
GAZİ ÜNİVERSİTESİ
Gazi
1926

MAKİNE ÖĞRENMESİYLE KAZAK DİLİNDE YENİ BİR TOPLULUK ANAHTAR KELİME ÇIKARIM MODELİ

Aiman ABİBULLAYEVA

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.

Aiman Abibullayeva

23/02/2023

MAKİNE ÖĞRENMESİYLE KAZAK DİLİNDE YENİ BİR TOPLULUK ANAHTAR KELİME ÇIKARIM MODELİ

(Doktora Tezi)

Aiman ABİBULLAYEVA

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Şubat 2023

ÖZET

Anahtar kelime çıkarımı; otomatik dizin oluşturma, özetleme, sınıflandırma, kümeleme ve otomatik filtreleme gibi birçok uygulama için çözülmesi gereken temel problemlerden biridir. Diğer dillerin yanı sıra, Kazakça'da internet üzerinden bilgiler her geçen gün muazzam bir şekilde artmaktadır. Büyük miktarda metni veya makaleyi işlemek için otomatik bir anahtar kelime çıkarımı sistemi büyük talep görmektedir. Bu tez çalışmasında Kazak haber sayfalarından anahtar kelime çıkarımı için yeni bir model önerilmektedir. Topluluk Token Sınıflandırma modülünde Rastgele Orman (Random Forest), Aşırı Gradyan Artırma (XgBoost), Oylama Sınıflandırması (Voting Classification) topluluk algoritmaları ve Karar Ağacı (Decision Tree) algoritması ayrı ayrı eğitilmiş ve test edilmiştir. Önerilen yöntem, anahtar kelime çıkarımı problemini bir dizi etiketleme problemi olarak çözüyor. Önerilen modelin eğitilmesi ve test edilmesi için Kazak ve Rusça haber sayfalarından veri setleri derlenmiştir. Bu veri kümeleri üzerinde istatistiksel ve grafik öznitelikler kullanılarak yeni Topluluk Anahtar Kelime Çıkarımı (T-AKÇ) modeli önerilmiştir. Modelin başarımını ölçmek için literatürde yaygın olarak kullanılan İngilizce dilinde haber içeriklerinden oluşan 500N-KPCrowd veri kümesi için sonuçlar alınmış ve yeni derlenmiş veri kümelerinden alınan sonuçlarla karşılaştırılmıştır. Önerilen model ile, 500N-KPCrowd ve Rus veri kümelerinde sırasıyla 0,71 ve 0,86 F1 skoru elde edilmiştir. Kazak veri kümesi için 0,97 en iyi F1 skoru ile literatürdeki en yüksek sonuca ulaşılmıştır.

Bilim Kodu	:	92432
Anahtar Kelimeler	:	Kazak dili, anahtar kelime çıkarımı, topluluk sınıflandırması, istatistiksel, grafik tabanlı.
Sayfa Adedi	:	89
Danışman	:	Prof. Dr. Aydın ÇETİN

A NOVEL ENSEMBLE KEYWORD EXTRACTION MODEL IN THE KAZAKH LANGUAGE WITH MACHINE LEARNING

(Ph. D. Thesis)

Aiman ABİBULLAYEVA

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

February 2023

ABSTRACT

Keyword extraction is one of the main problems to be solved for many text mining applications such as automatic indexing, summarization, classification, clustering and automatic filtering. The text data on the Internet in Kazakh is increasing gradually like the other languages daily. Automated keyword extraction is essential when dealing with large amounts of text or articles. In this thesis, a new ensemble model for keyword extraction from Kazakh news pages is proposed. The Ensemble Token Classification module, Random Forest, XgBoost, Voting Classification ensemble algorithms and Decision Tree algorithm are trained and tested separately. The proposed method solves the keyword extraction problem as a sequence labelling problem. Datasets from Kazakh and Russian news were compiled to train and test the proposed model. A new Ensemble Keyword Extraction Model (EnsembleKEM) was proposed using graphical and statistical features for these datasets. To measure the performance of the model, the results for the 500N-KPCrowd dataset, which consists of news content in English widely used in the literature, were used and compared with the results for the newly compiled datasets. Using the proposed model, F_1 skors of 0,71 and 0,86 were obtained in the 500N-KPCrowd and Russian datasets, respectively. The highest result in the literature was obtained with the best F_1 skor of 0,97 for the Kazakh dataset.

Science Code	:	92432
Keywords	:	Kazakh language, keyword extraction, ensemble classification, language free, statistical, graph-based.
Page Number	:	89
Supervisor	:	Prof. Dr. Aydın ÇETİN

TEŞEKKÜR

Doktora eğitimim süresince sağladığı burs desteği ile bu tezin gerçekleşmesine vesile olan Ahmet Yesevi Uluslararası Türk - Kazak Üniversitesi Mütevelli Heyet Başkanlığı ve çalışanlarına teşekkürlerimi sunarım.

Bu çalışmanın yürütülmesi sırasında desteğini, emeklerini esirgemeyen ve öğrencisi olmaktan her zaman gurur duyacağım tez danışmanım sayın Prof. Dr. Aydın ÇETİN'e sonsuz teşekkürlerimi sunarım. Tez izleme komitemde yer alarak değerli görüşleri ile araştırmanın şekillenmesini sağlayan Prof. Dr. Necaattin BARIŞÇI ve Prof. Dr. Ömer DEPERLİOĞLU hocalarıma teşekkürü bir borç bilirim.

Tez çalışmalarım süresince destekleriyle her zaman yanımda olan kıymetli arkadaşım Hüma Kılıç'a ve kıymetli çocukları Meryem ile Hüseyin'e teşekkür ederim.

Yoğun tez çalışmalarımda beni sürekli destekleyen sevgili eşim Sayan'a, hayatım boyunca beni her koşulda destekleyen, haklarını asla ödeyemeyeceğim canım annem ve babama teşekkürlerimi sunarım.

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil **Sayfa**

KISALTMALAR LİSTESİ

Bu çalışmada kullanılmış kısaltmalar ve açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklamalar
ABF-AKÇ	Anlamsal Bağlantı Farkındalık Anahtar Kelime Çıkarma
AİÇ	Anahtar İfade Çıkarımı
AKÇ	Anahtar Kelime Çıkarımı
AT-AKÇ	Ağaç Tabanlı Anahtar Kelime Çıkarma
AVT	Adlandırılmış-Varlık Tanıma
BDG	Belge Dizini Grafiği
BDMG	Bağlama Duyarlı Metin Grafiği
BGM	Belge Grafik Modeli
BKA	Bağımsız Komponent Analizi
BYA	Başlık Yönlendirmeli Ağ
ÇDBA-KRA	KRA'lerle Çok Düzeyli Bellek ağı
ÇKA	Çok Katmanlı Algılayıcı
ÇMAKÇ	Çoklu Makaleden Anahtar Kelime Çıkarma
ÇÜA	Çekişmeli Üretici Ağ
ÇY-GTB	Çift Yönlü Geçitli Tekrarlayan Birim
ÇY-UKSB	Çift Yönlü-Uzun Kısa Süreli Bellek
ÇY-UKSB+KRA	Çift Yönlü-Uzun Kısa Süreli Bellek – Koşullu Rastgele Alanlar
DAKÇ	Doküman Düzeyinde Anahtar Kelime Çıkarma
DÇYKG	Dönüştürücülerden Çift Yönlü Kodlayıcı Gösterimi
DDİ	Doğal Dil İşleme
G	Doğru Negatif

Kısaltmalar	Açıklamalar
G	Doğru Pozitif
DVM	Destek Vektör Makineleri
ESA	Evrişimli Sinir Ağı
GEA	Grafik Evrişimli Ağ
GTB	Geçitli Tekrarlayan Birim
GTBM	Grafik Tabanlı Benzerlik Matrisi
GT-YZS	Grafik Tabanlı Yoğunluk Zirveleri Sıralaması Yaklaşımı
HKKA	Hiperlink Kaynaklı Konu Arama
HO-AKÇ	Hızlı Otomatik Anahtar Kelime Çıkarımı
KGP	Kelime Grafik Puanı
KG-RYS	Kelime Gömmeleri ve rastgele yürüyüş Sıralama
Kopya-ESA	Kopya ESA
K-ÖSA	Kapsama Özyinelemeli Sinir Ağı
KP	Konuşmanın Parçası
KRA	Koşullu Rastgele Alanlar
LDA	Lineer Diskriminant Analizi
MGKA-AKÇ	Birden Fazla Makale için Grafik Konvolüsyon Ağı'na dayalı Anahtar Kelime Çıkarma
MT-AKÇ	Yayılma Tabanlı Anahtar Kelime Çıkarma
ÖSA	Özyinelemeli Sinir Ağı
SGA	Stokastik Gradyan Azaltması
SYO	Seç, Yönlendir ve Oluştur
T-AKÇ	Topluluk Anahtar Kelime Çıkarımı
TBA	Temel Bileşenler Analizi
TB-GTAKÇ	Tek Bir Belgede Anahtar İfadeleri Çıkarmak için Grafik Tabanlı Teknik

Kısaltmalar	Açıklamalar
TDK	Tekil Değer Kompozisyonu
TDY-AKÇ	Tek Derin Yürüme Anahtar Kelime Çıkarma
TF	Terim Frekansı
TF-TDF	Terim Frekansı-Ters Döküman Frakansı
TO-AKÇ	Tam Otomatik Anahtar Kelime Çıkarma
UKSB	Uzun Kısa Süreli Bellek
YN	Yanlış Negatif
YP	Yanlış Pozitif

1. GİRİŞ

Günlük hayatta internette bilgi taramak bilgisayar kullanıcıları için ortak bir faaliyet haline dönüşmüştür. Her geçen gün bilgi kalabalığı içerisinde istenilen bilgiye ulaşmak da giderek zorlaşmaktadır. Her gün internette binlerce internet haberi yayınlandığından ilgili dokümanın etkili bir şekilde alınması ve özetlenmesi zordur. Bu nedenle, belirli bir web sayfanın ana içeriğini sağlamak için anahtar kelime veya anahtar ifade çıkarımı tekniği kullanılır.

Anahtar kelime çıkarımı belgenin konusunu uygun şekilde temsil edebilecek bir belgeden anahtar terimler, anahtar ifadeler, anahtar bölümler veya anahtar kelimeler belirleme işlemidir. Anahtar kelime çıkarımı metin madenciliğinin önemli bir parçasıdır (Birdevrim, Boyacı ve S Al Thani, 2018). Metin işleme alt konuları arasında metnin özetlenmesi, metni karakterize edecek anahtar kelimelerin belirlenmesi gibi uygulamaları içerir. Anahtar kelime çıkarımı bilgi erişim sistemleri, dijital kütüphane araştırması, web içeriği yönetimi, belge kümeleme ve metin özetleme gibi çeşitli doğal dil işleme uygulamalarında kullanılmıştır. Manuel olarak atanan veya otomatik olarak seçilen anahtar kelime grupları kullanıcı için metin içeriği hakkında genel bir fikir oluşturmak için kullanılır.

Literatürde anahtar kelime;

- en önemli bilgileri içeren indeks terimleri (Kaur ve Gupta, 2010),
- okuyucu için içerik hakkında özet bir bilgi veren terim kümesi (Liu, Huang, Zheng ve Sun, 2010),
- bir dokümanın ana başlıklarını yakalayan kelimeler (Kim, Medelyan, Kan ve Baldwin, 2013),
- metnin ana fikrini veya başlığını yakalayan küçük parçalar (Awajan, 2014),
- metni en temiz şekilde gösteren kelime veya kelime grupları (Birdevrim ve diğerleri, 2018),
- bir dokümanda tartışılan ana başlığı yakalayan ifadeler (Basaldella, Antolli, Serra ve Tasso, 2018),
- belgenin içeriğinin tüm önemli yönlerini ifade eden kelimeler (Papagiannopoulou ve Tsoumakas, 2020),
- metinin içeriği hakkında bilgi veren kelimeler (Ünlü ve Çetin, 2019)

olarak tanımlanmıştır.

Genellikle anahtar kelimeler bir belgenin yazarları veya yayıncıları tarafından her gün manuel olarak seçilir. Anahtar kelimelerin profesyonel dizinleyiciler tarafından manuel olarak çıkarılması zaman alıcı, sıkıcı ve pratik değildir. Bu nedenle, anahtar kelimelerin dokümanlardan çıkarılması için otomatik teknikler belirlenmelidir.

Anahtar kelimeler ve ifadelerin çıkarılmasını otomatikleştirme sorununu çözmek için iki ana yaklaşım arasında ayrım yapmak gerekir: anahtar ifadelerin atanması (keyphrase assignment) ve çıkarılması (keyphrase extraction) (Nguyen ve Kan, 2007). Anahtar kelime çıkarımında belgedeki sözcükler, frekans ve uzunluk gibi özelliklere dayanarak görünüşte anlamlı olanları tanımlamak için analiz edilir. Anahtar kelime atamasında, anahtar kelimeler, terimlerin kontrollü bir terimler sözlüğünden seçilir ve belgeler, içeriğine göre kelime öğelerine karşılık gelen sınıflara sınıflandırılır. Anahtar kelime atamasında, anahtar kelimeler kontrollü bir terim sözlüğünden veya önceden tanımlanmış sınıflandırmadan seçilir ve belgeler içeriklerine göre sınıflara ayrılır. Anahtar kelime çıkarımında ise doğrudan metinden anahtar bilgilerin seçilmesini içermesidir. Her iki yaklaşımda makine öğrenmesi metotları kullanır ve anahtar ifadelere sahip eğitim dokümanlarına ihtiyaç duyulmaktadır.

Anahtar kelime çıkarımı konusunda önerilen bir çok yöntem ve algoritmalar vardır. Bu yöntemler denetimli ve denetimsiz yaklaşımlar olmak üzere iki ana başlık altında toplanabilir. Denetimli anahtar kelime çıkarımı algoritmaları tarih sırasına göre incelendiğinde 2010 yılına kadar makine öğrenmesi ve sınıflandırma algoritmaları ile çözülürken son yıllarda derin öğrenme temelli yöntemler yaygınlaşmaya başlanmıştır. Ancak, denetimsiz yaklaşımlara dayanan anahtar kelime çıkarımı algoritmaları da yıllar içinde geliştirilmiştir. Makine öğrenmesi olarak da bilinen denetimli öğrenme yaklaşımı, eğitim kümeleri kullanılarak anahtar kelimeler çıkarır ve geliştirilen model farklı bir veri kümesi üzerinden test edilir. Uygun bir model oluşturulduktan sonra, yeni dokümanlarda anahtar kelimeleri bulmak için kullanılır (Beliga, 2014). Öte yandan, denetlenen öğrenme yöntemlerinin geniş bir eğitim kümesi gerektirmesi nedeniyle modeli oluşturmak kolay değildir. Bu kümenin bulunmadığı durumlarda, denetimsiz ve yarı denetimsiz öğrenme metotları kullanılmaktadır. Her iki yaklaşım içinde literatürde anahtar kelime çıkarma algoritmaları İngilizce dilindeki dokümanlar için başarılı bir şekilde geliştirilmiş ve uygulanmıştır.

Literatürde anahtar kelime çıkarımı için algoritmalar kullanılarak modeller oluşturulmuş, elde edilen modellerin tahminleme başarımları karşılaştırılarak kullanılan veri kaynağında hangi algoritmanın daha başarılı modeller oluşturduğu irdelenmiştir. Bugüne kadar İngilizce ve diğer dillerde çok sayıda anahtar kelime çıkarımı konusunda çalışmalar yeterince olmasına karşın Kazakça dili için durum farklıdır.

Kazak dilinde anahtar kelime çıkarımı konusu henüz ele alınmamıştır. Kazakça dili için makine öğrenme ve derin öğrenme yöntemleri ile doğal dil işleme çalışmaları sınırlıdır. Kazakça metinlerden anahtar kelime çıkarımı için derin öğrenme ile eğitilmiş bir model henüz bulunmamaktadır. Kazakça dili için sınıflandırma ve kümeleme çalışmaları yeni başladı ancak bugüne kadar anahtar kelime çıkarımı konusunda hiç çalışma bulunmamaktadır. Bugüne kadar Kazak dilini temsil edecek bir külliyat olmadığından çok fazla çalışma yapılmamıştır. Ancak, Kazak dilinin yapısal sisteminin diğer dillerle karşılaştırıldığı, kişisel benzerlik ve farklılıklarını gösteren karşılaştırmalı-tipolojik yönde birçok bilimsel çalışma bulunmaktadır. Kazakistan Cumhuriyeti Kültür Bakanlığı Diller Komitesi tarafından bir Kazak külliyatı başlatılmıştır (CLMCRK, 2009). Ancak bu külliyatın boyutu küçüktür ve geliştirmenin çok erken aşamasında kaldığı için halka açık değildir.

Daha önce anahtar kelime çıkarımı için derin öğrenme çalışmaları, istatistiksel ve graf tabanlı çalışmalarına göre daha yakın döneme aittir. Metinler üzerinde istatistiksel incelemeler konusunda kullanılan anahtar kelime çıkarımı algoritmaları Terim Frekansı (TF-Term Frequency) (Hong ve Zhen, 2012), Terim Frekansı-Ters Döküman Frakansı (TF-TDF-Term Frequency-Inverse Document Frequency- TF-IDF) (Salton ve Buckley, 1988), Ki-kare (Kira ve Rendell, 1992) ve temel öğrenme algoritmaları Naïve Bayes (Domingos ve Pazzani, 1997), Destek Vektör Makineleri (DVM-Support Vector Machines-SVM) (Zhang, Xu, Tang, Li, 2006), Lojistik Regresyon (Kantardzic, 2011), Rastgele Orman (Random Forest) (Breiman, 2001) ile kapsamlı bir çalışmalar yapılmıştır. Önceki yıllarda daha çok makine öğrenmesi tabanlı sınıflandırma algoritmaları kullanılırken derin öğrenme alanında ilerlemeler kaydedildikçe çalışmaların yoğunluğu bu alana doğru kaymaya başlamıştır. Doğal dil işleme konusunda metin özetleme, sınıflandırma, soru-cevap gibi problemlerin çözümünde derin öğrenme algoritmaları arasında Özyinelemeli Sinir Ağı (ÖSA-Recurrent Neural Network-RNN) (Sutskever, Martens and Hinton, 2011), Uzun Kısa Süreli Bellek (UKSB - Long Short-Term Memory-LSTM) (Hochreiter ve Schmidhuber, 1997), Çift

Yönlü-Uzun Kısa Süreli Bellek - Koşullu Rastgele Alanlar (ÇY-UKSB+KRA - Bidirectioanal LSTM - Conditional Random Fileds-Bi-LSTM-CRF) (Basaldella ve diğerleri), Evrişimli Sinir Ağı (ESA-Convolutional Neural Network-CNN) (Kim, 2014) mimarileri etkileyici sonuçlar vermiştir. Son zamanlarda Google'ın doğal dil işleme eğitimi için kullanılan sinir ağı temelli bir tekniği Dönüştürücülerden Çift Yönlü Kodlayıcı Gösterimleri (DÇYKG - Bidirectional Encoder Representations from Transformers-BERT) (Devlin, Chang ve Toutanova, 2018) modelidir. DÇYKG anahtar ifade çıkarımı algoritmalarında sıklıkla kullanılmaktadır.

Anahtar kelime çıkarımını sekans etiketleme problemi olarak çözmek için ÇY-UKSB+KRA modeli (Alzaidy, Caragea ve Giles, 2019) önerilmiştir. Modelde, denetimsiz bir yöntem olan KRA çıkış katmanında kullanılmaktadır. ÇY-UKSB+KRA tabanlı Doküman Düzeyinde Anahtar Kelime Çıkarma (DAKÇ - Document-Level Attention Keyword Extraction-DAKE) (Santosh, Sanyal, Bhowmick ve Das 2020) modelinde, doküman düzeyinde mekanizması ile belge düzeyinde bağlamsal bilgiler ağa dahil edilmiştir. Birden fazla makaleden anahtar kelime çıkaran Çoklu Makaleden Anahtar Kelime Çıkarma (ÇMAKÇ - Multiple Article Keyword Extraction - M-GCKE) (Wang, Yang, Shan ve Chen, 2019) algoritması, anahtar kelimeleri daha doğru bir şekilde çıkarmak için, anahtar kelimeler arasındaki ilişkiyi tek bir kağıttan birden çok kağıda genişleten bir yöntemdir. Anahtar kelimeleri daha kesin olarak tahmin etmek için, ağdaki yapısal bilgileri ve düğüm nitelik bilgilerini öğrenmek için Grafik Evrişimli Ağ (GEA - Graph Convolutional Network-GCN) kullanılmıştır. Anahtar İfade Çıkarımı (AİÇ - Keyphrase Generation - KG) kütüphanesi (Chen et.al, 2019) iki kodlayıcı, bir çıkarıcı, bir kurtarıcı ve bir birleştirme modülünden oluşmaktadır.

Kazak haber metinlerden otomatik anahtar kelime çıkarımı hala yeni bir araştırma alanıdır. Bu alan, çevrimiçi Kazakça içeriğin boyutundaki önemli artış ve mevcut elektronik belgelere elle atanmış anahtar kelimelerin nadir olması nedeniyle daha fazla çaba gerektirir. Ulaşabildiğimiz kaynaklara göre literatürde Kazak dilinde yapılan çalışmalar sınırlı sayıda olup anahtar kelime çıkarımı üzerine bir yayın yer almamaktadır. Bu konuda literatürde yer çalışmalar da genel olarak sınıflandırma ve kümeleme çalışmalarıdır. Bu nedenlerle haber metinlerinden anahtar kelime çıkarma alanında çalışma yapılması ve önerilerde bulunması önem arz etmektedir.

Motivasyon

Bu tez çalışmasını motive eden sebepler şunlardır:

- Kazakistan Cumhuriyeti'nin kiril alfabesinden latin alfabesine geçilmesi durumunda arşiv belgelerine erişimde problemler yaşanabileceği öngörülmektedir. Belgelere erişim için ve belge bağlantılama yapılabilmesi için metinden anahtar kelime çıkarımı önemlidir. Belgelere ait anahtar kelimeler belgeye erişimi kolaylaştıracaktır.
- İngilizce başta olmak üzere birçok dilde anahtar kelime çıkarımı problemleri ile ilgili makine öğrenme yöntemleri kullanan çalışmalar yapılmış olmasına rağmen Kazakça için bu alanda henüz çalışma yapılmamıştır.
- Kazakçanın sondan eklemeli bir dil olması aynı kökten birden fazla kelime oluşturabilmesinden kaynaklanan seyreklik probleminden dolayı doğal dil işleme çalışmalarındaki ilerlemelerin aynı problemde çalışılan diğer dillerden geride kaldığı görülmüştür.

Kazakça'nın dil yapısı

Kazakça haber metinleri üzerinde gerçekleştirilen bu tezde, öncelikle Kazakça dilinin "Dil yapısından" bahsedilmiştir. Doğal dil işlemenin alt konularından biri olan anahtar kelime çıkarımına ilişkin bu tez Kazak diline dayanmaktadır. Kazakistan Cumhuriyetinin Ana dili olan Kazakça, Türk dili dünyasının güney Kıpçak grubundandır. Dolayısıyla bu özelliklerinin getirdiği bazı zorluklar doğal dil işleme konusunda da ortaya çıkmaktadır. Kazak dili düşük kaynaklı dillerden biridir ve sondan eklemeli diller grubuna aittir. Yazı tarihinde, Kazak dilinde kullanılan alfabe sistemi çeşitli tarihsel dönemlerden geçerek milli alfabe düzeyine ulaşmıştır.

Kazak halkının yüzyıllardır Arap grafiklerine dayalı alfabe sistemini kullandığı bilinmektedir. 1929'dan 1940'a kadar Latin alfabesine dayalı alfabe yazma sistemine dahil edildi ve 1940'tan beri Kiril alfabesi kullanılıyor. 2017 yılında, Kazak dilinin yeni Latin alfabesi, 26 Ekim'de Kazakistan Cumhuriyeti Cumhurbaşkanı'nın kararnamesi ile onaylandı. 2017-2025 yıllarında yeni bir alfabeye geçilmesi planlanmaktadır. Şu anda, Kiril'den Latince'ye geçiş konusu toplumda yaygın olarak tartışılmaktadır. Tüm ileri teknolojilerin dili haline gelen Latin alfabesine geçiş Kazak sanat ve kültürü için önem taşımaktadır.

Kazakistan'ın Latin alfabesine geçmesi Kazak dilinin dünya medeniyetinde hak ettiği konuma yükselmesi yanısıra hem sosyo-ekonomik, hem de siyasi açıdan önemlidir.

Alfabe değişikliğinin en önemli problemi eski mirasın unutulmasıdır. Alfabe değiştiğinde o alfabede yazılan metinlere erişim güçleşir ve zamanla geçmiş ile olan bağ azalmaya başlar. Örneğin, günümüzde insanlar benzer nedenlerle eski Türk veya sarı Uygur alfabesindeki eserleri okuyamamaktadır. Bu nedenle kiril alfabesinden latin alfabesine geçilmesi durumunda belge bağlantılama yapılabilmesi için anahtar kelime çıkarımı önem kazanmaktadır. Belgelere ait anahtar kelimelerin varlığında indeksleme, özetleme kümeleme, sınıflandırma ve benzeri yöntemlerle belgeye erişim kolaylaşmaktadır.

Kazak dilinin gramerinin kendine has özellikleri vardır. Kazak dili, hem Kazak halkının ana dili hem de Kazakistan Cumhuriyeti'nin devlet dilidir. Bu dil - sentetik sondan eklemeli (aglutinatif) dil türüne ait Kıpçak grubunun bir çağdaş Türk dilidir, zengin ve karmaşık bir morfolojiye sahiptir. Kazakça tüm dünyada 10 milyondan fazla insanın ana dilidir. Bu dil Türkçe ve Özbekçeden sonra en çok konuşulan üçüncü Türk dilidir. Söz varlığının %60'ı Türkçe kökenli kelimelerden oluşmaktadır. Kök genellikle temeli oluşturur ve ona en az iki veya üç ekin birleştirilmesi yöntemiyle kelimeler oluşturulur. Kök formuna ek eklenerek oluşturulduğu Kazak diline sondan eklemeli dil denir. Kazakçanın; eklemeli dil yapısına sahip olması sebebiyle, kelimelerin çoğunlukla bir kök ve kökün sonuna eklenmiş ekler tarafından oluşturulan bir yapısı vardır. Kökün sonuna eklenecek eklerle yeni bir kelime türetilebilir. Bu nedenle, sondan eklemeli bir dilde tek bir kelime, sondan eklemeli olmayan bir dilde birkaç kelimeden oluşan bir söz öbeğine karşılık gelebilir.

Genel morfolojik konfigürasyon tanımı şöyle görünür:
[kök] + [sonek] + [son ekler]

Çizelge 1.1. Kazak dilinde kelimeye ek ekleme örneği

Kazakça	Türkçe	İngilizce
üy	ev	home
üy+i+m	evim	my home
üy+i+m+de	evimde	in my home
üy+i+m+de+min	evimdeyim	I am in my home

Kazakça bir kelime olan “üy (ev)”, Türkçe kelime olan “ev” ve tek bir İngilizce kelime olan “home” Çizelge 1.1'de gösterildiği gibi tam bir cümleye karşılık gelebilir. Kazakça’da sınıflandırma ve kümeleme çalışmaları oldukça yenidir. Bu nedenle Kazak dili için yeterince büyük bir külliyat henüz oluşturulamamıştır. Bu nedenle Kazak dilinde yapılacak olan çalışmalar için bir veri setine ihtiyaç duyulmaktadır. Günlük yaşamı takip edebilmek, geçmişte olan olaylara dair kayıtlara erişebilmek için Kazak haber sitelerindeki yer alan metinlerden oluşan bir veri kümesi oluşturulması çalışmalar için bir başlangıç noktası olarak değerlendirilmiştir.

Bu tez çalışmasında, Kazakça anahtar kelime çıkarımı için haber sitelerinden elde edilen bir veri kümesi oluşturulmuştur. Günlük bilgi üreten Kazak haber sitelerinden elde edilen veri kümesi ile anahtar kelime çıkarımı için KazakhNews veri kümesi derlenmiş, istatistiksel ve grafiksel öznitelikler kullanılarak yeni bir Topluluk Anahtar Kelime Çıkarım (T-AKÇ) modeli önerilmiştir. Model KazakhNews veri kümesi kullanılarak eğitilmiş ve test edilmiştir. Ayrıca, karşılaştırma yapmak amacıyla anahtar kelime çıkarma konusunda diğer dillerde yapılmış çalışmalar incelenmiştir. Model, latin alfabesinden farklı aynı alfabeyi kullanan Rus dilinde anahtar kelime çıkarma işlemine tabi tutulmuştur. Bu amaçla model Rusça haber içeriklerinden yeni derlenen RussianNews veri kümesi için de eğitilmiş ve test edilmiştir. Ayrıca model literatürde sık kullanılan İngilizce haber içeriklerinden oluşan 500N-KPCrowd veri kümesi için de eğitilmiş ve test edilmiştir. Önerilen modelin başarım sonuçları üç farklı dil grubu için tablo halinde sunulmuştur.

Önerilen modelin literatüre olan temel katkıları şu şekilde özetlenebilir:

- Bu tezde sınıflandırma algoritmaları ile anahtar kelime çıkarımı bir dizi etikeleme görevi olarak ele alınmıştır. Bu yaklaşım mevcut çıkarım temelli algoritmalardan farklı olarak topluluk bir yöntem olarak tasarlanmıştır.
- Tasarlanan yöntemde mevcut yöntemlerde kullanılan özniteliklerin basit matematiksel ilişkileri yerine sınıflandırma algoritmaları aracılığıyla öğrenilen ağırlığa göre sonuca etki etmesi sağlanmıştır.
- Akademik literatürde erişebildiğimiz kaynaklar itibarıyle bu amaçla yapılmış herhangi bir yayın bulunmamaktadır. Çalışma bu yönüyle değerlendirildiğinde literatürde bir ilk olma özelliğini taşımaktadır.

- Kazak dilinde daha önce anahtar kelime çıkarımı konusunda çalışma olmadığı için bu araştırmanın akademik literatüre katkı sağlamasının yanı sıra Kazak dilinde Kiril alfabesi ile yazılan eserlere erişime de kolaylık sağlayacağı, bu şekilde kültürel mirasın korunmasınan destek olunacağı düşünülmektedir.
- Kiril alfabesinden latin alfabesine geçişte veritabanalarında geçmişe yönelik yapılacak araştırmalarda istenilen bilgiye daha kısa sürede (daha doğru ve etkin bir şekilde) erişim sağlanacaktır.
- Kiril alfabesinden Latin alfabesine geçişten sonra, anahtar kelimelerin türetilmesi belgeler arasında bağlantı kurılabilmesini sağlayacaktır. Bu sayede belgelere ulaşım ve içerik arama kolaylaşacaktır.

Tezin amacı

Bu tezin amacı, Kazakça metinden anahtar kelime çıkarımı için yeni bir model önermek ve elde edilen sonuçları daha önce var olan yöntemlerle kıyaslamaktır.

Tezin yapısı

Tezin bundan sonraki bölümü şu şekilde yapılandırılacaktır:

Bölüm 2'de literatürde anahtar kelime çıkarımı problemini çözmek amacıyla geliştirilen modeller ve yaklaşımlar detaylı bir şekilde incelenmiştir. Bölüm 3'te anahtar kelime çıkarımı için geliştirilen Topluluk Anahtar Kelime Çıkarımı (T-AKÇ) modeli verilmiştir. Bölüm 4'te çalışma ortamının detayları verildikten sonra modelin performans sonuçları tablo halinde sunulmuş, KazakhNews, RussianNews ve 500N-KPCrowd gibi veri kümelerinin performans sonuçlarıyla önerilen modelin sonuçları karşılaştırılmıştır. Bölüm 5'te tezin sonuçları ve önerileri sunulmaktadır.

2. LİTERATÜR İNCELEMESİ

Bu bölümde anahtar kelime çıkarımı problemleri üzerinde daha önce yapılmış çalışmalara değinilecektir. Bu amaçla öncelikle Kazakça ve Rusça yapılmış çalışmalar hem de İngilizce çalışmalar incelenecektir. Öncelikle anahtar kelime çıkarımı problemleri ile ilgili son yıllarda yapılmış güncel çalışmalar ve geliştirilen modeller anlatılacak, sonrasında Kazakça için yapılmış çalışmalara değinilecektir.

Kazakça dili için makine öğrenme, derin öğrenme yöntemleri ile doğal dil işleme çalışmaları sınırlıdır. Kazakça belgeler için anahtar kelime çıkarımı araştırılmamış bir konudur ve ilgili çok az sayıda yayın vardır. Kazakça dili açısından incelediğimizde makine öğrenme ve derin öğrenme yöntemleri ile sınırlı sayıda doğal dil işleme çalışması yapılmasına rağmen, Kazak dilinde anahtar kelime çıkarımı konusu Abibullayeva ve Çetin (2022) tarafından gerçekleştirilen bir çalışma dışında henüz ele alınmamıştır. Bu çalışmada, Kazakça veri kümesi için *bert-base-multilingual-uncased* dil modeli ile 0,24 F1Skoru başarım değerleri elde edilmiştir. Bir cümleden anahtar kelimeleri çıkarımı için DÇYKG Token Sınıflandırma Modeli kullanılmıştır. DÇYKG Token Sınıflandırma modeli ile dizi etiketleme tabanlı anahtar kelime çıkarımı algoritmaları kazakça veri setleri için eğitilmiş ve sonuçlar çıkarılmıştır.

Derin öğrenme yöntemlerinin çeşitli problemlere başarı ile uygulanması, anahtar kelime çıkarımı için bu yöntemlerin kullanılmaya başlamasında etkili olmuştur. Son zamanlarda, topluluk tabanlı makine öğrenme yöntemleri, daha iyi sonuçlar elde ettikleri için anahtar kelime çıkarımında çok dikkat çekmiştir. Topluluk yöntemi - aynı anda birden fazla makine öğrenmesi algoritmasını kullanarak problemi çözmeye çalışır. Topluluk yöntemi içerisindeki her bir metod, aynı girdi ile bağımsız olarak yürütülür ve daha sonra sonuçların birleştirilmesiyle çıktıya karar verilir. Topluluk yöntemi bu metodun içerisine dahil olan modellerden daha iyi performas sergilediği gözlemlenmiştir (Onan, Korukoğlu ve Bulut, 2016). Topluluk öğrenimi, makine öğrenimi araştırmasının gelecek vaat eden bir araştırma yönüdür. Yang, Zhang ve Li (2011), metin akışlarını sınıflandırmak için toplu öğrenmeye dayalı bir yaklaşım sunmuşlardır. Metin belgelerinin manuel olarak etiketlenmesine olan ihtiyacı ortadan kaldırmak için özellik olarak anahtar kelimeler kullanmıştır. Bu şemada, temel öğrenenler, anahtar kelimeler ve etiketlenmemiş belgelerle oluşturulmuştur. Ayrıca,

metin veri akışlarının kavram kaymasıyla başa çıkmak için bir topluluk algoritması sunulmuştur.

Anahtar kelime çıkarımı algoritmaları literatürde bir çok farklı dilde çalışılmıştır. Bu algoritmalar özellikle İngilizce ve diğer Avrupa dillerindeki belgeler için başarıyla geliştirilmiş ve uygulanmıştır. Bununla birlikte, Kazakça belgeler için anahtar kelime çıkarımı konusu az araştırılmış bir konudur. Kazakça metinler için her ne kadar anahtar kelime çıkarımı problemine yönelik model geliştirilmemiş olsa da doğal dil işleme problemlerini çözmek amacıyla geliştirilen yaklaşımlar bulunmaktadır.

Kessikbayeva ve Çiçekli (2014) sondan eklemeli dillerde doğal dil işleme ile ilgili görevler için çok önemli bir konu olan morfolojik analizi incelemişler ve Kazakçanın morfolojik açıdan ilk detaylı hesaplamalı analizi yapılmıştır. Bekbulatov ve Kartbayev (2014) Kazakça morfolojik bölümleme ile ilgili gazete, derlem üzerine araştırma sunmuş ve onu denetimsiz kural tabanlı dil işleme modelleriyle karşılaştırmıştır.

Myrzakhmetov ve Kozhirbayev (2018) çalışmalarında geleneksel n-gram ve UKSB tabanlı sinir ağlarını kullanarak bir gazete veri kümesi üzerinde dil modelleme deneyleri yürütmüş ve sinir tabanlı modellerin n-gram tabanlı modellerden daha iyi performans gösterdiğini belirtmiştir.

Nugumanova ve Mansurova (2019) monografta istatistiksel ve graf tabanlı yöntemleri incelenmiştir ve terim türetme görevi ile tematik modelleme görevi arasındaki ilişkiyi ele almışlardır. Ayrıca, terimlerin otomatik olarak tanınması konusunda genel bilgiler vererek terminoloji gibi karmaşık kavramların çalışma yöntemleri tartışarak Python kütüphaneleri ve R ekosisteminde örnekler uygulamışlardır.

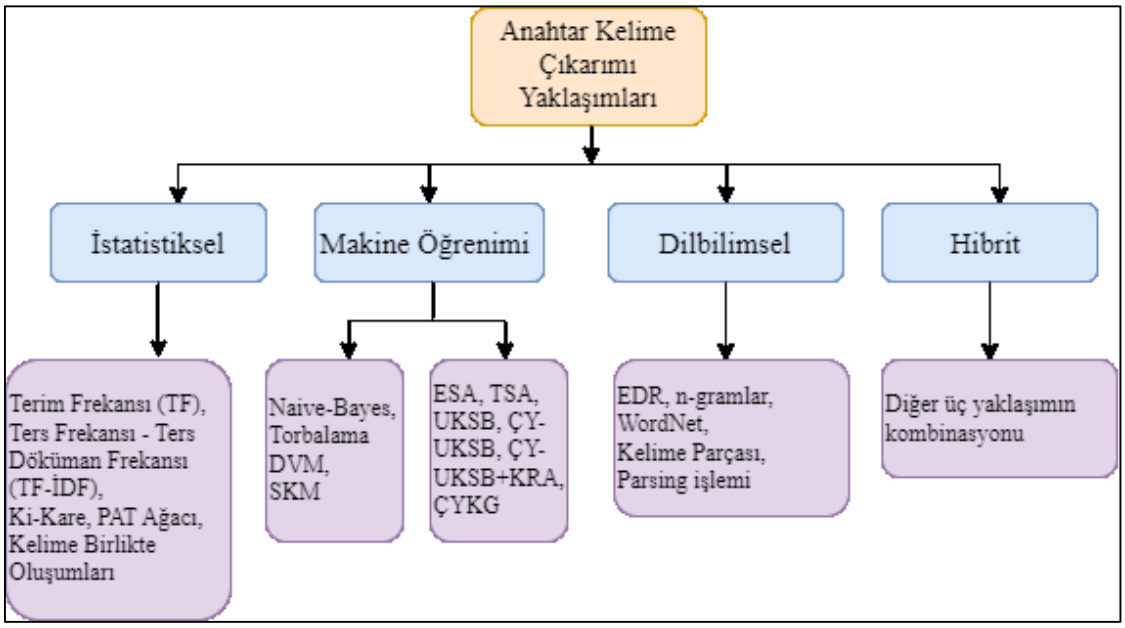
Raximova, Qasimova ve İsabaeva (2021) çalışmalarında Kazak dili için DÇYKG modeline dayalı bir soru-cevap sistemi geliştirmişlerdir. Kazak dili için hazırlanacak olan bir külliyata katkı sağlamak için 60,000 cümlelik bir külliyatı sorular ve cevapları ayrı dosyalarda oluşan İngilizceden çevrilerek bir araya getirmişlerdir.

Rusça dili için literatür incelendiğinde Rus haberleri olan Diyalog Değerlendirmesi 2021 (Dialogue Evaluation 2021) (Khaustov, Gorlova, Kalmykov ve Kabaev, 2021) veri kümesi

ile önceden eğitilmiş DÇYKG modeli ile iki yaklaşım sunulmuş ve karşılaştırılmıştır. İlk yöntem, denetimli gömme öğrenimini kullanırken, ikincisi sorunu ikili sınıflandırmaya indirgemektedir.

2.1. Anahtar Kelime Çıkarımı Yaklaşımları

Anahtar kelime çıkarımı yaklaşımları (Şekil 2.1) stil ve yönteme göre istatistiksel, dilbilimsel, makine öğrenme ve alana hibrit yaklaşımlar olarak 4 ana başlık altında incelenebilir (Siddiqi ve Sharan, 2015).



Şekil 2.1. Anahtar kelime çıkarımı yaklaşımları

2.1.1. İstatistiksel yaklaşımlar

İstatistiksel Yaklaşımlar eğitim verisi gerektirmeyen basit yöntemlerden oluşur ve bu yöntemler dilden ve etki alanından bağımsızdır. Metinler üzerinde istatistiksel öznitelikler kullanılarak kelime çıkarımı algoritmaları TF, TF-TDF, Ki-kare ve temel öğrenme algoritmalarının Naïve Bayes, DVM, Lojistik Regresyon, Rastgele Orman ile kapsamlı çalışmalar yapılmıştır. İlk anahtar kelime çıkarma çalışmaları 1950'lerde yapılmaya başlanmıştır ve bugün kullanılan birçok yöntem bu çalışmaları temel alarak yapılmaktadır. En çok kullanılan istatistiksel yaklaşım olan "Terim Frekansı" 1957 yılında Hans Peter Luhn çalışmasında ortaya çıkmıştır. Luhn (1957) çalışmasında "Bir kelime ya da kelime

grubu, makale içerisinde ne kadar fazla tekrarlanıyorsa, aynı oranda yazar o kelimeye önem veriyordur" fikrini vermiştir. Araştırma sonucunda bu yaklaşımın yanlış ve yetersiz sonuçlar verdiği fark edilmiştir. Daha sonra Spark Jones (1972) çalışmasında ters belge frekansı (TDF) kavramını açıklamış ve Luhn ve arkadaşlarının çalışmasında ortaya çıkan frekansı yüksek kelimeler listesinin aksine, normalize edilmiş daha doğru bir anahtar kelime listesi oluşturulmasını sağlamıştır. Jones çalışmasında, sık kullanılan kelimelerin ağırlıklarını azaltarak, az tekrarlanan kelimelerin ağırlık değerini artırmış ve daha homojen bir şema oluşturmuştur.

2.1.2. Dilbilimsel yaklaşımar

Dilbilimsel yaklaşımlar, kelime özelliklerini öğrenmek için cümlelerin ve belgelerin dilsel özelliklerini kullanır. Bu yaklaşım kelime analizi, sözdizimi analizi, konuşma analizi ve bu tür durumların analizinden oluşur. Bu yaklaşımın en büyük avantajı sadece bir dile değil birçok dile uygulanabilmesidir. Yani bağımsız olarak analize açık olduğu söylenebilir. Bu analizler dilbilimsel yöntemlere göre çok doğru sonuçlar vermese de istatistiksel analizlerde iyi sonuçlar vermiştir. Nguyen ve Kan (2007) gramerden anahtar kelimeleri çıkarmak için bir algoritma geliştirmişlerdir. Bu algoritma, anahtar kelimenin verilerden çıkarıldığı koşulları yakalayan bir algoritma türüdür.

2.1.3. Makine öğrenimi yaklaşımları

Makine öğrenmesi olarak da bilinen denetimli öğrenme yaklaşımı, eğitim kümeleri kullanılarak anahtar kelimeler çıkarır ve geliştirilen model farklı bir veri kümesi üzerinden test edilir. Uygun bir model oluşturulduktan sonra, yeni dokümanlarda anahtar kelimeleri bulmak için kullanılır. Öte yandan, denetlenen öğrenme yöntemlerinin geniş bir eğitim kümesi gerektirmesi nedeniyle modeli oluşturmak kolay değildir. Bu nedenle, denetimli yöntemler eğitim verileri gerektirir ve genellikle alana bağlıdır. Bu kümenin bulunmadığı durumlarda, denetimsiz ve yarı denetimsiz öğrenme metotları kullanılmaktadır.

2.1.4. Hibrit yaklaşım

Bu yaklaşım, genel olarak anahtar kelime çıkarımına yönelik daha iyi sonuçlar elde etmek için iki veya daha fazla yaklaşımın birleştirilmesidir. Ek olarak bazen terimlerin konumu, uzunluğu, düzen özellikleri, Hiper Metin İşaretleme Dili (HyperText Markup Language - HTML) ve benzeri etiketler, metin biçimlendirme bilgileri vb. gibi bilgileri de içerirler.

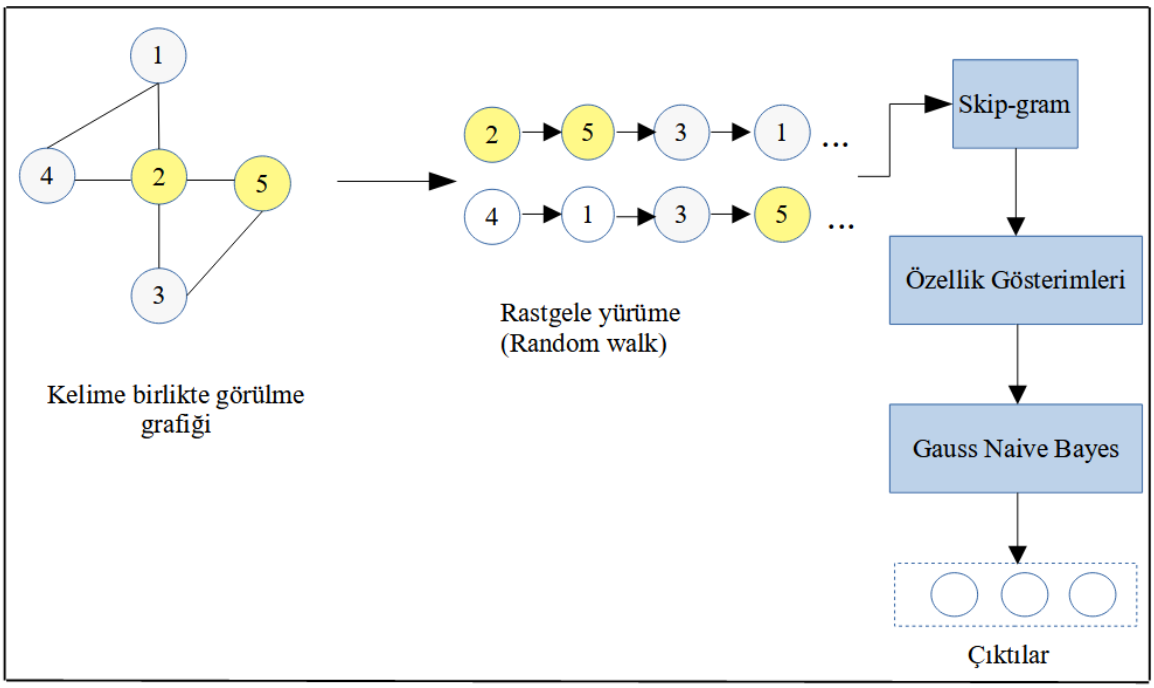
2.2. Denetimli Anahtar Kelime Çıkarımı Modelleri

Denetimli yöntemlerin temel amacı anahtar kelime çıkarımını bir ikili sınıflandırma görevine dönüştürmektedir ve etiketlenen kelime ya anahtar kelime veya değil olarak etiketlenir. Sınıfların önceden belirlenip bir eğitim seti ile sistemin eğitildiği sınıflandırma süreci "denetimli öğrenme" olarak adlandırılır. Denetimli yöntemlerin ana fikri, problemi bir sınıflandırma problemi olarak algılar ve çözüm geliştirir. Metindeki anahtar kelimeleri bulmak için Naïve Bayes, DVM, C4.5, Torbalama (Bagging) gibi algoritmalar mevcuttur. Makine öğrenmesi olarak da bilinen denetimli öğrenme yaklaşımı, eğitim kümeleri kullanılarak anahtar kelimeler çıkarır ve geliştirilen model farklı bir veri kümesi üzerinden test edilir. Uygun bir model oluşturulduktan sonra, yeni dokümanlarda anahtar kelimeleri bulmak için kullanılır. Öte yandan, denetlenen öğrenme yöntemlerinin geniş bir eğitim kümesi gerektirmesi nedeniyle modeli oluşturmak kolay değildir. Bu kümenin bulunmadığı durumlarda, denetimsiz ve yarı denetimsiz öğrenme metotları kullanılmaktadır (Birdevrim, Boyacı ve S Al Thani, 2018). Anahtar kelime çıkarımı için bir çok denetimli model önerilmiştir. Denetimli anahtar kelime çıkarımı algoritmaları tarih sırasına göre incelendiğinde 2010 yılına kadar makine öğrenmesi, sınıflandırma algoritmaları ile çözümlerin çoğunlukta olduğu görülmektedir. Uzun (2005) çalışmasında bir metindeki anahtar kelimelerin ayırt edici özelliklerini belirlemek ve bu bilgileri kullanarak metinden anahtar kelimeleri çıkarmak için makine öğrenme yöntemlerinden biri olan Naive-Baiyes yöntemini kullanmıştır. Zhang ve diğerleri (2006) çalışmalarında anahtar kelime çıkarımı için DVM temel yöntemlerden önemli ölçüde daha iyi performans gösterebileceğini göstermişler. Wang, Peng ve Hu (2006) bir cümlenin anahtar kelime olup olmadığını belirlemek için TF-TDF kullanarak ve çok katmanlı sinir ağına dayalı bir anahtar kelime çıkarma yaklaşımın kombinasyonu önermişler. Qingguo ve Chengzhi (2008) tarafından önerilen Çince aday anahtar kelime çıkarımı için K - en yakın komşu (K-Nearest Neighbor - KNN) modelini kullanarak etiketlenmiş veri kümesine dayalı bir dizi aday anahtar kelime

oluşturmuşlar. Krapivin, Autayeu, Marchese, Blanzieri ve Segata (2010) bilimsel makalelerden otomatik anahtar kelime çıkarma görevine yönelik çeşitli makine öğrenmesi yaklaşımlarını (DVM, Rastgele Orman) geliştirmek için DDİ (Doğal Dil İşleme) tekniklerini kullanmışlar. Sarkar, Nasipuri ve Ghose (2010) tarafından önerilen bir algoritmada anahtar ifade çıkarımı problemi, sınıflandırma problemi gibi değil bir sıralama problemi gibi düşünülerek Çok Katmanlı Algılayıcı (ÇKA- Multilayer perception - MLP) modeli kullanılarak çıkarılmıştır.

2.2.1. Tekli derin yürüme anahtar kelime çıkarma (TDY-AKÇ) modeli

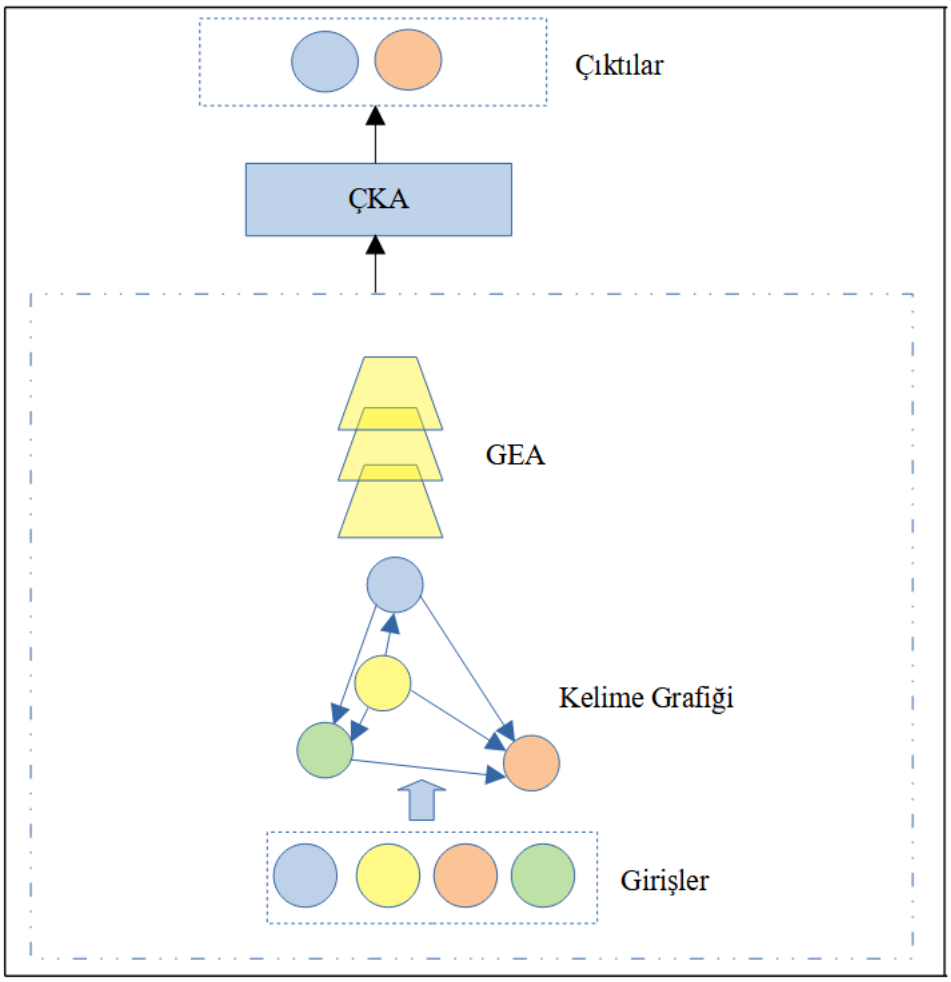
Tekli Derin Yürüme Anahtar Kelime Çıkarma (TDY-AKÇ - Single DeepWalk Keyword Extraction - S-DWKE) (Wang ve diğerleri, 2019) modeli bir giriş metni için kelime vektörlerini cümle içerisindeki kelimeleri, kelime birlikte görülme ağı üzerinden rastgele gezerek bulan bir derin öğrenme algoritması olarak adlandırılmıştır. Şekil 2.2'de TDY-AKÇ modeli TDY adımından sonra skip-gram ile öznitelikleri belirlemektedir. Son katmanda Gaussian Naive Bayes ile sınıflandırma gerçekleştirilmektedir.



Şekil 2.2. TDY-AKÇ modeli

Birden fazla makaleden anahtar kelime çıkaran Birden Fazla Makale için Grafik Konvolüsyon Ağı'na dayalı Anahtar Kelime Çıkarma (MGKA-AKÇ - Keyword Extraction

based on Graph Convolution Network for Multiple Papers - M-GCKE) algoritması, anahtar kelimeleri daha doğru bir şekilde yakalayabilmek için, bu kelimeler arasındaki ilişkiyi birden çok makaleye genişletmiştir. Şekil 2.3'te ağdaki yapısal bilgileri ve düğüm nitelik bilgilerini öğrenmek için bu modelde GEA kullanılmıştır. GEA aynı zamanda hem yapı bilgisi hem düğüm öznitelik bilgisi üzerinde öğrenme gerçekleştirebilir. Bu sayede modelin tahmini daha güçlü hale gelmiştir. GEA çıkışını doğrusal olmayan bir aktivasyon fonksiyonu ile işlemek için modele ÇKA katmanı eklenmiştir. Son katmanın çıktısı, softmax fonksiyonu ile bir olasılık değerine dönüştürülmüştür.

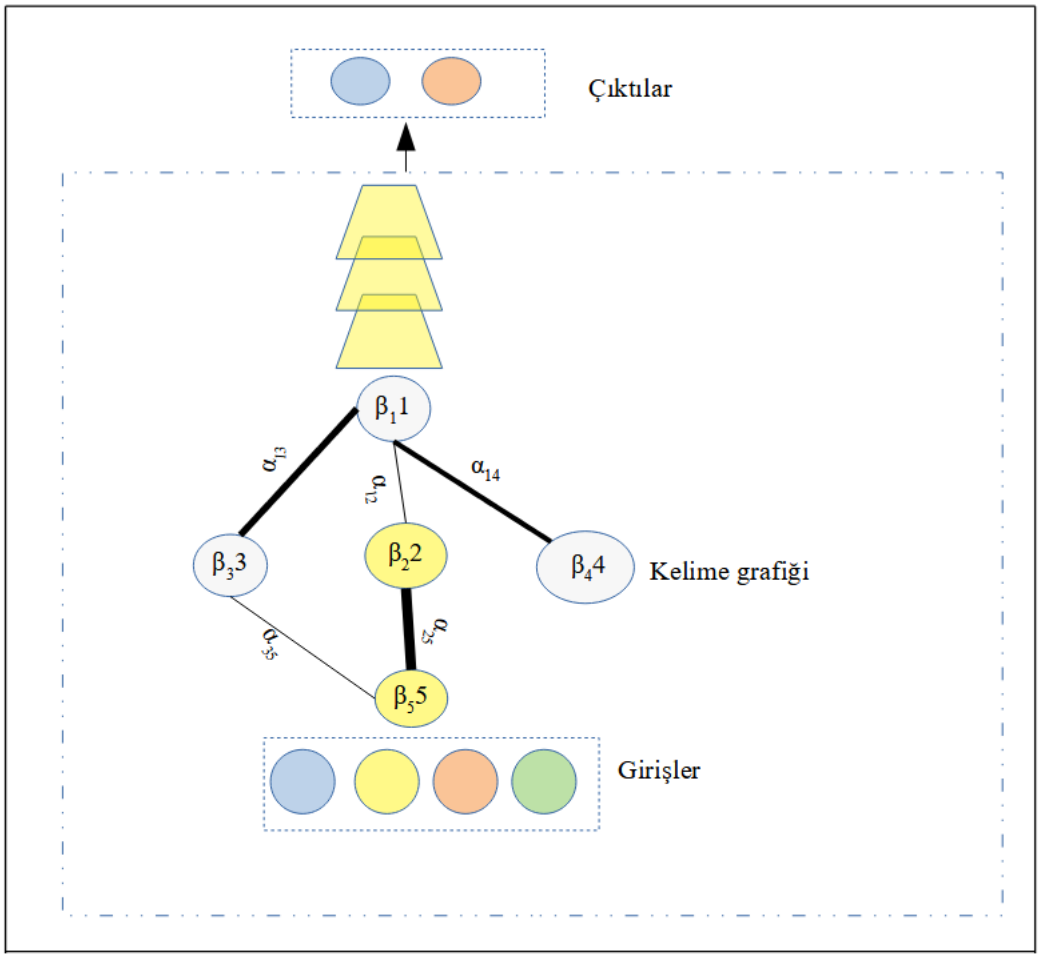


Şekil 2.3. MGKA-AKÇ modeli

2.2.2. Küresel-yerel modeli

GEA'lar, doğrudan grafik yapılarda denetimli sınıflandırma yapmak için yerelleştirilmiş konvolüsyon filtrelerinden yararlanan bir spektral kümeleme teknikleri sınıfıdır. Bu

yöntemler düğümlerin global önemini modelleme yeteneğinden yoksundur ve bu sebeple kritik bilgilerin kaçırılmasına neden olmaktadır. Küresel-yerel modeli (Glocal) (Prasad ve Kan, 2019) ölçeklendirilmiş gömme ağırlıklarını GEA'a dahil etmektedir. Metin Sıralamadan elde edilen küresel rastgele yürüme puanlarının modele dâhil edilmesi mevcut performansı artırmıştır. Şekil 2.4'te bulunan küresel-yerel için görülen β_i parametresi Metin Sıralama ile hesaplanmıştır.



Şekil 2.4. Küresel-yerel modeli

2.2.3. PhraseFormer modeli

PhraseFormer, grafik-tabanlı ve gömme tabanlı yöntemlerin bir kombinasyonudur (Nikzad-Khasmakhi ve diğerleri, 2021). Bu model DÇYKG kelime gömmesini kullanmaktadır. Anahtar kelime çıkarımı için önerilmiş hibrit bir modeldir ve sorunu 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.

2.2.4. Ortak ÖSA modeli

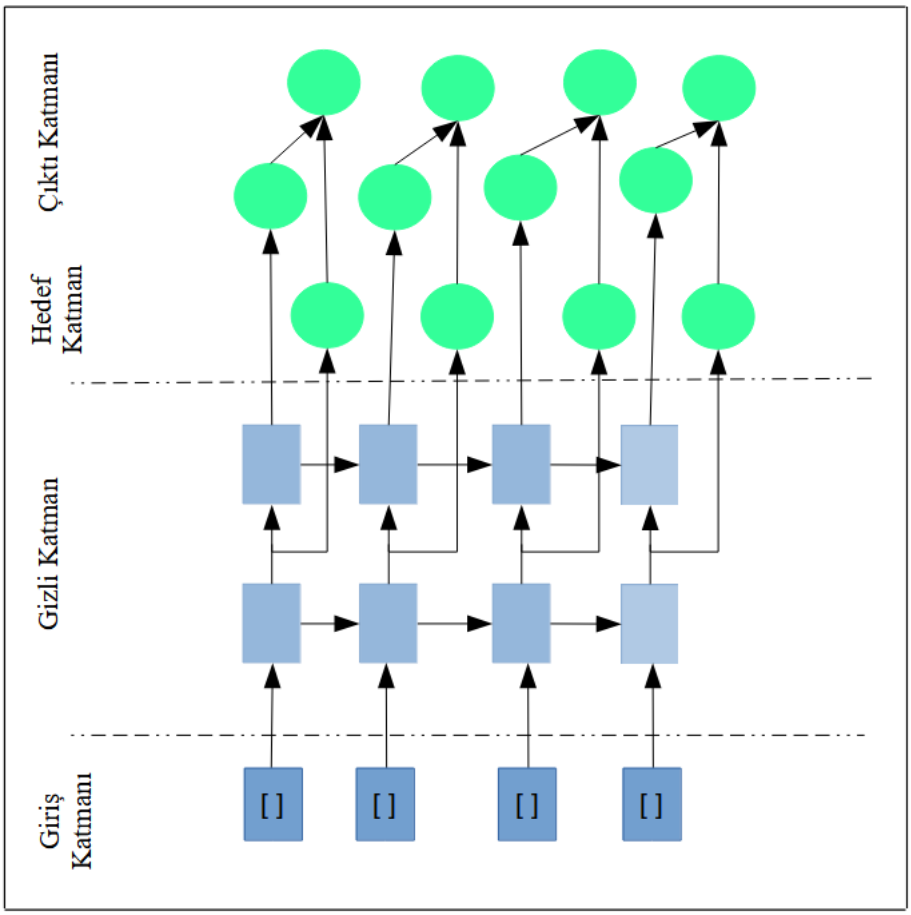
Twitter verileri üzerinde yapılan ÖSA tabanlı Ortak özyinelemeli sinir ağı (Ortak-ÖSA – Joint RNN) modelinde (Zhang, Wang, Gong ve Huang, 2021) anahtar kelime çıkarımı bir dizi etiketleme problemi olarak ele alınmıştır. Bu mimaride birden fazla gizli katmandan oluşan L katmanlı yığın ÖSA'ın özel bir versiyonu olan 2 katmanlı ÖSA kullanılmıştır. Birleşik katman olarak adlandırılan bu mimaride; anahtar kelime sıralama, anahtar kelime üretme ve anahtar ifade sıralama işlemleri birleştirilmiştir. 2 gizli katmanın birincisinde anahtar kelime bilgisi alınmaktadır. İkinci adımda birden fazla kelimeden oluşan anahtar ifadeleri tanımlamak için dizi etiketleme yöntemi kullanılır. Algoritmanın zayıf yönü bir metin girdisi için bir veya daha fazla kelimeden oluşan tek bir anahtar kelime veya anahtar ifade üretmesidir. Algoritma anahtar ifadenin tweet içerisinde olduğunu varsaymaktadır. OrtaK-ÖSA gizli katmanlarının hesaplanması:

$$h_t^1 = f_h(x_t, h_{t-1}^1),\ h_t^2 = f_h(h_t, h_{t-1}^2) \tag{2.1}$$

olarak yapılmaktadır. Çıktı katmanı:

$$\hat{y}_t^1 = f_0(h_t^1)(14),\ \hat{y}\,_t^2 = f_0(h_t^2) \tag{2.2}$$

şeklinde tanımlanmıştır. Şekil 2.5'te OrtaK-ÖSA modeli görülmektedir.



Şekil 2.5. OrtaK-ÖSA modeli

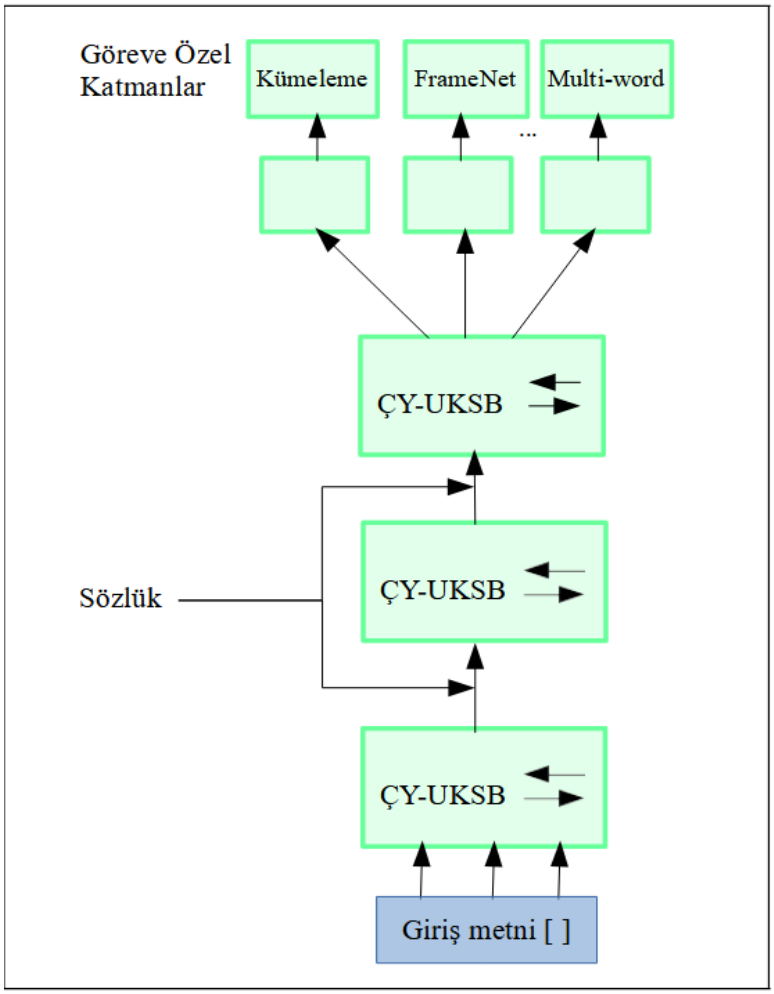
$\hat{y}_t^1$ sadece evet veya hayır değerlerini içerirken, $\hat{y}\,_t^2$ Tekil, Başlangıç, Orta, Son ve Değil etiketlerini içermektedir. Patlayan gradyan problemini çözmek için Stokastik Gradyan Azaltması (SGA- Stochastic Gradient Descent - SGA) algoritması kullanılmıştır. Kullanılan veri seti 110K tweetten oluşmaktadır.

2.2.5. OrtakÖSA+ modeli

Söz dizimsel ek bilgilerle birden fazla anahtar kelime dizisini ayıklamak için OrtakÖSA' nin giriş katmanı değiştirilerek OrtakÖSA+ (Mahfuzh, Soleman ve Purwarianti, 2019) algoritması üretilmiştir. Konuşmanın Parçası (KP - Part of Speech-PoS), Adlandırılmış-Varlık Tanıma (AVT - Named Entity Recognition (NER), cümle bağımlılık yapıları ayrı ayrı algoritmaya entegre edilmiştir. Mevcut OrtaK-ÖSA algoritmasından farklı olarak algoritma girdi metni için birden fazla anahtar ifade içerebilecek şekilde tasarlanmıştır.

2.2.6. Çok Görevli Öğrenim Çift Yönlü-Uzun Kısa Süreli Bellek (ÇGÖ-ÇY-UKSB) modeli

Sözlük tabanlı anahtar kelime çıkarımı için tasarlanan Çok Görevli Öğrenim ÇY-UKSB (ÇGÖ-ÇY-UKSB - Multi-Task Learning BiLSTM) mimarisi (Augenstein ve Sogaard, 2017) tüm gizli katmanların ortak kullanıldığı katı parametre paylaşımlı bir mimaridir. Şekil 2.6'da ÇGÖ-ÇY-UKSB mimarisi görülmektedir. Modelde, eğitim aşamasında her bir adımda rastgele bir t görevi seçilir, sonra rastgele bir eğitim örneği seçilir. Parametreler bir cümle için birlikte eğitilir. Her görev bağımsız bir sınıflandırma işleviyle ilişkilendirilir, ancak tüm görevler gizli katmanları paylaşır. Modelde beş yardımcı görev tanımlanmıştır. Bunlar (1) Chunking, (2) FrameNet, (3) Link tahmini, (4) Multi-word ve (5) Semantik süper duyu etiketlemedir. Bu mimari ile anahtar sözcük sınırı sınıflandırması yani bilimsel makalelerden anahtar kelimelerin çıkarılması (daha önceden tanımlanmış tiplerde) gerçekleştirilmiştir.

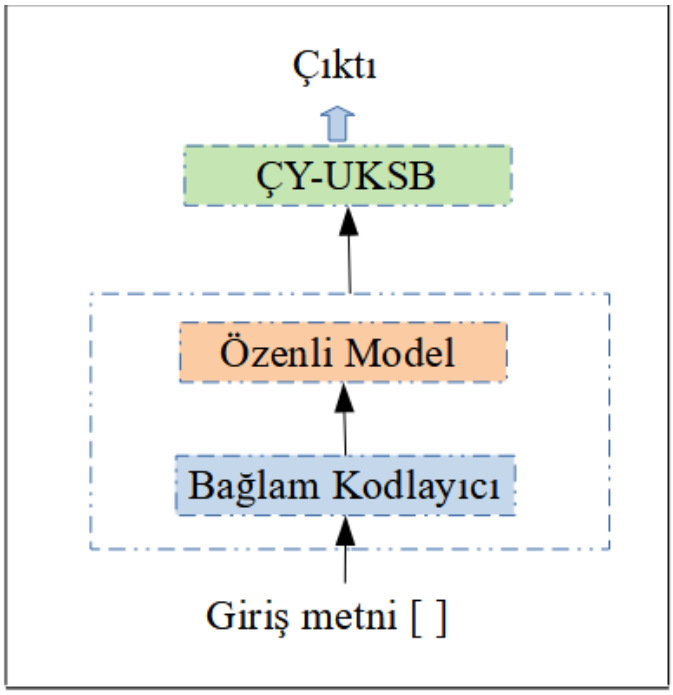


Şekil 2.6. ÇGÖ-ÇY-UKSB modeli

Diğer bir çoklu görev tabanlı mimari olan ÇYKB-Ortak Anahtar Terim Çıkarımı (Joint Keyphrase Extraction) (Sun, Xiong, Liu, Liu ve Bao, 2020) yüksek kaliteli ifadeleri tanımlamak için bir yığın ağı ve belgedeki özgünlüklerini öğrenmek için bir sıralama ağı kullanır. N-gramları tanımlamak için ÇYKB gömmesi ve ESA ağı kullanmıştır. Bu mimari, anahtar kelime kalitesi ile çıktı tahminini dengeleyerek bölme görevi ve sıralama görevi konusunda birlikte eğitilir.

2.2.7. Özenli model

Özenli Model (Attentive Model) (Passon, Comuzzo, Serra ve Tasso, 2019) bağlam ile daha bağlantılı kelimeler tanımlamak için geliştirilmiş, kodlama aşamasında bağlam bilgisini ve kelime gömmelerini birlikte kullanan bir mimaridir. Bu model Şekil 2.7'de görüldüğü gibi kodlama aşamasında iki bilgiyi birlikte kullanarak cümledeki kelimenin önemi algoritmaya dâhil edilmekte ve her kelimenin bir önem değeri bulunmaktadır. Özenli modülünden geçen girdiler, ÇY-UKSB katmanından geçirilerek; anahtar kelime, anahtar kelime başı, anahtar kelime içi veya anahtar kelime değil olarak etiketlenmektedir.



Şekil 2.7. Özenli model

2.2.8. Mesafe tabanlı anahtar kelime çıkarma modeli

Mesafe Tabanlı Anahtar Kelime Çıkarma (MT-AKÇ-Span Keyphrase Extraction-SKE) modeli (Mu ve diğerleri, 2020) anahtar kelimeler için tüm içerikten mesafe tabanlı bir özellik temsili çıkarır. Model çakışan anahtar kelimeleri bulabilmektedir ve ilk olarak KP ile aday ifadeler çıkarılmaktadır. Her aday anahtar ifade için başlangıç ve bitiş indekslerinde tutulur. Şekil 2.8'de MT-AKÇ modeli ikinci aşamada kelime vektörlerini temsil etmek için önceden eğitilmiş DÇYKG gömmesi kullanılır. Bu model mesafe tabanlı ilk çalışmadır.

$$\vec{t}_i = \overrightarrow{LSTM}(x_i^D), i \in [1, L], \tag{2.3}$$

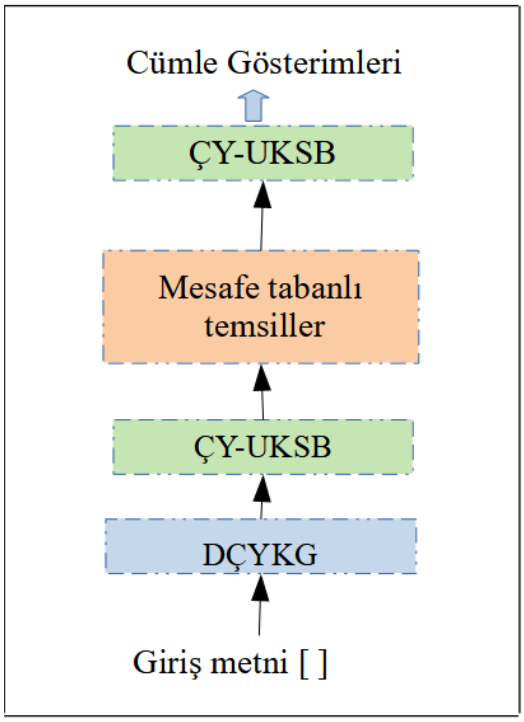
$$\overleftarrow{t}_i = \overleftarrow{LSTM}\,(x_i^D), i \in [L, 1] \tag{2.4}$$

Başlangıç ve bitiş belirteçlerinin ileri temsilleri olarak $\vec{t}_b$ ve $\vec{t}_e$, başlangıç ve bitiş belirteçlerinin geriye doğru temsilleri olarak $t^{\leftarrow}{}_b$ ve $t^{\leftarrow}{}_e$ kullanılmıştır. İfade gösterimi birleştirilmiş olarak üç tür vektörden oluşur: (i) özdeş gösterimler; (ii) eleman bazında ürün ve (iii) eleman bazında fark çarpımı. Nihai temsil $H \in R^{M \times 4d}$ için:

$$\vec{h}_i = \overrightarrow{LSTM}(s_i), i \in [1, M], \overleftarrow{t}_i = \overleftarrow{LSTM}\,(x_i^D), i \in [L, 1] \tag{2.5}$$

$$h_i = \left(\overleftarrow{h}_i, \vec{h}_i\right), i \in [1, M] \tag{2.6}$$

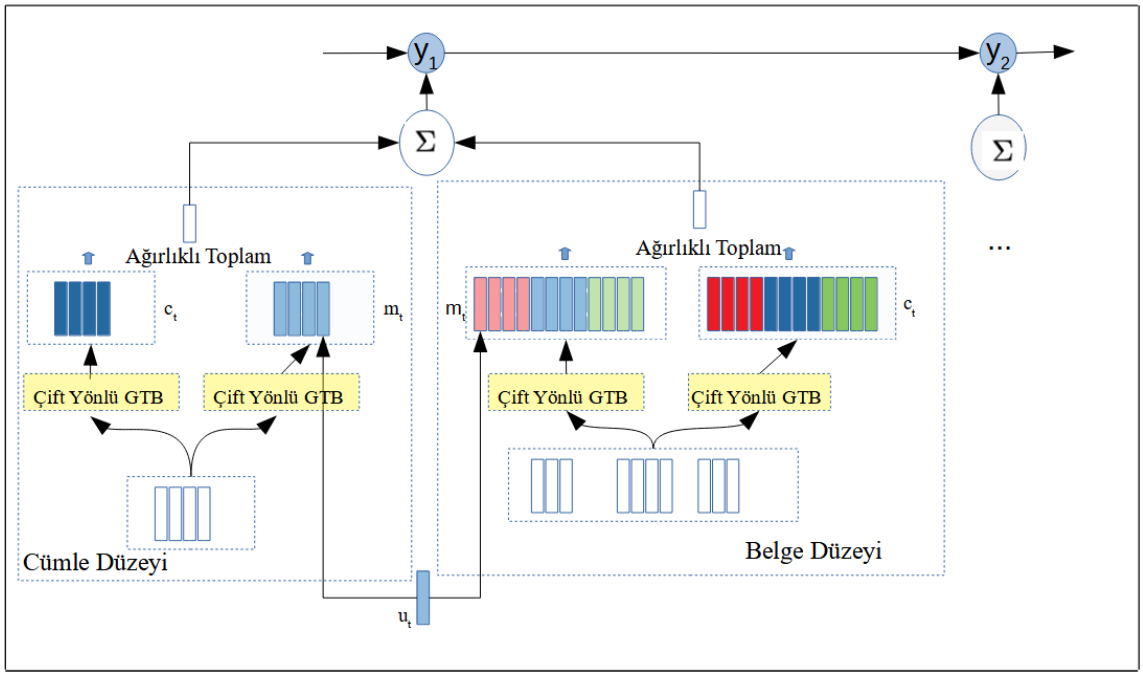
olarak hesaplanmaktadır.



Şekil 2.8. MT-AKÇ model

2.2.8. KRA'lerle çok düzeyli bellek ağı modeli

Metin verilerinde gizlenen uzun menzilli bağlamsal bilgileri yakalamak için bellek ağı kullanılarak geliştirilen KRA'lerle Çok Düzeyli Bellek ağı (ÇDBA-KRA - Multi-Level Memory network with CRFs - MLM-CRF) (Zhou, Zhang ve Zhu, 2020) modelinin çıkış katmanında KRA kullanmıştır. Metin sırasında gizlenen etkili bilgilerin tam olarak kullanılabilmesi için bellek ağının giriş belleği cümle seviyesi ve belge seviyesinde genişletilmiştir. Her seviye üç bileşen içerir: (1) metin dizisinin sözcük gömme katmanından bilgileri yakalayan giriş belleği vektörü m_t, (2) geçerli kelimenin temsili olan geçerli giriş gömmesi u_t, ve (3) m_t giriş belleğine benzer olan çıkış vektörü c_t. Şekil 2.9'da ÇDBA-KRA modeli mesafe sınırlaması olmadan giriş metni dizisinden uzun menzilli anlam ve yapı bilgilerini özetleyen bellek katmanının çıkış belleği gösterimi, dikkat ağırlıklarının belirlendiği çıkış gösterimleri üzerinde ağırlıklı bir toplamla hesaplanır.



Şekil 2.9. ÇDBA-KRA modeli

2.2.9. Öz-damıtım tabanlı ortak öğrenme yaklaşımı

Öz-Damıtım tabanlı Ortak Öğrenme Yaklaşımı (A Joint Learning Approach based on Self-Distillation – JLSD) (Lai, Bui, Kim ve Tran, 2020) DÇYKG, ÇY-UKSB+KRA modeli temel alınarak geliştirilmiştir. Bu model, öğretmen ve öğrenci olmak üzere iki modeli birlikte eğiterek çalışır. İlk olarak, öğretmen etiketli bir hedef veri seti kullanılarak eğitilir. Öğrenci modelinin başlangıç parametreleri öğretmen ile aynı şekilde ayarlanır. Kaynak veri kümesi olarak adlandırılan etiketlenmemiş veriler, öğretmen tarafından üretilen sözde etiketler ve önceden etiketlenmiş verilerle birlikte öğrenciye parçalar halinde girdi olarak verilir. Öğrenci öğretmenden daha iyi sonuçlar üretiyorsa, öğretmenin başlangıç parametreleri de öğrenci ile senkronize edilir. Bu sayede öğrencinin eğitim aşaması tamamlanmış olur. Eğitmen daha iyi sözde etiketler oluşturacağı ve bunları öğrenciye girdi olarak göndereceği için hem öğretmen hem de öğrenci daha iyi eğitilir.

2.2.10. Kopya ÖSA modeli

Kopya ÖSA modeli (CopyRNN) (Meng ve diğerleri, 2017) Şekil 2.10'da görüldüğü gibi seq2seq tabanlı bir mimaridir. Bu modelde kodlayıcı, ileri ve geri beslemeli iki gizli katman içeren çift yönlü bir Geçitli Tekrarlayan Birim (GTB - Gated Recurrent Unit -GRU) ve

çözücü ileri GTB'dan oluşur. ve dikkat mekanizması eklenerek oluşturulur. Modelin girdi verilerinin hangi kısımlarının eğitim sırasında daha önemli olduğunu anlamak için önem vektörü hesaplanır, böylece model bu bilgilere daha fazla dikkat eder. Dikkat mekanizması ve gizli katman eklenerek kodlayıcı bağlam vektörü oluşturulur ve hücre durumu şu şekilde hesaplanır:

$$h = (h_1, \dots, h_T);\ c_i = \sum_{j=1}^{T} \alpha_{ij} h_i \tag{2.7}$$

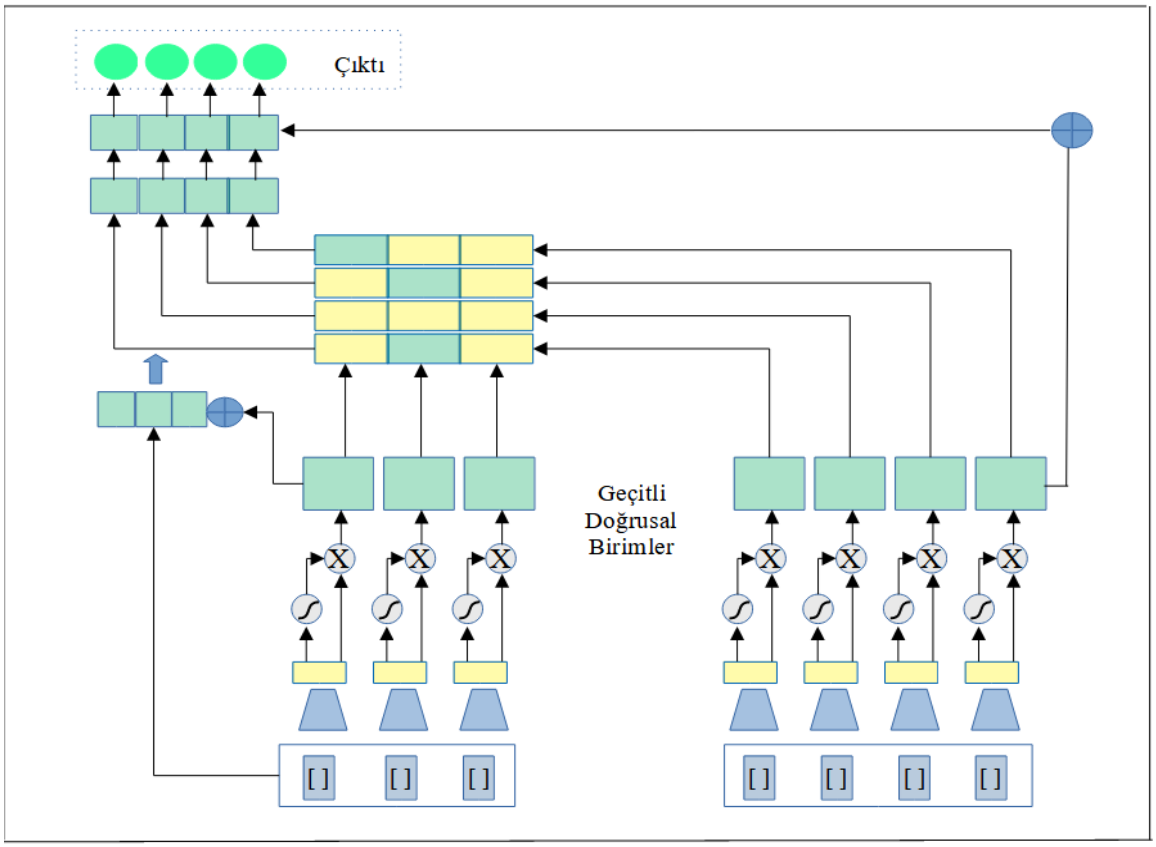
$$\alpha_{ij} = \frac{\exp(\alpha(s_{i-1}, h_j))}{\sum_{k=1}^{T} \exp(\alpha(s_{i-1}, h_k))} \tag{2.8}$$

Burada $a\left(s_{i-1}, h_j\right)$, s_{i-1} ve h_j arasındaki benzerliği ölçen bir fonksiyondur; yani, *j* konumun etrafındaki girişler ve *i* konumundaki çıkışın hangi derecede eşleştiğini ölçer.

Sözlükte bulunmayan kelimelerin de anahtar kelime olarak işaretlenebilmesi için kopyalama mekanizması kullanılmış ve çıkış olasılığı aşağıdaki gibi hesaplanmıştır:

$$p(y_t|y_1, \dots, {}_{t-1}, x) = p_g(y_t|y_1, \dots, {}_{t-1}, x) + p_c(y_t|y_1, \dots, {}_{t-1}, x) \tag{2.9}$$

burada p_g olasılığı anahtar kelime üretme olasılığını, p_c anahtar kelimenin girdi metninden kopyalanma olasılığını ifade eder. Kopyalama mekanizması modeli üretken bir model yapmaktadır.

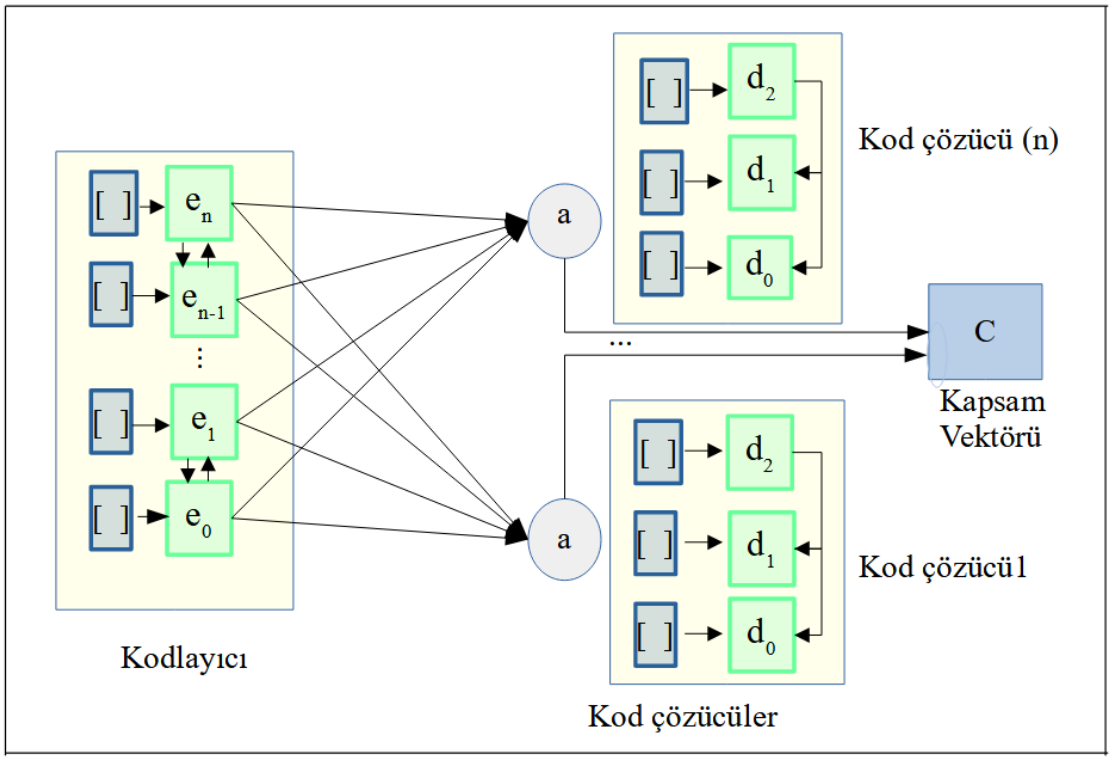


Şekil 2.10. Kopyalama ESA modeli

2.2.11. Kapsama özyinelemeli sinir ağı

Kapsama Özyinelemeli Sinir Ağı (K-ÖSA - Coverage Recurrent Neural Network -CovRNN) (Zhang ve Xiao, 2018) önem mekanizması ve kopyalama mekanizması ile birlikte anahtar kelime çıkarımı için önerilmiştir. Şekil 2.11'de K-ÖSA modeli bir kodlayıcı ile birden fazla kod çözücüden oluşur. Çıktı katmanında tekrarlama probleminin çözümü için kapsama mekanizması mimariye dâhil edilmiştir. Bu model c^t vektörü ile kaynak metin bilgisini hafızaya almıştır. c^t önceki adımdaki tüm çözücülerdeki önem dağılımının kümülatif toplamıdır ve Eşitlik 2.10 ile hesaplanır.

$$c^t = \sum_{t_s=0}^{t-1} a^{t_s} \tag{2.10}$$



Şekil 2.11. K-ÖSA modeli

2.2.12. Birleştirilmiş seq2seq ve Sınırlayıcı-Birleştirilmiş seq2seq modeli

Kaskat seq2seq (Concatenated seq2seq - CatSeq) modeli (Yuan ve diğerleri, 2018) birleştirilmiş olarak adlandırılan virgülle ayrılmış diziler olarak birden çok anahtar kelime türetir. Model Şekil 2.12'de gösterilen dikkat mekanizmasına sahip seq2seq üretici bir modelidir. Çözücü katmanında aşırı üretim sorununu çözmek için iki yeni eklenti kullanılmıştır. Bunlar; anlamsal kapsam ve ortogonal gösterim eklentileridir. KaskatSeq modeline eklentilerin dâhil edilmesiyle Ayırıcı-kaskat seq2seq (Delimiter-Concatenated seq2seq – CatSeqD) adlı yeni bir genişletilmiş model oluşturulmuştur.

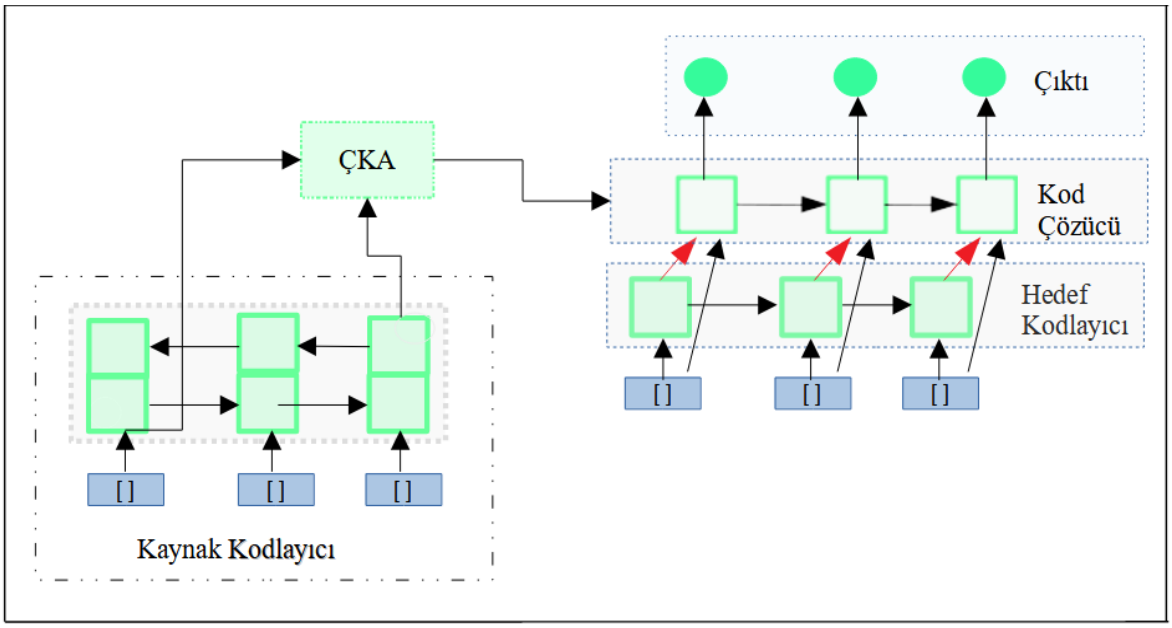
- Anlamsal Kapsam: Oluşturulan cümlelerin anlamsal temsillerine odaklanan semantik kapsama adı verilen bir mekanizma, özellikle, hedef kodlayıcı GTB_{SC} olarak adlandırılan başka bir tek yönlü tekrarlanan model sunulur ve kod çözücü tarafından oluşturulan belirteçleri gizli durumlara kodlar h_{SC}^{t} $\langle\,\rangle$ birleştirme anlamına gelir. GTB gizli durumu Kod çözücü katmanında Eşitlik 2.11'de olduğu gibi hesaplanır:

$$h^t = GTB(\langle x^t, h_{SC}^t \rangle, h^{t-1} \quad (2.11)$$

- Ortogonal Gösterim: Ayırıcıyı oluşturan kod çözücü durumlarının birbirinden farklı olmasını açıkça teşvik eden dikey bir düzenleme de önerilmiştir. L negatif log-olasılık kaybı (L_{NLL}), anlamsal kapsam kaybı (L_{SC}), ve ortogonal düzenleme kaybı (λ_{OR}) dahil; Eşitlik 2.12 ile hesaplanır:

$$L = L_{NLL} + \lambda_{OR} + \lambda_{SC} \cdot L_{SC} \quad (2.12)$$

burada λ_{OR}ve λ_{SC} hiper parametrelerdir.

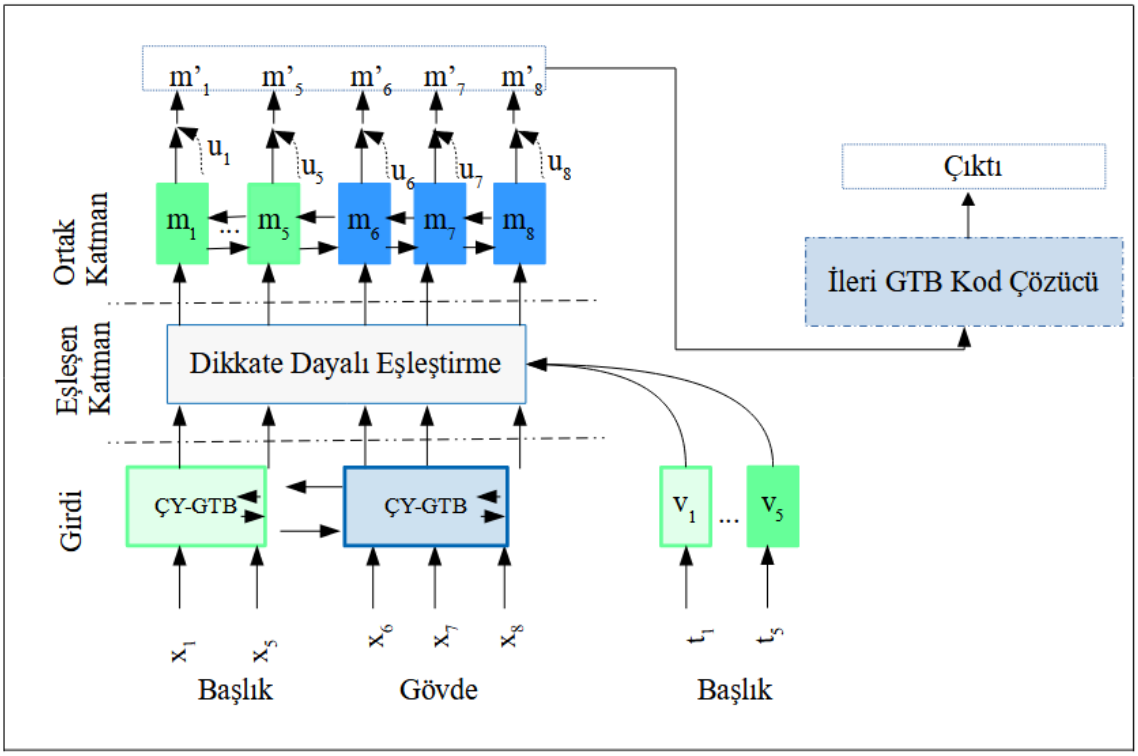


Şekil 2.12. Birleştirilmiş seq2seq modeli

2.2.13. Başlık yönlendirmeli ağ modeli

Başlık Yönlendirmeli Ağ (BYA - Title-Guided Network - TG-Net) (Chen, Gao, Zhang, King ve Lyu, 2019) Kopya-ESA modelinin bir uzantısıdır. Kopya-ÖSA ve Kopya-ESA'da başlık ve özet bilgisinin eşit ağırlıkta giriş metni olarak alınır. Kodlayıcı bu metinin tamamını kelime girdi vektörleri ile alıp işlemektedir. Bu da başlığın bir metinin içeriğini belirtmesi durumunu göz ardı etmektedir. Probleme çözüm olarak BYA'da başlık ve gövde metni için iki yönlü ayrı GTB kodlayıcı tanımlanmaktadır. Mevcut anahtar sözcükler için Başlık Bağlantılı kelime yüzdesi yaklaşık %60'a kadardır. Bir başlığın uzunluğunun genellikle ilgili kaynak metnin sadece %3-6'sı olduğu göz önünde bulundurularak, başlığın, anahtar sözcükler üretmek için gerçekten özetleyici ve değerli bilgiler içerdiği sonucuna

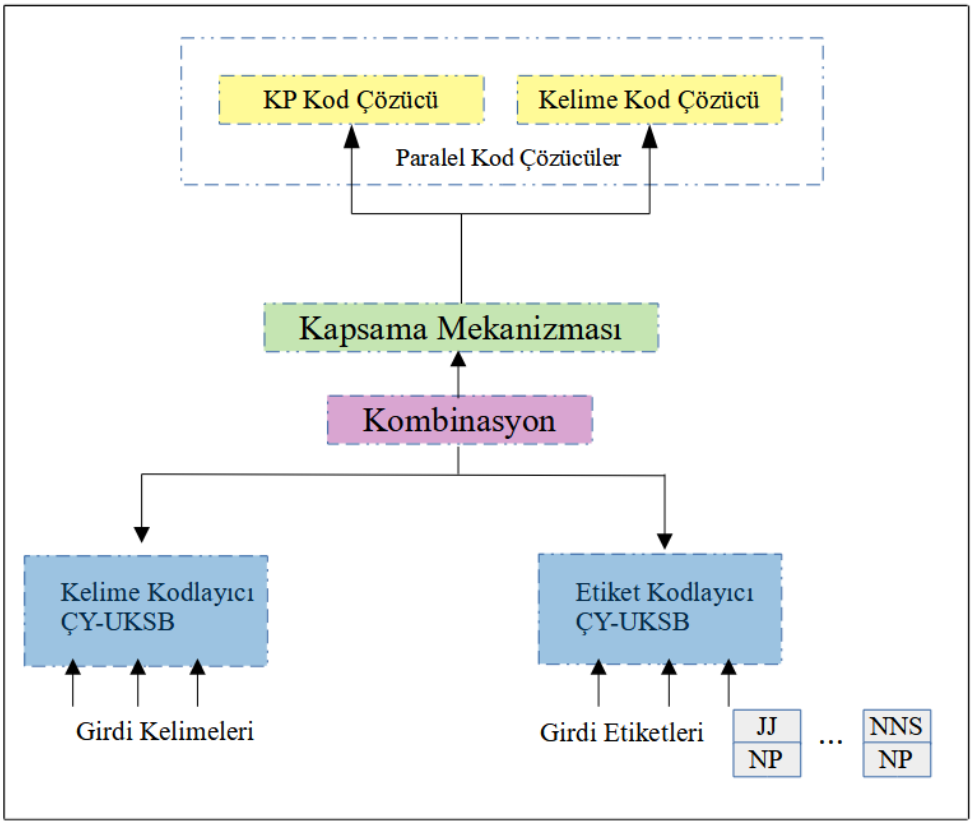
varılmaktadır. Şekil 2.13'te BYA modeli çözücü katmanında önem mekanizması ve kopyalama mekanizması eklenmiş Çift Yönlü GTB (ÇY-GTB - Bidirectional GRU-BiGRU) kullanılmıştır.



Şekil 2.13. BYA modeli

2.2.14. Paralel seq2seq kapsama dikkat modeli

Paralel seq2seq Kapsama Dikkat Modeli (Paraseq2seq-KDM - Parallel seq2seq Coverage Attention Model - ParaNetT+CoAtt) modelinde anahtar sözcüklerin çoğu, genellikle isimler ve sıfatlardan oluşmaktadır. Derin söz dizimsel detayları yakalayabilmek için paralel bir seq2seq modeli ParaNet (Zhao ve Zhang, 2019) önerilmiştir. Yöntem KP etiketi ve ifade etiketini söz dizimsel özellikleri yakalayabilmek için entegre etmiştir. Şekil 2.14'te ParaNet+CoAtt modeliyle önerilen yöntem görülmektedir. Kodlayıcı, hem kelime kodlayıcı hem KP ve ifade etiket birleşimini yapan tag kodlayıcı olmak üzere iki kodlayıcıdan oluşmaktadır. Gizli katmanlar paralel kodlayıcı aracılığı ile hesaplandıktan sonra kapsama önem vektörü hesaplanarak çözücüye gönderilir. Çözücü kısmı bir kopyalama mekanizmasına sahip kelime çözücü ve PoS çözücüden oluşan çok görevli öğrenme modülünden oluşmaktadır.



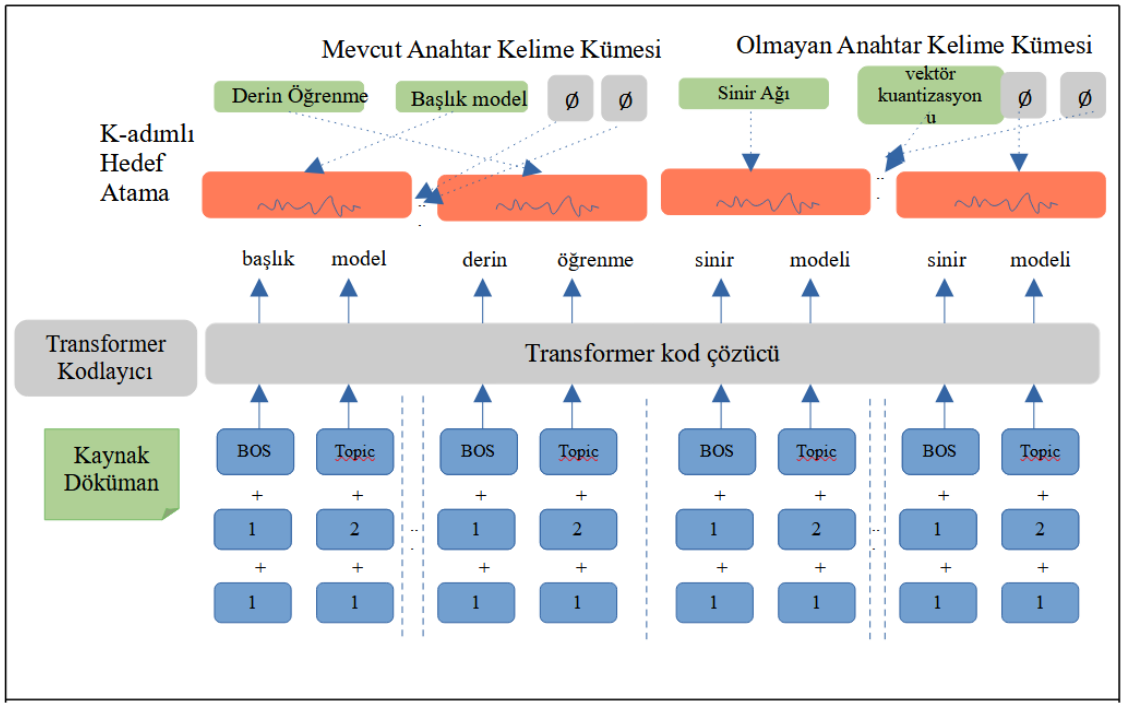
Şekil 2.14. Paralel seq2seq kapsama dikkat modeli

2.2.15. ONE2SET

ONE2SET modelinde (Ye, Gui, Luo, Xu ve Zhang, 2021) Şekil 2.15'te görüldüğü gibi, anahtar kelime üretimi için geliştirilmiş bir k adımlı hedef atama mekanizması önerilmiştir. Modelin eğitim aşamasında, hedef klasik seq2seq mimarisinde olduğu gibi gerçek anahtar kelimeler yerine bir küme olarak tanımlanmıştır. Transformatör tabanlı modelde, çözücüde kontrollü üretim için girdi olarak sabit öğrenilmiş kontrol kodları verilmiştir. Her t anında çözücüye gönderilen kontrol kodu n girişi şu şekilde tanımlanır:

$$d_t^n = e^{\omega}_{y_{t-1}^n} + e_t^p + c^n \tag{2.14}$$

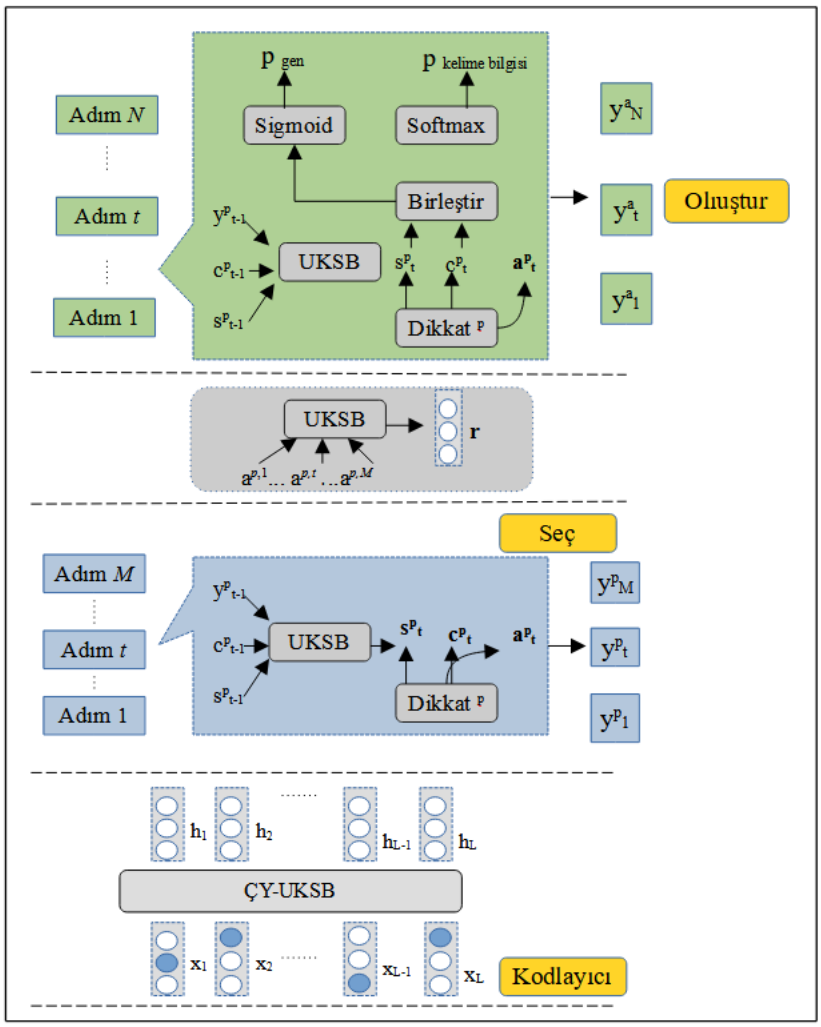
Burada $e^{\omega}_{y_{t-1}^n}$ y_{t-1}^n kelime için kelime gömmesini, e_t^p t-inci gömme için sinussel konumu belirtir. c^n, n-inci öğrenilmiş control kod gömmesinin değeridir. Çözücü tahminsel dağılım p_t^n'yi üretir,ve bu bir sonraki kelime çıktısı y_t^n için kullanılır.



Şekil 2.15. ONE2SET modeli

2.2.16. Seç, yönlendir ve oluştur modeli

Seç, Yönlendir ve Oluştur (*SYO - Select, Guide, and Generate – SGG)* (Zhao ve diğerleri, 2021) modeli temel olarak görünür anahtar kelimeleri çıkarma ve görünür anahtar kelimeler tarafından yönlendirilen görünmeyen anahtar kelimeler üretme ilkelerine dayanmaktadır. Seq2seq tabanlı modele çoklu görev öğrenme modülü eklenmiştir. Hiyerarşik bir yapıya sahip olan model, ilk katmanda seçici bir ağ, orta katmanda bir yönlendirici (kılavuz) ve üst katmanda üretici bir ağa sahiptir. Seçici, dikkat dağılımlarından görünür anahtar kelimeleri işaretler. Öte yandan, üretici, bir işaretleme mekanizması ile görünmeyen anahtar kelimeler üretir. Üreticinin daha önce seçici tarafından işaretlenen anahtar kelimeleri üretmemesi için, kılavuz işaretlenen anahtar kelimeleri ezberler ve üretim aşamasında en üst katmanı yönlendirir. Şekil 2.16'te SYO modeli kod çözme aşamasında, ÇY-UKSB tarafından üretilen gizli katmanlar, seçici ağ için giriş verilerini oluşturur.



Şekil 2.16. SYO modeli

2.2.17. Kaskat Seq-2RF1 / Kaskat SeqD-2RF1 modeli

Kaskat Seq2RF$_1$ (CatSeq-2RF$_1$)(Chan, Chen, Wang ve King, 2019) mimarisi, eğitim sırasında yeterince anahtar kelime üretmesi için kendini eleştiren bir politika kullanmaktadır. Önerilen bu pekiştirmeli öğrenme yaklaşımı kodlayıcı/çözücü tabanlı herhangi bir mimariye entegre edilebilmektedir. Yuan ve diğerleri (2018) tarafından geliştirilen Kaskat Seq ve Kaskat SeqD modellerine uygulanarak Kaskat Seq-2RF$_1$ ve Kaskat SeqD-2RF$_1$ modeli üretilmiştir. RL uygulandıktan sonra üretilen anahtar kelime sayısı ortalaması 4.3'ten 5.3'e yaklaşmıştır.

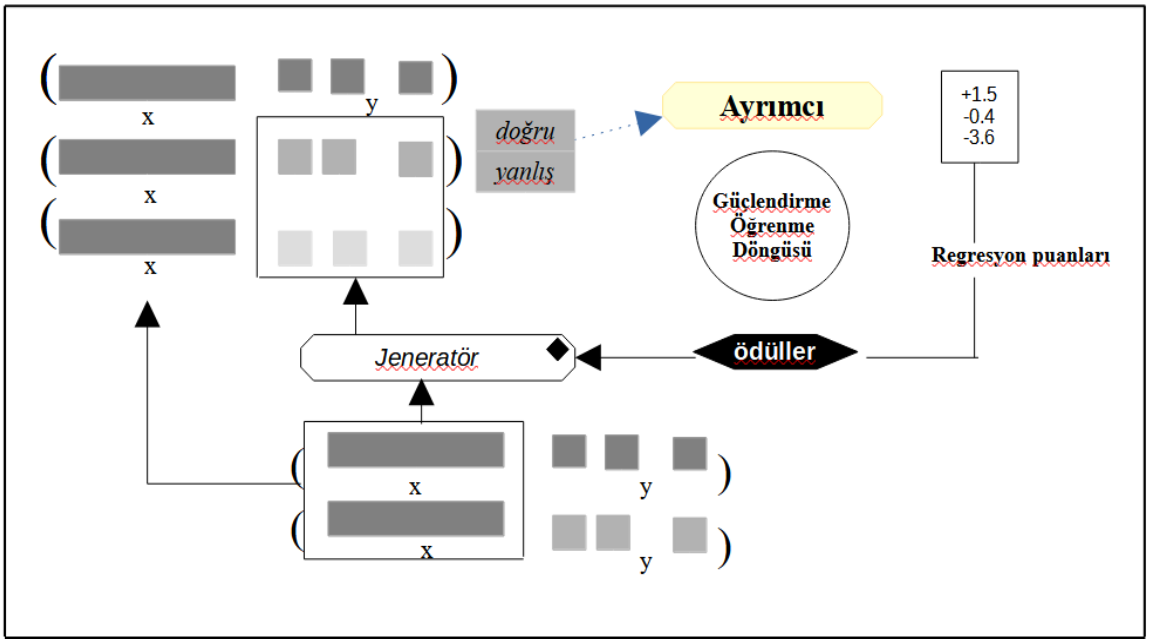
Adaptif ödül fonksiyonu RF$_1$, Eşitlik 2.15 ile hesaplanmaktadır:

$$RF_1 = \begin{cases} recall & if\ N<G, \\ F_1 & otherwise. \end{cases} \tag{2.15}$$

Eşitlik 2.15' te N tahmin edilen anahtar kelime sayısı, G gerçekte var olan anahtar kelime sayısıdır. Gerçekte üretmesi gerekenden az anahtar kelime ürettikçe tahminlerin Duyarlılık değeri ödül olarak sisteme verilir. Anahtar kelime olmayan kelimeler Duyarlılık değerini düşürmediğinden sistem yeteri kadar anahtar kelime üretilinceye kadar çalışır. Yeterli anahtar kelime üretilmesi durumda F_1-skor ödül olarak seçilir.

2.2.18. Çekişmeli üretici ağ ile etkili anahtar kelime çıkarımı

Ayrıştırıcı modülün üretim ağını beslediği bir ayrıştırıcı ve bir üretici ağından oluşan mimari literatürde Çekişmeli Üretici Ağ (ÇÜA - Generative Adversarial Networks - GAN) olarak adlandırılmıştır. CatSeqD mimarisi Lancioni, Mohamed, Portelli, Serra ve Tasso (2021) tarafından önerilen çekişmeli üretim ağındaki üretim katmanı olarak modele dahil edilmiştir. Ayrıştırıcı ile DÇYKG kullanılarak oluşturulan ifadeler, gerçek veya sahte olarak sınıflandırılmış, sonuçlar pekiştirmeli öğrenme kullanılarak üretici ağa beslenmiştir. Şekil 2.17'de bu mimarinin genel yapısını ve ayrıştırıcı modülün detaylarını göstermektedir.



Şekil 2.17. ÇÜA ile etkili anahtar kelime çıkarımı

2.2.19. Denetimli modellerin performans sonuçları

Anahtar Kelime Çıkarımı Modellerinin Mimari Açıdan Karşılaştırması Çizelge 2.1'de görüldüğü gibi, çıkarım yöntemleri problemi bir dizi etiketleme problemi veya sınıflandırma problemi olarak ele almaktadır. Çizelgede referanslar, mimari adı, yıl ve dil meta veri sütunları vardır. Bu sütunlar mimari hakkında temel bilgiler verir. Tüm algoritmalar İngilizce dili için eğitilmiş ve test edilmiştir. Veri kümesi sütunu, mimariyi test etmek için kullanılan yaygın olarak kullanılan veri kümesi adlarını içerir. Tablo aynı zamanda mimaride gömme sütununda hangi kelime gömmesinin kullanıldığını da içermektedir.

Tablonun Algoritma sütununda, geliştirilen yöntemin sınıflandırmaya mı, dizi etiketlemeye mi yoksa seq2seq'e mi dayalı olduğu açıklanmıştır. Mimari olarak dikkat mekanizması, kapsama mekanizması, kopyalama mekanizmasına benzer hangi mekanizmaların kullanıldığı açıklanmıştır. Alt mimari sütununda, yeni üretilen mimari için hangi alt mimarilerin kullanıldığı açıklanmıştır.

Çizelge 2.1'de, çıkarımsal ve üretici tabanlı 21 önerilen mimari incelenmiştir. 2016 yılından itibaren ilk zamanlarda geliştirilen modellerde ÖSA'nIn sıklıkla kullanıldığı görülmektedir. ÖSA mimarisinden kaynaklanan sorunları çözmek için geliştirilen UKSB ve GTB, 2017 yılından itibaren ÖSA modülü yerine kullanılmaktadır. Kelime yerleştirme olarak, Word2Vec ve GloVe başlangıçta ağırlıklı olarak mimariye dâhil edilirken, DÇYKG ve türevlerinin son yıllarda daha sıklıkla kelime gömmesi olarak kullanıldığı açıkça anlaşılmaktadır.

Çizelge 2.1. Denetimli anahtar kelime çıkarımı modellerinin mimari açıdan karşılaştırılması

Referans	Mimari Adı	Yıl	Veri kümesi	Algoritma	Alt mimari	Çıkarım/ Üretim	Gömme	Dil
Zhang Q. ve diğ.	OrtakÖSA	2016	Tweetler (110K)	Dizi Etiketleme	sÖSA tabanlı 2 katman ÖSA	Çıkarım	Word2Vec	İngilizce
Augenstein ve Soggard	Çok görevli öğrenim ÇY - UKSB	2017	SemEval-2017, diğer	Dizi Etiketleme	ÇY-UKSB + Çoklu Görev Öğrenimi	Çıkarım	SENNA	İngilizce
Meng ve diğ.	Kopya-ESA	2017	Inspec, Krapivin, NUS, KP20k, SemEval-2010, diğer	seq2seq + Kopyalama Mek. + Önem Mek.	ÇY-GTB Kodlayıcı	Üretim	-	İngilizce
Chen ve diğ.	Korelasyonel Tekrarlayan Sinir Ağı	2018	Krapivin, NUS, SemEval - 2010	seq2seq + Korelasyon Mek. + Önizleme Mek.	Kopya-ESA	Üretim	-	İngilizce
Yuan ve diğ.	KaskatSeq KaskatSeq D	2018	Inspec, Krapivin, NUS, KP20k, SemEval-2010	seq2seq + bağlamsal kopyalama + ortogonal düzenleme	ÇY-GTB Kodlayıcı + İleri G TB Çözücü ve Hedef Kodlayıcı	Üretim	-	İngilizce
Chan ve diğ.	KaskatSeq-2RF1 KaskatSeqD-2RF1	2019	Inspec, Krapivin, NUS, KP20k, SemEval-2010	seq2seq + Önem Mek. + Kopyalama Mek.	KaskatSeq + Takviyeli Öğrenme	Üretim	-	İngilizce
Alzaidy ve diğ.	ÇY-UKSB+ KRA	2019	KP20k, WWW, KDD	Dizi Etiketleme	ÇY-UKSB+KRA	Çıkarım	Glove	İngilizce
Wang B. ve diğ.	S-DWKE M-GCKE	2019	Inspec, Krapivin, SemEval-2010, WWW, diğer	Sınıflandırma GEA + ÇKA	Birlikte Oluşum Ağı + Rastgele yürüyüş + GaussNaive Bayes GEA +ÇKA	Çıkarım	-	İngilizce
Mahfuzh ve diğ.	OrtakÖSA+	2019	Twitter	Dizi Etiketleme	OrtaK-ÖSA + WE, KP NE, DS Büyütme	Çıkarım	Word2Vec	Endonezce
Chen ve diğ.	BYA	2019	Inspec, Krapivin, NUS, KP20k, SemEval-2010	seq2seq + Başlık Yönlendirmeli	Kopya-ESA	Üretim	-	İngilizce
Prasad ve Kan	Küresel-yerel modeli	2019	Schutz, Krapivin, SemEval - 2010	Dizi Etiketleme	Küresel-yerel Evrişim	Çıkarım	Glove	İngilizce
Wang ve diğ.	konuya duyarlı (topic-aware)	2019	Twitter, Weibo, StackExchange	seq2seq + Önem Mek. + Kopyalama Mek.	ÇY-GTB + GTB +Başlık Dikkatli Kodlayıcı	Üretim	-	İngilizce/ Çince
Zhao ve Zhang	Paralel seq2seq Kapsama Dikkat Modeli	2019	Inspec, Krapivin, NUS, KP20k, SemEval-2010	seq2seq + Kapsama Mek.	Multitask + Paralel Kodlayıcı / Çözücü	Üretim	-	İngilizce
Passon ve diğ.	Dikkatli Model	2019	Inspec	Dizi Etiketleme + Önem Mek.	ÇY-UKSB + Bağlama Kodlayıcı	Çıkarım	Glove	İngilizce
Mu ve diğ.	MT-AKÇ	2020	Inspec, Krapivin, NUS, KP20k, SemEval-2010	Dizi Etiketleme	ÇY-UKSB	Çıkarım	DÇYKG	İngilizce

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

Referans	Mimari Adı	Yıl	Veri kümesi	Algoritma	Alt mimari	Çıkarım/ Üretim	Gömme	Dil
Zhou ve diğ.	ÇDBA-KRA	2020	WWW, KDD	Dizi Etiketleme + Önem Mek.	Bellek Ağı + KRA	Çıkarım	Glove	İngilizce
Lai ve diğ.	JLSD	2020	Inspec	Dizi Etiketleme	ÇY-UKSB+KRA	Çıkarım	SciBERT	İngilizce
Zhao ve diğ.	SGG	2021	Inspec, Krapivin, NUS, SemEval-2010	Seq2Seq + Önem Mek.	ÇY-UKSB + Çoklu Görev Öğrenimi	Üretim	-	İngilizce
Lancioni ve diğ.	ÇÜA ile Etkili Anahtar Kelime Üretimi	2021	Inspec, Krapivin, NUS, SemEval-2010, KP20K	Seq2Seq +Önem Mek. + Kopyalama Mek.	catSeqD + RL + GANs	Üretim	DÇYKG	İngilizce
Ye ve diğ.	ONE2SET	2021	Inspec, Krapivin, NUS, SemEval-2010, KP20K	Seq2Seq + SetTrans	Transformer	Üretim	-	İngilizce
Nikzad-Khasmakhi ve diğ.	Phraseformer	2021	Inspec, SemEval-2010, SemEval-2017	Dizi Etiketleme + Kelime Grafiği	İleri Besleme Ağı	Çıkarım	DÇYKG	İngilizce

2.3. Denetimsiz Anahtar Kelime Çıkarımı Modelleri

Anahtar kelime çıkarımı problemi için önerilmiş denetimsiz öğrenme modellerinde kümeleme ve alt gruplara ayırma teknikleri kullanılmaktadır. Kümeleme girdi verileri homojen bir şekilde kümeleme ve alt gruplara ayırma işlemini veri analizi aracı olarak yürütür. Anahtar kelime çıkarımında denetimsiz yaklaşım, açıklamalı belgelere duyulan ihtiyacı ortadan kaldırır. Potansiyel anahtar kelimeleri seçmek için dil modelleme ve istatistiksel analizi kullanır. Bir anahtar kelime genellikle belgedeki kelime sıklığı, ilk oluşumunun konumu, kök ve konuşma parçası etiketi gibi özellikler temel alınarak seçilir (Matsuo ve Ishizuka (2004); Mihalcea ve Tarau (2004); Liu, Huang, Zheng ve Sun (2010). Denetimsiz yöntemler genel bir alandan bağımsızdır ve açıklamalı bir yapı oluşturmayı gerektirmiyor. Anahtar kelime çıkarımı bilgi erişim sistemleri, dijital kütüphane araştırması, web içeriği yönetimi, belge kümeleme ve metin özetleme gibi çeşitli doğal dil işleme uygulamalarında kullanılmıştır.

Anahtar kelime çıkarımı için mevcut denetimsiz yaklaşımlar dört gruba ayrılabilir:

- İstatistiksel tabanlı
- Grafik tabanlı sıralama
- Gömme Tabanlı
- Dil modelli tabanlı

Literatürde anahtar kelime çıkarımı için derin öğrenme çalışmaları, istatistiksel ve graf tabanlı çalışmalarına göre daha yakın döneme aittir. Doğal dil işleme alanlarında makine öğrenmesi ve derin öğrenme yaklaşımları sayesinde 2000'li yılların başında yapay zekaya dayalı yöntemler kullanılmaya başlanmıştır. Ancak, denetimsiz yaklaşımlara dayanan anahtar kelime çıkarımı algoritmaları da yıllar içinde geliştirilmiştir. Örneğin, Page, Brin, Motwani ve Winograd (1999) çalışmalarında ilk kez Metin Sıralama ile kelime grafiğindeki aday anahtar sözcük puanlarını hesaplamak için Sayfa Sıralaması algoritmasını kullanır ve anahtar sözcük elde etmek için puanları sıralar. Tomokiyo ve Hurst (2003) anahtar kelime çıkarımı için bir dil modeli yaklaşımı kullanmayı önermişler, Mihalcea ve Tarau (2004) anahtar kelimeleri bulmak için grafik tabanlı bir sıralama algoritması sunmuşlar.

Denetimsiz derin öğrenme yaklaşımı anahtar kelime seçmek için dil modelleme ve istatistiksel analizi kullanır. Wan and Xiao (2008) çalışmalarında denetimsiz anahtar kelime

çıkarımı yöntemlerini corpus'a bağımlı ve corpus'tan bağımsız olarak iki gruba ayırmışlardır. Korpus'tan bağımsız yöntemler, anahtar kelimelerin çıkarılacağı tek belgeden başka girdi gerektirmemektedir. Belgeden anahtar kelime çıkarımı tek kelimelik anahtar kelime çıkarımı, tek dilli çok belgeli anahtar kelime çıkarımı ve çok dilli çok belgeli anahtar kelime çıkarımı olarak gruplanır. Tek belgeden anahtar kelime çıkarımı yalnızca kelimelerin sıklığı kelime ve kelime konusu bilgileri arasındaki ilişki yoluyla yapılır. Bu tür mevcut yöntemlerin çoğu KeyCluster, Konu Sıralaması (TopicRank), Hızlı Otomatik Anahtar Kelime Çıkarımı (HO-AKÇ - Rapid Automatic Keyword Extraction-RAKE) ve değer istisnalar dışında grafik tabanlı yöntemlerdir. Metin Sıralama algoritmasını Matsuo ve Ishizuka (2004); Mihalcea and Tarau (2004) tarafından önerilen makalede ilk grafik tabanlı anahtar sözcük çıkarma yöntemi önerilmiştir. El-Beltagy ve Rafea (2009) tarafından denetimsiz bir KP-Miner modelinin İngilizce ve Arapça belgelerden otomatik anahtar sözcük çıkarımı için etkili olduğunu göstermişlerdir.

Yang, Zhao, Zhang ve Zhao, L (2008) Çince anahtar kelimeleri çıkarmak için ortalama terim sıklığı ve orantılı belge sıklığı kullanan Sayfa Sıralaması tabanlı bir algoritma geliştirmişlerdir. Liang, Huang, Li ve Lu (2009) çalışmalarında Çince haber makalelerinden anahtar kelimeleri çıkarmak için grafik tabanlı bir Metin Sıralaması öğrenme algoritması önermişlerdir. Başka bir çalışmada Konu Sıralaması algoritması (Bougouin, Boudin ve Daille, 2013) grafik ve kümelenme tabanlı yaklaşımları birleştirerek aday ifadeler önce kümelenir, ardından her düğümün bir kümeyi temsil ettiği bir grafik oluşturulur. Awajan (2014) makalesinde istatistiksel analiz ve dil bilgisini birleştiren Arapça belgelerden anahtar kelimeler çıkarmak için denetimsiz iki aşamalı bir yaklaşım sunmuştur. Anahtar kelime olarak değerlendirilebilecek tüm n-gramlarını tespit ettikten sonra morfolojik analizör kullanılarak analiz etmişlerdir. Bekbulatov and Kartbayev (2014) çalışmalarında Kazak morfolojik segmentasyonu ile ilgili gazete corpus üzerinde yapılmış araştırmaları sunmuş ve denetimsiz kural tabanlı dil işleme modelleriyle karşılaştırmıştır. Myrzakhmetov ve Kozhirbayev (2018) çalışmalarında geleneksel n-gram ve UKSB tabanlı sinir ağlarını kullanarak gazete veri seti üzerinde dil modelleme deneyleri yapılmış ve nöral tabanlı modellerin n-gram tabanlı modellerden daha iyi performans gösterdiğini belirtmişlerdir. Papagiannopoulou ve Tsoumakas (2020) çalışmalarında anahtar kelime çıkarımı için son derin öğrenme yöntemleri dahil olmak üzere çok sayıda denetimli ve denetimsiz yöntemlerinin ana özelliklerine göre kategorize ederek güçlü ve zayıf yönlerini incelemişlerdir. Başka bir çalışmada (El-Shishtawy ve Al-Sammak, 2012) aday terimleri

temsil etmek için kök formu yerine Arapça kelimelerin soyut formu kullanılmıştır. Makale, bir belgenin başlıklarını ve alt başlıklarını yakalamak için dil bilgisine dayanan anahtar sözcüklerin yeni özelliklerini tanıtmaktadır. Özellikler kullanılarak LDA kullanılarak anahtar kelime çıkarımı yapmışlardır.

Anahtar kelime çıkarımı için çeşitli makine öğrenimi tabanlı tekniğin işlevsel ayrıntıları incelendiğinde, grafiksel ve istatistiksel tabanlı yöntemlerin aday anahtar kelimeler belirlendikten sonra sıralama mekanizmasından geçirmektedir. İki grup arasındaki temel fark grafik tabanlı modeller sıralamadan önce kelime grafikleri oluşturmakta, istatistiksel tabanlı modeller geçmiş koleksiyonları kullanmaktadır.

2.3.1. İstatistiksel tabanlı yaklaşımlar

Literatüre baktığımızda anahtar kelime çıkarımı konusunda en yaygın kullanılan istatistiksel yaklaşım algoritmalarının TF-TDF (Salton ve Buckley, 1988), KP-Miner (El-Beltagy ve Rafea, 2008), Bir Diğer Anahtar Kelime Çıkarımı (BD-AKÇ - Yet Another Keyword Extraction - YAKE!) (Campos ve diğerleri, 2020), HO-AKÇ (Rose, Engel, Cramer ve Cowley, 2010), Hiperlink kaynaklı konu arama (HKKA - Hyperlink-induced topic search - HITS) (Kleinberg, 1999) olduğu görülmektedir. Literatürde yapılan anahtar kelime çıkarımı uygulamalarında çok eski çalışmalarda Terim Frekansı (TF), Ters Terim Frekansı gibi bilgiler kullanılarak sadece tek bir dokümana ait istatistiksel ölçümle anahtar kelime çıkarımı metodu kullanılmıştır. 1972 yılında terim frekansının ölçülü o korpusta bulunan diğer dokümanlarla ilişkilendirilerek TF-TDF çıkarılarak kullanılmaya başlanmıştır. Ramos (2003) TF-TDF yöntemini kullanarak metindeki her kelime oluşumu sıklığını ana anahtar kelime çıkarma ölçütü olarak değerlendirmişlerdir. Li, Fan ve Zhang (2007) Çin haber belgelerindeki aday kelime seçimi ve metin tabanlı seçimler için Çince metinler için TF-TDF, Tek-gram, Bi-gram ve Tri-gram yöntemlerini uygulayarak anahtar kelime çıkarımını gerçekleştirmiştir. Hong ve Zhen (2012) çalışmalarında Çin dili özelliklerini temel TF yöntemiyle birleştiren makalede genişletilmiş terim frekansı tabanlı bir yöntem Genişletilmiş TF (Extended TF) gerçekleştirmişlerdir. Yeom, Ko ve Seo (2019) araştırmalarında temel olarak denetimsiz yaklaşımlar olarak istatistiksel ve C-değeri (C-value) yöntemi, grafik tabanlı modeller etkili bir kombinasyon yöntemini kullanarak anahtar kelime çıkarımı problemi çözülmüştür. Özellikle, C-değeri iç içe geçmiş terimlerin

ayıklanmasını iyileştirmeyi amaçlayan, etki alanından bağımsız, çok kelimeli otomatik terim tanıma yöntemidir.

TF-TDF

Salton ve Buckley (1988) çalışmalarında belgedeki kelimeleri Sıklık - Ters Belge Frekansı (Frekans) Terim Sıklığı (TFF) ve her bir sözcük için belgede görünen ters belge frekansı Ters Belge Frekansı (TDF) sayarak sınıflandırma işlemi gerçekleştirmişlerdir. Terim Frekansı (TF) belgenin içerisinde terimin ne kadar sıklıkla olduğunu ifade eder. Bir belgedeki en sık kullanılan kelimeler anahtar kelime olamaz, o yüzden terimlerin ağırlığını azaltmak için ters belge frekansı kullanırlar. En temel ve en sık kullanılan denetimsiz anahtar kelime çıkarma yöntemi istatistik yönteme dayanan TF-TDF belgedeki bir terimin önemini gösteren ve metin içinde kelimenin önemli olup olmadığını belirlemek için kullanılan bir ağırlık faktörüdür:

$$TF(t,d) = \frac{f_t}{n} \tag{2.16}$$

TF(t,d) – terim sıkılığı

f (t), terim sıklığı

n, verilen belgedeki terimlerin toplam sayısı,

d. doküman.

Bir TDF derecelendirmesi, yaygın olarak kullanılan kelimelerin ağırlığını azaltır:

$$TDF(q_i) = log\frac{N-n(q_i)+0.5}{n(q_i)+0.5} \tag{2.17}$$

N- toplam doküman sayısıdır,

$n(q_i) -$ verilen q_i terimini içeren belge sayısıdır.

Sonuç olarak, bir belge çerçevesinde belirli bir terimin ağırlığına ilişkin bir tahmin elde edilir:

$$TF - TDF\ (t, d, n, N) = TF(t, d) \times TDF(n, N) \quad (2.18)$$

TF-TDF ağırlığı, bir kelimenin bir koleksiyondaki bir belgeye verdiği önemi değerlendirir ve daha yüksek TF-IDF puanları olan kelime, belgede önemlidir ve belgeyi özetleyebilir.

KP-Miner AKÇ mimarisi

TF-TDF ölçüsüne dayanan denetimsiz bir anahtar kelime çıkarımı sistemi olan KP-Miner, bir anahtar kelimenin önemini belirlemek için belgedeki kelime uzunluğu ve konumunu öneren El-Beltagy ve Rafea (2009) tarafından önerilmiştir.

Aday cümlelerin oldukça etkili bir filtreleme sürecini takip eder ve TF-TDF'e benzer bir puanlama işlevi kullanır. Özellikle, noktalama işaretleri / stopwords ile ayrılmayanlar, aynı zamanda en az izin verilen frekans faktörü ve bir dizi kelime olarak tanımlanan bir kesme sabiti göz önünde bulundurularak aday anahtarlar belirlenir. Daha sonra sistem, aday terimleri TF ve TDF puanlarının yanı sıra terim pozisyonu ve bileşik terimler için tek terim üzerinde bir artış faktörü dikkate alarak sıralar.

KP-Miner sisteminde anahtar kelime çıkarımı üç adımlı işlemle gerçekleşmektedir. Bunlar aday anahtar sözcük seçimi, aday anahtar sözcük ağırlık hesaplaması ve anahtar sözcük ayrıntılandırılması işlemlerinden oluşmaktadır. İlk adımda aday anahtar kelimeleri ortaya çıkarmak için bir dizi kural kullanılır. Daha sonra tek veya bileşik aday anahtar kelimelerinin ağırlığını hesaplamak için Eşitlik 2.19 kullanılır:

$$w_{ij} = tf_{ij} * tdf * B_i * P_f \quad (2.19)$$

Eşitlik 2.19'da w_{ij} – belgedeki t_{ij} teriminin ağırlığı, tf_{ij}- belgedeki terim sıklığı, B_i - arttırıcı faktör, P_f- pozisyonla ilişkili faktör terimidir.

Ters doküman frekansı *tdf* ise,

$$tdf = \frac{log_2 N}{n} \quad (2.20)$$

dir.

Ağırlık hesaplama aşaması gerçekleştirildikten sonra bu örtüşme iyileştirme yoluyla ele alınır. Son olarak, adaylar beş faktörün entegre edilmesiyle sıralanır: belgenin terim ağırlığı, belgenin terim sıklığı, TDF terimi, destekleyici bir faktör ve terim konumu.

Tam Otomatik Anahtar Kelime Çıkarma modeli

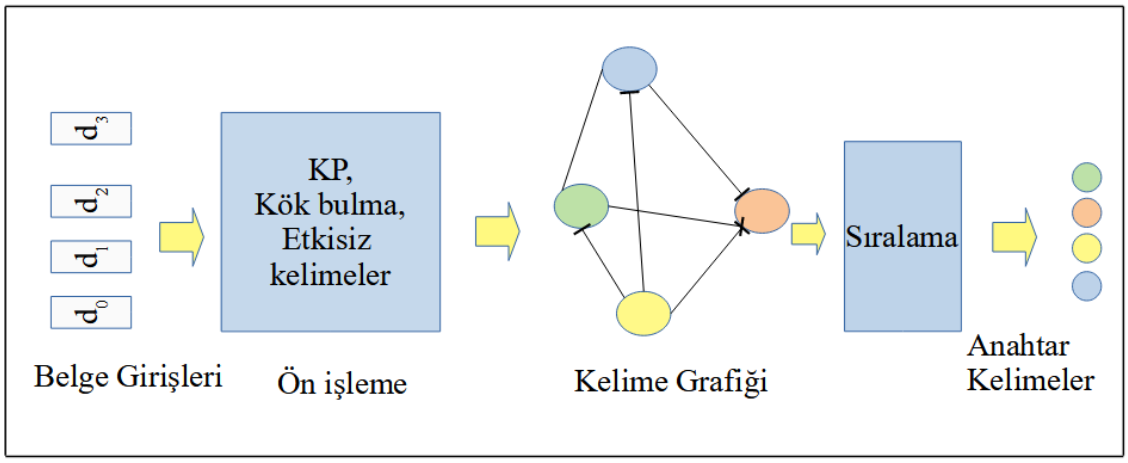
Tam Otomatik Anahtar Kelime Çıkarma (TO-AKÇ - Totally Automated Keyword Extraction- TAKE) tek belgeden anahtar kelimeler çıkarmak için denetimsiz ve alandan bağımsız bir yöntemdir. Pay (2016) çalışmasında belgeden otomatik anahtar kelime çıkarımı için TO-AKÇ yöntemi önermiştir. Ayrıca anahtar kelime olarak seçileceğine karar vermek için dinamik eşik fonksiyonları kullanarak TO-AKÇ ile yüksek duyarlılık değeri elde etmişlerdir.

TO-AKÇ dört adımdan oluşmaktadır:
1. aday anahtar kelime çıkarımı,
2. aday anahtar kelimelerin filtrelenmesi,
3. aday anahtar kelimelerin özellik hesaplaması,
4. eşik fonksiyonunu (threshold function) kullanarak aday anahtar kelimelerinin seçimi.

Her kelime için bir KP etiketi atamak için bir KP etiketleyici kullanır ve ardından tüm isim-öbeklerini (noun-phrase) aday anahtar kelimeler olarak işaretler. İsim-öbeklerini, sıfır veya daha fazla sıfat ve ardından bir veya daha fazla isimden oluşur. TO-AKÇ, aday anahtar kelime puanlarını HO-AKÇ ile aynı şekilde hesaplar. HO-AKÇ için aday anahtar kelimeler bulmak için bir durdurma listesine (stop-list) ihtiyaç duyar ve TO-AKÇ aday anahtar kelimeler listesine filtre uygulamak için bir durdurma listesine ihtiyaç duyar. Dinamik eşik fonksiyonu tarafından hesaplanan değerden daha yüksek bir puana sahip tüm aday anahtar kelime, anahtar kelime olarak çıkarılır.

2.3.2. Grafik tabanlı sıralama yaklaşımları

Anahtar kelime çıkarımı için istatistiksel yaklaşım yoluyla terimler hakkında istatistiksel bilgiler elde edilebilir, ancak kelimeler ve cümleler arasındaki ilişkiyi tanımlayamaz. Grafik tabanlı sıralama iki kelime arasındaki ilişkiyi açıklar ve konu tabanlı kümeleme anlamsal bilgiyi kelimelere ekler. Grafik tabanlı yaklaşım genel olarak giriş belgesinden bir grafik oluşturmak ve düğümlerini, grafik tabanlı bir sıralama yöntemi kullanarak önemlerine göre sıralanmaktadır. Grafik tabanlı anahtar kelime çıkarımı ile kelimeler ve kelimeler arasındaki ilişkiyi dikkate alır, ancak basit ikili eş oluşum (co-occurrence) sınırlıdır ve diller arasındaki belgelerin çoklu ilişkisini (multi) iyi ifade edemez. Grafik tabanlı sıralama algoritmaları, temel olarak grafik yapısından çizilen bilgilere dayanarak, grafik içindeki bir tepe noktasının önemine karar vermenin bir yoludur. Bu yaklaşımlar, iki kelime öbeği arasındaki ilişkiyi ölçmek için istatistiksel yöntemler kullanır ve iki kelime öbeği birbiriyle ilişkili ise kenar ekler. Sıralama iki kelime arasındaki ilişkiyi açıklar ve konu tabanlı kümeleme anlamsal bilgiyi kelimelere ekler.



Şekil 2.18. Grafik tabanlı anahtar kelime çıkarımı için kullanılan iş akışı

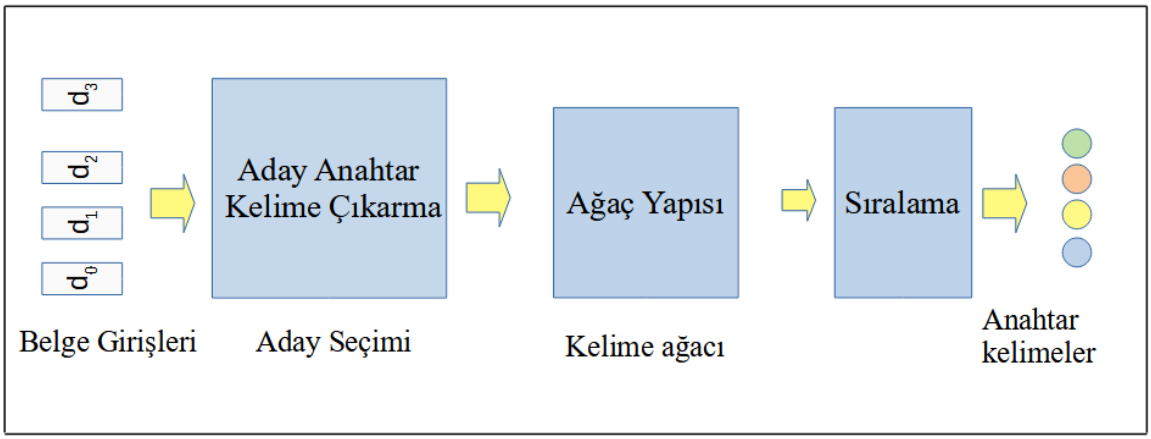
Şekil 2.18'de bir belge deposu veya belge veri kümesi verildiğinde grafik tabanlı anahtar kelime çıkarımı için kullanılan iş akışı görülmektedir. Grafik tabanlı yaklaşımlar genel olarak dört temel adımdan oluşur: aday kelime seçimi, sözcük grafik yapısı, aday sözcük sıralaması ve anahtar kelime çıkarımı. Konu modellerine dayalı anahtar kelime çıkarma, konu dağılımına çok bağımlıdır ve aynı alanda diller arası belgeleri temsil etmek zordur.

Literatürde Metin Sıralama modeli ilk olarak Mihalcea and Tarau (2004) tarafından önerilmiştir ve grafik tabanlı metin işleme, anahtar kelime ve cümle çıkarma görevlerine uygulanmıştır. Sonrasında, Huang, Tian, Zhou, Ling ve Huang (2006) otomatik anahtar sözcük çıkarımı için aynı zamanda bağlılık ve merkezlilik merkeziyetine dayanan denetimsiz bir yöntem kullanmışlardır. Wan and Xiao (2008) çalışmalarında tek bir belge anahtar kelime çıkarımını geliştirmek için daha fazla bilgi sağlamak amacıyla az sayıda en yakın komşu belgeyi kullanmayı (ExpandRank) önermektedir. Litvak ve Last (2008), belgelerden anahtar kelimelerin çıkarılması için denetimsiz bir hiperlink kaynaklı konu arama (HKKA) algoritması önermişlerdir. Litvak, Last, Aizenman, Gobits ve Kandel (2011) grafik tabanlı ve dilden bağımsız anahtar kelime çıkarıcı Dereceye Dayalı Çıkarım (DDÇ - Degree-based Extractor-DegExt) yöntemi tanıttılar. Ayrıca DDÇ'i en gelişmiş yaklaşımlarla karşılaştırarak hassasiyet, uygulama basitliği hem de hesaplama karmaşıklığı açısından üstün olduğu görülmüştür. Vega-Oliveros, Gomes, Milios ve Berton (2019) otomatik anahtar kelime çıkarımı için kelime sıralamasının optimal kombinasyonunu bulmayı amaçlayan çoklu merkezlilik dizi (Multi-Centrality Index (MCI) yaklaşımını sunmaktadır. Li ve diğerleri (2019) grafik tabanlı sıralama ve konu tabanlı kümeleme kullanarak anahtar kelime çıkarımı için denetimsiz bir yaklaşım önermişlerdir. Ayrıca, diğer algoritmalardan farklı olarak, yaklaşımımı gerçekleştirmek için sadece toplama içi kaynakları (Within-Collection Resources) kullanmışlardır.

Rabby, Azad, Mahmud, Zamli ve Rahman (2020), Ağaç Tabanlı Anahtar Kelime Çıkarma Tekniği (AT-AKÇ - Tree-based Keyphrase Extraction Technique – TeKET) olarak adlandırılan önerilen denetimsiz anahtar kelime çıkarma tekniği önermişlerdir. AT-AKÇ sınırlı istatistik bilgisi kullanan ve eğitim verisi gerektirmeyen, alandan bağımsız bir tekniktir. Şekil 2.19'da göründüğü gibi AT-AKÇ ile anahtar kelime çıkarma süreci üç temel adımı içerir.

Bunlar;

1. aday anahtar kelime seçimi veya ön işleme,
2. aday anahtar kelime işleme veya basitçe işleme ve
3. son anahtar sözcükleri veya son işlemlerin sıralanması ve seçilmesidir.

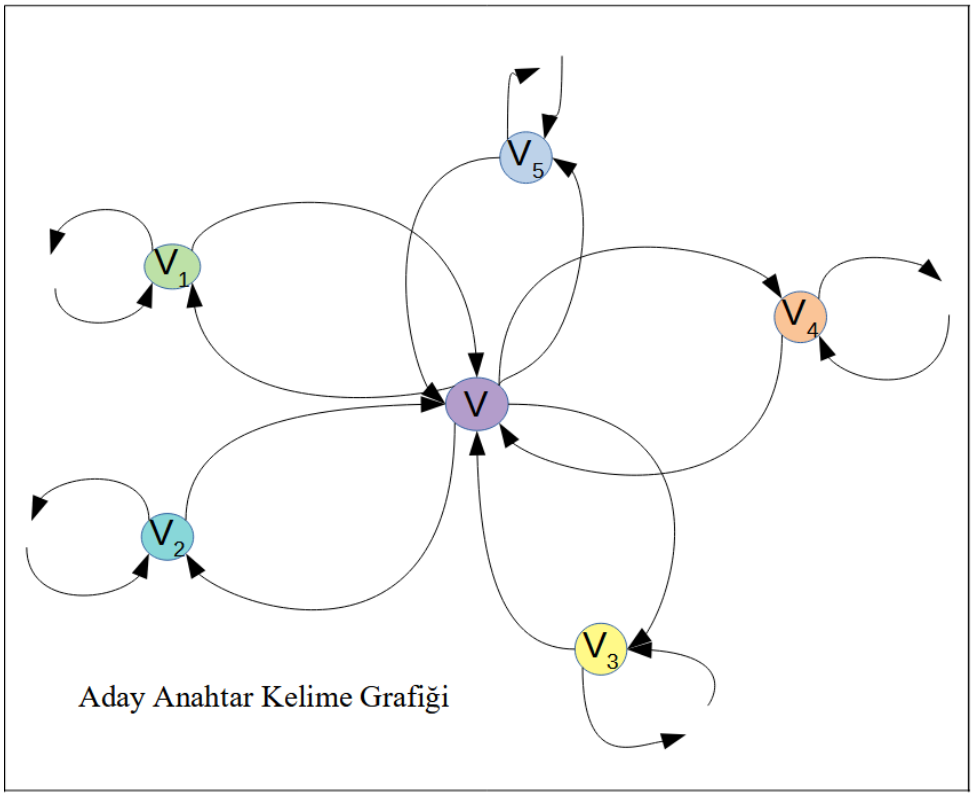


Şekil 2.19. AT-AKÇ modeli

Aday anahtar sözcükler çıkarıldıktan sonra, son anahtar kelime olma olasılığı daha düşük olan anahtar sözcük öbeklerini filtrelemek için bir temizleme sürecinden geçirilir. Bunun için aşağıdaki koşullar uygulanır: (i) alfabetik olmayan karakterler içeren herhangi bir aday anahtar kelime öbeği, (ii) tek alfabetik kelime(ler) içeren herhangi bir aday anahtar kelime öbeği ve (iii) herhangi bir aday anahtar kelime öbeğinin sıklığı. İlk iki koşul, genel olarak insan okuyucu için bir anlam ifade etmeyen aday anahtar kelime öbeklerini filtreler; ikincisi ise popüler olmayan tüm aday anahtar kelime öbeklerini listeden filtreler.

Metin Sıralama algoritması

Orijinal Metin Sıralama yöntemi Brin ve Page (1998) tarafından 1998 yılında sunulmuş olup en çok alıntı analizi, sosyal ağlar ve ağ ortamındaki bağlantıları analiz etmek amacıyla kullanılmaktadır. Daha sonra, Mihalcea ve Tarau (2004) tarafından ilk grafik tabanlı anahtar kelime çıkarma ve metin özetlemede kullanılan Sayfa Sıralaması tabanlı bir algoritmasını sunulmuştur. Metin Sıralama, her kelimeye bir pos etiketi atamak için konuşma etiketinin bir bölümünü kullanır. Sıfatları ve isimleri anahtar kelimelerin olası bileşenleri olarak görür. Bu kelimeler grafikte aday kelimeler grafiğe V düğüm olarak eklenir. Daha sonra, tüm kelimeler arasındaki bitişik ilişki E grafiği kelimesinin kenar kümesini oluşturur, böylece $G=(V, E)$ aday anahtar kelime grafiğini oluşturur. Kenarı oluştururken, a kelimesi b kelimesine bitişikse, kelime grafiğine iki yönlendirilmiş kenar $a \rightarrow b$ *ve* $b \rightarrow a$ eklenir. Yani, G grafiği kelimesi, Şekil 2.20'da gösterildiği gibi yönlendirilmiş bir grafiktir.



Şekil 2.20. Metin sıralama modeli

Şekil 2.20'deki Metin Sıralama modelinde G = (V, E) grafiği verildiğinde, t(u), u düğümünün Metin Sınıflandırma değerini gösterir. Daha sonra, t (u) Eşitlik 2.21 ile hesaplanabilir:

$$t(u) = d\sum_{d\in adj[u]} p(v \to u)t(v) + (1-d) * \frac{1}{v} \quad (2.21)$$

Her bir düğüme atanan ilk puan 1'e eşittir ve daha sonra Sayfa Sıralaması algoritması birleşene kadar çalışır. En üstteki üçüncü puanlama kelimeleri sonradan işleme için seçilir. Bu noktada, seçilen kelimelerin herhangi birinin yan yana görünüp görünmediğini belirlemek için orijinal metne dayalı olarak çok kelimeli anahtar kelimeler oluşturulur. Belgenin üçüncü en yüksek puanlama sözcükleri listesinde başka bir kelimenin yanında görünmeyen tüm kelimeler tek kelimeli anahtar kelimeler olarak, geri kalanı ise çok kelimeli anahtar kelimeler olarak çıkarılır. Metin Sınıflandırma için kullanılan Eşitlik 2.22'de hesaplanmaktadır:

$$WS(v_i) = (1-d) + d * \sum_{v_j\in İn(v_i)} \frac{w_{ji}}{\sum_{V_k\in Out(V_j)}} WS(v_j) \quad (2.22)$$

Metin Sınıflandırma algoritmasının çalışma adımları:

1. Tek kelimeler dikkate alınarak ve hiçbir n-gram kullanılmamaktadır. Konuşma Parçası ile tokenize edilir. Daha sonra çoklu kelimeler yeniden oluşturulur.
2. Sözdizimsel filtreler tüm sözcük birimlerinde kullanılır.
3. Ağırlıksız yönlendirilmemiş grafik elde etmek için, sözcüksel birimler bir arada ortaya geldiğinde, N kelimelik bir pencerede oluşturulur ve kenarı çizilir.
4. Sıralamak için, kelimeler Metin Sınıflandırma algoritmasını çalıştırır.
5. En iyi sözcükler alınır.
6. Çok kelimeli anahtar kelimeler, bitişik anahtar kelimeler daraltılarak oluşturulur.

Garg, Favre, Reidhammer ve Hakkani (2009) çalışmalarında kümeleme ve Metin Sınıflandırma kullanarak toplantı özeti çıkartan bir sistem önermişlerdir. Witt, Milz ve Seifert (2018) çalışmalarında temel anahtar kelime çıkarma yöntemine agnostik olan TF-TDF tabanlı puanlama ve anahtar kelimelerin yeniden sıralanmasını önermektedirler. Mevcut algoritmalar ile karşılaştırıldığında bu yöntem daha iyi performans göstermiş, puanlama ve toplama yaklaşımının doğru bir yaklaşım olduğunu ortaya çıkarmışlardır. Li, Huang Chen ve Wang (2019) kısa metin anahtar kelime çıkarımı için aday anahtar kelime grafiğini oluşturdular ve kelimeler arasındaki anlamsal bilgileri yakalamak için Word2Vec ve Doc2Vec kullanmışlardır. Düğümler arasındaki atlama olasılığını hesaplayarak oluşturulan anahtar kelimelerin ağırlıklarını ayarlayarak ağırlıklı skoru elde etmiş, üretilen anahtar kelimeleri sıralamışlardır. Azarafza, Feizi-Derakhshi ve Shendi (2020) Metin Sınıflandırma tabanlı mikrobloglardan Farsça anahtar kelime çıkarımı için Farsça'daki Telegram yayınından otomatik anahtar kelimeler çıkarmayı öneren bir yöntem sunmuşlardır.

Tek Sıralama algoitması

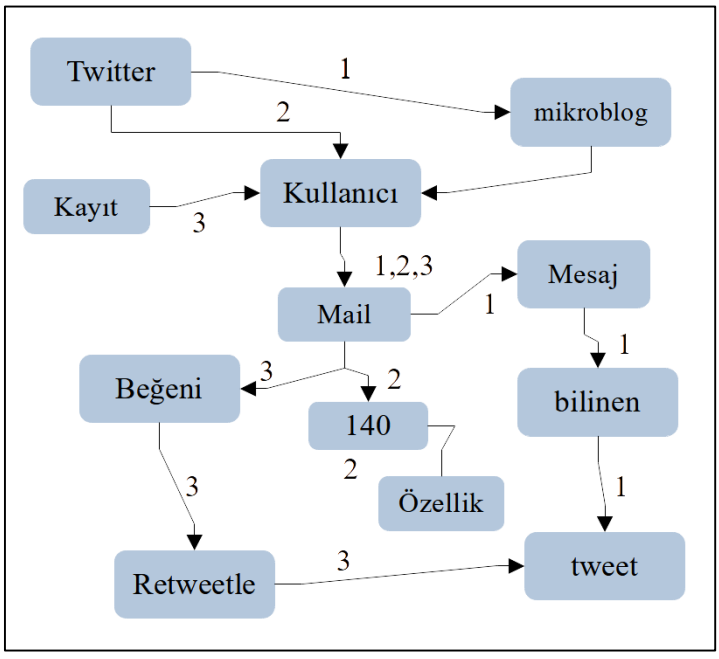
Tek Sıralama (Singlerank) (Wan ve Xiao, 2008) modeli grafikteki kenarları ağırlıklandırmak için eşzamanlı tekrarlama sayısını kullanan Metin Sınıflandırma modelinin bir uzantısıdır. Tek Sıralama grafik tabanlı sıralama algoritmasına dayalı olarak her bir belge için kelime puanlarını hesaplar. Tek Sıralama ağırlıkları kenarlara birleştiren Metin Sınıflandırma'nın genişletilmiş bir türüdür. Dolayısıyla, her kenar ağırlığı, karşılık gelen iki sözcüğün birlikte meydana gelme sayısına eşittir. Daha sonra V_i düğümü için skor fonksiyonu hesaplanır (Eşitlik 2.23).

$$WS(V_i) = 1 - \gamma + \gamma * \sum_{j \in N(V_i)} \frac{1}{N(V_i)} \frac{\#_c ij}{\sum_{V_k \in N(V_j) \#_c jk}} WS(V_j) \quad (2.23)$$

Burada $\#_c ij, i$ ve j kelimesinin eşdizimliliğidir. İşlem sonrası aşamada, bir metin belgesindeki her isim ve sıfat dizisi için kurucunun kelime puanları toplanır ve T'nin üst sıralarında yer alan adaylar anahtar sözcük olarak döndürülür.

Dereceye Dayalı Çıkarım modeli

Dereceye Dayalı Çıkarım (DDÇ) - grafik tabanlı, dilden bağımsiz ve alandan bağımsız bir yöntemdir. Grafik, basit sözdizimsel grafik gösterimleri ve belgenin yapısal özelliklerini dikkate alarak oluşturulmuştur. Litvak ve diğerleri (2011) çalışmalarında anahtar kelime çıkarımı için DDÇ yöntemini önermiştir. Ayrıca DDÇ'i en gelişmiş GenEx ve Metin Sınıflandırma yaklaşımlarla karşılaştığında hassasiyet, uygulama basitliği hem de hesaplama karmaşıklığı açısından üstün olduğu iddia edilmektedir. DDÇ algoritmasında (Litvak, Last ve Kandel, 2013) etkisiz kelimeler (stop words) önce silinir ve daha sonra köşeler arasındaki yayların yalnızca noktalama işaretleriyle ayrılmamış herhangi bir cümleye bitişik kelimeler için çizildiği bir grafik oluşturulur. En yüksek derecelere sahip köşeler, anahtar kelimedeki adaylara karşılık gelmektedir. Metin Sınıflandırma ile karşılaştırıldığında, bu algoritma hesaplama açısından daha az karmaşıktır. Şekil 2.21, sırasıyla bir örnek metni ve grafik temsilini göstermektedir.



Şekil 2.21. DDÇ modeli

Konum Sıralaması

Konum Sıralaması (PositionRank) (Florescu ve Caragea, 2017 (a)) modeli çeşitli aday ve son anahtar sözcükler elde etmek için sözcük oluşumlarının konum bilgilerini alan ve bunları taraflı bir Sayfa Sıralaması algoritmasına yerleştiren, anahtar kelimeleri çıkarmak için geliştirilen denetimsiz grafik tabanlı bir işlemdir. Özellikle, bir kelimenin tüm konumlarını taraflı ağırlıklı bir Sayfa Sıralamasına ekler. Son olarak, anahtar sözcükler puanlanır ve sıralanır.

Konum Sıralaması algoritması üç temel adımı içerir:
1. kelime düzeyinde grafik yapısı
2. Pozisyona Dayalı Sayfa Sıralaması tasarımı
3. aday ifadelerin oluşturulması.

Örneğin D hedef belgesinde kelime grafiği G = (V, E) olsun ve $v_i \in V$ tepe noktasına konuşma filtrelerini geçen her benzersiz kelime karşılık gelsin. Bu köşelere karşılık gelen kelimeler d'de bitişik belirteçlerin bir penceresi içinde ortaya çıkarsa, iki v_i ve v_j köşesi bir kenar $v_i, v_j \in E$ ile bağlanır. Daha sonra Sayfa Sıralaması algoritması kullanarak G 'deki köşeleri puanlanır. Yani, v_i köşesi için s skoru, Eşitlik 2.24'dü özyinelemeli hesaplayarak elde edilir:

$$S(v_i) = \alpha * p(v_i) + (1-\alpha) \sum_{v_j \in Adjv_i} \frac{w_{ji}}{\sum_{v_k \in Adjv_j} w_{jk}} s(v_j) \quad (2.24)$$

burada α – sönümleme faktörüdür ve $p(v_i)$ tepe noktasına atanan bir ağırlıktır.

Son olarak belgedeki bitişik konumlara sahip aday kelimeler, ifadelere birleştirilir ve kelime öbeğini içeren bireysel kelimelerin toplam puanları kullanılarak puanlanır.

Konum Sıralaması algoritması, bir terimin anahtar kelime olup olmadığını belirlemeden önce belgedeki terimlerin hem konumunu hem de sıklığını kontrol eder ve belgenin başında görünen ve diğer aday ifadelerden daha sık olan kelimelere öncelik verir. Yani, üst bölümlerde görünen kelime veya kelime grubunun, özet ve girişin, bu algoritmanın uygulanması ile bir anahtar sözcük olarak seçilme olasılığı daha yüksektir. Florescu ve

Caragea (2017 (a) bilimsel belgelerden anahtar kelimelerin çıkarılması için denetimsiz bir model olan Konum Sıralama ile ve üç veri seti üzerinde gerçekleştirdikleri ilk çalışmada %26,4 iyileştirme, devamında aynı model üzerinde %29,09'a varan iyileştirmeler elde etmişlerdir (Florescu ve Caragea 2017 (b).

Anlamsal Bağlantı Farkındalık AKÇ (ABF-AKÇ - Semantic Connectivity Aware Keyword Extraction- sCAKE) modeli

Duari ve Bhatnagar (2019) grafik oluşturma ve puanlama yöntemlerinin birleşimi olan belgedeki kelimelerin anlamsal bağlantısına dayalı yeni, parametresiz bir anahtar kelime çıkarma yöntemi ABF-AKÇ modeli önermişlerdir. ABF-AKÇ pragmatiklerin her bir cümlenden bir sonraki cümlesine geçmesine dikkat eden yeni cümle temelli grafik oluşturma yaklaşımına Bağlama Duyarlı Metin Grafiği (BDMG - Context-Aware Text Graph-CAG) dayanmaktadır. ABF-AKÇ algoritmasının çalışma üç aşamadan oluşur:

1. Aday Filtrasyonu: Aday anahtar kelimeleri KP etiketlemesinden sonra tutulan isimler ve sıfatlar olarak tanımlanır. Listenin gövdeli sürümü aday olarak kabul edilir ve orijinal metnin gövdeli sürümü ile birlikte bir sonraki aşamaya geçilir.
2. Grafik Yapısı: BDMG parametresiz bir grafik oluşturma yöntemi oluşturur. Bu yaklaşım, yazılı iletişim pragmatiklerini yakalar ve oluşum bağlamına bağlı olarak birbirleriyle yakından ilişkili kelimeleri birleştirir. BDMG grafiklerinde karşılık gelen grafiklere göre daha iyi performans gösterdiğini ortaya koymaktadır.
3. Kelime puanı: önerilen anlamsal bağlantı tabanlı kelime puanlama yöntemini (SCSkor) kullanarak adaylar için kelime puanı hesaplanır. (SCSkor), metindeki bağlamsal hiyerarşisini, anlamsal bağlanabilirliğini ve konumsal ağırlığını dikkate alarak kelimelerin alaka düzeyini hesaplar. Bu yöntem, semantik yönleri herhangi bir dilsel araç kullanmadan sadece kelime-kelime ilişkileri temelinde yakalamaya çalışır. Daha sonra adaylar kendi SCSkor'larına göre sıralanır ve kullanıcı en iyi k adaylarını çıkarabilir.

Grafik Tabanlı Tek Belgede Anahtar İfade Çıkarma (GTTB-AİÇ - Graph-Based Technique for Extracting Keyphrases in a Single-Document - GTEK) modeli

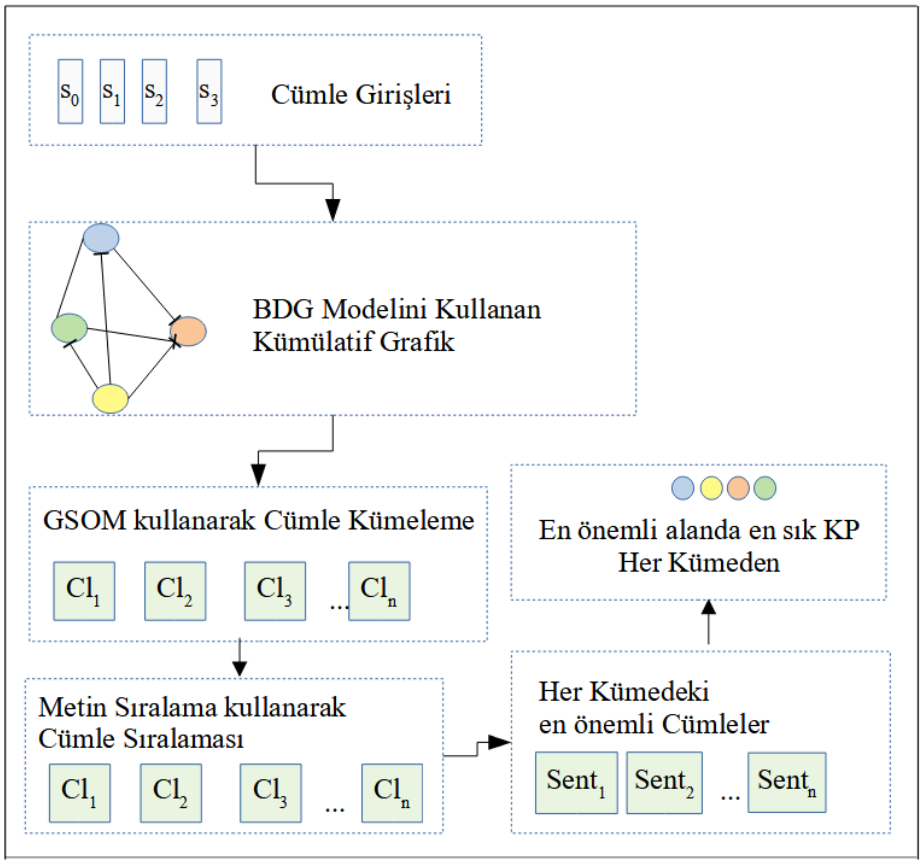
Alfarra, M. ve Alfarra, A. (2018) tarafından tek bir belgede anahtar sözcüklerin çıkarılması için grafik temelli bir GTTB-AİÇ yöntemi önerilmiştir. Yöntem, metnin sözcüklerini temsil

eden grafik modeline ve küme cümlelerini G-GSOM algoritmasına bağlıdır. GTTB-AİÇ , cümleleri grafik model kümeleri halinde gruplandırır ve sonra Metin Sınıflandırma algoritmasını kullanarak sıralar. Son olarak, yüksek sıralı cümlelerdeki en sık kullanılan ifadeler belge anahtar sözcükleri olarak seçilir.

GTTB-AİÇ modelin üç avantajlı faktöre dayanmaktadır:

1. Vektör uzay modelinden daha doğru olan grafik gösterimi.
2. En önemli anahtar sözcükleri kapsayacak şekilde en benzer cümleleri kümelendirme.
3. Cümleleri puanlamak ve sıralamak için Metin Sınıflandırma algoritmasının kullanımı.

İlk olarak, grafik tabanlı modeli tanımlamak için grafik teorisinden bazı temel kavramlar tanımlanır ve Belge Dizini Grafiği (BDG - Document Index Graph-DIG) modeli açıklanır. Şekil 2.22'de BDG modelin köşelerindeki ayrıntılı verileri açıklar. Grafikteki her köşe, sözcüğün cümle içindeki tüm oluşumu hakkındaki bilgileri grafikten cümlelerin çıkarılmasını kolaylaştırır.



Şekil 2.22. GTTB-AİÇ modeli

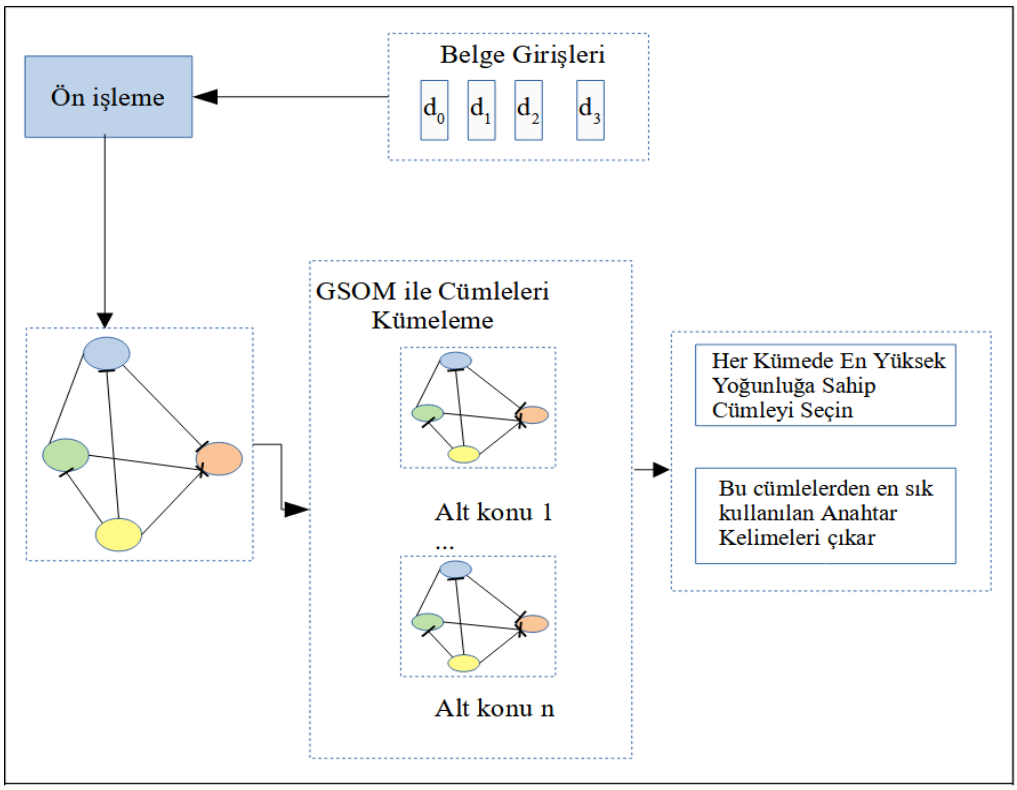
GTTB-AİÇ 'de tek bir belgeden anahtar sözcükleri çıkarmak Doküman Temsili, Cümle Kümelemesi, Cümle Sıralaması ve Anahtar Kelimelerinr Seçimi olmak üzere dört aşamaya ayrılabilir. Doküman Temsili, doküman, bir grafiğe dayanan bir grup köşe noktası olarak temsil eder. Cümle Kümelemesi, kümeleme algoritmasını kullanarak belgenin ana konularını keşfeder. Cümle Sıralamasında en önemli cümleleri tespit etmek için Metin Sınıflandırma algoritması kullanılarak sıralanır. Anahtar Kelimelerin Seçimi: aşamasında her kümenin anahtar sözcük adayları oluşturulur, ardından belgede en sık kullanılan anahtar kelime seçilir.

Grafik Tabanlı Yoğunluk Zirveleri Sıralaması Yaklaşımı (GT-YZS - Graph-based Density Peaks Ranking Approach for Extracting Keyphrases - GDREK)

Alfarra, Alfarra ve Salahedden (2019) yoğunluk zirveleri sıralama kullanılarak tekil veya çoklu dokümanlardan anahtar kelime çıkarımı için GT-YZS adıyla yeni bir metod önerilmiştir.

Önerilen algoritma birçok anahtar kelime çıkarımı yönteminde olan problemleri, metni tamamen tanımlayan küçük bir grup anahtar kelimeleri, metnin ana fikrini kapsayan anahtar kelimeleri, en önemli cümleler listesinden anahtar kelimeleri çıkarma ve cümleleri sıralamak için yoğunluk zirveleri kullanma yoluyla çözmeye çalışır.

Şekil 2.23’te görülebileceği gibi en sık kullanılan kelimeler ve sözcükler seçilen anahtar kelimeler olarak kabul edilir.



Şekil 2.23. GT-YZS-AKÇ modeli

GT-YZS-AKÇ algoritması,ön işleme (etkisiz kelime ve kök bulma işlemleri), metni Belge Grafik Modeli (BGM) kullanarak grafik olarak temsil etme, cümleleri G-GSOM kullanarak kümelemek ve cümleleri DP kullanarak paralel sıralama ve son olarak aday anahtar sözcükler listesinin çıkarılması ve ardından TF-TDF algoritması kullanılarak anahtar sözcüklerin seçilmesi aşamalarından oluşur.

GT-YZS-AKÇ, her cümlenin yoğunluğunu bulmak için tüm cümleler için bir Grafik Tabanlı Benzerlik Matrisi (GTBM) oluşturur. Cümlelerin yoğunluğu S_{ij} ile diğer tüm cümle arasındaki benzerlik değerlerinin toplamını metnin tamamındaki cümle sayısına bölünmesiyle (Eşitlik 2.25) hesaplanır.

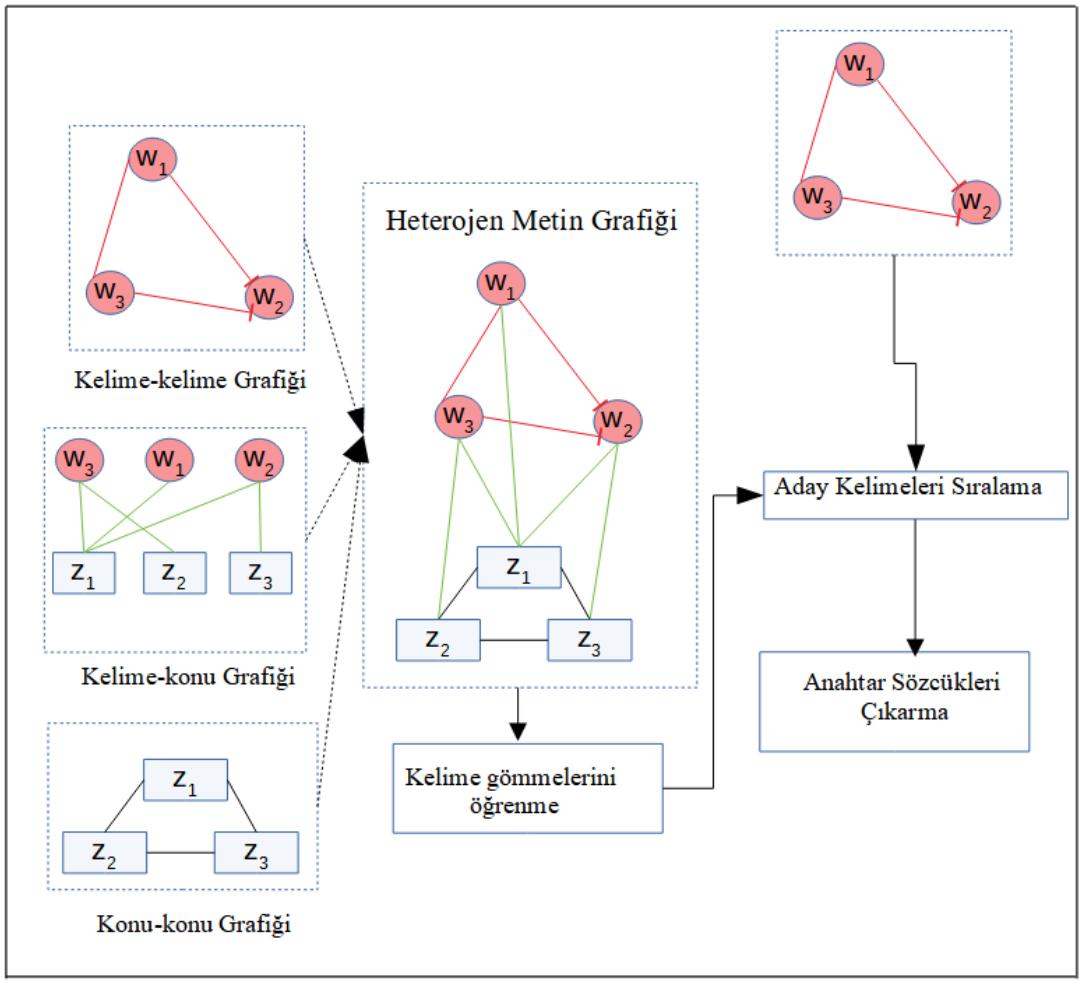
$$DS_i = \frac{\sum_{j=1, j\neq i}^{K}(sim_{ij})}{K} \tag{2.25}$$

Burada , DS_i – diğer tüm cümlelere göre cümle yoğunluğu, K- metindeki cümle sayısı, sim_{ij} – (i) cümlesinin (j) cümlesine benzerliğidir.

2.3.3. Gömme tabanlı yaklaşımlar

Kelime Gömme (Word Embedding) son yıllarda DDİ görevlerinde yaygın olarak uygulanmaktadır. Kelime gömme teknikleri, sözcükler arasındaki anlamsal ve bazen de sözdizimsel ilişkileri yansıtacak şekilde her kelimeyi düşük boyutlu sürekli bir alana vektör temsili ile gömmeye çalışır. Dağıtılmış kelime temsili olarak da bilinen kelime gömme, son yıllarda çeşitli metin madenciliğinde yaygın olarak kullanılmaktadır. Literatüre bakıldığında Yapay sinir ağı dil mimarilerine dayanan birçok yöntem önerilmiştir. Son zamanlarda Mikolov, Chen, Corrado ve Dean (2013) kelime gömmeyi büyük ölçekli bir metin topluluğundan öğrenmek için sürekli kelime torbası modeli ve Skip-gram yöntemini tanıtmışlardır. Bu iki kelime gömme modeli, etkinlik ve verimlilikleri nedeniyle hem akademi hem de endüstriden büyük ilgi görmüş ve kelime gömme yaklaşımlarının mevcut popülaritesini arttırmıştır. Bazı araştırmalarda kelime gömme teknikleri anahtar kelime çıkarımı için Sayfa Sıralaması tabanlı altyapıya entegre edilmiş olsa da, bu çalışmalarda kullanılan gömmeler genelde Wikipedia'daki SENNA modeli ve Skip-gram gibi modelleri ile öğrenilmektedir. Mahata, Kuriakose, Shah ve Zimmermann (2018), bilimsel makalelerden anahtar kelimeleri sıralamak ve çıkarmak için gömme ifadeleri kullanan denetimsiz bir yöntem olan anahtar kelimeden vektöre (Key2Vec) yaklaşımını sunmuşlardır. Başka bir çalışmada Mahata, Shah, Kuriakose, Zimmermann ve Talburt (2018) anahtar kelimeleri çıkarmak ve sıralamak için tema ağırlıklı kişiselleştirilmiş Sayfa Sıralaması algoritması ve sinirsel kelime gömmeleri bir kombinasyonunu kullanan denetimsiz bir teknik sunmuşlardır.

Zhang ve diğerleri (2019) çalışmalarında metin belgelerinden anahtar ifadeleri çıkarmak için rastgele yürüyüş tabanlı bir sıralama yöntemi önermişler, daha sonra çok çeşitli yararlı bilgileri rastgele yürüyüş modeline entegre etmek için gömme yöntemini kullanmışlardır. Kelime Gömmeleri ve rastgele yürüyüş Sıralama modeli (KG-RYS-Word Embeddings and random-walk Ranking model) yöntemin, anahtar kelime çıkarımı için temel ifade puanlama ve uzunluğa dayalı sıralı puanlama yöntemi kullanır.



Şekil 2.24. KG-RYS modeli

Şekil 2.24'te KG-RYS modeli anahtar sözcük çıkarımı için tüm Metin Sınıflandırma algoritmasına benzer heterojen bir metin grafiği oluşturulur. İkinci adımda bazı önemli bilgi türlerinin korunduğu kelime gömmeyi öğrenmek için heterojen metin grafiği gömme modeli önerilmiştir. Modifiye rastgele yürüyüş sıralama modeli aday sözcükleri puanlamak için tasarlanmıştır. Son adımda, ardışık kelimeler, deyimler veya n-gramlar yeni bir öbek puanlama modeli ile puanlanır ve en yüksek puan alan aday ifadeler / kelimeler belgenin son anahtar sözcükleri olarak verilir.

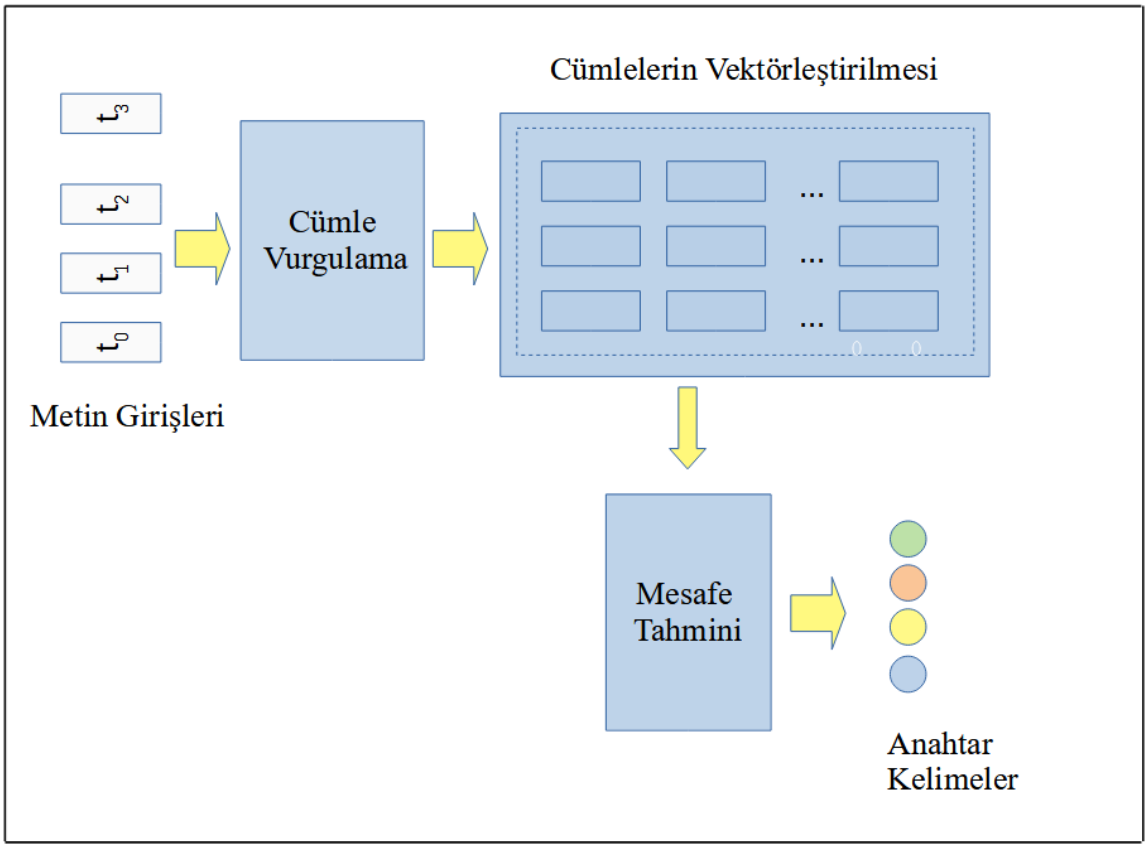
Gero ve Ho (2019), biyomedikal metinden anlamlı anahtar sözcükleri otomatik olarak tanımlamak için dört modülden oluşan bir yöntem olan yeni NamedKeys'i önermişlerdir. NamedKeys, biyomedikal belgelerden yazar tarafından atanan anahtar kelimeleri çıkarmak için adlandırılmış varlık tanıma, ifade yerleştirme, ifade kalitesi puanlama, sıralama ve kümeleme özelliklerini birleştirir. PubMed özetleri üzerindeki performans değerlendirmesi,

NamedKeys 'in mevcut son teknoloji anahtar kelime çıkarma modellerinde önemli gelişmeler sağladığını göstermektedir.

Gömme Sıralama modeli

Gömme Sıralama (EmbedRank) belgenin bağlanabileceği tüm bir korpus yerine yalnızca geçerli belgenin kendisini gerektirir (Bennani-Smires, Musat, Hossmann, Baeriswyl ve Jaggi, 2018). Gömme Sıralama bir veya daha fazla isim içeren ifadeleri yani KP sekanslarına dayalı aday ifadeler çıkarır.

Şekil 2.25'te Gömme Sıralama modeli görülmektedir. Modelde ilk aşamada metinden aday cümleleri çıkarılır. Daha sonra, karşılaştırma için cümle ve metin gömmeleri kullanımı (vektör boyutları eşit olarak kullanılır) ve son olarak adayların sıralamasına dayanarak (anahtar sözcüklerin vektörleri ile belge vektörü arasındaki vektör mesafesine göre), bir dizi N anahtar sözcük grubu seçilir.



Şekil 2.25. Gömme Sıralama modeli

2.3.4. Dilbilimsel yaklaşımlar

Dilbilimsel yaklaşımlar anahtar kelimeleri tanımlamak için kelime analizi, konuşma kısmı etiketleme, sözdizimsel analiz, dilbilgisi analizi gibi dil özelliklerini kullanan benzer yöntemler kullanılır. Bu yaklaşımlar esas olarak, etkili sonuçlar üretebilecek terimin anlambilimini inceler. Anahtar kelime çıkarmaya yönelik diğer yaklaşımlar temel olarak yukarıda belirtilen yöntemleri birleştirir veya anahtar kelime çıkarmanın görevlerinde konum, uzunluk, kelimelerin yeri, kelimelerin etrafındaki HTML etiketleri gibi bazı sezgisel bilgileri kullanır.

N-gram dil modelleri ile anahtar sözcük ayıklama (Tomokiyo ve Hurst, 2003), bir ön plan kümesinde (hedef belge) ve bir arka plan kümesinde (belge kümesi) hem unigram hem de n-gram dil modelleri oluşturur. Özellikle, tümcecik düzeyinde, her bir tümcecik için, öbeklik, ön plan korpus ve arka plan korpus üzerindeki unigram ve n-gram dil modelleri arasındaki ayrışma ve bilgisellik hesaplanır. Daha sonra, ifade ve bilgisellik, her bir cümle için nihai bir puan olarak toplanır. Son olarak, ifadeler bu puana göre sıralanır. Bir diğer çalışmada benzer yaklaşımla Haddoud, Mokhtari, Lecroq ve Abdeddaïm (2015) bilimsel makalelerden otomatik anahtar kelime çıkarımı için dil bilgisini kullanarak aday terimleri filtreleme işlemi gerçekleştirmişlerdir.

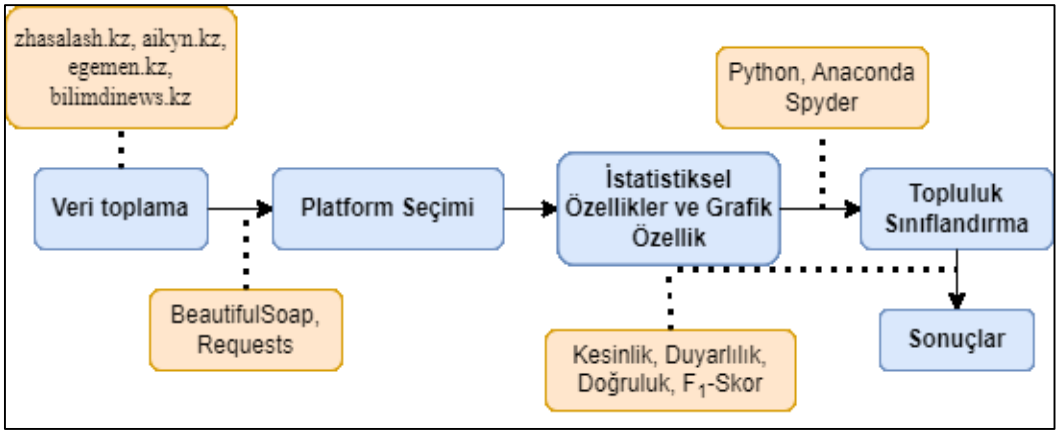
2.3.5. Denetimsiz modellerin performans sonuçlar

Çizelge 2.2. Denetimsiz Anahtar Kelime Çıkarımı Modellerinin Mimari Açıdan Karşılaştırılması

Referans	Mimari Adı	Yıl	Veri kümesi	Algoritma	Alt mimari	Çıkarım/Üretim	Dil
El-Beltagy & Rafea	KP-Miner	2010	SemEval-2	Graf-tabanlı	KP-Miner	Çıkarım	İngilizce& Arapça
Pay et al.	TO-AKÇ	2016	Inspec (DataSet I, DataSet II-validation set)	İstatistik tabanlı	TO-AKÇ, Etki alanından bağımsız	Çıkarım	İngilizce
Florescu & Caragea	Konum Sıralama	2017	KDD, WWW, NUS	Graf-tabanlı	Konum Sıralama, Konum Sıralama-fp	Çıkarım	İngilizce
Witt et al.	Metin Sıralamaş	2018	Inspec, SemEval2017, Scopus, KP20k	Graf-tabanlı	Topluluk Anahtar Kelime Çıkarımı	Çıkarım	İngilizce
Alfarra & Alfarra	TB-GTAKÇ	2018	UCST-News, Inspec	Graf-tabanlı	Tek Belgeler için	Çıkarım	İngilizce
Benani-Smires, et al.	Gömme Sıralama	2018	Inspec, NUS, DUC2001	Gömme tabanlı	Tek Belgeler için	Çıkarım	İngilizce
Alfarra et al.	GT-YZS-AKÇ	2019	UCST-News, Inspec	Graf-tabanlı	Yoğunluk zirveleri sıralama yaklaşımı	Çıkarım	İngilizce
Duari & Bhatnagar	ABF-AKÇ	2019	Inspec, Krapivin, NLM500, SemEval2010	Graf-tabanlı	Semantik Bağlantı Farkında Anahtar Kelime Çıkarma	Çıkarım	İngilizce

3. YÖNTEM VE ARAÇLAR

Bu bölümde, tez konusu Kazak dilinde haber sitelerinden anahtar kelime çıkarımı amacıyla kullanılan yöntem ve araçlar sunulmaktadır. Şekil 3.1'de tezde önerilen modelin geliştirilmesinde takip edilen sürece dair şema verilmiştir. Şemada sürece dair her bir aşama ve kullanılan yöntem araçlar görülmektedir.



Şekil 3.1. Model geliştirmede izlenen süreç

Tezde ilk olarak zhasalash.kz, aikyn.kz, egemen.kz, bilimdinews.kz haber sitelerinden BeautifulSoup ve Requests kütüphaneleri kullanılarak "veri kazıma (web scrapping)" aracılığıyla veriler toplanmıştır. Veri kazıma, web sayfasından belirli verileri almanın bir yoludur. Hedef sayfayı yazdığında kod bulunduğu sunucuya istek gönderir. Kod, tarama işinde belirtilen öğeleri çıkararak onu yükler. BeautifulSoup HTML veya XML dosyalarından veri ayıklamak için oluşturulmuş güçlü ve hızlı bir Python dilinde yazılmış kütüphanedir. Requests - web üzerindeki HTTP istekleri, kaynak kodlarını alınmasına olanak sağlayan bir modül olup, HTML, XML dosyalarında bir ayrıştırma ağacında gezinmek, aramak ve değiştirmek için basit yöntemler sağlar. BeautifulSoup karmaşık bir HTML belgesini Python nesneleri ağacına dönüştürür.

Sonrasında Platform olarak güçlü bilimsel karakteristik özelliklere sahip olması nedeniyle araştırmacılar tarafından yaygın olarak kullanılan Anaconda Spyder ortamı seçilmiştir. Anaconda Spyder makine öğrenimi için kullanabileceğimiz öğrenme modelleri ile ilgili çok çeşitli seçenekleri bizlere sunan özelliklerine sahip bir platformdur. Makine öğrenmesi ve derin öğrenme için gelişmiş kütüphanelere sahip Python dili modelin geliştirilmesinde

kullanılmak üzere seçilmiştir. Yapay zekâ kütüphaneleri olarak Torch ve Transformers kullanılmıştır. Yeni derlenen veri kümelerinin her bir içerik metni için istatistiksel ve grafiksel öznitelikler hesaplanmıştır. Hesaplanan bu öznitelikler ayrı ayrı Topluluk sınıflandırma modülünden geçirilerek dizi etiketleme görevi tamamlanmıştır.

Token Sınıflandırma modülünde Rastgele Orman (Random Forest), Aşırı Gradyan Artırma (XgBoost), Karar Ağacı (Decision Tree), Oylama Sınıflandırması (Voting Classification) topluluk sınıflandırma algoritmaları her bir veri kümesi için ayrı ayrı eğitilmiş ve test edilmiştir. Önerilen modelde istatistiksel puan ve grafiksel puanı hesaplayan iki alt modül bulunmaktadır. İlk kategoride kelimenin konumu (position of word), belge cümlelerinde kelime sıklığı (FreqSnt), kelimenin Terim Sıklığı Normalleşmesi (TFNorm) ve büyük harfler (Cast), ikinci kategoride Kelime Grafik Puanı (KGP - Word-Graph Skor - KGP) puanı hesaplanmıştır. Her bir kelime nodu kelime grafiği puanı KGP grafiksel puan ile hesaplandıktan sonra Temel Bileşenler Analizi (TBA - Principal Component Analysis - TBA) ile tek boyuta indirgenmiştir. İşlemde, Arasındalık Merkeziliği (Betweeness Centrality-BetwCnt), Yakınlık Merkeziliği (Closeness Centrality-ClosCnt), Öz vektör Merkeziliği (Eigenvector Centrality-EigCnt), Derece Merkeziliği (Degree Centrality-DegCnt), Sayfa Sıralama (PageRank), Kümeleme Katsayısı (Cluestering Coefficient-ClustCoef), Dış Merkezlik (Eccentricity-Eccent), Yapısal Delik (Structural Hole-StrucHole) değerleri kullanılmıştır. Bu 8 merkezilik, DDİ'de ve grafik temsil puanlamasında en çok kullanılanlardır. Son adımda modelin performans skorları tablo halinde sunularak değerlendirilmiştir. Oluşturulan model metinden anahtar kelime çıkarımı problemleri üzerinde denenmiş ve Kazak dili için elde edilen sonuçlar irdelenmiştir. Daha sonra modelin etkinliğini test edebilmek amacıyla Kiril alfabesi kullanan Rusça ve Latin alfabesi kullanan İngilizce dillerinde elde edilen sonuçlar değerlendirilmiştir.

3.1. Veri setleri

Kazak dili için bir külliyatın olmamasından öncelikle internetteki Kazak haber sitelerinde metin içeren veri toplanmıştır. Veri toplamak için kazak dilinde yayınlanan haber sitesinden BeautifulSoap ve Requests kütüphaneleri kullanılarak derleme gerçekleştirilmiştir. Verileri orijinal olarak Kazak dilinde yayınlanan web sayfalarından derlenmiştir. Bu web sayfalar çoğunlukla haber, tarih, edebiyat metinleri içermektedir. Çizelge 3.2'de elde bulunan veri setinde Kazak haber sitesindeki yayın başlığı, metin içeriği ve anahtar kelimeler

bulunmaktadır. Elde bulunan veri setinde Kazak haber sitesindeki yayın (ya da haberin) başlığı, metin içeriği ve anahtar kelimeler bulunmaktadır. KazakhNews veri kümesinde her bir haber için 5 tane görünür anahtar kelime yer almaktadır.

İkinci veri seti Rusça dilde yayınlanan "Коммерсантъ" günlük bir Rus sosyo-politik gazetesinden 500 veri alınarak elde edilmiştir. "Коммерсантъ" yayınevi, Rusya'daki en eski yayınevlerinden ve en yetkili medya yapılarından biridir. RussianNews olarak isimlendirilen bu veri kümesi başlık, özet ve anahtar kelimelerden oluşmaktadır. Her bir haber metni için 50 tane görünür anahtar kelime bulunmaktadır.

Üçüncüsü literatürde sıklıkla kullanılan 500N-KPCrowd (500N-KeyPhrasesCrowdAnnotated-Corpus) (Marujo, Gershman, Carbonell, Frederking ve Neto, 2013) veri kümesidir ve 400 adet haber içeriğinden oluşmaktadır. Veri kümesine ait anahtar kelimeler okuyucular tarafından seçilmiştir. Ortalama doküman başına 49.40 anahtar kelime düşmektedir. Bu anahtar kelimelerin %84.2' si görünür anahtar kelime olup geri kalanı metin içerisinde geçmeyen (absent) anahtar kelimelerdir.

Veri kümelerinin toplanmasından sonra, alanların hangilerinin kullanılacağı öznitelik belirleme aşamasında belirlendi ve çalışmaya veri temizleme ve bütünleştirme işlemleri gerçekleştirilmiştir. Temizlenen veri üzerinde çalışma yapmaya hazır hale getirilmiştir. Veri ön işleme iki farklı türde işlemi gerektirir. Bu farklı işlemlerin ilki veri kümesinin seçilmesi ve birleştirilmesi, ikincisi ise veri madenciliği için verilerin daha yararlı bir hale getirilmesi amacıyla verilerin işlenmesidir. Veri kümelerinde ön işlem (preprocessing) olarak etkisiz kelime çıkarme işlemleri gerçekleştirilmiştir. Çizelge 3.1'de veri kümelerine ait özet istatistikler, Çizelge 3.2'de ise KazakhNews veri kümesi örneği verilmiştir. Şekil 3.2'de de veri kümelerine ait metin örneği ve anahtar kelimeleri görülmektedir.

Çizelge 3.1. Veri kümesi özet istatistikleri

Veri kümesi	Dil	Alan	Başlık	#belge	#anahtar kelime
KazakhNews	Kazakça	Haber sitesi	Siyaset, edebiyet, v.s.	1000	5
RussianNews	Rusça	Haber sitesi	Sosyo-politik	500	50
500N-KPCrowd	İngilizce	Haber sitesi		400	49

Çizelge 3.2. KazakhNews veri kümesi örneği

Tegler	Başlık	Metin	Anahtar kelimeler
ПРАЙМЕРИ, Жаңалықтар, Қазақстан, Жамбыл облысы, Қордай ауданы, Мадияр Қарабаев, праймериз	БАҚ өкілі де бақ сынауда «Праймериз 2020»	«Nur Otan» партиясының праймериз науқаны қыза түсті. Жамбылда өтінім берушілердің қарасы күн санап артуда. Кеше көпбалалы ана, мүмкіндігі шектеулі азамат, өзге ұлт өкілдері құжаттарын тапсырса, бүгін әріптесіміз Мадияр Қарабаев та тәуекел етті. Мадияр Бақытбекұлы Жамбыл облысы Қордай ауданындағы «Қордай шамшырағы» газетінің тілшісі.	ең,келеді,мадияр,праймериз
ПРАЙМЕРИ, Жаңалықтар, Қазақстан, мәслихат, праймериз, Түркістан, Шымкент	Үміткер қиын-қыстау кезде қол ұшын созуға дайын – ПРАЙМЕРИЗ 2020	Шымкент қаласындағы Еңбекші ауданы бойынша алғаш болып құжат тапсырған «Ұлы Орда» жауапкершілігі шектеулі серіктестігінің директоры Қайрат Сүлейменов Түркістан облыстық мәслихатының ең жас депутаттарының бірі.	бойынша,болып,мәслихатының,праймериз, қайрат

Şekil 3.2. Veri kümeleri örnekleri a) Kazakh News, b) RussianNews, c) 500N-KPCrowd

3.2. Öznitelik Seçimi

Öznitelik çıkarımı üç adımdan oluşmaktadır. Birinci adımda istatistiksel, ikinci adımda grafiksel özellik çıkarılmakta, üçüncü adımda bu özelliklerin puanlanması gerçekleştirilmektedir. İstatistiksel özellikler ve grafik özelliği çıkarılma adımları paraleldir ve özellikleri aynı anda çıkarılmaktadır. Özelliklerin her biri, belgenin farklı görünümünü temsil etmektedir. Bu tezde, anahtar kelime çıkarımına yeni ve genel bir bakış getirebilmek

amacıyla bu özellkler birleştirilmektedir. Önerilen T-AKÇ modelinde istatistiksel ve grafiksel öznitelikleri kullanılarak Token Sınıflandırma işlemi gerçekleştirilmiştir. Çizelge 3.3'te token sınıflandırmada kullanılan öznitelikler ve hesaplanması verilmektedir. Çizelge 3.3'te görüldüğü gibi ilk kategoride kelimenin kelimenin terim sıklığı normalleşmesi, kelimenin konumu, belge cümlelerinde kelime sıklığı ve büyük harfler, ikinci kategoride kelime grafik puanı (KGP) hesaplanmıştır ve formül verilmiştir.

Çizelge 3.3. Token sınıflandırma için özniteliklerin hesaplanması

Açıklamalar	Formül
Kelimenin Terim Sıklığı Normalleşmesi (TFNorm)	lenTFs = TF[words] s avgTF = mean (validTFs) TFNorm = TF[word] / ((avgTF / lenTFs) + 1)
Kelimenin konumu (Poss)	ln (3 + mean (offsets-sentences[word]))
Belge cümlelerinde kelime sıklığı (FreqSnt)	len(offsets-sentences[word] / len(sentences)
Büyük harfler (Cast)	max (LetterTF[word], UpperTF[word]) / (1 + ln (TF[word]))
Kelime Grafik Puanı (KGP)	TBA(word-graph of input text)

3.2.1. İstatistiksel öznitelikler

İlk kategoride kelimenin Terim Sıklığı Normalleşmesi (Normalised Term Frequency TFNorm of word), kelimenin konumu (position of word (Poss)), kelimenin belge cümlelerindeki sıklığı (frequency of word in sentences of document (FreqSnt)), ve büyük harfler (Cast of word (Cast)) istatistiksel öznitelikleri hesaplanmıştır.

Terim Sıklığı Normalleşmesi değeri girdi kelimeye ait TF'in tüm kelimelere ait frekansların toplam ortalamasının kelime sayısına bölümünün bir fazlasına oranıdır. Kelimenin konumu skoru girdi kelimenin her cümle içerisindeki sırasının girdi kelimeyi içeren cümle sayısına oranını hesapladıktan sonra bu oranın 3 fazlasının doğal logaritması alınarak bulunmaktadır. Kelimenin belge cümlelerindeki sıklığı girdi kelimenin cümleler içerisinde görülme sayısının tüm cümle sayısına oranı ile hesaplanmaktadır. Son istatistiksel öznitelik büyük harfler ise kelimenin Upper veya letter olarak girdi paragrafında bulunma sayısının TFword' ün doğal logaritmasının bir fazlasına oranı ile hesaplanmaktadır.

3.2.2. Grafiksel öznitelikler

Grafikler, ilişkileri temsil etmek ve anlaşılmasını kolaylaştırmak için geliştirilmiştir. Grafikler düğümlerden ve $G = \{V, E\}$ köşelerinden oluşur ve oluşturulan ağırlıksız ve yönsüz grafiğin düğümleri kelimelerdir ve köşeler kelimelerin 3 gramıdır (trigram). Grafik oluşturulduktan sonra, her bir düğümü puanlamak için merkezilik ölçüsünü kullanarak tek boyuta indirgenmektedir.

İkinci kategoride KGP puanı hesaplanmıştır. Her bir kelime nodu kelime grafiği puanı grafiksel puan ile hesaplandıktan sonra TBA ile tek boyuta indirgemiştir. Her kelime düğümü için maksimum varyasyonun yönünü yakalamak için, KGP, Tekil Değer Kompozisyonu (TDK-Singular Value Decomposition-SVD), BKA ve LDA (Chandrashekar ve Sahin, 2014) gibi diğer doğrusal özellik dönüşümleri yerine 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, dışmerkezlik ve yapısal boşluk puanları sözcük birlikte görülme grafiği puanları olarak kullanılmıştır. İşlemde Arasındalık Merkeziliği Arasındalık Merkeziliği (Betweeness Centrality-BetwCnt), Yakınlık Merkeziliği (Closeness Centrality-ClosCnt), Öz vektör Merkeziliği (Eigenvector Centrality-EigCnt), Derece Merkeziliği (Degree Centrality-DegCnt), Sayfa Sıralama (PageRank), Kümeleme Katsayısı (Cluestering Coefficient-ClustCoef), Dış Merkezlik (Eccentricity-Eccent), Yapısal Delik (Structural Hole-StrucHole) değerleri kullanılmıştır. Merkeziyetlerine göre her düğüme birkaç puan atanır, böylece grafiğin her düğümüne yedi puan atanır. Ayrıca, bu adım her kelime için tek bir puan döndürmelidir; bu amaçla TBA dağıtılır. Bir TBA tekniği, yorumlanabilirliği artırırken ve bilgi kaybını en aza indirirken bir veri kümesinin boyutsallığını azaltır.

Büyük veri kümeleri giderek daha yaygın hale geliyor. Bu tür veri setlerinin boyuÖSAllığını azaltmak, yorumlanabilirliği artırmak ama aynı zamanda bilgi kaybını en aza indirmek için TBA kullanılmaktadır.

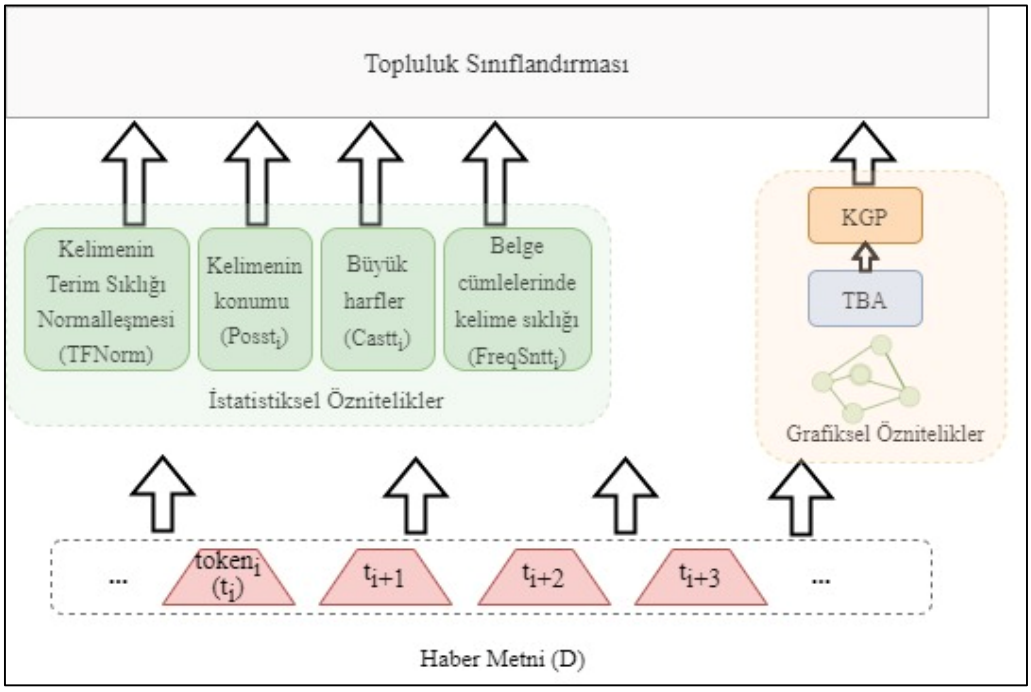
Kelime Grafik Puanı Grafik özniteliği (Zehtab-Salmasi ve diğerleri, 2021) olarak ikinci kategori altında hesaplanmıştır. Her kelime düğümü için maksimum varyasyonun yönünü yakalamak için, TDK, BKA ve LDA (Chandrashekar ve Sahin, 2014) gibi diğer doğrusal özellik dönüşümleri yerine TBA ile tek boyuta indirgenmiştir. Tek boyuta indirgendiğinde

özvektör, arasındalık, yakınlık, sayfa sıralama, kısıtlama, kümeleme, dışmerkezlik ve yapısal boşluk puanları sözcük birlikte görülme grafiği puanları olarak kullanılmıştır.

3.3. Topluluk Anahtar Kelime Çıkarma Modeli

Bu tez çalışmasında günlük bilgi üreten bir Kazak haber sitesinden elde edilen veri kümesi ile anahtar kelime çıkarımı için KazakhNews veri kümesi derlenmiş ve grafik ve metinsel öznitelikler kullanılarak yeni bir model Topluluk Anahtar Kelime Çıkarma (T-AKÇ) modeli önerilmiştir. Latin alfabesinden farklı alfabe kullanan diğer diller için de yüksek başarım elde eden modelimiz Rusça haber içeriklerinden yeni derlenen RussianNews veri kümesi için de eğitilmiş ve test edilmiştir. Ayrıca literatürde sık kullanılan İngilizce dilinde haber içeriklerinden oluşan 500N-KPCrowd veri kümesi için model eğitilmiş ve test edilmiş sonuçlar tablo halinde sunulmuştur. Bu kapsamda Kazakça dil yapısı hakkında bilgi verilmiştir ve örnekleri incelenmiştir. Anahtar Kelime çıkarma konusunda diğer dillerde yapılmış çalışmalardan bahsedilmiştir ve anahtar kelime çıkarma uygulama alanları ile ilgili bilgi verilmiştir.

Bu tez çalışmasında öncelikle önerilen T-AKÇ modeli için izlenen işlem adımları anlatılmıştır. Şekil 3.3'te de önerilen model için izlenen adımlar görülmektedir.



Şekil 3.3. Topluluk anahtar kelime çıkarımı (T-AKÇ) modeli

Bu çalışmada anahtar kelime çıkarımı bir dizi etiketleme problemi olarak ele alınmıştır. Veri kümelerine ait her bir Haber metni (D) içeriği n adet t_i tokenından oluşmaktadır:

$$D = [t_i,\ t_{i+1}, \dots, t_n]\,, \qquad i = 1,..,n \tag{3.1}$$

Şekil 3.3'te sunulan modelde her bir token t_i' ye ait istatistiksel puan ve grafiksel puanı hesaplayan iki alt modül bulunmaktadır. Bu modüllerde her iki kategoriye ait öznitelikler ayrı ayrı hesaplanmıştır. İlk kategoride kelimenin konumu (position of word (Poss), belge cümlelerindeki kelimelerin sıklığı (frequency of word in sentences of document (FreqSnt), kelimenin terim sıklığı normalleşmesi (TFNorm of word) ve büyük harf (Cast of word) istatistiksel öznitelikleri hesaplanmıştır.

İkinci kategoride KGP puanı hesaplanmıştır. Her bir kelime nodu kelime grafiği puanı grafiksel puan ile hesaplandıktan sonra TBA ile tek boyuta indirgenmiştir. İşlemde Arasındalık Merkeziliği (Betweeness Centrality-BetwCnt), Yakınlık Merkeziliği (Closeness Centrality-ClosCnt), Öz vektör Merkeziliği (Eigenvector Centrality-EigCnt), Derece Merkeziliği (Degree Centrality-DegCnt), Sayfa Sıralama (PageRank), Kümeleme Katsayısı (Cluestering Coefficient-ClustCoef), Dış Merkezlik (Eccentricity-Eccent), Yapısal Delik (Structural Hole-StrucHole) değerleri kullanılmıştır.

Haber metnini temsil eden her bir token t_i için hesaplanan istatiksel öznitelik değerleri dizi olarak

$$St_{t_i} = [\,TFNorm_{t_i}, Poss_{t_i},\ FreqSnt_{t_i},\ Cast_{t_i}] \tag{3.2}$$

dizisinde tutulmaktadır. Grafiksel öznitelikler ise

$$WGS_{t_i} = PCA(\,Betweeness_{t_i}, Closeness_{t_i},\ PageRank_{t_i}, Constraint_{t_i}, Clustering_{t_i}, Eccentricity_{t_i}, StructuralHole_{t_i}) \tag{3.3}$$

değişkeninde tutulmaktadır.

NT dokümanında bulunan tüm t_i tokenleri için St ve KGP değerleri ayrı ayrı,

$$X\_St = [\mathrm{st}_{t_i}, st_{t_{i+1}}, \dots, st_{t_n}], \qquad i = 1, .., n \tag{3.4}$$

$$X\_WGS = [\mathrm{KGP}_{t_i}, wgs_{t_{i+1}}, \dots, wgs_{t_n}], \qquad i = 1, .., n \tag{3.5}$$

olarak ifade edilir. Her iki kategori için X dizileri Token Sınıflandırma modülünden geçirilerek Y,

$$Y = [y_{t_i}, y_{t_{i+1}}, \dots, y_{t_n}], \qquad i = 1, .., n \tag{3.6}$$

dizisine göre eğitim ve test işlemi gerçekleştirilmektedir. Y dizisi her bir tokenı anahtar kelime (1) veya değil (0) olarak etiketlemektedir.

Önerilen modelin performans sonuçlarını ölçmek için, Token Sınıflandırma Modülünde Rastgele Orman (Random Forest), Aşırı Gradyan Artırma (XgBoost), Karar Ağacı (Decision Tree), Oylama Sınıflandırması (Voting Classification) algoritmaları ile test edilmiştir.

3.3.1. Rastgele orman algoritması

Rastgele orman algoritması (Random Forest) karar ağaçlarının çoklu hali olan yönetimsel bir makine öğrenmesi yöntemidir. Her bir karar ağacı, eğitim setinde verilen girdiye göre bütün rastgele düğümleri gezerek ve bu düğümlerdeki koşullara bağlı olarak ağacın en altındaki yaprağa ulaşır. Ağacın son düğümü olan yaprakta hedef olan cevap değişkeni bulunur. Rastgele orman algoritmasında, test setindeki hatayı objektif olarak hesaplamak için ekstra bir çapraz geçerleme işlemine gerek yoktur. Algoritma çalışma esnasında her ağacı, orijinal veriden farklı örnekler seçerek oluşturur. Oluşturulacak olan ağaçların genelde üçte birinde, rastgele seçilen örnek ağacın oluşturulma aşamasında dışarıda bırakılır. Bu sayede, ağaçlar oluşturulurken otomatik olarak test edilmiş olur. Bu rastgele örnekleme yöntemi, çeşitliliği artırarak ağaç içindeki düğümlere ait varyansı azaltır. Bu süreç aynı zamanda "özellik torbalama" (feature bagging) olarak da bilinmektedir. Bu çalışma kapsamında, karar ağaçlarındaki düğümlerde kelime ve kelime grupları, yaprak düğümünde ise İnternet sayfasının sınıfı ile her bir ağaç eğitilmiştir. Bu eğitim sırasında rastgele orman algoritmasının parametreleri olan ağaç derinliği, ağaç sayısı, gini-entropy gibi ölçüt

değişkenlerini ızgara araması (grid search) ile optimize edilmiştir. Çok sınıflı sınıflandırma probleminin çözümünde kullanılmıştır.

Rasatgele Orman algoritması için girdi olarak kullanılan öznitelikler istatistiksel öznitelikler ve kelime grafiği özniteliği altında gruplanmıştır. Birinci kategori olan istatistiksel öznitelikler kategorisi altında kelimenin kelimenin konumu belge cümlelerinde kelime sıklığ, kelimenin Terim Sıklığı Normalleşmesi ve büyük harfler ve KP puanları hesaplanmıştır. İkinci kategori girdi metni kullanılarak çizilen kelime grafiği puanıdır. Kelime grafiği puanı her bir kelime nodu için 8 merkeziliği değerlerinin TBA kullanılarak tek bir boyuta indirgenmesiyle üretilmiştir. Bu kategori altında tek öznitelik bulunmaktadır.

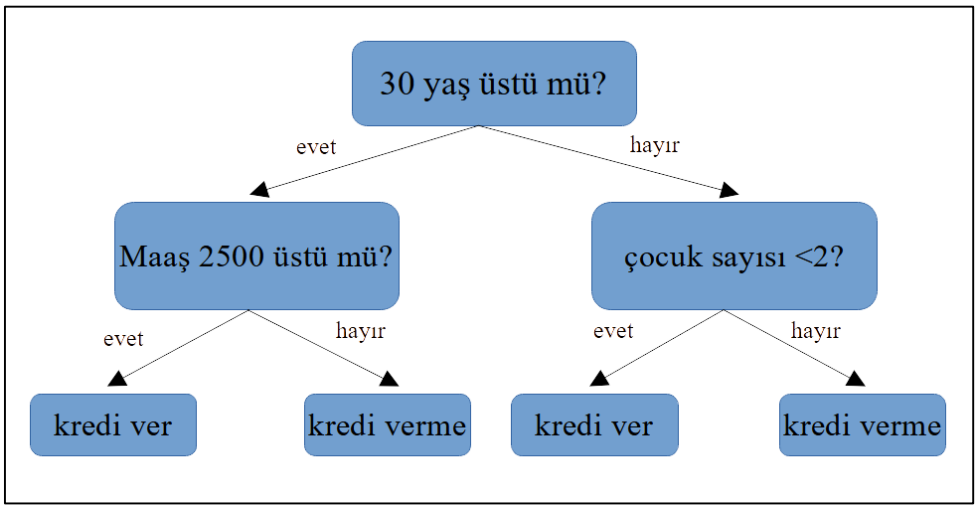
3.3.2. Aşırı gradyan artırma algoritması

Aşırı Gradyan Artırma algoritması (Extreme Gradient Boosting - XgBoost) kolektif öğrenme yöntemleri içindeki en güçlü modellerden biridir. Aşırı Gradyan Artırma ilk olarak 2011 yılında Tianqi Chen ve Carlos Guestrin tarafından önerilen modeldir (Chen ve Guestrin, 2016). Model, Artırılmış Ağaç modellerine dayalı bir öğrenme çerçevesidir. Geleneksel Artırılmış Ağaç modelleri yalnızca birinci türev bilgisini kullanır. Aşırı Gradyan Artırma algoritması diğer kolektif öğrenme modellerinden daha hızlıdır ve parallel computation prensibiyle çalışır. Onun bu özelliğide onu büyük veri setlerindeki en kullanışlı model yapıyor. Aşırı Gradyan Artırma algoritması max_depth, min_child_weight, learning_rate, colsample_bytree v.s. gibi hiperparametreleri kullanıyor. Bunların dışında düzelleştirme konusunda modeli geliştiren gamma, alpha ve lambda hiperparametreleri de vardır. Derin öğrenme algoritmalarıyla karşılaştırıldığında, Aşırı Gradyan Artırma'nın Merkezi İşlem Birimi (MİB) üzerinde çalışan küçük veri kümeleri için kullanımının daha kolay olduğu kabul edilmektedir.

3.3.3. Karar ağacı algoritması

Karar ağaçları (Decision Tree) hem sınıflandırma hem de regresyon problemleri için kullanılabilen denetimli bir makine öğrenme algoritmasıdır. Orijinal veri kümesini, verilerle yalnızca bir etiket ilişkilendirene kadar daha küçük ve daha küçük veri kümelerine böler. Karar ağaçlarını kullanmanın başlıca avantajları, yorumlanmalarının ne kadar kolay ve ne kadar hızlı eğitilmeleridir, ancak verilerdeki küçük bir değişiklik, ağacın yapısının

değişmesine ve dolayısıyla bazı tahminlerin de değişmesine neden olabilir. Şekil 3.4'te bir kökü ve birden fazla başka düğüm türü (ebeveyn, çocuk ve yaprak) olan Ağaç Veri Yapısına benzer.



Şekil 3.4. Basit bir karar ağacı örneği

3.3.4. Oylama sınıflandırması

Oylama Sınıflandırması (Voting Classification) algoritması kolektif öğrenme metotlarından biridir. Kolektif öğrenme metodu birden fazla öğrenme modelinin (bu durumda base model ya da weak model diye adlandırılır) beraber çalıştığında daha iyi bir doğruluk puanı elde etmesini konu alan makine öğrenmesi tekniğidir. Yalnızca doğruluk puanlarının birbirine yakın olması durumunda değil, puanı iyileştirmek istediğiniz her durumda başvurabilecek farklı kolektif öğrenme teknikleri bulunur. Oylama Sınıflandırması ise yukarıdaki örnekte belirtildiği gibi farklı algoritmaların doğruluk puanlarının birbirine yakın olduğu durumlarda tercih edilir. KazakhNews veri kümesi için Oylama Sınıflandırması modeli 0,974 F-Skoru ile en yüksek sonucu vermiştir.

4. BULGULAR VE DEĞERLENDİRME

Çalışma ortamı olarak deneysel çalışmalar Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz işlemcili, 8 GB RAM kapasiteli 2 çekirdekli bir bilgisayarda gerçekleştirilmiştir. Programlama dili olarak Python 3.7 versiyonu kullanılmıştır. Uygulama Anaconda Spyder ortamında gerçekleştirilmiştir.

Modelin başarımını doğru bir şekilde değerlendirmek için, doğruluk değerinin yanı sıra, F_1 skoru olarak tabir edilen değer de takip edilmiştir. Oluşturulan modellerin başarım derecelerini belirleyen doğruluk, kesinlik, duyarlılık ve F_1-skor gibi kriteler kullanılarak kullanılan algoritmaların başarıları değerlendirilir. F_1 skorunun hesaplanmasında GerçekPozitif (GP), Yanlış Pozitif (YP), Gerçek Negatif (GN) ve Yanlış Negatif (YN) değerleri kullanılmaktadır. GP, modelin tahmini ve gerçek değerlerin her ikisinde de olumlu sonuç vermesi; GN, modelin tahmini ve gerçek değerlerin her ikisinde de olumsuz sonuç vermesi; GP, model tahmini olumlu iken gerçek değerin olumsuz sonuç vermesi; YN, modelin tahmini olumsuz iken gerçek değerin olumlu sonuç vermesi şeklinde açıklanmaktadır. Bu durumda GP ve GN doğru sonuç, YP ve YN ise yanlış sonuç kabul edilmektedir.

Doğruluk değeri, modelin doğru tahmin ettiği GP ve GN değerlerinin, tahmin edilen tüm GP, GN, YP, YN değerlerine oranı ile hesaplanmaktadır.

$$Doğruluk = \frac{GP+GN}{\text{GP+ GN+ YP+ YN}} \tag{4.1}$$

Kesinlik (precision) değeri, modelin tahmin ettiği GP değerlerin sayısının, modelin ürettiği tüm olumlu sonuçlar olan GP ve YP değerlerinin sayısına oranıdır.

$$Kesinlik = \frac{GP}{GP+YP} \tag{4.2}$$

Duyarlılık değeri ise modelin tahmin ettiği GP değerlerinin sayısının, modelin üretmesi gereken tüm olumlu sonuçlar olan GP ve YN sayılarına oranı ile bulunabilir:

$$Duyarlılık = \frac{GP}{GP+YN} \tag{4.3}$$

F1 skoru ise kesinlik ve duyarlılık değerlerinin harmonik ortalaması olarak tanımlanabilir:

$$F_1 - skor = 2 \times \frac{Duyarlılık \times Kesinlik}{Duyarlılık + Kesinlik} \quad (4.4)$$

Literatürden farklı olarak her bir kelimenin öznitelikleri farklı sınıflandırma algoritmaları ile eğitilerek ağırlıklandırılmıştır. Modelde topluluk sınıflandırması modülü için Rastgele Orman, Aşırı Gradyan Artırma ve Oylama Sınıflandırması algoritmaları ayrı ayrı eğitilmiş ve test edilmiştir.

Topluluk sınıflandırması yöntemleri aslında sınıflandırma başarımını arttırmak amacıyla birkaç karar ağacını birleştiren yöntemlerdir. Burada ana ilke, bir grup zayıf öğrenicinin güçlü bir öğrenici oluşturmak için bir araya gelmesidir. Tezde, Karar Ağacı algoritması topluluk sınıflandırma algoritmalarının önerilen yaklaşımdaki etkinliklerini kıyaslayabilmek amacıyla ayrıca eğitilmiş ve test edilmiştir.

Modeli test etmek için kullanılmak üzere iki yeni veri kümesi KazakhNews ve RussianNews toplanarak araştırmacılara açık olarak sunulmuştur. Bu iki veri kümesine ilave olarak yine haber metinlerinden oluşan ve literatürde sıklıkla kullanılan 500N-KPCrowd veri kümesi için modelin performans sonuçları çıkarılmıştır. Model tüm veri kümelerinde Rastgele Orman ve Oylama Sınıflandırma algortimaları için yüksek sonuçlar elde etmiştir.

4.1. Topluluk AKÇ Modeli Deney Sonuçları

Bu tez çalışmasında önerilen T-AKÇ modeli Kazak haber web sayfalarndan ve Rusça haber sayfasından toplanan veri setleri ile eğitilmiş ve test edilmiştir. Ayrıca literatürde yaygın olarak kullanılan 500N-KPCrowd verikümesi ile karşılaştırılmıştır.

Token Sınıflandrma modülünde Rastgele Orman, Aşırı Gradyan Artırma, Oylama Sınıflandırması topluluk sınıflandırma algoritmaları ve Karar Ağacı algotritması her bir veri kümesi için ayrı ayrı eğitilmiş ve test edilmiştir.

Önerilen T-AKÇ modeli yeni derlenen KazakhNews ve RussianNews veri kümeleri ile 500N-KPCrowd veri kümesi üzerinde eğitilmiş ve test edilmiştir. Veri kümesinin bir kısmı ile modeli eğitilir, diğer bir kısmı ile modelin başarısını değerlendirilir. Sık kullanılan bir yaklaşımla %77'i eğitim için %33'i test için ayrılmıştır. Ancak burada veri parçalanırken verinin dağılımına bağlı olarak modelin eğitim ve testinde bazı sapmalar (bias) ve hatalar oluşabilmektedir. Bu sebeple k-çapraz doğrulama, veriyi belirlenen bir k sayısına göre eşit parçalara bölmektedir. Bu yöntemle her bir parçanın hem eğitim hem de test için kullanılması sağlanarak dağılım ve parçalanmadan kaynaklanan sapma ve hatalar en aza indirgenmektedir.

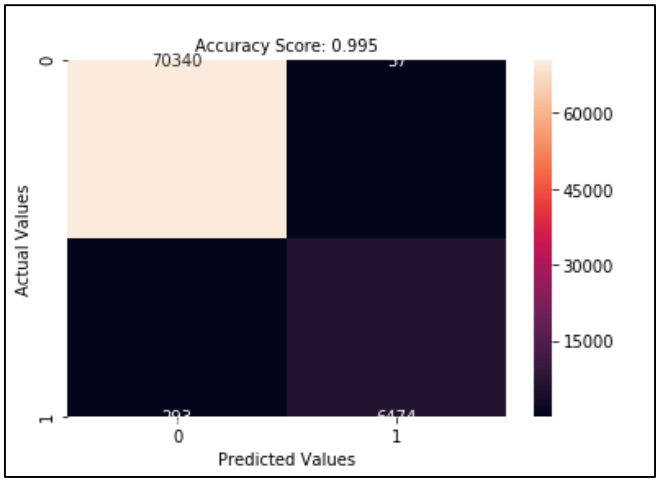
Deneyler, çapraz doğrulama yöntemi kullanılarak 20 kez çalıştırılmıştır. K-katlı çapraz doğrulama (k-fold cross-validation) olarak bilinen rastgele veri alt kümelerini kullanan çapraz doğrulama, sınıflandırma için kullanılan modellerin başarı oranını test etmenin güçlü bir yoludur. Bununla birlikte, bilinen istatistiksel özelliklere sahip verilerle test edilen modellerde k (number of subsets) değerlerinin doğrulama sonuçlarını nasıl etkilediğini araştıran çok az çalışma vardır. Marcot ve Hanea (2021) çalışmalarında istatistiksel özelliklere sahip veriler için k-katlı çapraz doğrulamada optimal k değerinin 10 olması önerilmiştir. Ancak anahtar kelime çıkarmada kullanılan veri kümeleri bilinen istatistiksel özelliklere sahip değildir.

Çizelge 4.1'de KazakhNews veri kümesi için önerilen modelin performans sonuçları bulunmaktadır. Tablo incelendiğinde istatistiksel özellikler ve grafiksel özelliiğin kombinasyonu ile Kazak veri kümesi için Aşırı Gradyan Artırma 0,88 ve Karar Ağacı 0,96 F_1-skor verirken Rastgele Orman ve Oylama Sınıflandırma modelleri ile 0,97 F_1-skor ile en iyi sonuçlara sahiptir.

Çizelge 4.1. KazakhNews veri kümesi için modellerin performans sonuçlarının karşılaştırılması

	Topluluk Sınıflandırma	*Metrikler*	*İstatistiksel Öznitelikler*	*Grafiksel Öznitelik*	*İstatistiksel Öznitelikler + Grafiksel Öznitelik*
T-AKÇ	Rastgele Orman	Doğruluk	0,994	0,987	0,995
		Kesinlik	0,978	0,904	0,988
		Duyarlılık	0,956	0,956	0,958
		F_1-skor	0,967	0,929	0,973
	Aşırı Gradyan Artırma	Doğruluk	0,978	0,958	0,981
		Kesinlik	0,919	0,838	0,936
		Duyarlılık	0,821	0,641	0,847
		F_1-skor	0,867	0,726	0,889
	Oylama Sınıflandırma	Doğruluk	0,994	0,987	0,995
		Kesinlik	0,980	0,905	0,988
		Duyarlılık	0,955	0,957	0,960
		F_1-skor	0,967	0,930	0,974
Karar Ağacı		Doğruluk	0,992	0,987	0,993
		Kesinlik	0,959	0,904	0,960
		Duyarlılık	0,957	0,957	0,961
		F_1-skor	0,958	0,930	0,961

Anahtar kelime çıkarımı algoritmalarının değerlendirilmesinde F_1-skoru ölçütü kullanılmaktadır. Bu skorun hesaplanmasında tahmin edilen değerlerin gerçek değeri/tahmin edilen değeri sayılarına bakılarak oluşturulan karmaşıklık matrisi kullanılmaktadır. Şekil 4.1'de karmaşıklık matrisi bulunmaktadır.



Şekil 4.1. Kazak veri kümesi için Rastgele Orman modelin karmaşıklık matrisi

Çizelge 4.2. Kazak verİ kümesi için karmaşıklık matrisi

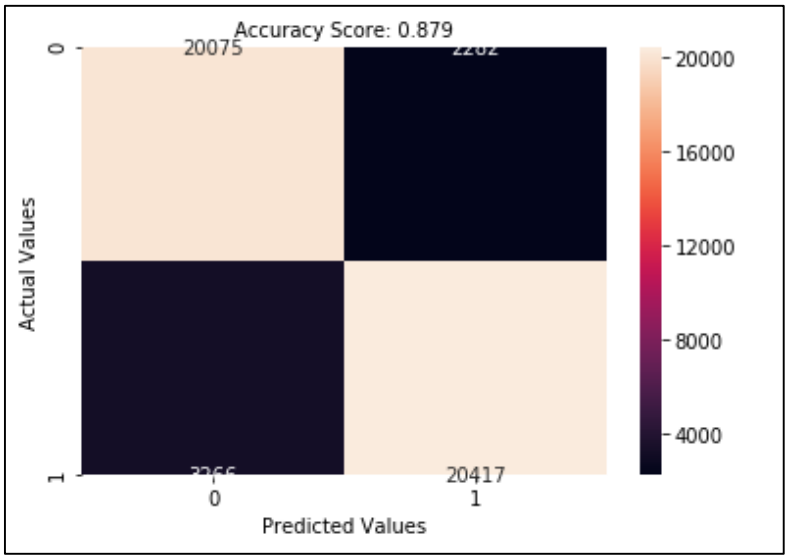
	Pozitif	*Negatif*
Pozitif	GN= 70340	YP=57
Negatif	YN= 293	GP=6474

Çizelge 4.2'de Karmaşıklık matrisi bulunmaktadır. Matris gerçekte anahtar kelime olan ve anahtar kelime olarak tahmin edilen GP, gerçekte anahtar kelime olmayan fakat anahtar kelime olarak tahmin edilen YP, gerçekte anahtar kelime olmayan fakat anahtar kelime olarak işaretlenen YN ve son olarak gerçekte anahtar kelime olmayıp anahtar kelime değil olarak işaretlenen GN değerlerini bulundurmaktadır.

RusçaNews veri kümesi için Çizelge 4.3'te görüldüğü gibi Karar Ağacı ile 0,85 sonucu verirken Rastgele Orman ile 0,87 Aşırı Gradyan Artırma ile 0,87 ve Oylama Sınıflandırma modeli ile 0,88 en yüksek F_1-skor elde edilmiştir.

Çizelge 4.3. RussianNews veri kümesi için modellerin performans sonuçlarının karşılaştırılması

	Topluluk Sınıflandırma	*Metrikler*	*İstatistiksel Öznitelikler*	*Grafiksel Öznitelik*	*İstatistiksel Öznitelikler + Grafiksel Öznitelik*
T-AKÇ	Rastgele Orman	Doğruluk	0,860	0,738	0,876
		Kesinlik	0,876	0,730	0,895
		Duyarlılık	0,847	0,779	0,861
		F_1-skor	0,861	0,754	0,877
	Aşırı Gradyan Artırma	Doğruluk	0,870	0,706	0,877
		Kesinlik	0,887	0,772	0,891
		Duyarlılık	0,857	0,610	0,868
		F_1-skor	0,872	0,681	0,879
	Oylama Sınıflandırma	Doğruluk	0,869	0,736	0,880
		Kesinlik	0,887	0,730	0,898
		Duyarlılık	0,854	0,775	0,866
		F_1-skor	0,870	0,752	0,881
Karar Ağacı		Doğruluk	0,855	0,737	0,845
		Kesinlik	0,861	0,730	0,848
		Duyarlılık	0,856	0,777	0,853
		F_1-skor	0,859	0,753	0,850



Şekil 4.2 Rusça veri kümesi için Oylama Sınıflandırma modelin karmaşıklık matrisi

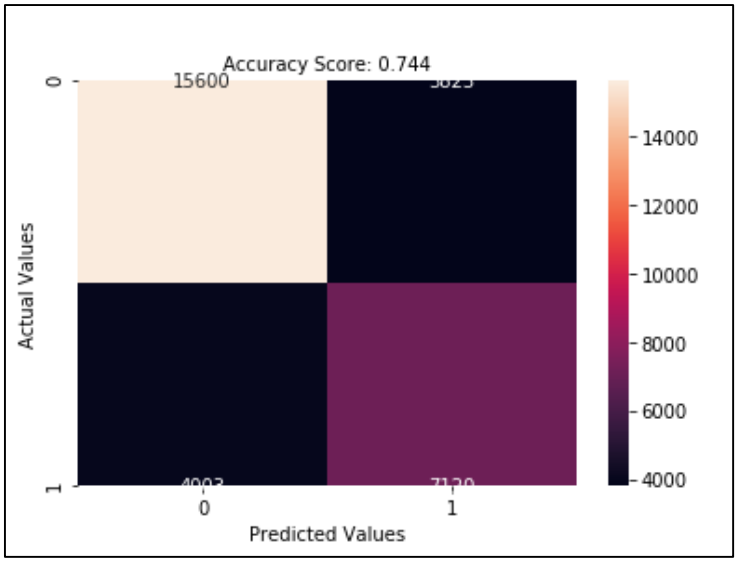
Çizelge 4.4 Rusça veri kümesi için karmaşıklık matrisi

	Pozitif	*Negatif*
Pozitif	GN= 20107	YP=2304
Negatif	YN= 3147	GP=20482

Çizelge 4.5'de 500N-KPCrowd veri kümesi için modelim performans sonuçları bulunmaktadır. Bu veri kümesi için Aşırı Gradyan Artırma 0,57 ve Karar Ağacı 0,67 sonucu verirken Oylama Sınıflandırma ve Rastgele Orman 0,70 en yüksek F_1-skor elde edilmiştir. Tüm veri kümeleri için Aşırı Gradyan Artırma algoritması dışındaki tüm topluluk sınıflandırma modelleri için her iki öznitelik grubunun birlikte kullanılması başarımı arttırmıştır.

Çizelge 4.5. 500N-KPCrowd veri kümesi için modellerin performans sonuçlarının karşılaştırılması

	Topluluk Sınıflandırma	*Metrikler*	*İstatistiksel Öznitelikler*	*Grafiksel Öznitelik*	*İstatistiksel Öznitelikler + Grafiksel Öznitelik*
T-AKÇ	Rastgele Orman	Doğruluk	0,791	0,738	0,803
		Kesinlik	0,725	0,643	0,779
		Duyarlılık	0,688	0,638	0,643
		F_1-skor	0,706	0,641	0,705
	Aşırı Gradyan Artırma	Doğruluk	0,746	0,706	0,750
		Kesinlik	0,754	0,714	0,763
		Duyarlılık	0,451	0,322	0,456
		F_1-skor	0,565	0,444	0,571
	Oylama Sınıflandırma	Doğruluk	0,796	0,740	0,807
		Kesinlik	0,748	0,648	0,801
		Duyarlılık	0,665	0,634	0,628
		F_1-skor	0,704	0,641	0,704
Karar Ağacı		Doğruluk	0,774	0,741	0,767
		Kesinlik	0,693	0,650	0,683
		Duyarlılık	0,686	0,633	0,676
		F_1-skor	0,689	0,641	0,679



Şekil 4.3. 500N-KPCrowd veri kümesi için Oylama Sınıflandırma modelin karmaşıklık matrisi

Çizelge 4.6. 500N-KPCrowd ver kümesi için karmaşıklık matrisi

	Pozitif	*Negatif*
Pozitif	GN= 15487	YP=3907
Negatif	YN= 4043	GP=7109

Çizelge 4.6'da Karmaşıklık matrisi bulunmaktadır. Çizelge 4.7'de 500N-KPCrowd veri kümesinin litertürdeki diğer çalışmalarla karşılaştırıldığında daha yüksek sonuç elde edilmiştir.

Çizelge 4.7. 500N-KPCrowd veri kümesinin karşılaştırılması

Model	500N-KPCrowd veri kümesi
CN-XGB	0,538
RaKUn	0,428
BiLSTM	0,29
UKE@15[1]	0,17
HybrydKEM-RF	0,69
T-AKÇ	0,705

5. SONUÇ VE ÖNERİLER

Bu tez çalışmasında Kazak haber metinlerinden anahtar kelime çıkarımı için yeni Topluluk Anahtar Kelime Çıkarımı (T-AKÇ) modeli önerilmiştir. Önerilen yöntemde anahtar kelime çıkarma problemi bir dizi etiketleme problemi olarak ele alınmış, metnin istatistiksel ve grafiksel öznitelikleri Topluluk Token Sınıflandırma modülünde işlenerek anahtar kelime çıkarımı gerçekleştirilmiştir. Çalışmada metnin iststistiksel ve grafksel özellikleri hem ayrı ayrı hem de birbirinin kombinasyonu şeklinde test edilmiştir. Modelin eğitim ve testinde kullanmak ve modelin farklı dillerdeki başarımını kıyaslamak amacıyla Kiril alfabesini kullanan iki yeni veri kümesi KazakNews ve RusNews oluşturulmuştur. Bu iki veri kümesine ek olarak literatürde yaygın olarak kullanılan ve haber metinlerini içeren Latin 500N-KPCrowd veri kümesi için modelin performans sonuçları elde edilmiştir. Modelin topluluk sınıflandırma modülü içerisinde Rastgele Orman (Random Forest), Aşırı Gradyan Artırma (XgBoost), Oylama Sınıflandırması (Voting Classification) sınıflandırma algoritmaları ve Karar Ağacı (Decision Tree) algoritması kullanılmıştır. Model her bir dil için farklı ML algorirmaları ve farklı veri setleri ile ayrı ayrı eğitilmiştir. Çalışmada, Rastgele Orman ve Oylama Sınıflandırma algoritmaları tüm veri setleri için birbirine oldukça yakın başarıma sahip olduğu görülmüştür. KazakhNews veri kümesi için en yüksek sonuç (0,97 F-skor) Rastgele Orman ve Oylama Sınıflandırma ile istatistiksel ve grafiksel özelliklerin birlikte kullanımı ile elde edilmiştir. Kiril alfabesini kullanan başka bir dil Rusça için topladağımız veri kümesi ile Oylama Sınıflandırma algoritmasında 0,88 F-skor ile literatürdeki sonuçlara göre en yüksek başarım elde edilmiştir. Latin harfini kullanan İngilizce veri kümesi olan 500N-KPCrowd ile yine en yüksek başarı Rastgele Orman ve Oylama Sınıflandırma algoritmları ile 0,70 F-skor sonuç alınmıştır. Önerilen yöntem, literatürdeki mevcut yöntemlerle karşılaştırıldığında, KazakNews veri kümesinde değerlendirme ölçütleri açısından çok yüksek performans göstermiştir.

KAYNAKLAR

Abibullayeva, A., Çetin, A. (2022). Keyword Extraction from Kazakh News Dataset with BERT. *El-Cezeri,* 9(4), 1193-1200.

Alfarra, M. R., 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.

Alfarra, M., Alfarra, A. M., and Salahedden, A. (2019, March). Graph-based Density Peaks Ranking Approach for Extracting KeyPhrases (GDREK). *2019 IEEE 7th Palestinian International Conference on Electrical and Computer Engineering (PICECE)*, Gaza, Palestine, 1-6.

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.

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

Awajan, A. A. (2014, October). Unsupervised Approach for Automatic Keyword Extraction from Arabic Documents. *Proceedings of the 26th Conference on Computational Linguistics and Speech Processing (ROCLING 2014)*,175-184.

Azarafza, M., Feizi-Derakhshi, M. R., and Shendi, M. B. (2020, March). Metin Sınıflandırma-based microblogs keyword extraction method for Persian language. *3rd International Congress on Science and Engineering*, Hamburg, Germany, 1-13.

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

Bekbulatov, E., Kartbayev, A. (2014, October). A study of certain morphological structures of Kazakh and their impact on the machine translation quality. *2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT)*, Astana, Kazakhstan, 1-5.

Beliga, S. (2014). Keyword extraction: a review of methods and approaches. *University of Rijeka, Department of Informatics, Rijeka,* 1(9), 1-9.

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

Birdevrim, A. S., Boyacı, A., and S Al Thani, D. A. (2018). İyileştirilmiş otomatik anahtar kelime çıkarımı BHO-AKÇ. *İstanbul Ticaret Üniversitesi Teknoloji ve Uygulamalı Bilimler Dergisi*, 1(1), 11-19.

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),* Nagoya, Japan, 543-551.

Breiman, L. (2001). Random forests. *Machine Learning,* 45(1), 5–32

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

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

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.*

Chen, W., Gao, Y., Zhang, J., King, I., and Lyu, M. R. (2019, July). -Guided Encoding for Keyphrase Generation. *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, The Chinese University of Hong Kong, 6268-6275.

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.*

Domingos, P., Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29,* 103-130.

Duari, S., Bhatnagar, V. (2019). sCAKE: semantic connectivity aware keyword extraction. *Information Sciences*, 477, 100-117.

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

El-Shishtawy, T., Al-Sammak, A. (2012). Arabic keyphrase extraction using linguistic knowledge and machine learning techniques. *arXiv preprint arXiv:1203.4605.*

Florescu, C., Caragea, C. (2017, February). A position-biased pagerank algorithm for keyphrase extraction. *Proceedings of the AAAI conference on artificial intelligence*. Vancouver, Association for Computational Linguistics. Canada, 31 (1), 4923-4924.

Florescu, C., & Caragea, C. (2017, July). Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. *Vol.(1): Long Papers,* 1105-1115.

Garg, N., Favre, B., Reidhammer, K., and Hakkani Tür, D. (2009). *Clusterrank: a graph based method for meeting summarization*. Idiap Research Institute. Rue Marconi, 1-6.

Gero, Z., Ho, J. C. (2019, September). Namedkeys: Unsupervised keyphrase extraction for biomedical documents. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 328-337.

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

Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation,* 9(8), 1735–1780 .

Hong, B., Zhen, D. (2012). An extended keyword extraction method. *Physics Procedia,* 24, 1120–1127 .

Huang, C., Tian, Y., Zhou, Z., Ling, C. X., AND huang, T. (2006, December). Keyphrase extraction using semantic networks structure analysis. *Sixth International Conference on Data Mining (ICDM'06)*, Hong Kong, China, 275-284.

İnternet: CLMCRK. (2009). The corpus of kazakh language. Web: http://web-corpora.net/KazakhCorpus/search / Erişim tarihi 29/08/2012

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*. New York: John Wiley & Sons. 37-42.

Kaur, J., Gupta, V. (2010). Effective approaches for extraction of keywords. *International Journal of Computer Science Issues (IJCSI)*, *7*(6), 144-148.

Kessikbayeva, G., Cicekli, I. (2014, June). Rule based morphological analyzer of Kazakh language. *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM*, Baltimore, Maryland, 46-54.

Khaustov, S.V., Gorlova, N.E., Kalmykov, A.V., and Kabaev, A.S. (2021). Bert for russian news clustering. 1, 1-6.https://doi.org/10.28995/2075-7182-2021-20-385-390

Kim Y. (2014). "Convolutional Neural Networks for Sentence Classification". Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, 1746–51.

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.

Kira, K., Rendell, L. A. (1992). A practical approach to feature selection. *Machine learning proceedings 1992*, Elsevier. Morgan Kaufmann, 249–256.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM),* 46(5), 604-632.

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.

Lai, T. M., Bui, T., Kim, D. S., and Tran, Q. H. (2020). A Joint Learning Approach based on Self-Distillation for Keyphrase Extraction from Scientific Documents. *arXiv preprint arXiv:2010.11980.*

Lancioni, G., Mohamed, S. S., Portelli, B., Serra, G., and Tasso, C. (2021). *Efficient Keyphrase Generation with GANs*, 1-12.

Li, J., Fan, Q. N., and Zhang, K. (2007). Keyword extraction based on tf/idf for Chinese news document. *Wuhan University Journal of Natural Sciences,* 12(5), 917-921.

Li, J., Huang, G., Chen, J., & Wang, Y. (2019). Dual CNN for relation extraction with knowledge-based attention and word embeddings. Computational intelligence and neuroscience, 1-11.

Li, T. F., Hu, L., Chu, J. F., Li, H. T., and Chi, L. (2019). An Unsupervised Approach for Keyphrase Extraction Using Within-Collection Resources. *IEEE Access,* 7, 126088-126097.

Liang, W., Huang, C., Li, M., and Lu, B. L. (2009, December). *Extracting keyphrases from chinese news articles using Metin Sınıflandırma and query log knowledge.* Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, (2), 733-740.

Litvak, M., Last, M. (2008, August). Graph-based keyword extraction for single-document summarization. In *Coling 2008: Proceedings of the workshop multi-source multilingual information extraction and summarization*, Beer-Sheva, 17-24.

Litvak, M., Last, M., Aizenman, H., Gobits, I., and Kandel, A. (2011). Degext -A language-independent graph-based keyphrase extractor. *Advances in intelligent web mastering–3*, Springer, Berlin, Heidelberg, 121-130.

Litvak, M., Last, M., and Kandel, A. (2013). Degext: a language-independent keyphrase extractor. *Journal of Ambient Intelligence and Humanized Computing,* 4(3), 377-387.

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*. Association for Computational Linguistics, Massachusetts, 366-376.

Lott, B. (2012). Survey of keyword extraction techniques. *UNM Education,* 50(10), 1-5.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development,* 1(4), 309-317.

Mahata, D., Kuriakose, J., Shah, R., and Zimmermann, R. (2018, June). Key2vec: Automatic ranked keyphrase extraction from scientific articles using phrase embeddings. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana, 634-639.

Mahata, D., Shah, R. R., Kuriakose, J., Zimmermann, R., and Talburt, J. R. (2018, April). Theme-weighted ranking of keywords from text documents using phrase embeddings. *2018 IEEE conference on multimedia information processing and retrieval (MIPR),* New Orleans, Louisiana, 184-189.

Mahfuzh, M., Soleman, S., and Purwarianti, A. (2019, September). Improving joint layer rnn based keyphrase extraction by using syntactical features. *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, Yogyakarta, Indonesia, 1-6.

Marujo, L., Gershman, A., Carbonell, J., Frederking, R., and Neto, J. P. (2013). Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. *arXiv preprint arXiv:1306.4886.*

Marcot, B. G., & 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.

Matsuo, Y., Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools,* 13(01), 157-169.

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). Metin Sınıflandırma: Brining order into texts. *Proceedings of EMNLP 2004, Association for Computational Linguistics.* Barcelona, 404-411.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mu, F., Yu, Z., Wang, L., Wang, Y., Yin, Q., Sun, Y., and Zhou, X. (2020). Keyphrase Extraction with Span-based Feature Representations. *arXiv preprint arXiv:2002.05407.*

Myrzakhmetov, B., Kozhirbayev, Z. (2018). Extended language modeling experiments for kazakh. *Академия наук Республики Татарстан*, 42.

Nguyen, T. D., Kan, M. Y. (2007, December). Keyphrase extraction in scientific publications. *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.*

Nugumanova, A., Mansurova, M. (2019). *Tabigi til matinderindegi terminderdi avtomatti turde tanu.* Monografiya, Oskemen, ShQMU.

Onan, A., Korukoğlu, S., and Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications,* 57, 232-247.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web.* Stanford InfoLab.

Papagiannopoulou, E., Tsoumakas, G. (2020). A review of keyphrase extraction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(2), e1339.

Passon, M., Comuzzo, M., Serra, G., and Tasso, C. (2019, January). Keyphrase extraction via an attentive model. *Italian Research Conference on Digital Libraries*, Springer, 304-314.

Prasad, A., 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,* (1), 1837-1846.

Qingguo, Z., Chengzhi, Z. (2008, December). Automatic chinese keyword extraction based on KNN for implicit subject extraction. *2008 International Symposium on Knowledge Acquisition and Modeling*, Wuhan, China, 689-692.

Rabby, G., Azad, S., Mahmud, M., Zamli, K. Z., and Rahman, M. M. (2020). AT-AKÇ: a tree-based unsupervised keyphrase extraction technique. *Cognitive Computation,* 12(4), 811-833.

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). Rutgers University, Piscataway, 29-48.

Raximova, D.R,. Qasimova, D.T, and İsabaeva D.N. (2021). "Qazaq tiline arnalgan BERT modeli negizinde suraq-jauap juyesin zertteu jane azirleu." Abay atındagı QazUPU-nin XABARSHISI. *Fizika-Matematika Gılımdarı» Seriyası,* 4(76), 1-5.

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

Salton, G., Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513-523.

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.

Sarkar, K., Nasipuri, M., and Ghose, S. (2010). A new approach to keyphrase extraction using neural networks. *arXiv preprint arXiv:1004.3274.*

Siddiqi, S., Sharan, A. (2015). Keyword and keyphrase extraction techniques: A literature review. *International Journal of Computer Applications,* 109(2), 18-23.

Spark Jones, K. (1972). A statistical interpretation of term importance in automatic indexing. *Journal of Documentation,* 28(1), 11-21.

Sun, S., Xiong, C., Liu, Z., Liu, Z., and Bao, J. (2020). Joint Keyphrase Chunking and Salience Ranking with BERT. *arXiv preprint arXiv:2004.13639*

Sutskever, I., Martens, J., and Hinton, G. E. (2011). Generating text with recurrent neural networks. *Proceedings of the 28th international conference on machine learning (ICML-11)*, Toronto, Canada, 1017-1024.

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

Uzun, Y. (2005). Keyword extraction using naive bayes. *Bilkent University, Department of Computer Science, Turkey* Web: www. cs. bilkent. edu. tr/~ guvenir/courses/CS550/Workshop/Yasin_Uzun. pdf. Son Erişim Tarihi: 01.03.2023.

Ünlü, Ö., 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)*, Ankara, 1-6.

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 & Management,* 56(6), 102063.

Wan, X., Xiao, J. (2008, July). Single document keyphrase extraction using neighborhood knowledge. *American Association for Artificial Intelligence*. (8), 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., Peng, H., and Hu, J. S. (2006). Automatic keyphrases extraction from document using neural network. *Advances in Machine Learning and Cybernetics: 4th International Conference, ICMLC 2005, Guangzhou, China, August 18-21, 2005, Revised Selected Papers*. Springer Berlin Heidelberg, 633-641.

Wang, J., Peng, H., and Hu, J. S. (2006). Automatic keyphrases extraction from document using neural network. *Advances in Machine Learning and Cybernetics: 4th International Conference, ICMLC 2005, Guangzhou, China, August 18-21, 2005, Revised Selected Papers*. Springer Berlin Heidelberg, 633-641.

Witt, N., Milz, T., and Seifert, C. (2018). Most important first–keyphrase scoring for improved ranking in settings with limited keyphrases. *Discovery Science: 21st International Conference, DS 2018, Limassol, Cyprus, October 29–31, 2018, Proceedings 21*, Springer International Publishing, 373-385.

Yang, S., Zhao, J., Zhang, X., and Zhao, L. (2008). Application of pagerank technique in collaborative learning. *Advances in Blended Learning: Second Workshop on Blended Learning, WBL 2008, Jinhua, China, Augustl 20-22, 2008. Revised Selected Papers*, Springer Berlin Heidelberg, 102-109.

Ye, J., Gui, T., Luo, Y., Xu, Y., and Zhang, Q. (2021). ONE2SET: Generating Diverse Keyphrases as a Set. *arXiv preprint arXiv:2105.11134.*

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 & 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.*

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

Zhang, K., Xu, H., Tang, J., and Li, J. (2006). Keyword extraction using support vector machine. *Advances in Web-Age Information Management: 7th International Conference, WAIM 2006, Hong Kong, China, June 17-19, 2006. Proceedings,* Springer Berlin Heidelberg, 85-96.

Zhang, Q., Wang, Y., Gong, Y., and Huang, X. J. (2016, November). Keyphrase extraction using deep recurrent neural networks on twitter. *Proceedings of the 2016 conference on empirical methods in natural language processing*, Shanghai, China, 836-845.

Zhang, Y., Liu, H., Wang, S., Ip, W. H., Fan, W., and Xiao, C. (2020). Automatic keyphrase extraction using word embeddings. *Soft Computing,* 24, 5593-5608.

Zhao, J., Zhang, Y. (2019, July). Incorporating linguistic constraints into keyphrase generation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 5224-5233.

Zhao, J., Bao, J., Wang, Y., Wu, Y., He, X., and Zhou, B. (2021). SGG: Learning to Select, Guide, and Generate for Keyphrase Generation. *arXiv preprint arXiv:2105.02544.*

Zhou, T., Zhang, Y., and Zhu, H. (2020). Multi-level memory network with crfs for keyphrase extraction. *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I 24*, Springer International Publishing, 726-738.

Gazili olmak ayrıcalıktır