
GAZİ ÜNİVERSİTESİ
Gazi
1926

DİZİ ETİKETLEME TEMELLİ YENİ BİR KARMA ANAHTAR KELİME ÇIKARIM MODELİ

Hüma KILIÇ

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ŞUBAT 2023

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.

Hüma KILIÇ
23/02/2023

DİZİ ETİKETLEME TEMELLİ YENİ BİR KARMA ANAHTAR KELİME ÇIKARIM MODELİ

(Doktora Tezi)

Hüma KILIÇ

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Şubat 2023

ÖZET

Anahtar kelime çıkarımı, metin içeriğinin kümelenmesi ve bağlanmasındaki büyük zorluklardan biridir. Literatürde, anahtar kelime ve anahtar ifade çıkarımı için çeşitli makine öğrenmesi yaklaşımları önerilmiştir. Bu tezde ilk olarak literatürde önerilen modeller ve performans sonuçları iki ana başlık altında sunulmuştur. Ancak, anahtar kelime çıkarımı modellerinin performans sonuçları hala beklentilerin altındadır. Bu tez kapsamında, yeni bir hibrit anahtar kelime çıkarma modeli olan HibritAKÇ önerilmiştir. Önerilen yöntem, anahtar kelime çıkarım problemini bir dizi etiketleme görevi olarak ele almaktadır. Naive Bayes, Destek Vektör Makinesi, Çok Katmanlı Algılayıcı ve Rastgele Orman sınıflandırma algoritmaları, modelin Token Sınıflandırma modülünde ayrı ayrı eğitilmiştir. Modelde metin, grafik, gömme ve küme öznitelikleri kullanılarak Token Sınıflandırma işlemi gerçekleştirilmiştir. Modelin performansı literatürde yaygın olarak kullanılan Inspec, Semeval-2017, 500N-KPCrowd veri kümeleri ve yeni derlenen TRDizinEn ve DergiParkEn veri kümeleri kullanılarak değerlendirilmiştir. Model, tüm veri kümeleri için ortalama 0,664 F1 skoruna ulaşmıştır. En yüksek F1-skor (0,74) TRDizinEn veri seti ile elde edilmiştir.

Bilim Kodu : 92432
Anahtar Kelimeler : Anahtar kelime çıkarımı, hibrit yöntem, dizi etiketleme
Sayfa Adedi : 82
Danışman : Prof. Dr. Aydın ÇETİN

A NOVEL SEQUENTIAL LABELING BASED HYBRID KEYWORD EXTRACTION MODEL

(Ph. D. Thesis)

Hüma KILIÇ

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

February 2023

ABSTRACT

Keyword extraction is one of the main problems in clustering and linking textual content. In literature, several machine learning approaches were proposed for keyword and keyphrase extraction. In this thesis, firstly, the models proposed in the literature and their performance results are presented under two main headings. However, the state-of-the-art performance results are still below the expectations. We propose a novel hybrid keyword extraction model, HybridKEM. The proposed method addresses the keyword extraction problem as a sequence labelling task. Naive Bayes, Support Vector Machine, Multi-Layer Perceptron and Random Forest classification algorithms were trained separately in the Token Classification module of the model. The Token Classification process was performed by using text, graphic, embedding, and set features in the model. The performance of the model was evaluated using the Inspec, Semeval-2017, 500N-KPCrowd datasets, which are widely used in studies in the literature, and two newly collected, TRDizinEn and DergiParkEn datasets. The model achieved an average F1-score of 0.664 for all datasets. The highest F1-score (0.74) was obtained with the TRDizinEn dataset.

Bilim Kodu	:	92432
Anahtar Kelimeler	:	Keyword extraction, hybrid method, sequence labeling.
Sayfa Adedi	:	82
Danışman	:	Prof. Dr. Aydın ÇETİN

TEŞEKKÜR

Doktora eğitimim ve Tez çalışmalarım boyunca her koşulda desteklerini esirgemeyen danışmanım Prof. Dr. Aydın ÇETİN'e sonsuz teşekkürlerimi sunarım. Tez izleme sürecinde tezin gelişimine sağladıkları katkılardan dolayı değerli Prof. Dr. Necaattin BARIŞÇI ve Prof. Dr. Ömer DEPERLİOĞLU'na, jüri üyeleri Prof.Dr. Nurettin DOĞAN ve Doç. Dr. Hüseyin POLAT'a teşekkür ederim. Doktora çalışmalarım süresince her zaman yanımda olan, varlıkları şükür sebebim yavrularım Meryem ve Hüseyin'e teşekkürü borç bilirim.

İÇİNDEKİLER

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış kısaltmalar ve açıklamaları ile birlikte aşağıda sunulmuştur.

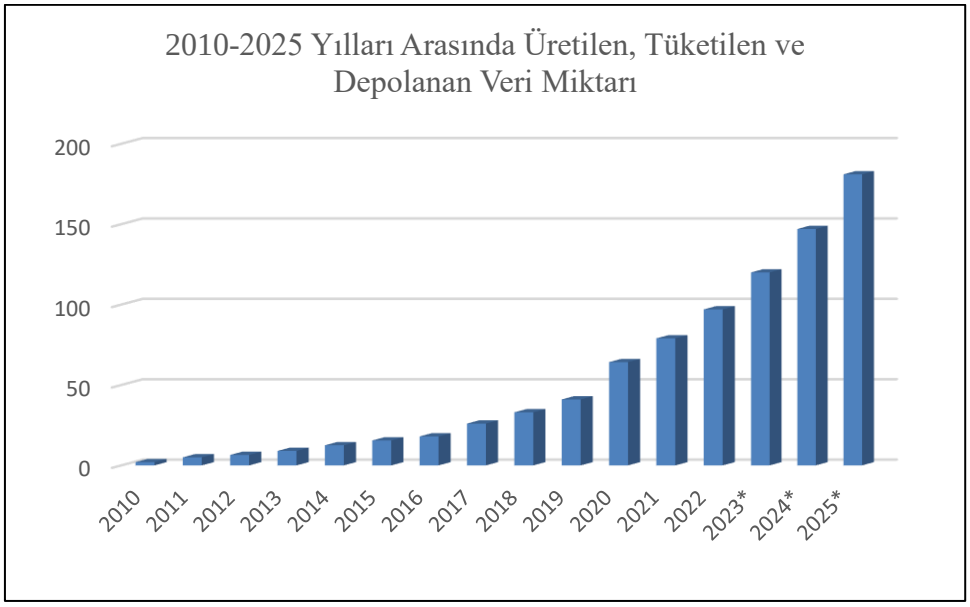
Kısaltmalar	Açıklamalar
500N-KPCrowd	500N-KeyPhrasesCrowdAnnotated-Corpus
AGA	Aşırı Gradyan Arttırma
AİK	Anahtar İfade Kümesi
AİÜ	Anahtar İfade Üretimi
AraMer	Arasındalık Merkeziliği
BBA	Bağımsız Bileşen Analizi
BilDÇYKG	Bilimsel DÇYKG
BSS	Başlıklı Sayfa Sıralama
CS	Cümle Sıklığı
CümVek	Cümleden Vektöre
ÇKA	Çok Katmanlı Algılayıcılar
ÇMİ-TB	Çok Merkezili İndis – Temel Bileşen
ÇY-UKSB	Çift Yönlü Uzun Kısa Süreli Bellek
DAKÇ	Denetimsiz Anahtar Kelime Çıkarımı
DAKO-ÖHKÇ	Derin Anahtar Kelime Oluşturma için Özel Hiyerarşik Kod Çözme
DD-AKÇ	Doküman Düzeyinde Anahtar Kelime Çıkarma
DDİ	Doğal Dil İşleme
DerMer	Derece Merkeziliği
DışMer	Dış Merkezlilik
DMKG	Dil Modelinden Kelime Gömmesi
DokDÇYKG	Doküman DÇYKG
DokVek	Dokümandan Vektöre
DT-AKTSE	Dönüştürücü Tabanlı Anahtar Kelime Tanımlama için Sinirsel Etiketleyici
DUC	Document Understanding Conferences
DVM	Destek Vektör Makinası
EA	Ekstra Ağaç

Kısaltmalar	Açıklamalar
ECG	Evrensel Cümle Gömücü
FBIS	Foreign Broadcast Information Service
GA	Genetik Algoritma
G-AİÜ	Görünmeyen Anahtar İfade Üretimi
G-AKÇ	Görünür Anahtar Kelime Çıkarımı
GEA	Grafiksel Evrişimsel Ağ
GloVe	Global Vektörler
GT-AKS	Grafik-Tabanlı Anahtar Kelime Sınıflandırması
GTB	Geçitli Tekrarlayan Birim
HibritAKÇ	Hibrit Anahtar Kelime Çıkarımı
HMT-UKSB	Hedef Merkez-Tabanlı UKSB
HO-AKÇ	Hızlı Otomatik Anahtar Kelime Çıkarımı
İÇ	İfade Çözücü
İTD	İnce Taneli Değerlendirme
İTK	İsim Tamlaması Kümesi
KÇ	Kelime Çözücü
KDD	Knowledge Discovery and Data Mining
KGS	Kelime Grafiği Skoru
KMSG	Kelime Merkeziliği Sabitlenmiş Gösterimi
KÖSA	Korelasyon ÖSA
KP	Konuşmanın Parçası
KRA	Koşullu Rastgele Alanlar
KümKat	Kümeleme Katsayısı
KV	Kelimeden Vektöre
LAA	Lineer Ayrımcı Analizi
LS	Lasso
ÖSA	Tekrarlayan Sinirsel Ağ
ÖzVMer	Özvektör Merkeziliği
SÇ	Sözcük Çantası
SR	Sarmalayıcı
S-SÇ	Sürekli SÇ

Kısaltmalar	Açıklamalar
ST-DAKÇ	Sıralama Tabanlı Denetimsiz Anahtar Kelime Çıkarımı
TBA	Temel Bileşenler Analizi
TDF	Ters Doküman Frekansı
TDGT-AKÇ	Tekil Dokümanlardan Grafik Tabanlı Anahtar Kelime Çıkarımı
TDK	Tekil Değer Kompozisyonu
TF	Terim Frekansı
UCST	University College of Science and Technology
YakMer	Yakınlık Merkeziliği
YapıDel	Yapısal Delik
ZB	Zeta Bayt

1. GİRİŞ

Tüm dünyada üretilen, tüketilen ve depolanan dijital veri miktarı hızla artmaktadır. Şekil 1.1' de yıllara göre artan veri miktarı incelendiğinde (Statista, 2022) 2010 yılında 2 Zetabayt (ZB) veri saklanmışken, 2016 yılında saklanan veri miktarının 18 ZB'a yükseldiği, 2022 yılında ise yaklaşık 100 ZB veri depolandığı görülmektedir. 2025 yılında bu miktarın 180 ZB civarına ulaşması beklenmektedir.



Şekil 1.1. Yıllara göre üretilen, tüketilen ve depolanan veri miktarı

Dijital içeriklerin hızla artmasıyla metinsel veri yığınları arasından aranılan bilgiye ulaşmak çözülmesi gereken bir problem haline gelmiştir. İstenilen bilgiye hızlı ve kolay bir şekilde ulaşabilmek için metinsel içeriklere anahtar kelime atanması gerekmektedir. Bununla birlikte anahtar kelime çıkarımı problemi özetleme, belge bağlantılama, kümeleme vb. sistemleri geliştirebilmek için de çözülmesi gereken temel problemdir.

Anahtar kelime çıkarma işlemi manuel veya otomatik olarak yapılabilir. Manuel anahtar kelime çıkarımı, dijital metin yığını için uzun zaman alır. Bu yöntem maliyet yönünden de uygun değildir. Bu nedenle, Doğal Dil İşleme (DDİ/Natural Language Processing-NLP) alanındaki araştırmacılar, bu süreci otomatikleştirmek için sürekli olarak yöntemler geliştirmeye odaklanmışlardır.

Literatürde anahtar kelime;

- verilen doküman hakkında ana fikri veren kısa metin parçaları (Chen W., Chan, Li, Bing ve King, 2019)
- dokümanı açıklayıcı kelime veya ifadeler (Alzaidy, Caragea ve Giles, 2019),
- okuyucunun ana fikri hızlıca kavramasını sağlayacak yoğun bilgiye sahip kelimeler (Zhang Q., Wang, Gong ve Huang, 2016),
- dokümandaki çekirdek bilgiyi içeren kısa ifadeler (Chen, Chan, Li ve King, 2020),
- uzun metinlerin ana fikirlerini açıklayan kısa, özet ifade veya kelimeler (Meng ve diğerleri, 2017),
- okuyuculara içeriğin yüksek düzeyde açıklamasını sağlayan önemli kelimeler grubu (Gupta ve Lehal, 2009),
- metinin içeriği hakkında bilgi veren kelimeler (Ünlü ve Çetin, 2019) olarak tanımlanmıştır.

Anahtar kelime çıkarımı modelleri önceden etiketlenmiş bir veri kümesine ihtiyaç duyan denetimli algoritmalar ve eğitim kümesine ihtiyaç duymayan denetimsiz algoritmalar olmak üzere iki ana başlık altında toplanmıştır.

Denetimsiz modeller tipik olarak "çıkarım tabanlıdır". Bu modeller, girdi metnine ait "görünür" anahtar kelimeleri bulmaya odaklanır. Bu algoritmaların çoğu problemi iki adımda ele alır. İlk adımda, aday ifadeler dilsel filtreler tarafından belirlenir veya tüm kelimeler giriş metninden çıkarılır. İkinci adımda, bu kelimeler veya aday ifadeler istatistiksel, grafik tabanlı, gömme tabanlı, dilsel veya hibrit öznitelikler kullanılarak derecelendirilir.

Denetimsiz modeller ilk olarak Terim Frekansı (TF/Term Frequency-TF), TF-Ters Doküman Frekansı (TDF/Inverse Document Frequency-IDF) (Ramos, 2003; El-Beltagy ve Rafea, 2009) gibi istatistikel öznitelikler kullanılarak geliştirilmiştir. Aynı yıllarda anahtar kelime çıkarımı problemi Konuşmanın Parçası (KP/Part of Speech-PoS), n-gram gibi dilsel özellikler (Tomokiyo ve Hurst, 2003; Haddoud, Mokhtari, Lecroq ve Abdeddaïm, 2015) kullanılarak ele alınmıştır. İstatistiksel ve dilsel modeller giriş terimleri hakkında güçlü dilsel ve istatistiksel bilgiler sağlar. Ancak bu yöntemler kelimeler ve cümleler arasındaki anlamsal

ilişkiyi betimleyemez. Sözcükler ve cümleler arasındaki anlamsal ilişkiyi tanımlamak için grafik tabanlı (Mihalcea ve Tarau, 2004; Zhao ve diğerleri, 2011; Alfarra ve Alfarra, 2018) ve gömme tabanlı (Benani-Smires, Musat, Hossmann, Baeriswyl ve Jaggi, 2018; Sun Y., Qiu, Zheng, Wang ve Zhang, 2020; Liang, Wu, Liu ve Li, 2021; Ajallouda, Fagroud, Zellou ve Lahmar, 2022) modeller önerilmiştir. Grafik tabanlı Metin Sıralama (TextRank) (Mihalcea ve Tarau, 2004), Başlıklı Sayfa Sıralama (BSS) (Zhao ve diğerleri, 2011) ve Tekil Dokümanlardan Grafik Tabanlı Anahtar Kelime Çıkarımı (TDGT-AKÇ) (Alfarra ve Alfarra, 2018) modelleri, girdi metnine ait kelime birlikte görülme grafiğini çizerek öznitelikleri çıkarır. Bu grafikler, kayan bir pencerede kelimelerin kenarlar olarak birlikte bulunma sayısını hesaplar ve merkezilikleri hesaplayarak anlamsal bilgileri yakalamayı amaçlar.

Gömme tabanlı modeller ise, büyük bir korpus ile oluşturulan kelime gömmelerini kullanarak anlamsal bilgileri yakalamaya çalışır. Bu modellerin temel prensibi, vektör uzayında dokümanı temsil eden doküman gömmesi ile kelime veya aday ifade gömmesi arasındaki benzerliği ölçmektir. Gömme Sıralama, Dokümandan Vektöre (DokVek/Document to Vector-Doc2Vec) (Lau ve Baldwin, 2016) ve Cümleden Vektöre (CümVek/Sentence to Vector-Sent2Vec) (Pagliardini, Gupta ve Jaggi, 2017) kullanarak belgeyi temsil eder. Cümle Sıralama modeli, bağlamsal bilgilere sahip olan ve Çift Yönlü Uzun Kısa Süreli Bellek (ÇY-UKSB/Bidirectional Long Short Term Memory-BiLSTM) ile eğitilmiş Dil Modelinden Kelime Gömmesi (DMKG/Embeddings from Language Model-ELMo) (Peters ve diğerleri, 2018) kullanır. Hem Gömme Sıralama hem de Cümle Sıralama modelleri, kosinüs mesafesini kullanarak benzerliği hesaplar. DMKG'da ÇY-UKSB'in eğitim dezavantajları nedeniyle, Denetimsiz Anahtar Kelime Çıkarımı (DAKÇ) (Liang ve diğerleri, 2021) modeli Dönüştürücülerden Çiftyönlü Kodlayıcı Gösterimi (DÇYKG /Bidirectional Encoder Representations from Transformers-BERT) (Devlin, Chang, Lee ve Toutanova, 2018) gömmelerini kullanır. Bu model, kosinüs benzerliğini mesafeye dönüştürmenin dezavantajını ortadan kaldıran Manhattan mesafesini kullanarak aday ifadeler ve doküman gömmesi arasındaki benzerliği hesaplar. Ajallouda ve diğerleri (2022) tarafından önerilen gömme tabanlı yeni bir model, girdi dokümanları temsil etmek için Evrensel Cümle Gömücüyü (ECG) (Cer ve diğerleri, 2018) kullanır.

Hibrit denetimsiz modeller (Zehtab-Salmasi, Feizi-Derakhshi ve Balafar, 2021; Shen, Wang, Meng ve Shang, 2022), birden fazla kategoriye ait öznitelikleri birlikte kullanır. Bu sayede her bir öznitelik kategorisinin avantajları birleştirilmeye çalışılmıştır. Zehtab-Salmasi ve

diğerleri (2021), metinsel özellikleri (dilsel ve istatistiksel) ve grafik tabanlı özellikleri birlikte kullanan FRAKE adlı denetimsiz bir hibrit model önermiştir. Otomatik Anahtar Üretme adlı başka bir hibrit model, Shen ve diğerleri (2022) tarafından önerilen ilk üretici tabanlı denetimsiz mimaridir. Bu model, adayları hem TF-TDF (istatistiksel) hem de DokVek (gömme tabanlı) özelliklerini birlikte kullanarak sıralar.

Anahtar kelime çıkarımı modellerinin ele alındığı ikinci alt başlık denetimli modeller başlığıdır. Denetimli modeller, makine öğrenimi, yapay zekâ veya derin öğrenme kullanılarak geliştirilmiştir ve etiketlenmiş eğitim verileri gerektirir. Bu modeller "üretim tabanlı" veya "çıkarım tabanlı" olabilir. Metin içerisinde geçmeyen *"görünmez"* anahtar kelimeleri yakalayabilmek için literatürde birçok "üretim tabanlı" mimari önerilmiştir (Meng ve diğerleri, 2017; Yuan ve diğerleri, 2018; Ye J., Cai, Gui ve Zhang, 2021; Wu ve diğerleri, 2021; Zhang, Jiang, Yang, Li ve Wang, 2022). Üretim tabanlı mimariler diziden diziye modellerdir.

Çıkarım tabanlı denetimli algoritmalar, metin içerisinde birebir bulunan "görünür" anahtar kelimeleri yakalamaya odaklanır. Çıkarım tabanlı modeller, problemi bir sınıflandırma veya dizi etiketleme problemi olarak ele alır (Sahrawat ve diğerleri, 2020; Duari ve Bhatnagar, 2020; Liu R., Lin ve Wang, 2020; Gero ve Ho, 2021; Nikzad-Khasmakhi ve diğerleri, 2021)

Giriş metnindeki tüm kelimelerin anahtar kelime veya anahtar kelime değil olarak etiketlenmesi dizi etiketleme olarak adlandırılmaktadır. Literatürde dizi etiketleme yoluyla anahtar kelime problemini ele alan birçok model bulunmaktadır. Basaldella ve diğerleri (2018), ÇY-UKSB ve Global Vektörler (GloVe/Global Vectors-GloVe) gömmelerine dayalı bir dizi etiketleme algoritması önermiştir. Alzaidy ve diğerleri (2019), uzun mesafeli bilgileri birleştirmek için bu modele bir Koşullu Rastgele Alanlar (KRA/Conditional Random Fileds-CRF) katmanı eklemiştir. Sahrawat ve diğerleri (2020) bilimsel yayınlar için ÇY-UKSB-KRA modelini geliştirirken bilimsel içerikle önceden eğitilmiş Bilimsel DÇYKG (BilDÇYKG) gömmelerini kullanmıştır. Bu modelde, bağlamsal bilgiler DÇYKG tabanlı gömmeler kullanılarak oluşturulur. Tüm bu algoritmalar, anahtar kelime içeren önemli cümleleri ve anahtar kelime içermeyen gürültülü cümlelerini eğitir ve test eder. Model performansını iyileştirmek için Liu ve ark. (Liu ve diğerleri, 2020), ÇY-UKSB-KRA mimarisine DÇYKG tabanlı bir cümle filtreleme modülü eklemiştir. Gero ve Ho (2021) gizli katmana merkezi bir ağırlık mekanizması eklemiştir.

Bu tez kapsamında öncelikle literatürde anahtar kelime çıkarımı problemini çözmek amacıyla geliştirilen modeller detaylı bir şekilde incelenmiştir. Anahtar kelime çıkarımında kullanılan kelime gömmeleri, veri kümeleri ve performans sonuçları hesaplama yöntemleri alt başlıklarda ele alınmıştır. Modellerin mimari detayları, test edildiği veri kümeleri, hangi dil için geliştirildiği gibi bilgiler tablo halinde sunulmuştur. Öerilen modellerin performans sonuçları literatürde kullanılan veri kümeleri için değerlendirilmiştir ve sonuçlar irdelenmiştir.

Tezin ikinci bölümünde anahtar kelime çıkarımı için bu tez kapsamında geliştirilen Hibrit Anahtar Kelime Çıkarımı (HibritAKÇ) modeli tanıtılmıştır. Bu bölümde çalışma ortamının detayları verildikten sonra modelin performans sonuçları tablo halinde sunulmuş, literatürde sık kullanılan Inspec, 500N-KeyPhrasesCrowdAnnotated-Corpus (500N-KPCrowd), Semeval-2017 gibi veri kümelerinin performans sonuçlarıyla önerilen modelin sonuçları karşılaştırılmıştır. Tartışmalar bölümünde önerilen modelin çıktıları görselleştirilmiş avantajları ve eksik yönleri irdelenmiştir.

2. LİTERATÜR İNCELEMESİ

Literatürde anahtar kelime çıkarımı için çok sayıda model önerilmiştir. Bununla birlikte önerilen modellerin performans sonuçlarına bakıldığında problemi çözme performansları hala beklenenin çok altındadır. Bu modeller temelde denetimli ve denetimsiz olmak üzere iki başlık altında toplanmaktadır. Denetimli modeller önceden etiketlenmiş bir eğitim kümesine ihtiyaç duyarken denetimsiz metotlar önceden derlenmiş bir veri kümesine ihtiyaç duymamaktadır. Çoğu denetimsiz algoritma bir korpustan ziyade tek bir girdi dokümanı kullanarak anahtar kelime çıkarımı görevini gerçekleştirmektedir.

Denetimsiz modeller üç alt başlık altında toplanmıştır. Bunlar istatistiksel, grafik-tabanlı ve gömme-tabanlı modeller olarak guruplandırılmıştır.

Denetimli modeller ise dört alt başlık altında incelenmiştir. Bu başlıklar çıkarım temelli, üretim temelli, hem çıkarım hem üretim temelli ve pekiştirmeli öğrenme modelleri olarak belirlenmiştir. Çıkarım temelli modeller görünür anahtar kelimeleri yakalamaya çalışmaktadır. Metin içerisinde bulunmayan fakat metinin anlamından çıkarılan görünmeyen anahtar kelimelere üretici tabanlı denetimli modeller odaklanmaktadır. Hem üretim hem çıkarım görevini birlikte gerçekleştiren kütüphaneler hem çıkarım hem üretim temelli olarak gruplandırılırken, pekiştirmeli öğrenme kullanarak geliştirilen modeller pekiştirmeli öğrenme başlığı altında derlenmiştir.

2.1. Kelime Gömmeleri

Kelime gömmeleri veya önceden eğitilmiş dil modeli olarak adlandırılan kelime gösterimleri, DDİ uygulamalarında sıklıkla kullanılmaktadır. Girdi kelimenin gömmesi tipik olarak bir vektör matrisidir. İki benzer kelimenin gömmesi önceden tanımlanmış vektör uzayında birbirine daha yakındır. Kelime gömmeleri literatürde önerilen hem denetimli hem denetimsiz modellerde kullanılmaktadır. Bu gömmeler hem üretim temelli hem çıkarım temelli denetimli algoritmalarda modelin birincil girdisidir.

Sözcük Çantası (SÇ/Bag of Words-BoW) olarak adlandırılan eski kelime gömmesi yaklaşımlarından biri, bir cümlede bulunan tüm sözcüklerin ikili frekans tabanlı matris

gösterimini içerir. Bununla birlikte, cümlede bulunan her kelimenin frekans değeri tek-sıcak bir vektör olarak tanımlandığından, SÇ'da artan bir hafıza ihtiyacı vardır. Ayrıca, tek-sıcak vektör yalnızca hedef kelimenin frekans değerine odaklandığından anlamsal olarak kelimeyi tanımlayamayabilir. SÇ algoritmasının bellek ve semantik bağlantı sorunları Kelimeden Vektöre (KV/Word to Vector-Word2Vec) gömmeleri ile çözülmüştür. KV (Mikolov, Chen, Corrado ve Dean, 2013) SÇ kullanarak kelime vektörlerini hesaplar. Kelimeler yoğunluk vektörleri ile temsil edilir ve iki kelime arasındaki anlamsal bağlantı kosinüs benzerliği ile ölçülür.

KV'de iki kelime yoğunluğu hesaplanırken kelimenin belli bir pencere boyutu içerisinde birlikte görülme frekansı göz önüne alırken, GloVe (Pennington, Socher ve Manning, 2014) kelime-bağlam birlikte görülme matrisini açıkça çarpanlara ayırır. Bu sayede GloVe istatistiksel bilgileri verimli bir şekilde kullanır. GloVe yalnızca sıfır olmayan öğeleri bir sözcük birlikte görülme matrisinde eğitir. Bu yaklaşımla GloVe, global matris çarpanlara ayırma ve yerel bağlam penceresi yöntemlerinin avantajlarına aynı anda sahiptir.

KV 'nin bir uzantısı olan hızlı metin (fastText) (Bojanowski, Grave, Joulin ve Mikolov, 2017), her kelimenin bir karakter n-gramları olarak temsil edildiği skip-gram modeline dayanmaktadır. Skip-gram modelinde amaç verilen bir kelimenin bağlamını tahmin etmek iken Sürekli SÇ (S-SÇ/Continuous BoW-C-BoW) modelinde amaç bir kelimeyi bağlamına göre tahmin etmektir. Bir kelime, daha düşük seviyeli gömmelerin bir kombinasyonu olarak temsil edilir. Böylece kelimelerin morfolojisi dikkate alınmayarak genelleme mümkün hale gelir. Fakat aynı zamanda büyük kelime hazinesi olan ve çok sayıda nadir kelimeye sahip diller için geniş hafıza ihtiyaçları nedeniyle birçok nadir kelimenin atılması gerekir.

Peters ve diğerleri (2018) tarafından önerilen DMKG, kelime kullanımının karmaşık özelliklerini ve dilsel bağlamlarda nasıl değiştiğini modelleyen derin bağlamsallaştırılmış bir kelime temsilidir. Kelime gömmeleri oluşturmak için birleşik dil modeli hedefiyle eğitilmiş ÇY-UKSB'den türetilen vektörler kullanılmıştır. Diğer yaklaşımlardan farklı olarak, DMKG, derin çift yönlü dil modelinin tüm iç katmanlarının bir fonksiyonundan oluşur.

Son zamanlarda sıklıkla kullanılan DÇYKG gömmeleri, DMKG ve birkaç Dönüştürücüyü birleştiren önceden eğitilmiş bir dil modelidir. Bu gömmeler, tüm katmanlarda hem sol hem

de sağ bağlamda ortak koşullandırma yoluyla etiketlenmemiş metinden derin çift yönlü gömmeleri önceden eğitmek için tasarlanmıştır. Soruna özel modeller oluşturmak için ek bir çıktı katmanında ince ayar yapabilmektedir. DÇYKG'nin bir kelimeye atadığı vektör, cümlenin bir fonksiyonu şeklinde tanımlanmaktadır. Bu, bir kelimenin bağlamlara göre farklı vektörlere sahip olabilmesini garanti etmektedir.

2.2. Denetimli Anahtar Kelime Çıkarımı

Denetimli öğrenme yaklaşımı eğitim setlerini kullanarak anahtar kelimeleri çıkarmı için bir model eğitir. Eğitilen bu model farklı bir veri seti üzerinde test edilerek performans sonuçları yeterli olması durumunda anahtar kelime çıkarımı için kullanılır. Uygun bir model elde edildikten sonra yeni metinler için anahtar kelime çıkarılırken kullanılır. Ancak denetimli bir modeli eğitmek için büyük veri seti gerekmektedir. Anahtar kelime çıkarımı için literatürde birçok denetimli yöntem önerilmiştir. Önceki yıllarda daha fazla sınıflandırma algoritması kullanılırken, yapay sinir ağları alanında ilerleme kaydedildikçe, çalışmaların yoğunluğu bu alana kaymaya başlamıştır. Derin öğrenme yöntemlerinin çeşitli problemlere başarılı bir şekilde uygulanması, anahtar kelime çıkarımı için bu yöntemleri popüler hale getirmiştir.

Bu bölümde anahtar kelime çıkarımı için literatürde önerilen denetimli modeller 4 başlık altında incelenmiştir. Birinci kategoride çıkarım tabanlı modeller, ikinci kategoride üretim temelli modeller bir araya toplanmıştır. Üçüncü kategori altında hem çıkarım hem üretim temelli olan modeller ele alınmıştır. Son kategori pekiştirmeli öğrenme yöntemi kullanılarak geliştirilen modelleri kapsamaktadır.

2.2.1. Çıkarım temelli modeller

Çıkarım temelli algoritmalar, görünür anahtar kelimeleri işaretlemeye odaklanır. İstatistiki ve dilsel filtreler kullanılarak önceden ayrıştırılan aday ifadeler veya metine ait anlamlı tüm belirteçler girdi olarak modele verilir. Aday ifadelerin girdi olarak verilmesi durumunda aday ifadeler makine öğrenme algoritmaları ile sınıflandırılır. Metine ait anlamlı tüm tokenler girdi olarak kullanılması durumunda ise yine makine öğrenme sınıflandırma algoritmaları ile dizi etiketleme görevi gerçekleştirilir.

Bu bölümde denetimli anahtar kelime çıkarımı modelleri detaylı olarak incelenmiştir. Bu modeller grafik tabanlı sınıflandırma ve Tekrarlayan Sinirsel Ağ (ÖSA/Recurrent Neural Network-RNN) tabanlı sınıflandırma olmak üzere iki alt başlıkta gruplanmıştır. Grafik tabanlı sınıflandırmada öncelikle girdi metnine ait Grafiksel Evrişimsel Ağ (GEA) veya kelime grafiği oluşturulur. Sonrasında sınıflandırma katmanında Naive Bayes, Destek Vektör Makinası (DVM), Çok Katmanlı Algılayıcı (ÇKA), Neural Network gibi algoritmalar kullanılarak dizi etiketleme veya sınıflandırma görevi gerçekleştirilir. ÖSA tabanlı modeller kategorisinde ise UKSB, Geçitli Tekrarlayan Birim (GTB/Gated Recurrent Unit-GRU), ÇY-UKSB modelleri kullanılarak sınıflandırma veya dizi etiketleme yapan derin öğrenme algoritmaları incelenmiştir.

Grafik temelli sınıflandırma

Bu bölümde grafik temelli sınıflandırma yapan 3 denetimli anahtar kelime çıkarma mimarisi incelenmiştir. Mimariler, kelime birlikte görülme grafiğini veya GEA oluşturduktan sonra problemi Naïve Bayes, Aşırı Gradyan Artırma (AGA), Sinirsel Ağ vb. gibi klasik makine öğrenmesi kullanarak çözmeye çalışmıştır. Girdi metni ile kelime birlikte görülme grafiği çizilerek kelimeler arasındaki semantik bilgi temsil edilmiştir.

Çeşitlilik grafik işaretçisi

Çeşitlilik Grafik İşaretçisi (ÇGİ/Diversity Graphic Pointer-DivGraphPointer) mimarisi (Sun Z., Tang, Du, Deng ve Nie, 2019); geleneksel grafik tabanlı sıralama yöntemlerinin ve sinir ağı tabanlı yaklaşımların avantajlarını birleştiren yeni bir yöntemdir. Kelime grafiği metindeki sözcükler arasındaki uzun mesafeli bağımlılığı modelleyerek aynı kelimeleri tek bir düğüme dönüştürür. Belgenin kelime grafiği oluşturulurken standart birlikte görülme değeri yerine cümlede kelimeler arasındaki uzaklığa dayalı bir ilişki metodu kullanılmıştır. Yönlendirilmiş bir kelime grafiği için yakınlık matrisleri $\overleftarrow{A}_{ij}$ and $\vec{A}_{ij}$ olmak üzere, w_i kelimesinden w_j kelimesine ağırlık Eşitlik 2.1 ve Eşitlik 2.2'de görüldüğü gibi hesaplanmaktadır.

$$\overleftarrow{A}_{ij} = \sum_{p_i \in \varphi(w_i)} \sum_{p_j \in \varphi(w_j)} relu\left(\frac{1}{p_i - p_j}\right) \tag{2.1}$$

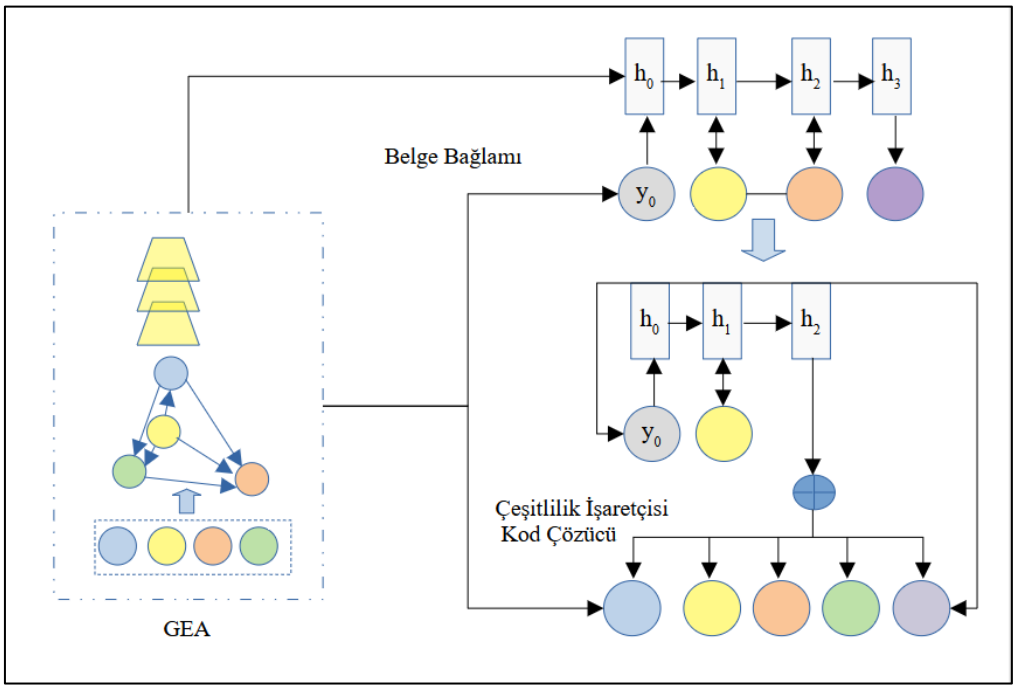
$$\vec{A}_{ij} = \sum_{p_i \in \varphi(w_i)} \sum_{p_j \in \varphi(w_j)} relu\left(\frac{1}{p_j - p_i}\right) \quad (2.2)$$

Bu iki eşitlikte relu yönsüz bilgiyi filtrelemek için kullanılan matematiksel bir fonksiyondur ve $\varphi(w_i)$, dokümandaki her kelime (w_i) için konumsal uzantıların (p_i) kümesidir. Kod çözme aşamasında kelime grafiğinden bir dizi farklı anahtar kelime üretmek için çeşitlendirilmiş bir işaretçi ağ önerilmiştir. Bu mimaride üretilen anahtar sözcük çeşitliliğini artırmak için iki mekanizma eklenmiştir. Bunlar tekrarlayan anahtar kelime problemini çözmek için önerilen kapsama mekanizması ve bağlam değişiklik mekanizmasıdır.

Bağlam değişiklik mekanizmasında, önceki üretilen anahtar kelimelerden yola çıkarak bağlam otomatik olarak güncellenir. Üretilen anahtar kelimelerin ortalama gösterimi $\overline{y^{(1:i-1)}}$ bağlam değişikliği $h_0^{(i)}$ ve kapsama mekanizması $c_j^{(i)}$ Eşitlik 2.3 ile hesaplanır:

$$h_0^{(i)} = tanh\left(W_h\left[c, \overline{y^{(1:i-1)}}\right] + b_s\right); c_j^{(i)} = c_j^{(i-1)} + \sum_t o_{t,j}^{(i-1)} \quad (2.3)$$

Bu eşitlikte $o_{t,j}^{(i)}$, t_{inci} anahtar kelimenin j_{inci} tek-sıcak vektördür. Şekil 2.1'de Çeşitlilik Grafik İşaretçisi modeli bulunmaktadır.



Şekil 2.1. Çeşitlilik grafik işaretçisi modeli

Karmaşık ağ tabanlı grafiksel modeller

Duari ve Bhatnagar (2020), Naive Bayes (NB) ve özel ağaç oluşturarak sınıflandırma yapan AGA sınıflandırıcılarını kullanarak iki ayrı model geliştirmiştir. Bu modeller anahtar kelime çıkarımını Karmaşık Ağ (KA/Complex Network-CN) tabanlı gerçekleştirmiş ve sırasıyla KA-NB ve KA-AGA olarak isimlendirilmiştir. Modeller tek bir belgenin kelimeleri için metinin karmaşık ağını oluşturmaktadır. Önce denetimsiz kelime vektörleri oluşturduktan sonra sınıflandırma yapılır. Bu sayede önerilen model dil bağımsız çalışabilmektedir. Metinden aday anahtar kelimeleri filtrelemek için istatistiksel σ-index, yani ardışık olaylarda sözcüğün aralık dağılımının normalleştirilmiş standart sapması kullanılmıştır. Bu sayede aday ifade belirleme aşamasında da dil bağımlılığı ortadan kaldırılmıştır.

Sınıflandırmada kullanılmak üzre grafik tabanlı 6 farklı öznitelik hesaplanmıştır:

- Düğümün kuvveti (ağırlığı),
- Özvektör merkezliliği (bu hesaplama ile önemli kelimelerle birlikte görülen kelimeler daha önemlidir)
- Sayfa Sıralama
- Poziyon Sıralama (anahtar kelimelerin daha çok makalenin başlarında bulunması ve buna göre hesaplama yapılması)
- Çekirdek
- Çekirdeksizlik, G ağını bir dizi maksimum bağlı G_k alt kümesine (k çekirdeği gösterir) ayıran bir ağ dejenerasyon özelliğidir, böylece G_k 'daki düğümler alt G_4 ve $G_k \subseteq G_{k+1}$ içinde en az k dereceye sahiptir. Bir düğümün çekiciliği, ait olduğu en yüksek çekirdektir.
- Kümeleme Katsayısı, düğümün çevresindeki kenar yoğunluğudur.

İfade formatlayıcı

İfade Formatlayıcı, grafik-tabanlı ve gömme tabanlı yöntemlerin bir kombinasyonudur (Nikzad-Khasmakhi ve diğerleri, 2021). Bu model DÇYKG kelime gömmesini kullanmaktadır. Anahtar sözcük çıkarımı için önerilmiş hibrit bir modeldir ve problemi bir dizi etiketleme problemi olarak ele almıştır.

Başlangıçta, belgeler iki ayrı modüle gönderilir. İlki birlikte oluşum grafiğinin oluşturulduğu grafik modülü, ikincisi ise metin öğrenmenin yapıldığı DÇYKG tabanlı modüldür. Her iki modülün çıktıları birleştirilir ve en son katman olan dizi etiketleme ağına girdi olarak gönderilir. Dizi etiketleme katmanı, bir ileri beslemeli ağ ve bir softmax katmanından oluşur.

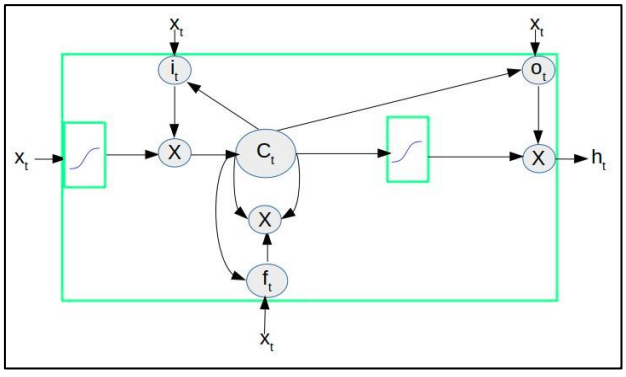
ÖSA temelli sınıflandırma

Son yıllarda derin öğrenme teknikleri kullanılarak anahtar kelime çıkarımı için geliştirilmiş birçok model bulunmaktadır. Modellerde başlangıçta ÖSA kullanılmıştır. Bir girdi metninin her bir kelimesini sırası ile işleyen ÖSA'da girdi kelimeler $\{x_1 \dots . x_n\}$ olarak tanımlanmaktadır. t. durumda ÖSA'nın gizli katmanı (t ∈ n) Eşitlik 2.4'teki gibi hesaplanır:

$$h_t = f_w\,(h_{t-1}\,,\, x_t\,) \qquad (2.4)$$

ÖSA her bir gizli katmanı x girdi değişkeni için bir küme olarak tutar. Formüldeki f fonksiyonu w parametreli lineer olmayan bir fonksiyondur. Üretilen her bir çıktı için *kayıp* değeri hesaplanır ve *kayıp* değeri ÖSA'ya girdi olarak verilir.

ÖSA'nın UKSB ve GTB versiyonları ile modelin tekrarlı yapısından kaynaklanan patlayan ve kaybolan gradyan problemleri çözülmüştür. UKSB ve GTB'da bir hücre hafızası diğer hücre hafızasına gizli katman aracılığı ile aktarılır. Şekil 2.2'de UKSB hücresi görülmektedir.



Şekil 2.2. UKSB hücresi (Graves ve Schmidhuber, 2005)

Bir UKSB her bir x_t tokeni için y_t çıktı vektörünü Eşitlik 2.5-2.10'a kadar olan eşitlikleri kullanarak 1'den n'e kadar tekrarlayarak hesaplar (Zhang X., Chen ve Huang, 2018).

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (2.5)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (2.6)$$

$$\tilde{C}_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_C) \quad (2.7)$$

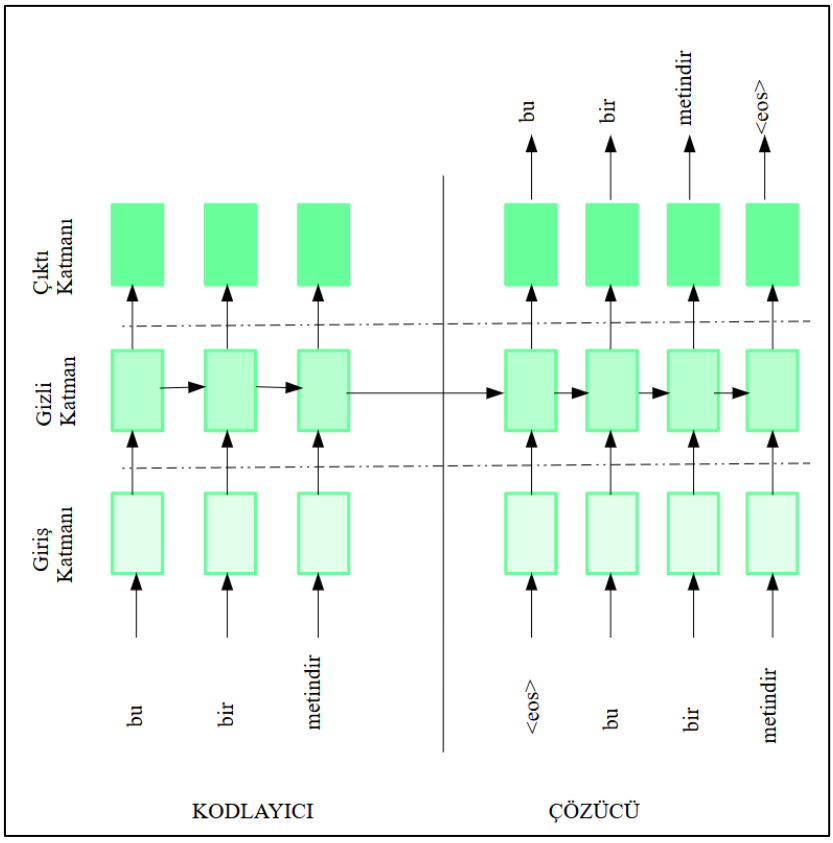
$$C_t = i_t \otimes \tilde{C}_t + f_t \otimes C_{t-1} \quad (2.8)$$

$$o_t = \sigma(W_{xo}\ x_t + W_{ho}h_{t-1} + b_o) \quad (2.9)$$

$$h_t = o_t \otimes tanh\,(C_t) \quad (2.10)$$

Burada σ lojistik sigmoid fonksiyonu *i, f, o* ve *c* sırasıyla girdi kapısı, unutma kapısı, çıktı kapısı ve hücre aktivasyon vektörüdür. *B'*ler öğrenilmiş sapmalar ve $\otimes$ matrix çarpımıdır.

Literatürde, iki ÖSA'dan oluşan kodlayıcı ve çözücü ağlarına sahip diziden diziye mimarisi derin öğrenme modellerinde sıklıkla kullanılmaktadır. Diziden diziye mimarisinin makine çevirisi, görüntü etiketleme, duygu analizi gibi problemleri çözmede başarılı skorlar elde ettiği görülmektedir. Bu sebeple diziden diziye mimarisi anahtar kelime çıkarımı için de sıklıkla kullanılmıştır. Şekil 2.3'te diziden diziye mimarisi görülmektedir.



Şekil 2.3. Diziden diziye kodlayıcı / çözücü mimarisi

Bu bölümde, anahtar kelime çıkarma problemini ÖSA ve diziden diziye mimarisi temelli bir sınıflandırma veya dizi etiketleme problemi olarak çözmeye çalışan çıkarım temelli modeller incelenmektedir.

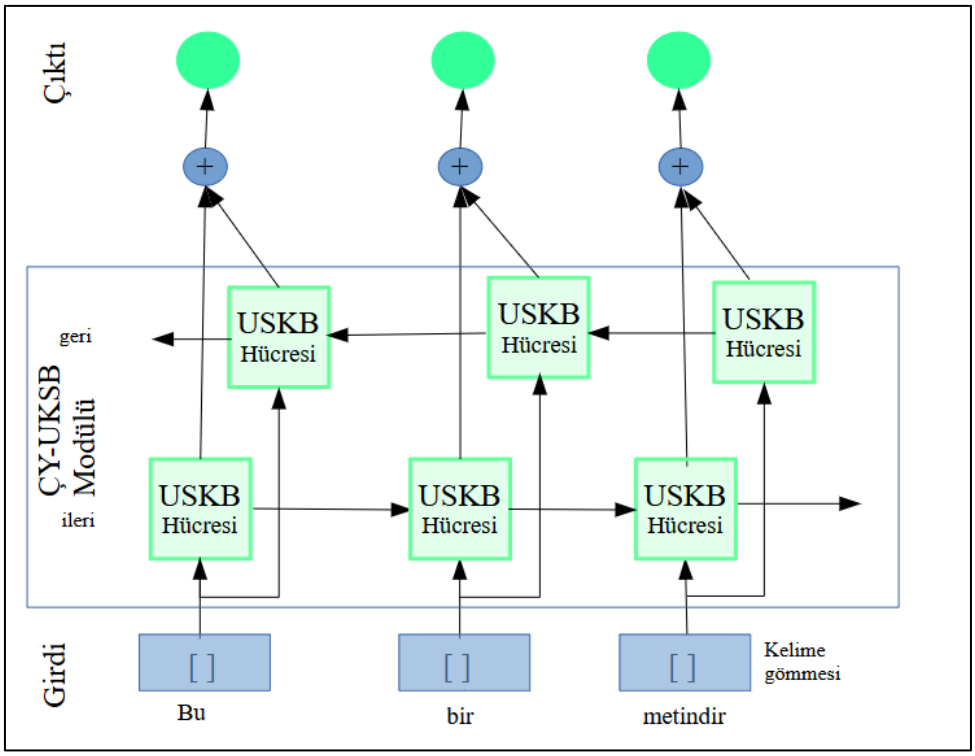
ÇY-UKSB modeli

ÇY-UKSB mimarisi Şekil 2.4'te görüldüğü gibi iki UKSB katmanından oluşmaktadır (Basaldella ve diğerleri, 2018). Bu katmanlar ileri ve geri besleme yapabilecek şekilde tasarlanmıştır. Örneğin, birkaç kelimeden oluşan bir ifade birinci katmandaki UKSB'den soldan sağa geçirilirken ikinci katmandaki UKSB'den sağdan sola geçirilmektedir. ÇY-UKSB gizli katmaları Eşitlik 2.11, Eşitlik 2.12 ve Eşitlik 2.13'teki formüller ile hesaplanır:

$$\vec{h}_t = H(W_{\overrightarrow{xh}} x_t + W_{\overrightarrow{h}\,\overrightarrow{h}} \vec{h}_{t-1} + b_{\vec{h}}) \quad (2.11)$$

$$\overleftarrow{h}_t = H(W_{\overleftarrow{xh}} x_t + W_{\overleftarrow{h}\,\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}) \quad (2.12)$$

$$y_t = W_{\vec{h}\,y} \vec{h}_t + W_{\overleftarrow{h}\,y} \overleftarrow{h}_t + b_y \quad (2.13)$$



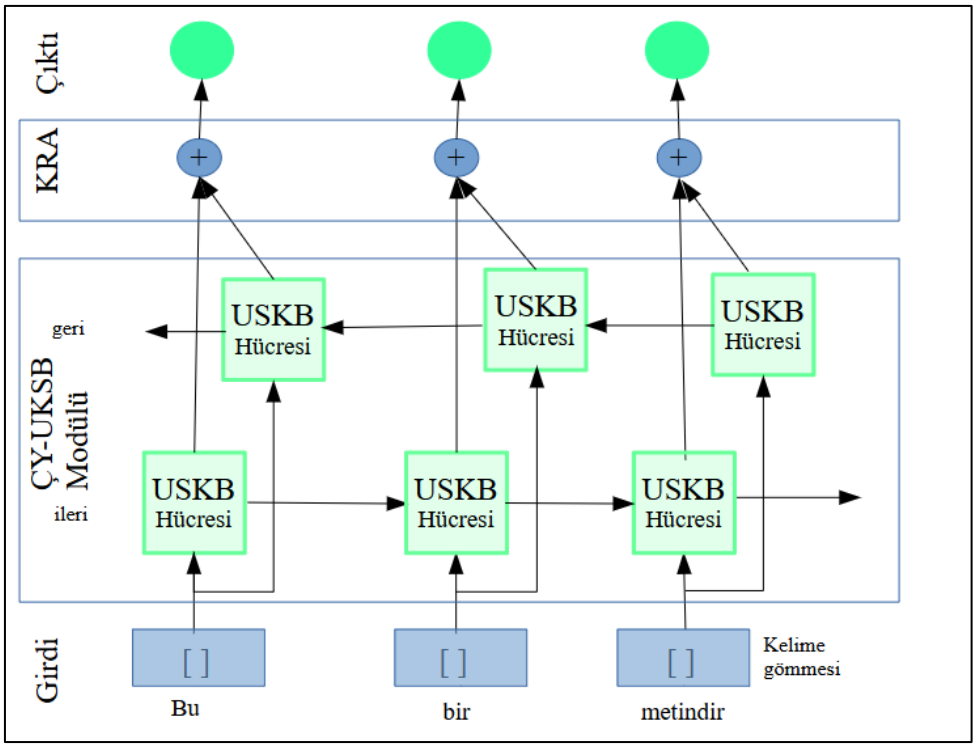
Şekil 2.4. ÇY-UKSB modeli

ÇY-UKSB-KRA

ÇY-UKSB-KRA modeli (Alzaidy ve diğerleri, 2019); anahtar kelime çıkarımını dizi etiketleme problemi olarak ele almaktadır. Şekil 2.5'te ÇY-UKSB-KRA modeli görülmektedir. Çıkış katmanında birçok dizi etiketleme görevinde başarılı olmuş denetimsiz bir yöntem olan KRA kullanılmaktadır. KRA tarafından tanımlanan x verilen y etiket dizisi üzerindeki koşullu olasılık dağılımı Eşitlik 2.14'te görüldüğü gibidir:

$$p(y|x;W,b) \propto exp\left(\sum_{i=1}^{n} W_{y_{i-1},y_1}^{T} x_i + b_{i-1,yi}\right) \quad (2.14)$$

x ve y anahtar kelime ve anahtar kelime değil olmak üzere iki farklı etikettir. Bir etiketten komşu bir etikete geçiş olasılıklarından oluşan bir geçiş parametresi matrisi aracılığıyla etiket bağımlılıklarını yakalar. Ağırlık vektörü (W) ve bias (b) modele ait parametrelerdir.



Şekil 2.5. ÇY-UKSB-KRA modeli

ÇY-UKSB-KRA tabanlı Doküman Düzeyinde Anahtar Kelime Çıkarma (DD-AKÇ) (Santosh ve diğerleri, 2020) modelinde, doküman düzeyinde bağlamsal bilgiler, doküman düzeyinde dikkat mekanizması ile ağa dâhil edilir. Önceden eğitilmiş DÇYKG gömmeleri kelime temsil vektörleri olarak kullanılmıştır.

Giriş cümlesinde (s_i) her bir kelime (w_{ij}) için doküman seviyesi önem mekanizması Eşitlik 2.15'teki gibi uygulanmaktadır:

$$e_{ij}^{l} = v^{\mathrm{T}}\, tan\,(W_1 h_{ij} + W_2 h_l' + b_1)\,,\ \alpha_{ij}^{l} = \frac{\exp\,(e_{ij}^{l})}{\sum_{p=1}^{m} \exp\,(e_{ij}^{p})} \qquad (2.15)$$

Burada W_1, W_2 ağırlıklar ve b_1 sapmadır. Her bir w_{ij} kelimesi için $\tilde{h}_{ij}$ doküman seviyesi önem ifadesi Eşitlik 2.16'daki gibi hesaplanmaktadır:

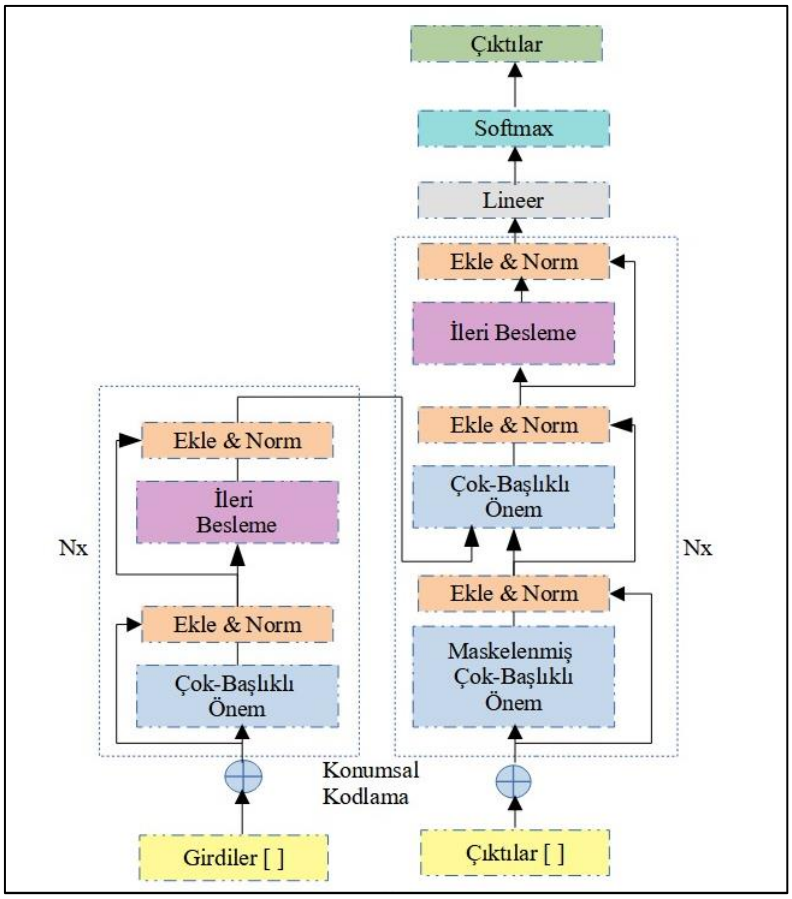
$$\tilde{h}_{ij} = \sum_{p=1}^{m} \alpha_{ij}^{l}\, h_l' \qquad (2.16)$$

ÇY-UKSB-KRA mimarisi tabanlı başka bir mimari bilimsel DÇYKG gömmelerini olan BilDÇYKG bağlam gömmeleri kullanılarak geliştirilmiş ÇY-UKSB-KRA BilDÇYKG (Sahrawat ve diğerleri, 2020) modelidir. Bu mimari ile yüksek performans elde edilmiştir.

DÇYKG

DÇYKG mimarisi Şekil 2.6'da görülen dönüştürücü (Vaswani ve diğerleri, 2017) ile birlikte yalnızca dikkat mekanizmalarına dayanan, tekrarlama ve normalizasyon katmanlarından oluşan basit bir mimaridir. Bu mimaride ÖSA'lardaki işlemleri paralel hale getirmek için cümledeki tüm kelimeler konumları ile birlikte girdi olarak gönderilirler. DÇYKG tüm katmanlarda hem sol hem de sağ bağlamda birlikte koşullandırılarak etiketlenmemiş metinin derin çift yönlü gömmelerini önceden eğitmek için tasarlanmıştır. DÇYKG kullanılarak önceden eğitilmiş model üretmek için özellik tabanlı ve ince ayarlamalı şeklinde iki yaklaşım kullanılmaktadır. Şekil 2.6'da görülen Transformatörlerin kodlayıcısı ve çözücüsü ayrılıp, birden fazla kodlayıcı ard arda bağlanarak DÇYKG mimarisini oluşturulur.

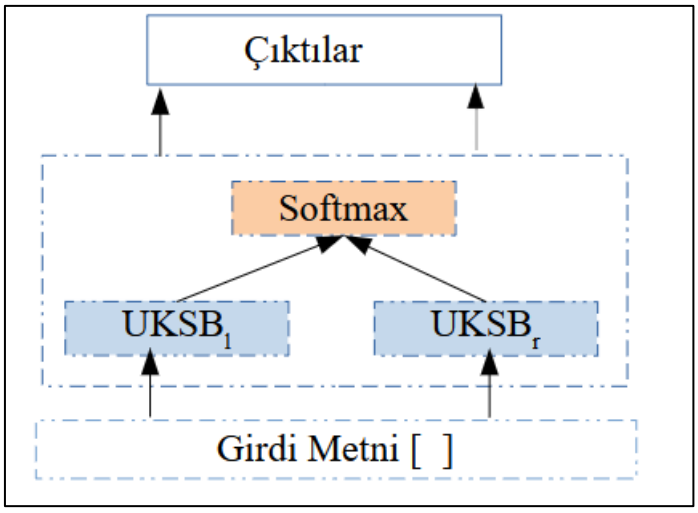
DÇYKG maskelenmiş dil modeli ve sonraki cümle tahmini görevlerini gerçekleştirir. Girdi gömmeleri; belirteç gömmeleri, segmentasyon gömmeleri ve konum gömmelerinden oluşur.



Şekil 2.6. Dönüştürücü modeli

Hedef merkez tabanlı UKSB

Hedef Merkez Tabanlı UKSB (HMT-UKSB), Zhang Y. Ve diğerleri (2020) tarafından önerilmiştir. Bu model Şekil 2.7'de gösterildiği gibi bağlamsal bilgileri dikkate alarak hedef kelimeyi kodlamayı öğrenir. Önerilen model, geleneksel UKSB modeli temelinde iki uzantı kullanılarak üretilmiştir. İlk uzantı, verilen hedef kelimenin hem tarihsel hem de bağlamsal bilgisini daha iyi kullanmak için gerçekleştirilmiştir. İkinci uzantı, HMT-UKSB modeline dayalıdır ve modelin ilgili metnin bilgilendirici kısımlarına odaklanma becerisine sahip olmasını sağlayan öz-dikkat mekanizmasına sahiptir. Bu model, verilen kelimenin hem önceki hem de sonraki bağlamlarını aynı anda modeller.

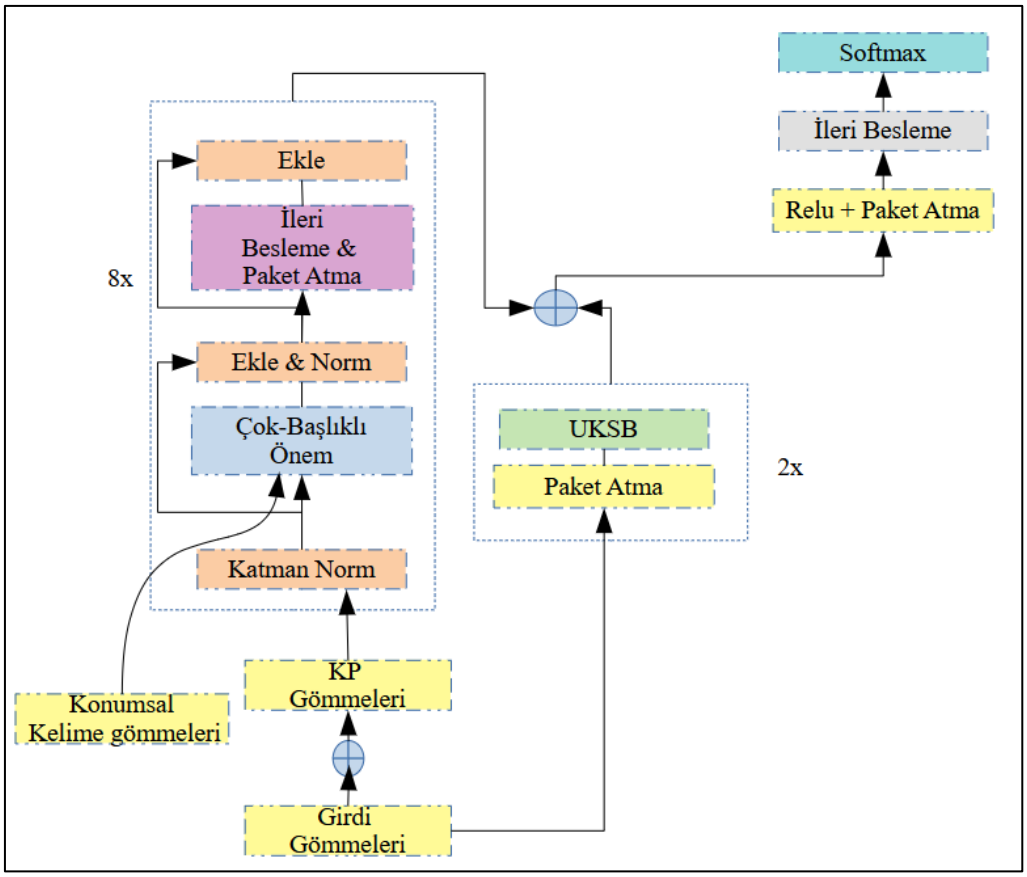


Şekil 2.7. Hedef merkez-tabanlı UKSB modeli

Dönüştürücü tabanlı sinirsel anahtar kelime etiketleyici

Dönüştürücü Tabanlı Sinirsel Anahtar Kelime Etiketleyici (DTS-AKE) modeli (Martinc, Škrlj ve Pollak, 2020) ön eğitimli dil modeli kullanarak geliştirilmiştir ve Şekil 2.8'de modelin mimarisi görülmektedir. İnce ayar safhasında bir UKSB ve silme mekanizmasından oluşan iki katmanlı kodlayıcı kullanır. Bu kodlayıcı dönüştürücünün çıkışına eklenir. Bir kelimenin anahtar kelime olma olasılığını yakalamak için eklentiler kullanılmıştır.

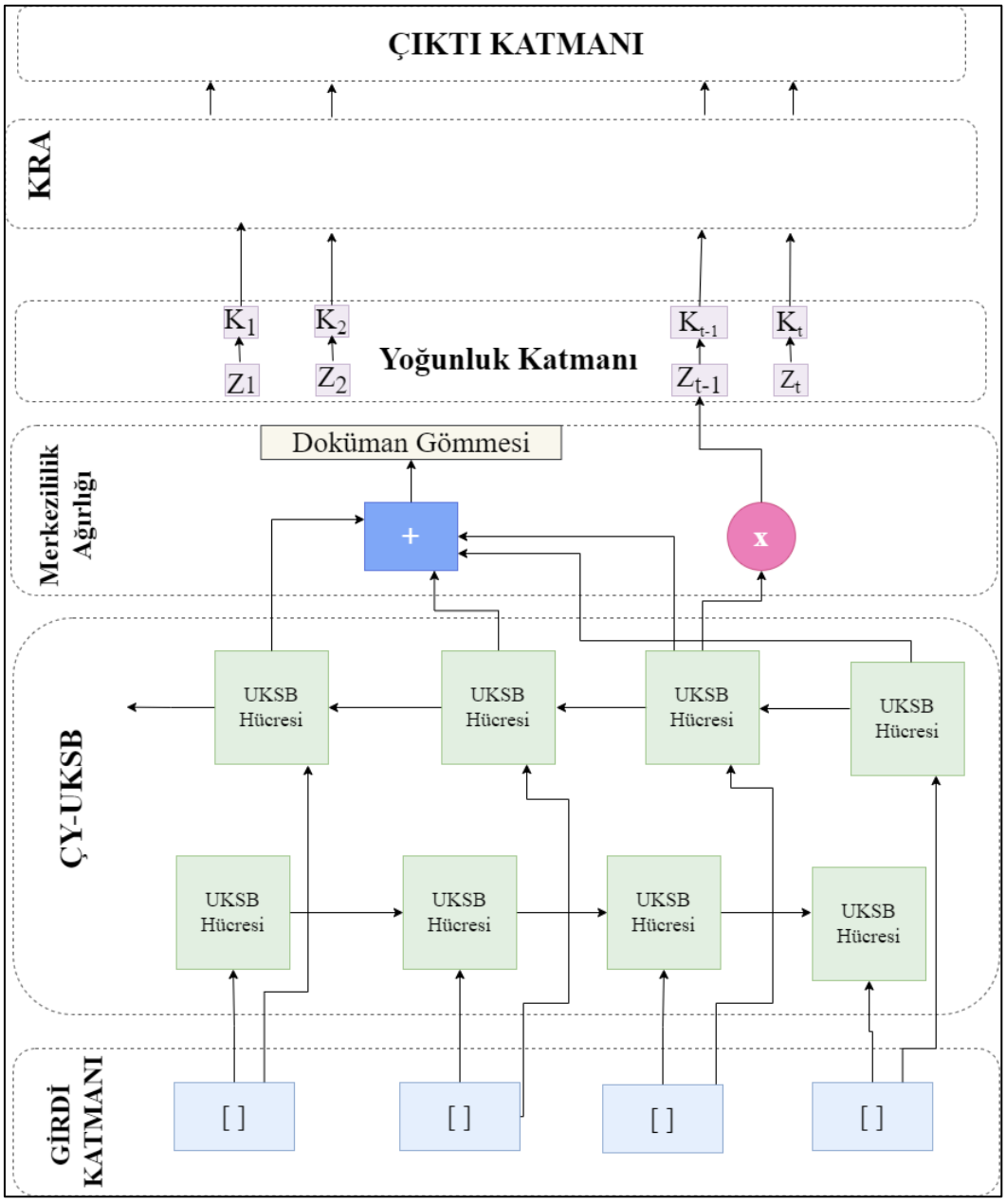
Bu modelde bir belirteç ile belirtecin konumu arasındaki ilişkiyi modellemeyi sağlayan dikkat mekanizması geliştirilmiştir. Standart nokta çarpımı önem mekanizmasına ek olarak, konum bilgisini gösteren bir K_{KP} vektörü önem hesaplamasına dâhil edilmiştir. Bu sayede model, her belirteç ve her konum arasındaki ilişkinin önemini doğrudan yakalayabilmektedir. Giriş aşamasında KP etiketleri de ilave bilgi olarak kodlayıcıya gönderilir.



Şekil 2.8. Dönüştürücü-tabanlı sinirsel anahtar kelime etiketleyici

Kelime merkeziliği sabitlenmiş gösterimi

Kelime Merkeziliği Sabitlenmiş Gösterimi (KMSG) (Gero ve Ho, 2021) modeli, Şekil 2.9'da gösterildiği gibi ÇY-UKSB-KRA modeline dayalı bir anahtar kelime çıkarımı mimarisidir. Klasik ÇY-UKSB-KRA modelinden farklı olarak, merkezilik sabiti benzerlik değeri ölçülerek hesaplanır. Her kelimenin belge ile kosinüs benzerliği, KRA katmanında hesaplanır. W_1, w_2, · · · , w_t sözcükleri için α_1, α_2, · · · , α_t değerleri, belge vektörü H kullanılarak her kelime gömmesinin kosinüs benzerliği ile hesaplanır. Inspec veri kümesi için DÇYKG, Pubmed veri kümesi için Biyoloji DÇYKG (BiyoDÇYKG) gömmeleri kullanılmıştır.



Şekil 2.9. Kelime merkeziliği sabitlenmiş gösterimi modeli

2.2.2. Üretim temelli modeller

Üretim temelli anahtar kelime çıkarımı modelleri metnin içerisinde bulunmayan fakat metin içerisinden anlamı çıkabilen “görünmeyen” anahtar kelimeleri üretmeye odaklanmaktadır. Bu modeller diziden diziye derin öğrenme mimarileridir. Bu bölümde, 5 diziden diziye üretici tabanlı model ayrıntılı olarak incelenmektedir.

Kopya ESA

Kopya ESA (Zhang, Fang ve Weidong 2017) ÖSA'ların önceki zaman adımlarındaki gizli katmanlara dayanmasından kaynaklı performans problemini çözmek için önerilmiştir. ESA kullanılması modele paralel olarak çalışabilme özelliğini kazandıran hiyerarşik bir çalışma şekli kazandırır. Hiyerarşik çalışmadan kastedilen elemanların yakınlığına göre daha önce veya daha sonra etkileşime girmesidir. Bu da uzun menzilli bağımlılıkları yakalamak için daha kısa bir yol sağlamaktadır. Her bir kod çözücü katmanına önem mekanizması eklenmiştir. Sözlük dışı kelimeleri ana metinden temin etmek için kopyalama mekanizması kullanılmıştır. Girişlerde düşük boyutlu w vektörü ile giriş elemanlarının pozisyonu modele girdi olarak verilmektedir. Hem kodlayıcı hem çözücü geçiş durumlarının giriş elemanlarından hesaplandığı evrişimsel bir yapıya sahiptir.

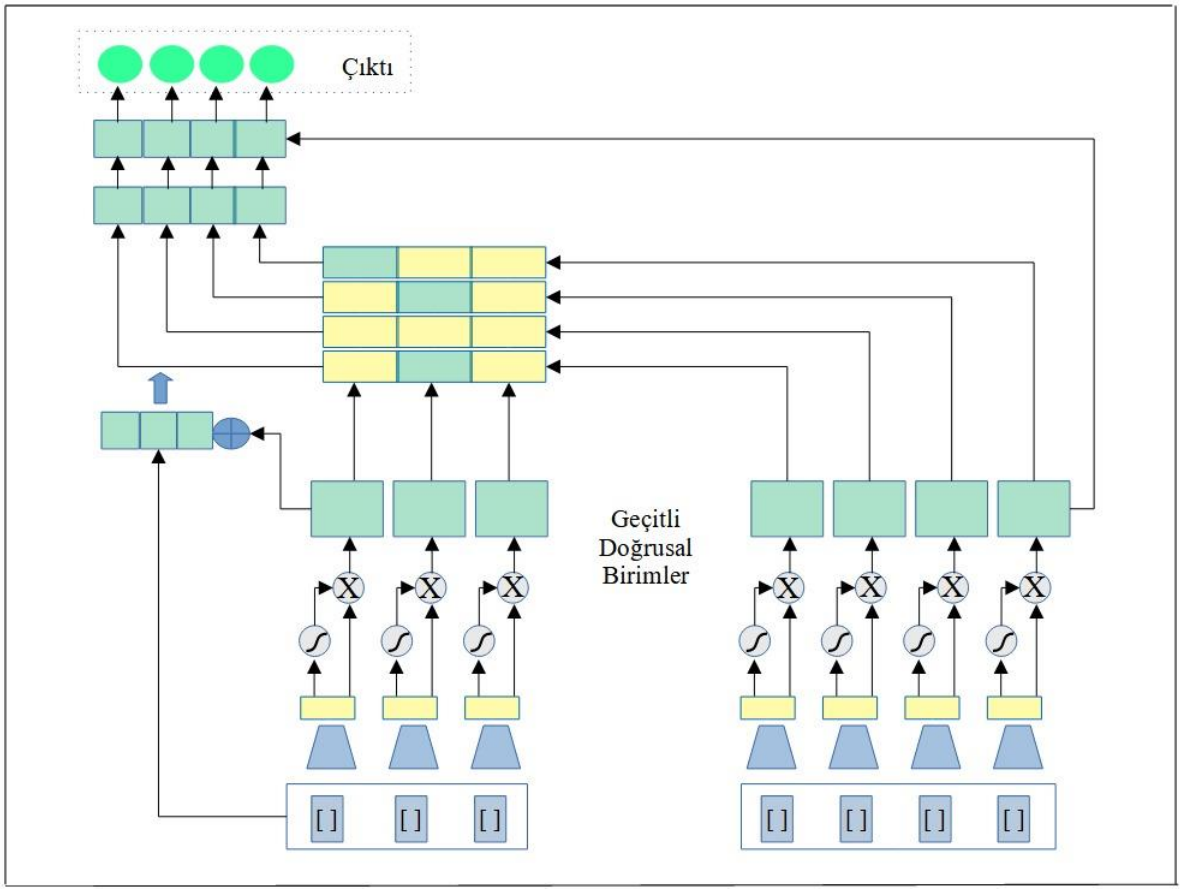
Derin ESA mimarisini uygulamak için evrişim katmanı eklenmiştir. Bu katmana ait gizli katman matrisi Eşitlik 2.17'de görüldüğü gibi hesaplanmaktadır.

$$h_i^l = v\left(W^l\left[h_{i-\frac{k}{2}}^{l-1}, \dots, h_{i+\frac{k}{2}}^{l-1}\right] + b_w^l\right) + h_i^{l-1} \tag{2.17}$$

Çıktı $Y = [A\ \ B] \in R^{2d}$ bir kapılama mekanizmasına sahiptir (Eşitlik 2.18):

$$v([A\ \ B]) = A \otimes \sigma(B) \tag{2.18}$$

Bu eşitlikte $[A\ \ B] \in R^{2d}$ nonlinear modülün girdisi, $\otimes$ elemanların çarpımı ve çıktı $v([A\ \ B]) = R^{2d}$ Y'nin sadece yarısıdır. Kapılar girdi bağımlılığını kontrol etmektedir. Kopya ESA Modeli Şekil 2.10'da görülmektedir.

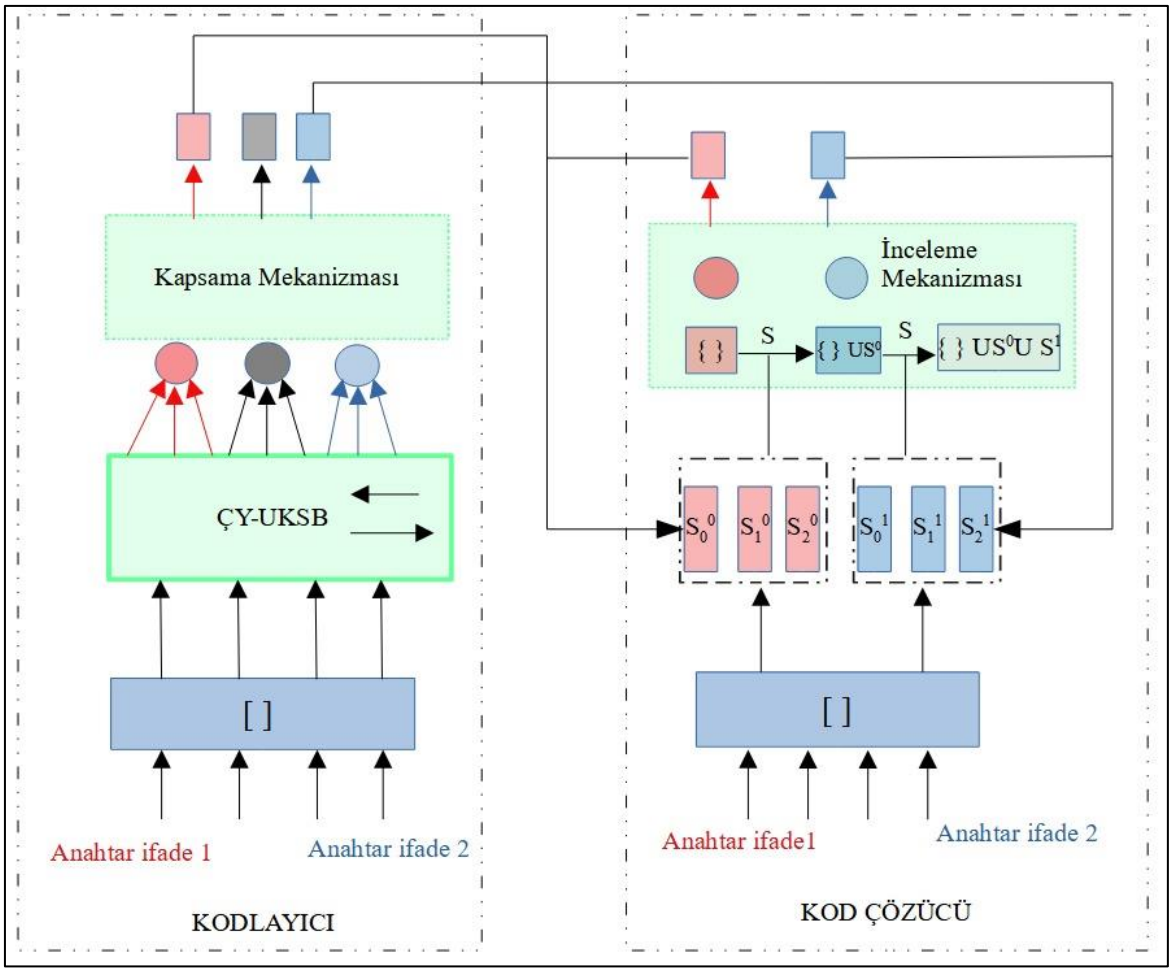


Şekil 2.10. Kopya evrişimsel sinir ağı modeli

Korelasyon ÖSA

Diziden diziye mimari ile anahtar kelimeler çıkarılırken, sonuç çıktılar arasındaki korelasyon ihmal edilir, bu da tekrarlama ve kapsama sorunlarıyla sonuçlanır. Bu sorunları çözmek için, birden fazla anahtar kelime arasındaki korelasyonu iki yolla yakalayan Korelasyon ÖSA (Chen J., Zhang, Wu, Yan ve Li, 2018) önerilmiştir. Şekil 2.11'de Korelasyon ÖSA modeli görülmektedir. Bu mimaride daha önceki anahtar kelime ile kaynak metinin özetlenip özetlenmediğini anlamak için kapsama vektörü kullanılmıştır. Aynı anlam ifade eden ve anahtar sözcüklerden daha fazla bilgi edinilmesini önleyen yinelenen ifadeleri ortadan kaldırmak için inceleme mekanizması kullanılmıştır.

Kapsama vektörü ile dikkat mekanizmasının kaynak metinde bir sonraki odaklanacağı yeri seçme kararı, önceki kararlarının bir hatırlatıcısı tarafından bildirilir, bu da dikkat mekanizmasının kaynak metindeki aynı konumlara daha kolay bir şekilde tekrar tekrar gitmesini önler. Böylece oluşturulan ifadeler kaynak belgedeki diğer konuları kapsar.



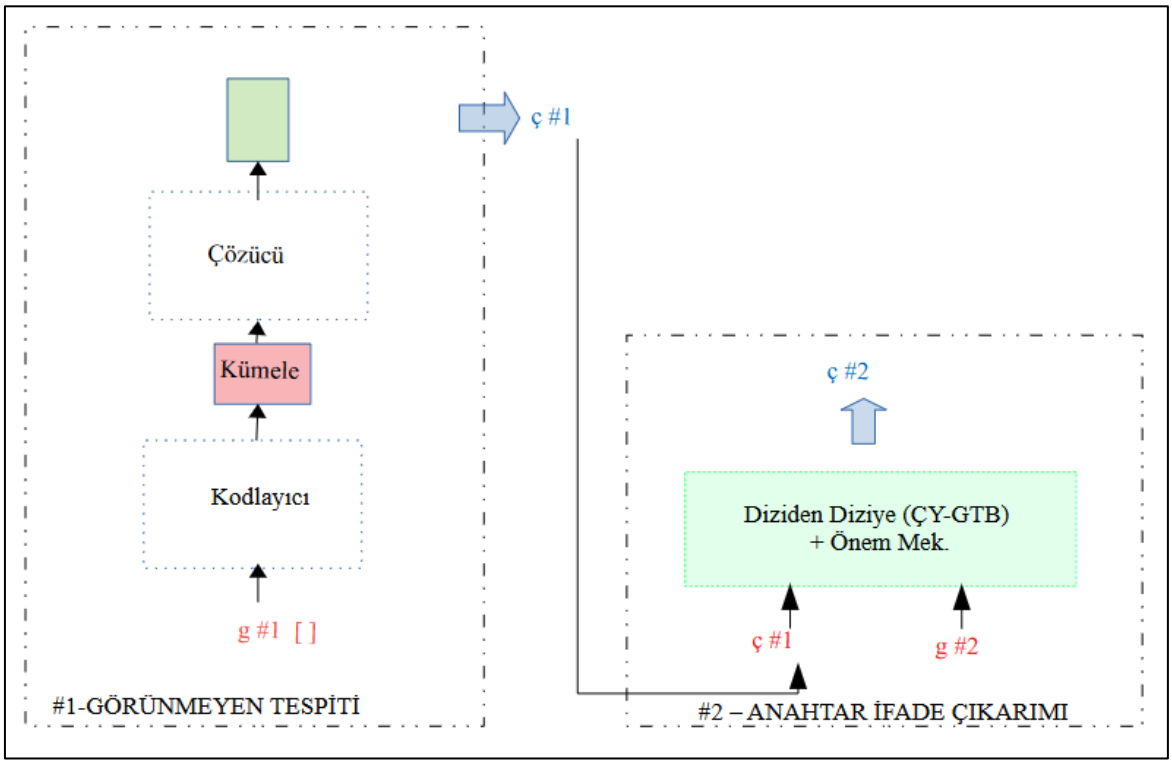
Şekil 2.11. Korelasyon ÖSA modeli

Derin görünmeyen modeli

Derin Görünmeyen metodu (Zahedi, Zahedi ve Fateh, 2019) dışarıdan beslenen (kelime gömmeleri) bilgilere ihtiyaç duymadan, verilen bilgi kaynaklarını önce bir Derin kümeleme ağı aracılığı ile kümeleyerek görünmeyen anahtar kelimelerin çıkarılmasını sağlar.

g#1: #1. bölümdeki ön işleme adımından sonra görünmeyen anahtar kelimeleri de içeren girdi vektörüdür.
g#2: #2. bölümdeki girdi vektörüdür ve görünür anahtar kelimeler ve onların eş anlamlılarını içerir.
ç#1: #1. bölümün çıktısıdır ve görünmeyen anahtar kelimeler ve bunların eş anlamlılarını içerir. #2. bölüme girdi olarak eklenir.
ç#2: Tahmin edilen anahtar ifadeleri de içeren algoritmanın sonuç çıktısıdır.

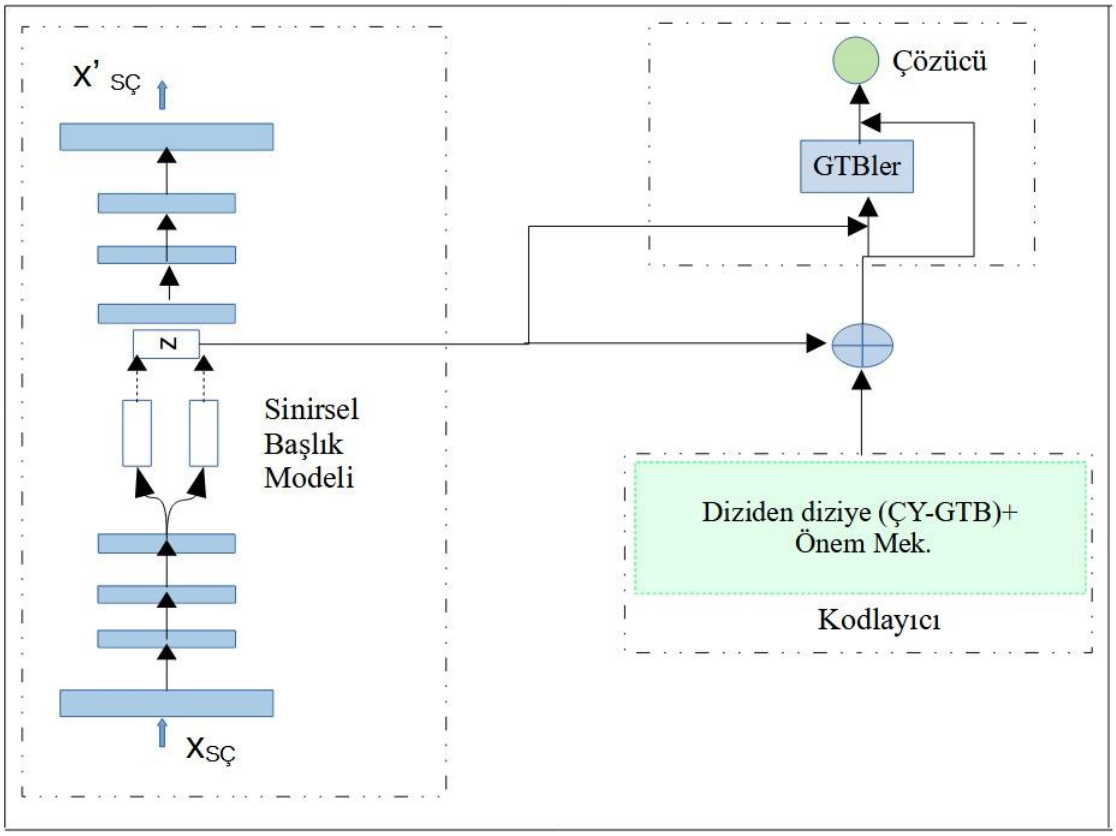
Şekil 2.12'de Derin Görünmeyen mimarisi görülmektedir.



Şekil 2.12. Derin görünmeyen modeli

Başlık-haberdar model

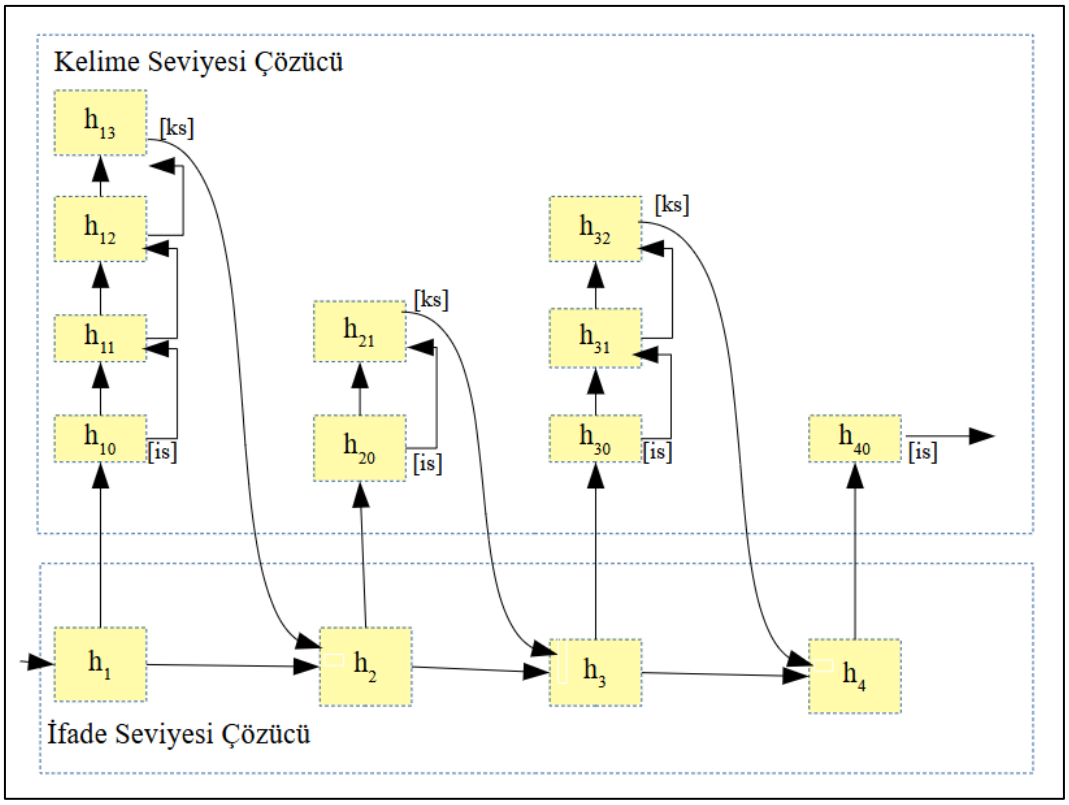
Başlık-haberdar modeli (Wang Y. ve diğerleri, 2019) gizli başlıkları aramak için tasarlanmıştır. Bu model sinir ağı ve diziden diziye tabanlı anahtar kelime üretimi modülü olmak üzere iki modülden oluşmaktadır. Şekil 2.13'te Başlık-haberdar modeli görülmektedir. Başlık için kullanılan sinir ağına x metinleri SÇ oluşturularak girdi olarak verilmektedir. $X_{SÇ}$ girişi önce bir SÇ kodlayıcı tarafından sürekli bir gizli değişken z'ye (x'in konusunu temsil eder) kodlanır. Daha sonra z üzerinde koşullandırılmış SÇ kod çözücü, x'i yeniden yapılandırmaya çalışır ve bir SÇ vektörü $X'_{SÇ}$ çıkarır. Sinirsel başlık modelinde theta, Gaussian softmax tarafından oluşturulur. Diziden diziye ikinci modül ÇY-GTB kodlayıcı ve ileri GTB çözücüden oluşmaktadır. Ayrıca bu modül önem ve kopyalama mekanizmasına sahiptir.



Şekil 2.13. Başlık haberdar model

Derin anahtar kelime oluşturma için özel hiyerarşik kod çözme

Üretim temelli bir model olarak önerilen Derin Anahtar Kelime Oluşturma için Özel Hiyerarşik Kod Çözme (DAKO-ÖHKÇ) (Chen W., Chan, Li ve King, 2020) ÇY-GTB kodlayıcı ve ileri GTB tabanlı bir çözücüden oluşmaktadır. Üretim temelli anahtar kelime çıkarımı yaklaşımları sıralı çözücü kullanıldığından tekrarlayan anahtar kelime üretme eğilimindedirler. DAKO-ÖHKÇ modelinde ise çözücü katmanı İfade Çözücü (İÇ) ve Kelime Çözücü (KÇ) olarak hiyerarşik iki aşamadan oluşmaktadır. Tekrarlama durumunu çözmek için eğitim aşamasında çalışan bir yumuşak çıkarım ve sonuç aşamasında çalışan bir sert çıkarım algoritması mimariye ayrı ayrı uygulanarak DAKO-ÖHKÇ-s ve DAKO-ÖHKÇ-h geliştirilmiştir. Şekil 2.14'te görüldüğü gibi İÇ ve KÇ aşamasında önem mekanizması eklenmiştir. Çözücüler ayrı ayrı is (ifade sonu) ve ks (kelime sonu) ifadeleri gelinceye kadar çalışır.



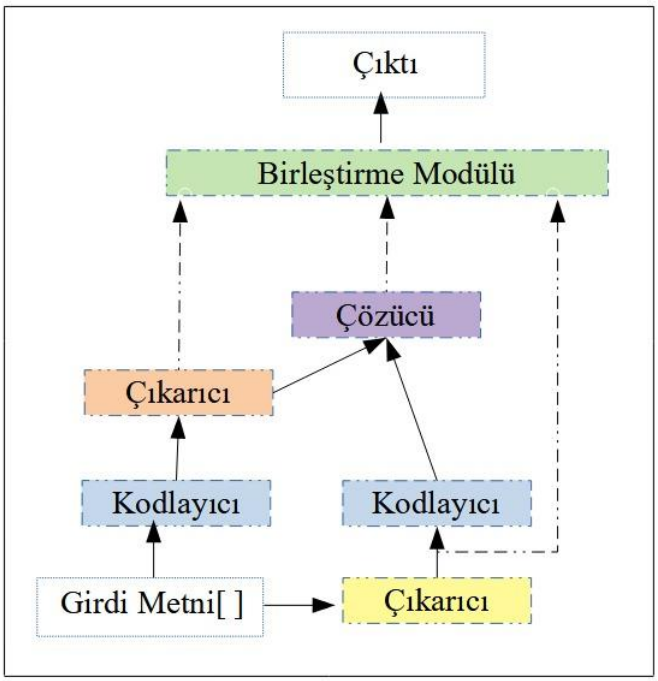
Şekil 2.14. Özel hiyerarşik kod çözücü

2.2.3. Hem üretim hem çıkarım temelli modeller

Literatürde anahtar kelimelerin üretim ve çıkarımını birlikte gerçekleştiren kütüphaneler geliştirilmiştir. Bu bölümde iki görevi birlikte gerçekleştiren modeller ele alınmıştır.

Anahtar ifade üretimi kütüphanesi

Anahtar İfade Üretimi (AİÜ) kütüphanesi (Chen W. ve diğerleri, 2019); iki kodlayıcı, bir çıkarıcı, bir kurtarıcı ve bir birleştirme modülünden oluşmaktadır. Kütüphanenin temel amacı hem metinden çıkarılabilen anahtar ifadeleri (görünen ifadeler) hem de metinden üretilebilen anahtar ifadeleri (görünmeyen ifadeler) bulmaktır. Kurtarıcı modül ilk K anahtar ifadeyi Jaccard benzerliğine göre çıkarır. Bu mimaride kodlayıcılar iki yönlü GTB'dan oluşmaktadır. Çözücü ise ileri GTB tabanlıdır. Şekil 2.15'te AİÜ Kütüphane modeli bulunmaktadır. Kurtarıcı modülü ürettiği çıktıyı ikinci kodlayıcıya girdi olarak vermektedir. Çıkrım ve üretim kayıp değerleri hesaplanıp toplanarak birleştirilmektedir.

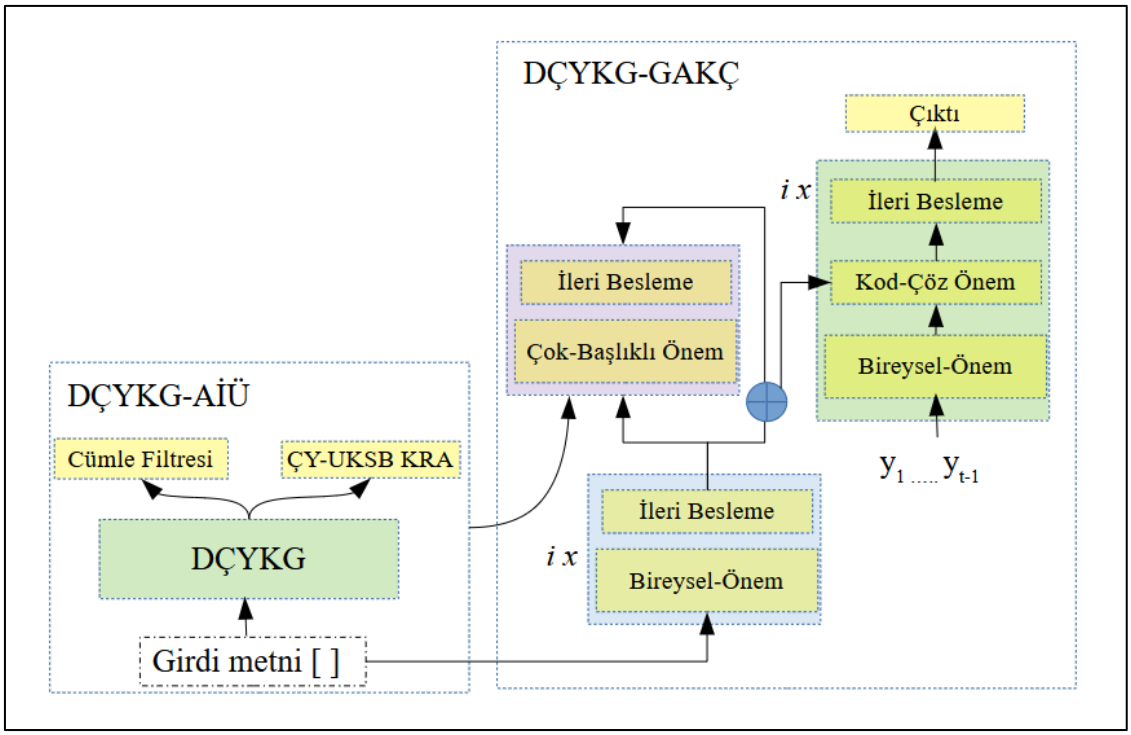


Şekil 2.15. Anahtar ifade üretim kütüphanesi

Görünmeyen anahtar ifade üretici / görünür anahtar kelime çıkarıcı

Üretici ve çıkarıcı anahtar kelime üretimini iki alt göreve ayrılmış tek bir mimari olarak modelleyen ve DÇYKG-Görünmeyen Anahtar İfade Üretici (GAİÜ) ve DÇYKG-Görünür Anahtar Kelime Çıkarıcı (GAKÇ) isimlendiren mimari (Liu R. ve diğerleri, 2020) Şekil 2.16'da görülmektedir. GAİÜ görünmeyen anahtar kelime çıkarımı için, ince ayarlı DÇYKG tarafından GAKÇ'den öğrenilen mevcut anahtar kelime bilgisini entegre eden bir mimaridir. GAİÜ'ye rehberlik etmek için mevcut anahtar sözcükler kullanılmıştır.

GAKÇ modeli ise problemi dizi etiketleme olarak ÇY-UKSB-KRA mimarisi kullanılarak ve daha önceden eğitilmiş DÇYKG modeli kullanılarak çözmeye çalışmaktadır. Performans artırmak için görünen anahtar kelime içermeyen cümleler filtrelenmiştir.



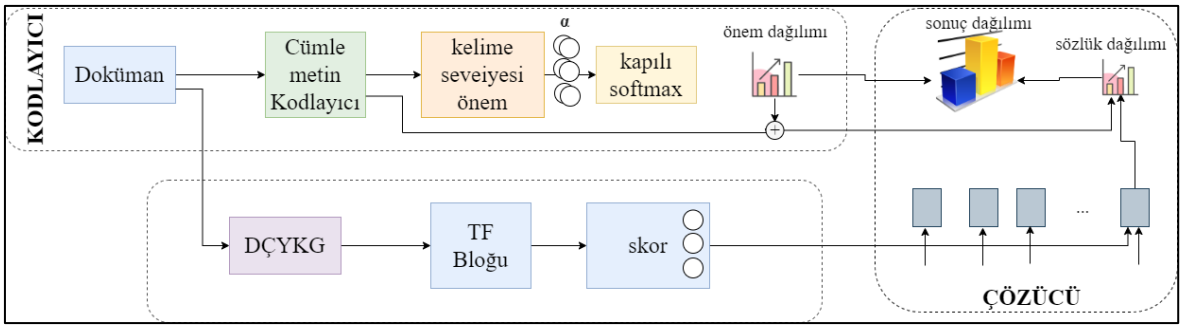
Şekil 2.16. Görünmeyen anahtar ifade üretici / görünür anahtar kelime çıkarıcı

2.2.4. Pekiştirmeli öğrenme temelli modeller

Son yıllarda anahtar kelime çıkarımı algoritmalarının başarımı pekiştirmeli öğrenme yöntemleri kullanılarak arttırılmaya çalışılmıştır. Bu bölümde DÇYKG modelinin performansını artırmak için önerilen iki pekiştirmeli öğrenme yöntemi ele alınmıştır.

Skor ağı

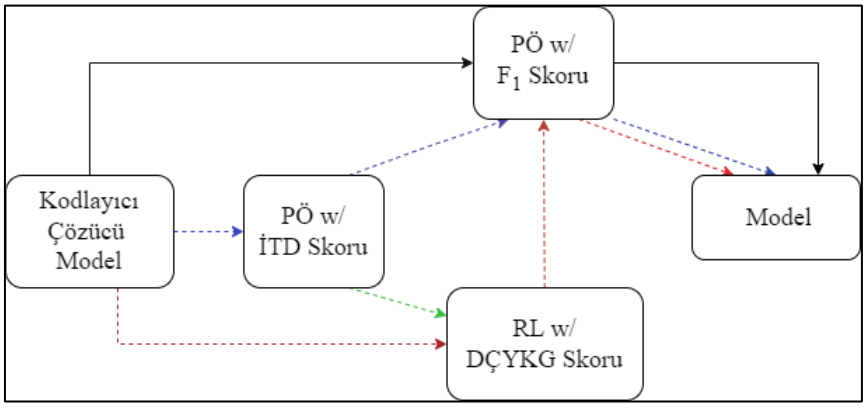
Liu R., Lin, Fu ve Wang (2020), DÇYKG tabanlı bir cümle puanlayıcı ile pekiştirmeli anahtar kelime çıkarma modeli Skor Ağı'nı önermiştir. ÇY-UKSB tabanlı diziden diziye bu modele kopyalama ve önem mekanizması eklenmiştir. Modelin giriş katmanından önce, cümle puanlayıcı bulunmaktadır. Ayrıca, eğitim girdilerinin daha iyi sunulmasını sağlayan bir ifade ön eşleştirmesine dayalı pekiştirmeli öğrenme mekanizması eklenmiştir. Şekil 2.17' de Skor Ağı mimarisi görülmektedir.



Şekil 2.17. Skor ağı modeli

İnce taneli değerlendirme

İnce Taneli Değerlendirme (İTD), Luo, Xu, Ye, Qiu ve Zhang (2021) tarafından önerilen pekiştirmeli öğrenmeye dayalı modeller için yeni bir puanlama yöntemidir. Ayrıca DÇYKG kullanılarak hesaplanan DÇYKG puanı da İTD ile birlikte pekiştirmeli öğrenme katmanında kullanılmıştır. CatSeq modeli temel mimari olarak seçmiştir. Şekil 2.18'de İTD modeli görülmektedir.



Şekil 2.18. İnce taneli değerlendirme

2.3. Sık Kullanılan Veri Kümeleri

Anahtar kelime çıkarma probleminin doğruluğunu ölçmek için literatürde sıklıkla kullanılan ve böylece algoritmaların performansını daha iyi karşılaştırmamızı sağlayan veri kümeleri vardır. Bunlar Inspec, NUS, Document Understanding Conferences (DUC), Krapivin, SemEval, Twitter, Knowledge Discovery and Data Mining (KDD), 500N-KPCrowd, World Wide Web (WWW) ve KP20k veri kümeleridir. Ayrıca, StackExchange ve Weibo gibi

mikrobloglardan derlenen veri kümeleri ve haberler, bilimsel makaleler gibi alanlardan üretilen diğer veri kümeleri bulunmaktadır.

Hulth (2003), 1998'den 2002'ye kadar yayınlanan Bilgi Teknolojisi dergilerinden Inspec veri kümesini derlemiştir. Veriler başlık, özet ve anahtar sözcükler alanlarını içermektedir. Bu veri kümesi 2000 içerikten oluşur. NUS (Nguyen ve Kan, 2007) veri kümesi 211 konferans makalesinden oluşmaktadır. Bu veri kümesi yazarlar ve okuyucular tarafından atanan iki ayrı anahtar kelime kümesine sahiptir. Veri kümesi makalelerin tam metinlerinden oluşmaktadır.

DUC (veya DUC-2001) veri seti (Wan ve Xaio, 2008) 308 haber makalesinden ve 2048 anahtar kelimeden oluşur ve nadiren kullanılır. Veri kümesi, Foreign Broadcast Information Service (FBIS), Los Angeles Times, Financial Times, San Jose Mercury News, Wall Street Journal ve Associated Press haber sitelerinden gelen haber verilerini içerir. Bir diğer haber veri kümesi Marujo ve diğerleri (2013) tarafından derlenen 500N-KPCrowd'dur. Bu veri kümesi, her kategoride 50 belge ile 10 farklı kategoride (sanat ve kültür, iş, suç, moda, sağlık, siyaset, siyaset dünyası, bilim, spor, teknoloji) 500 İngilizce haber içermektedir.

Meta verilerle birlikte tam metinleri içeren Krapivin veri seti (Krapivin, Autaeu, Marchese, Blanzieri ve Segata 2010) ACM tarafından 2003'ten 2005'e kadar yayınlanan toplam 2304 Bilgisayar Bilimleri makalesinden oluşmaktadır. Tüm anahtar sözcükler yazar tarafından seçilmiştir. SemEval-2010 (Kim ve diğerleri, 2013) ise 288 makaleden oluşan küçük bir veri setidir. Bildiriler yine ACM dijital kütüphanesinden derlenmiştir. Her makale altı ila sekiz sayfadan oluşmaktadır. Makaleler, dağıtık sistemler, çok ajanlı sistemler, bilgi çıkarımı ve ekonomi dâhil olmak üzere çeşitli alanlara aittir.

Zhang Q. ve diğerleri (2016) yaklaşık 147 bin tweet ve etiketten oluşan bir veri seti oluşturmuştur. Yazarlar, hashtag tahmini için bu veri kümesini derlemiştir. Veri kümesinde, bir metin parçası için yalnızca bir anahtar sözcük veya anahtar ifade içermektedir. Florescu ve Caragea (2017), ACM'nin KDD ve WWW'yi içeren iki üst düzey makine öğrenimi konferansından konferans makaleleri içeren iki veri seti derlemiştir. KDD 755 meta veriden, WWW ise 1330 meta veriden oluşmaktadır.

Literatürde en çok kullanılan veri seti (Meng ve diğerleri, 2017) KP20k'dir. 567830 Bu veri kümesi eğitim için 527830, doğrulama için 20K ve test için 20K Bilgisayar Bilimleri alanına ait makaleden oluşmaktadır. Makaleler Science Direct, Wiley, ACM Digital Library ve Web of Science'dan toplanmıştır. StackExchange veri kümesi ise Yuan ve diğerleri (2018) tarafından oluşturulmuş ve kullanılmıştır. Alfarra ve Alfarra (2018), University College of Science and Technology (UCST) tarafından sağlanan yeni bir haber veri seti oluşturmuştur. UCST korpusunda (330) İngilizce metin makalesi, (106) haber makalesi, (85) duyuru makalesi, (98) program açıklama makalesi ve (41) etkinlik açıklaması makalesi bulunmaktadır.

2.4. Denetimli Modellerin Mimari Karşılaştırması

Literatürde anahtar kelime çıkarımı için önerilen denetimli anahtar kelime çıkarımı modelleri iki alt başlıkta toplanmıştır: çıkarım temelli yöntemler ve üretim temelli yöntemler. Çıkarım yöntemleri yalnızca giriş metninde bulunan sözcükleri anahtar sözcük olarak etiketlerken, üretici yöntemler hem giriş metnindeki anahtar sözcükleri hem de metnin anlamından türetilebilecek anahtar sözcükleri üretir. Anahtar kelime çıkarımı için önerilen denetimli modeller Çizelge 2.1'de derlenmiştir.

Çizelgede referanslar, mimari adı, yıl ve dil meta veri sütunları vardır. Bu sütunlar model hakkında temel bilgiler verir. Çizelgede bulunan tüm algoritmalar İngilizce dili için eğitilmiş ve test edilmiştir. Veri kümesi sütunu, mimariyi test etmek için yaygın olarak kullanılan veri kümesi adlarını içerir. Kelime gömmesi sütununda modelin hangi kelime gömmesini kullandığı not edilmiştir.

Çizelgenin Algoritma sütununda, geliştirilen yöntemin sınıflandırmaya mı, dizi etiketlemeye mi yoksa diziden diziye'e mi dayalı olduğu belirtilmiştir. Ayrıca bu sütun modelde dikkat mekanizması, kapsama mekanizması, kopyalama mekanizması gibi alt mekanizmalardan hangilerinin kullanıldığını içermektedir. Alt mimari sütununda, ilgili model için hangi alt mimarilerin kullanıldığı bulunmaktadır.

Çizelge 2.1'de, son on yılda önerilmiş çıkarımsal ve üretici tabanlı 19 mimari incelenmiştir. 2016 yılından itibaren ilk zamanlarda geliştirilen modellerde ÖSA'nın sıklıkla kullanıldığı görülmektedir. ÖSA mimarisinden kaynaklanan sorunları çözmek için geliştirilen UKSB ve

GTB, 2017 yılından itibaren ÖSA modülü yerine tercih edildiği görülmektedir. Kelime gömmesi olarak ilk yıllarda ağırlıkla KV ve GloVe mimariye dâhil edilirken, DÇYKG ve türevlerinin son yıllarda daha sıklıkla kullanıldığı açıkça anlaşılmaktadır.

Çizelge 2.1. Denetimli anahtar kelime çıkarımı modellerinin mimari karşılaştırılması

Referans	Mimari Adı	Yıl	Veri kümesi	Algoritma	Alt mimari	Çıkarım Üretim	Gömme	Dil
Zhang ve diğ.	Kopya ESA	2017	Inspec, Krapivin, NUS, KP20k, SemEval-2010	Diziden Diziye + Kopyalama Mek. + Önem Mek.	ESA	Üretim	-	İngilizce
Basaldella ve diğ.	ÇY-UKSB	2018	Inspec	Dizi Etiketleme	ÇY-UKSB	Çıkarım	GloVe	İngilizce
Chen ve diğ.	Korelasyon ÖSA	2018	Krapivin, NUS, SemEval-2010	Diziden Diziye + Korelasyon Mek. + Önizleme Mek.	Kopya ÖSA	Üretim	-	İngilizce
Zahedi ve diğ.	Derin Görünmeyen	2019	Inspec, Krapivin, NUS, SemEval-2010	Diziden Diziye +Önem Mek.	İleri GTB Kodlayıcı Çözücü	Üretim	-	İngilizce
Alzaidy ve diğ.	ÇY-UKSB-KRA	2019	KP20k, WWW, KDD	Dizi Etiketleme	ÇY-UKSB +KRA	Çıkarım	GloVe	İngilizce
Chen ve diğ.	AİÜ Kütüphanesi	2019	Inspec, Krapivin, NUS, KP20k, SemEval-2010	Diziden diziye + Çıkarıcı ve Kurtarıcı	ÇY-GTB, Jaggard Benzerliği	Çıkarım Üretim	-	İngilizce
Sun ve diğ.	ÇGİ	2019	Inspec, Krapivin, NUS, KP20k, SemEval-2010	Diziden diziye + Bağlam Değiş. Mek. + Kapsama Mek.	İşaretçi ağ kod çözücü, GEA kodlayıcı	Çıkarım	hızlı metin	İngilizce
Wang ve diğ.	BaşlıkHhaberdar	2019	Twitter, Weibo, StackExchange	Diziden diziye + Önem Mek. + Kopyalama Mek.	ÇY-GTB +Başlık Dikkatli Kodlayıcı	Üretim	-	İngilizce/ Çince
Lui ve diğ.	DÇYKG–AKE DÇYKG-PKE	2020	KP20k, NUS, Krapivin	Diziden Diziye Dizi Etiketleme	ÇY-UKSB-KRA Dönüştürücüler	Üretim Çıkarım	DÇYKG	İngilizce
Sahrawat ve diğ.	BilDÇYKG ile ÇY-UKSB-KRA	2020	Inspec, SemEval-2010	Dizi Etiketleme	ÇY-UKSB-KRA	Çıkarım	BilDÇYKG	İngilizce
Chen ve diğ.	DAKO-ÖHKÇ-s DAKO-ÖHKÇ-h	2020	Inspec, Krapivin, KP20k, SemEval-2010	Diziden Diziye	ÇY-GTB+ Hiyerarşik Çözücü	Üretim	-	İngilizce

Çizelge 2.1. (devam) Denetimli anahtar kelime çıkarımı modellerinin mimari karşılaştırılması

Referans	Mimari Adı	Yıl	Veri kümesi	Algoritma	Alt mimari	Çıkarım Üretim	Gömme	Dil
Duari ve Bhatnagar	KA-NB KA-AGA	2020	Krapivin, SemEval-2010, Inspec, WWW, KDD	Sınıflandırma	Naive Bayes AGA	Çıkarım	-	Dil Bağımsız
Santosh ve diğ.	DD-AKÇ	2020	KP20k	Dizi Etiketleme + Belge Düzeyi Önem Mek.	ÇY-UKSB-KRA	Çıkarım	DÇYKG	İngilizce
Martinc ve diğ.	DTS-AKE	2020	Inspec, Krapivin, NUS, KP20k, SemEval-2010, DUC, diğer	Dizi Etiketleme + Önem Mek. + Silme Mek.	Transformers + UKSB	Çıkarım	DÇYKG	İngilizce
Liu ve diğ.	Skor Ağı	2020	Inspec, Krapivin, NUS, SemEval-2010, KP20K	Diziden Diziye + DÇYKG Cümle Puanlama	ÇY-UKSB	Üretim	DÇYKG	İngilizce
Gero ve Ho	KMSG	2021	Inspec, Pubmed	Dizi Etiketleme	ÇY-UKSB-KRA + Kelime Merkeziliği	Çıkarım	DÇYKG BioDÇYKG	İngilizce
Luo ve diğ.	İTD	2021	Inspec, Krapivin, KP20k	Diziden Diziye + 2RL	catSeq + FG +FB	Üretim	DÇYKG	İngilizce
Nikzad-Khasmakhi ve diğ.	İfade Formatlayıcı	2021	Inspec, SemEval-2010, SemEval-2017	Dizi Etiketleme + Kelime Grafiği	İleri Besleme Ağı	Çıkarım	DÇYKG	İngilizce

2.5. Sonuç Değerlendirme Yöntemleri

Anahtar kelime çıkarımı algoritmalarının değerlendirilmesinde F-skor ölçütü kullanılmaktadır. Bu ölçütün hesaplanmasında tahmin edilen değerlerin gerçek değeri / tahmin edilen değeri sayılarına bakılmaktadır. F-skor haricinde Kesinlik ve Duyarlılık ölçüm parametreleri kullanılmaktadır.

Kesinlik değeri algoritma tarafından pozitif olarak doğru bir şekilde etiketlenmiş çıktıların (DoğruPozitif), algoritma tarafından anahtar kelime olarak etiketlenen çıktı sayısına (DoğruPozitif + YanlışPozitif) bölünmesi ile bulunmaktadır. Duyarlılık değeri ise doğru bir şekilde pozitif olarak etiketlenmiş çıktıların (DoğruPozitif) gerçekte anahtar kelime kümesine (DoğruPozitif + YanlışNegatif) bölümü ile bulunmaktadır. F-skor değeri, Kesinlik ve Duyarlılık değerlerinin β parametresine bağlı olarak Eşitlik 2.19'daki gibi hesaplanmaktadır.

$$F - skor = (1 + \beta^2) \frac{Kesinlik + Duyarlılık}{\beta^2 * Kesinlik * Duyarlılık} \quad (2.19)$$

F_1-skor Eşitlik 2.19'da görülen $\beta^2 = 1$ için Kesinlik ve Duyarlılık değerlerinin harmonik ortalaması ile hesaplanmaktadır. Modelin performansını ölçmek için çıkarılan k adet anahtar kelime için F_1@k Eşitlik 2.20'de görüldüğü gibi hesaplanmaktadır. Bu skor Kesinlik@k ve Duyarlılık@k değerlerinin harmonik ortalaması ile hesaplanmaktadır. Hesaplamalar yapılırken sadece k veya k'dan çok anahtar kelime gözönünde bulundurulmaktadır.

$$F_1@k = 2 * \frac{Kesinlik@k + Duyarlılık@k}{Kesinlik@k * Duyarlılık@k} (\beta=1) \quad (2.20)$$

2.6. Denetimli Modellerin Performans Sonuçları

Literatürde anahtar kelime çıkarımı için önerilen denetimli modellerin farklı metrikler kullanılarak değerlendirilmesi ile elde edilen sonuçlar Çizelge 2.2'de derlenmiştir. Bu çizelgede modellerin en sık kullanılan veri setleri için F_1, F_1@5, F_1@10 ve F_1@M ölçümleri için hesaplanan performans değerleri yer almaktadır. Bazı modeller için F_1@k değeri k = 4 ve k = 8 için hesaplanmış ve sonuçları F_1@5 ve F_1@10'a çok yakın olduğu için aynı sütun

altına yazılmıştır. Önerilen bazı algoritmaların performansını ölçmek için F_1@M skoru kullanılmıştır.

Denetimli modeller için performans sonuçları incelendiğinde, algoritmaların farklı veri kümeleri için farklı başarı puanları elde ettiği görülmektedir. Örneğin, Inspec veri setinde F_1@5 ölçümünde en iyi sonuçları veren KA-NB, KA-AGA mimarileri, Krapivin ve SemEval veri setlerinde düşük performans göstermiştir. ÇGİ algoritması, SemEval, NUS, Krapivin veri kümeleri için F_1@5 ve F_1@10 metrikleri için en yüksek performans değerlerine ulaşmıştır. DÇYKG-PKE mimarisi, F1@M ölçümü için Krapivin ve KP20k veri kümelerinde çok iyi sonuçlar elde etmiştir. KA-AGA çıkarım algoritması Inspec veri kümesi için en yüksek skoru üretmiştir. ÇY-UKSB-KRA modeli Inspec veri kümesi için GloVe gömmeleri ile 0,457 F_1-skor elde ederken, BilDÇYKG kullanılarak geliştirilen BilDÇYKG ile ÇY-UKSB-KRA modeli 0,593 F_1-skoruna ulaşmıştır. F_1@M ile F_1@5 veya F_1@10'u tam olarak karşılaştırmak mümkün değildir fakat mimarinin performansı hakkında genel bilgi vermesi için tabloya eklenmiştir.

Çizelge 2.3’te literatürde önerilen denetimli modellerin görünmeyen anahtar kelimeleri üretme performans sonuçları görülmektedir. Performans sonuçları R@10, R@50 ve F_1@M metriklerine göre raporlanmıştır. DÇYKG-AKG modeli NUS ve KP20k verikümeleri için en yüksek skorları elde etmiştir. Inspec, Krapivin ve Semeval-2010 verikümeleri için Skor Ağı modeli en yüksek sonuçlara ulaşmıştır.

Çizelge 2.2. Denetimli modeller için performans sonuçları

Model	Inspec			Krapivin			NUS			SemEval-2010			KP20k		
	F_1@10	F_1@5	F_1	F_1@10	F_1@5	F_1	F_1@10	F_1@5	F_1	F_1@10	F_1@5	F_1	F_1@10	F_1@5	F_1
Kopya ESA	0,346	0,285	-	0,272	0,314	-	0,33	0,34	-	0,308	0,295	-	0,288	0,351	-
ÇY-UKSB	0,422	-	0,428	-	-	-	-	-	-	-	-	-	-	-	0,167
Korelasyon ÖSA	0,279	-	-	0,278	0,318	-	0,33	0,36	-	0,32	0,32	-	0,218	-	-
Derin Görünmeyen	-	-	0,384	-	-	0,329	-	-	0,364	-	-	0,289	-	-	-
ÇY-UKSB-KRA	-	-	0,457	0,316@M	-	-	0,351@M	-	-	-	-	0,111	-	-	0,356
AİÜ Kütüphanesi	0,284	0,257	-	0,25	0,272	-	0,286	0,29	-	0,223	0,202	-	0,282	0,317	-
ÇGİ	0,417	0,386	-	0,297	0,363	-	0,402	0,46	-	0,389	0,401	-	0,292	0,368	-
DÇYKG-PKE	-	-	-	0,407@M	-	-	-	-	-	-	-	-	0,437@M		-
BilDÇYKG ile ÇY-UKSB-KRA	-	-	0,593	-	-	-	-	-	-	-	-	0,357	-	-	-
DAKO-ÖHKÇ-s	0,278@M	0,235	-	0,338@M	0,278	-	-	-	-	0,322@M	0,276	-	0,372@M	0,307	-
DAKO-ÖHKÇ-h	0,291@M	0,253	-	0,347@M	0,286	-	-	-	-	0,335@M	0,284	-	0,374@M	0,311	-
KA- AGA	0,547	-	0,607	-	0,277	0,309	-	-	-	0,303	-	0,464	-	-	-
KA-NB	0,511	-	0,501	-	0,267	0,341	-	-	-	0,283	-	0,395		--	-

Çizelge 2.2. (devam) Denetimli modellerin görünmeyen anahtar kelime üretme performansı

Model	Inspec			Krapivin			NUS			SemEval-2010			KP20k		
	F_1@10	F_1@5	F_1	F_1@10	F_1@5	F_1	F_1@10	F_1@5	F_1	F_1@10	F_1@5	F_1	F_1@10	F_1@5	F_1
DD-AKÇ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0,423
DTS-AKE	0,504	0,423	-	0,269	0,276	-	0,345	0,33	-	0,341	0,301	-	0,287	0,288	-
Skor Ağı	0,395	0,365	-	0,324	0,379	-	0,385	0,43	-	0,348	0,353	-	0,345	0,415	-
KMSG	-	-	0,586	-	-	-	-	-	-	-	-	-	-	-	-
İTD	0,308@M	0,266	-	0,375@M	0,304	-	-	-	-	-	-	-	0,391@M	0,327	-
İfade Formatlayıcı	-	-	0,698	-	-	-	-	-	-	-	-	0,486	-	-	-

Çizelge 2.3. Denetimli modellerin görünmeyen anahtar kelime üretme performansı

Model	Inspec			Krapivin			NUS			SemEval-2010			KP20k		
	F1@M	R@50	R@10	F1@M	R@50	R@10	F1@M	R@50	R@10	F1@M	R@50	R@10	F1@M	R@50	R@10
Kopya ESA	-	0,107	0,05	-	0,205	0,119	-	0,12	0,062	-	0,074	0,044	-	0,225	0,147
Korelasyon ÖSA	-	-	-	-	-	0,108	-	-	0,059	-	-	0,041	-	-	-
Derin Görünmeyen	-	0,11	0,051	-	0,259	0,128	-	0,173	0,093	-	0,063	0,041	-	-	-
AİÜ Kütüphanesi	-	-	0,0455	-	0,252	0,1355	-	0,19	0,1005	-	-	0,0405	-	0,249	0,135
Başlık Haberdar	-			-			-			-	-	-	-		
DÇYKG-AKG	-	-	-	-	0,268	-	-	0,219	-	-	-	-	-	0,282	-
DAKO-ÖHKÇ-s	0,021	-	-	0,033	-	-	-	-	-	0,016	-	-	0,029	-	-
DAKO-ÖHKÇ-h	0,022	-	-	0,043	-	-	-	-	-	0,017	-	-	0,032	-	-
Skor Ağı	-	0,13	-	-	0,276	-	-	0,201	-	-	0,088	-	-	0,272	-

2.7. Denetimsiz Anahtar Kelime Çıkarımı

Anahtar kelime çıkarımı için geliştirilen denetimsiz yöntemler etiketli bir veri kümesine ihtiyaç duymamaktadır. Bu yaklaşımlar problemi bir kümeleme veya gruplandırma problemi olarak ele almaktadır. Literatürde önerilen denetimsiz modeller modelin kullandığı özniteliklere göre isimlendirilmiş 3 grup altında toplanmaktadır. Bunlar:

- İstatistiksel tabanlı
- Grafik tabanlı
- Gömme tabanlı modellerdir.

2.7.1. İstatistiksel tabanlı modeller

Anahtar kelime çıkarımının yapıldığı ilk yıllarda TF, TDF gibi bilgiler kullanılarak tek bir doküman üzerinde gerçekleştirilmiştir. İlerleyen yıllarda TF ve TDF yöntemleri birleştitilerek yeni bir model kullanılmaya başlanmıştır (Ramos, 2003). Daha yakın yıllarda istatistiksel ve c-değer yöntemi kullanılarak anahtar kelime çıkarımı gerçekleştirilmiştir (Yeom Ko ve Seo, 2019). Bu bölümde literatürde geliştirilen istatistiksel yöntemler sırası ile ele alınmıştır.

Hızlı otomatik anahtar kelime çıkarımı

Hızlı Otomatik Anahtar Kelime Çıkarımı (HO-AKÇ/Rapid Automatic Keyword Extraction-RAKE) (Rose, Engel, Cramer ve Cowley, 2010), otomatik anahtar kelime elde etmek için tasarlanan denetimsiz bir algoritmadır. HO-AKÇ algoritması cümleleri sıralamak için bir kelimenin hem sıklığını hem de derecesini kullanmıştır. Ardından, eşik fonksiyonuna göre anahtar kelimeleri seçmeden önce aday anahtar kelimelerdeki belirli metrikleri hesaplanmıştır. Bu modelin avantajı dilden bağımsız ve alandan bağımsız olmasıdır. HO-AKÇ için girdi parametreleri durdurma kelimeleri ‘ve’, ‘veya’, ‘de’ gibi bir grup öbek sınırlayıcı ve bir dizi kelime sınırlayıcı içerir. İki durma listesi kelimesi ve/veya noktalama işareti arasında görülen herhangi bir kelime dizisi aday anahtar kelimeler olarak işaretlenir. Daha sonra aday anahtar kelimeler listesindeki her kelimenin sıklığı ve derece değerleri hesaplanır. Burada bir kelimenin derecesi, aday anahtar kelimeler listesinde göründüğü toplam kelime sayısı olarak kabul edilir. Sonra her kelimeye frekans üzerinde bir derece

puanı atanır ve her aday anahtar kelimenin birikimli puanı, içerdiği kelimelerin puanlarının toplanmasıyla hesaplanır. En son adımda en yüksek puan alan aday anahtar kelimeler, anahtar kelime olarak çıkarılır. Özellikle, HO-AKÇ, terim dizilerini aday anahtar kelimeler olarak görür ve buna dayanarak bir terimler eşleşmesi matrisi oluşturur. Daha sonra, her aday anahtar kelime için, matristeki terimlerin derecesi ve sıklığına bağlı olarak üye kelime puanlarının toplamı olarak hesaplanır (Eşitlik 2.21):

$$S(w) = \frac{der(w)}{frek(w)} \tag{2.21}$$

Eşitlik 2.21'de *S(w)* üye kelimeleri, *frek(w)* terim sıklığını, *der(w)* terim derecesini ifade etmektedir. Sonraki adımda her aday anahtar kelime için bir puan hesaplanır ve üye kelime puanlarının toplamı olarak tanımlanır (Tomokiyo ve Hurst, 2003).

Sandul ve Mikhailova (2018) Rusça metinlerde Anahtar kelime çıkartma problemini çözmek için HO-AKÇ algoritmasının ve istatistiksel yaklaşımın etkinliği incelenmektedir. Ek olarak, makalede, HO-AKÇ algoritması sonucunda elde edilen ifadelerin ağırlıklandırılması için için Γ-endeksi kullanan yeni Hibrit yöntemi önermişlerdir.

Bir diğer anahtar kelime çıkarımı

İstatistiksel özelliklere dayanan tek bir belgeden anahtar kelimeleri (hem tek kelimeli hem de çok kelimeli terimler) çıkarımı için bir denetimsiz algoritmadır. Campos ve diğerleri (2018) çalışmalarında anahtar kelimeleri tanımlamak ve sıralamak için tek bir belgeden çıkarılan metin istatistiksel özelliklerine dayanan Bir Diğer Anahtar Kelime Çıkarımı (BD-AKÇ) algoritmasını önermişlerdir.

Model her anahtar kelimeye tek bir puan atamak için sezgisel olarak birleştirilen anahtar kelime özelliklerini yakalayan bir dizi öznitelik tanımlar. Bu öznitelikler arasında kasa, konum, frekans, bağlama ilişkin ilişki ve belirli bir terimin dağılımı yer alır. BD-AKÇ dört ana aşamadan oluşmaktadır:

1. Metin ön işleme ve aday terim tanımlaması.

2. Kelime konumu, kelime sıklığı gibi istatistiksel özellik kümesi kullanılarak her bir terimin özelliklerinin çıkarılması.
3. Sezgisel olarak tüm bu özelliklerin tek bir ölçüde birleştirilmesi ve her aday anahtar kelimeye bir S(w) ağırlık atanması. S (w) değeri ne kadar küçük olursa, (w) kelimesi o kadar önemli olmaktadır. Bu ağırlık, bir sonraki bölümde anahtar kelime oluşturma sürecini besleyecektir (Eşitlik 2.22):

$$S(w) = \frac{W_{İlgi} * W_{Pozisyon}}{W_{BüyükHarf} + \frac{W_{Frek}}{W_{İlgi}} + \frac{W_{CümleFark}}{W_{İlgi}}} \quad (2.22)$$

4) Aday anahtar kelimeleri n-gram yapım metodolojisi yoluyla oluşturur ve önemlerine göre puanlar atamaktadır (Eşitlik 2.23):

$$S(kw) = \frac{\prod_{w \in kw} S(w)}{TF(kw) * (1 + \sum_{w \in kw} S(w))} \quad (2.23)$$

Bu eşitlikte *S (kw)*, en fazla 3 terim büyüklüğüne sahip bir aday anahtar kelimenin puanıdır ve bu aday anahtar kelimenin ilk teriminin puanının, kalan terimlerin sonraki puanlarıyla çarpılmasıyla belirlenir. Son olarak oluşturulan potansiyel olarak alakalı anahtar kelimelerin bir listesi çıkarılır, böylece *S(kw)* puanı ne kadar düşük olursa anahtar kelime o kadar önemli olmaktadır (Bojanowski ve diğerleri, 2017).

5) Son adımda verileri sıralama ve tekilleştirme uygulanarak benzer anahtar kelimeler karşılaştırılır. Son anahtar kelimeler listesi ilgi düzeylerine göre sıralanır.

2.7.2. Grafik tabanlı modeller

Anahtar kelime çıkarımında kullanılan istatistiki öznitelikler yoluyla istatistiksel bazı bilgiler elde edilmektedir. Fakat bu modeller kelimelerin birbirleri ile ve kelimelerin cümlelerle ilişkisi hakkında bir bilgi içermez. Girdi metninden çizilen “kelime grafiği” kullanılarak çıkarılan grafik-tabanlı öznitelikler kelimelerin birbirleyle ve cümle ile bağlantılarını tanımlamaktadır. Grafik tabanlı denetimsiz modeller bu öznitelikleri kullanarak çıkarılan aday ifadeleri sıralamaktadır.

Sayfa sıralama

Sayfa Sıralama (PageRank) (Brin ve Page, 1998), Google arama motorunun temel bileşeni olan ve web sayfalarının sıralamasını hesaplamak üzere geliştirilmiş bir algoritmadır. Tüm web sayfalarını büyük bir yönlendirilmiş grafik olarak tanımlamaktadır. Bu grafikte, bir düğüm bir web sayfasıdır. A web sayfasında B web sayfasına bağlantı varsa, A'dan B'ye yönlendirilmiş bir kenar olarak gösterilebilir. Grafiğin tamamını oluşturduktan sonra, aşağıdaki formüle göre web sayfaları için ağırlıklar atanmıştır. Ağırlıklandırılmış tanımlamada, vi ∈ V düğümüne ilişkin Sayfa Sıralama derecesi Eşitlik 2.24 ile verilmektedir (Zhang Y. ve diğerleri, 2019):

$$S(V_i) = (1-d) + d * \sum_{j \in G(v_i)} \frac{1}{|Ç(V_j)|} S(V_j) \qquad (2.24)$$

Bu eşitlikte $S(V_i)$ bir web sayfasının ağırlığını, d damp faktörünü, G V_i i'ye gelen linkleri, $Ç(V_j)$ j'den giden linkleri ve $|Ç(V_j)|$ giden bağlantıların sayısını ifade etmektedir. En iyi değerleri alan düğüm noktaları merkez düğüm olarak seçilmektedir.

Wang, Liu ve Wang (2007) çalışmalarında metni WordNet'ten gelen synset ile ağırlıklı semantik grafik olarak temsil etmektedir. Daha sonra kelime anlamında anlam ayrımı yapmak için Sayfa Sıralama tabanlı metinden anahtar kelimeler çıkaran yeni bir algoritma sunmuşlardır. Deneysel sonuçlar önerilen algoritmanın pratik ve etkili olduğunu göstermektedir. Başka bir çalışmada Liu ve diğerleri (2010) otomatik anahtar kelime çıkarımı için Eş anlamlı Sayfa Sıralama algoritması kullanılmıştır. İlk olarak tek bir belgedeki içerik ağırlıklı eşanlamlı ortak oluşum ağı olarak temsil edilir daha sonra Sayfa Sıralama algoritması bu eşanlamlı ağda her eşanlamlı grup için sıralama yapmak için kullanılır.

Başlıklı sayfa sıralama

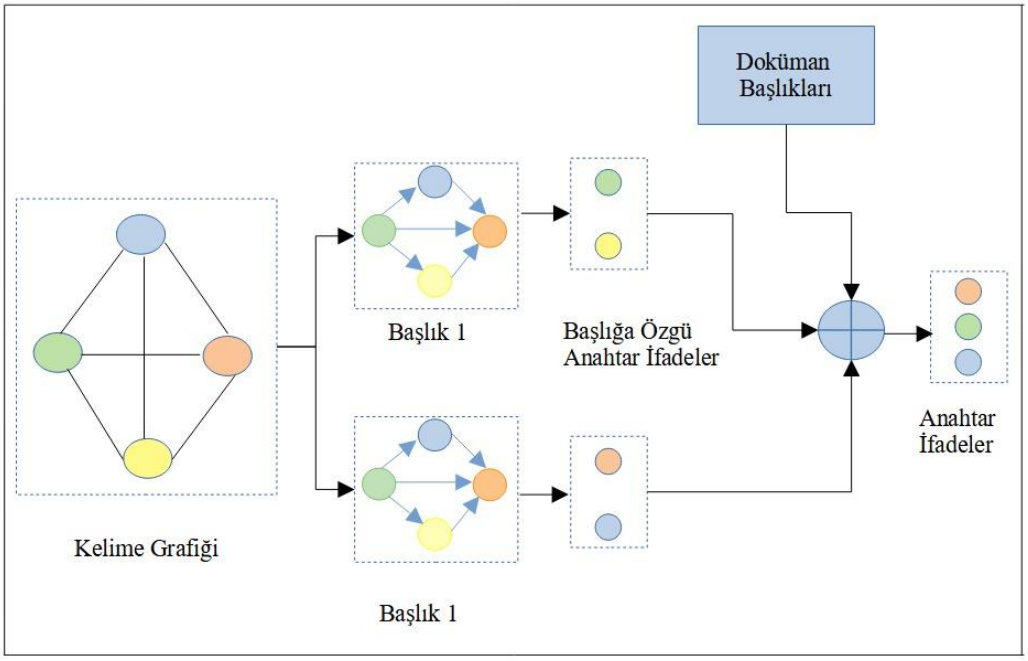
BSS konuya duyarlı puan yayılımı tanıtarak Liu ve diğerleri (2010) tarafından anahtar kelime sıralaması için önerilmiştir. Başka bir çalışmada Zhao ve diğerleri (2011) Twitter içeriğini özetlemek ve analiz etmek için başlıksal anahtar kelime çıkarmayı önermişlerdir.

BSS gibi konu temelli kümeleme yaklaşımları, Merin Sıralama ve konu temelli kümelemeyi bir araya getirip her konu için Metin Sıralama sonuçlarını toplayarak öbekleri oluşturur. Büyük ölçekli bir belge koleksiyonundan kelime konuları elde etmek ve kelimelerin konularını sağlayabilmek için WordNet ya da denetimsiz makine öğrenme teknikleri kullanmaktadır.

Anahtar kelime çıkarımı için BSS iki aşamadan oluşur:

- Kelime ve belge konularını elde etmek için bir konu tercüman oluşturulur;
- Belgeler için anahtar kelime çıkarımı için BSS gerçekleştirilir.

Kelime ve belgelerin konularını elde etmek için bir konu başlık tercümanı oluşturduktan sonra, BSS aracılığıyla belgeler için anahtar kelime çıkarımı yapılır. Bir doküman d verildiğinde, BSS kullanılarak anahtar kelime çıkarımı işlemi, Şekil 2.19'da de gösterilen aşağıdaki dört adımdan oluşmaktadır.



Şekil 2.19. Başlıklı sayfa sıralama modeli

1. D dökümanı için bir kelime grafiği oluşturulur.
2. Farklı konulara göre her kelimenin önem puanlarını hesaplamak için BSS gerçekleştirilir.

3. Kelimelerin konuya özel önem puanlarını kullanarak, her bir konu için adayların temel ifadelerini ayrı ayrı sıralanır.
4. d dokümanı konularını dikkate alınarak üst sıralarda yer alan anahtar sözcükler anahtar sözcükler olarak seçilir.

Anahtar kelime çıkarımı için BSS'de önce her bir aday anahtar kelimelerin sıralama puanlarını ayrı ayrı hesaplanır. Yani her bir konu z için Eşitlik 2.25'te görüldüğü gibi D belgesinin konu dağılımını her konu z için $pr(z|d)$ olarak belirtilir.

$$R_z(p) = \sum_{w_i \in p} R_z(w_i) \tag{2.25}$$

Her aday anahtar kelimesi p için, son sıralama puanı Eşitlik 2.26'daki gibi hesaplanır:

$$R(p) = \sum_{z=1}^{K} R_z(p) \times pr(z|d) \tag{2.26}$$

Aday cümleleri sıralama puanlarının azalan düzeninde sıraladıktan sonra, d belgesindeki en üstteki sözcükler anahtar sözcük olarak seçilir.

Başlık sıralama

Bougouin, Boudin ve Daille (2013) Metin Sıralama ve Metin Sıralama tabanlı yöntemlerin iyileştirilmesi olan Başlık Sıralama modelini sunmuşlardır. Başlık Sıralama bir belgenin en önemli konularından anahtar sözcükleri çıkarmayı amaçlayan denetimsiz bir yöntemdir. Konular benzer anahtar sözcük adaylarının kümeleri olarak tanımlanır. Bir belgeden anahtar kelime çıkarımı aday anahtar kelime çıkrıldıktan sonra, adayların kümelenmesi adımını içermektedir.

Başlık Sıralama; belgeleri grafik olarak temsil eder ve belgedeki konuları köşeler, kenarları köşeler arasındaki anlamsal ilişkilerin gücüne göre ağırlıklandırır. Örneğin grafik yapısı $G = (V, E)$ ve V bir dizi köşe noktasıdır ve kenarları E, V×V bir alt kümesi olarak göz önüne alındığında, köşeler konulardır ve iki konu t_i ve t_j arasındaki kenar yani anlamsal

ilişkilerinin gücüne göre ağırlıklandırılır. Eğer anahtar kelime adayları belgede genellikle birbirine yakın görünüyorsa t_i ve t_j arasında güçlü bir anlamsal ilişki vardır. Bu nedenle, kenarlarının w_{ij} ağırlığı Eşitlik 2.27'deki gibi tanımlanır:

$$w_{i,j} = \sum_{c_i \in t_i} \sum_{c_j \in t_j} mes(c_i, c_j) \tag{2.27}$$

$$mes(c_i, c_j) = \sum_{p_i \in poz(c_i)} \sum_{p_j \in poz(c_j)} \frac{1}{|p_i - p_j|} \tag{2.28}$$

Eşitlik 2.28'de $mes(c_i, c_j)$ belgedeki c_i ve c_j aday anahtar sözcüklerinin ofset pozisyonları arasındaki ve $poz(c_i)$aday anahtar sözcük c_i'sinin tüm ofset pozisyonlarını temsil ettiği karşılıklı mesafeleri ifade eder. Grafik oluşturduktan sonra konuları sıralamak için grafik tabanlı sıralama modeli Metin Sıralama kullanılır. Bu model, "puanlama" kavramına dayalı konulara bir anlam puanı atar: yüksek puan alan konular bağlantılı konularının puanına daha fazla katkıda bulunur. Konu puanı t_iEşitlik 2.29'da görüldüğü gibi hesaplanmaktadır.

$$S(t_i) = (1 - \lambda) + \lambda \times \sum_{t_j \in V_i} \frac{w_{j,i} \times S(t_j)}{\sum_{t_k \in V_j} w_{j,k}} \tag{2.29}$$

V_i, t_i ve λ için oy veren konulardır ve genellikle 0,85 olarak tanımlanan bir sönümleme faktörüdür (Brin ve Page, 1998).

Başlık Sıralama'nın metin Sıralama'ya göre birçok avantajı vardır. Sezgisel olarak, kelimeler yerine konuları sıralamak, bir belgenin ana konularını kapsayan anahtar kelime kümesini tanımlamanın daha kolay bir yolududur. Diğer bir avantajı, konular arasındaki anlamsal ilişkileri daha iyi yakalayan eksiksiz bir grafiğin kullanılmasıdır. Bunu yapmak için, en üst sıradaki kümelerin her birinden bir anahtar kelime adayı seçilir. Aday anahtar kelimeleri kümeleme fazlalığı ortadan kaldırır ve aynı zamanda sınırları güçlendirir. Sıralamanın etkinliği büyük ölçüde grafiğin kısalığına ve belgedeki anlamsal ilişkileri doğru bir şekilde temsil etme yeteneğine bağlıdır.

Sıralama tabanlı denetimsiz anahtar kelime çıkarımı

Škrlj, Repar ve Pollak (2019) çalışmalarında Sıralama Tabanlı Denetimsiz Anahtar Kelime Çıkarımı (St-DAKÇ) algoritmasını, benzer köşeleri toplamak amacıyla meta köşeler kavramını ilk kez kullanarak geliştirmişlerdir. Önerilen algoritmada ilk sözcüksel grafik, meta-köşe noktalarının, yani mevcut köşe noktalarının toplamlarının eklenmesiyle genişletilir ve oluşturulur. İkinci oluşturulan grafik yedek filtreler ile birlikte bir grafik yük merkeziyet ölçüsü kullanılır. St-DAKÇ algoritmasının bir parçası olarak, kabul edilen belirteçler için hesaplanan yük merkezi puanlarına dayalı olarak anahtar kelimeleri çıkartır.

Algoritma üç adımdan oluşur: ilk olarak grafik belirli bir sıralı belirteç kümesinden oluşturulur, sonra ortaya çıkan grafik genellikle çok seyrektir, çünkü çoğu kelime nadiren birlikte görülür. Bu adımın sonucu hem köşe hem de kenar sayısının daha az olduğu daha küçük, daha yoğun bir grafiktir. So adımda aday anahtar ifadeler sıralanır.

2.7.3. Gömme tabanlı modeller

Grafik tabanlı anahtar kelime çıkarımı ile kelimeler ve kelimeler arasındaki ilişkiyi dikkate alır, ancak basit ikili birlikte oluşum ile sınırlıdır ve diller arasındaki belgelerin çoklu ilişkisini iyi ifade edememektedir. Kelime gömmeleri, kelimeler arasındaki anlamsal ve sözdizimsel ilişkileri korpus için oluşturulan vektör uzayında temsil etmektedir. DDİ problemlerini çözmek için kelime gömmeleri son yıllarda sıklıkla kullanılmaktadır.

Cümle sıralama

Sun ve diğerleri (2020) Cümle Sıralama adı verilen önceden eğitilmiş bir dil modeline dayalı olarak denetimsiz anahtar kelime çıkarımı için yeni bir yöntem önermişlerdir. Cümle Sıralama, cümle gömme modeli olan Cümle ve otoregresif önceden eğitilmiş dil modeli DMKG'yu birleştirir ve kısa belgeler için anahtar kelime çıkarma görevini gerçekleştirir.

Cümle Sıralama'nın bir anahtar kelime çıkarımının genel adımları:

- Doküman belirteçlere ayrılır ve konuşma kısmı etiketli bir dizi belirtece etiketlenir.

- Diziden isim cümleleri çıkarılır ve belgeden çıkarılan isim tamlamaları aday anahtar sözcükler olarak kabul edilir.
- Belirteç dizisi önceden eğitilmiş bir dil modeline yerleştirilir, her belirtecin temsili çıkarılır. Bu durumda, temsil, farklı özelliklere sahip çok katmanlı kelime gömme olabilir.
- Cümle isim tamlaması gömmelerine ve belge gömmeye çevirilir. Bu noktada, aynı sayıda katman ve boyuta sahiptirler.
- Aday ifade gömme ve belge gömme arasındaki kosinüs mesafesini hesaplanır. Bu mesafe, aday anahtar sözcüklerle belge konusu arasındaki benzerliği tanımlamaktadır. Son olarak anahtar sözcükler en benzer aday anahtar sözcüklerin İlk n tanesi seçilerek bu aday anahtar kelimelerini sıralar.

Belirli bir d dokümanı için, d için gömme $v_d$dir. Aday anahtar kelimesi isim tamlamalarının düğümleri v_{NP}'dir. Cümle Sıralama, v_d ve v_{NP} arasındaki benzerlik veya korelasyon puanı olarak tanımlanır (Eşitlik 2.30):

$$Cümle\ Sıralama = \left(v_{NP_i}, v_d\right) = Sim\left(v_{NP_i}, v_d\right) \tag{2.30}$$

Benzerlik genellikle kosinüs mesafesi ile hesaplanmaktadır. Cümle Sıralama değeri 0 ile 1 arasındadır, 1'e ne kadar yakın olursa, aday anahtar kelimesi belgenin konusuyla o kadar alakalı olmaktadır. Tersine, değer 0'a ne kadar yakın olursa, ifade konu ile o kadar alakasız olmaktadır (Campos ve diğerleri, 2018).

Çizelge 2.4. Denetimsiz modellerin performans sonuçları

Model	Inspec			SemEval2010			SemEval2017		500N-KPCrowd	WWW		KDD	
	F_1@10	F_1@5	F_1	F_1@10	F_1@5	F_1	F_1@10	F_1@5	F_1	F_1@k	F_1	F_1@k	F_1
BSS	-	0.242	-	-	-	-	-	-	0,158	-	-	-	-
HO-AKÇ	-	-	-	-	-	-	-	-	-	-	-	-	-
Başlık Sıralama	-	0,227	0,279	-	-	0,121	0,226	0,171	0,172	-	-	-	-
BD-AKÇ	0,19	0,157	0,316	-	-	0,123	0,181	0,118	0,101	-	0,172	-	0,156
ÇMİ-TB1	-	-	0,488	-	-	0,38	-	-	0,518	-	-	-	-
ÇMİ-ÖS	-	-	0,485	-	-	0,354	-	-	0,511	-	-	-	-
Cümle Sıralama	0,388	0,291	-	-	-	-	0,328	0,225	-	-	-	-	-
St-DAKÇ	-	-	0,054	-	-	0,091	-	0,112	0,428	-	-	0,06	0,046

3. YÖNTEM VE ARAÇLAR

Üretilen, tüketilen ve depolanan verinin hızla artması ile birlikte büyük verilerden bilgiyi filtreleme problemi ortaya çıkmıştır. Bu problemin çözümü için otomatik anahtar kelime ataması yapılmaktadır. Literatürde anahtar kelime ataması probleminin çözümü için birçok denetimli ve denetimsiz model önerilmiştir. Bu modeller tezin bir önceki bölümünde mimari detayları ve performans sonuçları ile ele alınmıştır. Fakat önerilen modellerin performans sonuçları hala beklentilerin çok altındadır. Anahtar kelime çıkarımı probleminin zorluğu, farklı veri kümelerine ait değişken karakteristik özelliklerden kaynaklanmaktadır. Bu tez kapsamında anahtar kelime çıkarımı için yeni hibrit bir yöntem önerilmiştir. Önerilen bu hibrit model ile farklı kategorilere ait çok sayıda seçici özniteliğin hibrit olarak eğitilmesi yaklaşımı kullanılarak anahtar kelime çıkarımı problemine ait zorluklar aşılmaya çalışılmıştır.

3.1. Hibrit Anahtar Kelime Çıkarımı

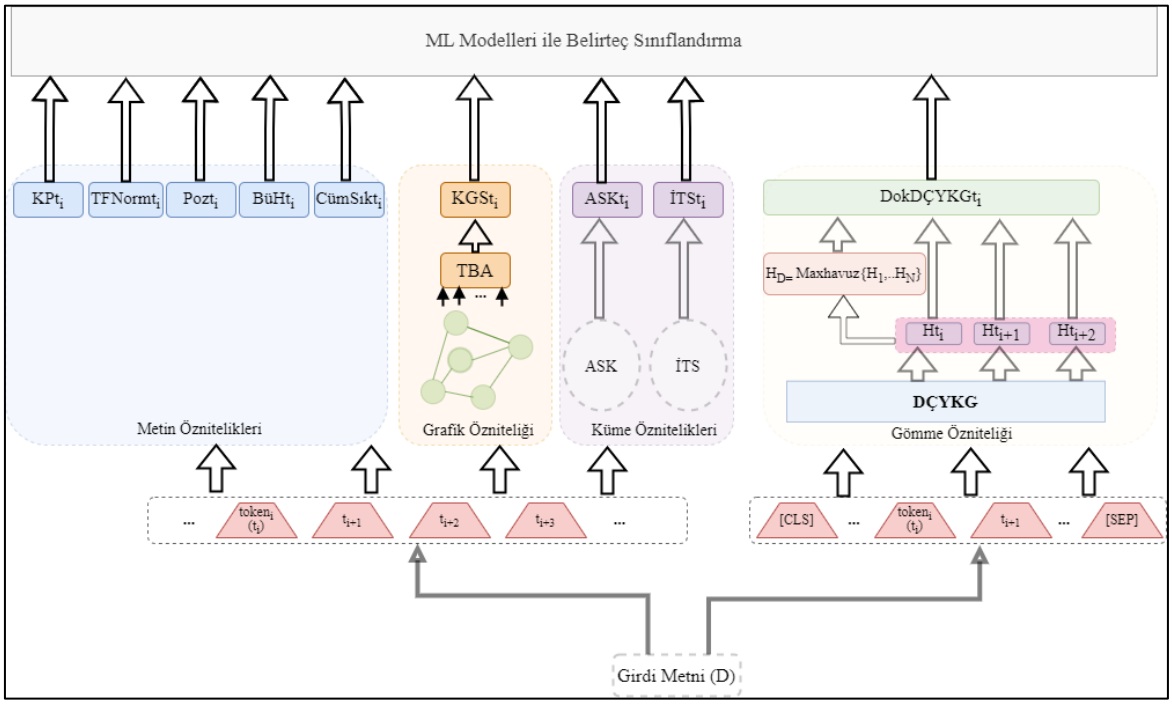
Bu tezde, anahtar kelime çıkarımı, bir dizi etiketleme görevi olarak ele alınmıştır. Geliştirilen model, birçok potansiyel seçici özniteliği hibrit kullanan ilk dizi etiketleme modelidir. Çok sayıda seçici öznitelik kullanılarak veri kümelerine ait değişken özelliklerden kaynaklanan zorluk aşılmaya çalışılmıştır. Önerilen model, metin (dilbilimsel, istatistiksel), grafik tabanlı, gömme tabanlı ve küme özelliklerinin bir arada kullanıldığı hibrit bir yöntem olarak tasarlanmıştır. Literatürde yaygın olarak kullanılan özelliklere ek olarak dilsel özellikleri yakalamak için Anahtar Sözcük Kümesi (ASK) ve İsim Tamlama Kümesi (İTK) öznitelikleri oluşturulmuştur. Şekil 3.1.'de önerilen HibritAKÇ modeli görülmektedir. Önerilen modelde metin puanları, grafiksel puan, DÇYKG (Devlin ve diğerleri, 2018) doküman bezerlik puanı ve küme puanlarını hesaplayan dört alt modül bulunmaktadır.

Modelde, veri kümesinde bulunan her bir girdi Dokümanı (D),

$$\mathrm{D} = [\mathrm{t}_i,\ \mathrm{t}_{i+1}, \dots, \mathrm{t}_n]\,, \qquad i = 1, .., n \tag{3.1}$$

dokümanı temsil eden belirteç dizisi şeklinde gösterilmektedir. Eşitlik 3.1'de n, D dokümanındaki toplam token sayısını göstermektedir. Her bir kelime belirteçi ti için bu öznitelik skorları ayrı ayrı hesaplanarak Eşitlik 3.2'de tanımlanan dizide tutulmaktadır.

$$X_{t_i} = [\text{DokDÇYKG}_{t_i}, \text{BüH}_{t_i}, \text{KP}_{t_i}, \text{TFNorm}_{t_i}, \text{Poz}_{t_i}, \text{CümSık}_{t_i}, \text{KGP}_{t_i}, \text{ASK}_{t_i}, \text{İTK}_{t_i}] \quad (3.2)$$



Şekil 3.1. Belirteç sınıflandırma tabanlı HibritAKÇ

Eşitlik 3.2'de yer alan her bir özniteliğe dair tanımlama ve formüller detaylı olarak Çizelge 3.1'de verilmiştir. D dokümanında bulunan tüm belirteçler için X matrisleri Eşitlik 3.3'te görüldüğü şekilde dizi olarak tutulmaktadır.

$$X = [X_{t_i}, X_{t_{i+1}}, \dots, X_{t_n}], \qquad i = 1, .., n \quad (3.3)$$

Tüm X değerleri Dizi Sınıflandırma modülünden geçirilerek Eşitlik 3.4'te görülen Y dizisine göre eğitim ve test görevini gerçekleştirmektedir.

$$Y = [y_{t_i}, y_{t_{i+1}}, \dots, y_{t_n}], \qquad i = 1, .., n \quad (3.4)$$

Y dizisi her bir tokenin anahtar kelime (1) veya değil (0) etiketlerinden oluşmaktadır. Belirteç Sınıflandırma modülünde DVM, ÇKA, NB ve Rassal Orman (RO) sınıflandırma algoritmaları ayrı ayrı eğitilmiştir.

3.1.1. Öznitelikler

Literatürde istatistiksel, grafik tabanlı, gömme tabanlı ve dilsel özellikler çok sayıda modelde ayrı ayrı kullanılmıştır. Belirteç Sınıflandırma tabanlı bu model için aralarından güçlü özellikler seçilmiş ve hibrit olarak kullanılmıştır. Bu yaklaşım, geliştirilen modelin daha yüksek performans elde etmesini sağlamaktadır. Geliştirilen model için girdi olarak kullanılan öznitelikler metin, grafik, küme ve gömme öznitelikleri olmak üzere dört kategoriye ayrılmıştır. Her özniteliğin açıklamaları ve formülleri Çizelge 3.1'de verilmiştir.

Çizelge 3.1. Belirteç sınıflandırma için kullanılan özniteliklerin hesaplanması

Kategori	Öznitelik	Formül
Metin Öznitelikleri	KP	Konuşmanın parçası(kelime) − > ‘İsim’ = 1 , ‘Sıfat’ = 0.5, ‘Fiil’ = 0.35, ‘Diğer’ = 0.25.
	TFNorm	uzunlukTFler = TF[kelime]’ler ortTF = orta(geçerliTFs) TFNorm = TF[kelime] / ((ortTF / uzunlukTFs) + 1)
	Pozisyon (Poz)	ln (3 + orta(ofsetler-cümleler[kelime]))
	Kelimenin cümlelerdeki sıklığı (CümSık)	uzunluk(ofsetler-cümle[kelime] / uzunluk(cümleler)
	Büyük harf (BüH)	max(KüçükTF[kelime], BüyükTF[kelime]) / (1 + ln (TF[kelime]))
Grafik Özniteliği	Kelime Grafik Puanı (KGP)	TBA(girdi metninin kelime grafiği)
Küme Özniteliği	Anahtar Sözcük Kümesi (ASK)	0: ASK içerisinde değil, 1: ASK içerisinde, 2: Bir anahtar ifadenin parçası
	İsim Tamlaması Kümesi (İTK)	0: İTK içinde değil, 1: İTK içinde, 2: İsim tamlamasının parçası
Gömme Özniteliği	DÇYKG kelime/doküman benzerliği (DokDÇYKG)	H_i = DÇYKG(t_i), H_D = Maxhavuzlama({H_1, H_2, …, H_N}), Kelime/Dok Benzerliği = $\frac{1}{\|\|H_D - H_i\|\|}$

Birinci kategori altında metin öznitelikleri, Konuşma Bölümü (KP), kelimenin Terim Sıklığı Normalleşmesi (TFNorm), kelimenin konumu (Poz), kelimenin belge cümlelerindeki sıklığı

(CümSık) ve büyük harfler (BüH) puanları hesaplanmıştır. KP puanı isim, sıfat, fiil ve diğer KP türleri için sırasıyla 1, 0,5, 0,35, 0,25 olarak tanımlanmıştır. TFNorm puanı, girdi belirteçlerin tüm kelimelere TF oranıdır. Poz özniteliği, girdi tokenların konumunun ortalamasını kullanarak hesaplanmaktadır. CümSık özelliği, cümlelerdeki giriş tokenlarının görüldüğü cümle sayısının toplam cümle sayısına oranıdır. Son olarak, BüH puanı, büyük harf girdi belirteçi veya harf girdi belirteç sayısı (maksimum) ile TF'nin (kelime) doğal logaritması (ln) arasındaki oranı hesaplamaktadır.

Kelime Grafik Puanı (KGP), Grafik özniteliği (Zehtab-Salmasi ve diğerleri, 2021) olarak ikinci kategori altında hesaplanmıştır. Her kelime düğümü için maksimum varyasyonun yönünü yakalamak için, KGP, Tekil Değer Kompozisyonu (TDK), Bağımsız Komponent Analizi (BKA), ve Lineer Ayrımcı Analizi (LAA) (Chandrashekar ve Sahin, 2014) gibi diğer doğrusal özellik dönüşümleri yerine Temel Bileşenler Analiz (TBA) (Zehtab-Salmasi ve diğerleri, 2021; Vega-Oliveros, Gomes, Milios ve Berton, 2019) ile tek boyuta indirgenmiştir. Tek boyuta indirgendiğinde özvektör, arasındalık, yakınlık, Sayfa Sıralama, kısıtlama, kümeleme, eksantriklik ve yapısal boşluk puanları sözcük birlikte görülme grafiği puanları olarak kullanılmıştır.

Üçüncü kategoride, kelimelerin küme puanları, ASK ve İTK olmak üzere iki farklı kümeye (öznitelikler) eşlenmektedir. Bu puanlar her küme için 0, 1 veya 2 değerlerini alabilmektedir. ASK, TRDizinEn'in 25000 özetten oluşan büyük bir sürümü olan TRDizinLargeEn veri kümesindeki anahtar sözcüklerden oluşmaktadır. İTK ise, "<NN.*|JJ>*<NN.*>" düzenli ifadesiyle çıkarılan tamlamalardan oluşur.

Dördüncü kategori altında hesaplanan kelimenin Doküman DÇYKG (DokDÇYKG) puanı, önceden eğitilmiş DÇYKG kelime gömmeleri kullanılarak hesaplanır. Her bir belirteç t_i için, kelime gömme vektörünün temsili H_i'dir ve belge gömme vektörünün temsili H_D olarak ifade edilir. Kelimenin DokDÇYKG puanı hesaplanırken H_D ile H_i arasındaki Manhattan Mesafesi ölçülmektedir.

Belirteç Sınıflandırma modülünde, özniteliklerin ağırlıkları makina öğrenmesi modelleri ile hesaplanmıştır. RO'den elde edilen özniteliklerin Inspec, 500N-KPCrowd, Semeval-2017, TRDizinEn ve DergiParkEn veri setleri için ağırlıkları Çizelge 3.2'de görülmektedir. Çizelge

3.2'de görüldüğü gibi DocDÇYKG, TFNorm, Poz ve KGP her bir veri seti için daha yüksek ağırlıklara sahipken, KP, BüH, CümSık, ASK ve İTK ağırlıkları düşük ağırlıklara sahiptir.

Çizelge 3.2. Özniteliklerin RO'a göre ağırlıkları

Öznitelik	Inspec	500N-KPCrowd	Semeval-2017	TRDizinEn	DergiParkEn
KP	0.020	0.029	0.033	0.015	0.017
TFNorm	0.200	0.153	0.200	0.212	0.182
Poz	0.173	0.195	0.169	0.181	0.148
BüH	0.018	0.041	0.015	0.039	0.109
CümSık	0.096	0.144	0.109	0.116	0.122
KGP	0.188	0.215	0.193	0.199	0.170
ASK	0.030	0.033	0.035	0.044	0.029
İTK	0.015	0.023	0.020	0.016	0.015
DokDÇYKG	0.258	0.168	0.226	0.178	0.208

3.2. Veri Kümesi

Anahtar kelime problemi için önerilen algoritmaların test edilmesi için literatürde farklı veri kümeleri yer almaktadır. Bunlar arasında Inspec (Hulth, 2003), 500N-KPCrowd (Marujo ve diğerleri, 2013), Semeval-2017 (Augenstein ve Sogaard, 2017) veri kümeleri sık kullanılmaktadır. Bu çalışmada sık kullanılan veri kümelerine ilave olarak yeni iki özet veri kümesi, TRDizinEn ve DergiParkEn veri kümeleri oluşturularak çalışmaya eklenmiştir.

Çizelge 3.3'te veri kümeleri ve özellikleri sunulmaktadır. Çizelgede testlerde kullanılan tüm veri kümeleri için alan, Doküman Sayısı (DokSay), toplam Anahtar Kelime Sayısı (AKSay), görünür anahtar kelime sayısı (Görünür), görünmeyen anahtar kelime sayısı (Görünmez), görünür anahtar kelimelerin oranı (%Görünür), her doküman için ortalama anahtar kelime sayısı (DokOA) bulunmaktadır. Veri kümeleri eğitim, test ve doğrulama alt kümelerinden oluşmaktadır. Bu çalışmada her veri kümesi için sadece eğitim veri kümesinde bulunan dokümanlar kullanılmıştır.

Çizelge 3.3. Veri kümeleri

Veri Kümesi	Alan	DokSay	AKSay	Görünür	Görünmez	%Görünür	DokOA
DergiParkEn	Mühendislik	725	3178	2136	1042	67.2	4,38
TRDizinEn	Karışık	1075	6718	2577	2864	38.4	6,24
Inspec	Bilgisayar Bil.	1000	9788	7493	2295	76.5	9,78
500N-KPCrowd	Karışık	400	19834	18024	1810	90.8	49,58
Semeval-2017	Karışık	350	6732	6732	0	100	19,23

DergiParkEn veri kümesi web kazıma ile elde edilmiş DergiPark makalelerinden oluşan 725 özet bulunan bir veri kümesidir. DergiParkEn veri kümesi Mühendislik alanına, TRDizinEn ise 1075 dokümandan oluşan karma alana ait verilerden oluşur. TRDizinEn ve DergiParkEn veri kümeleri anahtar kelimeler alan uzmanı atayıcı ve/veya yazarlar tarafından atanmıştır. DergiParkEn veri kümesi TRDizinEn veri kümesi ile benzer özellikler taşımaktadır. HibritAKÇ modelinin alan spesifik çalışıp çalışmadığını gözlemlemek amacıyla farklı alana ait veri kümeleri oluşturulmuştur.

Inspec içerisinde 1000 adet doküman bulunmaktadır. Bu dokümanlar Bilgisayar Bilimleri alanına ait makale özetlerinden oluşmaktadır. Bu veri kümesine ait anahtar kelimeler sadece profesyonel indeksleyiciler tarafından atanmıştır. 500N-KPCrowd veri kümesi 400 adet haber içeriğinden oluşmaktadır. Veri kümesine ait anahtar kelimeler okuyucular tarafından seçilmiştir. Semeval-2017 veri kümesi ise 350 doküman içermektedir. Bu veri kümesi karışık alana ait paragraflardan oluşmaktadır. Semeval-2017 veri kümesine ait paragraflara anahtar kelime ataması öğrenciler ve indeksleyicilerin ortak kararına göre yapılmıştır.

3.3. Uygulama

Bu tez kapsamında geliştirilen modele ait deneysel çalışmalar, Intel Core i5 10210U 2.11 GHz işlemci, 8 GB RAM'e sahip 4 çekirdekli bir bilgisayarda gerçekleştirilmiştir. Programlama dili olarak Python kullanılmıştır. Uygulama, Anaconda Spyder ortamında gerçekleştirilmiş, yapay zekâ kütüphaneleri olarak Torch ve Transformers kullanılmıştır. Önerilen modelin performans sonuçlarını ölçmek için, Belirteç Sınıflandırma Modülünde bir dizi klasik makine öğrenimi algoritması incelenmiştir. DVM sınıflandırıcılar arasından DVM, Bayes sınıflandırması için Gaussian NB, yapay sinir ağı modelleri arasından ÇKA ve

topluluk sınıflandırması için RO seçilmiştir. Her model arasından, testler sırasında daha iyi sonuçlar veren seçilmiş parametreler belirlenmiştir. DVM için kernel fonksiyonu olarak sigmoid kullanılmış ve c hata teriminin ceza parametresi 3 olarak tanımlanmıştır. Naive Bayes'in genişletilmiş versiyonu olarak Gauss Naive Bayes seçilmiştir. ÇKA, 128 nöronlu tek bir gizli katman olarak oluşturulmuştur. ÇKA modeli için maksimum iterasyon sayısı 5000 olarak belirlenmiş ve öğrenme katsayısı 0,1 olarak alınmıştır. Aktivasyon fonksiyonu olarak Relu seçilmiş ve geriye yayılımdaki ağırlıkları optimize etmek için stokastik gradyan iniş fonksiyonu kullanılmıştır. RO algoritması için n-tahmin edici değişkeni 300 olarak ayarlanmıştır.

3.4. HibritAKÇ Performans Sonuçları

Önerilen HibritAKÇ modeli, DergiParkEn ve TRDizinEn adlı yeni derlenen veri kümeleri ile eğitilmiş ve test edilmiştir. Model ayrıca literatürde yaygın olarak kullanılan Inspec, 500N-KPCrowd ve Semeval-2017 veri kümeleri üzerinde eğitilmiş ve test edilmiştir. Deneyler, çapraz doğrulama yöntemi kullanılarak 20 kez çalıştırılmıştır. Bilinen istatistiksel özelliklere sahip veriler için k-katlı çapraz doğrulamada optimal k değerinin (Marcot ve Hanea, 2021) 10 olması önerilmiştir. Ancak anahtar kelime çıkarmada kullanılan veri kümeleri bilinen istatistiksel özelliklere sahip değildir. Sınıflandırma hata oranının kararlılığını sağlamak için literatürde (Argamon ve Levitan, 2005; Ghosh, Saha ve Molakathaala, 2022; Hafeez ve Kathirisett, 2022) olduğu gibi k çapraz doğrulama değeri 20 olarak seçilmiştir.

Performans sonuçlarını karşılaştırmak için farklı kategorilerde yüksek puan alan anahtar kelime çıkarım modelleri kullanılmıştır. Bir aday sınıflandırma modeli Karmaşık Ağ tabanlı Aşırı Gradyan Artırma (CN-AGA) (Duari ve Bhatnagar, 2020), denetimli dizi etiketleme modeli ÇY-UKSB (Basaldella ve diğerleri) ve ÇY-UKSB-KRA modeli (Sahrawat ve diğerleri, 2020) test için seçilmiştir. Ek olarak, önerilen modelin performans karşılaştırması için denetimsiz gömülü tabanlı bir aday sıralama modeli DAKÇ (Liang ve diğerleri, 2021)'nin performans sonuçları çizelgeye eklenmiştir. Çizelge 3.4'te önerilen modelin F_1-skor sonuçları diğer dört model ile karşılaştırmaktadır.

Çizelge 3.4. HibritAKÇ performans sonuçlarının karşılaştırılması

Model		Inspec	500N-KPCrowd	SemEval-2017	TRDizinEn	DergiParkEn
CN-AGA		0,607	0,538	-	-	-
ÇY-UKSB-KRA		0,59	-	0,52	-	-
ÇY-UKSB		0,42	0,29	0,28	0,26	0,22
DAKÇ @15		0,41	0,17	0,37	0,17	0,11
HibritAKÇ	NB	0,34	0,36	0,51	0,27	0,39
	DVM	0,28	0,14	0,46	0,10	0,36
	ÇKA	0,35	0,34	0,65	0,27	0,55
	RO	0,50	0,69	0,69	0,74	0,70

Çizelge 3.4'te görüldüğü gibi ÇY-UKSB modeli ile Inspec veri seti için 0,42 F_1-skor elde edilirken, RO ile 0,50 F_1-skor elde edildi. Sahrawat ve diğerleri (2020), ÇY-UKSB-KRA ile Inspec ve Semeval-2017 veri setleri için sırasıyla 0,59 ve 0,52 F_1-skor bildirmiştir. Duari ve Bhatnagar (2020), KA-AGA ile Inspec ve 500N-KPCrowd veri kümeleri için sırasıyla 0,607 ve 0,538 F_1-skor elde etmiştir. Önerilen model ile 500N-KPCrowd, Semeval-2017, TRDizinEn ve DergiParkEn veri setleri için sırasıyla 0,69; 0,69; 0,74 ve 0,70 F_1-skorları elde edilmiştir. Çizelge 3.4'te görüldüğü gibi, önerilen yaklaşım veri kümelerinin çoğu için neredeyse yakın sonuçlar elde etmiştir. Çizelge 3.4 ayrıca topluluk modellerinin hem önerilen modelde hem de literatürdeki diğer çalışmada (KA-AGA) daha yüksek sonuçlara ulaştığı sonucu elde edilmektedir.

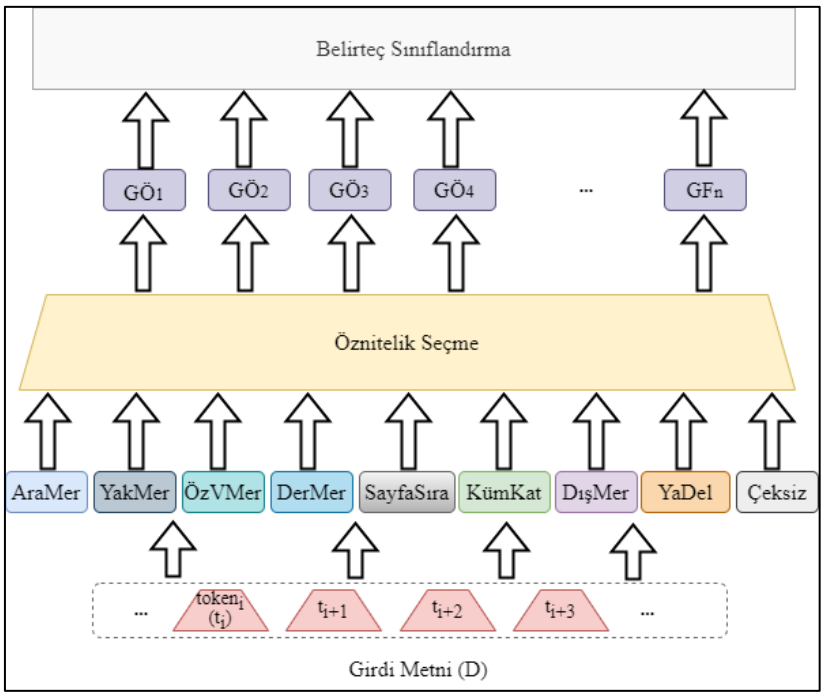
3.5. Grafik Tabanlı Öznitelikler Kullanılarak Yeni bir Model Geliştirilmesi

Literatürde grafik-tabanlı anahtar kelime çıkarımı denetimsiz ve denetimli olmak üzere iki şekilde gerçekleştirilebilmektedir. Denetimsiz grafik tabanlı yöntemlerde önce dil tabanlı filtreler ve konuşmanın parçası bilgisine dayalı aday anahtar ifadeler çıkarılır. İkinci adımda ise girdi metninden çizilen kelime grafiğinden bu aday ifadelerin grafik tabanlı puanları hesaplanır ve sıralanır. Denetimli modellerde ise grafik tabanlı öznitelikler kullanılarak tüm belirteçler için dizi etiketleme veya aday anahtar ifadeler için sınıflandırma yapılarak anahtar kelimeler çıkarılmaktadır. Her bir kelime için öznitelikler kullanılarak dizi etiketleme, sadece aday anahtar ifadeler için öznitelikler çıkarılarak sınıflandırma gerçekleştirilmektedir.

Literatürde önerilen grafik tabanlı modellerde kelime birlikte görülme grafiğinden elde edilen Arasındalık Merkeziliği (AraMer), Yakınlık Merkeziliği (YakMer), Öz Vektör Merkeziliği (ÖzVMer), Derece Merkeziliği (DerMer), Sayfa Sıralama, Kümeleme Katsayısı (KümKat), Dış Merkezlilik (DışMer), Yapısal Delik (YaDel), ve Çekirdeksizlik (Çeksiz) özniteliklerinin sıklıkla kullanıldığı görülmektedir. Grafik-tabanlı modellerde en sık kullanılan öznitelik Sayfa Sıralama'dır (Brin ve Page, 1998). Sayfa Sıralama temel alınarak BSS (Liu, Huang, Zheng ve Sun, 2010; Zhao ve diğerleri, 2011), Pazisyon Sıralama (Florescu ve Caragea, 2017) gibi algoritmalar önerilmiştir. Sayfa Sıralama temel alınarak geliştirilen metin Sıralama (Mihalcea ve Tarau, 2004) algoritması da modellerde (Wan ve Xiao, 2008; Bougouin ve diğerleri, 2013; Alfarra ve Alfarra, 2018; Prasad ve Kan, 2019) sıklıkla kullanılmaktadır. Bazı çalışmalarda bir tek öznitelik yerine birden fazla grafik-tabanlı özniteliği (Vega-Oliveros ve diğerleri, 2019; Zehtab-Salmasi ve diğerleri., 2021; Beliga, Meštrović ve Martinčić-Ipšić, 2014; Duari ve Bhatnagar, 2020) birlikte kullanılmıştır.

AraMer, köşelerin bilgi aktarma kapasitesi ile ilgilidir. Bir j köşesi için AraMer, j içeren tüm köşe çiftleri arasındaki en kısa yolların sayısı ile hesaplanır (Das, Samanta ve Pal, 2018). YakMer, her köşeden ağın geri kalanına giden en kısa yolların uzunluklarına göre tanımlanır (Das ve diğerleri, 2018). YakMer, j'den tüm köşelere giden en kısa yolların ortalamasının tersidir. ÖzVMer veya $vertex_j$ prestiji, yakınsama sağlanana kadar yinelemeli olarak hesaplanır (Zaki ve Meira, 2014). Bir düğümün DerMer'i, düğümün yerel düzeyde gömülmüşlüğünü ölçer (Barrat, Barthelemy, Pastor-Satorras ve Vespignani, 2004). Sayfa Sıralama, ağda rastgele bir yürüyüşle tanımlanan çok sayıda adımdan sonra belirli bir j köşesine ulaşma olasılığını ifade eder (Brin ve Page, 1998). Bir düğümün KümKat değeri, komşuluğundaki kenar yoğunluğu kullanılarak hesaplanır. Ağda üçgenlerin (üçüncü dereceden döngüler) varlığını ölçer (Pastor-Satorras, Castellano, Van Mieghem ve Vespignani, 2015). Bir j köşesinin Eksant değeri, diğer köşelere giden en kısa yolların tümü üzerindeki en uzun mesafedir (Das ve diğerleri, 2018). Alt Eksant, j tepe noktasının daha merkezi olduğunu gösterir. Ağdaki bazı köşeler kaldırılırsa, yapısal bir boşluk oluşur. YaDel köşeleri, köşe grupları arasında anahtar görevi görür (Vega-Oliveros, Berton, Andrade Lopes ve Rodrigues, 2015). Daha merkezi köşeler daha yüksek YaDel değerine sahiptir. Çeksiz, G ağını maksimum bağlı alt graflara böler ve ait olduğu en yüksek çekirdeği alan bir düğümün değerini hesaplar (Seidman, 1983).

Literatürde sık kullanılan grafik tabanlı 9 öznitelik dahil edilen bu modelde anahtar kelime çıkarımı yine bir dizi etiketleme problemi olarak ele alınmıştır. Şekil 3.4'te Grafik-tabanlı Anahtar Kelime Sınıflandırması (Gt-AKS) modeli görülmektedir. Önerilen model düğüm grafik puanlarını girdi dokümanına ait her bir kelime için hesaplamakta ve bu özniteliklere göre sınıflandırma yapmaktadır. Girdi metni için tüm öznitelikler yerine Öznitelik Seçme Modülü ile ayıklanmış Grafik Öznitelikleri (GÖ) kullanılarak performans sonuçları elde edilmiştir. Bu modül içerisinde ekstra ağaç, lasso, genetik algoritma ve sarmalama algoritmaları ile öznitelik grupları oluşturulmaktadır.



Şekil 3.2. Grafik tabanlı anahtar kelime sınıflandırma

3.6. Grafik Tabanlı Anahtar Kelime Sınıflandırma Modeli Performans Sonuçları

Çizelge 3.5'te Ekstra Ağaç (EA), Lasso (LS), Genetik Algoritma (GA), ve Sarmalayıcı (SR) öznitelik seçme yöntemleri ile her bir veri kümesi için seçilmiş öznitelikler sunulmuştur. Çizelge incelendiğnde YaDel özniteliğinin bütün veri kümeleri için tüm öznitelik seçme yöntemleri ile seçildiğini görülmektedir. Yine çizelgeye göre ÖzVMer ve Sayfa Sıralama özniteliklerinin için tüm veri kümeleri için neredeyse tüm öznitelik seçme yöntemi ile seçildiğini görülmektedir. Eksant ve Çeksiz özniteliklerinin çok az case'de öznitelik seçme modülü tarafından işaretlenmiştir.

Çizelge 3.5. Veri kümeleri için seçilen grafik tabanlı öznitelikler

Öznitelik	Inspec				Semeval-2017				500N-KPCrowd				DergiParkEn				TRDizinEn			
	EA	LS	GA	SR	EA	LS	GA	SR	EA	LS	GA	SR	EA	LS	GA	SR	EA	LS	GA	SR
AraMer	✓	✓		✓	✓			✓	✓	✓		✓	✓			✓	✓	✓		✓
YakMer	✓	✓		✓	✓			✓	✓			✓	✓			✓	✓			✓
ÖzVMer	✓	✓		✓	✓	✓		✓	✓	✓		✓	✓	✓		✓	✓	✓		✓
DerMer	✓				✓				✓		✓		✓	✓	✓		✓			
Sayfa Sıralama	✓	✓		✓	✓	✓		✓	✓	✓		✓	✓	✓		✓	✓	✓		✓
KümKat		✓				✓					✓			✓				✓	✓	
DışMer		✓																		
YaDel	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Çeksiz		✓								✓	✓			✓	✓					

Çizelge 3.6'da önerilen modelin performans sonuçları görülmektedir. Ayrıca, literatürde kullanılan Çok Merkezili İndis – Temel Bileşen (ÇMİ-TB*) (Vega-Oliveros ve diğerleri, 2019) derecelendirme için modelin sınıflandırma performansı hesaplanmış, tabloya eklenmiştir. Sonuçlar incelendiğinde tüm verikümeleri için AraMer, ÖzVMer, Sayfa Sıralama ve YaDel özniteliklerinin yüksek performans gösterdiği görülmektedir. Buna karşı Çeksiz özniteliğinin tüm veri kümeleri için düşük bir ağırlığa sahip olduğu görülmektedir. KümKat özniteliği Inspec, DergiParkEn ve TRDizinEn veri kümeleri için düşük başarım göstermesine rağmen Semeval-2017, ve 500N-KPCrowd için ortalama sonuçlar üretmiştir. DışMer ve Çeksiz öznitelikleri çoğu veri kümesi için tek başına anlamlı bir başarım skoru elde edememiştir. ÇMİ-TB özniteliği irdelendiğinde Inspec, Semeval-2017 veri kümeleri için özniteliklerin tekil olarak sınıflandırılmasından daha yüksek başarım elde ettiği görülmektedir. Diğer 500N-KPCrowd, DergiParkEn, TRDizinEn veri kümeleri için sırasıyla Sayfa Sıralama, YaDel ve DerMer özniteliklerinin ÇMİ-TB özniteliğinden daha yüksek başarım elde ettiği görülmektedir. Neredeyse tüm veri kümeleri için özniteliklerin hepsinin birlikte kullanıldığı Tüm Öznitelikler veya Öznitelik Seçme Modulü ile ayıklanan özniteliklerin kullanıldığı yeni bir model olan Gt-AKS modelimizin en yüksek başarımlara ulaştığı görülmektedir. Inspec verikümesi için literatürde önerilen ÇMİ-TB derecelendirme modelinin en yüksek başarıma ulaştığı görülmektedir.

Çizelge 3.6. Grafik tabanlı özniteliklerin RO ile performans sonuçları

Öznitelik / Öznitelik Grubu	Inspec			Semeval-2017			500N-KPCrowd			DergiParkEn			TRDizinEn		
	Kesinlik	Duyar	F_1	Kesinlik	Duyar	F_1	Kesinlik	Duyar	F_1	Kesinlik	Duyar	F_1	Kesinlik	Duyar	F_1
Aramer	0,453	0,441	0,447	0,62	0,607	0,613	0,647	0,63	0,639	0,632	0,626	0,629	0,665	0,675	0,67
YakMer	0,413	0,19	0,26	0,567	0,569	0,568	0,611	0,571	0,59	0,546	0,231	0,325	0,547	0,186	0,278
ÖzVMer	0,452	0,458	0,455	0,612	0,622	0,617	0,644	0,631	0,637	0,634	0,642	0,638	0,664	0,684	0,674
DerMer	0,51	0,14	0,22	0,649	0,482	0,554	0,701	0,417	0,523	0,605	0,146	0,235	0,634	0,108	0,185
Sayfa Sıralama	0,452	0,452	0,452	0,616	0,619	0,618	0,661	0,651	0,656	0,635	0,642	0,639	0,665	0,679	0,672
KümKat	0,479	0,047	0,085	0,635	0,505	0,563	0,696	0,219	0,333	0,595	0,086	0,15	0,756	0,067	0,122
DışMer	0,5	0,001	0,001	0,545	0,504	0,524	0,569	0,109	0,183	0,5	0	0,001	0	0	0
YaDel	0,564	0,369	0,446	0,679	0,497	0,574	0,748	0,52	0,614	0,708	0,613	0,657	0,774	0,652	0,708
Çeksiz	0	0	0	0,527	0,729	0,611	0,565	0,01	0,02	0	0	0	0	0	0
ÇMİ-TB RF	0,456	0,46	0,458	0,615	0,622	0,619	0,639	0,641	0,64	0,638	0,632	0,635	0,671	0,677	0,674
Tüm Öznitelikler	0,64	0,38	0,477	0,697	0,594	0,641	0,801	0,611	0,693	0,81	0,623	0,704	0,886	0,67	0,763
ÇMİ-TB* Sıralama	0,493	0,484	0,488	-	-	-	0,526	0,512	0,518	-	-	-	-	-	-

Çizelge 3.6. (devam) Grafik tabanlı özniteliklerin RO ile performans sonuçları

Öznitelik / Öznitelik Grubu	Inspec			Semeval-2017			500N-KPCrowd			DergiParkEn			TRDizinEn		
	Kesinlik	Duyar	F_1	Kesinlik	Duyar	F_1	Kesinlik	Duyar	F_1	Kesinlik	Duyar	F_1	Kesinlik	Duyar	F_1
Gt-AKS (EA)	0,625	0,381	0,474	0,683	0,599	0,639	0,802	0,612	0,694	0,821	0,621	0,707	0,886	0,674	0,766
Gt-AKS (LS)	0,647	0,37	0,0471	0,673	0,587	0,627	0,799	0,605	0,688	0,8	0,634	0,707	0,9	0,663	0,764
Gt-AKS (GA)	0,564	0,369	0,446	0,678	0,5	0,576	0,72	0,637	0,676	0,763	0,64	0,696	0,775	0,654	0,709
Gt-AKS (SR)	0,636	0,378	0,474	0,695	0,572	0,628	0,81	0,605	0,692	0,81	0,626	0,707	0,909	0,664	0,765

4. BULGULAR VE DEĞERLENDİRME

Literatürde önerilen denetimsiz çıkarımsal hibrit modellerde, istatistiksel, grafik veya gömme tabanlı özellikler, veri setine bağlı olarak basit matematiksel ifadelerle birleştirilmektedir. Bu tezin ikinci bölümünde bu modeller mimari detayları ile birlikte tek tek ele alınmıştır. Denetimsiz modellere ait performans sonuçları Çizelge 2.4'te derlenmiştir. Çizelge incelendiğinde ÇMİ-TB sıralama modelinin tüm veri kümeleri için en iyi performans sonuçlarına ulaştığı görülmektedir. Bu model Inspec, Semeval-2010 ve 500N-KPCrowd veri kümeleri için sırasıyla 0,488; 0,310 ve 0,518 F_1-skor elde etmiştir.

Denetimli hibrit İfade Formatlayıcı modelinde, derin öğrenme modülünün girdi gömmesi, grafik tabanlı gömme ile sözcük gömme toplanarak hesaplanmaktadır. Çizelge 2.2'de denetimli anahtar kelime çıkarımı modellerinin performans sonuçları derlenmiştir. Sonuçlar incelendiğinde Inspec veri kümesinde F_1@5 ölçümünde en iyi sonuçları problemi bir aday ifade sınıflandırma problem olarak ele alan KA-NB, KA-AGA mimarileri elde etmiştir. Fakat bu modeller, Krapivin ve SemEval-2010 veri setlerinde düşük performans göstermiştir. ÇGİ mimarisi, SemEval-2010, NUS, Krapivin veri kümeleri için F_1@5 ve F_1@10 metrikleri için en yüksek performans değerlerine ulaşmıştır. DÇYKG-PKE mimarisi, F_1@M ölçümü için Krapivin ve KP20k veri kümelerinde çok iyi sonuçlar elde etmiştir. KA-AGA çıkarım algoritması Inspec veri kümesi için en yüksek skoru üretmiştir. ÇY-UKSB-KRA modeli Inspec veri kümesi için GloVe gömmeleri ile 0,457 F_1-skor elde ederken, BilDÇYKG kullanılarak geliştirilen BilDÇYKG ile ÇY-UKSB-KRA modeli 0,593 F_1-skoruna ulaşmıştır.

Mevcut çıkarım tabanlı hibrit algoritmalardan farklı olarak, bu tez kapsamında önerilen denetimli model HibritAKÇ, özniteliklerin basit matematiksel ilişkiler yerine, sınıflandırma algoritmaları aracılığıyla öğrenilen ağırlıklara göre sonucu etkilemesini sağlamaktadır. Bu tezde, birçok seçici özellik kullanan yeni bir dizi etiketleme modeli HibritAKÇ önerilmiştir. Belirteç sınıflandırma için metinsel, grafik tabanlı, gömme ve set öznitelikleri hibrit olarak kullanılmıştır. Bu şekilde mevcut modellerin performans sonuçlarında görülen en büyük güçlüğü, yani bir veri kümesinde yüksek sonuç veren bir algoritmanın diğer veri kümesinde düşük sonuç vermesini sorunu çözülmeye çalışılmıştır. Deneysel sonuçlar incelendiğinde, HibritAKÇ modelinin tüm veri kümelerinde yüksek performans elde ettiği görülmektedir.

Çizelge 3.4'te HibritAKÇ dizi etiketleme modelinin performans sonuçları görülmektedir. Çizelgede görüldüğü gibi 500N-KPCrowd, Semeval-2017, TRDizinEn ve DergiParkEn verikümeleri için sırasıyla 0,69; 0,69; 0,74 ve 0,70 RO sınıflandırma ile en yüksek F_1-skorlara ulaşılmıştır. Inspec verikümesi için ise 0,50 F_1-skor elde edilmiştir. Bu veri kümesi için KA-AGA mimarisi 0.607 F_1-skor ile en yüksek başarıma ulaşmıştır.

HibritAKÇ modeli, diğer metin madenciliği yaklaşımları için kullanılabilecek oldukça başarılı anahtar kelimeler çıkarmaktadır. Şekil 4.1, geliştirilen model tarafından etiketlenen anahtar kelimeleri Inspec veri kümesinden bir özet için simüle etmektedir. Gerçek pozitifler yeşil renkliyken, yanlış negatifler kırmızı renklidir. Mor, yanlış pozitif olan anahtarları temsil etmektedir.

Extending Kamp's theorem to model time granularity

In this paper, a generalization of Kamp's theorem relative to the functional completeness of the until operator is proved. Such a generalization consists in showing the functional completeness of more expressive temporal operators with respect to the extension of the first-order theory of linear orders MFO with an extra binary relational symbol. The result is motivated by the search of a modal language capable of expressing properties and operators suitable to model time granularity in omega-layered temporal structures

Keywords: *Kamp's theorem; functional completeness; until operator; temporal operators; first-order theory; linear orders; binary relational symbol; omega-layered temporal structures; model time granularity*

Şekil 4.1. Inspec veri kümesi için etiketlenmiş bir örnek

Şekil 4.2'de haber veri kümesi olan 500N-KPCrowd veri kümesinden bir örnek için modelin etiketleme sonucu görülmektedir. Hem bilimsel özetler hem de haber içerikleri sonuçların detaylı incelenebilmesi için birlikte sunulmaktadır. Deneysel sonuçlar önerilen modelin etki alanından ve veri kümesinin karakteristik diğer özelliklerinden bağımsız bir şekilde etiketleme yapabildiğini göstermektedir.

German arrested in stadium bomb plot

DORTMUND, Germany, April 1 (UPI) A 25 year old German man has been arrested for allegedly burying a cache of bombs near a German soccer stadium in a blackmail plot, authorities say. The unnamed German national was arrested in Cologne on Tuesday after allegedly placing the explosives in a parking garage near the Westfalenstadion in Dortmund, home of the Borussia Dortmund team, the Federal Office of Criminal Investigation told The Local news agency. The bombs were safely defused, and three more were found at the man's home in Krefeld, officials said. Investigators said they began tracking the man after he e-mailed the German Embassy in Pakistan, offering information about two planned attacks in Germany by a group. The warning appeared to be a blackmail bid and was worded like an unsolved attempted blackmail case last year. "The suspect apparently acted alone with a general criminal motive," a federal spokesman said. "There are absolutely no ties to terrorist or Islamist organizations." Authorities say he admitted placing the bombs. Dortmund police spokesman Michael Stein told the BBC: "We expect no security threat at all for the upcoming Bundesliga match on Saturday. Visitors are invited to come to Dortmund. They will be safe here."

Keywords: *Dortmund, unsolved attempted blackmail case, blackmail bid, safely defused, Cologne, Tuesday, cache of bombs, Westfalenstadion, explosives, parking garage, arrested, Germany, Pakistan, 25yearold German, burying, no ties to terrorist or Islamist organizations, soccer stadium, Federal Office of Criminal Investigation, German Embassy, began, terrorist, information, national, unnamed, stadium, DORTMUND, blackmail, Criminal Investigation, plot, German arrested, man, organizations Authorities, safe, bombs, cache, Bundesliga, worded, German, invited, Federal, unsolved, Saturday, apparently, April*

Şekil 4.2. 500N-KPCrowd veri kümesi için etiketlenmiş bir örnek

Daha önce önerilen grafik tabanlı modeller incelendiğinde ÇMİ-TB sıralama modelinin Inspec verikümesi için 0,488 ve 500N-KPCrowd veri kümesi için 0,518 F_1-skora ulaştığı görülmektedir. Bu model 9 grafik tabanlı özniteliği her bir belirteç için hesapladıktan TBA ile tek boyuta indirgedikten sonra sıralama yapmaktadır. Yine grafik tabanlı 6 öznitelik kullanılarak geliştirilen KA-AGA mimarisi aday ifadeleri istatistiksel bir filtre yardımı ile çıkardıktan sonra bu ifadeleri sınıflandırarak anahtar kelime çıkarımını gerçekleştirmektedir. KA-AGA mimarisi Inspec verikümesi için 0,607 F_1-skora ulaşırken 500N-KPCrowd verikümesi için 0,452 F_1-skor elde etmiştir.

Literatürde sıklıkla kullanılan 9 grafik tabanlı öznitelik kelimelerin cümle içerisindeki bağlam bilgisini içermektedir. Bu tez kapsamında grafik tabanlı öznitlikler ile Öznitelik Seçme modülü ile veri kümesine en uygun öznitelikler seçilerek ikinci bir model GtAKS önerilmiştir. Geliştirilen bu model 9 grafik tabanlı özniteliği kullanmak yerine veri kümelerinin karakteristik özelliklerine uygun olanları seçmiştir. Bu öznitelikler belirteç sınıflandırma modülüne girdi olarak verilerek anahtar kelime etiketleme görevi gerçekleştirilmiştir. GtAKS olarak isimlendirilen bu modelin performans sonuçları Çizelge

3.6'da verilmiştir. Performans sonuçları incelendiğinde GtAKS-EA ile 500N-KPCrowd, DergiParkEn ve TRDizinEn için sırasıyla 0,694; 0,707 ve 0,766 en yüksek skorları elde edilmiştir. Semeval-2017 veri kümesi için 0,641 skoru bütün grafik tabanlı öznitelikler birlikte kullanılarak elde edilmiştir. Inspec verikümesi için 0,488 skor ile ÇMİ-TB sıralama mimarisi en yüksek başarımı elde etmiştir.

Bu tez kapsamında web kazıma ile elde edilen DergiParkEn veri kümesinin Türkçe bir örneği olan DergiParkTR verikümesi karışık alana ait 600 dokümandan oluşacak şekilde derlenmiştir. GtAKS modeli kullanılarak bu veri kümesi ile 0,673 skoru elde edilmiştir. Bu model ile DergiPark İngilizce veri kümesi ile 0,704 skoru tüm öznitelikler kullanılarak elde edilmiştir. Sonuçların yakınlığı göz önüne alındığında GtAKS modeli dil karakteristik özelliği dahil olmak üzere veri kümesinin özelliklerinden bağımsız çalışmaktadır.

5. SONUÇ VE ÖNERİLER

Anahtar kelime çıkarma probleminin bazı zorlukları vardır. İlk olarak, kullanılan veri kümeleri değişken uzunlukta ve sayıda anahtar kelimeye sahiptir. Bu nedenle, önerilen bir model bir veri seti için iyi sonuçlar verirken başka bir veri seti için kötü sonuçlar vermektedir. Bu dezavantajın üstesinden gelmek için, bu tezde HibritAKÇ adlı yeni bir hibrit yöntem önerilmiştir. Model, makine öğrenimi kullanan Token Sınıflandırmasına dayanmaktadır ve anahtar kelime çıkarma problemini bir dizi etiketleme problemi olarak ele almaktadır. Metinsel, grafik tabanlı, gömülü tabanlı ve küme öznitelikleri olmak üzere dört özellik grubu oluşturulmuştur. HibritAKÇ modeli bu özelliklere dayalı olarak tüm tokenler için öğrenme gerçekleştirmektedir.

Literatürden farklı olarak farklı kategorilerdeki birçok öznitelik bir arada hibrit kullanılmakta ve öznitelikler eğitim veri setine göre sınıflandırma algoritması ile ağırlıklandırılmaktadır. RF Token Sınıflandırmasına sahip HibritAKÇ modeli, test edilen beş makine öğrenimi modeli arasında en yüksek puanı elde etmiştir. Modelin farklı veri setleriyle tutarlılığını kontrol etmek için, haber içeriklerinden oluşan başka bir veri seti (500N-KPCrowd) ve bilimsel makale özetlerinden oluşan dört farklı veri seti kullanılmıştır. HibritAKÇ, 500N-KPCrowd, Semeval-2017, TRDizinEn ve DergiParkEn veri setleri için sırasıyla 0,69; 0,69; 0,74 ve 0,70 F1 skorlarına ulaşmıştır. Sonuçlar, modelin veri seti uzunluğu, belge başına anahtar kelime sayısı ve anahtar kelime uzunluğu gibi veri seti özelliklerinden bağımsız çalıştığını göstermektedir. Ayrıca bu tez kapsamında grafik tabanlı özniteliklerin içerisinden uygun öznitelikler seçilerek oluşturulan Gt-AKS önerilmiştir. Önerilen yöntem anahtar kelime çıkarımı problemini yine bir dizi etiketleme problemi olarak ele almıştır. Sınıflandırmada kullanılan öznitelikler sık kullanılan 9 öznitelik arasından seçilmiştir. Sonuçlar incelendiğinde her bir veri kümesi için farklı bir grup grafik tabanlı özniteliğin daha ağırlıklı olarak sonuca etki ettiği bilgisine ulaşılmıştır.

KAYNAKLAR

Ajallouda, L., Fagroud, F. Z., Zellou, A., and Lahmar, E. B. (2022). KP-USE: An Unsupervised approach for key-phrases extraction from documents. *International Journal of Advanced Computer Science and Applications*, 13(4), 283-289.

Alfarra, M. R., and Alfarra, A. (2018, October). *Graph-Based Technique for Extracting Keyphrases in a Single-Document (GTEK).* 2018 International Conference on Promising Electronic Technologies (ICPET), Deir El-Balah, Palestine, 92-97.

Alzaidy, R., Caragea, C., and Giles, C. L. (2019, May). *Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents.* The World Wide Web Conference, 2551-2557.

Argamon, S., and Levitan, S. (2005, June). *Measuring the usefulness of function words for authorship attribution.* Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, 1-3.

Augenstein, I., and Søgaard, A. (2017). Multi-task learning of keyphrase boundary classification. *arXiv preprint arXiv:*1704.00514.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11), 3747-3752.

Basaldella, M., Antolli, E., Serra, G., and Tasso, C. (2018, January). Bidirectional lstm recurrent neural network for keyphrase extraction. In *Italian Research Conference on Digital Libraries*, Springer, 180-187.

Beliga, S., Meštrović, A., and Martinčić-Ipšić, S. (2014). Toward selectivity based keyword extraction for Croatian news. *arXiv preprint arXiv:*1407.4723.

Benani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., and Jaggi, M. (2018). Simple unsupervised keyphrase extraction using sentence embeddings. *arXiv preprint arXiv:*1801.04470.

Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7), 107-117.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135-146.

Bougouin, A., Boudin, F., and Daille, B. (2013, October). *Topicrank: Graph-based topic ranking for keyphrase extraction.* International Joint Conference on Natural Language Processing (IJCNLP), 543-551.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., and Jatowt, A. (2018, March). *A text feature based automatic keyword extraction method for single documents.* European Conference on Information Retrieval, Springer, 684-691.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., and Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509, 257-289.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... and Kurzweil, R. (2018). Universal sentence encoder. *arXiv preprint arXiv:*1803.11175.

Chan, H. P., Chen, W., Wang, L., and King, I. (2019). Neural keyphrase generation via reinforcement learning with adaptive rewards. *arXiv preprint arXiv:1906.04106.*

Chandrashekar, G., and Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, 40(1), 16-28.

Chen, J., Zhang, X., Wu, Y., Yan, Z., and Li, Z. (2018). Keyphrase generation with correlation constraints. *arXiv preprint arXiv:1808.07185*.

Chen, W., Chan, H. P., Li, P., Bing, L., and King, I. (2019). An integrated approach for keyphrase generation via exploring the power of retrieval and extraction. *arXiv preprint arXiv:*1904.03454.

Chen, W., Chan, H. P., Li, P., and King, I. (2020). Exclusive Hierarchical Decoding for Deep Keyphrase Generation. *arXiv preprint arXiv:2004.08511*.

Das, K., Samanta, S., and Pal, M. (2018). Study on centrality measures in social networks: a survey. *Social Network Analysis and Mining*, *8*(1), 1-11.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duari, S., and Bhatnagar, V. (2020). Complex network based supervised keyword extractor. *Expert Systems with Applications*, 140, 112876.

El-Beltagy, S. R., and Rafea, A. (2009). KP-Miner: A keyphrase extraction system for English and Arabic documents. *Information Systems*, 34(1), 132-144.

Florescu, C., Caragea, C. (2017, February). *A position-biased pagerank algorithm for keyphrase extraction.* Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, 31(1), 4923-4924.

Gero, Z., and Ho, J. (2021, June). *Word centrality constrained representation for keyphrase extraction.* Proceedings of the 20th Workshop on Biomedical Language Processing, 155-161.

Ghosh, S., Saha, C., and Molakathaala, N. (2022). NeuraGen-A Low-Resource Neural Network based approach for Gender Classification. *arXiv preprint arXiv:*2203.15253.

Graves, A., and Schmidhuber, J. (2005, July). *Framewise phoneme classification with bidirectional LSTM networks.* 2005 IEEE International Joint Conference on Neural Networks, 4, 2047-2052.

Gupta, V., and Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, 1(1), 60-76.

Haddoud, M., Mokhtari, A., Lecroq, T., and Abdeddaïm, S. (2015, June). *Accurate Keyphrase Extraction from Scientific Papers by Mining Linguistic Information*. CLBib@ ISSI, 12-17.

Hafeez, S., and Kathirisetty, N. (2022, March). *Effects and comparison of different data pre-processing techniques and ml and deep learning models for sentiment analysis: Svm, knn, pca with svm and cnn.* 2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR), 1-6.

Hulth, A. (2003, July). *Improved automatic keyword extraction given more linguistic knowledge.* Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 216-223.

İnternet: Statista Research Department Amount of data created, consumed, and stored 2010-2025. URL: https://www.statista.com/statistics/871513/worldwide-data-created/, Son Erişim Tarihi: 27.05.2022.

Kılıç, Ö. Ü., and Çetin, A. (2019, October). *A Survey on Keyword and Key Phrase Extraction with Deep Learning.* 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Samsun, Türkiye, 1-6.

Kim, S. N., Medelyan, O., Kan, M. Y., and Baldwin, T. (2013). Automatic keyphrase extraction from scientific articles. *Language Resources and Evaluation*, 47(3), 723-742.

Krapivin, M., Autayeu, A., Marchese, M., Blanzieri, E., and Segata, N. (2010, June). *Keyphrases extraction from scientific documents: improving machine learning approaches with natural language processing.* International Conference on Asian Digital Libraries, Springer, Berlin, Heidelberg, 102-111.

Lau, J. H., and Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:*1607.05368.

Liang, X., Wu, S., Li, M., and Li, Z. (2021). Unsupervised keyphrase extraction by jointly modeling local and global context. *arXiv preprint arXiv:*2109.07293.

Liu, Z., Huang, W., Zheng, Y., and Sun, M. (2010, October). *Automatic keyphrase extraction via topic decomposition.* Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 366-376.

Luo, Y., Xu, Y., Ye, J., Qiu, X., and Zhang, Q. (2021). Keyphrase Generation with Fine-Grained Evaluation-Guided Reinforcement Learning. *arXiv preprint arXiv:*2104.08799.

Marcot, B. G., and Hanea, A. M. (2021). What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?. *Computational Statistics*, 36(3), 2009-2031.

Martinc, M., Škrlj, B., and Pollak, S. (2020). TNT-KID: Transformer based neural tagger for keyword identification. *arXiv preprint arXiv:*2003.09166.

Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., and Chi, Y. (2017). Deep keyphrase generation. *arXiv preprint arXiv:*1704.06879.

Mihalcea, R., Tarau, P. (2004). *TextRank: Brining order into texts.* Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 404-411.

Mikolov T, Chen K, Corrado G, Dean J (2013) *Efficient estimation of word representations in vector space*. Proceedings of ICLR workshop.

Nguyen, T. D., and Kan, M. Y. (2007, December). *Keyphrase extraction in scientific publications.* In International Conference on Asian Digital Libraries. Springer, Berlin, Heidelberg, 317-326.

Nikzad-Khasmakhi, N., Feizi-Derakhshi, M. R., Asgari-Chenaghlu, M., Balafar, M. A., Feizi-Derakhshi, A. R., Rahkar-Farshi, T., ... and Ranjbar-Khadivi, M. (2021). Phraseformer: Multimodal key-phrase extraction using transformer and graph embedding. *arXiv preprint arXiv:*2106.04939.

Pagliardini, M., Gupta, P., and Jaggi, M. (2017). Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:*1703.02507.

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of Modern Physics*, 87(3), 925-979.

Prasad, A., and Kan, M. Y. (2019, June). *Glocal: Incorporating global information in local convolution for keyphrase extraction.* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 1837-1846.

Pennington, J., Socher, R., and Manning, C. D. (2014, October). *Glove: Global vectors for word representation.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532-1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:*1802.05365.

Ramos, J. (2003, December). *Using tf-idf to determine word relevance in document queries.* Proceedings of the First Instructional Conference on Machine Learning, 242(1), 29-48.

Rose, S., Engel, D., Cramer, N., and Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, 1, 1-20.

Sahrawat, D., Mahata, D., Zhang, H., Kulkarni, M., Sharma, A., Gosangi, R., ... and Zimmermann, R. (2020, April). *Keyphrase extraction as sequence labeling using contextualized embeddings.* European Conference on Information Retrieval. Springer, 328-335.

Sandul, M. V., and Mikhailova, E. G. (2018). *Keyword Extraction from Single Russian Document.* CEUR Workshop Proc., 2135, 30-36.

Santosh, T. Y. S. S., Sanyal, D. K., Bhowmick, P. K., and Das, P. P. (2020, April). *DAKE: Document-Level Attention for Keyphrase Extraction.* European Conference on Information Retrieval. Springer, 392-401.

Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks*, *5*(3), 269-287.

Shen, X., Wang, Y., Meng, R., and Shang, J. (2022, June). Unsupervised deep keyphrase generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10), 11303-11311.

Sun, Y., Qiu, H., Zheng, Y., Wang, Z., and Zhang, C. (2020). SIFRank: A New Baseline for Unsupervised Keyphrase Extraction Based on Pre-Trained Language Model. *IEEE Access*, 8, 10896-10906.

Sun, Z., Tang, J., Du, P., Deng, Z. H., and Nie, J. Y. (2019, July). Divgraphpointer: A graph pointer network for extracting diverse keyphrases. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 755-764.

Škrlj, B., Repar, A., Pollak, S. (2019, October). RaKUn: Rank-based Keyword extraction via Unsupervised learning and Meta vertex aggregation. *In International Conference on Statistical Language and Speech Processing.* Springer, 311-323.

Tomokiyo, T., and Hurst, M. (2003, July). A language model approach to keyphrase extraction. *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatmen*, 33-40.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 5998-6008.

Vega-Oliveros, D. A., Berton, L., de Andrade Lopes, A., and Rodrigues, F. A. (2015, July). *Influence Maximization Based on the Least Influential Spreaders.* SocInf@ IJCAI, Buenos Aires, Argentina, 3-8.

Vega-Oliveros, D. A., Gomes, P. S., Milios, E. E., and Berton, L. (2019). A multi-centrality index for graph-based keyword extraction. *Information Processing and Management*, 56(6), 102063.

Wan, X., and Xiao, J. (2008, July). *Single Document Keyphrase Extraction Using Neighborhood Knowledge.* AAAI, Chicago, 855-860.

Wang, B., Yang, B., Shan, S., and Chen, H. (2019). Detecting Hot Topics From Academic Big Data. *IEEE Access*, 7, 185916-185927.

Wang, J., Liu, J., and Wang, C. (2007, May). *Keyword extraction based on pagerank.* Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, 857-864.

Wang, Y., Li, J., Chan, H. P., King, I., Lyu, M. R., and Shi, S. (2019). Topic-aware neural keyphrase generation for social media language. *arXiv preprint arXiv:*1906.03889.

Wu, H., Liu, W., Li, L., Nie, D., Chen, T., Zhang, F., and Wang, D. (2021). UniKeyphrase: A Unified Extraction and Generation Framework for Keyphrase Prediction. *arXiv preprint arXiv:*2106.04847.

Yeom, H., Ko, Y., and Seo, J. (2019). Unsupervised-learning-based keyphrase extraction from a single document by the effective combination of the graph-based model and the modified C-value method. *Computer Speech and Language*, 58, 304-318.

Yuan, X., Wang, T., Meng, R., Thaker, K., Brusilovsky, P., He, D., and Trischler, A. (2018). One size does not fit all: Generating and evaluating variable number of keyphrases. *arXiv preprint arXiv:*1810.05241.

Zahedi, A. G., Zahedi, M., and Fateh, M. (2019). A deep extraction model for an unseen keyphrase detection. *Soft Computing*, 24, 1-10.

Zaki, M. J., Meira Jr, W., and Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. New York, Cambridge University Press, 593.

Zehtab-Salmasi, A., Feizi-Derakhshi, M. R., and Balafar, M. A. (2021). FRAKE: Fusional Real-time Automatic Keyword Extraction. *arXiv preprint arXiv:*2104.04830.

Zhang, Q., Wang, Y., Gong, Y., and Huang, X. J. (2016, November). *Keyphrase extraction using deep recurrent neural networks on twitter.* 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, 836-845.

Zhang, X., Chen, F., and Huang, R. (2018). A combination of RNN and CNN for attention-based relation classification. *Procedia Computer Science*, 131, 911-917.

Zhang, Y., Fang, Y., and Weidong, X. (2017, November). *Deep keyphrase generation with a convolutional sequence to sequence model.* 2017 4th International Conference on Systems and Informatics (ICSAI), Hangzhou, China, 1477-1485.

Zhang, Y., Jiang, T., Yang, T., Li, X., and Wang, S. (2022). *HTKG: Deep Keyphrase Generation with Neural Hierarchical Topic Guidance.* 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 1044-1054.

Zhang, Y., Tuo, M., Yin, Q., Qi, L., Wang, X., and Liu, T. (2020). Keywords extraction with deep neural network model. *Neurocomputing*, 383, 113-121.

Zhao, W. X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E. P., and Li, X. (2011, June). *Topical keyphrase extraction from twitter*. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1, 379-388.

Gazili olmak ayrıcalıktır