özniteliklere ait cümle öznitelik değerleri hesaplanmıştır. Buna göre her bir cümle için 42 öznitelik değeri atanmış ve bu

özniteliklere ait χ^2 değerlerine Çizelge 4.7'de yer verilmiştir. $4 \leq n \leq 42$ olmak üzere LSTM-NN'nin sözdizimsel öznitelikle gerçekleştiriminde maksimum χ^2 değerine sahip n öznitelikten oluşan öznitelik kümesi için ayrı deneysel çalışmalar yapılmıştır.

Sözdizimsel özniteliklerle LSTM-NN gerçekleştirimi Şekil 5.2'de sunulmuştur.



Şekil 5.2. Sözdizimsel öznitelikler üzerinde LSTM-NN mimarisi

Kullanılacak her bir öznitelik kümesinin LSTM-NN ile gerçekleştiriminde iki boyutlu öznitelik vektörü özyerleşik katmanına verilmiştir. Burada sözdizimsel özniteliklerin aldıkları gerçek değerler sabit boyutlu özyerleşikler olarak belirlenmiştir. Buna göre özyerleşik boyu bir niteliğin aldığı tekil değerlerin sayısı kadardır.

Sözdizimsel özniteliklerin özyerleşikler ile temsilinin ardından ortaya çıkan sözdizimsel öznitelik vektörü LSTM-NN'nin girdisini oluşturmuştur. LSTM katmanı sıralı cümlelerin modellemesinde görev yaparken ardından gelen yapay sinir ağı cümle sınıflandırma amacıyla kullanılmaktadır. Sınıflandırmanın son katmanında kullanılan üyelik fonksiyonu sigmoid, diğer katmanlardaki üyelik fonksiyonu RELU olarak belirlenmiştir.

LSTM-NN'nin eğitiminde en iyileştirilecek ölçütler doğruluk ve ROC-AUC'tur. En iyileştirme yöntemi Adam [120] ve kayıp fonksiyonu ikili çapraz entropi olarak seçilmiştir. Minimum hata farkının üst üste 10 kez 1.00E-07 değerinin altına düşmesi halinde eğitim erken durdurulmaktadır.

Sözdizimsel öznitelikler üzerinden yapılan LSTM-NN tabanlı deneylerde, yararlanılan öznitelikler en yüksek χ^2 değerine sahip 4 öznitelikten başlanmak üzere 42 özniteliğe kadar artırımlı olarak değiştirilmiştir. Her bir öznitelik kümesinin kullanımı için ayrı ayrı beş-katlı çapraz geçerleme yapılmıştır. Burada çapraz geçerlemenin eğitim ve test kümeleri, rastgele

seçilen 100 doküman eğitimde 25 doküman testte kullanılacak şekilde belirlenmiştir. Çizelge 5.4 ve Çizelge 5.5'te çapraz geçerleme sonucunda elde edilen metin özetlerinin ortalama ROUGE değerleri görülmektedir.

Çizelge 5.4. Sözdizimsel özniteliklerle LSTM-NN'nin kaynak metne göre özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
n	D	A	D	A	D	A	D	A	D	A	D	A
4	95.0	24.3	92.7	17.5	95.0	24.3	92.0	23.3	89.7	16.6	92.0	23.3
5	95.8	24.6	93.4	17.7	95.8	24.6	94.4	24.6	92.1	17.6	94.4	24.6
6	97.4	25.3	95.0	18.3	97.4	25.3	96.8	24.7	94.4	17.9	96.8	24.7
7	97.2	25.1	94.8	18.1	97.2	25.1	95.2	24.5	92.9	17.7	95.2	24.5
8	97.0	25.0	94.6	18.1	97.0	25.0	98.4	24.3	96.2	17.4	98.4	24.3
9	98.2	26.7	95.7	19.4	98.2	26.7	99.2	26.5	96.7	19.1	99.2	26.5
10	97.2	26.3	94.7	19.1	97.2	26.3	98.4	25.5	95.8	18.4	98.4	25.5
11	97.0	26.2	94.5	19.0	97.0	26.2	98.4	25.9	95.8	18.7	98.4	25.9
12	97.8	27.2	95.3	19.9	97.8	27.2	98.4	26.6	95.8	19.3	98.4	26.6
13	97.6	26.5	95.1	19.4	97.6	26.5	99.2	25.4	96.7	18.2	99.2	25.4
14	97.6	26.6	95.1	19.4	97.6	26.6	97.6	25.2	95.0	18.0	97.6	25.2
15	97.8	25.6	95.3	18.7	97.8	25.6	98.4	25.7	96.0	18.9	98.4	25.7
16	97.8	26.2	95.3	19.2	97.8	26.2	99.2	26.7	96.5	19.7	99.2	26.7
17	98.4	27.7	95.8	20.4	98.4	27.7	98.4	27.6	95.6	20.1	98.4	27.6
18	98.6	27.9	96.0	20.6	98.6	27.9	97.6	27.3	94.9	19.9	97.6	27.3
19	98.0	27.6	95.4	20.4	98.0	27.6	98.4	27.6	95.6	20.0	98.4	27.6
20	98.6	27.1	96.1	20.0	98.6	27.1	98.4	26.2	96.0	19.4	98.4	26.2
21	98.6	28.2	95.9	20.9	98.6	28.2	98.4	28.7	95.7	21.2	98.4	28.7
22	99.4	28.2	96.7	20.8	99.4	28.2	99.2	27.9	96.4	20.6	99.2	27.9
23	99.8	28.3	97.0	20.9	99.8	28.3	99.2	28.7	96.5	21.2	99.2	28.7
24	98.8	27.4	96.1	20.1	98.8	27.4	98.4	27.3	95.7	20.0	98.4	27.3
25	94.8	26.7	92.4	19.8	94.8	26.7	96.8	26.7	94.3	19.9	96.8	26.7
26	98.4	26.0	95.9	19.1	98.4	26.0	97.6	25.6	95.2	19.0	97.6	25.6
27	99.8	34.6	96.6	26.4	99.8	34.6	100	34.8	96.8	26.2	100	34.8
28	100	35.3	96.8	27.0	100	35.3	100	35.8	96.7	27.1	100	35.8

Çizelge 5.4. (devam) Sözdizimsel özniteliklerle LSTM-NN'nin kaynak metne göre özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
n	D	A	D	A	D	A	D	A	D	A	D	A
29	99.6	35.6	96.4	27.2	99.6	35.6	99.2	35.1	96.0	26.8	99.2	35.1
30	99.6	34.1	96.5	26.0	99.6	34.1	99.2	35.0	95.9	26.7	99.2	35.0
31	100	34.5	96.8	26.3	100	34.5	98.4	34.8	95.2	26.7	98.4	34.8
32	100	35.3	96.8	27.1	100	35.3	100	37.8	96.9	29.0	100	37.8
33	100	34.4	96.8	26.2	100	34.4	100	34.9	96.7	26.5	100	34.9
34	100	36.6	96.7	28.1	100	36.6	99.2	38.1	96.0	29.2	99.2	38.1
35	100	35.7	96.8	27.4	100	35.7	100	35.7	96.7	27.2	100	35.7
36	100	35.9	96.7	27.5	100	35.9	100	38.1	96.7	29.3	100	38.1
37	100	36.7	96.7	28.2	100	36.7	100	36.9	96.6	28.3	100	36.9
38	100	37.0	96.7	28.4	100	37.0	100	40.1	96.7	31.0	100	40.1
39	99.8	35.6	96.5	27.3	99.8	35.6	100	36.6	96.9	28.2	100	36.6
40	100	36.6	96.7	28.1	100	36.6	100	37.1	96.7	28.6	100	37.1
41	100	37.6	96.6	29.0	100	37.6	100	38.9	96.8	30.1	100	38.9
42	100	36.4	96.7	28.0	100	36.4	100	37.9	96.7	29.2	100	37.9

Çizelge 5.4'te yer alan ROUGE değerleri sırasıyla kaynak dokümanlar referans alınarak hesaplanmıştır. Çizelgelerde ilk kolon girdideki öznitelik sayısını işaret etmektedir. Ardından modelin eğitim ve testindeki ortalama ROUGE değerleri sunulmuştur.

Çizelge 5.4'e göre ROUGE değerleri, artan öznitelik sayısıyla artmış; kaynak dokümanlara en yakın özetler 38 sözdizimsel özniteliğin kullanımından elde edilmiş ve 38'in üzerindeki özniteliklerin metin özetlerinin başarısına yüksek etkisi olmamakla beraber ROUGE değerlerini bir miktar düşürdüğü görülmüştür. Bu bulgu, ROUGE hesaplamasında aday cümle kümelerinin kesişimi olan $\text{Ext}_{\cap}$ çıkarmalı özetlerini referans alan ve sonuçlarının Çizelge 5.5'te sunulmuş olan deneyler için de benzerdir.

Çizelge 5.5. Sözdizimsel özniteliklerle LSTM-NN'nin hedef özete göre özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)						
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L		
n	D	A	D	A	D	A	D	A	D	A	D	A	Δ
4	73.1	48.9	66.3	40.8	70.6	47.3	54.8	35.3	43.6	25.6	49.8	31.9	2
5	79.4	51.4	73.8	44.0	77.3	50.2	58.1	39.3	46.9	29.5	53.2	36.1	2
6	81.1	52.0	76.1	44.9	79.5	50.9	55.4	37.3	44.0	27.5	50.4	34.0	2
7	78.9	54.6	73.6	47.2	76.9	53.4	58.8	38.9	48.2	29.5	54.4	35.8	2
8	80.2	55.2	74.3	47.7	77.9	53.9	59.8	37.9	49.0	27.9	55.1	34.8	2
9	82.9	54.4	78.0	47.4	81.2	53.4	59.2	41.0	47.5	30.5	54.3	37.4	1
10	82.0	54.8	77.0	47.8	80.4	53.8	58.2	39.6	46.8	29.7	53.6	36.5	1
11	80.7	55.7	76.0	48.7	79.0	54.5	57.2	38.9	45.3	29.0	51.9	35.7	2
12	85.5	57.6	81.7	51.0	84.2	56.8	57.5	39.9	46.3	29.8	52.8	36.5	2
13	85.9	61.3	81.7	54.6	84.6	60.5	58.0	37.6	46.8	27.6	53.2	34.4	3
14	86.5	57.7	82.9	51.2	85.4	57.0	59.0	39.2	48.5	29.6	54.5	36.0	2
15	87.0	58.9	83.2	52.6	85.8	58.1	60.5	40.1	49.0	30.0	55.5	36.6	2
16	87.4	60.7	83.6	54.6	86.3	60.0	58.9	41.8	46.6	31.8	53.6	38.3	2
17	89.3	61.9	86.0	56.0	88.3	61.3	56.1	40.3	44.5	29.3	51.2	36.5	1
18	92.1	56.0	90.0	50.6	91.7	55.7	57.2	40.5	46.0	30.0	52.4	36.8	2
19	88.7	60.0	85.3	54.2	87.6	59.4	56.3	41.5	44.5	30.9	50.9	37.5	2
20	89.0	57.0	86.0	51.3	88.3	56.5	57.1	39.0	45.1	28.7	51.5	35.2	3
21	87.6	63.9	83.7	57.8	86.5	63.1	55.8	42.0	43.8	31.2	50.4	38.1	3
22	58.6	28.3	48.6	20.9	54.3	26.3	55.8	40.9	43.8	30.6	50.7	37.3	2
23	64.5	51.6	54.6	42.0	60.8	48.5	58.0	42.8	46.3	32.1	52.9	39.0	3
24	88.6	52.1	84.5	46.3	87.2	51.4	57.0	41.7	45.7	31.7	52.2	38.4	2
25	90.4	60.6	87.3	54.9	89.7	60.2	55.6	39.6	44.0	30.0	50.8	36.2	1
26	92.6	62.8	89.8	57.1	92.0	62.4	56.5	38.9	43.7	29.3	51.0	35.4	2
27	93.8	80.0	91.7	76.0	93.6	79.8	56.0	50.9	43.8	39.1	50.9	46.4	1
28	95.3	79.6	93.3	75.9	95.1	79.5	55.4	51.0	42.5	38.4	49.5	46.0	1
29	94.2	79.2	92.1	75.4	94.0	79.0	55.6	50.4	43.1	38.3	50.8	45.9	1
30	93.9	78.7	91.8	74.9	93.8	78.6	55.4	49.9	42.8	37.9	50.3	45.4	2
31	95.5	85.6	94.0	82.5	95.4	85.4	56.4	51.2	44.5	40.0	51.8	47.0	1
32	95.8	80.8	94.0	77.0	95.6	80.7	56.7	55.4	45.4	44.0	52.2	51.2	1
33	94.4	84.9	92.5	81.7	94.3	84.8	56.9	50.9	44.5	38.9	51.9	46.6	0
34	94.2	88.3	92.5	85.5	94.1	88.2	55.6	54.3	44.0	42.1	51.0	49.7	0

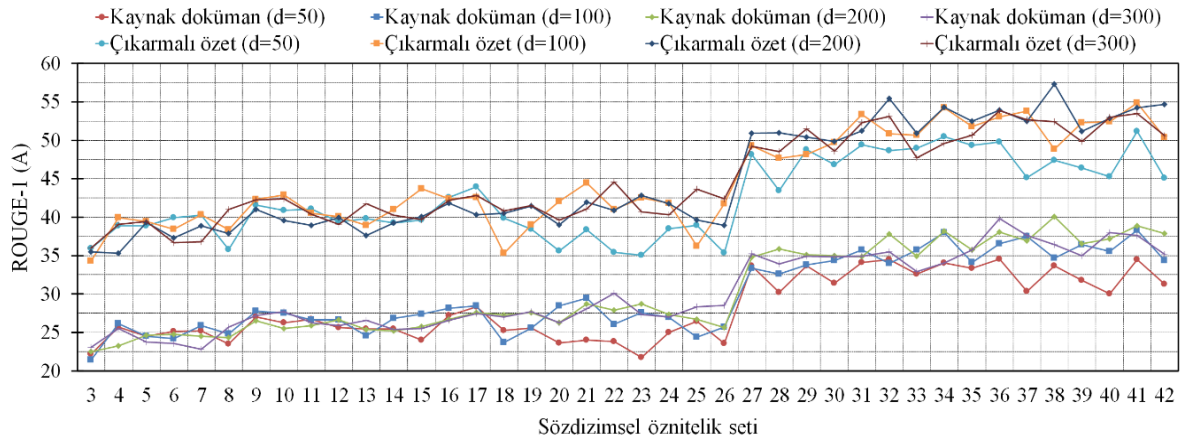
Çizelge 5.5. (devam) Sözdizimsel özniteliklerle LSTM-NN'nin hedef özete göre özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)						
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L		
n	D	A	D	A	D	A	D	A	D	A	D	A	Δ
35	92.7	90.2	90.4	87.8	92.4	89.9	57.1	52.5	44.7	40.2	51.9	47.7	0
36	96.5	90.0	95.0	87.7	96.5	89.9	54.7	54.0	42.5	41.5	49.8	49.5	0
37	93.8	90.0	91.7	87.4	93.5	89.8	54.8	52.5	42.2	40.1	49.6	47.8	3
38	90.8	89.5	88.0	86.5	90.6	89.3	55.2	57.3	43.9	45.1	50.8	52.7	1
39	97.5	90.1	96.4	87.8	97.4	90.0	54.0	51.2	40.9	38.9	48.3	46.2	3
40	95.2	92.4	93.6	90.5	95.1	92.3	55.2	52.8	42.8	40.8	50.3	48.3	1
41	97.5	85.2	96.2	82.0	97.4	85.1	54.5	54.2	42.2	41.8	49.4	49.4	0
42	95.6	90.6	93.9	88.2	95.5	90.5	55.9	54.7	43.6	42.4	51.2	50.2	1

Çizelge 5.5'de yer alan ROUGE değerleri aday cümle kümelerinin kesişimi olan $\text{Ext}_{\cap}$ çıkarmalı özetleri referans alınarak hesaplanmıştır. Çizelgelerde ilk sütun girdideki öznitelik sayısını işaret etmektedir. Ardından modelin eğitim ve testindeki ortalama ROUGE değerleri sunulmuştur. Son sütundaki Δ değeri otomatik özetteki cümle sayısı ile referans alınan $\text{Ext}_{\cap}$ çıkarmalı özetlerinin içerdiği cümle sayısının farkını temsil etmektedir.

Özet değerlendirmesinde aday cümle kümelerinin kesişimi olan $\text{Ext}_{\cap}$ çıkarmalı özetlerinin referans alındığı deneyler arasından en yüksek ROUGE değerleri 38 sözdizimsel öznitelik üzerinden tanımlanan LSTM-NN modelinin çıktısı olan çıkarmalı özetlerden elde edilmiştir. Burada ROUGE-1 %57.3, ROUGE-2 %45.1 ve ROUGE-L %52.7 olarak ölçülmüştür. Bu ROUGE skorları $\Delta = 1$ koşulu ile elde edilmiş olduğundan anlamlıdır.

Çizelge 5.5'te yer alan ROUGE-1 anma değerlerinin grafiği Şekil 5.3'teki grafikte sunulmuştur. ROUGE değerleri kaynak doküman (grafikteki alt gruptaki eğriler) ve aday cümle kümelerinin kesişiminden oluşan $\text{Ext}_{\cap}$ çıkarmalı özetler (grafikteki üst gruptaki eğriler) için ayrı ayrı görülmektedir.



Şekil 5.3. Sözdizimsel öznitelikler üzerinde LSTM-NN ROUGE-1 anma değerleri

Şekil 5.3'te 38 öznitelik üzerinde çalışan LSTM-NN'nin düşük Δ değerine rağmen yüksek ROUGE değerine sahip olduğu açıkça görülmektedir. Sözdizimsel özniteliklerin kullanımı ile elde edilen en yüksek ROUGE-1 anma değeri %57.3 olarak ölçülmüştür.

Çizelge 5.6. Sözdizimsel özniteliklerle LSTM-NN'nin cümle sınıflandırma başarısı

	Eğitim başarısı (%)				Test başarısı (%)			
n	D	A	Doğruluk	ROC-AUC	D	A	Doğruluk	ROC-AUC
4	67.4	32.0	80.1	81.9	41.7	20.4	74.9	67.0
5	72.2	34.3	81.2	83.4	46.3	22.4	75.8	65.8
6	74.5	36.3	81.9	84.2	42.1	21.3	74.9	65.6
7	74.5	35.7	81.8	83.9	44.9	22.2	75.5	65.5
8	74.6	35.7	81.8	84.0	44.3	21.7	75.5	65.3
9	75.8	39.4	82.6	85.2	42.9	23.0	74.8	64.6
10	76.3	38.9	82.6	85.3	44.4	22.8	75.2	64.0
11	76.2	38.7	82.5	85.2	42.6	22.3	74.7	64.5
12	78.5	42.2	83.5	86.5	43.2	23.6	74.9	63.4
13	80.3	41.5	83.6	86.6	43.0	22.1	75.0	63.3
14	79.9	41.3	83.5	86.4	44.5	23.2	75.3	63.7
15	84.5	43.0	84.5	87.9	43.3	23.6	75.0	63.8
16	83.2	44.0	84.6	87.9	43.1	24.6	74.9	62.8
17	81.3	45.8	84.6	88.2	39.4	22.7	73.7	61.7
18	82.3	47.1	85.0	88.8	40.7	23.4	74.0	60.3
19	82.4	46.2	84.8	88.5	39.4	23.7	73.8	62.1
20	83.9	45.4	84.8	88.4	40.5	21.2	73.9	62.4

Çizelge 5.6. (devam) Sözdizimsel özniteliklerle LSTM-NN'nin cümle sınıflandırma başarısı

	Eğitim başarısı (%)				Test başarısı (%)			
n	D	A	Doğruluk	ROC-AUC	D	A	Doğruluk	ROC-AUC
21	81.0	47.0	84.7	88.2	39.8	23.6	73.7	62.3
22	84.0	48.4	85.4	89.4	40.6	24.2	74.1	60.8
23	82.4	48.0	85.1	88.6	40.2	24.5	73.7	63.9
24	77.6	43.5	83.5	86.2	42.5	24.4	74.7	63.4
25	74.5	43.4	83.7	86.4	40.4	22.5	73.6	64.3
26	82.2	42.9	84.1	86.8	41.7	21.9	74.4	63.8
27	89.9	74.1	91.9	97.1	37.5	32.0	71.8	61.8
28	87.2	74.2	91.3	96.8	37.7	33.3	71.7	61.9
29	88.0	74.8	91.5	96.8	37.8	33.2	71.6	61.5
30	91.6	73.5	92.1	97.4	36.9	32.1	71.4	60.9
31	91.2	75.5	92.4	97.4	40.3	34.7	72.5	63.1
32	91.7	78.2	93.1	98.0	40.5	38.1	72.5	64.0
33	92.7	75.9	92.8	97.3	38.9	33.7	72.2	61.0
34	89.6	81.6	93.3	98.2	38.2	36.7	71.3	64.6
35	89.2	77.0	92.3	97.3	39.2	35.1	72.0	64.1
36	90.5	80.1	93.3	98.0	37.8	35.7	71.2	62.1
37	92.1	83.7	94.4	98.7	36.1	34.2	70.6	61.5
38	91.8	84.9	94.6	98.7	37.8	38.5	70.8	63.1
39	90.3	78.9	92.9	97.9	36.2	32.5	70.6	60.0
40	91.1	81.8	93.7	98.3	37.0	34.2	70.9	62.7
41	88.1	83.6	93.4	98.0	35.5	34.6	70.0	63.1
42	88.4	78.8	92.4	97.1	38.1	35.8	71.4	64.0

Çizelge 5.6'da sözdizimsel özniteliklerin kullanımının standart sınıflandırma performans değerlendirme ölçütleri açısından karşılaştırması görülmektedir. Burada da referans alınan cümle kümeleri aday cümle kümelerinin kesişiminden oluşan $\mathrm{Ext}_{\cap}$ çıkarmalı özetlerdir. Yani bu özetlerde yer alan cümleler, özete dâhil edilmeye değer cümleler sınıfında diğer cümleler ise özete dâhil edilmeye değer olmayan cümleler sınıfına eşleştirilmiştir.

Çizelge 5.6'da ROC-AUC tipi bir başarı değerlendirmesinde modeller arasında önemli değişikliklerin olduğu gözlenememektedir. Çünkü tüm modellerden elde edilen sonuçlar çok

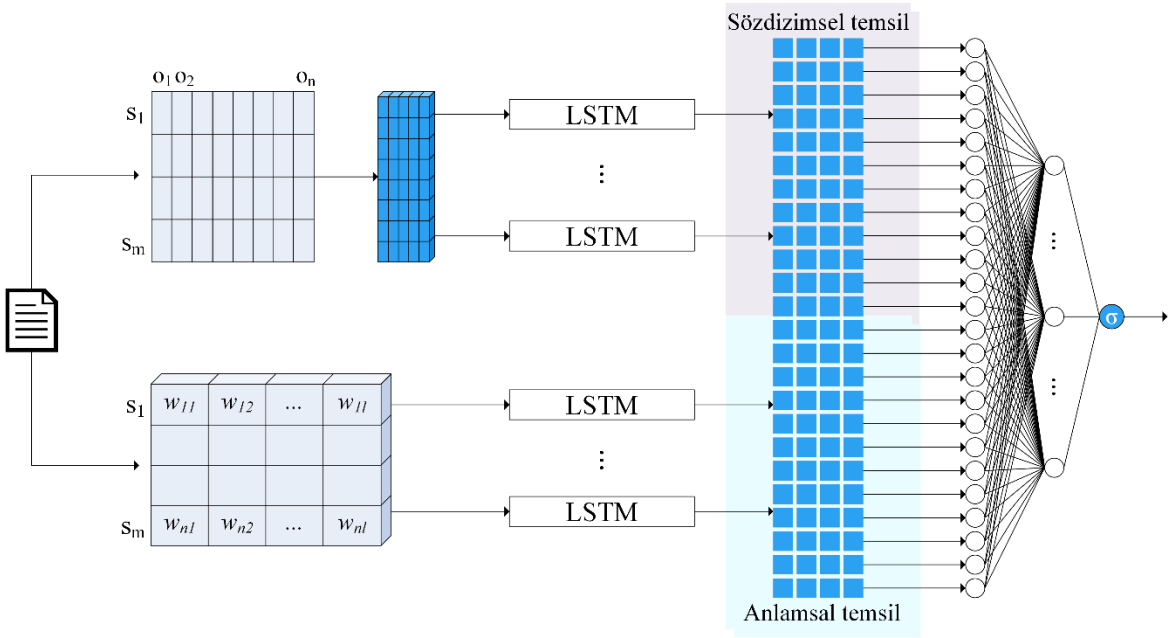
dar bir aralıkta ve birbirine yakın değerlerdir. Öte yandan, 0.5 sınıflandırma eşiğinin sabit kabul edilmesinden elde edilen duyarlılık, anma ve doğruluk değerleri de ölçülmüştür. Burada metin özetlemede anma değerleri en çok dikkate alınan performans değerlendirme ölçütüdür. ROUGE değerlerinden elde edilen bulguya paralel olarak sınıflandırma başarısı açısından yapılmış olan değerlendirmede de en başarılı özetler 38 sözdizimsel öznitelikli deneysel çalışmalardan elde edilmiştir.

5.3. Hibrit Öznitelik Kümesi Üzerinde LSTM-NN

Bu tez çalışmasında gerek SIGIR2018 veri kümesi, gerekse literatürde en sık rastlanan DUC veri kümesi için kelime özyerleşiklerine dayalı anlamsal özniteliklerin ve metnin ve cümlelerin yapısına bağlı sözdizimsel özniteliklerin tek başına kullanımının metin özetlemedeki başarıya etkisi farklı özetleme yöntemleriyle derinlemesine incelenmiş ve bu yaklaşımlarla metin özetlemenin başarı limitleri ölçülmüştür.

SIGIR2018 veri kümesinin ham kelime özyerleşikleri ile metin özetleme başarısı Bölüm 5.1'de ele alınmış ve aday cümle kümelerinin kesişimi olan çıkarmalı metin özetleri referans alındığında bu özyerleşiklerden elde edilen ortalama ROUGE-1 anma değerleri %60.5 olarak hesaplanmıştır. Öte yandan, otomatik metin özetlemede dikkate alınan ve metnin yazım stili ve yapısıyla bağlantılı olan sözdizimsel özniteliklerin kullanımından elde edilen maksimum ROUGE-1 anma değeri ortalama %57.3 olarak hesaplanmış ve Bölüm 5.2'de ayrıntılı bir şekilde analiz edilmiştir.

Bu bölümde, tez çalışmasında LSTM-NN mimarisi üzerinde anlamsal ve sözdizimsel özniteliklerin birlikteliğini içeren hibrit öznitelik kullanımı önerilmiştir. Önerilen hibrit yaklaşım Şekil 5.4'te görülmektedir.



Şekil 5.4. Hibrit öznitelikler üzerinde LSTM-NN mimarisi

Şekil 5.4'teki LSTM-NN yapısına göre anlamsal öznitelik vektörünün LSTM ile modellenmesi sonucu LSTM hücrelerinden elde edilen gizli durumlar ile sözdizimsel özniteliklerin LSTM çıktıları bir birleştirme katmanına giriş olarak verilmiş ve bu iki vektör ilgili oldukları cümlelerle eşleşecek şekilde birleştirilmiştir.

Şekil 5.4'te sözdizimsel ve anlamsal özniteliklerin kullanımı sonucunda LSTM katmanının sonunda oluşan bu vektörler sözdizimsel temsil ve anlamsal temsil olarak isimlendirilmiştir. Birleşme sonucunda oluşan hibrit öznitelik vektörü Bölüm 4.3.1 ve Bölüm 4.3.2'deki LSTM-NN modelleri ile aynı olacak şekilde LSTM-NN'nin yapay sinir ağı katmanlarına sınıflandırma amacıyla verilmiştir. Özetle, sözdizimsel öznitelikler üzerinden geliştirilen LSTM-NN ile anlamsal öznitelikler üzerinden geliştirilen LSTM-NN, iki tip özniteliği eş zamanlı olarak birer LSTM ile işleyecek şekilde geliştirilmiş ve sözdizimsel ve anlamsal LSTM çıktılarının birleşimi olan vektör yapay sinir ağında sınıflandırma amacıyla kullanılmıştır.

Gerçekleştirimde en iyileştirilecek ölçütler olarak doğruluk ve ROC-AUC seçilmiştir. En iyileştirme yöntemi Adam [120] ve kayıp fonksiyonu ikili çapraz entropi olarak belirlenmiştir. Minimum hata farkının üst üste 10 kez 1.00E-07 değerinin altına düşmesi halinde eğitim erken durdurulmaktadır.

Şekil 4.5(c)'de sunulmuş olan hibrit öznitelikler üzerinde gerçekleştirilen LSTM-NN tabanlı metin özetlemede, kelime özyerleşiklerinin kullanıldığı anlamsal özyerleşiklerin işlendiği katmanlarda yine önceden eğitilmiş GloVe (d = 50, d = 100, d = 200 ve d = 300) ve word2vec (d = 100) kelime vektörleri kullanılmıştır. Buna karşılık gelen sözdizimsel özniteliklerin işlenmesinde ilk olarak bir özyerleşik katmanında ölçeklendirme yapılmaktadır ve sözdizimsel özyerleşikler anlamsal özyerleşiklerin boyuna eşitlenmektedir. Bu koşullarda, birleştirme katmanının sonucunda elde edilen vektörler $d+d$ boyunda olmaktadır.

Bölüm 4.3.2'de sözdizimsel öznitelikler üzerinden yapılan LSTM-NN tabanlı deneyler çalışmalarda, yararlanılan öznitelikler en yüksek χ^2 değerine sahip 4 öznitelikten başlanmak üzere 42 özniteliğe kadar artırımlı olarak değiştirilmiştir ve her bir öznitelik kümesinin kullanımı için ayrı ayrı beş-katlı çapraz geçerleme yapılmıştır. Burada 38 öznitelik ve üstündeki sözdizimsel özniteliklerden yüksek özetleme başarısı elde edildiği görülmüştür. Bu nedenle, bu bölümde performansı incelenecek olan hibrit özniteliklerin sözdizimsel kısmında [38, 42] aralığında sözdizimsel öznitelik içeren kümeler kullanılmıştır. Öte yandan anlamsal öznitelikler için gerçekleştirilen ve Bölüm 4.3.1'de sunulan deneysel çalışmaların tümü bu bölümde de yinelenmiştir.

Sözdizimsel ve anlamsal özniteliklerin birlikte kullanımının metin özetleme başarısına etkisini ölçmek adına, SIGIR2018 veri kümesindeki dokümanlar üzerinde beş-katlı çapraz geçerleme yapılmıştır. Çapraz geçerlemenin eğitim-test kümeleri yine rastgele seçilen 100 doküman eğitimde 25 doküman testte kullanılmak üzere ayrılmıştır.

Çizelge 5.7 ve Çizelge 5.8'de hibrit öznitelik kümesinin kullanımı ile LSTM-NN'nin beş-katlı çapraz geçerlemeye göre ROUGE ölçütleri açısından metin özetleme başarısı sunulmuştur.

Çizelge 5.7. Hibrit LSTM-NN'nin kaynak metne göre özetleme başarısı

		Eğitim başarısı (%)						Test başarısı (%)					
		1		2		L		1		2		L	
n		D	A	D	A	D	A	D	A	D	A	D	A
38	G_{50}	100	38.5	96.6	29.9	100	38.5	99.2	43.0	96.2	34.9	99.2	43.0
	G_{100}	100	38.5	96.6	29.9	100	38.5	100	44.7	96.9	36.3	100	44.7
	G_{200}	100	38.4	96.6	29.8	100	38.4	100	43.9	97.0	35.5	100	43.9
	G_{300}	100	38.4	96.6	29.8	100	38.4	100	41.5	96.9	33.5	100	41.5
38	W_{50}	100	38.9	96.6	30.2	100	38.9	100	38.2	96.8	30.3	100	38.2
	W_{100}	100	38.5	96.6	29.8	100	38.5	99.2	40.5	96.1	32.5	99.2	40.5
	W_{200}	100	38.4	96.6	29.8	100	38.4	99.2	39.9	96.2	32.0	99.2	39.9
	W_{300}	100	38.5	96.6	29.8	100	38.5	100	37.5	97.0	29.8	100	37.5
39	G_{50}	100	38.4	96.6	29.7	100	38.4	98.4	46.0	95.4	37.6	98.4	46.0
	G_{100}	100	38.4	96.6	29.8	100	38.4	100	43.4	96.7	35.1	100	43.4
	G_{200}	100	38.4	96.6	29.8	100	38.4	97.6	43.7	94.5	35.2	97.6	43.7
	G_{300}	100	38.4	96.6	29.7	100	38.4	100	44.9	97.0	36.8	100	44.9
39	W_{50}	100	38.5	96.6	29.9	100	38.5	99.2	41.1	96.3	33.1	99.2	41.1
	W_{100}	100	38.2	96.6	29.6	100	38.2	100	35.9	97.0	28.4	100	35.9
	W_{200}	100	38.1	96.6	29.4	100	38.1	100	39.5	96.9	31.5	100	39.5
	W_{300}	100	38.4	96.6	29.8	100	38.4	100	43.4	96.8	35.2	100	43.4
40	G_{50}	100	38.5	96.6	29.9	100	38.5	100	42.7	96.7	34.4	100	42.7
	G_{100}	100	38.4	96.6	29.7	100	38.4	100	44.9	97.0	36.6	100	44.9
	G_{200}	100	38.5	96.6	29.9	100	38.5	100	41.7	96.8	33.5	100	41.7
	G_{300}	100	38.4	96.6	29.8	100	38.4	99.2	42.4	96.1	34.2	99.2	42.4
40	W_{50}	100	38.6	96.6	30.0	100	38.6	99.2	36.3	96.5	28.7	99.2	36.3
	W_{100}	100	38.4	96.6	29.8	100	38.4	98.4	38.0	95.3	30.2	98.4	38.0
	W_{200}	100	38.4	96.6	29.8	100	38.4	99.2	40.2	96.1	32.3	99.2	40.2
	W_{300}	99.4	35.5	96.3	27.4	99.4	35.5	100	36.8	96.9	29.0	100	36.8
41	G_{50}	100	38.4	96.6	29.8	100	38.4	100	41.6	96.9	33.3	100	41.6
	G_{100}	100	38.2	96.6	29.6	100	38.2	100	38.3	96.8	30.2	100	38.3
	G_{200}	100	38.4	96.6	29.8	100	38.4	100	42.6	97.1	34.3	100	42.6
	G_{300}	100	38.4	96.6	29.8	100	38.4	98.4	45.1	95.5	37.0	98.4	45.1

Çizelge 5.7. (devam) Hibrit LSTM-NN'nin kaynak metne göre özetleme başarısı

		Eğitim başarısı (%)						Test başarısı (%)					
		1		2		L		1		2		L	
n		D	A	D	A	D	A	D	A	D	A	D	A
41	W_{50}	100	38.7	96.6	30.1	100	38.7	100	43.5	97.0	35.5	100	43.5
	W_{100}	100	38.5	96.6	29.8	100	38.5	100	38.9	97.0	31.0	100	38.9
	W_{200}	100	38.4	96.6	29.8	100	38.4	99.2	41.3	96.1	33.2	99.2	41.3
	W_{300}	100	38.5	96.6	29.8	100	38.5	100	45.2	97.0	37.1	100	45.2
42	G_{50}	100	38.4	96.6	29.8	100	38.4	100	43.6	96.9	35.1	100	43.6
	G_{100}	100	38.4	96.6	29.8	100	38.4	100	45.1	96.9	37.0	100	45.1
	G_{200}	100	38.4	96.6	29.8	100	38.4	100	40.6	97.0	32.7	100	40.6
	G_{300}	100	38.4	96.6	29.8	100	38.4	99.2	41.8	96.1	33.8	99.2	41.8
42	W_{50}	100	38.6	96.6	29.9	100	38.6	100	41.0	96.9	32.7	100	41.0
	W_{100}	100	38.4	96.6	29.8	100	38.4	100	37.0	96.8	28.9	100	37.0
	W_{200}	100	38.5	96.6	29.8	100	38.5	100	45.4	96.8	37.3	100	45.4
	W_{300}	99.8	38.3	96.4	29.7	99.8	38.3	100	46.0	97.0	37.4	100	46.0

Çizelge 5.7'de bulunan ROUGE değerleri kaynak dokümanlar referans alınarak hesaplanmıştır. Çizelgelerde ilk sütun girdideki sözdizimsel öznitelik sayısını, ikinci sütun ise ilgili cümlelerde yer alan kelimelerin kullandığı kelime özyerleşik türlerini ve boyunu işaret etmektedir. Burada G, GloVe özyerleşiklerine, W ise word2vec özyerleşiklerine karşılık gelmektedir. Çizelgelerin diğer sütunlarında modelin eğitim ve testindeki ortalama ROUGE değerleri sunulmuştur.

Çizelge 5.7'de ROUGE hesaplaması kaynak dokümanların referans alınmasıyla yapılmıştır. Bu sayede ortaya çıkan özet metinlerin, kaynak dokümanı temsil etmedeki yeterliliği ve özetleme süreci sonunda kaynak dokümandaki bilginin ne oranda korunduğu ölçülebilir ve yorumlanabilir hale gelmiştir. Burada, duyarlılık değerlerinin %100 seviyesinde olması, ROUGE hesaplamalarında referans alınan metnin dokümanın kendisi olmasından kaynaklanmaktadır. Anma değerleri, özette bulunması gereken söz öbeklerinin ne oranda otomatik özetlere dâhil edildiğini ölçtüğünden anlamlıdır.

Çizelge 5.7'de ROUGE-1, ROUGE-2 ve ROUGE-L değerlerine göre elde edilen anma değerlerinin sözdizimsel ve anlamsal özniteliklerin tek başına kullanımından oldukça

yüksek olduğu görülmektedir. Burada, beş-katlı çapraz geçerlemenin test veri kümesi üzerinden elde edilen ortalama ROUGE değerleri ROUGE-1 için %46.0, ROUGE-2 için %37.6 ve ROUGE-L için %46.0 olarak ölçülmüştür. Elde edilen bu değerler, anlamsal özniteliklerin tekil kullanımına kıyasla yaklaşık olarak %5 oranında; sözdizimsel özniteliklerin birlikte kullanımına göre yaklaşık %6 oranında artışa karşılık gelmektedir. Çizelge 5.8'de hibrit öznitelik uzayı üzerinde yapılan metin özetlemenin aday cümle kümelerinin kesişimi olan $Ext_{\cap}$ çıkarmalı özetlere göre ROUGE sonuçları sunulmuştur.

Çizelge 5.8. Hibrit LSTM-NN'nin hedef özete göre özetleme başarısı

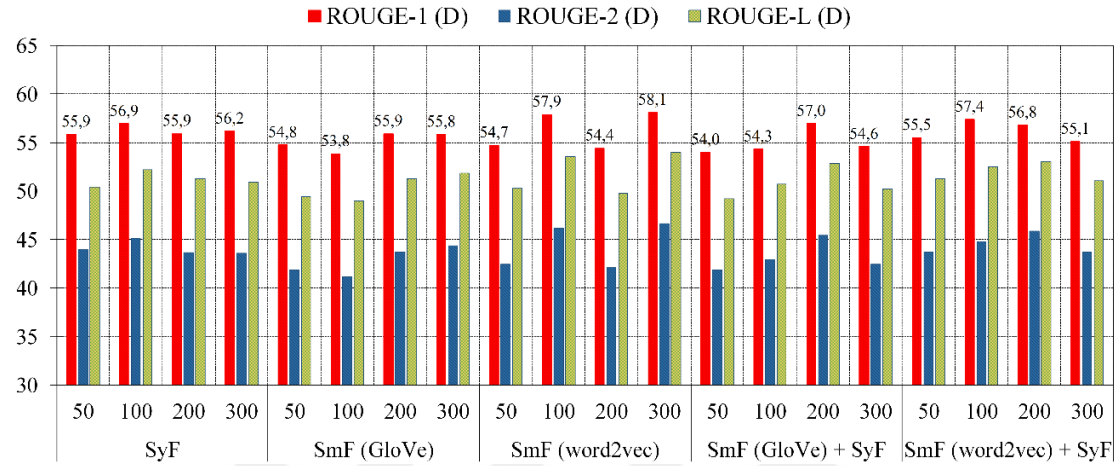
		Eğitim başarısı (%)						Test başarısı (%)						Δ
		1		2		L		1		2		L		
n		D	A	D	A	D	A	D	A	D	A	D	A	
38	G_{50}	99.5	99.5	99.3	99.4	99.5	99.5	53.0	58.7	40.9	47.7	48.7	54.7	4
	G_{100}	99.7	99.8	99.6	99.8	99.7	99.8	52.5	60.4	40.6	48.9	48.4	55.9	2
	G_{200}	100	99.9	100	99.8	100	99.9	55.9	63.0	44.4	52.2	52.1	59.0	4
	G_{300}	100	100	100	99.9	100	100	56.5	60.3	44.5	49.6	52.4	56.3	4
38	W_{50}	98.8	99.8	98.5	99.7	98.8	99.8	55.7	54.1	43.5	43.0	50.9	49.8	1
	W_{100}	99.8	99.8	99.8	99.8	99.8	99.8	59.0	60.1	48.1	49.8	54.7	56.4	4
	W_{200}	100	99.8	100	99.8	100	99.8	59.4	60.7	48.3	50.3	55.6	57.1	2
	W_{300}	99.7	99.8	99.7	99.8	99.7	99.8	58.3	55.6	46.8	45.3	54.2	51.9	3
39	G_{50}	99.9	99.5	99.8	99.4	99.9	99.5	51.9	61.1	40.4	49.7	47.5	56.4	3
	G_{100}	99.9	99.9	99.9	99.8	99.9	99.9	53.7	58.9	41.8	47.2	49.5	54.5	2
	G_{200}	100	99.9	100	99.9	100	99.9	53.4	60.4	42.0	49.3	49.1	55.8	4
	G_{300}	100	99.8	100	99.8	100	99.8	54.1	61.4	42.5	50.6	50.2	57.5	4
39	W_{50}	99.4	99.4	99.2	99.2	99.4	99.4	56.2	58.0	44.5	47.5	51.9	54.1	4
	W_{100}	99.9	99.3	99.8	99.2	99.9	99.3	57.6	52.3	45.3	41.7	52.8	48.3	2
	W_{200}	99.3	98.4	99.1	98.0	99.3	98.4	58.2	57.0	46.8	45.9	54.0	52.9	4
	W_{300}	100	100	100	99.9	100	100	57.4	63.8	45.9	53.3	53.6	60.1	4
40	G_{50}	99.5	99.6	99.4	99.5	99.5	99.6	55.3	60.6	43.0	49.1	50.8	56.3	5
	G_{100}	99.9	99.7	99.9	99.6	99.9	99.7	55.5	62.7	44.1	51.4	51.4	58.3	4
	G_{200}	99.8	99.9	99.7	99.9	99.8	99.9	55.4	59.2	42.9	47.7	50.8	54.6	3
	G_{300}	99.9	99.8	99.9	99.7	99.9	99.8	56.8	61.7	45.6	51.2	52.8	57.8	3

Çizelge 5.8. (devam) Hibrit LSTM-NN'nin hedef özete göre özetleme başarısı

		Eğitim başarısı (%)						Test başarısı (%)						
		1		2		L		1		2		L		Δ
n		D	A	D	A	D	A	D	A	D	A	D	A	
40	W_{50}	99.2	99.5	99.1	99.4	99.2	99.5	57.0	52.8	44.9	41.7	52.1	48.3	1
	W_{100}	99.9	99.8	99.9	99.8	99.9	99.8	55.8	55.4	44.1	45.0	51.7	51.7	3
	W_{200}	99.9	99.7	99.9	99.7	99.9	99.7	56.5	59.0	44.0	48.2	52.1	55.1	4
	W_{300}	94.9	88.8	93.5	87.3	94.4	88.4	61.5	56.9	51.0	47.1	57.8	53.4	2
41	G_{50}	99.6	99.5	99.5	99.4	99.6	99.5	57.0	59.0	45.2	47.8	52.8	55.1	4
	G_{100}	99.9	99.1	99.8	98.9	99.9	99.1	55.1	53.9	42.6	42.1	50.4	49.4	0
	G_{200}	100	99.9	100	99.8	100	99.9	54.8	59.2	43.0	47.8	50.3	54.8	4
	G_{300}	100	99.9	100	99.9	100	99.9	53.9	63.1	42.9	52.6	50.3	59.3	5
41	W_{50}	99.1	99.7	98.9	99.6	99.1	99.7	55.1	61.2	43.5	50.7	50.8	57.2	3
	W_{100}	99.9	99.8	99.9	99.8	99.9	99.8	56.6	57.3	43.9	46.5	52.1	53.3	2
	W_{200}	100	99.9	100	99.9	100	99.9	56.0	60.4	44.7	49.9	52.2	56.6	3
	W_{300}	99.9	99.9	99.8	99.9	99.9	99.9	56.5	64.2	45.6	54.2	52.8	60.5	4
42	G_{50}	99.9	99.8	99.9	99.8	99.9	99.8	54.0	59.1	41.9	47.0	49.2	54.2	4
	G_{100}	99.9	99.8	99.9	99.8	99.9	99.8	54.3	61.8	43.0	50.7	50.7	57.9	2
	G_{200}	100	100	100	100	100	100	57.0	58.2	45.5	47.5	52.9	54.3	4
	G_{300}	100	99.9	100	99.9	100	99.9	54.6	58.9	42.5	47.5	50.2	54.5	4
42	W_{50}	99.4	99.7	99.3	99.6	99.4	99.7	55.5	58.1	43.7	46.6	51.3	54.0	3
	W_{100}	100	99.9	100	99.9	100	99.9	57.4	53.9	44.8	42.1	52.5	49.4	1
	W_{200}	99.9	99.9	99.8	99.8	99.9	99.9	56.8	65.2	45.8	55.1	53.1	61.4	5
	W_{300}	98.5	98.3	98.1	97.9	98.5	98.3	55.1	63.1	43.7	51.4	51.1	58.7	5

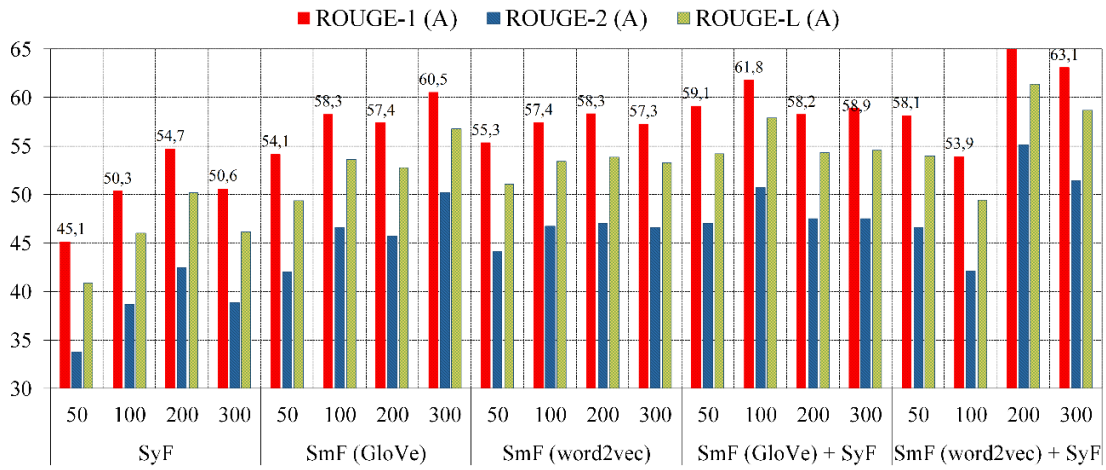
Çizelge 5.8'deki Δ değeri referans alınan $\text{Ext}_{\cap}$ çıkarmalı özetlerinin içerdiği cümlelerin sayısı ile LSTM-NN'den elde edilen otomatik özetlerde yer alan cümlelerin sayısının farkını ifade eder. Anlamsal özniteliklerin tek başına kullanımı ile elde edilen en yüksek ROUGE-1 anma değerinin %60.5, sözdizimsel özniteliklerin tek başına kullanımı ile elde edilen en yüksek ROUGE-1 anma değerinin %57.3 olmuştur. Bu iki özniteliğin birlikte kullanımından elde edilen hibrit öznitelik uzayı üzerindeki deneylerde ise beş-katlı çapraz geçerleme test başarısının %65.2'ye ulaştığı Çizelge 5.8'de görülmektedir. Çizelgeden, ROUGE-2 ve ROUGE-L için de benzer tespitler yapmak mümkündür. Buna göre hibrit öznitelik kümesinin kullanımının, özetleme modelinde radikal bir değişikliğe gidilmediği durumda

bile anlamsal ve sözdizimsel özniteliklerin tekli kullanımına göre ROUGE değerlerini yüksek oranda etkilediği görülmektedir. Şekil 5.5, Şekil 5.6 ve Şekil 5.7'de sözdizimsel ve anlamsal özniteliklerin hibrit kullanımının bu özniteliklerin tekil kullanımına etkisi sırasıyla ROUGE-1 duyarlılık, anma ve F-ölçütü değerleri açısından grafiksel olarak sunulmuştur.



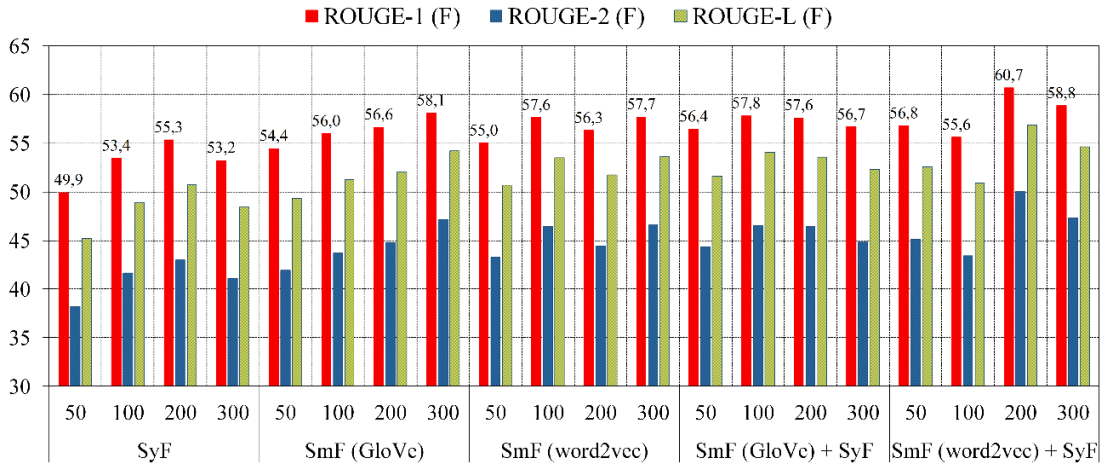
Şekil 5.5. LSTM-NN özetlerin SIGIR2018 verisinde ROUGE-1 duyarlılık değerleri

Şekil 5.5'te tüm sözdizimsel özniteliklerin kullanıldığı durumda, beş-katlı çapraz geçerlemenin ortalama test başarısı duyarlılık değerleri açısından sunulmaktadır ve ROUGE değerleri aday kümelerinin kesişimi olan $\text{Ext}_{\cap}$ çıkarmalı özetleri temel alınarak hesaplanmıştır. SyF, sözdizimsel öznitelikleri, SmF ise GloVe veya word2vec kelime özyerleşiklerine bağlı anlamsal öznitelikleri göstermektedir. Burada özyerleşik boyu yine 50, 100, 200 ve 300 olmak üzere değişkendir. Sunulan ROUGE duyarlılık değerlerine göre özniteliklerin birlikte kullanımı modelin ürettiği özetlere önemli oranda etki etmemektedir. Yani özete dâhil edilmeye değer olarak seçilen cümlelerin hedef özetlerde yer alma oranı büyük oranda değişiklik göstermemektedir. Duyarlılık ölçütü dikkate alındığında word2vec özyerleşikleri üzerinden yapılan deneylerde GloVe özyerleşiklerine göre daha yüksek değerler elde edildiği gözlenebilmektedir. Ancak tek başına duyarlılık ölçütü özet performansının değerlendirilmesinde yeterli değildir. Literatürdeki çalışmalar göz önüne alındığında metin özetlerinin anma değerleri açısından değerlendirildiği görülmektedir.



Şekil 5.6. LSTM-NN özetlerin SIGIR2018 verisinde ROUGE-1 anma değerleri

Şekil 5.6'da beş-katlı çapraz geçerlemenin ortalama test başarısı anma değerleri açısından sunulmaktadır ve ROUGE değerleri aday kümelerinin kesişimi olan $\text{Ext}_{\cap}$ çıkarmalı özetleri temel alınarak hesaplanmıştır. SyF, sözdizimsel öznitelikleri, SmF ise GloVe veya word2vec kelime özyerleşiklerine bağlı anlamsal öznitelikleri göstermektedir. Burada özyerleşik boyu yine $d=50$, $d=100$, $d=200$ ve $d=300$ olmak üzere değişkendir. Deneysel çalışmalarda yine sözdizimsel özniteliklerin tümü kullanılmıştır. Burada sözdizimsel veya anlamsal özniteliklerin birlikte kullanımının, bu özniteliklerin tekil kullanımına oranla özet başarısına belirgin katkı sağladığı açıkça görülmektedir.

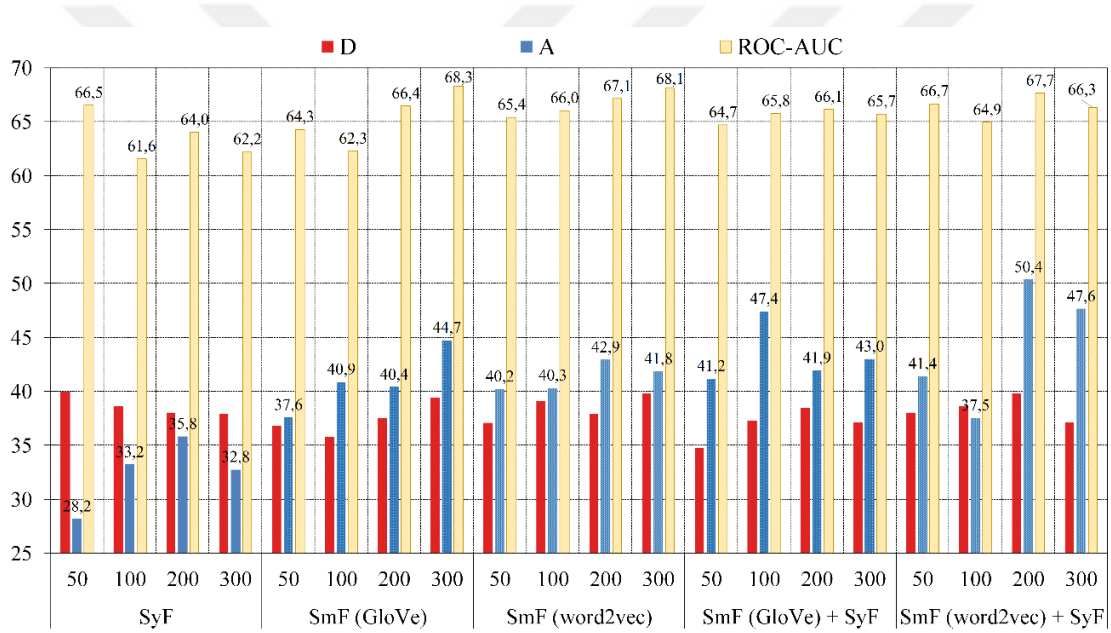


Şekil 5.7. LSTM-NN özetlerin SIGIR2018 verisinde ROUGE-1 F değerleri

Şekil 5.7'de beş-katlı çapraz geçerlemenin ortalama test başarısı F-ölçütü üzerinden sunulmaktadır ve ROUGE değerleri aday kümelerinin kesişimi olan $\text{Ext}_{\cap}$ çıkarmalı özetleri temel alınarak hesaplanmıştır. SyF, sözdizimsel öznitelikleri, SmF ise GloVe veya word2vec

kelime özyerleşiklerine bağlı anlamsal öznitelikleri göstermektedir. Burada özyerleşik boyu yine $d=50$, $d=100$, $d=200$ ve $d=300$ olmak üzere değişkendir. Deneysel çalışmalarda yine sözdizimsel özniteliklerin tümü kullanılmıştır. Burada değerlendirmede esas alınan nokta duyarlılık değerinde düşmeye neden olmadan anma değerini ve bununla birlikte F değerini yüksek tutabilmek olduğundan özniteliklerin birlikte kullanımının ortaya çıkan otomatik metin özetlerini yüksek oranda geliştirmiş olduğu söylenebilir.

Şekil 4.10'da hibrit özniteliklerin kullanımının cümle sınıflandırma başarısına olan etkisi duyarlılık ve anma açısından ve sınıflandırma eşik değerinden etkilenmeyen ROC-AUC ölçütleri açısından değerlendirilmiştir.



Şekil 5.8. LSTM-NN özetlerin SIGIR2018 verisinde cümle sınıflandırma

Şekil 5.8'de, ROC-AUC değerleri arasında büyük farklılıklar gözlenmemektedir. Ancak 0.5 değerinin sınıflandırma eşik değeri olarak kabul edildiği ölçümlerde dikkate alınan esas ölçüt olan anma değerinin hibrit öznitelik kullanımında yükselmiş olduğu açıkça görülmektedir.

Otomatik metin özetleme problemini cümlelerdeki kelimelerin anlamlarıyla ve birbiriyle olan ilişkisiyle tanımlamanın yanında, metindeki yapısal özelliklerin ve cümlelerdeki sözdizimsel özniteliklerin de dikkate alınması gerekliliğini öne süren bu tez çalışmasından elde edilen deneysel sonuçlar, literatürde sıklıkla çalışılmış olan mevcut metin özetleme yöntemleriyle de kıyaslanmıştır. Buradaki kıyaslamalarda ROUGE değerleri dikkate

alınmıştır ve ROUGE hesaplamalarında kullanılmış olan referans özetler, aday cümle kümelerinin kesişimi olan $Ext_{\cap}$ çıkarmalı özetlerdir. Yapılan çok sayıdaki farklı parametrelere dayalı deneyler arasından tüm metin özetleme yöntemlerinin en yüksek başarı elde ettiği deneysel çalışmalar dikkate alınmıştır. Kıyaslamalar ilgili metin özetleme yaklaşımının en yüksek sınırları dikkate alınarak yapılmıştır. Bu yöntemler ve bu yöntemlerle geliştirilmiş otomatik özetlerden elde edilen ROUGE değerleri Çizelge 5.9'da sunulmuştur.

Çizelge 5.9. SIGIR2018 veri kümesi üzerinde LSTM-NN yöntem karşılaştırması

	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
Model	D	A	D	A	D	A	D	A	D	A	D	A
Bulanık sistem	50.8	67.4	40.2	55.0	47.3	62.7	50.3	66.7	39.3	54.2	46.6	61.9
Sinir ağı	93.2	93.0	91.2	90.9	93.0	92.9	51.5	53.6	39.1	41.7	46.9	48.9
SummaRuNNer	83.6	92.9	80.6	90.6	83.6	92.9	53.4	62.6	42.5	51.0	49.5	58.3
LSTM-NN	99.9	99.9	99.8	99.8	99.9	99.9	56.8	65.2	45.8	55.1	53.1	61.4

Çizelge 5.9'da, LSTM-NN'nin yapay sinir ağı ve SummaRuNNer'dan daha başarılı otomatik özetler üretebildiği görülmektedir. Bulanık sistem çözümü anma değerleri açısından LSTM-NN'den daha başarılı olmuştur. Ancak bu modelin duyarlık değerleri dikkate alınırsa özete dâhil edilmemesi gereken çok sayıda cümlenin özete dâhil edildiği görülmektedir. Buna rağmen kural tabanlı sistemlerin, kuralların doğru oluşturulabildiği durumlarda oldukça başarılı olduğu görülmektedir.

Yapılan tüm deneysel çalışmalar dikkate alındığında bulanık sistem için minimum ve maksimum ROUGE-1 anma değerleri sırasıyla %47.8 ve %66.7 olmuştur. Bu değerler sinir ağı açısından %13.2 ve %54.3 iken SummaRuNNer için %58.3 ve %62.6'dır. Önerilen yaklaşımın ise minimum ROUGE-1 anma değeri %52.3 ve maksimum anma değeri %65.2 olarak ölçülmüştür. Önerilen LSTM-NN'nin çeşitli şekillerde gerçekleştiriminden elde edilen en başarısız özetler üzerinden hesaplanan ROUGE değerleri bile diğer yöntemlerle elde edilen başarı düzeyine erişmektedir. LSTM-NN, birçok durumda diğer yöntemlerin önüne geçen performans değerlerine sahiptir. LSTM-NN'nin bu yöntemlerin önünde performans gösterebilmesini sağlayan en önemli özelliği, özetlemede kullanılabilecek ve

ortaya genel özetler çıkmasını sağlayacak öğrenme kanallarının geniş bir kısmını özetleme sürecine dâhil ediyor olmasıdır.

6. DUC VERİSİ ÜZERİNDE OTOMATİK METİN ÖZETLEME

Bu bölümde, SIGIR2018 veri kümesi geliştirilmeden önce, tez kapsamında literatürde sıklıkla değerlendirilmeye alınmış olan DUC veri kümesi üzerinde yapılmış olan çıkarmalı metin özetleme çalışmalarına yer verilmiştir. Bu çalışmalar tez çalışmasının öncül çalışmaları olarak değerlendirilmektedir. Bu bağlamda öncelikle çeşitli sözdizimsel özniteliklerin kullanımı ile bir bulanık sistem gerçekleştirimi yapılmıştır. Ardından problem çözümünde sözdizimsel özniteliklerle anlamsal özniteliklerin kullanımının etkisinin karşılaştırılması için ESA tabanlı bir cümle sınıflandırma yaklaşımının gerçekleştirimi yapılmıştır.

Bölüm 2.4.1'de DUC veri kümesi ile ilgili genel özelliklere yer verilmiştir. Bu tez çalışmasında DUC veri kümeleri içerisinden DUC2002 veri kümesi kullanılmıştır. DUC2002'de 59 haber başlığı ve her bir başlık altında yaklaşık 10 doküman ve toplamda yaklaşık 275 cümle bulunmaktadır. Veri kümesinde toplam 567 doküman bulunmakta, dokümanların her biri ortalama 27.8 cümle içermektedir. Her bir haber başlığı altında yaklaşık 200 kelime ve 400 kelime içeren iki çıkarmalı özet bulunmaktadır. Bu çıkarmalı özetler, ilgili başlık altındaki dokümanların içlerindeki önemli cümlelerin okuyucular tarafından elle seçilmesiyle elde edilmiştir. Bu tez çalışmasında, yaklaşık 400 kelime içeren çıkarmalı özetler, hedef özet niteliğinde değerlendirilmiştir. Bu özetlerde ortalama 22 cümle bulunmaktadır.

6.1. Metin Özetlemede Sıkıştırma Oranı Üzerine Çalışmalar

DUC2002 veri kümesi üzerinde yapılan ilk çalışmada metin içerisindeki önemli cümlelerin tahmin edilmesinde kullanılması gereken sözdizimsel öznitelikler araştırılmıştır. Bunun için, literatürde kullanılagelmiş özniteliklerin cümlelerin öznitelik derecelerini hesaplamada kullandıkları yöntemler derlenmiş ve Çizelge 6.1'de yer alan öznitelikler elde edilmiştir. Çizelge 6.1'de kullanılmış olan 28 sözdizimsel öznitelik ve dokümanlardaki cümleler için bu özniteliklerin hesaplanmasında kullanılan eşitliklere yer verilmiştir. DUC2002 veri kümesindeki 400 kelimelik çıkarmalı özetler özete değer cümleleri içermektedir ve bu sayede tüm cümlelerin özete değer olup olmadığını adresleyen etiketleri elde edilebilmiştir. Her bir cümle için Çizelde 6.1'deki gibi hesaplanan öznitelik değerlerinin, cümle etiketleri ile birleştirilmesiyle DUC2002 veri kümesinn sayısal temsili iki boyutlu bir matris olarak

oluşturulmuştur. Ardından bu matris, Chi kare (χ^2) testine tabi tutulmuş ve özniteliklere ait önem değerleri elde edilmiştir. Bu değerler Çizelge 6.1'de sunulmuştur.

Çizelge 6.1. DUC2002 veri kümesinden elde edilen sözdizimsel öznitelikler

Sıra numarası	Öznitelik	Eşitlik	χ^2 Değeri
1	Pozisyon	Eş. 2.7	349.72
2	Başlık	Eş. 2.4	71.76
3	Pozisyon (metne göre)	Eş. 2.9	59.26
4	Başlık	Eş. 2.5	23.69
5	Cümle-cümle ilişkisi (1-gram) (cosinus)	Eş. 2.16	23.68
6	Cümle-cümle ilişkisi (1-gram) (jaccard)	Eş. 2.16	21.43
7	Uzunluk	Eş. 2.11	18.09
8	TF-ISF (4-gram)	Eş. 2.3	16.20
9	Cümle-cümle ilişkisi (2-gram) (cos)	Eş. 2.16	13.26
10	TF-ISF (3-gram)	Eş. 2.3	12.84
11	Cümle-cümle ilişkisi	Eş. 2.17	10.70
12	TF-ISF (2-gram)	Eş. 2.3	9.90
13	TF	Eş. 2.1	9.45
14	Uzunluk	Eş. 2.10	8.86
15	TF-ISF (1-gram)	Eş. 2.3	7.70
16	Dokümandaki cümle sayısı	-	5.87
17	Şebeke bağlantı sıklığı	-	4.76
18	Cümle-cümle ilişkisi (4-gram)	Eş. 2.16	4.41
19	Cümle-cümle ilişkisi (3-gram)	Eş. 2.16	4.21
20	Toplam benzerlik	-	4.05
21	TF	Eş. 2.2	3.74
22	Pozisyon (metne göre)	Eş. 2.8	3.63
23	Tamlamalar (PP)	Eş. 2.15	3.15
24	Tamlamalar (NP)	Eş. 2.15	1.37
25	Tamlamalar (AP)	Eş. 2.15	1.09
26	Tamlamalar (VP)	Eş. 2.15	0.54
27	Uzunluk	Eş. 2.12	0.46
28	Özel isimler	Eş. 2.15	0.43

Bölüm 4.2.1'de SIGIR2018 veri kümesi için de yapıldığı şekilde, DUC2002 veri kümesindeki sözdizimsel özniteliklerin çıkarmalı metin özetleme problemine etkisinin ölçülebilmesi adına, her bir cümlenin öznitelik değerlerini ve özete değer olup olmadığı bilgisini içeren matris, öznitelikler içerisinden en yüksek χ^2 değerine sahip olan beş tanesi ile dokuz tanesi arasında artımlı olarak değiştirilmiştir. Bu sayede aynı veriyi ifade eden farklı öznitelik sayısına sahip beş veri kümesi elde edilmiştir. Burada Çizelge 6.1'deki önem değerlerine göre yapılan sıralamaya dikkat edilmiştir. Örneğin beş özniteliğin olduğu veri kümesinde pozisyon (Eş. 2.7), başlık (Eş. 2.4), pozisyon (Eş. 2.9), başlık (Eş. 2.5) ve kosinüs benzerliğine dayanan cümle-cümle ilişkisi (1-gram) (Eş. 2.16) sözdizimsel öznitelikleri yer almaktadır.

Elde edilen veri kümeleri, bulanık sistem gerçekleştiriminin gereği olarak bir kural üretme adımına tabi tutulmaktadır. Burada iki kural üretme yaklaşımından faydalanılmış ve elde edilen kuralların doğruluğu, nihai özetlerin doğruluğu üzerinden değerlendirilmiştir. Kural üretiminde kullanılan yaklaşımlar WM ve E2E yöntemleridir. Bu iki yöntem birbirinden veri kullanım miktarı açısından ayrılmaktadır. WM yöntemi ile kural üretiminde veri kümesindeki tüm kayıtlar üzerinden öncelikle aday kurallar, ardından nihai kural listesi oluşturulur. E2E'deki kuralların üretilmesinde ise yalnızca özniteliklerin değer aralıklarına ve belirli değer aralıklarında özete değer cümlelerin verideki dağılımına ihtiyaç duyulur. Bu özelliğiyle E2E'deki kural üretme stratejisi verideki kayıtlar üzerinden bir kural çıkarma veya kural en iyileştirmesi yapmadığından veri bağımlılığı çok düşük bir kural üretme yöntemidir.

Tanımlanan sözdizimsel özniteliklerin her biri bulanık sistemin girdi sözel değişkenini oluşturmuştur ve her bir sözel değişken 3 üçgen bulanık küme ile tanımlanmıştır. Bulanık küme parametrelerinin tanımlanmasında WM'in önerdiği şekilde sözel değişkenin tanımlı olduğu aralık, iki kümenin orta noktalarında belirsizliğinin maksimum olacağı şekilde 3 eşit bulanık kümeye bölünmüştür. Çıkış sözel değişkeni ise "düşük önem" ve "yüksek önem" olarak iki yamuk formunda bulanık küme ile ifade edilmiş ve bu bulanık kümelerin sınır parametreleri (0.0, 0.0, 0.2, 0.8), (0.2, 0.8, 1.0, 1.0) olarak tanımlanmıştır.

Girdi ve çıktı sözel değişkenlerinin ve bunlara ait bulanık küme parametrelerinin tanımlanmasının ardından E2E ve WM ile bulanık kurallar üretilmiş ve ayrı bulanık sistemlere bulanık kural tabanı olarak sunulmuştur.

Gerçekleştirilen E2E ve WM kurallarına dayalı Mamdani bulanık çıkarsama sisteminin kural tetikleme aşamasında minimum, kural birleştirmede maksimum operatörü kullanılmıştır. Durulaştırma adımında ise bölgenin merkezi yöntemi ile birleştirme adımının çıktısı olan bulanık kümenin [0, 1] aralığında keskin bir değere dönüştürülmesi sağlanmıştır.

E2E ve WM kurallarına dayalı iki tip bulanık çıkarsama sisteminin yanında, tümüyle veri bağımlı bir yapay sinir ağı gerçekleştirimi de yapılmış ve yöntemlerin DUC2002 veri kümesindeki davranışı ölçülmüştür. Bu yapay sinir ağında 2 gizli katman tanımlanmıştır ve katmanlarda (n girişteki sözdizimsel özniteliklerin sayısı olmak üzere) n, $2n+1$, n ve 1 sinir hücresi bulunmaktadır. Ara katmanlarda RELU, çıkış katmanında sigmoid aktivasyon fonksiyonu kullanılmıştır. Eğitimde Limitli-Hafıza Broyden-Fletcher-Goldfarb-Shanno (BFGS) [121] en iyileştirme algoritması ile logaritmik kayıp fonksiyonu üzerinden yapılmıştır. Erken durdurmada minimum hata farkı 1.00E-08 ve beklenen iterasyon sayısı 10 olarak atanmıştır. Gerçekleştirim, python programlama dilinde scikit-learn [122] kütüphanesi ile yapılmıştır.

Yapay sinir ağı tümüyle veri üzerinden öğrendiğinden, performansının ölçülmesinde beş-katlı çapraz geçerleme uygulanmıştır. Bunun için her dokümanlardan en az bir cümle içerecek ve cümlelere ait her bir etiket değerinin dağılımının veri kümesindeki tüm cümlelerin dağılımına oranına göre örneklem alınmasını sağlayan tabakalı beş-katlı çapraz geçerleme yapılmıştır. Çalışmanın bu bölümünde bulanık sistemler üzerine çapraz geçerleme uygulanmamıştır. Çünkü E2E, kural üretmede verideki kayıtları doğrudan kullanmamaktadır; bu istisnai durum dolayısıyla bulanık sistem gerçekleştirimlerinde veri kümesinin tümü otomatik özet geliştirme sürecinde değerlendirilmiştir.

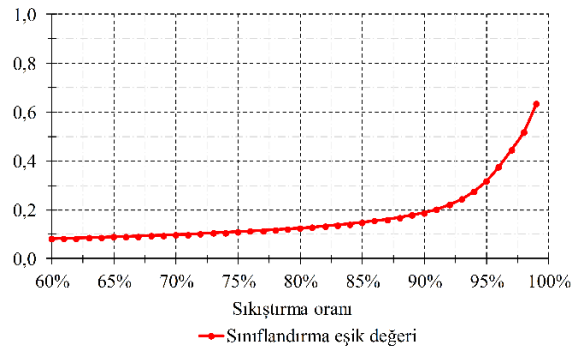
Gerek yapay sinir ağının gerekse bulanık sistemlerin gerçekleştiriminde, model çıktısı [0, 1] aralığındaki önem dereceleridir. Cümlelere atanan önem dereceleri üzerinden metin özetlerinin oluşturulmasında iki yaklaşım kullanılmıştır. Bunların ilkinde problem standart bir cümle sınıflandırma yaklaşımı olarak ele alınmış ve 0.5 sınıflandırma eşik değerinin üstünde önem derecesine sahip olan cümleler özete dâhil edilmiştir. Çizelge 6.2'de DUC2002 veri kümesi üzerinde uygulanmış olan yapay sinir ağının ve iki tip kural üretme yöntemiyle uygulanmış olan bulanık sistemden elde edilen çıkarmalı özetlerin DUC2002 veri kümesinin elle üretilmiş özetlerine göre ROUGE değerleri sunulmuştur.

Çizelge 6.2. DUC2002'de yapay sinir ağı ve bulanık sistemlerin ROUGE sonuçları

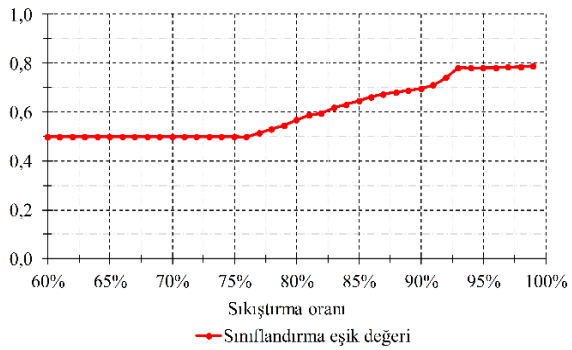
			ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-W	
n	Yöntem	Deney	D	A	D	A	D	A	D	A
5	Yapay sinir ağı	Eğitim	19.5	4.4	11.8	2.4	17.8	5.0	14.9	1.3
	Yapay sinir ağı	Test	9.2	2.0	2.1	0.4	8.2	2.3	5.4	0.5
	Bulanık sistem	E2E	45.4	46.4	31.7	32.6	42.2	42.9	31.9	13.1
	Bulanık sistem	WM	40.6	41.4	25.4	26.2	37.1	37.7	26.5	10.9
6	Yapay sinir ağı	Eğitim	44.4	21.1	26.0	12.9	40.0	21.2	30.5	5.7
	Yapay sinir ağı	Test	14.7	3.2	3.9	0.8	13.1	3.6	8.8	0.8
	Bulanık sistem	E2E	45.4	46.4	31.7	32.6	42.2	43.0	31.8	13.1
	Bulanık sistem	WM	43.0	43.9	28.9	29.7	39.8	40.6	29.3	12.1
7	Yapay sinir ağı	Eğitim	44.4	24.1	28.3	14.8	40.5	23.5	31.5	6.5
	Yapay sinir ağı	Test	17.5	4.3	4.2	1.0	16.0	4.8	10.5	1.0
	Bulanık sistem	E2E	45.9	46.9	31.9	32.8	43.0	43.7	32.3	13.3
	Bulanık sistem	WM	49.1	49.2	36.6	36.6	47.5	47.5	36.3	14.5
8	Yapay sinir ağı	Eğitim	51.1	26.3	31.9	16.1	46.0	25.8	35.7	7.2
	Yapay sinir ağı	Test	17.5	4.1	4.5	1.0	15.9	4.7	10.4	1.0
	Bulanık sistem	E2E	45.9	46.8	32.4	33.3	42.8	43.6	32.9	13.5
	Bulanık sistem	WM	53.4	53.4	41.6	41.6	51.9	51.9	40.4	16.1
9	Yapay sinir ağı	Eğitim	48.6	31.4	32.2	21.1	43.7	30.0	33.6	8.6
	Yapay sinir ağı	Test	18.9	4.7	4.6	1.1	16.7	5.1	10.9	1.0
	Bulanık sistem	E2E	46.0	47.0	32.6	33.4	42.6	43.4	32.7	13.4
	Bulanık sistem	WM	65.5	66.3	58.3	59.1	66.1	66.7	56.1	22.7

Çizelge 6.2'deki deney sonuçlarına göre, yapay sinir ağının cümle sınıflandırma başarısı hem eğitim hem de test başarısı açısından bulanık sistemlerin ikisinin de gerisinde kalmıştır. Kural üretiminde WM algoritmasını kullanan bulanık sistemler, E2E'nin kurallarının sağladığı çıkarsama başarısının önüne geçmiş ve 9 özniteliğin kullanıldığı deneylerde %66.3 anma değerine ulaşabilmiştir. Bu değer E2E ile üretilen kurallarla geliştirilen bulanık sistem için %47.0 düzeyindedir. Yine de E2E tabanlı kurallarla beslenen bulanık sistemde bile, tümüyle veri bağımlı olan yapay sinir ağından daha başarılı çıkarmalı metin özetleri üretilebilmiştir.

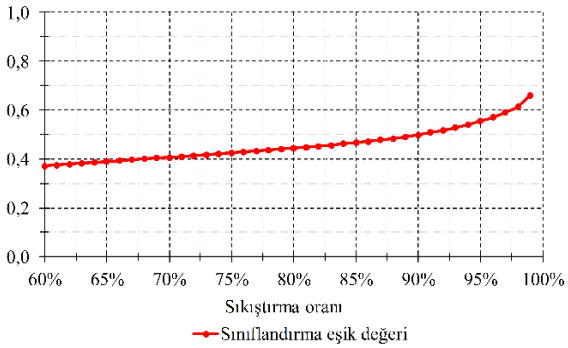
Önem dereceleri tanımlanan cümleler üzerinden metin özetlerinin oluşturulmasında kullanılan ikinci yaklaşımda, cümle önem dereceleri büyükten küçüğe sıralanmış ve en önemli k cümlenin özete dâhil edilmesi sağlanmıştır. Bu durumda değişen k değerine göre metin özetlerinin değerlendirme sonuçları da farklılık gösterecektir. Bu nedenle, yapay sinir ağı ve iki tip bulanık sistem için değişen k değerine göre ortaya çıkan özetlerin ayrı ayrı değerlendirmeleri yapılmıştır. Buna göre metin özetleri üretilirken, k değerinin türetilmesinde [0, 100) aralığında değişen sıkıştırma oranından faydalanılmıştır. Burada minimum sıkıştırma oranında kaynak metindeki tüm cümleler özete dâhil edilirken maksimum sıkıştırma oranında, özette hiçbir cümle bulunmamaktadır. Söz konusu k değeri ilgili sıkıştırma oranına ters orantılı olarak değişmektedir. Hesaplanan k değeri üzerinden de ilgili metin özetleme modelinin ürettiği değerler içerisinden hangisinin sınıflandırma eşiği olarak kabul edilmesi gerektiği bulunmuştur.



(a) Yapay sinir ağı



(b) E2E kuralları ile bulanık sistem



(c) WM kuralları ile bulanık sistem

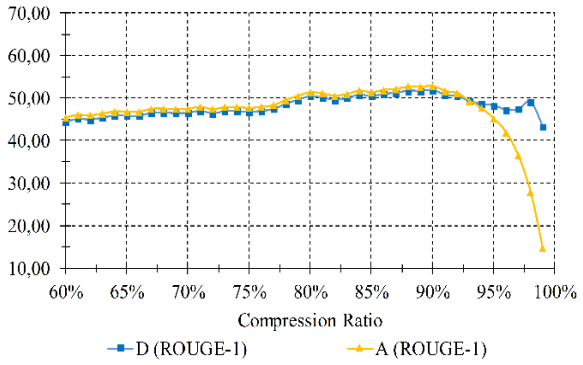
Şekil 6.1. Sıkıştırma oranı, sınıflandırma eşik değeri ilişkisi

Şekil 6.1'de, bu çalışmada DUC2002 veri kümesi için gerçekleştirilmiş olan yapay sinir ağı ve sırasıyla E2E ve WM kurallarına dayalı Mamdani bulanık sistemlerinin üretmiş oldukları

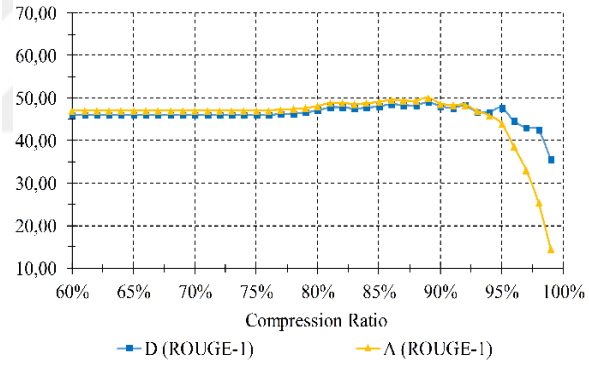
çıktılar üzerinden belirlenen sınıflandırma eşik değerinin, artan sıkıştırma oranına göre değişimi gözlenmektedir.

Şekil 6.1(a) kısmında yapay sinir ağı için artan sıkıştırma oranına göre eşik değerdeki değişim sunulmuştur ve burada görülen eşik değer dağılımına göre Çizelge 6.2'deki ROUGE değerlerinin düşük olması makuldür. Yapay sinir ağı için 0.5 eşik değeri üzerinden ROUGE değeri hesaplamak, bu noktada kötümser bir yaklaşımdır ve bu nedenle, ROUGE değerlerinin tamamı değişen sıkıştırma oranlarına göre de incelenmiştir ve modellerin özetleme başarısı farklı sıkıştırma oranları üzerinden de tartışılmıştır.

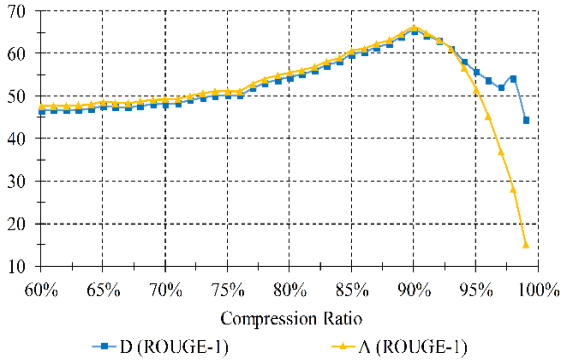
Şekil 6.1(b) ve Şekil 6.1(c)'de bulanık sistemler için tanımlanması gereken sınıflandırma eşik değerinin, değişen sıkıştırma oranına göre değişimi gösterilmiştir. Buna göre E2E için de 0.5 sınıflandırma eşiğinden elde edilen metin özetleri en başarılı özetler olmayacaktır. Ancak WM için 0.5 eşiği makul seviyededir.



(a) Yapay sinir ağı



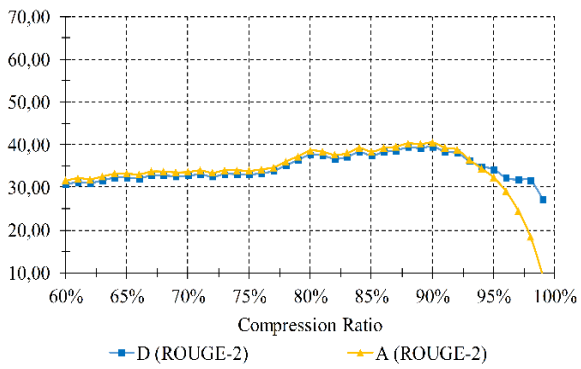
(b) E2E kuralları ile bulanık sistem



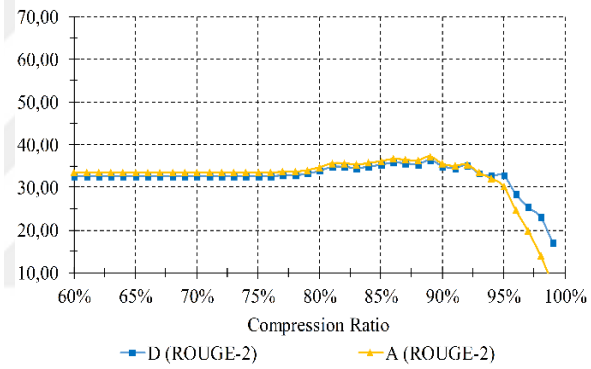
(c) WM kuralları ile bulanık sistem

Şekil 6.2. Sıkıştırma oranı, ROUGE-1 anma değeri ilişkisi

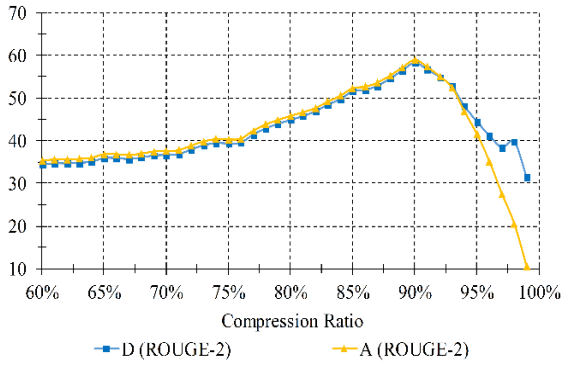
Şekil 6.2, Şekil 6.3 ve Şekil 6.4'te yapay sinir ağı ve sırasıyla E2E ve WM kurallarına dayalı Mamdani bulanık sistemlerinin üretmiş oldukları cümle önem dereceleri üzerinden belirlenen sınıflandırma eşik değerinin değişimine göre hesaplanan ROUGE değerlerinin, artan sıkıştırma oranına göre değişimi gözlenmektedir. Buna göre Şekil 6.2'deki ROUGE-1 değerleri incelendiğinde yapay sinir ağının %90 sıkıştırma oranıyla elde ettiği metin özetlerinin %52.81 anma değerine eriştiği gözlenmektedir. Bu noktada yapay sinir ağı için sınıflandırma eşik değeri 0.19'dur. E2E ile üretilmiş kuralların kullanımıyla oluşturulan bulanık sistemde %90 sıkıştırma oranına göre elde edilen özetlerin ROUGE-1 değeri %50.02'dir ve bu değer 0.69 olan bir sınıflandırma eşiğine göre elde edilmiştir. WM ile üretilen kuralların kullanımıyla elde edilen bulanık sistemde ise %90 sıkıştırma oranına göre sınıflandırma eşiği 0.501 ve ROUGE-1 değeri %66.15'tir.



(a) Yapay sinir ağı



(b) E2E kuralları ile bulanık sistem

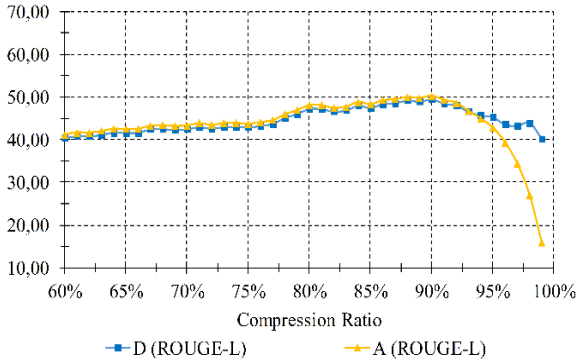


(c) WM kuralları ile bulanık sistem

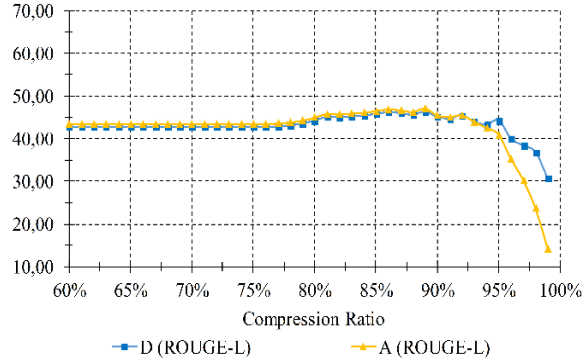
Şekil 6.3. Sıkıştırma oranı, ROUGE-2 anma değeri ilişkisi

Şekil 6.3'te yer alan ROUGE-2 anma değerleri, ROUGE-1 değerlerinden elde edilen bulguya paralel sonuçlar ortaya koymuştur. Tüm ROUGE değerleri için tüm modeller için artan sıkıştırma oranına göre ROUGE anma değerleri hafifçe yükselmektedir ve elde edilen

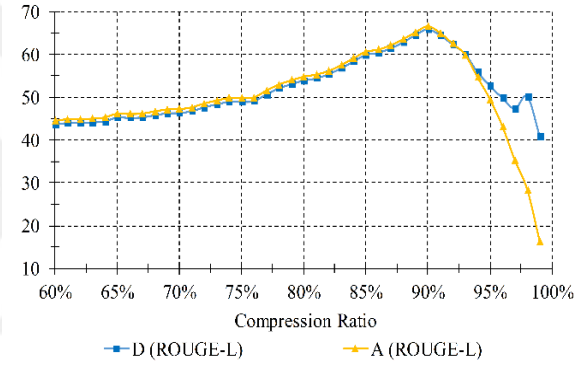
maksimum ROUGE değerleri yaklaşık %90 sıkıştırma oranından elde edilmektedir; %90'un üzerindeki sıkıştırma oranlarında ise ROUGE değerleri belirgin bir şekilde azalmaktadır.



(a) Yapay sinir ağı



(b) E2E kuralları ile bulanık sistem



(c) WM kuralları ile bulanık sistem

Şekil 6.4. Sıkıştırma oranı, ROUGE-L anma değeri ilişkisi

Şekil 6.4'te yer alan ROUGE-L anma değerleri, ROUGE-1 ve ROUGE-2 değerlerinden elde edilen bulguya paralel sonuçlar ortaya koymuştur. Tüm ROUGE değerleri için tüm modeller için artan sıkıştırma oranına göre ROUGE anma değerleri hafifçe yükselmektedir ve elde edilen maksimum ROUGE değerleri yaklaşık %90 sıkıştırma oranından elde edilmektedir; %90'un üzerindeki sıkıştırma oranlarında ise ROUGE değerleri belirgin bir şekilde azalmaktadır.

6.2. DUC Veri Kümesinde LSTM-NN

Bu bölümde, DUC2002 veri kümesi üzerinde, sözdizimsel özniteliklerin ve kelime özyerleşiklerine dayalı anlamsal özniteliklerin kullanılmasının metin özetleme başarısına etkisi incelenmiştir. Bunun için öncelikle DUC2002 içerisindeki 59 başlık altında hazırlanmış olan haber makaleleri aynı başlık altındaki metinler bir dokümanın farklı

paragrafları olacak şekilde birleştirilmiş ve veri kümesi 59 dokümandan oluşan bir forma dönüştürülmüştür. Ardından her bir dokümandaki her bir cümle için sözdizimsel öznitelik değerleri hesaplanmıştır. Burada Çizelge 4.7'de bulunan sözdizimsel özniteliklerden ve burada önerilmiş olan öznitelik hesaplama eşitliklerinden faydalanılmıştır.

Veri kümesindeki dokümanlar, her bir başlık altında hazırlanan metinlerin birleşiminden oluştuğundan bir cümlenin tüm metindeki pozisyonunun ölçülmesi yapılmamış ve Çizelge 4.7'de bulunan ve 28 ve 35 sıra numarasına sahip olan metne göre pozisyon öznitelikleri bu bölümdeki çalışmaya dâhil edilmemiştir. Buna göre dokümanlardaki her bir cümle için 40 öznitelik değerinin hesaplaması yapılmıştır. Veri kümesindeki 400 kelime içeren çıkarmalı özetlerin barındırdığı cümleler biliniyor olduğundan cümlelerin özete dâhil edilip edilmemesini belirten sınıf etiketleri de belirlenebilmiştir.

Veri kümesinden anlamsal öznitelik vektörünün elde edilmesi için 100 ve 200 elemanlı önceden eğitilmiş GloVe ve word2vec kelime özyerleşikleri kullanılmıştır. Şekil 4.5(a)'da bir dokümandaki cümlelerin metin formundan vektörel temsili görülmektedir. Buna göre her satır cümleleri (n yığın boyutu olmak üzere S_i öyle ki $0<i<n$) ifade etmektedir ve her sütun ilgili satıra karşılık gelen cümledeki her bir tekil kelimeye karşılık gelmektedir. Kelimeler [1,100] ve [1,200] boyutlu kelime özyerleşikleri ($w_{i,j}$) ile ifade edildiğinden söz konusu giriş matrisleri 100 ve 200 derinliğine sahiptir. Her bir yığındaki maksimum sıralı dizi boyu, bir cümledeki maksimum tekil kelime sayısı kadardır ve maksimum sıralı dizi boyundan daha az sayıda kelime içeren cümleler için sıralı dizinin doldurulması, cümledeki kelime özyerleşiklerinin ardına, kelime özyerleşikleri ile eşit boyda sıfır vektörleri yerleştirilmesiyle yapılmıştır.

Cümle önem derecelendirme yöntemi olarak sözdizimsel özniteliklerin kullanıldığı durumda yapay sinir ağı, bulanık çıkarsama sistemleri ve bu tez çalışmasında önerilen LSTM-NN olmak üzere üç model üzerinde çalışılmıştır. Sözdizimsel öznitelikler Çizelge 4.7'deki sıra ve hesaplama şekillerine göre kullanılmıştır. Anlamsal öznitelikler ise yalnızca LSTM-NN modeli üzerinde test edilmiştir. Burada yapay sinir ağı ve bulanık çıkarsama sistemlerinin de gerçekleştirilmesindeki amaç aynı öznitelik kümesinin aynı veride kullanımı ile temel alınacak performans değerleri elde edebilmek ve adil bir performans karşılaştırması sunabilmektir.

Uygulanan yapay sinir ağı modeli, 1 giriş katmanı, 3 gizli katman ve 1 çıkış katmanı içermektedir ve n girişteki öznitelik sayısı olmak üzere sırasıyla n, n, $2n$, n, 1 sinir hücresinden oluşan tam bağlı çok katmanlı algılayıcıdır. Çıktı katmanındaki sigmoid temelli sinir hücresinin haricinde katmanlarda aktivasyon fonksiyonu olarak RELU kullanılmıştır. ROC-AUC ve doğruluk ölçütlerini en iyileştirmek için Adam kullanılmıştır. Problem ikili sınıflandırma problemi olarak ele alınmış olduğundan kullanılan kayıp fonksiyonu ikili çapraz entropidir. Modelin eğitimi, minimum hata farkı 10 iterasyon boyunca 1.00E-07 değerinden düşük olduğunda erken durdurmaya maruz kalır.

Modelin eğitiminde en iyileştirilecek ölçütler, tüm deneylerde olduğu gibi doğruluk ve ROC-AUC olarak belirlenmiştir. En iyileştirme yöntemi Adam ve kayıp fonksiyonu ikili çapraz entropi olarak seçilmiştir. Minimum hata farkının üst üste 10 kez 1.00E-07 değerinin altına düşmesi halinde eğitim erken durdurulmaktadır.

Bulanık çıkarsama sisteminin gerçekleştiriminde çıkarılan en yüksek önem derecesine ait 4 sözdizimsel öznitelikten 10 sözdizimsel özniteliğe kadar olan öznitelik kümeleri artırımlı olarak kullanılmıştır. Kullanılan sözdizimsel öznitelikler Çizelge 4.7'deki önem derelerine göre sırasıyla uzunluk (Eş. 2.1), göreceli cümle pozisyonu, uzunluk (Eş. 2.11), tamlamalar (PP) (Eş. 2.15), tamlamalar (NP) (Eş. 2.15), TF-ISF (4-gram) (Eş. 2.3), TF-ISF (3-gram) (Eş. 2.3), tamlamalar (AP) (Eş. 2.15), TF-ISF (2-gram) (Eş. 2.3) olarak belirlenmiştir. Bulanık çıkarsama sistemi Mamdani tipi bir bulanık çıkarsama yapmaktadır. Kural tabanı ve sözdizimsel özniteliklere ait bulanık kümeler WM algoritmasına göre oluşturulmuştur.

LSTM-NN'nin gerçekleştiriminde sözdizimsel öznitelikler, anlamsal öznitelikler ve hibrit öznitelikler olmak üzere üç farklı öznitelik kümesinden yararlanılmıştır. Bölüm 5'te ayrıntılarıyla yer aldığı şekilde bu tez çalışmasında sözdizimsel ve anlamsal özniteliklerin birlikte kullanımının metin özetleme başarısına etkisi incelenmektedir. Bu doğrultuda, DUC veri kümesi için yapılan deneylerde de özniteliklerin tekil kullanımının birlikte kullanımına olan etkisi incelenmiştir. Bu sayede geliştirilen LSTM-NN modelin literatürde yaygın kullanımı olan bir veri kümesindeki davranışı da ölçülmüştür.

Sözdizimsel ve anlamsal özniteliklerin metin özetleme başarısına etkisinin karşılaştırılması adına, geliştirilen yapay sinir ağının ve LSTM-NN tabanlı metin önem derecelendirme modelinin özetleme başarısı, beş-katlı çapraz geçerleme ile ölçülmüştür. Çapraz

geçerlemenin her bir iterasyonunda 48 adet başlık altında hazırlanmış dokümanlar eğitim, 11 başlık altında toplanmış dokümanlar ise test için ayrılmıştır.

Gerçekleştirilmiş olan yapay sinir ağı ve LSTM-NN modellerinden cümlelere ait önem dereceleri elde edilmiştir. Dokümanlardaki tüm cümleler önem derecelerine göre sıralanıp %92 sıkıştırma oranına göre en önemli cümleler seçilmiş ve özete dâhil edilmiştir. Burada sıkıştırma oranının hesaplanmasında kullanılan yaklaşım veri kümesindeki hedef çıkarmalı özetlerdeki cümle sayısının dokümandaki cümle sayısına oranının 0.08 olmasından kaynaklıdır. Bu bağlamda, otomatik metin özetlerinin de hedef çıkarmalı özetlerle aynı sayıda cümle içermesi garanti edilmiş olur.

Çizelge 6.3'te elde edilen yalnızca sözdizimsel öznitelik kullanımı ile yapılan yapay sinir ağı gerçekleştiriminden elde edilen otomatik özetlerin ROUGE değerlerine yer verilmiştir.

Çizelge 6.3. DUC2002 veri kümesinde yapay sinir ağının metin özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
Referans	D	A	D	A	D	A	D	A	D	A	D	A
Kaynak dok.	100	19.5	96.7	11.6	100	19.5	100	20.3	96.5	11.8	100	20.3
Özet dok.	85.7	84.9	81.9	80.1	85.6	84.8	43.9	44.9	25.4	24.6	40.9	41.8

Çizelge 6.3'te görüldüğü gibi yapay sinir ağından sözdizimsel özniteliklerin kullanımı ile elde edilen metin özetlerinin kaynak dokümanın tümü referans alındığında hesaplanan ROUGE-1 anma değeri %20.3'tür. Referans alınan dokümanın, metnin hedef özetleri olması halinde bu değer %44.9 olarak ölçülmüştür. Bu değer literatürde DUC2002 veri kümesi üzerinde yapılmış olan mevcut makine öğrenmesi yöntemleriyle rekabet edebilir düzeydedir.

Bulanık sistemlerin %92 sıkıştırma oranına göre gerçekleştiriminin beş-katlı çapraz geçerlemedeki özetleme başarısını yansıtan ROUGE değerleri Çizelge 6.4'te ve Çizelge 6.5'te sunulmuştur.

Çizelge 6.4. DUC veri kümesinde bulanık sistemin kaynak metne göre özetleme başarısı

Öznitelik sayısı	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
	D	A	D	A	D	A	D	A	D	A	D	A
4	100	20.4	97.5	12.2	100	20.4	100	20.5	97.5	12.2	100	20.5
5	100	21.4	97.0	13.0	100	21.4	100	21.3	97.1	13.0	100	21.3
6	100	21.3	97.1	13.0	100	21.3	100	21.6	97.1	13.1	100	21.6
7	100	21.3	97.2	13.0	100	21.3	100	21.4	97.2	13.0	100	21.4
8	100	21.1	97.2	12.8	100	21.1	100	21.0	97.1	12.8	100	21.0
9	100	21.4	97.1	12.9	100	21.4	100	21.7	97.1	13.1	100	21.7
10	100	21.5	97.0	13.0	100	21.5	100	21.6	97.0	13.1	100	21.6

Çizelge 6.4'te bulanık sistemlerden elde edilen özetlerin kaynak dokümana göre sunulan ROUGE ölçümlerine göre beş ve daha çok sözdizimsel özniteliğin kullanımı ile yüksek anma değerlerine erişildiği görülmektedir. Burada elde edilen anma değerleri %21 civarındadır.

Çizelge 6.5. DUC veri kümesinde bulanık sistemin hedef özete göre özetleme başarısı

Öznitelik sayısı	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
	D	A	D	A	D	A	D	A	D	A	D	A
4	44.0	45.4	24.9	25.2	40.9	42.1	44.4	46.1	25.9	26.3	41.4	43.0
5	47.3	50.8	30.0	32.2	44.6	47.9	46.1	49.2	28.4	30.2	43.4	46.3
6	49.3	53.1	32.6	35.0	46.8	50.3	47.3	51.3	30.3	32.4	44.8	48.4
7	50.2	53.9	33.9	36.1	47.8	51.2	48.0	51.6	31.0	32.7	45.5	48.8
8	50.5	53.8	34.3	36.4	48.2	51.3	48.1	50.8	30.7	32.2	45.5	47.9
9	50.2	54.3	34.3	36.8	47.9	51.7	45.8	49.8	28.3	30.0	42.9	46.6
10	50.2	54.6	34.4	37.0	48.0	52.1	45.0	48.8	27.1	28.9	41.9	45.4

Çizelge 6.5'te bulanık sistemlerden elde edilen özetlerin hedef çıkarmalı özet dokümana göre sunulan ROUGE ölçümlerine göre yedi sözdizimsel özniteliğin kullanımı ile elde edilen ROUGE-A anma değeri %51.6 düzeyindedir ve bu özetlerin başarısı yapay sinir ağı ile elde edilen özetlerin başarısından belirgin oranda yüksektir.

DUC2002 veri kümesi üzerinde, bu tez çalışmasında önerilmiş olan LSTM-NN mimarisinin anlamsal özniteliklerin, sözdizimsel özniteliklerin ve bu özniteliklerin birlikte kullanımından türetilmiş olan hibrit özniteliklerin kullanımı sonucunda elde edilen özetlerin ROUGE analizi, kaynak dokümanların referans alınması ile ayrı ayrı gerçekleştirilmiştir. Analiz sonuçları Çizelge 6.6'da sunulmuştur.

Çizelge 6.6. DUC veri kümesinde LSTM-NN'nin kaynak metne göre özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
Öznitelik	D	A	D	A	D	A	D	A	D	A	D	A
G_{100}	100	16.3	98.1	9.4	100	16.3	100	15.0	95.0	8.2	100	15.0
G_{200}	100	18.0	98.3	10.8	100	18.0	100	22.8	96.9	13.8	100	22.8
W_{100}	100	17.2	99.0	10.2	100	17.2	100	13.3	95.4	7.3	100	13.3
W_{200}	100	17.1	99.2	10.2	100	17.1	100	11.8	92.4	6.3	100	11.8
SyF_{100}	100	17.0	96.8	9.9	100	17.0	100	18.9	96.6	10.8	100	18.9
SyF_{200}	100	18.2	96.9	10.7	100	18.2	100	20.5	96.6	11.9	100	20.5
$G_{100}+SyF_{100}$	100	20.3	96.9	12.6	100	20.3	100	21.0	96.7	12.6	100	21.0
$G_{200}+SyF_{200}$	100	20.3	96.9	12.6	100	20.3	100	21.9	96.6	13.1	100	21.9
$W_{100}+SyF_{100}$	100	20.3	96.9	12.6	100	20.3	100	17.0	96.3	9.7	100	17.0
$W_{200}+SyF_{200}$	100	20.3	96.9	12.6	100	20.3	100	23.1	96.9	14.0	100	23.1

Çizelge 6.6'da kaynak dokümanların referans alınmasından elde edilen ROUGE değerleri G, GloVe ve W ise word2vec kelime özyerleşiklerini ve SyF sözdizimsel öznitelikleri temsil edecek şekilde gösterilmektedir. Alt indisle gösterilen değerler özyerleşik boyunu ifade etmektedir. Kaynak dokümana göre elde edilen en yüksek ROUGE-1 anma değeri word2vec kelime özyerleşiklerinin sözdizimsel özniteliklerle birlikte kullanımından elde edilmiş ve %23.1 olarak ölçülmüştür.

Çizelge 6.7'de DUC veri kümesi üzerinde LSTM-NN'den elde edilen özetlerin referans alınan çıkarmalı özetlere göre ROUGE değerleri sunulmuştur.

Çizelge 6.7. DUC veri kümesinde LSTM-NN'nin hedef özete göre özetleme başarısı

	Eğitim başarısı (%)						Test başarısı (%)					
	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-1		ROUGE-2		ROUGE-L	
Öznitelik	D	A	D	A	D	A	D	A	D	A	D	A
G_{100}	40.3	33.6	17.9	14.0	37.1	30.9	40.6	30.4	17.3	11.7	37.4	27.8
G_{200}	59.4	56.2	44.9	42.8	57.3	54.3	41.3	46.6	22.8	25.1	38.4	43.3
W_{100}	39.9	35.2	17.5	15.0	36.5	32.3	39.4	26.7	13.2	8.5	36.2	24.4
W_{200}	39.8	35.1	17.6	15.2	36.5	32.2	39.2	23.6	13.4	7.6	36.5	21.7
SyF_{100}	100	87.6	99.5	84.7	100	87.6	47.0	44.3	28.5	25.2	44.2	41.7
SyF_{200}	89.4	81.0	85.6	75.1	88.9	80.5	43.6	45.0	24.5	24.1	40.6	42.0
G_{100}+SyF_{100}	100	100	100	100	100	100	43.1	44.9	23.7	24.7	40.4	42.0
G_{200}+SyF_{200}	100	100	100	100	100	100	43.5	48.0	25.3	28.2	40.8	45.1
W_{100}+SyF_{100}	100	100	100	100	100	100	46.8	39.0	24.5	20.8	43.3	36.3
W_{200}+SyF_{200}	100	100	100	100	100	100	44.1	51.4	26.6	31.6	41.5	48.5

Çizelge 6.7'deki ROUGE ölçümleri DUC veri kümesindeki hedef çıkarmalı metin özetlerinin referans alınmasından elde edilmiştir. Sözdizimsel özniteliklerin tekil kullanımından elde edilen ROUGE-1 anma değeri %44.4 ve %45.0 olarak ölçülmüştür. Bu özelliğiyle yapay sinir ağının özetleme başarısına yakın değerler elde edilmiştir. Çizelgede GloVe kelime özyerleşikleri dikkate alındığında anlamsal özniteliklerin tekil kullanımından elde edilen maksimum ROUGE-1 anma değeri 46.6'dır. Bu değer word2vec kelime özyerleşikleri için %23.6 olarak hesaplanmıştır. Bu noktada, LSTM-NN word2vec özyerleşiklerinin tekil kullanımında yeterli olgunlukta metin özetleri elde edememiştir. Ancak, LSTM-NN'nin metin özetlemede sözdizimsel ve anlamsal öznitelikleri birlikte kullanmasıyla GloVe özyerleşikleri üzerindeki başarı ortalama %2.2 artarak %48.0'a, word2vec özyerleşikleri üzerindeki başarı ise %51.4'e ulaşmıştır.

DUC2002 veri kümesi çoklu belge metin özetleme veri kümesidir. Ancak LSTM-NN, çoklu belge için geliştirilmiş bir çözüm değildir. Bu özelliğiyle LSTM-NN'nin DUC2002 veri kümesindeki başarısının literatürdeki mevcut çalışmalarla karşılaştırılması adil bir kıyaslama sunmayabilir. Ancak yine de LSTM-NN, literatürdeki derin öğrenmeye dayalı 2 mevcut çalışma ile kıyaslanmış ve sonuçlar Çizelge 6.8'de sunulmuştur.

Çizelge 6.8. DUC2002 veri kümesi üzerinde LSTM-NN yöntem karşılaştırması

			ROUGE-1		ROUGE-2		ROUGE-L	
Ref.	Yöntem	Özyerleşik	D	A	D	A	D	A
[34]	SummaRuNNer	word2vec	-	46.6	-	23.1	-	43.0
[15]	Otokodlayıcı	word2vec	-	51.7	-	27.5	-	44.6
-	Bulanık sistem	sözdizimsel	48.0	51.6	31.0	32.7	45.5	48.8
-	Yapay sinir ağı	sözdizimsel	43.9	44.9	25.4	24.6	40.9	41.8
-	LSTM-NN	GloVe	43.5	48.0	25.3	28.2	40.8	45.1
-	LSTM-NN	word2vec	44.1	51.4	26.6	31.6	41.5	48.5

Çizelge 6.8'e göre SummaRuNNer yönteminden elde edilen metin özetleme anma değerleri ROUGE-1, ROUGE-2 ve ROUGE-L için sırasıyla %46.6, %23.1, %43.0'dır. Otokodlayıcı ve takip eden doğrusal cümle önem derecelendirme fonksiyonuna bağlı olan bir diğer derin öğrenme yönteminden %51.7 ROUGE-1, %27.5 ROUGE-2 ve %44.6 ROUGE-L değerleri üretilmiştir. Öte yandan, LSTM-NN ile aynı sözdizimsel öznitelik kümesinin bir alt kümesini kullanan bulanık sistemden elde edilen ROUGE-1 değeri %51.6, ROUGE-2 değeri %32.7 ve ROUGE-L değeri %48.8'dir. Sözdizimsel özniteliklerle beslenen yapay sinir ağının ise %44.9 ROUGE-1, %24.6 ROUGE-2 ve %41.8 ROUGE-L değerleri üretilmiştir. Temel alınan bu yöntemlerden elde edilen özetlere ait ROUGE değerlendirmesine göre, LSTM-NN'nin DUC2002 veri kümesindeki özetleme başarısı literatürdeki çalışmalarla rekabet edebilir boyuttadır. Ancak LSTM-NN'nin çoklu belge metin özetleme problemine uygun olarak tasarlanmamış olması özet başarısının daha yüksek seviyelere erişmesine engel niteliğindedir.

7. TARTIŞMA

Metin özetleme, orijinal belgenin ana içeriğini ve genel anlamını olabildiğince koruyarak kaynak belgenin özet adı verilen daha kısa bir sürüm halinde yeniden yapılandırılması işlemidir. Metin özetleme, bilgi erişim sistemleri açısından, sayısı ve karmaşıklığı artan metin dokümanları için bir boyut indirgeme yöntemi olması dolayısıyla; okur açısından ise bir konuda yazılmış olan bir veya birden çok kaynaktan elde edilen metinleri okumaya gerek duymadan içerisindeki önemli bilgiyi çıkarmayı sağladığı için değerli ve gereklidir. Bu işlemin insanlar tarafından elle yürütülmesi muhakkak ki zor ve maliyetli olması sebebiyle uygulanabilir değildir. Bu nedenle özetlemenin otomatik olarak gerçekleştirilmesi, bilgisayar bilimlerinde, metnin içindeki temel bilgiyi daha küçük boyuttaki bir metin içinde ifade etmeyi hedefleyen yeni bir araştırma problemi olarak değerlendirilmektedir.

Otomatik metin özetleme yaklaşımları soyutlamalı ve çıkarmalı metin özetleme olmak üzere iki şekilde ele alınmaktadır. Soyutlamalı metin özetlemede bir metnin içerisindeki temel kavramlar ve kavramlar arasındaki anlamsal ilişkiler ortaya çıkarılarak doğal dilin kurallarına ve yapısına uygun yeni bir özet metin üretilir. Öte yandan, çıkarmalı metin özetlemede metni oluşturan bölümler (genellikle cümleler) için çeşitli özellikler doğrultusunda önem dereceleri atanmakta ve bu önem dereceleri dikkate alınarak bölümler özet metne doğrudan aktarılmaktadır. Çıkarmalı metin özetleri, doğası gereği üzerinde çalışılan doğal dilin kurallarına uygun ve anlamlıdır. Bölümlere önem dereceleri atama ve önem derecesine göre metinden bölüm seçme süreçleri bütünüyle dilden bağımsızdır ve uzun metinler için uygulanabilirliği soyutlamalı metin özetlemeye nazaran daha yüksektir.

Literatürde gerek soyutlamalı gerekse çıkarmalı metin özetleme alanlarında doğrudan özetleme görevi için üretilmiş çok az sayıda veri kümesi bulunmaktadır ve metin özetleme çalışmalarının tamamına yakını bu veri kümeleri üzerinde özetleme yöntemlerinin başarılarını değerlendirmektedir. Bu veri kümelerinin çoğu insanlar tarafından yazılmış özet dokümanları içermemektedir. Bunun yerine çeşitli kurallarla dokümanlara ait hedef özetlerin hangi bilgiyi içermesi gerektiğine karar vermiş ve özetleme yaklaşımlarını bu algoritmik yapıyı öğrenir hale getirmeyi hedeflemişlerdir. İnsanlar tarafından geliştirilmiş hedef özetlere sahip veri kümesinin çok az sayıda olması, bu alanda yeni bir özetleme verisinin geliştirilmesine olan ihtiyacı ortaya çıkarmıştır. Çünkü az sayıdaki veri kümesi üzerinden yapılan analizler, üzerinde çalışılan özetleme yönteminin başarısını yeterince

yansıtamamakta ve karşılaştırmalı gözlemleri sınırlandırmaktadır. Bu amaçla, bu tez çalışmasında otomatik metin özetleme problemi için özelleşmiş olan, akademik yayın ve bu yayınlara ait soyutlamalı ve çıkarmalı metin özetlerini barındıran yeni bir veri kümesi SIGIR2018 adıyla üretilmiştir. Bunun için 2018 yılında gerçekleşen 41. Uluslararası ACM SIGIR Bilgi Erişiminde Araştırma ve Geliştirme Konferansı'nda yayınlanmış bildirilerin tam metni indirilmiştir. Önerilen veri kümesinde hedef soyutlamalı özetler, yayının kendi yazarları tarafından hazırlanmış olan özet bölümlerinden elde edilmiştir. Hedef çıkarmalı metin özetleri ise bilgisayar mühendisliğinden üç okurun metinlerde özete dâhil etmeye değer buldukları cümleleri seçmesi ile oluşturulmuştur. Okurlar, diğer okuyucuların seçtikleri cümlelerden haberdar olmadan metinlerdeki cümleler içerisinden kendileri için en çok önem arz eden cümleleri seçerek kendi çıkarmalı özetlerini oluşturmuşlardır. Her okurun seçtiği cümleler o okurun aday cümle kümesi olarak değerlendirilmiştir ve aday cümle kümelerindeki cümle sayısı kaynak metnin cümle sayısının üçte biri ile sınırlandırılmıştır. Bu sürecin sonunda tüm dokümanlar için üç aday cümle kümesi meydana gelmiştir. Ardından, her bir doküman için aday cümle kümelerinin kesişimi ve birleşimi üzerinden iki hedef çıkarmalı özet elde edilmiştir. Böylece veri kümesindeki her yayın için bir soyutlamalı ve iki çıkarmalı hedef metin özeti meydana gelmiştir. Aday cümle kümelerinin kesişiminden elde edilen çıkarmalı özetler ve hedef soyutlamalı özetler kaynak dokümandaki cümlelerin yaklaşık olarak %23'ünü içermektedir; öte yandan bu oran aday cümle kümelerinin birleşiminden elde edilen çıkarmalı özetler için %50 civarındadır.

SIGIR2018 veri kümesinden elle üretilmiş olan metin özetlerinin kaynak metni ne oranda yansıttığının ve özetlerin doğru ve tutarlı oluşunun değerlendirilmesinde, metin özetlemede en sık kullanılan özet değerlendirme metodu olan ROUGE ölçümlerinden yararlanılmıştır. ROUGE, iki metin arasında n-gram eşleştirmesine dayanan ve özet değerlendirmede anma odaklı yardımcı bir metottur. Genel uygulanmasında, geliştirilen otomatik özet metni, referans anılacak hedef özet metni ile n-gram tabanlı eşleşmesine göre analiz edilerek hesaplanır. Ancak bu tez çalışmasında, geleneksel ROUGE uygulamasının yanında referans metni olarak kaynak metinler de değerlendirilmeye alınmış ve bu sayede, geliştirilmiş olan özet metnin kaynak metinle eşleşme oranı ölçülmüştür.

Hazırlanmış olan soyutlamalı özetle iki çıkarmalı metin özeti, kaynak metnin referans kabul edildiği şekilde ROUGE analizine tabi tutulmuştur ve aday cümle kümelerinin kesişiminden elde edilen çıkarmalı özetlerden %38.4, aday cümle kümelerinin birleşiminden elde edilen

özetlerden %64.9 ROUGE değeri elde edilmiştir. Aday cümle kümelerinin birleşiminden elde edilen çıkarmalı özetlerin ROUGE değerindeki baskınlığının sebebi kaynak dokümandaki cümlelerin neredeyse yarısını içeriyor olması ve bu çıkarmalı özetlerin cümle sayısı açısından avantajlı konumda bulunmasıdır. Öte yandan, içerdiği cümle sayısı aday cümle kümelerinin kesişimi olan çıkarmalı özetlerle örtüşen soyutlamalı özetler için bu değer %23.7 olarak ölçülmüştür. Buna göre çıkarmalı özetlerin kaynak metni temsil yeteneği yayının orijinal yazarları tarafından hazırlanmış olan soyutlamalı özetlerden daha yüksektir.

Bu tez çalışmasında çıkarmalı metin özetleme problemi, metindeki önemli cümlelerin seçilmesi ve en önemli cümlelerin özete dâhil edilmesi süreci olarak ele alınmıştır. Burada önemli cümlelerin seçimi için en önemli nokta önem tanımının mümkün olduğunca doğru ve genel yapılabiliyor olmasıdır. Bunun için literatürde metin özetleme çalışmalarında cümle öneminin belirlenmesi için kullanılmış iki öğrenme kanalı olan sözdizimsel öznitelikler ve anlamsal öznitelikler derinlemesine araştırılmıştır. Önerilen SIGIR veri kümesi üzerinde bu özniteliklerin, literatürdeki mevcut çalışmalardaki gibi tek başına kullanımının metin özetlemeye etkisi araştırılmıştır ve bu iki yaklaşım üzerinden elde edilen metin özetleri gerek birbirleriyle gerekse literatürdeki diğer popüler çalışmalarla karşılaştırılmıştır. Bunun için mevcut çalışmalar içerisinden sözdizimsel öznitelikler üzerinde çalışan bulanık kural tabanlı sistemlerin ve veri bağımlı çok katmanlı algılayıcı yapay sinir ağının gerçekleştirimi yapılmıştır. Sözdizimsel özniteliklerin kendi içinde metin özetlemeye bireysel katkısının ölçülebilmesi adına sözdizimsel öznitelikler 4 öznitelikten 42 özniteliğe kadar genişletilmiştir. Farklı öznitelik kümesi seçimi için yapılan farklı deneyler söz konusu olmakla birlikte, aday cümle kümelerinin kesişim kümesi olan çıkarmalı özetler referans alındığında elde edilen en yüksek ROUGE değerleri bulanık sistem için %66.7, yapay sinir ağı için %40.1 olarak ölçülmüştür.

SIGIR2018 veri kümesinde anlamsal özniteliklerin literatürde mevcut olan özetleme yöntemleriyle testinde, literatürde sıklıkla rastlanan SummaRuNNer isimli modelin gerçekleştirimi yapılmıştır. Bu model, GRU tabanlı hiyerarşik bir derin öğrenme yaklaşımıdır ve cümleleri meydana getiren terimlerin kelime özyerleşiklerinden ve birkaç sözdizimsel öznitelikten beslenerek cümlelere önem derecesi atama işi yapar. Bu metodun SIGIR2018 veri kümesinde üç tip uygulaması yapılmış ve elde edilen maksimum ROUGE değeri %62.3 olarak ölçülmüştür. Bu bağlamda sözdizimsel öznitelik kullanımı ile bir

bulanık sistem gerçekleştirimi yapmak, anlamsal özniteliklerle derin öğrenme gerçekleştirimi yapmakla rekabet edebilir sonuçlar üretmiştir.

Literatürdeki mevcut metin özetleme yöntemlerinin SIGIR2018 veri kümesinde uygulanmasında son olarak, denetimsiz öğrenme yaklaşımı olan TextRank yönteminden faydalanılmıştır. Burada metinden sözdizimsel veya anlamsal öznitelikler çıkarılmamış; cümlelerin birbiri ile ilişkisi üzerinden bir cümle gruplaması yapılarak yüksek ilişkiye sahip olan cümleler önemli olarak değerlendirilmiştir. Bu yöntemden elde edilen en yüksek ROUGE değeri ise %47.9 olarak ölçülmüş; bu değer uygulanmış olan diğer makine öğrenmesi yöntemlerinin belirgin oranda gerisinde kalmıştır.

Bu tez çalışmasında çıkarmalı metin özetleme problemi için özete dâhil edilmeye değer cümlelerin seçimini sağlayan özyineli sinir ağı ile yapay sinir ağının birlikte kullanımından elde edilen bir çözüm mimarisi, Uzun-Kısa Dönem Bellek-Sinir Ağı (LSTM-NN) adıyla önerilmiştir. Bu mimari üzerinde öncelikle önceden eğitilmiş kelime özyerleşiklerine dayalı anlamsal özniteliklerin tek başına kullanılmasının SIGIR2018 veri kümesindeki özetleme başarısı ölçülmüştür. Bu yaklaşımla, aday cümle kümelerinin kesişimi olan çıkarmalı metin özetlerine göre %60.5 ROUGE değerine erişilebilmiştir. Akabinde, yine bu çalışmada önerilmiş olan ve literatürdeki birçok çalışmada yer alan öznitelik hesaplamalarını barındıran geniş sözdizimsel öznitelik listesi üzerinde LSTM-NN uygulaması yapılmış ve %57.3 ROUGE değeri elde edilmiştir. Bu değer, LSTM-NN'nin anlamsal öznitelik kullanımı ile gerçekleştiriminden elde edilen ROUGE başarısının yaklaşık %3 gerisindedir.

Sözdizimsel ve anlamsal özniteliklerin tekil kullanımının ardından tez çalışmasında önerilen yaklaşım olan hibrit öznitelik kullanımının gerçekleştirimi SIGIR2018 veri kümesi üzerinde yapılmıştır. Burada temel alınan varsayım, özetleme sürecinin, özetleyen mekanizmanın (insan veya makine) niyeti ve metnin çeşitli karakteristik özelliklerine göre farklı öğrenme kanallarından besleniyor ve etkileniyor olması ve burada tüm dokümanlar için genel bir özet metin meydana getirmede tüm bu kanalların dikkate alınmasının gerekliliğidir. Bu doğrultuda, LSTM-NN, sözdizimsel ve anlamsal özniteliklerin birlikte kullanımına imkân verecek hibrit bir yaklaşımla genişletilmiş ve ortaya çıkan çıkarmalı özetlerin performansı gerek özet değerlendirme ölçütleri gerekse standart sınıflandırma performans değerlendirme ölçütleri açısından ölçülmüştür. Performans ölçümleri sonucunda özniteliklerin birlikte kullanımının metin özetlerinin aday cümle kümelerinin kesişimi olan çıkarmalı özetlere göre

hesaplanan ROUGE değerlerini %65.2 değerine kadar yükselttiği görülmüştür. Buna göre, hibrit öznitelik uzayı kullanımıyla LSTM-NN, hem özniteliklerin tekli kullanımının hem de SIGIR2018 veri kümesinde uygulanmış mevcut makine öğrenmesi yaklaşımlarının üstünde performans göstermiş olmakla birlikte bu çalışmaya en yakın araştırma olarak değerlendirilen SummaRuNNer özetleme başarısının da önüne geçmiştir.

LSTM-NN'nin kaynak metne göre ROUGE ölçüm sonuçları incelendiğinde %40.1 anma değerine eriştiği gözlenmiştir. Bu rakam elle üretilmiş olan aday cümle kümelerinin kesişimine karşılık gelen çıkarmalı özetler için %38.4'tür. Bu durumda LSTM-NN'nin metni temsil etmede elle çıkarılan özetlere nazaran da daha başarılı olduğu sonucuna da varılmıştır.

Tez çalışması kapsamında yapılan ek araştırmalarda, yapay sinir ağlarının ve bulanık çıkarsama sistemlerinin metin özetlemedeki başarısı, değişen sıkıştırma oranı açısından incelenmiştir. Buradaki çalışmalarda, literatürde en yaygın kullanıma sahip DUC2002 özetleme veri kümesi kullanılmıştır. Özet değerlendirmesi, hedef çıkarmalı özetlerin referans alınmasıyla hesaplanan ROUGE ölçütleri açısından yapılmıştır. Uygulanan iki yöntem için de en başarılı ROUGE anma değerlerine %90 sıkıştırma oranının kullanımı ile erişilebilmiştir. Buna göre, yapay sinir ağının %90 sıkıştırma oranıyla elde ettiği metin özetlerinin ROUGE-1 anma değeri %52.81 olarak ölçülmüştür. Bu noktada yapay sinir ağı için sınıflandırma eşik değeri 0.19'dur.

Bulanık sistemler üzerinde yapılan sıkıştırma oranı analizlerinde ise dokuz özniteliğin kullanıldığı deneysel çalışmalar sonucunda %90 sıkıştırma oranına göre sınıflandırma eşiği 0.501 ve ROUGE-1 anma değeri %66.15 olarak ölçülmüştür. Bu değer artan sıkıştırma oranı ile azalmaktadır.

Ek araştırmalar içerisinde ayrıca, önerilen LSTM-NN'nin sıklıkla kullanılan DUC özetleme veri kümesi üzerinde model değerlendirmesi yapılmıştır. Bunun için ilk olarak, yeni ve kapsamlı bir sözdizimsel öznitelik kümesi oluşturulmuştur ve bu özniteliklerin tekil kullanımının metin özetleme başarısına etkisi gözlenmiştir. Buna göre, sözniteliklerin tekil kullanımından elde edilen ROUGE-1 anma değeri %44.4 ve %45.0 olarak ölçülmüştür. GloVe kelime özyerleşikleri dikkate alındığında anlamsal özniteliklerin tekil kullanımından elde edilen maksimum ROUGE-1 anma değeri 46.6'dır. Bu değer word2vec kelime

özyerleşikleri için %23.6 olarak hesaplanmıştır. Bu noktada, LSTM-NN word2vec özyerleşiklerinin tekil kullanımında yeterli olgunlukta metin özetleri elde edememiştir. Ancak, LSTM-NN'nin metin özetlemede sözdizimsel ve anlamsal öznitelikleri birlikte kullanmasıyla GloVe özyerleşikleri üzerindeki başarı ortalama %2.2 artarak %48.0'a, word2vec özyerleşikleri üzerindeki başarı ise %51.4'e ulaşmıştır. Deneysel çalışmalardan elde edilen sonuçlar metin özetleme probleminin sözdizimsel ve anlamsal özniteliklerin birlikteliği üzerinden tartışılması gerektiğini bir kez daha ortaya koymaktadır.

8. SONUÇ

Otomatik metin özetleme, orijinal belgenin ana içeriğini ve genel anlamını olabildiğince koruyarak kaynak belgenin özet adı verilen daha kısa bir sürüm halinde yeniden yapılandırılması işlemidir. Literatürde soyutlamalı ve çıkarmalı metin özetleme olmak üzere iki tür otomatik özetleme yaklaşımı vardır. Bu tez çalışmasında incelenmiş olan çıkarmalı metin özetlemede metindeki cümleler için çeşitli özellikler doğrultusunda önem dereceleri atanmakta ve bu önem dereceleri dikkate alınarak bölümler özet metne doğrudan aktarılmaktadır.

Bu tez çalışmasında, çıkarmalı metin özetlerinin üretilmesinde cümlelere önem derecelerinin atanmasında kullanılan özellikler sözdizimsel ve anlamsal olmak üzere iki tipte ele alınmıştır. Metnin türü, yapısı ve özetlemedeki amaç doğrultusunda, önem tanımının değişmesi dolayısıyla önemi belirleyen özelliklerin de farklılık göstereceği üzerinde durulmuştur. Buna göre yalnızca sözdizimsel veya yalnızca anlamsal öznitelikler üzerinden yapılacak bir önem derecelendirmenin genel özetler üretmede yetersiz kalabileceği yargısıyla, metin özetleme problemine sözdizimsel ve anlamsal özniteliklerin birlikte kullanımını öneren bir çözüm mimarisi olan LSTM-NN geliştirilmiştir.

Tez kapsamında geliştirilmiş olan LSTM-NN, literatürde yaygın kullanıma sahip olan DUC çoklu belge veri kümesi ile bu tez çalışmasında sunulmuş olan SIGIR2018 veri kümesi üzerinde ROUGE analizine dayanan değerlendirmelere tabi tutulmuştur. Bunun yanı sıra, bu veri kümeleri üzerinde bulanık çıkarsama sistemlerine, yapay sinir ağlarına, derin öğrenme tabanlı SummaRuNNer isimli modele dayanan metin özetleme yöntemlerinin gerçekleştirimi yapılmış ve yapılan deneysel çalışmalarla üretilen LSTM-NN metin özetleri bu modellerin çıktısı olan özetlerle karşılaştırılmıştır. Elde edilen sonuçlar doğrultusunda tez çalışmasının temel bulguları aşağıda özetlenmiştir:

Metin özetleme problemi tek başına tamamıyla anlamsal bir problem olarak düşünülmemeli yalnızca cümleleri oluşturan kelimelerin anlamları üzerinden değerlendirilmemelidir.

Üzerinde çalışılan metnin yapısallığı, karakteristik özellikleri, özetlemenin niyeti ve kapsamı cümlelerin önem tanımı için belirleyici unsurlardır.

Cümlelere ait önem derecelerinin belirlenmesinde sözdizimsel öznitelikler, kelime özyerleşiklerine dayanan anlamsal özniteliklerden daha belirleyici olabilmektedir. Ancak bu noktada genel bir yargıya varılamaz. Özetlemenin niyeti ve özetlenen metnin yapısallığı bu konudaki bulguyu değiştirebilmektedir.

LSTM-NN'nin sözdizimsel ve anlamsal özniteliklerin birlikte kullanımından elde edilen özetleri, bu özniteliklerin tekli kullanımından elde edilen özetlerine nazaran gerek kaynak metni yansıtma gerekse hedef özete yaklaşma açısından daha başarılıdır. Buna göre genel bir çıkarmalı metin özetlemede amaç önemli cümlelerin tespitidir ve bu tespitin yapılabilmesinde sözdizimsel ve anlamsal özniteliklerin birlikte değerlendirilmesi gerekir.

Özniteliklerin birlikteliğini temel alan LSTM-NN'nin ürettiği özetlerin başarısı literatürdeki mevcut çalışmalarla rekabet edebilir düzeydedir. Ayrıca, yapılan çok sayıdaki deneysel çalışmada bu özetlerin bilindik ve yaygın kullanıma sahip birçok çalışmadan kaynak metnin yerini tutma ve hedef özet metne yaklaşma açısından daha yüksek performans değerlerine erişebildiği gözlenmiştir.

Tez çalışmasından elde edilen akademik çıktılar Ek-1, Ek-2, Ek-3 ve Ek-4'te sunulmuştur. Bölüm 6.1'de bulanık çıkarsama sistemlerinin ve yapay sinir ağlarının DUC2002 veri kümesindeki metin özetleme performansı ve sıkıştırma oranı konusundaki çalışmalar "Knowledge-Based Systems" isimli dergide "Multi-document extractive text summarization: A comparative assessment on features" başlıklı makale ile yayınlanmıştır. Makalenin ilk sayfası Ek-1'de sunulmuştur. Söz konusu çalışmada bulanık sistemin gerçekleştiriminde kural üretme aşamasında kullanılmış olan yöntem, "End-to-End Hiyerarchical Fuzzy Inference Solution" başlıklı bildiride önerilmiş ve çalışma "2018 IEEE World-Congress on Computational Intelligence" isimli konferansta sunulmuştur. Bildirinin ilk sayfası Ek-2'de verilmiştir.

Bölüm 5'te ayrıntılarına yer verilmiş olan metin özetleme öznitelikleri ESA tabanlı bir mimari ile bulanık çıkarsama sistemleri üzerinden DUC2002 veri kümesinin kullanımı ile karşılaştırılmış ve bulgular "Investigation of Text Summarization Features by Fuzzy Systems and Convolutional Neural Networks" başlıklı bildiri çalışmasında derlenmiştir. Çalışma, "SEMANTiCS 2020" isimli konferansa aday bildiri olarak gönderilmiştir. Ek-3'te gösterilmiş olan bildiri hakem değerlendirme sürecindedir.

Bölüm 4'te önerilmiş olan veri kümesi ve Bölüm 5'te bu veri kümesi üzerinde yapılan metin özetleme çalışmalarından tez kapsamında önerilmiş olan mimari, "Candidate Sentence Selection for Extractive Text Summarization" başlığıyla Information "Processing & Management" isimli dergiye gönderilmiştir. Ek-4'te ilk sayfasına yer verilmiş olan makale düzeltme sürecindedir.

Geliştirilen çözüm mimarisi LSTM-NN, çoklu belge metin özetleme problemi için özelleşmiş bir çözüm sunmamaktadır. Bu sebeple çoklu belge metin özetleme veri kümesindeki performansı tek belge metin özetleme veri kümesindeki performansından nispeten daha zayıftır. Çalışmanın ilerleyen kısımlarında modelin bu çoklu belge metin özetlemeye nasıl uyarlanabileceği üzerine çalışmalar yapılacak bu sayede heterojen veri kaynaklarından toplanan belgelerin, bu belgelerdeki temel yargıları yansıtacak şekilde merkezi bir noktadan özetlenebilmesi sağlanacaktır.

KAYNAKLAR

1. Mutlu, B., Sezer, E. A. and Akcayol, M. A. (2019). Multi-document extractive text summarization: A comparative assessment on features. *Knowledge-Based Systems*. 183(1), 1-13.

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

3. Nallapati, R., Zhou, B., Dos Santos, C. N., Gulcehre, and C., Xiang, B. (2016). *Abstractive text summarization using sequence-to-sequence RNNs and beyond*. 20th Special Interest Group on Natural Language Learning Conference on Computational Natural Language Learning, Berlin, Germany.

4. Lopyrev, K. (2015). Generating news headlines with recurrent neural networks. *Journal of Controlled Release*. 230(1), 1-9.

5. Nasar, Z., Jaffry, S. W. and Malik, M. K. (2019). Textual keyword extraction and summarization: state-of-the-art. *Information Processing & Management*. 56(102088), 1-31.

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

7. Miao, L., Cao, D., Li, J. and Guan, W. (Jan. 2020). Multi-modal product title compression. *Information Processing & Management*. 57(1), 102-123.

8. Zajic, D., Dorr, B. J., Lin, J. and Schwartz, R. (2007). Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing and Management*. 43(6), 1549-1570.

9. Krahmer, E., Marsi, E. and Van Pelt, P. (2008). *Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion*. 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, Unites States.

10. Mohamed, A. and Rajasekaran, S. (2008). Query-based summarization based on document graphs. *Engineering*. 1(8), 142-146.

11. Zhu, S., Cheng, X. and Su, S. (2020). Knowledge-based question answering by tree-to-sequence learning. *Neurocomputing*. 372(1), 64-72.

12. Verma, P. and Verma, A. (2020). A review on text summarization techniques. *Journal of Scientific Research*. 64(1), 251-257.

13. Xie, Y., Zhou, T., Mao, Y. and Chen, W. (2020). Conditional self-attention for query-based summarization. *arXiv:2002.07338*. 1(1), 1-14.

14. Xu, Y. and Lapata, M. (2020). Query focused multi-document summarization with distant supervision. *arXiv:2004.03027*. 1(1), 1-11.

15. Joshi, A., Fidalgo, E., Alegre, E. and Fernández-Robles, L. (2019). SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders. *Expert Systems with Applications*. 129(1), 200-215.

16. Hu, Y. H., Chen, Y. L. and Chou, H. L. (2017). Opinion mining from online hotel reviews - a text summarization approach. *Information Processing & Management*. 53(2), 436-449.

17. Demirci, M. S. and Sağıroğlu, S. (2017). Sosyal ağ verilerinin kullanım alanları üzerine kapsamlı bir inceleme. *Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji*. 5(2), 1-21.

18. Litvak, M., Vanetik, N., Last, M. and Churkin, E. (2016). *MUSEEC: A multilingual text summarization tool*. Association for Computational Linguistics - 2016 System Demonstrations, Berlin, Germany.

19. Sherry, A., and Bhatia, P. (2015). *Multilingual text summarization with UNL*. International Conference on Advances in Computer Engineering and Applications, United States.

20. Steinberger, J. and Ježek, K. (2009). *Update summarization based on latent semantic analysis*. International Conference on Text, Speech and Dialogue, Ljubljana, Slovenia.

21. Jing, H. and McKeown, K. R. (2000). *Cut and paste based text summarization*. North American chapter of the Association for Computational Linguistics, Seattle, Washington.

22. Moratanch, N. and Chitrakala, S. (2016). *A survey on abstractive text summarization.* IEEE International Conference on Circuit, Power and Computing Technologies, Nagercoil, India.

23. Yang, M., Wang, X., Lu, Y., L., J., Shen, Y. and Li, C. (Jun. 2020). Plausibility-promoting generative adversarial network for abstractive text summarization with multi-task constraint. *Information Sciences*. 521(1), 46-61.

24. Song, S., Huang, H. and Ruan, T. (2019). Abstractive text summarization using LSTM-CNN based deep learning. *Multimedia Tools and Applications*. 78(1), 857-875.

25. Lloret, E. and Palomar, M. (2012). Text summarisation in progress: a literature review. *Artificial Intelligence Review*. 37(1), 1-41.

26. Kabadjov, M., Poesio, M. and Steinberger, J. (2005). *Task-based evaluation of anaphora resolution: the case of summarization*. Recent Advances in Natural Language Processing 05 Workshop on Crossing Barriers in Text Summarization Research, Borovets, Bulgaria.

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

28. Kumar Meena, Y., Gopalani, D., Meena, Y. K. Y. K., Gopalani, D., Kumar Meena, Y. and Gopalani, D. (2015). Evolutionary algorithms for extractive automatic text summarization. *Procedia Computer Science*. 48(3), 244-249.

29. Ferreira, R. *et al.* (2013). Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications*. 40(14), 5755-5764.

30. Oliveira, H. *et al.* (2016). Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. *Expert Systems with Applications*. 65(1), 68-86.

31. Suanmali, L., Salim, N. and Binwahlan, M. S. (2009). Fuzzy logic based method for improving text summarization. *arXiv:0906.4690*. 2(1), 1-6.

32. Fattah, M. A. and Ren, F. (2009). GA, MR, FFNN, PNN and GMM based models for automatic text summarization. *Computer Speech and Language*. 23(1), 126-144.

33. Goularte, F. B., Nassar, S. M., Fileto, R. and Saggion, H. (2019). A text summarization method based on fuzzy rules and applicable to automated assessment. *Expert Systems with Applications*. 115(1), 264-275.

34. Nallapati, R., Zhai, F. and Zhou, B. (2017). *Summarunner: A recurrent neural network based sequence model for extractive summarization of documents*. Thirty-First Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence, San Francisco, California, USA.

35. Davis, J. (2006). *Effective Academic Writing 3*. (Third edition). New York: Oxford University Press, 87-92.

36. Orrú, T., Rosa, J. L. G. and De Andrade Netto, M. L. (2006). SABio: An automatic portuguese text summarizer through artificial neural networks in a more biologically plausible model. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 3960(1), 11-20.

37. Wang, W. M., Li, Z., Wang, J. W. and Zheng, Z. H. (2017). How far we can go with extractive text summarization? Heuristic methods to obtain near upper bounds. *Expert Systems with Applications*. 90(1), 439-463.

38. Meena, Y. K. Y. K. and Gopalani, D. (2014). *Analysis of sentence scoring methods for extractive automatic text summarization*. 2014 International Conference on Information and Communication Technology for Competitive Strategies, New York, USA.

39. Mahajani A., Pandya V., Maria I., Sharma D. (2019). A comprehensive survey on extractive and abstractive techniques for text summarization. In YC. Hu, S. Tiwari, K. Mishra, M. Trivedi (Eds.), *Ambient Communications and Computer Systems Advances in Intelligent Systems and Computing*. Springer, pp. 339-351.

40. Ucan, A. and Akcapinar Sezer, E. (2019). *A new approach on emotion analogy by using word embeddings*. Signal Processing and Communications Applications Conference, Sivas, Türkiye.

41. Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). *Efficient estimation of word representations in vector space*. International Conference on Learning Representations, International Conference on Learning Representations 2013 - Workshop Track, Scottsdale, USA.

42. Harris, Z. S. (1954). Distributional structure. *WORD*. 10(3), 146-162.

43. Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2006). Distributed representations of words and phrases and their compositionality. *Neural Information Processing Systems*. 1(1), 1-9.

44. Li Y., Yang T. (2018). Word embedding for understanding natural language: a survey. In S. Srinivasan (Eds.), *Guide to Big Data Applications Studies in Big Data*. Springer, pp. 83-104.

45. Pennington, J., Socher, R. and Manning, C. D. (2014). *GloVe: Global vectors for word representation*. Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar.

46. Rada Mihalcea and Paul Tarau, Mihalcea, R., Tarau, P. and Rada Mihalcea and Paul Tarau. (2004). *Textrank: Bringing order into text*. 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain.

47. Page, L., Brin, S., Motwani, R. and Winograd, T. (1999). The PageRank citation ranking: bringing order to the web. *World Wide Web Internet And Web Information Systems*. 54(66), 1-17.

48. Barrios, F., López, F., Argerich, L. and Wachenchauzer, R. (2016). Variations of the similarity function of textrank for automated summarization. *arXiv:1602.03606*. 1(1), 1-8.

49. Kruengkrai, C. and Jaruskulchai, C. (2003). *Generic text summarization using local and global properties of sentences*. IEEE International Conference on Web Intelligence, Halifax, NS, Canada.

50. Mitra, M., Singhal, A. and Buckley, C. (1997). *Automatic text summarization by paragraph extraction*. Association for Computational Linguistics Workshop on Intelligent and Scalable Text Summarization, Madrid, Spain.

51. Fang, C., Mu, D., Deng, Z. and Wu, Z. (2017). Word-sentence co-ranking for automatic extractive text summarization. *Expert Systems With Applications*. 72(1), 189-195.

52. Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*. 104(2), 211-240.

53. Baxendale, P. B. (1958). Machine-made index for technical literature-an experiment. *IBM Journal of Research and Development*. 2(1), 354-361.

54. Binwahlan, M. S., Salim, N. and Suanmali, L. (2009). Fuzzy swarm based text summarization. *Journal of Computer Science*. 5(5), 338-346.

55. Babar, S. A. A. and Thorat, S. A. A. (2014). Improving text summarization using fuzzy logic & latent semantic analysis. *International Journal of Innovative Research in Advanced Engineering*. 1(4), 170-177.

56. Babar, S. A. A. and Patil, P. D. (2015). Improving performance of text summarization.

Procedia Computer Science. 46(1), 354-363.

57. Gong, Y. and Liu, X. (2001). *Generic text summarization using relevance measure and latent semantic analysis*. Special Interest Group on Information Retrieval (SIGIR) Conference on Research and development in information retrieval, New Orleans, Louisiana, USA.

58. Chuang, W. and Yang, J. (2000). *Extracting sentence segments for text summarization: a machine learning approach*. Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval, Athens, Greece.

59. Moradi, M. and Ghadiri, N. (2016). A bayesian approach to biomedical text summarization. *arXiv*:*1605.02948*. 1(10), 1-50.

60. Shah C., Jivani A. (2019). An Automatic Text summarization on naive bayes classifier using latent semantic analysis. In R. Shukla, J. Agrawal, S. Sharma, G. S. Tomer (Eds.), *Data, Engineering and Applications*. Springer, pp. 171-180.

61. Najibullah, A. (2015). *Indonesian text summarization based on naïve bayes method*. International Seminar and Conference 2015: The Golden Triangle, Indonesia-India-Tiongkok.

62. Kumbhar, S., Bordia, A., Kapse, S. and Paatil, F. (2018). Comparison of extractive text summarization methods. *International Journal of Advances in Electronics and Computer Science*. 5(7), 166-171.

63. Galanis, D., Lampouras, G. and Androutsopoulos, I. (2012). *Extractive multi-document summarization with integer linear programming and support vector regression*. 24th International Conference on Computational Linguistics, Mumbai, India.

64. İnternet: Li, S., Ouyang, Y., Wang, W. and Sun, B. (2007). Multi-document summarization using support vector regression. URL:http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.137.6243, Son Erişim Tarihi: 16.01.2020.

65. Begum, N., Fattah, M. A. and Ren, F. (2009). Automatic text summarization using support vector machine. *International Journal of Innovative Computing, Information and Control*. 5(7), 1987-1996.

66. Neto, Joel Larocca, Alex A. Freitas, and C. A. K., Neto, Joel Larocca, Alex A. Freitas, and Kaestner, C. A. A. (2002). Automatic text summarization using a machine learning approach. *Advances in Artificial Intelligence*. 1(1), 205-215.

67. Osborne, M. (2002). *Using maximum entropy for sentence extraction*. Workshop on Automatic Summarization, Phildadelphia, Pennsylvania, USA.

68. Hannah M. E. (2020). A classification-based summarization model using supervised learning. In M. Pant, T. Sharma, S. Basterrech, C. Banerjee (Eds.), *Computational Network Application Tools for Performance Management. Asset Analytics (Performance and Safety Management)*. Springer, pp. 87-93.

69. Jayashree, R., Srikanta Murthy, K. and Anami, B. S. (2014). *An artificial neural*

network approach to text document summarization in the Kannada language. 13th International Conference on Hybrid Intelligent Systems, USA.

70. Lamsiyah, S., El Mahdaouy, A., El Alaoui, S. O. and Espinasse, B. (2019). *A supervised method for extractive single document summarization based on sentence embeddings and neural networks*. International Conference on Advanced Intelligent Systems, Cairo, Egypt.

71. Ren, P. *et al.* (2018). Sentence relations for extractive summarization with deep neural networks. *Association for Computing Machinery (ACM) Transactions on Information Systems (TOIS)*. 36(4), 1-32.

72. Cheng, J. and Lapata, M. (2016). *Neural summarization by extracting sentences and words*. Annual Meeting of the Association for Computational Linguistics, Berlin, Germany.

73. Nallapati, R., Zhou, B. and Ma, M. (2016). Classify or select: neural architectures for extractive document summarization. *arXiv:1611.04244*. 1(1), 1-13.

74. Narayan, S., Cohen, S. B. and Lapata, M. (2018). *Ranking sentences for extractive summarization with reinforcement learning*. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, USA.

75. Ayhan, S. and Erdoğmuş, Ş. (2014). Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi. *Eskişehir Osmangazi Üniversitesi İİBF Dergisi*. 9(1), 175-198.

76. Mourya, A. K., ShafqatUlAhsaan, and Kaur, H. (2020). Performance and evaluation of different kernels in support vector machine for text mining. *Lecture Notes in Networks and Systems*. 109(2020), 264-271.

77. Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*. 52(1), 273-292.

78. Sun, A., Lim, E. P. and Liu, Y. (2009). On strategies for imbalanced text classification using SVM: a comparative study. *Decision Support Systems*. 48(1), 191-201.

79. Fan, B., Fan, W., Smith, C. and Garner, H. S. (2020). Adverse drug event detection and extraction from open data: a deep learning approach. *Information Processing and Management*. 57(1), 1-14.

80. Kastrati, Z., Imran, A. S. and Yayilgan, S. Y. (2019). The impact of deep learning on document classification using semantically rich representations. *Information Processing and Management*. 56(5), 1618-1632.

81. Aydogan, E. and Akcayol, M. A. (2016). *A comprehensive survey for sentiment analysis tasks using machine learning techniques*. 2016 International Symposium on INnovations in Intelligent SysTems and Applications (INISTA), Sinaia, Romania.

82. Turhan, C. G. and Bilge, H. Ş. (2020). Scalable image generation and super resolution using generative adversarial networks. *Journal of the Faculty of Engineering and*

Architecture of Gazi University. 35(2), 953-966.

83. Guzel, M., Kok, I., Akay, D. and Ozdemir, S. (2019). ANFIS and deep Learning based missing sensor data prediction in IoT. *Concurrency Computation*. 32(2), 1-15.

84. Bilge, Y. C., Ikizler-Cinbis, N. and Cinbis, R. G. (2019). *Zero-shot sign language recognition: can textual data uncover sign languages?* British Machine Vision Conference (BMVC), Cardiff, UK.

85. Cao, Z., Wei, F., Li, S., Li, W., Zhou, M. and Wang, H. (2015). *Learning summary prior representation for extractive summarization*. 53rd Annual Meeting of the Association for Computational Linguistics, Beijing, China.

86. Wang, L. X. and Mendel, J. M. (1992). Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man, and Cybernetics*. 22(6), 1414-1427.

87. Koupaee, M. and Wang, W. Y. (2018). Wikihow: A large scale text summarization dataset. *Interinstitutional Center for Computational Linguistics (NILC) Technical Report*. 1(2018), 1-5.

88. İnternet: Pardo, T. A. S. and Rino, L. H. M. TeMario: a corpus for automatic text summarization. URL: https://www.semanticscholar.org/paper/TeMario%3A-a-corpus-for-automatic-text-summarization-Pardo-Rino/bb687d5aaa92b9a552dfe698c46e6b75317d45f4, Son Erişim Tarihi: 05.06.2020.

89. Pottker, H. (2003). News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies*. 4(4), 501-511.

90. Fisher, S. and Roark, B. (2006). *Query-focused summarization by supervised sentence ranking and skewed word distributions*. Document Understanding Conferences, New York, USA.

91. Li, J., Sun, L., Kit, C. and Webster, J. (2007). *A query-focused multi-document summarizer based on lexical chains*. Document Understanding Conferences, New York, USA.

92. Hachey, B., Murray, G. and Reitter, D. (2005). The embra system at duc 2005: query-oriented multi-document summarization with a very large latent semantic space. Document Understanding Conferences, New York, USA.

93. See, A., Liu, P. J. and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *Association for Computational Linguistics*. 1(1), 1-20.

94. Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L, Kay, W., Suleyman, M. and Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems*. 1(1), 1-14.

95. Chowdhury, T., Kumar, S. and Chakraborty, T. (2020). *Neural abstractive summarization with structural attention*. International Conference on Artificial Intelligence, Berkeley.

96. Zhang, R. and Tetreault, J. (2019). *This email could save your life: introducing the task*

of email subject line generation. 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy.

97. Kupiec, J. (1995). *A trainable document summarizer*. International Association of Computational Machinery (ACM) Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval, Seattle, USA.

98. Sharma, E., Li, C. and Wang, L. (2019). *BIGPATENT: A large-scale dataset for abstractive and coherent summarization*. 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy.

99. Steinberger, J. and Ježek, K. (2012). Evaluation measures for text summarization. *Computing and Informatics*. 28(2), 251-275.

100. Ramanujam, N. and Kaliappan, M. (2016). An automatic multidocument text summarization approach based on naïve bayesian classifier using timestamp strategy. *The Scientific World Journal*. 1(2016), 1-10.

101. Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T. and Sundheim, B. (1999). *The TIPSTER SUMMAC text summarization evaluation*. European Chapter of the Association for Computational Linguistics, Valencia, Spain.

102. Hachey, B. and Grover, C. (2004). *Sentence classification experiments for legal text summarisation*. Annual Conference of Legal Knowledge and Information Systems, Berlin, Germany.

103. Fuentes, M., Alfonseca, E. and Rodríguez, H. (2007). *Support vector machines for query-focused summarization trained and evaluated on pyramid data*. ACL 2007, 45th Annual Meeting of the Association for Computational Linguistics, Prague.

104. Shen, D., Sun, J. T., Li, H., Yang, Q. and Chen, Z. (2007). *Document summarization using conditional random fields*. International Joint Conference on Artificial Intelligence (IJCAI), Australia.

105. Wu, X. and Zong, C. (2010). *An map based sentence ranking approach to automatic summarization*. International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China.

106. Ouyang, Y., Li, W., Li, S. and Lu, Q. (Mar. 2011). Applying regression models to query-focused multi-document summarization. *Information Processing & Management*. 47(2), 227-237.

107. Suanmali, L., Binwahlan, M. S. and Salim, N. (2009). *Sentence features fusion for text summarization using fuzzy logic*. International Conference on Hybrid Intelligent Systems, Shenyang, China.

108. Kumar, Y. J., Salim, N., Abuobieda, A. and Albaham, A. T. (2014). Multi document summarization based on news components using fuzzy cross-document relations. *Applied Soft Computing*. 21(1), 265-279.

109. Hannah, M. E., Geetha, T. V. and Mukherjee, S. (2011). Automatic extractive text

summarization based on fuzzy logic: a sentence oriented approach. *Lecture Notes in Computer Science*. 1(1), 530-538.

110. Yousefi-Azar, M. and Hamey, L. (2017). Text summarization using unsupervised deep learning. *Expert Systems with Applications*. 68(10), 93-105.

111. Sinha, A., Yadav, A. and Gahlot, A. (2018). Extractive text summarization using neural networks. *arXiv:1802.10137*. 1(1), 1-5.

112. Cao, Z., Li, W., Li, S., Wei, F. and Li, Y. (2016). *AttSum: joint learning of focusing and summarization with neural attention*. International Conference on Computational Linguistics: Technical Papers, Osaka, Japan.

113. Zhang, Y., Er, M. J. and Zhao, R. (2015). *Multi-document extractive summarization using window-based sentence representation*. IEEE Symposium Series on Computational Intelligence, Cape Town, South Africa.

114. Yin, W. and Pei, Y. (2015). *Optimizing sentence modeling and selection for document summarization*. International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina.

115. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*. 12(1), 2493-2537.

116. Jain, A., Bhatia, D. and Thakur, M. K. (2017). *Extractive text summarization using word vector embedding*. 2017 International Conference on Machine Learning and Data Science (MLDS), Noida, India.

117. Kiros, R. *et al.* (2015). *Skip-thought vectors*. Advances in Neural Information Processing Systems, Montreal, Quebec, Canada.

118. Wu, Q., Iyer, N. and Wang, H. (2018). *Learning contextual bandits in a non-stationary environment*. 41st International Association of Computational Machinery (ACM) Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval, USA.

119. Glorot, X. and Bengio, Y. (2010). *Understanding the difficulty of training deep feedforward neural networks*. Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy.

120. Kingma, D. P. and Ba, J. (2015). *ADAM: a method for stochastic optimization.* International Conference for Learning Representations, San Diego, USA.

121. Liu, D. C., Nocedal, J., Dong, C. L. (Northwestern U. and Jorge, N. (Northwestern U. (189AD). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*. 45(1–3), 503-528.

122. Pedregosa, F. *et al.* (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research*. 12(1), 2825-2830.

EKLER

EK-1. Knowledge-Based Systems

Knowledge-Based Systems 183 (2019) 104848

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Multi-document extractive text summarization: A comparative assessment on features☆

Begum Mutlu [a], Ebru A. Sezer [b,*], M. Ali Akcayol [a]

[a] Gazi University, Department of Computer Engineering, Ankara, 06570, Turkey
[b] Hacettepe University, Department of Computer Engineering, Ankara, 06800, Turkey

ARTICLE INFO

Article history:
Received 27 February 2019
Received in revised form 5 July 2019
Accepted 13 July 2019
Available online 16 July 2019

Keywords:
Multi-document text summarization
Feature space
Sentence scoring-extraction
Artificial neural network
Fuzzy inference system

ABSTRACT

Text summarization is the process of generating a brief version of a text that preserves the salient information of the text. For information retrieval, it is a good dimension reduction solution. In addition, it reduces the required reading time. This study focused on extracting informative summaries from multiple documents using commonly used hand-crafted features from the literature. The first investigation focused on the generation of a feature vector. The features were the number of sentences, term frequency, similarity with the title, term frequency-inverse sentence frequency, sentence position, sentence length, sentence–sentence similarity, bushy-path results, phrases of the sentence, proper nouns, n-gram co-occurrence, and length of the document. Secondly, several combinations of these features were examined and a shallow multi-layer perceptron and two differently modeled fuzzy inference systems were used to extract salient sentences from texts in the Document Understanding Conference (DUC) dataset. The summarization performances of these models were evaluated using original classification performance metrics, and recall-oriented understudy for gisting evaluation (ROUGE)-n. This study recommended the use of fuzzy systems based on a feature vector and a fuzzy rule set for extractive text summarization. The extraction methods were evaluated against a changing compression ratio. Results of experiments showed that the implemented neural model tended to incorrectly infer sentences that were not considered salient by human annotators. However, for distinguishing between summary-worthy and summary-unworthy sentences, the fuzzy inference systems performed better than the utilized neural network, as well as better than the existing fuzzy inference-based text summarization approaches in the literature.

1. Introduction

Today, a tremendous amount of data is available on the Internet. With the proliferation of the Internet, it has become increasingly difficult to efficiently reach information from such massive amounts of data. This issue is not only related to the volume of the data, but also to the diversity of subjects, qualities, and idioms [1,2].

It is a complex and exhaustive process to gather and perceive any main information from such a massive amount of resources in sufficient time for humans. Fortunately, these processes can be automatically performed by information retrieval methods. However, the rising quantity of information causes some performance issues, such as insufficient solutions and unwieldy applications of information retrieval tasks. The use of higher-technology machines may reduce the losses caused by these issues. However, they may cost more. As a more applicable alternative, dimension reduction can be employed in raw data to handle these issues and to accelerate the implementations of these tasks. As regards the domain of text processing, automatic text summarization is a good practice for dimension reduction.

Text summarization is the process of generating a brief version of a single document or document set, so that the resulting summaries preserve the fundamental information of the original document(s). This study addresses two motivations regarding automatic summary generation: *summarization for machine*, and *summarization for human*. First, insofar as information retrieval tasks, a reduction approach is satisfactory if it can be applied efficiently. This addresses *summarization for machine*, and more specifically, the use of summarization as a dimension reduction for information retrieval purposes. Second, in *summarization for human*, the reading becomes easier, and people can obtain the salient information of the text more quickly. For example, instead of reading through an entire document, people can understand the text more quickly, and in an easier way.

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2019.07.019.

* Corresponding author.
E-mail address: ebruakcapinarsezer@gmail.com (E.A. Sezer).

https://doi.org/10.1016/j.knosys.2019.07.019
0950-7051/

EK-2. IEEE World Congress on Computational Intelligence (WCCI) 2018

End-to-End Hierarchical Fuzzy Inference Solution

Begum Mutlu
Department of Computer Engineering
Gazi University
Ankara, Turkey 06570
Email: begummutlu@gazi.edu.tr

Ebru A. Sezer
Department of Computer Engineering
Hacettepe University
Ankara, Turkey 06800
Email: ebruakcapinarsezer@gmail.com

M. Ali Akcayol
Department of Computer Engineering
Gazi University
Ankara, Turkey 06570
Email: akcayol@gazi.edu.tr

***Abstract*—Hierarchical Fuzzy System (HFS) is a popular approach for handling curse of dimensionality problem occurred in complex fuzzy rule-based systems with various and numerous inputs. However, the processes of modeling and reasoning of HFS have some critical issues to be considered. In this study, the effect of these issues on the accuracy and stability of the resulting system has been investigated, and an end-to-end HFS framework has been proposed. The proposed framework has three main steps such as single system modeling, rule partitioning and HFS reasoning. It is fully automated, generic, almost independent from data, and applicable for any kind of inference problem. In addition, the proposed framework preserves accuracy and stability during the HFS reasoning. These judgments have been ensured by a number of experimental studies on several datasets about software faulty prediction (SFP) problem with a large feature space. The main contributions of this paper are as follows: *(i)* it provides the entire HFS implementation from problem definition to calculation of final output, *(ii)* it increases the accuracy of recently proposed rule generation scheme in the literature, *(iii)* it presents the only possible fuzzy system solution for SFP problem containing a large feature space with reasonable accuracy.**

I. Introduction

Fuzzy rule-based systems suffer from *curse of dimensionality* when the number of input parameters and/or fuzzy sets are increased. This problem is both experienced in the system modeling and reasoning processes. Regarding the former, linear increase in the number of inputs causes an exponential growth in fuzzy rule base. It also complicates the production of each individual fuzzy rule for human expert, since it becomes difficult to consider the impact of input on the output. In addition, reasoning process requires much more computational time and memory as the system becomes more complex.

As an effective solution to *curse of dimensionality*, Hierarchical Fuzzy Systems (HFSs) have been proposed in [1]. In HFSs, the complex, high dimensional fuzzy system is decomposed into lower dimensional subsystems each of which covers a small portion of corresponding problem. The outputs provided from these subsystems are then combined in an hierarchical manner in order to reach the final decision. This strategy has been successfully implemented in several applications from numerous research fields [2]–[10]. Besides its applications, HFSs have been also investigated in terms of HFS structures, HFS modeling and HFS accuracy.

How to combine the subsystems, and how to fed the subsystems with inputs determines the HFS structure. In order to reach less costly and most accurate HFS, several HFS structures are investigated as aggregated (parallel), incremental (serial) and hybrid (a customized combination of aggregated and incremental structures) [11]–[13]. Here, accuracy and stability are very important because any change in the HFS structure cause the system output to differ from the preceding HFS's output and surely the output provided from the single system. At the first glance, the reason behind this differentiation is based on rule base modeling. If the single system rules cannot be partitioned accurately to the layers of HFS, the HFS behaviors differ from the original one. This issue is also encountered in such cases where the system structure is transformed into another one. In order to satisfy accuracy and stability, the HFS rules should be equivalent to single system rule base, and it should be adaptive to changing HFS structures [14]. In theory, this strategy eliminates the *curse of dimensionality* of single system by preserving its accuracy. In practice, it is a very challenging procedure for human expert. Because just like the difficulties in modeling single system, the input-output linguistic variables, the relations between them, and interrelation between subsystems in the inner layers should be also taken into consideration in HFSs' rule base modeling. Therefore, some automatic rule partitioning methods are proposed to convert a complex single system to equivalent HFS [13], [15], [16]. Nevertheless, inaccuracy and instability remain. In other words, conversion of single system rules into equivalent HFS rules cannot overcome this issue alone. There is another reason which triggers these problems, misleading data transmission between layer [11]. In Mamdani style hierarchical fuzzy inference process, redundantly repeated defuzzification-fuzzification steps in the inner layers manipulate the fuzziness level of transferred data. This is an essential problem since data proceed the HFS layers as losing its own features [12], [17]. It is a strict requirement of a complete implementation of HFS that, all of these issues should be considered and entirely solved in both modeling steps and decision making process.

HFSs' modeling is directly related to the mathematical representation of input/output parameters, and generation of rule base. It addresses the problems which are how to produce single system (specification of membership functions (MF) and generation of rule base), and how to partition and share the resulting rule base between HFS subsystems. On the other hand, accurate data transmission between layers of HFS

978-1-5090-6020-7/18/$31.00 ©2018 IEEE

EK-3. International Conference on Semantic Systems (SEMANTiCS)

Investigation of Text Summarization Features by Fuzzy Systems and Convolutional Neural Networks

Begum Mutlu[1][0000−0003−1960−2143], Ebru A. Sezer[2][0000−0002−9287−2679], and M. Ali Akcayol[1][0000−0002−6615−1237]

[1] Gazi University, Ankara 06570, Turkey
begummutlu@gazi.edu.tr
[2] Hacettepe University, Ankara 06570, Turkey

Abstract. In extractive text summarization, the text units of source documents (the sentences) are scored by their salience, and the most important text units are included in the automatic summary. The features obtained from the corresponding text unit are used to define the importance. Regarding these features, the studies have been focused on two approaches: using syntactic/hand-crafted features and using semantic features by word embeddings. However, the most distinctive feature space is still unknown. To that end, this study investigates the extractive text summarization based on the feature space used to score the sentences. An ample syntactic feature space was generated by combining multiple metrics for several summarization features, and the contribution of this feature space on extractive text summarization was compared with the state-of-the-art use of semantic features based on word embeddings. This comparison was performed by collaborating 3 different sentence scoring approaches: a fuzzy rule-based system, a shallow multi-layer perceptron, and a convolutional neural network. The most salient sentences obtained from these approaches were included in the model summaries basing syntactic and semantic features, respectively, and the resulting summaries were evaluated by both recall-oriented understudy for gisting evaluation metrics and by standard performance evaluation methods for sentence classification. The obtained results showed that the proposed syntactic features are more capable of determining the summary-worthy sentences, and the text summarization problem is mostly related to these syntactic features instead of the semantic relations of sentences.

Keywords: Extractive text summarization, feature space for summarization, fuzzy rule-based systems, convolutional neural network

1 Introduction

Text summarization is a dimensionality reduction practice for text documents, that is also interpretable for human. Because the conventional dimensionality reduction methods focus on reducing the dimensions of corresponding data by

EK-4. Information Processing & Management

Candidate Sentence Selection for Extractive Text Summarization

Begum Mutlu

Department of Computer Engineering, Gazi University, Ankara, 06570, Turkey

Ebru A. Sezer*

Department of Computer Engineering, Hacettepe University, Ankara, 06800, Turkey

M. Ali Akcayol

Department of Computer Engineering, Gazi University, Ankara, 06570, Turkey

Abstract

Text summarization is a process of generating a brief version of documents by preserving the fundamental information of documents as much as possible. Although the majority of the text summarization researches have been focused on supervised learning solutions, there are a few datasets indeed generated for summarization task, and most of the existing summarization datasets do not have human-generated goal summaries, which is vital for both summary generation and evaluation. Therefore, a new dataset was presented for abstractive and extractive summarization tasks in this study. This dataset contains academic publications, the abstracts written by the authors, and extracts in two sizes, which were generated by human readers in this research. Then, the resulting extracts were evaluated to ensure the validity of the human extract production process. Moreover, the extractive summarization problem was reinvestigated on the proposed summarization dataset. Here the main point taken into account was to analyze the feature vector to generate more informative summaries. To that end, a comprehensive syntactic feature space was generated for the pro-

*This is to indicate the corresponding author.
Email address: ebruakcapinarsezer@gmail.com (Ebru A. Sezer*)

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : MUTLU, Begüm
Uyruğu : T.C.
Doğum tarihi ve yeri : 16.01.1990, Mersin
Medeni hali : Bekâr
Telefon : 0(312) 582 31 30
Faks : 0(312) 230 65 103
e-mail : begummutlu@gazi.edu.tr

Eğitim

Derece	Eğitim Birimi	Mezuniyet Tarihi
Doktora	Gazi Üniversitesi / Bilgisayar Mühendisliği	Devam Ediyor
Yüksek lisans	Hacettepe Üniversitesi / Bilgisayar Mühendisliği	2014
Lisans	Gazi Üniversitesi / Bilgisayar Mühendisliği	2012
Lise	Seyhan Rotary Anadolu Lisesi	2008

İş Deneyimi

Yıl	Yer	Görev
2014-Halen	Gazi Üniversitesi	Araştırma Görevlisi
2013-2014	Netcad Yazılım A.Ş.	ARGE Uzmanı

Yabancı Dil

İngilizce

Yayınlar

1. Mutlu, B., Sezer, E. A. and Akcayol, M. A. (2019). Multi-document extractive text summarization: a comparative assessment on features. Knowledge-Based Systems. 183(1), 1–13.

2. Mutlu, B., Haciömeroğlu, M., Guzel, M. S., Dikmen, M. and Sever, H. (2014). Silhouette extraction from street view images. *International Journal of Advanced Robotic Systems*. 11(1), 1-12.

3. Mutlu, B., Sezer, E. A. and Nefeslioglu, H. A. (2017). A defuzzification-free hierarchical fuzzy system (DF-HFS): Rock mass rating prediction. *Fuzzy Sets and Systems*. 307(1), 50-66.

4. Akcayol, M. A., Utku, A., Aydoğan, E. and Mutlu, B. (2018). A weighted multi-attribute based recommender system using extended user behavior analysis. *Electronic Commerce Research and Applications*. 28(1), 86-93.

5. Mutlu, B., Nefeslioglu, H. A., Sezer, E. A., Ali Akcayol, M. and Gokceoglu, C. (2019). An experimental research on the use of recurrent neural networks in landslide susceptibility mapping. *ISPRS International Journal of Geo-Information*, 8(12), 1-21.

6. Ozer, B. C., Mutlu, B., Nefeslioglu, H. A., Sezer, E. A., Rouai, M., Dekayir, A. and Gokceoglu, C. (2020). On the use of hierarchical fuzzy inference systems (HFIS) in expert-based landslide susceptibility mapping: the central part of the Rif Mountains (Morocco). *Bulletin of Engineering Geology and the Environment*. 79(1), 551-568.

Hobiler

Yoga, Pilates, Pastacılık

GAZİ GELECEKTİR...